UNIVERSIDADE DA BEIRA INTERIOR
Engenharia

Decision Support Systems for Risk Assessment in Credit Operations Against Collateral

Germanno Gurgel do Amaral Teles

Tese para obtenção do Grau de Doutor em
Engenharia Informática
(3º ciclo de estudos)

Orientador: Prof. Doutor Joel José Puga Coelho Rodrigues

Covilhã, Julho de 2020

Dedication

To my parents, Francy Gurgel and José Teles. To my Family, Whana, Liz and Lara Teles

Acknowledgments

First of all, I would like to thank my wife Whana Teles, for being my refuge in challenging moments and for giving me two beautiful daughters, Liz and Lara and for being my strength to face this long voyage.

I want to thank and express my gratitude to my Professor and Supervisor, Joel José Puga Coelho Rodrigues, that carefully advised me in the course of these last four years. His patience, generosity, and knowledge were crucial to the development of this work.

I would also like to thank my parents, Francy Gurgel and José Teles, to my brother Guilherme Teles, my systers Milena Teles, and Mirella Battista, they give me good energies, even being so far away, are always helpful and never hesitated.

To the Brazilian National Council for Scientific and Technological Development (CNPq) and the Science Without Borders program, through the grant number 200450/2015-8, for all the support and trust that was given to me and my doctorate project. This work is partially funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020.

To all my colleagues of the Next Generation Networks and Applications (NetGNA) research group, for the shared moments of happiness and knowledge, especially to Professor Mario Moreira, Professor Gilberto Junior, Professora Simone Ferreira, Professor José Victor Sobral, for all support, advices, guidance, and friendship.

I would also like to thank all my colleagues of Banco do Nordeste, for their support and constant words of encouragement during the last four years.

To all the Professors and employees from the University of Beira Interior (UBI), and the Instituto de Telecomunicações (IT), Covilhã delegation.

All Brazilian friends that live in Covilhã. They have been my family.

Finally, to all people that, in some way, have led me to reach this extraordinary achievement in my life.

Foreword

This thesis describes the research work performed in the scope of the doctoral research programme and presents its main contributions and achievements. This doctoral programme and inherent research activities were carried out at the Next Generation Networks and Applications Group (NetGNA) research group of the IT department, University of Beira Interior, Covilhã, Portugal and Instituto de Telecomunicações, delegation of Covilhã, Portugal. This work is partially funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020. The research work was supervised by Prof. Dr. Joel José Puga Coelho Rodrigues and also financially supported by the National Council for Scientific and Technological Development (CNPq) through the grant contract 200450/2015-8.

List of Publications

Papers included in the thesis resulting from this 4-year doctoral research programme

1. **Classification Methods Applied to Credit Scoring with Collateral**

   Germanno Teles, Joel J. P. C. Rodrigues, Kashif Saleem and Sergei Kozlov

   IEEE Systems Journal, IEEE, ISSN: 1932-8184, September 2019.

   DOI: doi.org/10.1109/JSYST.2019.2937552

2. **Machine Learning and Decision Support System on Credit Scoring**

   Germanno Teles, Joel J. P. C. Rodrigues, Kashif Saleem,Sergei Kozlov and Ricardo A. L. Rabêlo

   Neural Computing and Applications, Springer, ISSN:1433-3058, vol.32, n.14, pp.9809–9826, October 2019.

   DOI: doi.org/10.1007/s00521-019-04537-7

3. **Decision Support System on Credit Operation Using Linear and Logistic Regression**

   Germanno Teles, Joel J. P. C. Rodrigues, Sergei Kozlov, Ricardo A. L. Rabêlo and Victor Hugo C. Albuquerque

   Expert Systems, Wiley, ISSN:1468-0394, pp. e12578, May 2020.

   DOI: doi.org/10.1111/exsy.12578

4. **Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction**

   Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabêlo and Sergei Kozlov

   Journal of Artificial Intelligence and Systems, Institute of Electronics and Computer, ISSN: 2642-2859, vol.2, pp. 118-132, March 2020.

   DOI: doi.org/10.33969/ais.2020.21008

5. **Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation**

   Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabêlo and Sergei Kozlov

   Software: Practice and Experience, Wiley, ISSN:1097-024X, pp. 1-9, May 2020.
   DOI: doi.org/10.1002/spe.2842

Presented posters resulting from this doctoral research programme included in the thesis as appendixes

6. **Decision Support Systems to Predict a Sufficiency of Collateral for Credit Risk Operations**

   Germanno Teles, Joel J. P. C. Rodrigues

   Poster Ciência 2018 - Science and Technology in Portugal Summit.

7. **Intelligent Decision Support System on Credit Scoring**

   Germanno Teles, Joel J. P. C. Rodrigues

   Poster IEEE UBI Student Branch.

8. **Using Natural Language Processing with Sentiment Analysis to Improve Credit Score on Collateral Reports**

   Germanno Teles, Joel J. P. C. Rodrigues

   Poster LxMLS 2019, 9th Lisbon Machine Learning School

Resumo

Com a crise econômica global, que atingiu seu auge no segundo semestre de 2008, e diante de um mercado abalado pela instabilidade econômica, as instituições financeiras tomaram medidas para proteger os riscos de inadimplência dos bancos, medidas que impactavam diretamente na forma de análise nas instituições de crédito para pessoas físicas e jurídicas. Para mitigar o risco dos bancos nas operações de crédito, a maioria destas instituições utiliza uma escala graduada de risco do cliente, que determina a provisão que os bancos devem fazer de acordo com os níveis de risco padrão em cada transação de crédito. A análise de crédito envolve a capacidade de tomar uma decisão de crédito dentro de um cenário de incerteza e mudanças constantes e transformações incompletas. Essa aptidão depende da capacidade de analisar situações lógicas, geralmente complexas e de chegar a uma conclusão clara, prática e praticável de implementar.

Os modelos de *Credit Score* são usados para prever a probabilidade de um cliente propor crédito e tornar-se inadimplente a qualquer momento, com base em suas informações pessoais e financeiras que podem influenciar a capacidade do cliente de pagar a dívida. Essa probabilidade estimada, denominada pontuação, é uma estimativa do risco de inadimplência de um cliente em um determinado período. A mudança constante afeta várias seções bancárias, pois impede a capacidade de investigar os dados que são produzidos e armazenados em computadores que frequentemente dependem de técnicas manuais.

Entre as inúmeras alternativas utilizadas no mundo para equilibrar esse risco, destaca-se o aporte de garantias na formalização dos contratos de crédito. Em tese, a garantia não "garante" o retorno do crédito, já que não é computada como pagamento da obrigação dentro do projeto. Tem-se ainda, o fato de que esta só terá algum êxito se acionada, o que envolve a área jurídica da instituição bancária. A verdade é que, a garantia é um elemento mitigador do risco de crédito. As garantias são divididas em dois tipos, uma garantia individual (patrocinadora) e a garantia do ativo (fiduciário). Ambos visam aumentar a segurança nas operações de crédito, como uma alternativa de pagamento ao titular do crédito fornecido ao credor, se possível, não puder cumprir suas obrigações no prazo. Para o credor, gera segurança de liquidez a partir da operação de recebimento. A mensuração da recuperabilidade do crédito é uma sistemática que avalia a eficiência do mecanismo de retorno do capital investido em garantias.

Para tentar identificar a suficiência das garantias nas operações de crédito, esta tese apresenta uma avaliação dos classificadores inteligentes que utiliza informações contextuais para avaliar se as garantias permitem prever a recuperação de crédito concedido no processo de tomada de decisão antes que a operação de crédito entre em *default*. Os resultados observados quando comparados com outras abordagens existentes na literatura e a análise comparativa das soluções de inteligência artificial mais relevantes, mostram que os classificadores que usam garantias como parâmetro para calcular o risco contribuem para o avanço do estado da arte, aumentando o comprometimento com as instituições financeiras.

Palavras-chave

Sistemas Inteligentes de Apoio à Decisão, Redes Neurais, *Support Vector Machine*, *Fuzzy*, *Big Data*, Finanças, Operação de Crédito, Aprendizado de Máquina, Árvore de decisão, *Credit Risk*, *Credit Score*, Regressão Linear e Regressão Logística.

Extended Abstract in Portuguese

Introdução

Esta seção resume os 4 anos de trabalho de investigação no âmbito da tese de doutoramento intitulada "*Decision Support System for Risk Assessment in Credit Operations Against Collateral*". Esta tese foca-se no estudo e proposta de estratégias e metodologias de análise de dados para identificar a suficiência das garantias nas operações de crédito, esta tese apresenta uma avaliação dos classificadores inteligentes que utiliza informações contextuais para avaliar se as garantias permitem prever a recuperação de crédito concedido no processo de tomada de decisão antes que a operação de crédito entre em *default*. Na primeira etapa é descrito o enquadramento da tese, definido o problema abordado e os principais objetivos do estudo. Em seguida, a hipótese de investigação é descrita e são apresentadas as principais contribuições deste trabalho para o avanço do estado da arte.

Enquadramento do Tema

A implementação do Novo Acordo de Basiléia (Basiléia III) traz como desafio a estimativa de parâmetros críticos para a modelagem do risco de crédito, como a Perda por Inadimplência; a probabilidade de inadimplência e a exposição dado o incumprimento. A pesquisa avançou e os aspectos fundamentais para a implementação dos parâmetros já estão equacionados. A perda por inadiplência exigida pela Basiléia III tem sido objeto de intenso debate do setor financeiro no Brasil e no exterior.

Nos esforços para reduzir a complexidade dos padrões internacionais da contabilidade existente que trata de instrumentos financeiros, especialmente as Normas Internacionais de Contabilidade, e em resposta à crise financeira de 2008, o Conselho Internacional de Normas Contábeis, juntamente com a Contabilidade Financeira Standards Board, estão revisando esses padrões. As Normas Internacionais de Contabilidade estabelecem procedimentos para transações contábeis e de divulgação envolvendo instrumentos financeiros. A norma também contém definições relacionadas a esses instrumentos e determina procedimentos contábeis específicos para o reconhecimento inicial, avaliação baixa e subsequente desses itens.

Gitman [1] se pergunta sobre a atividade de verificação de crédito de uma empresa e procura determinar se deve ser concedido crédito a um cliente e quais limites quantitativos devem ser impostos. A classificação de risco no contexto bancário pode ser vista sob três aspectos: primeiro, o risco do cliente que indica a capacidade atual de incorrer em uma dívida do cliente. Gitman [1] usa análises de crédito sobre o risco do cliente para o modelo dos cinco C's. O risco da proposta é o segundo aspecto, que avalia o objetivo, a finalidade, o valor e o prazo do crédito e sua adequação e ponderação das garantias. E o terceiro ponto, que indica a qualidade (suficiência e liquidez) que as garantias têm para mitigar o efeito.

A orientação da análise de risco do cliente possui cinco dimensões principais da capacidade creditícia do cliente. São eles: Caráter, referente ao histórico de conformidade do requerente com suas obrigações financeiras e contratuais; Capacidade, referente ao potencial do requerente para pagar o crédito solicitado; Capital, referente à saúde financeira do requerente; Garantia, referente à quantidade de bens disponíveis pelo requerente para garantir crédito; Condições, relativas às condições econômicas e da indústria existentes, bem como

elementos especiais que podem afetar o solicitante e o credor.

A mensuração da recuperabilidade é uma tecnologia que avalia a eficiência do mecanismo de retorno do capital investido em bens, combinando o tempo necessário para recuperar uma propriedade em um determinado sistema "fluxos de caixa a valor presente", em que todos os custos retornam efetivamente ao caixa durante toda a vida útil do ativo. Na prática, se não houver recuperabilidade, isso poderá gerar uma diminuição nos ativos por perda. O rastreamento do relacionamento das empresas com o mercado bancário ocorre a partir de setores que têm acesso ao banco de dados das instituições, que armazena e fornece carteira de crédito mensal de diversas empresas e pessoas físicas encaminhadas aos bancos. Dessa forma, o controle das instituições bancárias sabe o que pode ou não conceder com mais precisão, reduzindo perdas futuras de empresas insolventes

Os modelos de pontuação de crédito são usados para prever a probabilidade de um cliente propor crédito tornar-se inadimplente a qualquer momento, com base em suas informações pessoais e financeiras que podem influenciar a capacidade do cliente de pagar a dívida. Essa probabilidade estimada, denominada pontuação, é uma estimativa do risco de inadimplência de um cliente em um determinado período. Essa crescente preocupação não foi causada em grande parte pelas fraquezas das técnicas existentes de gerenciamento de riscos que se revelaram pela recente crise financeira e pela crescente demanda por crédito ao consumidor. A mudança constante afeta várias seções bancárias porque impede a capacidade de investigar os dados que são produzidos e armazenados em computadores que frequentemente dependem de técnicas manuais [2].

O teste de redução ao valor recuperável é usado para demonstrar e medir a perda de recuperabilidade do valor contábil de um ativo de longa duração. Uma perda por redução ao valor recuperável ocorre quando o valor contábil excede o valor recuperável de um ativo ou grupo de ativos, a longo prazo, incluindo o valor do dinheiro ao longo do tempo.

As informações existentes sobre credit scoring são principalmente de um estudo sobre a evolução dos indicadores financeiros para algumas empresas, que não tiveram sucesso ou continuaram suas atividades durante o período avaliado. A falha ou o sucesso da estrutura de gerenciamento é avaliado por um indicador conhecido como pontuação de corte, que é definido por uma combinação linear de indicadores financeiros.

Os modelos de pontuação exemplificam uma maneira de reconhecer, quantificar e controlar o risco corporativo de falência [3]. Seu caráter multidimensional pesquisa um diagnóstico financeiro da entidade permite a avaliação do risco com mais facilidade e de forma correta.

Delimitação do Problema

A Resolução Brasileira Nº. 3721, de 30/04/2009, pelo Conselho Monetário Nacional, instituído no art. 4, item XI, alínea c, "avaliação periódica da adequação das garantias" ou seja, a estrutura de gerenciamento de risco de crédito deve prever o estabelecimento de critérios e procedimentos claramente definidos e documentados, acessíveis aos envolvidos na gestão de concessão e crédito processo para gerenciar a suficiência das garantias em períodos subsequentes. No sistema bancário, os modelos predominantes de pontuação de crédito são desenvolvidos a partir de janelas estáticas e mantidos inalterados por anos. Nesse cenário, os dois mecanismos básicos de memória, memória de curto e longo prazo, são fundamentais para o aprendizado, mesmo se você estiver usando classificadores de modelos de decisão, se este não for o modelo certo.

A essência do negócio bancário, na tomada de decisões de crédito, são os modelos de

Probabilidade de Inadimplência (PD), para determinar o custo de capital e o contrato de preço. Além disso, os bancos centrais e a regulamentação internacional evoluíram drasticamente para um cenário em que o uso desses modelos favorece a obtenção de padrões de firmeza para a avaliação do risco de crédito no sistema bancário. No setor bancário, a avaliação do risco de crédito geralmente depende de modelos de pontuação de crédito, modelos de PD.

As garantias são divididas em dois tipos, uma garantia individual (patrocinadora) e a garantia do ativo (fiduciário). Ambos visam aumentar a segurança nas operações de crédito, como uma alternativa de pagamento ao titular do crédito fornecido ao seu credor, se possivelmente não puder cumprir suas obrigações no prazo. Para o credor, gera segurança de liquidez a partir da operação de recebimento. A mensuração da recuperabilidade do crédito é uma sistemática que avalia a eficiência do mecanismo de retorno do capital investido em garantias.

As garantias fiduciárias geram um alto custo operacional e são difíceis de seguir "in loco" para os procedimentos de reavaliação de um ativo, por isso precisam de atenção especial, dado o grande volume de mercadorias e a insuficiente capacidade técnica e operacional. De acordo com a resolução, não há obrigação ou abordagem de investigação que use características de garantia como variável em um sistema para apoiar a tomada de decisão, mesmo produzindo um alto custo em operações de crédito. Com base nessa dificuldade, é necessária uma avaliação conceitual de um modelo inteligente de verificação de recuperabilidade da concessão de crédito contra garantias.

O trabalho tem como objetivo demonstrar a importância das variáveis qualitativas e quantitativas do processo de concessão de crédito de bancos e assim propor uma metodologia alternativa para instituições financeiras, baseadas na suficiência e principalmente sua liquidez. Descobrir a expectativa de recuperação da garantia real, utilizando classificadores inteligentes. Identificar padrões de maturidade para atualizar os dados de bens no Garantia das instituições financeiras. Mapear todas as variáveis das garantias. Criar um mecanismo para gerenciamento de recuperação em Sistemas de Crédito Pontuação utilizando parâmetros de maturidade dos processos de suporte e apoio ao classificador aplicados a machine learning.

Objetivos de Investigação

O principal objetivo desta tese é a proposta e a avaliação de desempenho de um modelo inteligente que utiliza informações contextuais para avaliar se as garantias permitem prever a recuperação de crédito concedido no processo de tomada de decisão antes da operação de crédito entrar em default.
Para alcançar este objetivo principal, foram definidos os seguintes objetivos parciais:

- Revisão do estado da arte em pontuação de crédito, tecnologias e abordagens inteligentes de sistemas de suporte à decisão, que utilizam técnicas de mineração de dados e aprendizado de máquina;
- Avaliação de desempenho dos sistemas de pontuação de crédito existentes para escolher o melhor que será usado como referência para avaliar e validar as contribuições propostas;
- Proposta, projeto e construção de um novo sistema inteligente de suporte à decisão para pontuação de crédito;
- A avaliação de desempenho das abordagens propostas através de experimentos reais envolvendo operações de crédito contra garantias;

- Proposta de um modelo de pontuação de crédito inteligente para estimar a recuperabilidade de uma operação de crédito com garantia como ativo.

Principais Contribuições

A primeira contribuição desta tese é uma revisão do estado da arte na identificação dos classificadores mais utilizados para análise de risco de crédito e seu desempenho nos sistemas de suporte à decisão. Inicialmente, o trabalho apresenta uma visão geral do processo que observa a avaliação do conteúdo e avalia sistematicamente a comunicação gravada, sendo apresentada em dois grupos de critérios, primeiro uma análise exploratória para detectar os classificadores e uma revisão sistemática dos trabalhos que utilizam o crédito. Modelos de pontuação para estimar a recuperabilidade de uma operação de crédito com garantia como ativo. Isso foi feito para preparar os classificadores e fornecer consistência à análise com ou sem garantias [4]. Esta pesquisa foi publicada em textit IEEE Systems Journal [5].

A segunda contribuição é comparar o desempenho da pontuação de crédito de conjuntos nebulosos e árvores de decisão com base em uma rede neural artificial para prever o valor recuperado usando uma amostra de 1890 tomadores de empréstimos. que usam operações de crédito. Esta contribuição foi aceita no *Neural Computing and Applications*, da Springer.

A terceira contribuição foi enviada em uma edição especial intitulada “Future Hybrid Artificial Intelligence and Machine Learning for Smart Expert Systems” para a revista *Expert Systems* (Wiley). Este trabalho tem como objetivo entender como os modelos preditivos podem fornecer diferentes estimativas de recuperação esperada com base nos mesmos conjuntos de dados. Ele compara a eficiência da regressão logística com a de uma regressão linear na previsão de se a recuperação é devida em uma operação de crédito.

A quarta contribuição desta tese propõe determinar a melhor combinação de parâmetros a serem usados com RNAs e a abordagem do BN para lidar e avaliar com precisão o risco de crédito. A principal contribuição deste estudo é o modelo de aprendizado de máquina proposto ou a combinação deles, que é uma abordagem rara ao problema de mensuração do risco de crédito. O estudo destaca a lacuna existente que impede os sistemas de inteligência abordarem questões de modelagem bancária. Esta contribuição foi submetida a uma revista internacional.

Finalmente, a última contribuição introduz uma avaliação da estabilidade dos dois modelos com base nas variáveis escolhidas. O método usado pelos bancos na tomada de decisões sobre empréstimos não é claro; no entanto, a implementação de modelos lineares clássicos em sistemas bancários está adequadamente documentada. Para este artigo, a abordagem elástica transparente foi usada como referência. Esta pesquisa mostra vantagens da abordagem de RF sobre o algoritmo SVM: velocidade e simplicidade operacional, e o SVM possui o benefício de uma maior precisão de classificação que a RF. Esta contribuição foi submetida a uma revista internacional.

Principais Conclusões

Esta tese abordou a condição real e essencial para os sistemas de apoio à decisão, desde o Acordo de Basiléia até os dias atuais. Este trabalho deixa claro a importância do sistema DSS e a ”inteligência” envolvida neles. Uma evolução e a importância do DSS nas instituições bancárias de variáveis qualitativas e quantitativas de um banco para concessão de processos de crédito, descrevendo os muitos classificadores que podemos usar. Portanto, a abordagem proposta fornece uma maneira e uma análise para avaliar o risco de crédito.

O capítulo 2 apresentou o trabalho de pesquisa intitulado “Classification Methods Applied to Credit Scoring with Collateral”. O artigo fornece uma revisão aprofundada do estado da arte da análise e inclui 84 estudos neste trabalho para propor uma metodologia estatística de uso para realizar uma meta-análise a fim de comparar os resultados dos métodos de classificação. O resultado mostra que o SVM é o classificador mais usado para as pontuações de crédito e, embora o sistema tenha um bom desempenho, ele não aplica abordagens com garantias. Este artigo apresenta uma revisão e avaliação detalhadas dos métodos classificadores de pontuação de crédito existentes. É realizada uma extensa revisão da literatura com foco nos métodos de classificação aplicados na pontuação de crédito. Ele se concentra principalmente em métodos para classificar um solicitante de operações de crédito com a suficiência de garantias, mas também consideraremos brevemente outros problemas associados no setor de crédito ao seu provável comportamento de pagamento (por exemplo, ’inadimplência’ ou ’não inadimplente’ com reembolsos) . A análise deste trabalho propõe uma metodologia estatística de uso para realizar uma meta-análise para comparar os resultados dos métodos de classificação. Ele mostra alguns casos que consideram várias distribuições de probabilidade e também dados de sobrevivência. Também elabora que a garantia não é a primeira abordagem para a pontuação de crédito. Existe uma estatística satisfatória disponível para a garantia, a distribuição posterior de probabilidade depende dos dados apenas por meio dessa estatística e, portanto, em muitos casos, podemos reduzir nossos dados sem perda de informações. O resultado geral mostra que a garantia não é a primeira abordagem para a pontuação de crédito, mas quando usada pode ser um valor alto nos métodos dos modelos de classificação.

O capítulo 3, intitulado “Machine Learning and Decision Support System on Credit Scoring”, está comparando o desempenho da pontuação de crédito de conjuntos nebulosos e árvores de decisão com base em uma rede neural artificial para prever o valor recuperado. Este artigo é um estudo inicial de garantias como uma variável no cálculo da pontuação de crédito. A lógica difusa faz algumas suposições implícitas que podem dificultar ainda mais o processo de tomada de decisão pelos concedentes de crédito. O estudo conclui que ambos os modelos permitem modelar a incerteza no processo de pontuação de crédito. Portanto, são necessários mecanismos que ajudem a caracterizar cenários tão complexos. A literatura sugere que a aplicação de múltiplos critérios para tomada de decisão pode facilitar a resolução desses problemas. Outros modelos, como lógica fuzzy, redes neurais artificiais e modelos de árvore de decisão, explicitamente consideram os relacionamentos subjacentes e reconhecem as incertezas (como riscos operacionais). Além dos métodos com vários critérios, ferramentas complementares adicionais, como conjuntos nebulosos ou simulações numéricas, estão sendo cada vez mais usadas no processo de pontuação de crédito. Embora a lógica difusa seja mais difícil de implementar, ela modela com mais precisão a incerteza.

No capítulo 4, intitulado “Support System on Credit Operation Using Linear and Logistic Regression”, o objetivo é entender como os modelos preditivos podem fornecer estimativas diferentes da recuperação esperada com base nos mesmos conjuntos de dados. A análise preditiva, que é o método de obter conhecimento dos conjuntos de dados existentes para decidir guias e prever resultados e tendências futuras, inclui técnicas de classificação e técnicas de regressão. Técnicas de classificação como análise de árvore de decisão, análise estatística, redes neurais, máquinas de vetores de suporte, raciocínio baseado em casos, classificadores bayesianos, algoritmos genéticos e conjuntos aproximados ajudam a identificar padrões em grandes conjuntos de dados não estruturados e a gerar conjuntos de agrupamentos. As técnicas de regressão incluem regressão linear e regressão logística. Um modelo simples de regressão logística pode ser facilmente estendido a um modelo de regressão logística múltipla, integrando mais de uma variável de previsão, o que indica uma dificuldade crescente na obtenção de várias observações com um número crescente de variáveis independentes.

O capítulo 5, intitulado “Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction”, compara redes bayesianas com redes neurais artificiais para prever o valor recuperado em uma operação de crédito. O estudo explora esse problema e descobre que as RNAs são uma ferramenta mais eficiente para prever o risco de crédito do que a abordagem ingênua Bayesiana (NB). As RNAs foram usadas para estudar o sistema nervoso e a maneira como o cérebro processa as informações. Uma RNA requer um algoritmo de processamento para modelar o cérebro de humanos e compreende um grande número de nós interconectados (neurônios) trabalhando como um sistema para resolver problemas de reconhecimento de padrões ou classificação de dados. O presente estudo demonstra que vários algoritmos podem ser usados em paralelo para resolver o problema em questão, que no caso apresentado aqui é a concessão de empréstimos. Várias estratégias para identificar a escolha de recursos (ou variáveis), algoritmo e critérios podem fornecer uma solução. Por exemplo, no novo Big Data e na era digital, a transparência é crítica. Estratégias baseadas em aprendizado profundo também são necessárias para treinar dados sobre o aplicativo, e algoritmos de aprendizado de máquina e seu uso devem ser regulados para garantir a precisão. A abordagem comparativa foi usada para gerar os melhores resultados das iterações. Os resultados mostram que os modelos de RNA e RN fornecem resultados confiáveis, mas o modelo de RNA é mais valioso para prever o risco de crédito.

No Capítulo 6,“Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation”, mostrou vantagens da abordagem de RF em relação ao algoritmo SVM: velocidade e simplicidade operacional, e o SVM tem o benefício de maior precisão de classificação que o RF. O objetivo dessa abordagem é implementar a técnica de aprendizado de Máquina de Vetor de Suporte para prever a probabilidade de inadimplência. Os dados precisavam de treinamento, cada um consistindo em valores para o conjunto de variáveis de entrada e saída. As variáveis foram escolhidas e apenas aquelas cujo comportamento era previsível foram incluídas. Para o atual estudo comparativo, as variáveis de dados foram consideradas os principais indicadores de risco e os empréstimos foram considerados os sujeitos. Os dados coletados para este estudo eram de um banco e consistiam apenas em empréstimos de curto prazo, uma vez que representam a parcela mais significativa dos empréstimos. Um conjunto de dados de 1890 arquivos de crédito foi obtido e os sujeitos foram classificados como clientes menos arriscados ou arriscados. A variável mais constante é a probabilidade de inadimplência juntamente com uma variável dummy, Y é igual a zero para um cliente menos arriscado e um para clientes arriscados. Isso significa que Y = 1 quando o pagamento está atrasado e Y =

0 quando o pagamento é efetuado em tempo útil. As classificações SVM foram implementadas para prever os potenciais membros da classe. O objetivo principal deste estudo foi comparar as técnicas de aprendizado de máquina SVM e RF com base no desempenho em instituições de crédito e mais precisamente na determinação do valor de recuperação. O conceito de risco de crédito é caracterizado por uma imprecisão que complica a formulação de uma definição limitada para a identificação de fatores de risco, e formas funcionais adequadas para aproximar e prever seu valor não é tarefa fácil. De maneira semelhante, o alcance e a complexidade do conceito de risco de crédito tornam obsoletos os modelos matemáticos tradicionais. Este estudo comparou o desempenho de duas abordagens, a SVM e a RF, abordando o problema da avaliação de risco de crédito. No estudo, as variáveis foram escolhidas em relação aos dados coletados dos registros dos bancos. Apesar das inúmeras capacidades de máquinas de vetores de suporte e algoritmos de previsão de florestas aleatórias, a preocupação com a estimativa de risco de crédito mal foi abordada com relação aos algoritmos de aprendizado de máquina e muito menos a uma combinação deles. O estudo atual, portanto, preenche lacunas existentes que permeiam sistemas inteligentes de questões extremamente intrigantes de modelagem de bancos. O foco principal foi a ideia de insolvência como uma técnica de caracterização do risco de crédito. O artigo compara os algoritmos SVM e RF para prever o valor recuperado em uma tarefa de crédito. A execução dos sistemas inteligentes projetados utiliza testes e algoritmos para autenticação do modelo projetado.

O trabalho teve como objetivo demonstrar a importância de variáveis qualitativas e quantitativas no processo de concessão de crédito aos bancos e, assim, propor uma metodologia alternativa para as instituições financeiras, baseada na suficiência e, principalmente, na sua liquidez. Descubrir a expectativa de recuperação da garantia real, usando classificadores inteligentes e identificar padrões de maturidade para atualizar dados de ativos na Garantia das instituições financeiras, assim como o mapeamento todas as variáveis de garantia e criar um mecanismo para gerenciar a recuperação em sistemas de pontuação de crédito usando parâmetros de maturidade dos processos de suporte e classificador aplicados ao aprendizado de máquina.

O principal objetivo desta tese foi criar um modelo que utilize informações contextuais para avaliar se as garantias permitem a recuperação de crédito concedido no processo de tomada de decisão que pode ajudar os tomadores de decisão na operação de crédito. A perda por redução ao valor recuperável está relacionada à desvalorização de ativos ou valores mobiliários devido à falta de compradores ou excesso de oferta, como pode ocorrer no setor imobiliário. Em certas áreas, os preços de apartamentos novos podem não estar “a par” dos preços anteriores. Assim, diremos que os novos preços estão em uma situação de impairment. O Modelo de Avaliação de Risco de Clientes adotado pelos Bancos tem como função classificar os clientes de acordo com o nível de risco, proporcionando maior segurança às decisões de crédito e outros serviços financeiros e bancários, sem se tornar um elemento que inibe o poder competitivo da instituição no mercado. Para tanto, a pesquisa se baseia em estudos anteriores, buscando detectar uma abordagem viável. Os sistemas investigados em cada subtópico nos apresentam técnicas de mineração de dados como uma solução necessária e mostram a necessidade de encontrar o melhor classificador que será usado para atender aos objetivos desta pesquisa.

Perspectivas de Trabalhos Futuros

Para concluir esta tese, são sugeridas as seguintes direções de investigações futuras que resultaram do trabalho desenvolvido:

- Realizar uma nova pesquisa literária para analisar outra maneira de medir uma meta-análise, como taxa de risco ou diferença de risco; além disso, pesquisas de alta qualidade são publicadas em periódicos científicos; outras formas de publicação podem ser incluídas nesta lista em futuras investigações. Não obstante essas limitações, nossa revisão sistemática fornece informações importantes sobre a literatura de pesquisa sobre técnicas de classificação aplicadas à pontuação de crédito e como essa área vem se movendo ao longo do tempo;

- A força particular das árvores de decisão baseadas em redes neurais artificiais é sua tendência a ajudar a compreender decisões seqüenciais e dependências de resultados. O modelo pode desempenhar um papel complementar a outras ferramentas de pontuação, como ativos *fuzzy*, em que as classes que ele cria podem ser usadas como *fuzzy sets*. No entanto, um algoritmo de árvore de decisão requer que o atributo de destino tenha apenas valores discretos. Outra desvantagem é que ele apresenta um desempenho ruim em termos de interações complexas nas quais as árvores de decisão são redesenhadas toda vez que novos dados são adicionados ao modelo. Além disso, as árvores de decisão são sensíveis ao conjunto de treinamento, atributos irrelevantes e ruído. Criar um modelo usando *fuzzy* pode ser integrado, por exemplo, redes neurais, levando a maior precisão de previsão;

- Como vários fatores podem afetar a acessibilidade e o endividamento excessivo, é um desafio fazer previsões para o futuro. A avaliação de acessibilidade geralmente é baseada em dados de aplicativos, relatórios de crédito e estimativa de gastos. Pouca informação sobre os modelos implementados de acessibilidade está disponível em domínio público, exceto pelas soluções oferecidas pelas agências de crédito. Há ainda menos informações sobre modelos para operações de crédito. A literatura existente sobre modelos de acessibilidade e superendividamento também é escassa. No entanto, pode ser preferível uma abordagem dinâmica para a avaliação da acessibilidade, que leve em consideração possíveis mudanças nas receitas e despesas e permita prever o futuro. Crie uma validação abrangente dos métodos sugeridos com outro banco de dados de pontuação de crédito e seja capaz de fazer comparações;

- O método proposto é útil para identificar uma nova maneira que tenha uma forte possibilidade de ser mais avançada do que os classificadores anteriores semelhantes. Segundo, no método proposto, propriedades e funções foram categorizadas manualmente em grupos representativos, mas um tópico futuro automatizará essa tarefa usando tecnologias semânticas. Aplique o conceito de processamento de linguagem natural (NLP) para analisar o sentimento adicionado à pontuação de crédito.

Referências

[1] L. J. Gitman *et al.*, *Princípios de administração financeira*. Harbra São Paulo, 1997.

[2] T. Harris, "Credit scoring using the clustered support vector machine," *Expert Systems with Applications*, vol. 42, no. 2, pp. 741–750, 2015.

[3] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking state-of-the-art classification algorithms for credit scoring," *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.

[4] R. H. Kolbe and M. S. Burnett, "Content-analysis research: An examination of applications with directives for improving research reliability and objectivity," *Journal of consumer research*, vol. 18, no. 2, pp. 243–250, 1991.

[5] G. Teles, J. J. Rodrigues, K. Saleem, and S. A. Kozlov, "Classification methods applied to credit scoring with collateral," *IEEE Systems Journal*, 2019.

[6] G. Teles, J. J. Rodrigues, K. Saleem, S. Kozlov, and R. A. Rabêlo, "Machine learning and decision support system on credit scoring," *Neural Computing and Applications*, vol. 32, no. 14, pp. 9809–9826, 2020.

[7] G. Teles, J. J. Rodrigues, S. A. Kozlov, R. A. Rabêlo, and V. H. C. Albuquerque, "Decision support system on credit operation using linear and logistic regression," *Expert Systems*, p. e12578.

[8] G. Teles, J. J. Rodrigues, R. A. Rabêlo, and S. A. Kozlov, "Artificial neural network and bayesian network models for credit risk prediction," *Journal of Artificial Intelligence and Systems*, vol. 2, pp. 118–132, 2020.

[9] G. Teles, J. J. Rodrigues, R. A. Rabêlo, and S. A. Kozlov, "Comparative study of support vector machines and random forests machine learning algorithms on credit operation," *Software: Practice and Experience*, 2020.

Abstract

With the global economic crisis, which reached its peak in the second half of 2008, and before a market shaken by economic instability, financial institutions have taken steps to protect the banks' default risks, which had an impact directly in the form of analysis in credit institutions to individuals and to corporate entities. To mitigate the risk of banks in credit operations, most banks use a graded scale of customer risk, which determines the provision that banks must do according to the default risk levels in each credit transaction. The credit analysis involves the ability to make a credit decision inside a scenario of uncertainty and constant changes and incomplete transformations. This ability depends on the capacity to logically analyze situations, often complex and reach a clear conclusion, practical and practicable to implement.

Credit Scoring models are used to predict the probability of a customer proposing to credit to become in default at any given time, based on his personal and financial information that may influence the ability of the client to pay the debt. This estimated probability, called the score, is an estimate of the risk of default of a customer in a given period. This increased concern has been in no small part caused by the weaknesses of existing risk management techniques that have been revealed by the recent financial crisis and the growing demand for consumer credit.The constant change affects several banking sections because it prevents the ability to investigate the data that is produced and stored in computers that are too often dependent on manual techniques.

Among the many alternatives used in the world to balance this risk, the provision of guarantees stands out of guarantees in the formalization of credit agreements. In theory, the collateral does not ensure the credit return, as it is not computed as payment of the obligation within the project. There is also the fact that it will only be successful if triggered, which involves the legal area of the banking institution. The truth is, collateral is a mitigating element of credit risk. Collaterals are divided into two types, an individual guarantee (sponsor) and the asset guarantee (fiduciary). Both aim to increase security in credit operations, as an payment alternative to the holder of credit provided to the lender, if possible, unable to meet its obligations on time. For the creditor, it generates liquidity security from the receiving operation. The measurement of credit recoverability is a system that evaluates the efficiency of the collateral invested return mechanism.

In an attempt to identify the sufficiency of collateral in credit operations, this thesis presents an assessment of smart classifiers that uses contextual information to assess whether collaterals provide for the recovery of credit granted in the decision-making process before the credit transaction become insolvent. The results observed when compared with other approaches in the literature and the comparative analysis of the most relevant artificial intelligence solutions, considering the classifiers that use guarantees as a parameter to calculate the risk contribute to the advance of the state of the art advance, increasing the commitment to the financial institutions.

Keywords

Smart Decision Support Systems, Neural Networks, Support Vector Machine, Fuzzy, Big Data, Finance, Credit Operation, Machine Learning, Decision Tree, Credit Risk, Credit Score, Linear Regression and Logistic Regression.

Contents

List of Figures

Chapter 3

Machine Learning and Decision Support System on Credit Scoring

Chapter 4

Decision Support System on Credit Operation Using Linear and Logistic Regression

Chapter 5

Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction

Chapter 6

Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation

List of Tables

Chapter 2

Classification Methods Applied to Credit Scoring With Collateral

Chapter 3

Machine Learning and Decision Support System on Credit Scoring

Chapter 4

Decision Support System on Credit Operation Using Linear and Logistic Regression

Chapter 5

Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction

Chapter 6

Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation

Acronyms

AI	Artificial Intelligence
ANFIS	Adaptive Network-based Inference System
ANN	Artificial Neural Network
API	Application Programming Interface
AUC	Area Under Curve
BD	Big Data
BI	Business Intelligence
BN	Bayesian Network
CI	Computational Intelligence
CNPq	National Council for Scientific and Technological Development
CR	Credit Scoring
CRC	Credit Score Classifier
DBMS	Database Management System
DM	Data Mining
DSS	Decision Support System
DT	Decision Tree
FCT	Foundation for Science and Technology
FL	Fuzzy Logic
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GA	Genetic Algorithm
FSBRC	Fuzzy System-Based Route Classifier
FSELC	Fully Simplified Exponential Lifetime Cost
FUZZY OF	Fuzzy-based Objective Function
GDPR	General Data Protection
GPS	Global Positioning System
ICT	Information and Communication Technologies
ID	Identifier
KES	Knowledge-based Expert Systems
LR	Linear Regression
LMA	Levenberg–Marquardt Algorithm
LxMLS	Lisbon Machine Learning School
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naive Bayes
NetGNA	Next Generation Networks and Applications Group
NLP	Natural Language Processing
NN	Neural Network
OF	Objective Function
OF0	Objective Function Zero
OF-FL	Objective Function Fuzzy Logic
OS	Operational System
PD	Probability of Default
PLUM	Polytomous Universal Model

Acronyms

PTCC	Percentage True Correctly Classified
RAM	Random Access Memory
RF	Randon Forest
ROC	Receiver Operating Characteristic
RV	Recoveed Value
SA	Sentiment Analysis
SH	Smart Home
SQL	Structured Query Language
SoC	System-on-Chip
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

Chapter 1

Introduction

This section summarizes the four years of research work under the Ph.D. thesis entitled "Decision Support System for Risk Assessment in Credit Operations Against Collateral". This thesis focuses on the study and proposal of data analysis strategies and methodologies to identify the sufficiency of collateral in credit operations. Credit granted in the decision-making process before the credit operation defaults. In the first stage, the thesis framework is described, the problem addressed, and the main objectives of the study are defined. Then, the research hypothesis is described, and the main contributions of this work to state of the art are presented.

1.1 Focus and Scope

The implementation of the New Basel Capital agreement (Basel III) brings as a challenge the estimation of critical parameters for the credit risk modeling, such as the Loss Given Default (LGD); the Probability of Default (PD); Exposure at Default (ED). Research has advanced and the fundamental aspects for the implementation of the parameters are already equated, LGD required by Basel III has been the subject of intense debate by the financial industry in Brazil and abroad.

In efforts to reduce the complexity of international standards of existing accounting dealing with financial instruments, especially the International Accounting Standards (IAS 39), and in response to the 2008 financial crisis, the International Accounting Standards Board (IASB), together with the Financial Accounting Standards Board (FASB), are reviewing such standards. The IAS 39 establishes procedures for accounting and disclosure transactions involving financial instruments. The standard also contains definitions related to such instruments and determines specific accounting procedures for the initial recognition, low and subsequent evaluation of these items.

Gitman [1] wonders about the credit check activity of a company and seeks to determine whether it should be granted credit to a client and what quantitative limits that should be imposed. The risk classification in the banking context can be seen from three aspects, first, the client risk that indicates the current capacity to incur a debt of the client. Gitman [1] uses credit analyses on client risk for the five C's model. Proposal risk is the second aspect, which evaluates the objective, purpose, value and credit term and their suitability and weighting of guarantees it is the third point, that indicates the quality (sufficiency and liquidity) that the guarantees have to mitigate the effect.

The guidance of the analysis of client risk has five key dimensions of the client's creditworthiness. They are: Character, regarding the applicant's history of compliance with its financial and contractual obligations; Capacity, referring to the applicant's potential to pay off the credit requested; Capital, regarding the financial health of the applicant; Collateral, referring to the amount of goods available by the applicant to ensure credit; Conditions, relating to economic and industry conditions existing as well as special elements which may affect both the applicant and the lender.

The measurement of recoverability is a technology that evaluates the efficiency of the mechanism of the capital return invested in goods, matching the time required to recover a property on a given system “cash flows at present value”, in which all costs returns effectively to the cashier for the entire lifetime of the asset. In practice, if recoverability doesn’t exist this may generate a decrease in assets for loss. The tracking of relationship of companies with the banking market occurs from sectors (Credit Risk Center) that have access to the institutions database which stores and provides monthly loan portfolio of several companies and individuals sent to the banks. With that, control of banking institutions know what may or may not grant with greater precision, reducing future losses from insolvent companies

Credit scoring models are used to predict the probability of a customer proposing to credit default at any given time, based on your personal and financial information that may influence the ability of the client to pay the debt. This estimated probability, called the score, is an estimate of the risk of default of a customer in a given period. This increased concern has been in no small part caused by the weaknesses of existing risk management techniques that revealed itself by the recent financial crisis and the growing demand for consumer credit. The constant change affects several banking sections because it prevents the ability to investigate the data that is produced and stored in computers that are too often dependent on manual techniques [2].

The impairment test is used to demonstrate and measure the loss of recoverability of the carrying amount of a long-lived asset. An impairment loss occurs when the carrying amount exceeds the recoverable amount of an asset or group of assets, long-term, including the value of money over time.

The existing information on credit scoring is mostly of a study on the evolution of financial indicators for some companies, which were not successful or continued their activity during the evaluated period. The failure or the success of management structure is assessed by indicator known as the cutoff score which is defined by a linear combination of financial indicators.

The scoring models exemplify a way to recognize, quantify and control the corporate risk of bankruptcy [3]. Its multidimensional character surveys a financial diagnostic of the entity and allows a relevant ranking of companies, considering certain financial outcomes that integrated into a score function. Currently, there is no universal scoring model which could be used by all financial institutions, because each institution preserves its approach to dealing with clients.

Model-based DSS integrate different kinds of mathematical and analytical models for simulation and prediction of trends. These systems exploit resolution and detail of simulation models, avoiding limitations of the approximations often used for optimization. The key issue is the choice of appropriate models and software, and the definition of data format. The following subsection is dedicated to present problem definition.

1.2 Problem Definition

The Brazilian Resolution Nº. 3721 of 30/04/2009, by the National Monetary Council (NMC), establishes in Art. 4, item XI, point c, the ”periodic assessment of the adequacy of guarantees”, i.e., the credit risk management structure should predict the establishment of criteria and procedures clearly defined and documented, accessible to those involved in the granting and credit management process to manage the sufficiency of collateral in subsequent

periods. In the banking system, predominating credit scoring models are developed from static windows and kept unchanged for years. In this setting, the two basic mechanisms of memory, short-term and long-term memory are fundamental to learning, even if you are using decision model classifiers, if it is not the right model.

The essence of the banking business, in credit decision-making, are Probability of Default (PD) models, to determine the cost of capital and in price agreement. Moreover, central banks and international regulation have dramatically evolved to a setting where the use of these models favors to achieve firmness standards for credit risk valuation in the banking system. In banking, credit risk assessment usually depends on credit scoring models, PD models.

The guarantees are divided into two types, a guarantee with individual (sponsor) and guarantee of the asset (fiduciary). Both aim to enhance security in credit operations, as a payment alternative to the credit holder provided to your lender if possibly can't honor its obligations on time. For the creditor, it generates liquidity security from receiving operation. Measurement of credit recoverability of credit is a systematic that evaluates the efficiency of the mechanism of return of capital invested in collateral.

Fiduciary guarantees generate a high operating cost and are difficult to follow "in loco" for the revaluation procedures of an asset, that is why it needs particular attention, given the large volume of goods and insufficient technical, operational capacity. According to the resolution, there is no obligation or approach to investigation that, that uses characteristics of collateral as a variable in a system to support decision making even producing a high cost in credit operations. Based on that difficulty, a conceptual evaluation of an intelligent model for recoverability verifying of credit grant against collateral is necessary.

1.3 Research Objectives

The main objective of this thesis is the proposal and performance evaluation of an intelligent model that uses a context-aware information to evaluate whether guarantees allow predicting credit granted recovery on decision-making process before the credit operation comes into default.

To reach this main goal, the following partial objectives were defined:

- Review of the state of the art on smart decision support system credit scoring, technologies and approaches that use data mining, and machine learning techniques;
- Performance evaluation of existing credit scoring systems in order to choose the best one that will be used as a reference to evaluate and validate the proposed contributions;
- Proposal, design, and construction of a new smart decision support system for credit scoring;
- Performance evaluation of the proposed approaches thought real experiments involving credit operations against collateral;
- Proposal of a smart credit scoring model to estimate the recoverability of a credit operation with a collateral as asset.

1.4 Main Contributions

The first contribution of this thesis is a review of the state of the art on identifying the most used classifiers for credit risk analysis and their performance on the decision support systems (DSSs). The work initially introduces an overview of the process that observes the evaluation of the content and systematically evaluates the recorded communication is presented in two groups of criteria, first an exploratory analysis was performed to detect the classifiers second a systematic review of papers that use the credit scoring models to estimate the recoverability of a credit operation with collateral as an asset. This was done to prepare the classifiers and to provide consistency to the analysis with or without collateral[4]. This survey was published on *IEEE Systems Journal* [5]

The second contribution is comparing the credit scoring performance of fuzzy sets and decision trees based on an artificial neural network to predict the recovered value using a sample of 1890 borrowers.This work aims to investigate collateral as a variable in the calculation of credit scores in systems that use credit operations.This contribution was published in the *Neural Computing and Applications*, from *Springer* [6].

The third contribution was submited in a special issue entitled “Decision support system on credit operation using linear and logistic regression” from *Expert Systems* (Wiley). This work aims to understand how predictive models can provide different estimations of expected recovery based on the same data sets, it compares the efficiency of the logistic regression with that of a linear regression in predicting whether recovery is proper in a credit operation was published on *Wiley*[7].

The fourth contribution of this thesis proposes the determination of the best combination of parameters to use with ANNs and the BN approach to handle and precisely evaluate credit risk. The main contribution of this study is its proposed machine learning model or the combination of them, which is a rare approach to the problem of credit risk measurement. The study highlights the existing gap that prevents intelligence systems from addressing bank modeling concerns. This contribution was published on *Journal of Artificial Intelligence and Systems* [8] .

Finally, the last contribution introduces an evaluation of the stability of the two models based on the variables chosen. The method used by banks in making loans decisions is unclear; however, implementation of classical linear models in banking systems is adequately documented. For this paper, the transparent elastic approach was used as a benchmark. This research shows advantages of the RF approach over the SVM algorithm which are its speed and operational simplicity, and SVM has the benefit of higher classification accuracy than RF. This contribution was published on *Software: Practice and Experience* [9] .

1.5 Thesis Statement

This Thesis proposes several machine learning classifiers for improvement of credit risk operations on financial institutions using ensemble learning classifiers. An output is a score model that explains the chance of a given entity, a private individual or a company, becoming a

defaulter (insolvent) in a future period. The current approaches do not use the set of entries of collateral, increasing the impairment of Financial Institutions. Furthermore, this study claims that a new smart credit scoring model based on decision support systems should predict the right amount of collateral to compensate the credit operation.

1.6 Document Organization

This thesis comprises seven chapters, which are organized as follows. The first chapter introduces and focuses on the topic of the study, presents the work's motivation, identifies the problem delimitation, defines the thesis objectives, highlights the main contributions, and includes the thesis statement. The document's organization is also included in this chapter. Except for this, and the concluding chapter, all other chapters are based on papers published in, or submitted to, international journals.

Chapter 2 presents the survey paper entitled "Classification Methods Applied to Credit Scoring with Collateral." The paper provides an in-depth review of the state-of-the-art of the analysis includes 84 studies in this work to propose a using statistical methodology to conduct a meta-analysis to compare the results of classification methods.

Chapter 3, entitled "Machine Learning and Decision Support System on Credit Scoring", compares the credit scoring performance of fuzzy sets and decision trees based on an artificial neural network to predict the recovered value. This paper is an initial study of collateral as a variable in the calculation of the credit score.

Chapter 4, entitled "Support System on Credit Operation Using Linear and Logistic Regression," aims to understand how predictive models can provide different estimations of expected recovery based on the same data sets.

Chapter 5, entitled "Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction", compares Bayesian networks with artificial neural networks for predicting recovered value in a credit operation. The study explores this problem and finds that ANNs are a more efficient tool for predicting credit risk than the Naïve Bayesian (NB) approach

In Chapter 6 "Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation", shows advantages of the RF approach over the SVM algorithm which are its speed and operational simplicity, and SVM has the benefit of higher classification accuracy than RF.

Chapter 7 summarizes the main conclusions of the thesis and suggests future work that can emerge from the conducted work.

In the Appendices three contributions are presented, three poster presentations in international conferences. In Appendix A, entitled "Decision Support Systems to Predict a Sufficiency of Collateral for Credit Risk Operations". An evaluation of existing credit scoring classifiers methods to choose the best one that will be used as a reference to evaluate and validate the proposed contributions.

Appendix B, entitled “Intelligent Decision Support System on Credit Scoring,” introduces an Intelligent Decision support systems (IDSS) and Knowledge-based Expert Systems (KES) are two of the strongest general paths on Machine Learning (ML)

Appendix C (“Using Natural Language Processing with Sentiment Analysis to Improve Credit Score on Collateral Reports”) proposes an evaluation of existing credit scoring classifiers methods to choose the best one that will be used as a reference to evaluate and validate the proposed contributions.

References

[1] L. J. Gitman *et al.*, *Princípios de administração financeira*. Harbra São Paulo, 1997.

[2] T. Harris, “Credit scoring using the clustered support vector machine,” *Expert Systems with Applications*, vol. 42, no. 2, pp. 741–750, 2015.

[3] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, “Benchmarking state-of-the-art classification algorithms for credit scoring,” *Journal of the operational research society*, vol. 54, no. 6, pp. 627–635, 2003.

[4] R. H. Kolbe and M. S. Burnett, “Content-analysis research: An examination of applications with directives for improving research reliability and objectivity,” *Journal of consumer research*, vol. 18, no. 2, pp. 243–250, 1991.

[5] G. Teles, J. J. Rodrigues, K. Saleem, and S. A. Kozlov, “Classification methods applied to credit scoring with collateral,” *IEEE Systems Journal*, 2019.

[6] G. Teles, J. J. Rodrigues, K. Saleem, S. Kozlov, and R. A. Rabêlo, “Machine learning and decision support system on credit scoring,” *Neural Computing and Applications*, vol. 32, no. 14, pp. 9809–9826, 2020.

[7] G. Teles, J. J. Rodrigues, S. A. Kozlov, R. A. Rabêlo, and V. H. C. Albuquerque, “Decision support system on credit operation using linear and logistic regression,” *Expert Systems*, p. e12578.

[8] G. Teles, J. J. Rodrigues, R. A. Rabêlo, and S. A. Kozlov, “Artificial neural network and bayesian network models for credit risk prediction,” *Journal of Artificial Intelligence and Systems*, vol. 2, pp. 118–132, 2020.

[9] G. Teles, J. J. Rodrigues, R. A. Rabêlo, and S. A. Kozlov, “Comparative study of support vector machines and random forests machine learning algorithms on credit operation,” *Software: Practice and Experience*, 2020.

Chapter 2

Classification Methods Applied to Credit Scoring With Collateral

This chapter consists in the following paper:

Classification Methods Applied to Credit Scoring With Collateral

Germanno Teles, Joel J. P. C. Rodrigues, Kashif Saleem and Sergei A. Kozlov

IEEE Systems Journal, IEEE, ISSN: 1932-8184, September 2019.

DOI: doi.org/10.1109/JSYST.2019.2937552

According to Journal Citation Reports published by Thomson Reuters in 2019, this journal scored ISI journal performance metrics as follows:

ISI Impact Factor (2019): 3.987

Journal Ranking (2019): 31/156 (Information Systems)

Journal Ranking (2019): 56/266 (Electrical & Electronic)

Journal Ranking (2019): 17/83 (Operations Research & Management Science)

Journal Ranking (2019): 24/90 (Telecommunications)

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Classification Methods Applied to Credit Scoring With Collateral

Germanno Teles, Joel J. P. C. Rodrigues, *Senior Member, IEEE*, Kashif Saleem, and Sergei A. Kozlov

***Abstract*—Credit operations are indispensable in the organizational development of financial institutions. However, misconduct in these operations occurs, and this can lead to financial loss. These consequences are caused by incorrectly granting credit or incorrectly assigning customer ratings and can compromise a credit portfolio. The result shows that support vector machine is the most commonly used classifier for credit scores, and while the system performs well, it does not apply approaches with collateral. The analysis includes 84 studies in this article to propose using statistical methodology to conduct a meta-analysis to compare the results of classification methods. It shows some cases that consider various probability distributions and also survival data. It also elaborates that collateral is not the first approach for credit scoring. The credit scoring system can then give several starting credit scores according to the classifier the user wants to use.**

***Index Terms*—Basel Accords, Bayesian network (BN), collateral, credit scoring, decision support system, neural network (NN), support vector machine (SVM).**

I. Introduction

NO INDUSTRY is risk free. Risks can be minimized, managed, shared, accepted, and transferred, but not ignored [1]. The risk of default (credit risk) was considered by Angbazo [2] among the determinants of pure spread, along with interest rate volatility. According to Angbazo, the optimal spread represents an insurance not only against interest rate volatility, but also against the risk of default by the borrowers. Thus, a positive relationship between the spread and the credit risk is expected. However, similar to risk aversion, the impact of the credit risk on spread formation may differ depending on the nature of the loan. [2] The banking sector is regarded as one of the most uncertain industries as credit operations pose some of the greatest risks [3]. The complexity of the banking industry and the credit issuing procedures generate more significant risks. Credit operations against collateral can help mitigate the risks involved.

At the beginning of 2004, the Basel Accord was reinforced by the Basel banking committee, expanding the applications of the credit score to manage and monitor the risk of operations, along with the existing policies and procedures in granting of credit. The work carried out by the institutions on risk classifications was incorporated into the institutions themselves with the Basel III agreement from 2013, providing far more valuable estimates of the risk of default. Reference [4] described the obligations and methodologies that must be followed and the forms that must be published for any significant change to be produced by the institution [4].

The analysis of client risk has five key dimensions to establish a client's creditworthiness. They are as follows.

1) Character, regarding the applicant's history of compliance with its financial and contractual obligations.
2) Capacity, referring to the applicant's potential to pay back the requested credit.
3) Capital, regarding the financial health of the applicant.
4) Collateral, referring to the amount of goods available with the applicant to ensure credit.
5) Conditions, relating to current economic and industry conditions as well as special factors that may affect both the applicant and the lender.

Credit scoring models are used to predict the probability of a customer requesting credit to become a defaulter at any given time, based on the costumer personal and financial information that may influence the ability of the customer to pay the debt [5], [6]. This estimated probability, called the score. A significant part of the increased concern has been caused by the weaknesses of existing risk management techniques that have been revealed by the recent financial crisis and the growing demand for consumer credit. The constant change affects several banking sections because it affects the ability to investigate the data that are produced and stored in computers, which are too often dependent on manual techniques [7].

Glass [8] defines meta-analysis as an analysis of analyses, that is, a statistical analysis that aims to combine results already found in previous analyses of different studies of the same interest. The meta-analysis aims to combine studies performed under different conditions, with different levels of precision and by groups of researchers from different regions and formations. Thus, broader conclusions are expected than those obtained by the studies that compose the systematization [9].

Manuscript received June 25, 2019; revised August 15, 2019; accepted August 21, 2019. This work was supported in part by National Funding from the Fundação para a Ciência e a Tecnologia under Project UID/EEA/50008/2019, in part by the Government of the Russian Federation under Grant 08-08, in part by the Brazilian National Council for Research and Development (CNPq) under Grant 309335/2017-5, and in part by Ciência sem Fronteiras of CNPq, Brazil, under process number 200450/2015-8. *(Corresponding author: Joel J. P. C. Rodrigues.)*

G. Teles is with the Instituto de Telecomunicações, Universidade da Beira Interior, Covilha 6201-001, Portugal (e-mail: germanno.teles@ubi.pt).

J. J. P. C. Rodrigues is with the Instituto de Telecomunicações, Universidade da Beira Interior, Covilha 6201-001, Portugal, with the Federal University of Piauí, Teresina 64049-550, Brazil, with the Center of Excellence in Information Assurance, King Saud University, Riyadh 11451, Saudi Arabia, and also with ITMO University, St. Petersburg 197101, Russia (e-mail: joeljr@ieee.org).

K. Saleem is with the Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh, Saudi Arabia (e-mail: ksaleem@ksu.edu.sa).

S. A. Kozlov is with ITMO University, St. Petersburg 197101, Russia (e-mail: kozlov@mail.ifmo.ru).

Digital Object Identifier 10.1109/JSYST.2019.2937552

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

This article presents a review and evaluation of the existing credit scoring classifier methods. An extensive review of the literature focusing on classification methods applied in the credit scoring is performed. Then, the main contributions of this article are as follows.

1) A review of the state of the art on identifying the most used classifiers for credit risk analysis and their performance on the decision support systems (DSSs).
2) Comparison analysis of the most relevant solutions considering the classifiers that use collateral as a parameter for calculating the risk.
3) Identification of the promising open research issues for further studies.

A research process that observes the evaluation of the content and systematically evaluates the recorded communication is presented in [10]. Thus, the following two groups of criteria are considered in this article to choose studies connected with the credit scoring to stay involved in this article.

1) An exploratory analysis was performed to detect the classifiers that have patterns and other elements that could compromise the results of this article. This was done to prepare the classifiers and to provide consistency to the analysis with or without collateral.
2) A systematic review of papers that use the credit scoring models to estimate the recoverability of a credit operation with collateral as an asset.

The remainder of this article is organized as follows. Section II introduces the conceptual classification scheme for the systematic literature review and highlights some important practical aspects of credit scoring. Section III shows the conceptual classification design for the systematic literature review. Section IV addresses a behavior of the credit scoring system related work, followed by a summary of the credit scoring classifiers. Section V discusses the credit scoring classifiers and identifies open research topics on the topic. Finally, this article presents a summary of lessons learned and conclusion in Section VI.

II. Credit Scoring

Currently, the credit scoring is considered an important tool to prequalify borrowers and assist managers in making better risk decisions for the business. The use of models that enables the managers/authorities to decide whether or not to grant credit is considered in an objective, standardized, and impartial manner, which is not guaranteed in the judgmental analysis. This enables the client to be treated in a personalized way, regardless of the service channel.

The way of banks classify risks is with the use of the proposal risk. It assesses the objective, purpose, value, credit term, and the applicant's suitability. Weighting of guarantees is the last aspect that indicates quality (sufficiency and liquidity) that the guarantees must mitigate effect.

Several techniques that apply statistics and artificial intelligence are applied in the credit scoring, such as decision trees (DT), neural networks (NN), linear discriminant analysis, logistic regression (LR), Bayesian networks (BN), and support vector machine (SVM). However, the use artificial intelligence is still limited, despite being a powerful tool for pattern recognition. This is because of its "black box" nature. It does not know either the dependency on the relationships between the variables of the model, or the contribution of each variable.

According to Ala'raj [11], the all those classifiers introduce on credit score approach are well-known base classifiers in this domain are used o his work they show results, analysis, and statistical tests demonstrate the ability of the proposed combination method to improve prediction performance against all base classifier.

Jadhav [12] considered classification as an example of supervised learning as training data associated with class labels and focus on study of various classification techniques, showing its advantages and disadvantages.

The work of Moro *et al.* [13] presents a study on business intelligence on a bank institution showing the potentially benefit business, increasing the visibility and recognition of research achievements.

The behavior of an AI can be reinforced by the work of Abid *et al.* [14] that use this classifier on analyses of costumers' loans default payment allowing providing an effective decision support system for banks.

There are two types of models for measuring and estimating the probability of a customer becoming a defaulter. They are the following [15].

1) Credit Scoring—Obtained from the registration information provided by customers, such as type of residence, income level, age, occupation, level of education, relationships with financial institutions, and consultations with credit bureaus.
2) Behavioral Scoring—It is a scorsing system based on behavioral analysis and uses the information that the company already has regarding the customer on the renovation, maintenance, and granting of a new credit line. It may include information related to consumer habits, payment, and income commitment, amongst others.

Credit scoring models are used to estimate the probability of a customer proponent to credit become default during a particular period, based on the costumer's personal and financial information, which may influence the ability of the customer to repay the debt. This estimated probability, called the score, has a value between 0 and 100, and is an estimate of the risk of default of a customer for a given period.

III. Methodology

A systematic review is an option for cataloguing and classifying important scientific contributions to an area on a systematic, qualitative, and quantitative information of the content in the literature [16]. It consists of an observational study method applied to systematically estimate the content of registered information.

The systematic review limits the study qualification to journal papers, especially considering "credit scoring" as a keyword related to "machine learning," "decision support system," "classification," or "statistic" topics.

It included 139 papers in the study. It was reduced to the systematic review to 84 documents according to three interrogatories on the conceptual summary over the methods: What is the type of the primary classification method? Is the paper using

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

collateral as an independent variable? How many times was that paper cited?

It was used as a statistical method with meta-analysis and compare findings in each of the methods. The statistical analysis which combines the results of several independent studies considered by the analyst to be combinable [17].

The meta-analysis was divided into two groups cited before 2014 and after 2015, assuming that the Basel III deadline to implementation. By performing a meta-analysis, it will be capable of reducing the difficulties of understanding due to sampling variation and quantifying between study variations. Meta-analysis can also assist in answering questions not posed by individual studies. It will imply to understand the comparative effectiveness of multiple interventions for the same condition and settle debates emerging from conflicting studies.

With the BN approach, it is observed that 12 studies are shown since 2001, where the credit score can be calculated, sometimes using the collateral [18]–[29]. The NN method appears more frequently applying on the credit score since 2000 with a great number of citation [30]–[41]. The methodical distributions of citation are evidence reviewed, during the period using fuzzy on the studies [42]–[53]. It was observed that in some periods, the number of citations is equal to zero. Those studies are automatically excluded from the meta-analyses [25], [54]–[64]. A certain level of capacity control applied with SVMs produces a high level of quality for the performance of the out-of-sample attributes [65]–[76].

Making the LR model more broadly applicable requires the consideration of suggestions of methods of including previously rejected applications for credit in the building-up step of the model [60], [77]–[88]. The linear regression, despite the few studies, can be identified on a little using collateral [89]–[101].

The Forest plot shows that the point estimate describes each study and the 95% confidence interval. So, the square in the middle for each study is the point estimate. Also, the size of the square differs between studies so that the more extensive study takes a more substantial, and the square is more significant because it takes more weight in the meta-analysis.

The analysis of the credit scoring with collaterals from various viewpoints reach to a variation of questions and research methods. For example, the term "collateral" exists 1 610 000 results and 534 000 "credit scoring" on Google Scholar. So, to obtain a significant meta-analysis, it defined criteria for separating the relevant studies of the large amount of previous work. Therefore, studies with hybrid methods with collateral as the independent variable were excluded. Some studies have been excluded due to lacking statistical information.

It has initiated our search for relevant publications queering journals, conferences, and some books pairing terms of "credit score" with "decision support system," "bank intelligent system," "Basel Accords," "BN," "NN," "SVM," "Fuzzy," "decision tree," "linear regression," or "LR." After regarding the abstracts of the studies returned by our search query, it downloaded those that show any promise of containing empirical estimates of the utilized collateral on credit score.

IV. Classification Methods in Credit Scoring

Credit operations are most frequent in the banking industry. Therefore, it is important to analyze the DSS for assessing the risk factors on credit operations against collateral in the banking industry. The credit risk assessment in credit operations is largely used in all banks globally. Therefore, the assessment of the credit risk has a lot of disagreement and even a gloomy process [102]. There should be a variety of risk methods that should be used for assessing risk. Also, the credit risk is one of the main functions of the banking industry organizations, such as banks categorizes the customers based on their profile. It may range from the financial contextual of the customers to the individual aspects of the clients [103].

It is a very effective decision tool that can act as a DSS for the assessment of the risk factors that are involved against collateral. Owing to its active nature, noisy and incomplete data can be used to extract inferences when they are provided. The most remarkable feature of the BN is its ability to provide both qualitative and the quantitative information. Therefore, it can be applied in all domains [104]. With regard to issuing and paying back credits, both kinds of features are involved in the objective and the subjective factors. Every bank has their own opinion, which cannot be changed [105]. However, the conclusion may change when the banker knows the borrower very well. Furthermore, the risk of each type of credit varies as they are different. For instance, the required conditions for providing working capital credit are that the estimated credit risk of the transaction must be below the admissible level and the current credit potential of the customer and the ability of the customer to provide collateral must be acceptable. Each of these factors has a particularly special meaning [106]. For example, the current ability of the customer to pay back the borrowed credit is used to determine whether the customer can pay back the requested credit and interest on time [107]. When determining the credit risk class, the economic and financial strength of the person must be stated [108].

In attempting to understand how risks relating to credit operations can be assessed, the manner in which loan applications are categorized into either bad or good applications by financial institutions must be considered. The bad applications are the applications that are rejected because of the low probability of the applicants returning the loans. Institutions, therefore, must employ loan officers to make credit decisions or recommendations for the organization. Every institution must have rules in place to govern the evaluation in the worthiness of loan applications. Several reasons have been suggested for why human beings make poor decisions when it comes to evaluating the creditworthiness of an individual. Some of the reasons are that the business data repository stores historical data from the hidden data, there is a considerable gap of data when the decision depends on the knowledge of skilled employees, and the fact that humans are inclined to arbitrariness.

The determination of loan applications tends to be flawed with regard to human beings. The use of a knowledge discovery system is key to make decisions regarding loan applications. This will enable an easier assessment of the risk [109].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

For the development of a model, there are two main segments of the system as shown in the neuro-based construction model. The initial section is the part that deals with the decision making by considering the basic decision rules [110]. The high accuracy in decision making is a symbol that all the risk factors have been considered; therefore, it is easy to access all the possible risks that can be incurred during the credit operations against the collateral [111]. Through this, it becomes easier to perform credit operations with confidence and to avoid any factors that can amount to risk in the process [112]. Experts select the actions that are either created by the DT or developed using the standards of the organization that is being borrowed from.

In the fuzzification, the input from the databases is converted into a particular fuzzy set using three different input variables [113]. Each variable is assigned a percentage scale that results in the construction of an input membership function which highlights the threshold on the variables that are helpful in the development of the system-based rules.

According to Abdou [114], in credit scoring, the fuzzy credit rating techniques are based on a hierarchical method. Thus, the process begins with the user (could be a credit officer) connecting to the provided credit interface, which is connected to the database. The database has all the clients' information that feeds the fuzzifier that subsequently begins the fuzzy process.

Researchers indicate that the performance of NNs, SVM, and BNs is higher than any of the individual classifiers applied independently. Similar experiments have yielded that the performance of the three base learners improves the overall accuracy of the sample classification while also enabling the identification of the most vital features in the dataset of choice [30]. Therefore, the combined assessment method proves to be a useful method of feature selection and classification problems. The combination of methods is based on the fact that each method represents its characteristics and features. Each method has its requirements, which makes the application of a combined method broad in the credit scoring classification [116].

A. Analysis of the BN

Developing BNs is the main challenge [117]. In this section, the steps and other requirements to develop a functional BN are provided. Once the tool has been developed, it is very easy to use it, as it can consider both the objective and the subjective factors that are involved [118]. Furthermore, this system can make decisions even with incomplete data. It assesses the existing information and makes the necessary calculations before making the final decision [119]. The only challenge involved in using a BN-based tool is that an expert is required to operate it. It requires the expert to be very competitive and look at all the data that have been presented to them [120]. If the bank sets its criteria, it should be consistent and constant with regard to all the branch offices of the bank.

The financial institution or the institution involved should determine the objective and subjective factors with regard to its goal and strategy [121]. From the analysis of the BN as a decision tool that can be used as a DSS, it can be seen that it is very effective and can be used to show the risks involved in the credit operations. Most of the risks are associated with the collaterals.

Bayesian learning is computationally costly. This holds true even when the network structure is already provided. Moreover, BNs tend to perform poorly on high-dimensional data. Finally, BN models can be hard to interpret and require Copula functions to separate out effects between different parts of the network.

B. NNs in Decision Support System

The need for a decision, in most cases, is subject to the urgency in an environment with large amounts of information. NNs can be used in specialized systems to improve the decision making in corporations. "This is because they serve the operations, the management, and the planning levels of an organization and aids in the decision-making process that can change rapidly and cannot be easily predetermined" [122]. The maturation of the artificial intelligence techniques and a blend of Internet empowering and entry have contributed to the improvement of decision-making support under uncertain and risky conditions. This assists in providing the ability to improve the quality of the decision by making suggestive solutions compared to those that are made by humans alone [123].

In a second step for all groups, the DTs are built to obtain the rules that are fed into the NN for displaying the clients that are not expected to repay the loan [124]. When these techniques are used together with the established rules, it becomes possible to make a quality decision that will assess all the risk factors that are involved.

NNs are limited in that they require a large dataset, the black box of the node so if you want to know what causes the output, despite the training dataset is too large.

C. DT Neuro-Based Model Design

Developing a DT model is proposed to be lined up with the intelligent decision-making process in a complicated procedure in computational and organizational decision making such that it can be in line with the current complex and dynamic environments [125]. The risks involved in the credit operations have intensified because of the changing nature of the economic environment [126]. Therefore, DT is more efficient in working with the involved complexities, together with the increasing volume of data that are needed in the decision-making process [127].

For most of the papers that has been mentioned, instability is the main obstacle. The information fed to the model must be well defined to change variables, duplicate data must be excluded, and altering the sequence midway can lead to significant changes and might require redrawing of the tree.

D. Application of Fuzzy and Fuzzy Logic in Risk Assessment

The method of fuzzy logic, if necessary, is capable of handling the previously described linguistic terms. In situations where the specialists validate and acquire knowledge, the diffuse logic provides a structure capable of identifying relevant issues making inferences on the evaluation of the risks. This model makes

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

inferences and uses estimation from unclear information and data. For highly complex models, diffuse logic theory can be used to identify risks. Therefore, the application of fuzzy logic can be sufficient in assessing the credit risks associated with credit operations because it is able to make accurate decisions from the available information.

E. SVM in Risk Assessment

SVM has been extensively used in the suppression of the subjective ad hoc elements in the process of decision making and strengthening of sustainability and accountability, as well as reduction of corruption [128]. The application of the SVM is based on the fact that it enables the classification of arbitrarily complex cases through the shaping of arbitrarily complex sets of decisions. SVMs can identify special data points from decision sets in a parsimonious manner [129]. Capacity control is used to avoid overfitting, with penalties applied to the number of the support vectors of the models.

F. Logistic Regression

LR represents one of the most prominent methods used for conducting the credit scoring. It represents one of the most successful methods despite being one of the earliest. The method provides for both the dependencies structure of the explanatory variable as well as the statistical significance of the variables.

The LR model is extremely suitable for the determination of good credit versus bad credit paid back versus default such that it is well suited to predictive modeling subsequent to datasets containing a binary indicator variable.

Some of studies do not have any citation but this is acceptable because some studies are published after 2014. A simple linear function is also obtained through the application of additional information to the model parameter, and it can be associated with each of the predictor class values after coarse coding [130]. Therefore, the final credit score, which also determines the credit risk, is a simple sum of individual use of the scores that can be obtained from the scorecard.

G. Linear Regression

Linear regression has developed to become one of essential components in the data analysis concerned with the description of the relationship between response variables and other independent variables. In the credit scoring applications, linear regression has been extensively used as the two class problems that can be represented by a dummy variable [131].

Linear regression can be used in credit analysis for assessing factors, such as guarantees, default rates, and historical payments by a customer, which can then be compared with the cutoff point of a bank.

V. Comparison Analysis and Discussion

In this section, it has presented a more detailed analysis and discussion of the most relevant solutions described above.

The problems of credit risks have been consistently addressed in artificial intelligence journal publications [132]. The primary concern has been that personal data are used and there is also a growing fear that an algorithm could replace human beings in decision making. These questions are genuine, and this article emphasizes the appropriate algorithms with regard to decision making in lending institutions. Algorithms can be implemented to simplify a process while at the same time increasing its fluidity as well as increasing speed [133]. Algorithms include a set of code modeled to achieve set objectives. For example, an algorithm designed to perform a recruitment process introduces several discriminatory conditions based on individual profiles. A similar approach is adopted by lending institutions when making lending decisions to banks. It is, therefore, of the essence to understand the underlying challenges and find ways to manage the use of algorithms.

As a kind of classification technique implemented in a credit scoring model, the choice of method is frequently associated with the subjectivity of the state-of-the-art methods. A correct prediction shows whether a credit spread to a candidate will probably end in profit for the lending institution.

Each study reviewed by an estimator effect or the result, it has used odds ratio to measure of impact is a weighted average of the results of the individual studies, so the full measure of effect is represented by the blue diamond.

In this article, it is interested in identifying the importance of collateral among the credit scores on classifications methods. A related methodology was adopted by Hamdaoui *et al*. [134] in the context of a meta-analysis to choose potential indicators of heterogeneity in previous observational results concerning the influence of guarantee.

It was used as a random effects model in assuming there is a distribution of true effects is different, and it assumed there is a variation of them. This variation is called effects heterogeneity that refers to the credit score classification methodological diversities among a set of studies.

The critical part of the random effects model is figuring out the weight, and the weight equals the inverse of the variance. However, instead, the variance has two components. One is the within study variance. The second is the between study variance. The density matches the inverse of the variance, but the between study variance modifies the variance.

Some key aspects emerge from Table I. First, it is of common practice to use a small number of datasets, many of which contain only a few cases and independent variables, separately of the classification are adopted the possibility of happening of inconsistencies among predictions presented by the different classifier.

Second, the number of classifiers per study varies some statistical hypothesis testing is often neglected or employed inappropriately. Common mistakes include using parametric tests or performing multiple comparisons without controlling the family wise error level. Most studies rely on a single performance measure or measures of the same type. In general, performance measures split into three types.

The one without a collateral set of studies misses an essential aspect of scorecard performance because financial institutions require the probability of default.

It was reviewed that the validation method, the SVM is the most used with randomization, the partitioning of the datasets in two different branches. The fitting of the model is, therefore,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
SUMMARY OF CREDIT SCORING CLASSIFIERS

Classifiers	Representative references with collateral	Representative references without collateral
Bayesian Network	Kao *et al.* Vedala *et al.* Yang *et al.*	Giudici *et al.* Baesens *et al.* Mira *et al.* Maltritz *et al.* Biçer *et al.* Hsieh *et al.* *Zhang et al.* *Mileris et al.* *Bekiroglu et al.*
Neural Networks	*Malhotra et al.*	*Li et al.* West *et al.* Atiya *et al.* Witkowska *et al.* Wang *et al.* West *et al.* Baesens *et al.* Paliwal *et al.* Tsai *et al.* Lisboa *et al.* Khashman *et al.*
Decision Tree	Lin *et al.*	*Zhang et al.* Hsieh *et al.* Zurada *et al.* Li *et al.* Tuffery *et al.* Yap *et al.* Wu *et al.* Wang *et al.* Kabari *et al.* Marque *et al.* Bhuvaneswari *et al.*
Fuzzy logic	Hoffmann *et al.* Lahsasna *et al.* Yu *et al.* Lahsasna *et al.*	*Tah et al.* Hoffmann *et al.* Kogut *et al.* Tang *et al.* Laha *et al.* Bojadziev *et al.* Laha *et al.* Bekiros *et al.*
Support Vector Machine	Chen *et al.* Xu *et al.*	Van *et al.* Huang *et al.* Wang *et al.* LI *et al.* Martens *et al.* Huang *et al.* Wang *et al.* Hao *et al.* Zhou *et al.* Min *et al.*
Logistic Regression	Hamadi *et al.* Yap *et al.*	Galindo *et al.* Zekic-Susac *et al.* Bandyopadhyay *et al.* Gouvea *et al.* Vlah *et al.* Figini *et al.* Nikolic *et al.* Ferreira *et al.* Triki *et al.* Sohn *et al.*
Linear Regression	Lee *et al.* Jiang *et al.* Sinha *et al.* Martens *et al.* Smith *et al.*	Srinivisan *et al.* Desai *et al.* Hand *et al.* Hsieh *et al.* Huang *et al.* Majeske *et al.*

based on the training set, which aims at predicting the cases in the test set. The model can generate the data if the second dataset demonstrates an excellent performance such that overfitting is wholly avoided at the first subset with the training set.

The results can be summarized as follows on Table I, seven classification methods for risk credit measure which can be applied to decision support systems, and only three, linear regression, fuzzy, and BN show some papers using the collateral as an independent variable.

The general condition to authors is that they must demonstrate proof of the effectiveness of their score process on any simulated as real datasets as well. "There are many times in which the number of datasets used is small or engages the use of some other real datasets that present proof of the rating method owing to the hassle associated with acquiring information at the credit score." A study conducted by Bijak and Thomas [135] is an example that employed 16 different accessible datasets to testify the measures of performance and which were used in the application for a credit card.

VI. CONCLUDING REMARKS

A. Lessons Learned

In this article, we reviewed an essential facet of the business as it can drive business into creating value or transform the entire business. As much as data can be used toward the completion of simple tasks, the real power of data is drawn from the use of analytical tools used to extract useful knowledge.

There are at least two datasets that mostly used the Australian and German in the credit score determination and evaluation [72]. These credit scoring datasets are the most common. Besides, there is the Benelux credit dataset, which originates from Belgium, the Netherlands, and Luxembourg institutions. These are the three types of datasets that are available publicly. Different subsets under the Benelux include the Bene1 and Bene2, as well as the behavioral dataset. The Australian credit approval dataset is a multivariate dataset, which allows several instances. It is categorical, considers integers and real numbers. This type of dataset is extensively used for classification purposes. It is a new dataset because of the mix of attributes including nominal with both small, and significant numbers of values, as well as continuous.

The other type is the Credit Fraud dataset, which falls under the German credit dataset category [72]. This type is also multivariate, allowing several instances to be considered. It provides two datasets, which include the original dataset and a modified dataset that has been edited to include several indicator variables aimed at making it suitable for the algorithm unable to cope with categorical variables. The attributes can either be qualitative or numerical while the dataset demands the use of a cost matrix, whereby the columns represent the predicted classification while the rows represent the accrual classification. The other type of dataset is the data mining dataset, which is used for data mining, which allows the description of several types of datasets depending on the experiment under consideration. They perform

different tasks, and the size depends on the type of data being mined.

The distributed file system requires analytics of data to have a solid IT infrastructure that can support the work. Database management is an important part of data analytics [136]. An ecosystem exists that mainly deals with database systems including NoSQL and object-oriented-types. Database management systems that are well known include Oracle, SQL, and Sybase. Dataset selection faces the challenge of allocating the number of input variables.

The access to information is considered secure nowadays due to the modernization of the web as well as the creation of large data storage center. However, data availability on the credit history of businesses and customers is still facing challenges in terms of access. Careful consideration is required before releasing any data containing confidential information on applicants [137].

B. Conclusion

In the banking system, the predominant credit scoring models are generated from static windows of time and kept unchanged, possibly for years. In this setting, the two primary mechanisms of memory, short-term memory and long-term memory, are fundamental to learning, even if you are using the decision model classifiers which cannot be the right model.

In review, it was conducted a study of the current and essential components for some decision support systems starting from the Basel Agreement until now. This article clarifies the meaning of the DSS system and the "intelligence" involved in them with development at the financial institutions of the qualitative and quantitative variables into credit operations for granting credit process by describing the many classifiers that it can use. Therefore, the proposed approach provides a way and analysis for evaluating the credit risk. Most of these papers do not take into account endogenous variables, which may decrease their significance.

This article describes the importance of using the DSS to help decision makers in credit operations. This article is based on recent studies and identifies approaches and technologies used in several areas. This article, regarding the use of these systems on the credit scoring is vast and is divided into subtopics, such as knowledge-based systems, data-based systems, and model-based systems. During the preparation of this article, several papers regarding these subtopics are analysed, enumerating the different methods.

A detailed meta-analysis of the classifiers presented in this article used in the credit score within some areas of knowledge, the model of a smart system to give support to the decision making still presents several challenges one of them being the approach of collateral.

The meta-analysis reports that this heterogeneity also has an important impact on the primary outcomes. For future works, it can analyze some other way to measure the meta-analysis such as risk ratio or risk difference. The authors also get a comparison of the random model to the fixed model, to check if there is no difference between them. It can introduce some studies with hybrids (combined) methods on the credit score.

The primary objective of this article lies in the creation of a model that uses context-aware information to evaluate if guarantees allow credit granted recovery on decision-making process capable of helping decision makers in credit operation. For this objective, this article is based on previous studies, attempting to detect a viable approach. The systems investigated in each subtopic present the data mining and machine learning techniques as a necessary solution, as well as show us the necessity of finding the best classifier that will be used for meeting the objectives of this article.

REFERENCES

[1] M. Latham, "Constructing the team joint review of procurement and contractual arrangements in the United Kingdom construction industry; final report," 1994, p. 129. http://www.opengrey.eu/item/display/10068/491035. doi: GB_1994:13111.

[2] L. Angbazo, "Commercial bank net interest margins, default risk, interest-rate risk, and off-balance sheet banking," *J. Banking Finance*, vol. 21, no. 1, pp. 55–87, Jan. 1997.

[3] F. J. L. Iturriaga and I. P. Sanz, "Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks," *Expert Syst. Appl.*, vol. 42, pp. 2857–2869, 2015.

[4] R. VM, S. Kumar, and J. Kumar, "Basel II to Basel III—The way forward," Infosys, Bengaluru, India, 2013. [Online]. Available: https://static1.squarespace.com/static/537e8bcbe4b09ac6c31f0ae6/t/53dabea8e4b0cfc1a4b306fe/1406844584460/

[5] C. Tuckwell and A. Mendonça, "The global crisis and unconventional monetary policy: ECB versus Fed," CEsA - Center for African, Asian and Latin American Studies, Working Paper no. 141, 2016. [Online]. Available: https://EconPapers.repec.org/RePEc:cav:cavwpp:wp141

[6] K. P. Wong, "On the determinants of bank interest margins under credit and interest rate risks," *J. Banking Finance*, vol. 21, no. 2, pp. 251–257, 1997.

[7] T. Harris, "Credit scoring using the clustered support vector machine," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 741–750, 2015.

[8] G. V. Glass, "Primary, secondary, and meta-analysis of research," *Educ. Res.*, vol. 5, no. 10, pp. 3–8, 1976.

[9] R. H. Fagard, J. A. Staessen, and L. Thijs, "Advantages and disadvantages of the meta-analysis approach," *J. Hypertens. Suppl.*, vol. 14, no. 2, pp. S9–S12; discussion S13, 1996.

[10] R. H. Kolbe and M. S. Burnett, "Content-analysis research: An examination of applications with directives for improving research reliability and objectivity," *J. Consum. Res.*, vol. 18, no. 2, pp. 243–250, 1991.

[11] M. Ala'raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring," *Knowl.-Based Syst.*, vol. 104, pp. 89–105, 2015.

[12] S. D. Jadhav and H. P. Channe, "Comparative study of K-NN, naive Bayes and decision tree classification techniques," *Int. J. Sci. Res.*, vol. 14611, no. 1, pp. 2319–7064, 2016.

[13] S. Moro, P. Cortez, and P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1314–1324, 2015.

[14] L. ABID, S. Zaghdene, A. Masmoudi, and S. Z. Ghorbel, "Bayesian network modeling: A case study of credit scoring analysis of consumer loans default payment," *Asian Econ. Financial Rev.*, vol. 7, no. 9, pp. 846–857, 2017.

[15] J. R. Securato, *Crédito - Análise e Avaliação do Risco*, 2nd ed. São Paulo, Brazil: Saint Paul, 2012.

[16] R. Quansah *et al.*, "Association of arsenic with adverse pregnancy outcomes / infant mortality:," *Environmental Health Perspective*, vol. 123, no. 5, pp. 412–422, 2015.

[17] M. F. Huque, "Experiences with meta-analysis in NDA," *Biopharmaceutical Sect. Amer. Statist. Assoc.*, vol. 2, no. 1, pp. 28–33, 1988.

[18] P. Giudici, "Bayesian data mining, with application to benchmarking and credit scoring," *Appl. Stoch. Model. Bus. Ind.*, no. Jul. 2000, pp. 69–81, 2001.

[19] B. Baesens, M. Egmont-Petersen, R. Castelo, and J. Vanthienen, "Learning Bayesian network classifiers for credit scoring using Markov chain Monte Carlo search," in *Proc. Object Recognit. Supported User Interact. Service Robots*, 2002, vol. 3, pp. 49–52.

[20] A. Mira and P. Tenconi, "Bayesian estimate of credit risk via MCMC with delayed rejection," *Stoch. Anal. Random Fields Appl. IV*, pp. 277–291, 2004.

[21] D. Maltritz and A. Molchanov, "Economic determinants of country credit risk: A Bayesian approach," in *Proc. 12th New Zealand Finance Colloq.*, 2008.

[22] I. Biçer, D. Seviş, and T. Bilgiç, "Bayesian credit scoring model with integration of expert," in *Proc. Int. Conf. 24th Mini EURO Conf. "Continuous Optim. Inf. Technol. Financial Sector*, Feb. 2010, pp. 324–329.

[23] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 534–545, Jan. 2010.

[24] J. L. Zhang and W. K. Härdle, "The Bayesian additive classification tree applied to credit risk modelling," *Comput. Statist. Data Anal.*, vol. 54, no. 5, pp. 1197–1205, 2010.

[25] R. Mileris, "Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier," *Econ. Manag.*, vol. 15, no. 1, pp. 1078–1084, 2010.

[26] B. Bekiroglu, H. Takci, and U. C. Ekinci, "Bank credit risk analysis with Bayesian network decision," *Int. J. Adv. Eng. Sci. Technol.*, vol. 9, no. 2, pp. 273–279, 2011.

[27] L.-J. Kao, C.-C. Chiu, and F.-Y. Chiu, "A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring," *Knowl.-Based Syst.*, vol. 36, pp. 245–252, 2012.

[28] R. Vedala and B. R. Kumar, "An application of naive bayes classification for credit scoring in E-lending platform," in *Proc. Int. Conf. Data Sci. Eng.*, 2012, pp. 81–84.

[29] J. Yang and Y. Zhou, "Credit risk spillovers among financial institutions around the global credit crisis: Firm-level evidence," *Manage. Sci.*, vol. 59, no. 10, pp. 2343–2359, 2013.

[30] E. Y. Li, "Artificial neural networks and their business applications," *Inf. Manag.*, vol. 27, no. 5, pp. 303–313, Nov. 1994.

[31] D. West, "Neural network credit scoring models," *Comput. Oper. Res.*, vol. 27, no. 11–12, pp. 1131–1152, 2000.

[32] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," *IEEE Trans. Neural Netw.*, vol. 12, no. 4, pp. 929–935, Jul. 2001.

[33] R. Malhotra and D. K. Malhotra, "Evaluating consumer loans using neural networks," *Omega*, vol. 31, no. 2, pp. 83–96, 2003.

[34] D. Witkowska, W. Kaminski, K. Kompa, and I. Staniec, "Neural networks as a supporting tool in credit granting procedure," *Inf. Technol. Econ. Manag.*, vol. 2, no. 1, 2004.

[35] S. Wang, "Forecasting foreign exchange rates with artificial neural networks: a review," *Int. J. Inf. Technol. Decis. Mak.*, vol. 3, no. 1, pp. 145–165, 2004.

[36] D. West, S. Dellana, and J. Qian, "Neural network ensemble strategies for financial decision applications," *Comput. Oper. Res.*, vol. 32, no. 10, pp. 2543–2559, Oct. 2005.

[37] B. Baesens, T. Van Gestel, M. Stepanova, and J. Vanthienen, "Neural network survival analysis for personal loan data," *J. Oper. Res. Soc.*, vol. 9, no. 56, pp. 1089–1098, 2005.

[38] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 2–17, Jan. 2009.

[39] M.-C. Tsai, S.-P. Lin, C.-C. Cheng, and Y.-P. Lin, "The consumer loan default predicting model—An application of DEA–DA and neural network," *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11682–11690, Nov. 2009.

[40] P. J. G. Lisboa *et al.*, "Partial logistic artificial neural network for competing risks regularized with automatic relevance determination," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1403–1416, Sep. 2009.

[41] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6233–6239, 2010.

[42] J. H. M. Tah and V. Carr, "A proposal for construction project risk assessment using fuzzy logic," *Constr. Manag. Econ.*, vol. 18, no. 4, pp. 491–500, 2000.

[43] F. Hoffmann, B. Baesens, J. Martens, F. Put, and J. Vanthienen, "Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring," *Int. J. Intell. Syst.*, vol. 17, no. 11, pp. 1067–1083, 2002.

[44] B. Kogut, J. P. MacDuffie, and C. Ragin, "Prototypes and strategy: Assigning causal credit using fuzzy sets," *Eur. Manag. Rev.*, vol. 1, no. 2, pp. 114–131, 2004.

[45] T.-C. Tang and L.-C. Chi, "Predicting multilateral trade credit risks: comparisons of logit and fuzzy logic models using ROC curve analysis," *Expert Syst. Appl.*, vol. 28, no. 3, pp. 547–556, Apr. 2005.

[46] A. Laha, "Developing credit scoring models with SOM and fuzzy rule based k -NN classifiers," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2006, pp. 692–698.

[47] F. Hoffmann, B. Baesens, C. Mues, T. Van Gestel, and J. Vanthienen, "Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms," *Eur. J. Oper. Res.*, vol. 177, no. 1, pp. 540–555, 2007.

[48] G. Bojadziev and M. Bojadziev, *Fuzzy Logic for Business, Finance, and Management*. Singapore: World Scientific, 2007.

[49] A. Laha, "Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring," *Adv. Eng. Inform.*, vol. 21, no. 3, pp. 281–291, Jul. 2007.

[50] A. Lahsasna, R. N. Ainon, and T. Y. Wah, "Credit risk evaluation decision modeling through optimized fuzzy classifier," in *Proc. Int. Symp. Inf. Technol.*, 2008, pp. 1–7.

[51] L. Yu, S. Wang, and K. K. Lai, "An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring," *Eur. J. Oper. Res.*, vol. 195, no. 3, pp. 942–959, Jun. 2009.

[52] A. Lahsasna, R. N. Ainon, and T. Y. Wah, "Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier," *Maejo Int. J. Sci. Technol.*, vol. 4, no. 1, pp. 136–158, 2010.

[53] S. D. Bekiros, "Fuzzy adaptive decision-making for boundedly rational traders in speculative stock markets," *Eur. J. Oper. Res.*, vol. 202, no. 1, pp. 285–293, 2010.

[54] D. Zhang, X. Zhou, S. C. H. Leung, and J. Zheng, "Vertical bagging decision trees model for credit scoring," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 7838–7843, Dec. 2010.

[55] J. Zurada, "Could decision trees improve the classification accuracy and interpretability of loan granting decisions?," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.*, 2010, pp. 1–9.

[56] W.. Li and J.. Liao, "An empirical study on credit scoring model for credit card by using data mining technology," in *Proc. 7th Int. Conf. Comput. Intell. Secur.*, 2011, pp. 1279–1282.

[57] S. Tuffery, *Data Mining and Statistics for Decision-Making*. Hoboken, NJ, USA: Wiley, 2011.

[58] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13274–13283, Sep. 2011.

[59] V. Chorniy and G. A. Bayesian, "Networks and stochastic factor models," *SSRN Electron. J*, vol. 2015, pp. 1–21, 2015. Available at SSRN: https://ssrn.com/abstract = 2688324 or http://dx.doi.org/10.2139/ssrn.2688324

[60] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," *Knowl.-Based Syst.*, vol. 26, pp. 61–68, 2012.

[61] L. G. Kabari and E. O. Nwachukwu, "Credit risk evaluating system using decision tree – neuro based model," *Int. J. Eng. Res. Technol.*, vol. 2, no. 6, pp. 2738–2745, 2013.

[62] V. Garcı, J. S. Sa, and A. I. Marque, "A literature review on the application of evolutionary computing to credit scoring," *J. Oper. Res. Soc.*, vol. 9, pp. 1384–1399, 2013.

[63] U. Bhuvaneswari, P. J. D. Paul, and S. Sahu, "Financial risk modelling in vehicle credit portfolio," in *Proc. Int. Conf. Data Mining Intell. Comput.*, 2014, pp. 1–8.

[64] G. Lin, C. Shen, Q. Shi, A. Van Den Hengel, and D. Suter, "Fast supervised hashing with decision trees for high-dimensional data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1971–1978.

[65] I. T. Van Gestel, B. Baesens, I. J. Garcia, and P. Van Dijcke, "A support vector machine approach to credit scoring," *Bank En Financiewezen*, vol. 2, pp. 73–82, 2003.

[66] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study," *Decis. Support Syst.*, vol. 37, no. 4, pp. 543–558, Sep. 2004.

[67] P.-Y. Hao, M.-S. Lin, and L.-B. Tsai, "A new support vector machine with fuzzy hyper-plane and its application to evaluate credit risk," in *Proc. 8th Int. Conf. Intell. Syst. Des. Appl.*, 2008, pp. 83–88.

[68] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 6, pp. 820–831, Dec. 2005.

[69] S. LI, W. Shiue, and M. Huang, "The evaluation of consumer loans using support vector machines," *Expert Syst. Appl.*, vol. 30, no. 4, pp. 772–782, May 2006.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[70] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1466–1476, 2007.

[71] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines," *Expert Syst. Appl.*, vol. 33, no. 4, pp. 847–856, Nov. 2007.

[72] Y. Wang, "Building credit scoring systems based on Support-Based support vector machine ensemble," in *Proc. 4th Int. Conf. Nat. Comput.*, 2008, vol. 5, pp. 323–327.

[73] W. Chen, C. Ma, and L. Ma, "Mining the customer credit using hybrid support vector machine technique," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7611–7616, May 2009.

[74] X. Xu, C. Zhou, and Z. Wang, "Credit scoring algorithm based on link analysis ranking with support vector machine," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 2625–2632, Mar. 2009.

[75] L. Zhou and K. K. Lai, "Multi-agent ensemble models based on weighted least square SVM for credit risk assessment," in *Proc. WRI Global Congr. Intell. Syst.*, 2009, pp. 559–563.

[76] Z. M. Z. Min, "Credit risk assessment based on fuzzy SVM and principal component analysis," in *Proc. Int. Conf. Web Inf. Syst. Min.*, 2009, pp. 125–127.

[77] J. Galindo and P. Tamayo, "Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications," *Comput. Econ.*, vol. 15, no. 1/2, pp. 107–143, 2000.

[78] M. Zekic-Susac, N. Sarlija, and M. Bensic, "Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models," in *Proc. 26th Int. Conf. Inf. Technol. Interfaces*, Jul. 2004, pp. 265–270.

[79] A. Bandyopadhyay, "Predicting probability of default of Indian corporate bonds: Logistic and Z-score model approaches," *J. Risk Finance*, vol. 7, no. 3, pp. 255–272, 2006.

[80] M. A. Gouvea, "Credit risk analysis applying logistic regression, neural networks an genetic algorithms models," in *Proc. 18th Annu. Conf. Prod. Oper. Manage. Soc.*, 2007, Paper 007-0210.

[81] S. Vlah, N. Sarlija, K. Soric, and V. Vojvodić Rosenzweig, "Logistic regression and multicriteria decision making in credit scoring," in *Proc. 10th Int. Symp. Oper. Res. Slovenia*, 2009, pp. 175–184.

[82] S. Figini and P. Uberti, "Model assessment for predictive classification models," *Commun. Statist. - Theory Methods*, vol. 39, no. 18, pp. 3238–3244, 2010.

[83] M. Hamadi and A. K. Abdelmoula, "Credit-risk evaluation of a Tunisian commercial bank: Logistic regression vs neural network modelling," *Int. J. Accounting Inf. Manag.*, vol. 19, no. 2, Jun. 2011, doi: 10.1108/ijaim.2011.36619baa.005.

[84] N. Nikolic, N. Zarkic-Joksimovic, D. Stojanovski, and I. Joksimovic, "The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statements," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5932–5944, Nov. 2013.

[85] P. H. Ferreira, F. Louzada, and C. Diniz, "Credit scoring modeling with state-dependent sample selection: A comparison study with the usual logistic modeling," *Sci. Electron. Libr. Online*, vol. 35, no. 1, pp. 39–56, 2015.

[86] I. Triki, "Credit scoring models for a Tunisian microfinance institution: Comparison between artificial neural network and logistic regression," *Rev. Econ. Finance*, vol. 6, pp. 61–78, 2016.

[87] S. Y. Sohn, D. H. Kim, and J. H. Yoon, "Technology credit scoring model with fuzzy logistic regression," *Appl. Soft Comput.*, vol. 43, pp. 150–158, Jun. 2016.

[88] Z. Yulia, O. Krasotkina, and V. Mottl, "Sparse logistic regression with supervised selectivity for predictors selection in credit scoring," in *Proc. 7th Symp. Inf. Commun. Technol.*, 2016, pp. 167–172.

[89] V. Srinivisan and Y. H. Kim, "Credit granting a comparative analysis of classificatory procedures," *J. Finance*, vol. 42, no. 3, pp. 655–683, 1987.

[90] V. S. Desai, J. N. Crook, and G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," *Eur. J. Oper. Res.*, vol. 95, no. 1, pp. 24–37, Nov. 1996.

[91] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: A review," *R. Statist. Soc.*, pp. 523–541, 1997.

[92] N.-C. Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customers," *Expert Syst. Appl.*, vol. 27, no. 4, pp. 623–633, Nov. 2004.

[93] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, and C.-J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," *Comput. Statist. Data Anal.*, vol. 50, no. 4, pp. 1113–1130, Feb. 2006.

[94] M.-H. Jiang and X.-C. Yuan, "Personal credit scoring model of non-linear combining forecast based on GP," in *Proc. 3rd Int. Conf. Nat. Comput.*, 2007, vol. 4, pp. 408–414.

[95] A. P. Sinha and H. Zhao, "Incorporating domain knowledge into data mining classifiers: An application in indirect lending," *Decis. Support Syst.*, vol. 46, no. 1, pp. 287–299, Dec. 2008.

[96] D. Martens, T. Van Gestel, M. De Backer, R. Haesen, J. Vanthienen, and B. Baesens, "Credit rating prediction using ant colony optimization," *J. Oper. Res. Soc.*, vol. 61, no. 4, pp. 561–573, 2010.

[97] P. A. Smith, "Methods to estimate losses using linear regression analysis-linear regression analysis can refine and improve ALLL calculations and provide a framework for understanding portfolio risk," *RMA J.*, pp. 60–66, Nov. 2011.

[98] G. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst. Man, Cybern. Part B, Cybern.*, vol. 42, no. 2, pp. 513–29, Apr. 2012.

[99] C. Bravo, S. Maldonado, and R. Weber, "Granting and managing loans for micro-entrepreneurs: New developments and practical experiences," *Eur. J. Oper. Res.*, vol. 227, no. 2, pp. 358–366, 2013.

[100] K. D. Majeske and T. W. Lauer, "The bank loan approval decision from multiple perspectives," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1591–1598, 2013.

[101] L. C. Thomas, "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers," *Int. J. Forecast.*, vol. 16, no. 2, pp. 149–172, 2000.

[102] M. Hossain, M. A. Malek, and N. C. Das, "Tenant farmers' access to credit and extension services: BRAC tenant farmer development project in Bangladesh," *Center Int. Res. Japanese Economy*, vol. 7, no. 40, pp. 7–9, 2014.

[103] S. M. Sadatrasou, M. R. Gholamian, and K. Shahanaghi, "An application of data mining classification and bi-level programming for optimal credit allocation," *Decis. Sci. Lett.*, vol. 4, pp. 35–50, 2015.

[104] J. Huang, "Feature selection in credit scoring—A quadratic programming approach solving with bisection method based on Tabu search," Ph.D. dissertation, Texas A&M International Univ., Laredo, TX, USA, 2015.

[105] A. Capponi, "Systemic risk, policies, and data needs," INFORMS Tut. Oper. Res. Forthcom., 2016, pp. 1–23.

[106] R. Zepeda, "Enhancing Islamic finance through risk benchmarking," *Capco Inst. J. Financial Transformation*, vol. 38, pp. 17–34, 2013.

[107] A. King, J. C. Liechty, C. V. Rossi, and C. Taylor, "Frameworks for systemic risk monitoring," in *Handbook of Financial Data and Risk Information I*. Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 105–147.

[108] S. Aliakbari, *Corporate Credit Risk and Economic Performance*. London, U.K.: Brunel University London, 2016.

[109] Y. Y. Haimes, *Risk Modeling, Assessment, and Management*. Hoboken, NJ, USA: Wiley, 2015.

[110] A. Ç. Tolga, F. Tuysuz, and C. Kahraman, "A fuzzy multi-criteria decision analysis approach for retail location selection," *Int. J. Inf. Technol. Decis. Making*, vol. 12, no. 4, pp. 729–755, 2013.

[111] Y. H. Ju and S. Y. Sohn, "Updating a credit-scoring model based on new attributes without realization of actual data," *Eur. J. Oper. Res.*, vol. 234, no. 1, pp. 119–126, 2014.

[112] L. Hatzilygeroudis and J. Prentzas, "Fuzzy and neuro-symbolic approaches in personal credit scoring: Assessment of bank loan applicants," in *Innovations in Intelligent Machines-4*. Berlin, Germany: Springer, 2014, pp. 319–339.

[113] A. Capotorti and E. Barbanera, "Credit scoring analysis using a fuzzy probabilistic rough set model," *Comput. Statist. Data Anal.*, vol. 56, no. 4, pp. 981–994, 2012.

[114] H. A. Abdou, "Genetic programming for credit scoring: The case of Egyptian public sector banks," *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11402–11417, Nov. 2009.

[115] K. Falangis and J. J. Glen, "Heuristics for feature selection in mathematical programming discriminant analysis models," *J. Oper. Res. Soc.*, vol. 61, no. 5, pp. 804–812, 2010.

[116] S. Efromovich, "Oracle inequality for conditional density estimation and an actuarial example," *Ann. Inst. Statist. Math.*, vol. 62, no. 2, pp. 249–275, 2010.

[117] A. V Thakor, "The financial crisis of 2007 – 2009: Why did it happen and what did we learn?" *Rev. Corp. Finance Stud.*, vol. 4, no. 2, pp. 1–51, 2015.

[118] A. Bandyopadhyay, *Managing Portfolio Credit Risk in Banks*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[119] A. Moro, S. Nolte, and A. Diaz, "Entrepreneur's wealth, firm performance and cost of capital: A Bayesian approach to the capital structure of entrepreneurial ventures," *SSRN Electron. J.*, vol. 2014, pp. 1–39, 2014, Available at SSRN: https://ssrn.com/abstract=1967697 or http://dx.doi.org/10.2139/ssrn.1967697
[120] D. Kumar, A. Hossain, and M. C. Gope, "Role of micro credit program in empowering rural women in Bangladesh: A study on Grameen Bank Bangladesh Limited," *Asian Bus. Rev.*, vol. 3, no. 6, pp. 114–120, 2013.
[121] J. Yang and Y. Zhou, "Credit risk spillovers among financial institutions around the global credit crisis: Firm-Level evidence," *Manage. Sci.*, vol. 59, no. 10, pp. 2343–2359, 2013.
[122] D. D. Wu, D. L. Olson, and C. Luo, "A decision support approach for accounts receivable risk management," *IEEE Trans Syst Man, Cybern Syst*, vol. 44, no. 12, pp. 1624–32, Dec. 2014, doi: 10.1109/TSMC.2014.2318020.
[123] T. Macaulay, "Critical infrastructure: Understanding its component parts, vulnerabilities," in *Operating Risks, and Interdependencies*. Boca Raton, FL, USA: CRC Press, 2010.
[124] J. Lam, *Enterprise Risk Management: From Incentives to Controls*. Hoboken, NJ, USA: Wiley, 2014.
[125] M. Imran, M. Zulfqar, H. Rasheed, S. Tayyaba, W. Ashraf, and Z. Ahmad, "Fuzzy logic based flow controller of dam gates," *J. Eng. Res. Technol.*, vol. 1, no. 3, pp. 83–90, 2014.
[126] A. M. Grace and S. O. Williams, "Comparative analysis of neural network and fuzzy logic techniques in credit risk evaluation," *Int. J. Intell. Inf. Technol.*, vol. 12, no. 1, pp. 1–16, 2016.
[127] D. Zhang, W. Xu, Y. Zhu, and X. Zhang, "Can sentiment analysis help mimic decision-making process of loan granting? A novel credit risk evaluation approach using GMKL mode," in *Proc. 48th Hawaii Int. Conf. Syst. Sci.*, 2015, pp. 949–958.
[128] R. Pears and R. Oetama, "Boosting prediction accuracy of bad payments in financial credit applications," in *Rare Association Rule Mining and Knowledge Discovery*. Hershey, PA, USA: IGI Global, 2010, pp. 255–269.
[129] S. Finlay, "Are we modelling the right thing? The impact of incorrect problem specification in credit scoring," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9065–9071, 2009.
[130] R. A. Baxter, M. Gawler, and R. Ang, "Predictive model of insolvency risk for Australian corporations," in *Proc. 6th Australian Conf. Data Mining Anal.*, 2007, vol. 70, pp. 21–28.
[131] L. Feng, Y. Yao, and B. Jin, "Research on credit scoring model with SVM for network management," *J. Comput. Inf. Syst.*, vol. 6, no. 11, pp. 3567–3574, 2010.
[132] D. Sweeney, D. Anderson, and T. Williams, *Statistics for Business and Economics*, 7th ed. London, U.K.: Thomson Learning EMEA, 2007.
[133] A. C. Antonakis and M. E. Sfakianakis, "Assessing naïve Bayes as a method for screening credit applicants," *J. Appl. Statist.*, vol. 36, no. 5, pp. 537–545, 2009.
[134] M. Hamdaoui, "Financial liberalization and systemic banking crises: A meta-analysis," *Int. Econ.*, vol. 152, pp. 26–54, Dec. 2017.
[135] K. Bijak and L. C. Thomas, "Does segmentation always improve model performance in credit scoring?," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 2433–2442, 2012.
[136] R. DeYoung, W. S. Frame, D. Glennon, and P. Nigro, "The information revolution and small business lending: The missing evidence," *J. Financial Serv. Res.*, vol. 39, no. 1–2, pp. 19–33, 2011.
[137] S. Finlay, "Multiple classifier architectures and their application to credit risk assessment," *Eur. J. Oper. Res.*, vol. 210, no. 2, pp. 368–378, Apr. 2011.

Germanno Teles received the master's degree in applied computing from the State University of Ceará, Fortaleza, Brazil, in 2013. He is currently working toward the Ph.D. degree in informatics engineering with Instituto de Telecomunicações, University of Beira Interior, Covilha, Portugal.

He is an IT Specialist with the Bank of Northeast, Fortaleza, Brazil. He is a member of the Next-Generation Networks and Applications Group supervised by Professor J. J. P. C. Rodrigues.

Joel J. P. C. Rodrigues (S'01–M'06–SM'06) received the five-year B.Sc. degree (licentiate) in informatics engineering from the University of Coimbra, Coimbra, Portugal, in 1995, the M.Sc. and Ph.D. degrees in informatics engineering degree from the Universidade da Beira Interior (UBI), Covilha, Portugal, in 2002 and 2006, respectively, the Habilitation in computer science and engineering from the University of Haute Alsace, Mulhouse, France, in 2014, and the Academic Title of Aggregated Professor in Informatics Engineering from UBI in 2015.

He is a Professor with the Federal University of Piauí, Teresina, Brazil, and a Senior Researcher with Instituto de Telecomunicações, Covilha, Portugal. He has authored or coauthored more than 750 papers in refereed international journals and conferences, three books, two patents, and one ITU-T recommendation.

Prof. Rodrigues is the Editor-in-Chief of an international journal and an Editorial Board Member of several journals. He is the Leader of the Internet of Things Research Group (CNPq), the Director for Conference Development—the IEEE ComSoc Board of Governors, the IEEE Distinguished Lecturer, the Technical Activities Committee Chair of the IEEE ComSoc Latin America Region Board, the Past Chair of the IEEE ComSoc TCs on eHealth and on Communications Software, and the Steering Committee Member of the IEEE Life Sciences Technical Community.

Kashif Saleem received the Ph.D. degree in electrical engineering and the M.E. degree in electrical engineering—electronics and telecommunication from Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2007 and 2011, respectively.

He is currently an Assistant Professor with the Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia. He has authored several research publications and handles ICT related funded research projects in the Middle East and European Union. His research interests include ubiquitous computing, biologically inspired algorithms, the Internet of Things, machine-to-machine communications, wireless mobile networks, wireless sensor networks, and mobile *ad hoc* networks.

Sergei A. Kozlov received the graduate engineer degree (with Hons.) in quantum electronics from Leningrad Institute of Fine Mechanics and Optics (now ITMO University), Leningrad, Russia, in 1982, and the Ph.D. and Dr. Sci. Phys. and Maths. degrees from the Saint Petersburg State University, Saint Petersburg, Russia, in 1986 and in 1997, respectively.

Since 2002, he has been a Full Professor, the Head of the Department of Photonics and Optoinformatics, and the Dean of the Faculty of Photonics and Optoinformatics, ITMO University. Since 2013, he has been the Head of the International Institute of Photonics and Optoinformatics, ITMO University. He has authored 250 articles in various publications.

Chapter 3

Machine Learning and Decision Support System on Credit Scoring

This chapter consists in the following paper:

Machine Learning and Decision Support System on Credit Scoring

Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A.L. Rabêlo, Kashif Saleem and Sergei Kozlov

Neural Computing and Applications, Springer, ISSN:1433-3058, vol.32, n.14, pp.9809-9826, October 2019.

DOI: doi.org/10.1007/s00521-019-04537-7

According to Journal Citation Reports published by Thomson Reuters in 2019, this journal scored ISI journal performance metrics as follows:

ISI Impact Factor (2019): 4.774

Journal Ranking (2019): 23/136 (Computer Science, Artificial Intelligence)

Neural Computing and Applications
https://doi.org/10.1007/s00521-019-04537-7

REVIEW ARTICLE

Machine learning and decision support system on credit scoring

Gernmanno Teles[1] · Joel J. P. C. Rodrigues[1,2,3,4] · Kashif Saleem[5] · Sergei Kozlov[4] · Ricardo A. L. Rabêlo[2]

Received: 8 February 2019 / Accepted: 5 October 2019

Abstract
Among the numerous alternatives used in the world of risk balance, it highlights the provision of guarantees in the formalization of credit agreements. The objective of this paper is to compare the achievement of fuzzy sets with that of artificial neural network-based decision trees on credit scoring to predict the recovered value using a sample of 1890 borrowers. Comparing with fuzzy logic, the decision analytic approach can more easily present the outcomes of the analysis. On the other hand, fuzzy logic makes some implicit assumptions that may make it even harder for credit-grantors to follow the logical decision-making process. This paper leads an initial study of collateral as a variable in the calculation of the credit scoring. The study concludes that the two models make modelling of uncertainty in the credit scoring process possible. Although more difficult to implement, fuzzy logic is more accurate for modelling the uncertainty. However, the decision tree model is more favourable to the presentation of the problem.

Keywords Machine learning · Decision trees · Fuzzy logic · Credit scoring · Performance evaluation

1 Introduction

Decision-making is an important factor in achieving success in selecting borrowers using large amounts of information and knowledge. Identifying less risky borrowers is crucial in successful credit scoring. The financial sector is not only volatile but also very competitive. The probability of failure to deliver the desired outcomes is high for individual borrowers, and the lenders must take appropriate steps to confront the situation and mitigate the associated risks to achieve favourable debt repayment outcomes [1, 2]. The selection process for the borrowers should lead to the identification of safer borrowers who have the ability and willingness to pay back debt within the set repayment period. Some credit scoring methods currently in existence are criticized for being inadequate in addressing the borrower's ability to repay debt with minimal risk within the given time [3].

Probability models are widely used in the quantification and assessment of risk. These models have become integral in the informed decision-making processes related to uncertainty in many application areas [4–12]. However, such frameworks based on the conventional theory may not be able to describe some uncertainties in the most appropriate manner because imprecise data, insufficient experience datasets, and the complex cause-effect relationships make it difficult to assess the different risk types using only the conventional probability models [13–15]. Of particular concern is the misunderstanding of the cause of the risk and its characteristics [7, 16].

Credit scoring is a collection of many different tasks, processes, and requirements that need to be considered in their entirety [2, 17–23]. Consequently, it is difficult and arduous to make credit lending decisions in such environments [24–33]. Therefore, mechanisms that help to characterize such complex scenarios are needed. The literature suggests that the application of multiple criteria for decision-making can facilitate the resolution of these issues [9]. Other models, such as fuzzy logic, artificial neural networks, and decision tree models explicitly consider the

✉ Gernmanno Teles
germanno.teles@ubi.pt

[1] Instituto de Telecomunicações, Universidade da Beira Interior, 6201-001 Covilhã, Portugal

[2] Federal University of Piauí, Teresina-PI, Brazil

[3] College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh 12372, Saudi Arabia

[4] ITMO University, St. Petersburg 197101, Russia

[5] Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh 11653, Saudi Arabia

Published online: 24 October 2019

underlying relationships and recognize the uncertainties (such as operational risks). In addition to the multi-criteria methods, additional complementary tools such as fuzzy sets or numerical simulations are increasingly being used in the credit scoring process. These tools have been applied in lender decision-making as far as dealing with the aspects of financial risks are concerned. The capability of these tools to deal with uncertainty enables the credit grantor to deal with the issues in a manner that conventional methods would not [6, 34]. The advantage of our approach is that it uses features that are independent to use collateral as a parameter for calculating risk.

The concept of Fuzzy sets also known as the possibility theory was first introduced by Lukasiewicz (cited in Rescher [35]. Later, Zadeh [36] extended the work on this theory to identify Fuzzy sets as a mathematical model for the characterization and quantification of uncertainty [37]. Since then, the fuzzy set theory has been used as a method for modelling of risk in many applications, including credit scoring in the financial sector [32]. The usefulness of this model is such that it can be used to characterize and quantify uncertainty using fuzzy sets in the absence of sufficient data to estimate uncertainty through the conventional statistical estimation of frequencies [38]. The basis of the fuzzy set theory is on a group of elements or data that share some common characteristics within their memberships [39, 40]. As sufficient data on borrowers may not be readily available to the lender, the application of the fuzzy set theory can play an essential role in the quantification of uncertainty in borrower selection decisions.

Then, this work aims to investigate study of collateral as a variable in the calculation of the credit scoring applied to systems that are using credit operations. The main contribution of this work includes a compare the performance of fuzzy sets with that of artificial neural network-based decision trees on credit scoring to predict the recovered value using a sample of 1890 borrowers.

The paper is organized as follows. Section 2 presents theoretical background, introducing the related works and discussing used fuzzy methods and decision trees procedures on classifying credit scoring approach. Section 3 presents results analysis and shows which model is superior in terms of accuracy and which is superior in terms of fitting the data correctly. Section 4 performs the discussion about the issues of fuzzy and decision tree models bring forward a comparison of those two models. Finally, Section 5 provides the conclusion and suggestions for further works.

2 Theoretical background

In the decision trees technique, a set of rules presented as a tree are used to make decisions [41]. For example, one can build a classification tree for credit risk based on a person's income, age, among other parameters. The benefit of this model is that, unlike most of the credibility models, the tree expresses the reasoning process behind the framework. The usual algorithm for decision tree building includes classifying sets under a root node by assigning all the training data to the best splitting attribute (see Fig. 1).

The primary objective of a decision tree is to divide a group of data into fewer portions. On a qualitative data used to build a person's income, for instance, the root of the tree asks if Income < 92.5. If the answer is "yes", it move to the next child, and it pass to the left child of the node; if "no", it moves to the right. Proceeding in this way, it finally succeeds at a final node.

The data comprehend a piece of information on the credit history of customers of a lender. The population consists of existing customer accounts of the lender. A sample of customers is chosen randomly from the data available on their performance. The main characteristics of the customer information are the recovered and whether the value rates are overdue. The data used in the analysis are obtained from the lender's database.

The underlying analysis techniques include fuzzy logic- and artificial neural network-based decision trees to predict credit recovery using a real dataset of 1890 records from a bank in Brazil.

2.1 Credit scoring using fuzzy set theory and fuzzy logic

Definition 1 Suppose that (X, Y, Z) is the target system.

X is a limited set object;

$$X = (x_1 + x_2 \ldots, x_n) \tag{1}$$

Y is the object attribute set

$$Y = (y_1 + y_2 \ldots y_n) \tag{2}$$

Z is the target set

$$Z = (z_1 + z_2 \ldots z_n) \tag{3}$$

and the fuzzy target set is Z_i.

Fuzzy logic is applied because y_j and the degree of effect of y_j on target Z_i in the target system are uncertain.

Definition 2 "Generated real objected states target set is defined as target state set z where one or more than one objected attributes are inputted to the system. And the objected states from the math model of the system are defined as expected states" z^0 [10, 31, 42–44].

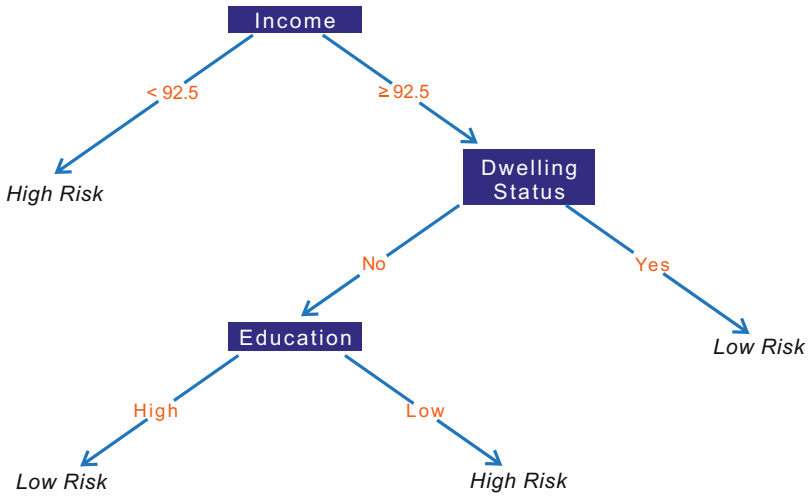


Fig. 1 Example of a decision tree model

The fuzzy relation is

$$R : Y \times Z \rightarrow [0, 1] \tag{4}$$

If there is $(y, z) \in Y \times Z, R$, (y, z) is the degree of relation from attribute y to state z. R is the binary relation from set Y to set Z.

Fuzzy set on Y and fuzzy set on Z are denoted by A and B, respectively. Then R can be expressed as implication relations;

$$R = A \rightarrow B \tag{5}$$

Then,

$$R(y, z) = (A \rightarrow B)(y, z) \tag{6}$$

And the algorithm regulated,

$$(A \rightarrow B)(y, z) = (1 - A(y)) \curlyvee B(z) \tag{7}$$

Definition 3 Suppose $\lambda \in \left[\frac{1}{2}, 1\right]$

(i) if $\exists(y, z)$ makes $R(y, z) \geq \lambda$, the effect relation R to (y, z) is regarded as fuzzy λ true, or as fuzzy λ false.

(ii) if $\forall(y, z)$ makes $R(y, z) \geq \lambda$, the effect relation R is regarded as fuzzy λ true, or as fuzzy λ false.

To calculate the effect characteristic of intersection values y,

$$y = \bigcap_{i=1}^{n} U_i \tag{8}$$

Since Eq. (6) and $\forall U \in E$, membership of fuzzy set A on Y is,

$$Y(y_j) = \begin{cases} 1, & y_j \in U \\ 0, & y_j x \notin U \end{cases} \quad (j = 1, 2 \ldots, m) \tag{9}$$

According to formula (7),

$$\begin{aligned} &(A_i \cap A_j \rightarrow B_i \cup B_j)(y, z) = [1 - (A_i \cap A_j)(y)] \curlyvee [(B_i \cup B_j)(y)] \\ &= [(1 - A_i(y)) \curlyvee B_i(z)] \curlyvee [(1 - A_j(y)) \curlyvee B_j(z)] \\ &= (A_i \rightarrow B_i)(y, z) \curlyvee (A_j \rightarrow B_j)(y, z) \end{aligned} \tag{10}$$

Therefore, $\forall \lambda \in \left[\frac{1}{2}, 1\right]$,

$$\begin{aligned} &(A_i \rightarrow B_i)(u, p) \geq \lambda (A_j \rightarrow B_j)(u, p) \geq \lambda \\ &\Leftrightarrow (A_i \cap A_j \rightarrow B_i \cup B_j)(u, p) \geq \lambda \end{aligned}$$

2.2 Fuzzy data analysis procedure

Case–control matching is a traditional procedure used to join records in the "case" representation with related records in a typically much larger "control" example based on a set of essential variables. So to explains the fuzzy extension command for Statistical Package for the Social Sciences (SPSS) that implements this method and some recent improvements to it, which allows the input of a custom function, is used. First, the data is reduced down to only the variables used [38, 42, 45–48]. It was used on this dataset to reduce the select bias and improve the internal validity.

The Polytomous Universal Model (PLUM), an extension of the general linear model to ordinal categorical data, was used to fit the logistic model predicting the probability of the treatment. It uses contract value, collateral value, main value delay, the balance value, tax rate value, tax interest value, client size, seniority level, per cent used, duration in years, duration in days, and delay in days as predictors of the recovered value.

```
*Fitting logit model via PLUM.
PLUM HalfwayHouse WITH NonViol SidewalkCafe TypeC_D
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5) PCONVERGE(1.0E-6)
SINGULAR(1.0E-8)
  /LINK=LOGIT
  /PRINT=FIT PARAMETER SUMMARY
  /SAVE=ESTPROB.
```

The model is not good but as can be seen, the balance value, tax rate value, tax interest value, client size, seniority level, per cent used, and duration are not associated with credit repayment. Now a custom function with which to restrict matches based on the probability of the treatment and period in months is created. In this case, another file is made in python and named PerFun.py in which the following functions are placed:

```
#These functions are for SPSS's fuzzy case control matching
import math
#period under 12, and caliper within 0.02
def PerFun(demander,supplier):
  dy = math.pow(demander[1] - supplier[1],2)
  dz = math.pow(demander [2] - supplier [2],2)
  period = math.sqrt(dy + dz)
  p = abs(d[0] - s[0]) #diference in month
  if per < 12 and p < 0.02:
    t = 1 #for credit default
  else:
    t = 0
  return t
#period over 12, but under 24
def PerBuf(demander,supplier):
  dx = math.pow(demander[1] - supplier[1],2)
  dy = math.pow(demander[2] - supplier[2],2)
  period = math.sqrt(dx + dy)
  p = abs(d[0] - s[0]) #diference in month
  if period  12 and p < 0.02:
    t = 1 # for credit default
  else:
    t = 0
  return t
```

The fuzzy logic algorithm above returns either the value of 1 for credit default or 0 otherwise [2, 9, 32, 39, 46, 49, 50]. Also, the algorithm takes only a fixed set of vectors from the dataset. The first two elements of the first function PerFun account for the period of repayment, and the last element is the probability of treatment. Next, the function returns the euclidean distance [12, 38]. If the period is under 12 months, it returns the value of 0. The second function sets the boundaries of the period under consideration with values not too far away from 12 to 24 months of repayment default.

The program code is then run in Python as follows.

```
import PerFun

#test case
x = [0,0,0.02]
y = [0,23,0.02]
z = [0,24,0.02]
print PerFun.PerFun(x,y)
print PerFun.PerFun(x,z)
END PROGRAM.
```

To this end, a custom function in order to use the fuzzy command is made. The custom function helps one to do more complicated functions such as the buffer function, which takes the probability of the treatment along with the two spatial points of the period variable. The custom function returns the values for all the predictor variables. To conduct the credit score analysis involves a little more data involvement. The cases and controls of the second data of the just matched credit defaulters are reshaped in the long format before merging with the original.

```
*custom function

FUZZY BY=EST2_1 XMonths YMonths ID=PerID IDVARS=Match1 Match2 Match3 GROUP=Period
CUSTOMFUZZ = "PerFun.PerFun"
   EXACTPRIORITY=FALSE
MATCHGROUPVAR=PGroup
/OPTIONS REPLACEMENT=FALSE SHUFFLE=TRUE MINIMIZEMEMORY=TRUE SEED=10.
```

```
*Reshaping and merging the data back as well as the outcome analysis.

DATASET COPY PerMatch.
DATASET ACTIVATE PerMatch.
SELECT IF Period = 1.
CASES /MAKE PerID FROM PerID Match1 Match2 Match3
        /INDEX Type
        /KEEP PGroup.
```

```
* merging with the original.
SORT CASES BY PerID.
MATCH FILES FILE = *
  /TABLE = 'Period_Data'
  /BY PerID.

* merging with the original
SORT CASES BY PerID.
MATCH FILES FILE = *
  /TABLE = 'Period_Data'
  /BY PerID.
*Analysis
T-TEST GROUPS=Period(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Viol
  /CRITERIA=CI(.95).
```

2.3 Decision trees procedure

The customer data of a lending institution are imported into RapidMiner Software to produce a decision tree for addressing the current problem. RapidMiner is a data science software program built by the organization of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics [51]. The steps for applying decision trees for credit solutions include;

1. Importing the data from a spreadsheet,
2. Splitting data into training and testing samples,
3. Training the decision tree, and
4. Applying the model to the testing samples and evaluating the performance.

Importing the data is easy using RapidMiner's built-in import operator to automatically load the data from a spreadsheet into the software interface (Fig. 2).

As with all data modelling techniques, the dataset must be split into training and testing sets to ensure that the model works well on both sets [11, 13, 50]. The standard practice is to split the available data into a training set and a validation set [46, 52, 53]. "Typically 70–80% of the original data of the dataset is split into the training set and the remainder is set aside for validation" [26, 30, 54]. So, in this step, stratified sampling is chosen with a split ratio of 0.7 (70% training) to ensure that all variables have similar distributions of class values. Finally, the XL operator output is connected to the Split Validation operator input (the process causes some errors to appear because the process is still incomplete) (Fig. 3).

As discussed earlier, five simple steps are used to create decision trees by developing information contained in the reduced data. The reduced data reduces the uncertainties contained in the data before splitting thereby increasing information through classification. So, the greatest increase in information is achieved by reducing data as much as possible. The three main options available in RapidMiner Studio for decision tree splitting are.

1. *Information gain* is calculated as the difference between the information before and after splitting the data. The problem with this parameter arises when the variables are very few with a large number of classes. In this case, the variables have the tendency of becoming root nodes and it becomes a challenge when the case is extreme.
2. *Gain ratio* This is a better parameter than the information gain whereby it overcomes the problem with information gain because it considers the probable number of branches before splitting the data.
3. *Gini index* provides supplementary information to the gain ratio.

However, there are other parameters such as the "minimal gain" value whose default value is 0.1 but it can be anything from 0 upwards.

The size of the dataset also determines the other decision tree parameters such as the "minimal size for a split", "maximal depth", and "minimal leaf size" indices.

Finally, the training ports and the model ports are joined together (Fig. 4).

The last step involves checking the model validity using performance operators such as accuracy, precision, recall, receiver operating characteristic (ROC) and area under curve (AUC) charts. The common methods of evaluating the performance of classification models include confusion matrices and Gain and Lift charts. A confusion matrix is a table that is able to compare the model predicted and actual classes from the labelled data within the validation set. The main evaluation criteria in the confusion matrix are accuracy, sensitivity, and specificity. Whereas accuracy is the indicator of overall model effectiveness, sensitivity measures the rate of true positives and specificity the rate of false positives (Fig. 5).

3 Results analysis

The two models explored in this study show that both can be applied effectively to credit scoring. Decision trees can be used for the classification where it has discrete data (continuous or nominal values). In the case where both the input attributes and the output decision are continuous, decision trees can be used to generate fuzzy rules. For

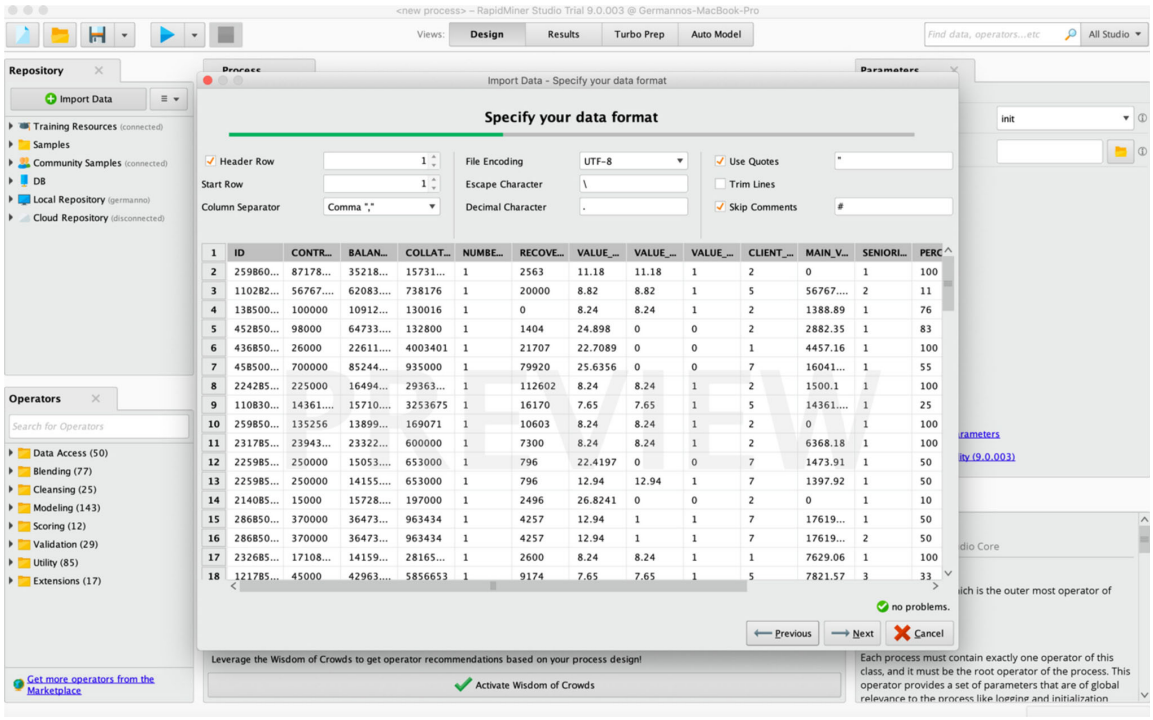


Fig. 2 Step 1: showing how to import external data

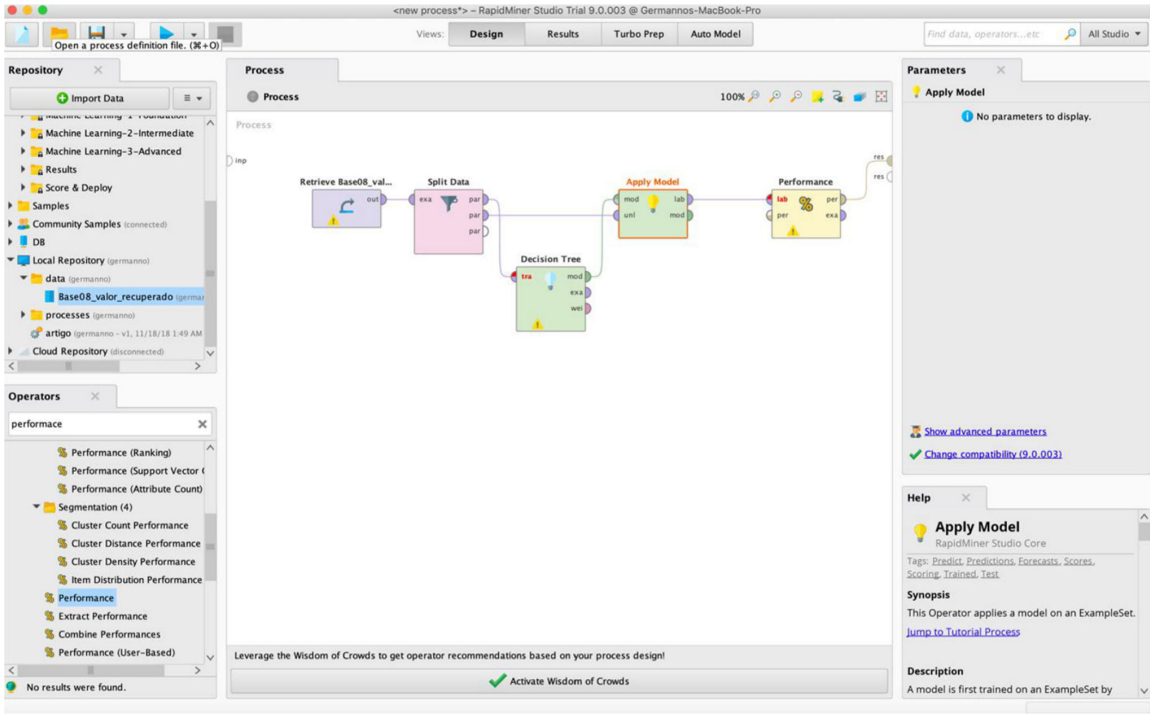


Fig. 3 Step 2: splitting and validating data

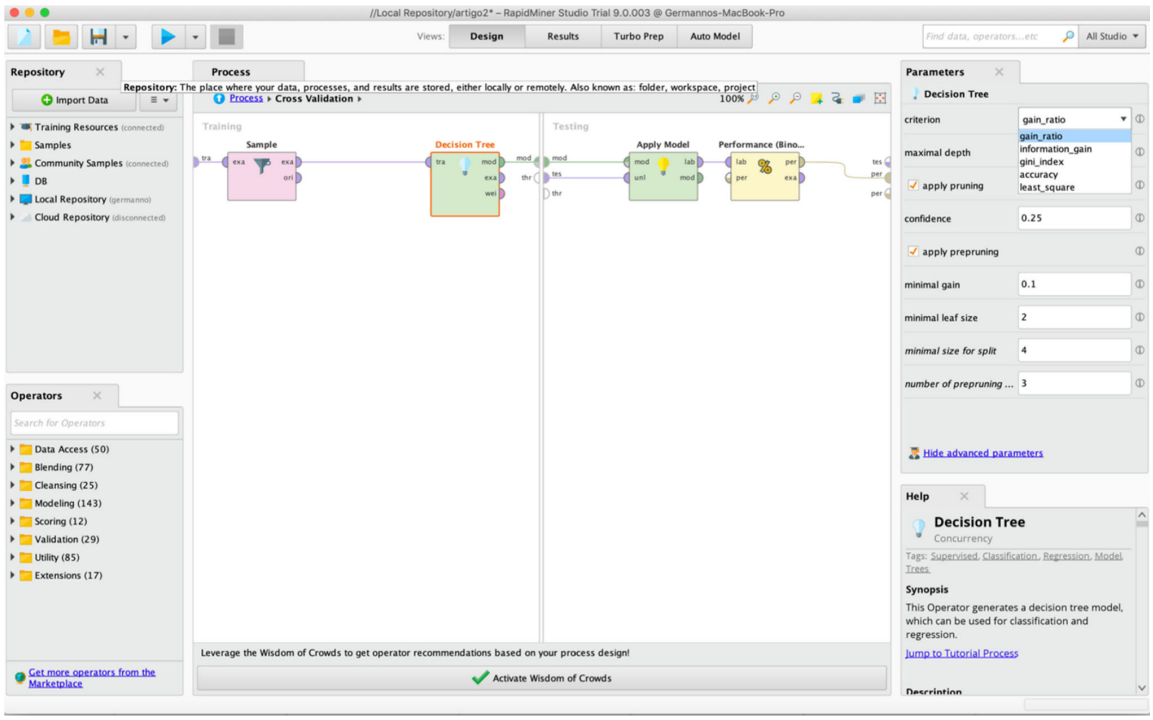


Fig. 4 Step 3: setting up the decision tree parameters

example, if the sets provide output values that provide at most 2 sets, each training example will belong to 1 or 2 output classes or fuzzy sets.

Table 1 includes the Chi-Square goodness of fit test of 392.741 on 1 df (degrees of freedom), significant beyond .001. Therefore, the null hypothesis that adding the variables to the model has not significantly increased our ability to predict defaulters is rejected. Table 2 is the model summary that includes the Pseudo-R^2. It sees that the − 2 log likelihood statistic is .000. The − 2 log likelihood measures how poorly the model predicts the decisions whereby the smaller the statistic, the better the model. Adding the variables reduced the − 2 log likelihood statistic by 392.741 − .000 = 392.741. As such, the result shows that the model is superior in terms of accuracy great. It can also see that Nagelkerke's R^2 is 1.00, which indicates that the model is superior in terms of fits the data perfectly. Cox and Snell's R^2 is .26, which can be explained as 26% probability of the prediction is explained by the logistic model.

Table 3 illustrates "The cut value is .500" whereby the likelihood of a case being classified into the "default" category is greater than .500; otherwise, the case is classified as in the "no default" category. The classification tables show the percentage accuracy in classification (PAC) of cases correctly classified as "no default" with the independent variables added. It also indicates the sensitivity, which is the percentage of cases that had the observed characteristic that was correctly predicted by the model. It also includes the specificity, which is the percentage of cases that did not have the marked characteristic and were also accurately predicted as not having the observed feature. Table 3 shows that 1278/1278 = 100% of the subjects in default were observed. In other words, the sensitivity of the prediction was 100%. It also sees that 45/45 = 100% were correctly classified into the "no default category." Thus, the specificity of prediction of non-occurrences correctly predicted was 100%.

Table 4 shows the contribution of each predictor variable to the model and its statistical significance. The results show that all the independent variables did not add significantly to the model ($p > .05$). Although insignificant, the largest effect size was due to the number of collateral ($b = -24.53$, $p = 1.00$), followed by duration in months ($b = -13.56$, $p = 1.00$), value tax interest rate ($b = 6.56$, $p = 1.00$), and duration in years ($b = 2.71$, $p = 1.00$). The results show that contract value, balance value, recovered value, main value delay, and collateral values did not have any effect at all, while DELAY in days had a very minimal effect.

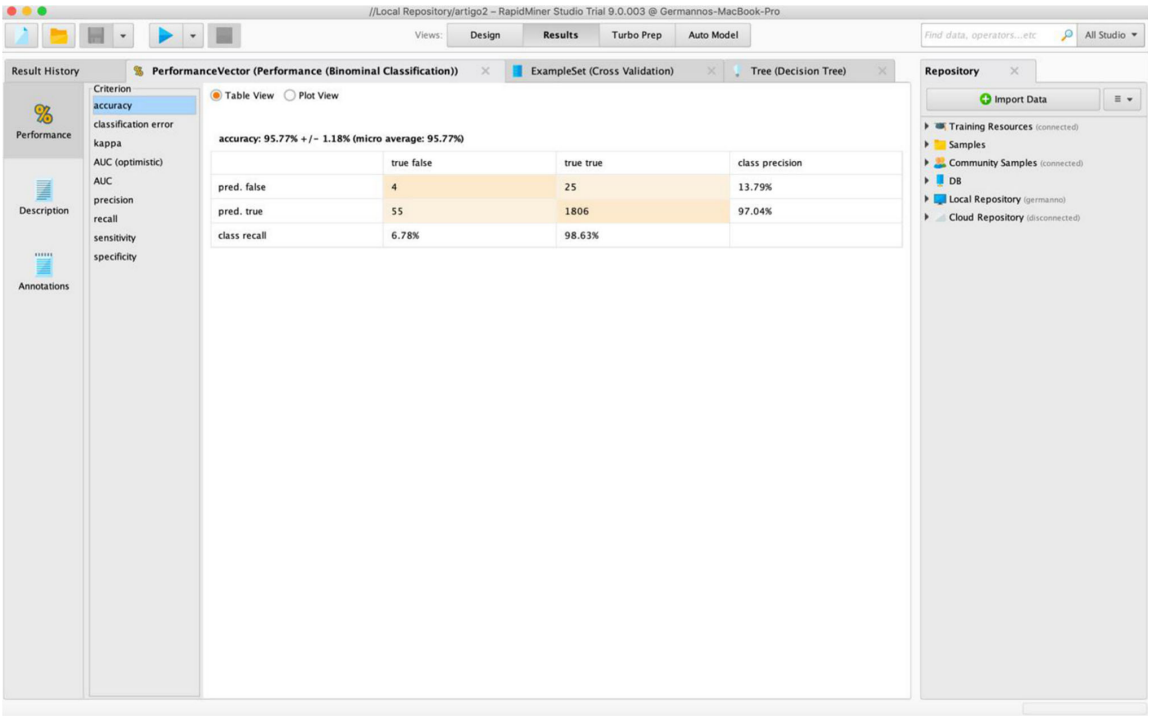


Fig. 5 Step 4: evaluating the performance

Table 1 Omnibus tests of model coefficients

	Chi square	*df*	Sig.
Step 1			
Step	392.741	14	.000
Block	392.741	14	.000
Model	392.741	14	.000

Table 2 Model summary statistic functions

Step	− 2 log likelihood	Cox and Snell R^2	Nagelkerke R^2
1	.000[a]	.257	1.000

[a]Estimation terminated at iteration number 20 because maximum iterations have been reached. The final solution cannot be found

Table 5 shows bootstrapping results that descend robust approximations of regular errors and confidence periods for the regression coefficients. It reveals that value tax rate and value tax interest rate significantly added to the model ($p < .05$). In other words, number of collateral ($b = -24.53$, $p = .05$), value tax rate ($b = -2.34$, $p = .02$) and value tax interest rate ($b = 6.56$, $p = .02$) play an influential role in determining whether the clients will

Table 3 Classification table of percentage accuracy in classification (PAC)[a]

Observed	Predicted		
	OVERDUE		Percentage correct
	No default	In default	
Step 1			
OVERDUE			
No default	45	0	100.0
In default	0	1278	100.0
Overall percentage			100.0

[a]The cut value is .500

default on their repayments. Again, contract value, balance value, recovered value, main value delay, and collateral values, did not have any effect at all, while delay in days had a very minimal effect.

Tables 6, 7, 8, and 9 present results of the expert dataset results, which show the results are more or less the same as those of the training dataset. Table 6 shows that 1837/1837 = 100% of the subjects "in default" were observed. The overall percentage of the correct subjects that were identified by the model in the expert dataset or the sensitivity of the prediction was 97.2%. Comparatively,

Table 4 Variables statistical significance contribution

	B	SE	Wald	*df*	Sig.	Exp (*B*)	95% CI for EXP (*B*)	
							Lower	Upper
Step 1[a,b]								
CONTRACT_VALUE1	.000	.003	.000	1	.988	1.000	.993	1.007
BALANCE_VALUE1	.000	.004	.000	1	.986	1.000	.993	1.007
COLLATERAL_VALUE1	.000	.000	.000	1	.996	1.000	1.000	1.000
NUMBER_OF_COLLATERAL1	− 24.529	25,929.218	.000	1	.999	.000	.000	–
RECOVERED_VALUE1	.000	.002	.000	1	.995	1.000	.996	1.004
VALUE_TX_RATE1	− 2.337	253.262	.000	1	.993	.097	.000	3.649E + 214
VALUE_TX_INTEREST_RATE1	6.557	587.232	.000	1	.991	704.368	.000	–
VALUE_RATE_OVERDUE1	2.644	359.890	.000	1	.994	14.069	.000	3.068E + 307
CLIENT_SIZE1	.930	965.552	.000	1	.999	2.534	.000	–
MAIN_VALUE_DELAY1	.000	.011	.000	1	.985	1.000	.979	1.021
SENIORITY_LEVEL1	− .716	1316.866	.000	1	1.000	.489	.000	–
PERCENT_USED1	− .116	27.920	.000	1	.997	.891	.000	5.188E + 308
DURATION_IN_YEARS1	2.712	4453.449	.000	1	1.000	15.054	.000	–
DURATION_IN_MONTHS1	− 13.557	3412.583	.000	1	.997	.000	.000	–
DURATION_IN_DAYS1	.451	119.137	.000	1	.997	1.569	.000	4.031E + 101
DELAY_IN_DAYS1	− .019	2.233	.000	1	.993	.981	.012	78.131
Constant	23.385	27,376.297	.000	1	.999	14,325,648,374.858		

[a]Variable(s) entered on step 1: CONTRACT VALUE, BALANCE VALUE, COLLATERAL VALUE, NUMBER OF COLLATERAL, RECOVERED VALUE, VALUE TX RATE, VALUE TX INTEREST RATE, VALUE RATE OVERDUE, CLIENT SIZE, MAIN VALUE DELAY, SENIORITY LEVEL, PERCENT USED, DURATION IN YEARS, DURATION IN MONTHS, DURATION IN DAYS, DELAY IN DAYS

[b]Variable(s) entered on step 1: CONTRACT VALUE, BALANCE VALUE, COLLATERAL VALUE, NUMBER OF COLLATERAL, RECOVERED VALUE, VALUE TX RATE, VALUE TX INTEREST RATE, VALUE RATE OVERDUE, CLIENT SIZE, MAIN VALUE DELAY, SENIORITY LEVEL, PERCENT USED, DURATION IN YEARS, DURATION IN MONTHS, DURATION IN DAYS, DELAY IN DAYS

this model predicted 97.2% of the expert dataset correctly as compared with 100% in the training dataset (see Table 3). Table 7 shows the contribution of each predictor variable to the model and its statistical significance in the expert dataset. Like in the training dataset (i.e., Table 3), the results show that all the independent variables did not add significantly to the model ($p > .05$). Although insignificant, the largest effect size was due to the number of collateral ($b = -34.79$, $p = 1.00$), followed by duration in years ($b = -14.97$, $p = 1.00$), per cent used ($b = 12.68$, $p = 1.00$), and value tax interest rate ($b = 11.65$, $p = 1.00$). The results show that contract value, balance value, recovered value, main value delay, client size, and collateral values did not have any effect at whatsoever while the rest had negligible effects. Tables 8 and 9 show the correlation of the various variables in the training set and the expert dataset with the recovered value. In Table 8, contract value, value tax interest rate, client size, duration in years, and per cent used were all negatively correlated with recovered value. All the other variables had a positive correlation with the recovered value. Similarly, in Table 9, contract value, value tax interest rate, client size, duration in years, and per cent used were all negatively correlated with recovered value. All the other variables had a positive correlation with the recovered value.

Table 10 shows the Kolmogorov–Smirnov test results for the various independent-samples tests. The nonparametric tests show that the contract value is not the same across the categories of decisions ($p > .05$). Therefore, the null hypothesis is accepted. The same is observed for the number of collateral ($p > .05$) and collateral value ($p > .05$). There are significant differences in the means of all the other variables, in which case the null hypotheses are rejected and the alternative hypotheses accepted.

Table 11 shows the results of the decision approach. The decision tree results in Table 11 show that the model correctly predicted 97.4% of defaulters. This is slightly higher than what was predicted by the previous model that correctly predicted 97.2% of defaulters (see Table 6).

Table 5 Bootstrap for variables with standard errors and confidence intervals for the regression coefficients[a]

	B	Bootstrap				
		Bias	SE	Sig. (2-tailed)	95% confidence interval	
					Lower	Upper
CONTRACT VALUE	.000	.000[b]	.000[b]	.071[b]	.000[b]	.000[b]
BALANCE VALUE	.000	.000[b]	.000[b]	.059[b]	.000[b]	.000[b]
COLLATERAL VALUE	.000	.000[b]	.000[b]	.099[b]	.000[b]	.000[b]
NUMBER_OF COLLATERAL	− 24.529	5.933[b]	11.456[b]	.053[b]	− 41.839[b]	− 1.944[b]
RECOVERED VALUE	.000	.000[b]	.000[b]	.109[b]	.000[b]	.000[b]
VALUE TX RATE	− 2.337	− .159[b]	1.135[b]	.016[b]	− 5.631[b]	− 1.090[b]
VALUE TX INTEREST RATE	6.557	− 1.083[b]	.953[b]	.016[b]	3.764[b]	7.455[b]
VALUE RATE OVERDUE	2.644	− 1.145[b]	.674[b]	.245[b]	.315[b]	2.764[b]
CLIENT SIZE	.930	.369[b]	1.930[b]	.703[b]	− 3.101[b]	4.680[b]
MAIN VALUE DELAY	.000	.000[b]	.000[b]	.059[b]	.000[b]	.000[b]
SENIORITY LEVEL	− .716	− .948[b]	1.598[b]	.712[b]	− 5.223[b]	1.266[b]
PERCENT USED	− .116	.032[b]	.079[b]	.156[b]	− .243[b]	.069[b]
DURATION IN YEARS	2.712	− .029[b]	6.348[b]	.687[b]	− 10.105[b]	17.086[b]
DURATION IN MONTHS	− 13.557	7.428[b]	7.107[b]	.400[b]	− 22.353[b]	6.407[b]
DURATION IN DAYS	.451	− .250[b]	.230[b]	.447[b]	− .187[b]	.734[b]
DELAY IN DAYS	− .019	.009[b]	.006[b]	.393[b]	− .021[b]	.000[b]
Constant	23.385	− 1.731[b]	15.498[b]	.128[b]	− 5.105[b]	52.794[b]

[a]Unless otherwise noted, bootstrap results are based on 1223 bootstrap samples

[b]Based on 1035 samples

Table 6 Classification table of fuzzy model[a]

Observed	Predicted		
	OVERDUE		Percentage correct
	No default	In default	
Step 1			
OVERDUE			
No default	0	53	.0
In default	0	1837	100.0
Overall percentage			97.2

[a]The cut value is .500

4 Discussion

4.1 Issues with decision trees in credit scoring

Decision trees clearly have many advantages. The experience with decision tree modelling in this study shows that it is an excellent tool for representing graphics decision alternatives efficiently, i.e., visual representation of complex alternatives can be expressed quickly and clearly. The visual appearance is particularly important in understanding subsequent decisions and outcome dependencies. Therefore, the model can be used to compare the relationships between the changing input values and the various decision alternatives. Furthermore, it appears that decision tree models can play a complementary role to other scoring tools. For example, decision can be used to easily explain complex alternatives to non-professional users in ways that fuzzy logic cannot. Not only do decision trees are capable of representing any discrete-value classifier but also are capable of handling datasets with errors and missing values. However, three disadvantages were encountered notably; the algorithm requires that the target attribute has only discrete values, poor performance with the presence of more complex interactions, and over-sensitivity to the training set, irrelevant attributes, and to noise. Poor performance is mostly associated with the need to redraw the tree every time the model is updated, i.e., data classified by an already-trained Tree is added to the Tree as a training data point. Unlike in most other supervised learning algorithms, the addition of training instances is not incremental. Put in other words, training of decision trees cannot occur online, but rather only in batch mode. This limitation became obvious when the classifier was updated. This is significant because for other supervised learning algorithms, for example, they begin classifying data once they

Table 7 Contribution of each predictor variable

	B	SE	Wald	*df*	Sig.	Exp (*B*)	95% CI for EXP (*B*)	
							Lower	Upper
Step 1[a]								
CONTRACT VALUE	.000	.003	.001	1	.978	1.000	.994	1.006
BALANCE VALUE	.000	.003	.001	1	.977	1.000	.994	1.006
COLLATERAL VALUE	.000	.000	.000	1	.998	1.000	1.000	1.000
NUMBER_OF COLLATERAL	− 34.789	15,751.026	.000	1	.998	.000	.000	–
RECOVERED VALUE	.000	.002	.000	1	.991	1.000	.997	1.003
VALUE TX RATE	− 3.641	137.979	.001	1	.979	.026	.000	7.357E + 115
VALUE TX INTEREST RATE	11.645	362.399	.001	1	.974	114,119.379	.000	–
VALUE RATE OVERDUE	.070	323.855	.000	1	1.000	1.072	.000	4.974E + 275
CLIENT SIZE	.000	.008	.001	1	.972	1.000	.986	1.015
MAIN VALUE DELAY	− .001	605.169	.000	1	1.000	.999	.000	–
SENIORITY LEVEL	− .023	14.433	.000	1	.999	.977	.000	1.885E +12
PERCENT USED	12.683	1642.600	.000	1	.994	322,122.914	.000	–
DURATION IN YEARS	− 14.970	1813.071	.000	1	.993	.000	.000	–
DURATION IN MONTHS	.468	55.962	.000	1	.993	1.597	.000	6.892E + 47
DURATION IN DAYS	− .012	1.110	.000	1	.992	.988	.112	8.708
Constant	36.288	15,941.435	.000	1	.998	5.752E + 15		

[a]Variable(s) entered on step 1: CONTRACT VALUE, BALANCE VALUE, COLLATERAL VALUE, NUMBER OF COLLATERAL, RECOVERED VALUE, VALUE TX RATE, VALUE TX INTEREST RATE, VALUE RATE OVERDUE, CLIENT SIZE, MAIN VALUE DELAY, SENIORITY LEVEL, PERCENT USED, DURATION IN YEARS, DURATION IN MONTHS, DURATION IN DAYS, DELAY IN DAYS

have been trained. The data can also be used to prepare the already-trained classifier. However, the entire dataset needs to be trained with decision trees, i.e., the original dataset plus new instances need to be retrained.

As the data is classified stepwise one node a time until the terminal node, only two possibilities are possible (left–right) with decision trees. Therefore, there are other relationships between some variables that the decision trees just are not able to learn. This is significant in the sense that in the logistic regression, for example, it is able to see the individual impacts of the variables in the model which is impossible with the decision trees.

A practical limitation that was encountered with the use of decision tree modelling technique is that it is not possible to use it in regression mode most probably because it is predominantly used for prediction of discrete outcomes.

4.2 Issues with fuzzy logic in credit scoring

This study finds fuzzy logic as the most convenient method for risk analyses. However, it was important to gain a comprehensive understanding of the system processes in order to conduct credit scoring using fuzzy logic. This included identifying the sources of defaulting correctly and consistently as well as identifying the input data for credit scoring. The most important feature of the credit scoring based on fuzzy logic principles is that the entire process leads to the creation of mechanisms that can effectively reduce the risk of default. Because of the analysis's exact output and mechanisms to ameliorate the risks, it can repeat the scoring process on a regular basis with a valuable output.

Because of the application of quantitative data in the fuzzy logic algorithms and methods, subjectivity is reduced to acceptable levels. Therefore, it can better control the process of creating relationships and dependencies between input data and credit scoring. However, this should not be misconstrued to imply that subjectivity is eliminated entirely from the process of risk analysis.

Risk analysis basically implies an assessment of risk, which is an activity associated with the measurement of the strength of the overall system. Consequently, it provides the information required for planned improvement of the company's operation based on the information obtained from credit scoring. One of the benefits of using fuzzy logic in credit scoring is that the whole system is very flexible [8]. Although every situation that can be solved by fuzzy logic can be solved using other methods, fuzzy logic is the most efficient of all. Unlike in decision trees, the modification of the fuzzy logic system requires only adding some

Table 8 Results of negative correlation with recovered value

	Constant	CONTRACT VALUE	BALANCE VALUE	COLLATERAL VALUE	NUMBER OF COLLATERAL	RECOVERED VALUE	VALUE TX RATE	VALUE TX INTEREST RATE
Constant	1	0.127	− 0.134	− 0.148	− 0.988	− 0.044	− 0.051	0.131
CONTRACT VALUE	0.127	1	− 0.989	− 0.446	− 0.113	− 0.195	− 0.066	0.745
BALANCE VALUE	− 0.134	− 0.989	1	0.465	0.12	0.158	0.073	− 0.791
COLLATERAL VALUE	− 0.148	− 0.446	0.465	1	0.158	0.099	− 0.07	− 0.692
NUMBER_OF COLLATERAL	− 0.988	− 0.113	0.12	0.158	1	0.024	0.019	− 0.154
RECOVERED VALUE	− 0.044	− 0.195	0.158	0.099	0.024	1	0.094	− 0.121
VALUE TX RATE	− 0.051	− 0.066	0.073	− 0.07	0.019	0.094	1	− 0.038
VALUE TX INTEREST RATE	0.131	0.745	− 0.791	− 0.692	− 0.154	− 0.121	− 0.038	1
VALUE RATE OVERDUE	− 0.13	− 0.304	0.306	0.261	0.025	0.16	− 0.301	− 0.163
CLIENT SIZE	0.125	0.474	− 0.499	− 0.127	− 0.053	− 0.271	− 0.45	0.305
MAIN VALUE DELAY	− 0.128	− 0.002	0.005	− 0.19	0.005	0.055	0.128	0.168
SENIORITY LEVEL	− 0.212	− 0.414	0.429	0.632	0.132	0.169	0.103	− 0.399
PERCENT USED	0.02	− 0.067	0.071	0.549	0.067	− 0.02	− 0.41	− 0.334
DURATION IN YEARS	0.088	0.216	− 0.238	− 0.735	− 0.136	− 0.022	0.048	0.456
DURATION IN MONTHS	− 0.092	− 0.214	0.237	0.719	0.136	0.023	− 0.021	− 0.443
DURATION IN DAYS	− 0.062	− 0.386	0.409	0.674	0.113	0.035	0.328	− 0.687

	VALUE RATE OVERDUE	CLIENT SIZE	MAIN VALUE DELAY	SENIORITY LEVEL	PERCENT USED	DURATION IN YEARS	DURATION IN MONTHS	DURATION IN DAYS
Constant	− 0.13	0.125	− 0.128	− 0.212	0.02	0.088	− 0.092	− 0.062
CONTRACT VALUE	− 0.304	0.474	− 0.002	− 0.414	− 0.067	0.216	− 0.214	− 0.386
BALANCE VALUE	0.306	− 0.499	0.005	0.429	0.071	− 0.238	0.237	0.409
COLLATERAL VALUE	0.261	− 0.127	− 0.19	0.632	0.549	− 0.735	0.719	0.674
NUMBER_OF COLLATERAL	0.025	− 0.053	0.005	0.132	0.067	− 0.136	0.136	0.113
RECOVERED VALUE	0.16	− 0.271	0.055	0.169	− 0.02	− 0.022	0.023	0.035
VALUE TX RATE	− 0.301	− 0.45	0.128	0.103	− 0.41	0.048	− 0.021	0.328
VALUE TX INTEREST RATE	− 0.163	0.305	0.168	− 0.399	− 0.334	0.456	− 0.443	− 0.687
VALUE RATE OVERDUE	1	− 0.437	0.299	0.481	− 0.004	0.131	− 0.14	− 0.232
CLIENT SIZE	− 0.437	1	− 0.241	− 0.405	0.255	− 0.048	0.035	− 0.057

Table 8 (continued)

	VALUE RATE OVERDUE	CLIENT SIZE	MAIN VALUE DELAY	SENIORITY LEVEL	PERCENT USED	DURATION IN YEARS	DURATION IN MONTHS	DURATION IN DAYS
MAIN VALUE DELAY	0.299	− 0.241	1	0.368	− 0.653	0.193	− 0.148	− 0.422
SENIORITY LEVEL	0.481	− 0.405	0.368	1	0.043	− 0.421	0.432	0.227
PERCENT USED	− 0.004	0.255	− 0.653	0.043	1	− 0.634	0.584	0.37
DURATION IN YEARS	0.131	− 0.048	0.193	− 0.421	− 0.634	1	− 0.998	− 0.531
DURATION IN MONTHS	− 0.14	0.035	− 0.148	0.432	0.584	− 0.998	1	0.519
DURATION IN DAYS	− 0.232	− 0.057	− 0.422	0.227	0.37	− 0.531	0.519	1

other variables rather than changing all or most of what has already been done [9, 16].

The experience associated with the use of fuzzy logic had its own drawbacks as well. These limitations are most likely associated with the principles of the method, i.e., the rules of combining membership functions into the min–max rule for conjunctive (AND) and disjunctive (OR) reasoning. The major drawback with these rules is that they are not robust enough. This view is augmented by the findings of other studies that proposed different rules of combining the AND and OR clauses. Some studies have proposed that instead of applying the minimum or the maximum of the membership functions, the geometric mean should be considered. There are many possibilities of rules, but all-in-all, they are just arbitrary. One of the ways that these limitations were minimized was by using enough training data to train the fuzzy logic system to choose the best rule for the classification. Lack of adequate data can have deleterious consequences on the robustness of the output. As such, fuzzy logic seems to be best suited for big data classification and predictive analytics.

Perhaps the biggest challenge encountered with the use of fuzzy logic is associated with the fact that the rules assign similar importance to all factors that need to be aggregated. For instance, it is possible that the role of the value tax rate is not of the same importance to financial defaulting as the role of seniority level. Therefore, this problem is solved by not insisting on all membership functions in the fuzzy logic model taking values between 0 and 1.

Consequently, fuzzy logic models can be integrated with neural networks leading to superior prediction accuracy. In other words, the incorporation of qualitative data into the system increases the scope of the model. Furthermore, instead of using standalone fuzzy logic models, it can be used as a module in credit rating whereby the retail customers can be rated and investigated further using artificial intelligence techniques to make the model self-learning. Integration of fuzzy logic models with other neural networks has been used in credit scoring to isolate risky borrowers from those with high creditworthiness successfully.

4.3 Comparison of the two models

The study suggests that fuzzy logic is preferable over the decision trees method. However, care should be taken in the interpretation of the effectiveness of this model. First, the identification of the defaulters was done using dynamic modelling, which was assumed to be perfectly discriminatory. This is probably not true in real life.

All the assumptions were stated a priori in the decision tree modelling. However, assigning equal weights, both to

Table 9 Results of positive correlation with the recovered value

	Constant	CONTRACT VALUE	BALANCE VALUE	COLLATERAL VALUE	COLLATERAL	RECOVERED VALUE	TX RATE	TX INTEREST RATE
Constant	1	0.127	− 0.134	− 0.148	− 0.988	− 0.044	− 0.051	0.131
CONTRACT VALUE	0.127	1	− 0.989	− 0.446	− 0.113	− 0.195	− 0.066	0.745
BALANCE VALUE	− 0.134	− 0.989	1	0.465	0.12	0.158	0.073	− 0.791
COLLATERAL VALUE	− 0.148	− 0.446	0.465	1	0.158	0.099	− 0.07	− 0.692
NUMBER_OF COLLATERAL	− 0.988	− 0.113	0.12	0.158	1	0.024	0.019	− 0.154
RECOVERED VALUE	− 0.044	− 0.195	0.158	0.099	0.024	1	0.094	− 0.121
VALUE TX RATE	− 0.051	− 0.066	0.073	− 0.07	0.019	0.094	1	− 0.038
VALUE TX INTEREST RATE	0.131	0.745	− 0.791	− 0.692	− 0.154	− 0.121	− 0.038	1
VALUE RATE OVERDUE	− 0.13	− 0.304	0.306	0.261	0.025	0.16	− 0.301	− 0.163
CLIENT SIZE	0.125	0.474	− 0.499	− 0.127	− 0.053	− 0.271	− 0.45	0.305
MAIN VALUE DELAY	− 0.128	− 0.002	0.005	− 0.19	0.005	0.055	0.128	0.168
SENIORITY LEVEL	− 0.212	− 0.414	0.429	0.632	0.132	0.169	0.103	− 0.399
PERCENT USED	0.02	− 0.067	0.071	0.549	0.067	− 0.02	− 0.41	− 0.334
DURATION IN YEARS	0.088	0.216	− 0.238	− 0.735	− 0.136	− 0.022	0.048	0.456
DURATION IN MONTHS	− 0.092	− 0.214	0.237	0.719	0.136	0.023	− 0.021	− 0.443
DURATION IN DAYS	− 0.062	− 0.386	0.409	0.674	0.113	0.035	0.328	− 0.687

	VALUE RATE OVERDUE	CLIENT SIZE	VALUE DELAY	SENIORITY LEVEL	PERCENT USED	DURATION IN YEARS	DURATION IN MONTHS	DURATION IN DAYS
Constant	− 0.13	0.125	− 0.128	− 0.212	0.02	0.088	− 0.092	− 0.062
CONTRACT VALUE	− 0.304	0.474	− 0.002	− 0.414	− 0.067	0.216	− 0.214	− 0.386
BALANCE VALUE	0.306	− 0.499	0.005	0.429	0.071	− 0.238	0.237	0.409
COLLATERAL VALUE	0.261	− 0.127	− 0.19	0.632	0.549	− 0.735	0.719	0.674
NUMBER_OF COLLATERAL	0.025	− 0.053	0.005	0.132	0.067	− 0.136	0.136	0.113
RECOVERED VALUE	0.16	− 0.271	0.055	0.169	− 0.02	− 0.022	0.023	0.035
VALUE TX RATE	− 0.301	− 0.45	0.128	0.103	− 0.41	0.048	− 0.021	0.328
VALUE TX INTEREST RATE	− 0.163	0.305	0.168	− 0.399	− 0.334	0.456	− 0.443	− 0.687
VALUE RATE OVERDUE	1	− 0.437	0.299	0.481	− 0.004	0.131	− 0.14	− 0.232
CLIENT SIZE	− 0.437	1	− 0.241	− 0.405	0.255	− 0.048	0.035	− 0.057

Table 9 (continued)

	VALUE RATE OVERDUE	CLIENT SIZE	VALUE DELAY	SENIORITY LEVEL	PERCENT USED	DURATION IN YEARS	DURATION IN MONTHS	DURATION IN DAYS
MAIN VALUE DELAY	0.299	− 0.241	1	0.368	− 0.653	0.193	− 0.148	− 0.422
SENIORITY LEVEL	0.481	− 0.405	0.368	1	0.043	− 0.421	0.432	0.227
PERCENT USED	− 0.004	0.255	− 0.653	0.043	1	− 0.634	0.584	0.37
DURATION IN YEARS	0.131	− 0.048	0.193	− 0.421	− 0.634	1	− 0.998	− 0.531
DURATION IN MONTHS	− 0.14	0.035	− 0.148	0.432	0.584	− 0.998	1	0.519
DURATION IN DAYS	− 0.232	− 0.057	− 0.422	0.227	0.37	− 0.531	0.519	1

Table 10 Kolmogorov–Smirnov test results regarding the fuzzy model

Hypothesis test summary

	Null hypothesis	Test	Sig.	Decision
1	The distribution of CONTRACT_VALUE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.910	Retain the null hypothesis
2	The distribution of BALANCE_VALUE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.039	Reject the null hypothesis
3	The distribution of COLLATERAL_VALUE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.163	Retain the null hypothesis
4	The distribution of NUMBER_OF_COLLATERAL1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.810	Retain the null hypothesis
5	The distribution of RECOVERED_VALUE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
6	The distribution of VALUE_TX_RATE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
7	The distribution of VALUE_TX_INTEREST_RATE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
8	The distribution of VALUE_RATE_OVERDUE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
9	The distribution of CLIENT_SIZE1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.002	Reject the null hypothesis
10	The distribution of MAIN_VALUE_DELAY1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
11	The distribution of SENIORITY_LEVEL1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.043	Reject the null hypothesis
12	The distribution of PERCENT_USED1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.042	Reject the null hypothesis
13	The distribution of DURATION_IN_YEARS1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
14	The distribution of DURATION_IN_MONTHS1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
15	The distribution of DURATION_IN_DAYS1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis
16	The distribution of DELAY_IN_DAYS1 is the same across categories of OVERDUE	Independent-samples Mann–Whitney *U* test	.000	Reject the null hypothesis

Asymptotic significances are displayed. The significance level is .05

Table 11 Confusion matrix of a model developed using the decision tree method

	Predicted		
	0	1	$\sum$
Actual			
0	NA	2.6%	35
1	NA	97.4%	1290
$\sum$		1325	1325

defaulters and non-defaulters, may not truly reflect the true values of the research population. Each estimate was assumed to be independent and equally weighted. These assumptions were addressed by conducting the sensitivity analysis to affirm the robustness of the findings.

The fuzzy logic made no assumptions about the probabilities even though it was assumed that the expert estimates were comparable in scale thus limiting its discriminatory effect on dimensions. However, a sensitivity analysis cannot be performed to ascertain the robustness of the fuzzy logic model. As such, if the expert data analysis were not accurate and verifiable, this could result in a large variation in the model's outcome.

As compared with fuzzy logic, it is conceivable that a decision analytic approach can more easily present the outcomes of the analysis in a decision tree or a diagram. Furthermore, the decision analytic approach explicitly states the hypotheses and can be easily verified using sensitivity tests. However, it can easily discourage its use especially for those who are not familiar with probabilistic concepts. Nevertheless, fuzzy logic makes some implicit assumptions that may make it even harder for creditgrantors to follow the logical decision-making process.

5 Conclusion and future work

This paper compares the performance of fuzzy sets with that of artificial neural network based decision trees on a credit scoring to predict the recovered value. The specific objective is to determine the best model for predicting the recovered value. The findings of the study show that artificial neural network-based decision trees are excellent for representing graphics decision alternatives efficiently. The particular strength of artificial neural network-based decision trees is in their tendency to help to comprehend sequential decisions and outcome dependencies. The model can play a complementary role to other scoring tools such as fuzzy assets whereby the classes it creates can be used as fuzzy sets. However, a decision tree algorithm requires that the target attribute has only discrete values. Another disadvantage is that it performs poorly in terms of complex interactions in which the decision trees are redrawn every time new data is added to the model. Also, decision trees are over-sensitive to the training set, irrelevant attributes, and to noise. Coupled with the fact that data is classified stepwise one node a time until the terminal node, it is difficult to add a regression function to the model making it is predominantly a classification model rather than a predictive one.

On the other hand, the findings of the fuzzy logic show it to be the most convenient method for credit scoring. Our ability to add the regression feature on the fuzzy logic models increases its predictive capabilities. Because of this predictive capability, the data miner can follow an entire process that leads to the creation of mechanisms that can effectively reduce the risk of default. Unlike the decision trees, fuzzy logic allows for additive analysis of data using the selected model. The model is also useful in credit scoring because it helps in the better control of the process of creating relationships and dependencies between the datasets. In addition, it leads to the reduction of subjectivity to acceptable levels thanks in part for the application of quantitative data in the fuzzy logic algorithms. Furthermore, risk analysis based on fuzzy logic models provides the information required for planned operational improvement based on the information obtained from credit scoring.

One of the most important concerns about fuzzy logic credit scoring is that the rules of combining membership functions are too simplistic which makes the model not robust at all. This limitation is minimized by using other arithmetic functions such as the mean instead of the minimum or the maximum of the membership functions. However, the classification gains more credibility with increasing the size of the training data to train the fuzzy logic system to choose the best rule for the classification. Lack of adequate data has adverse consequences on the robustness of the output. It finds that the biggest challenge encountered with the use of fuzzy logic is associated with the fact that the rules assign similar importance to all factors that need to be added together. The solution to this problem can be found in using different values for the different membership functions in the fuzzy logic model rather than just taking values between 0 or 1.

It observes that both fuzzy logic- and artificial neural network-based decision trees have benefits as well as drawbacks. However, the most successful model for credit scoring in financial institutions is the one that incorporates various qualitative consumer aspects in addition to the quantitative ones. Indeed, most of these models rely on quantitative data for classification and modelling of creditworthiness.

In conclusion, the two models made it possible to model uncertainty in the credit scoring process. Although more difficult to implement, fuzzy logic was the best for modelling the uncertainty. However, the decision tree model is

more favourable to the presentation of the problem. The assumptions were slightly different for both models. It was difficult to determine which model produces the best results. The two models explored in this study show that both can be applied effectively to credit scoring, although decision trees can complement fuzzy logic models by generating fuzzy rules for these models.

For future works, fuzzy logic models can be integrated, for example, with neural networks leading to superior prediction accuracy. Incorporating qualitative data expands the scope of the model thereby increasing its robustness. In this regard, it is not surprising that standalone fuzzy logic models are hardly used in contemporary credit scoring practices, but rather they are integrated with neural networks for machine learning of consumer behaviour. For example, the fuzzy support vector machine [55, 56] may provide more favourable outcomes than the original version of the algorithm. This practice of integrating standalone classification and predictive models with other neural networks has been used in credit scoring to successfully evaluating creditworthy from non-creditworthy borrowers.

Acknowledgements This work was supported by the National Funding from the FCT - Fundação para a Ciência e a Tecnologia through the UID/EEA/50008/2019 Project; by the Government of the Russian Federation, Grant 08-08; by Brazilian National Council for Research and Development (CNPq) via Grant No. 309335/2017-5; by Ciência sem Fronteiras of CNPq, Brazil, process number 200450/2015-8; and by the International Scientific Partnership Program ISPP at King Saud University through ISPP #0129.

References

1. Akkoç S (2012) An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: the case of Turkish credit card data. Eur J Oper Res 222:168–178. https://doi.org/10.1016/j.ejor.2012.04.009
2. Lahsasna A, Ainon RN, Wah TY (2010) Credit scoring models using soft computing methods: a survey. Int Arab J Inf Technol 7:115–123
3. Chen W, Xiang G, Liu Y, Wang K (2012) Credit risk evaluation by hybrid data mining technique. Syst Eng Procedia 3:194–200. https://doi.org/10.1016/j.sepro.2011.10.029
4. Ibn UF, Panford JK, Ben H-J (2014) Fuzzy logic approach to credit scoring for micro finances in Ghana (A case study of KWIQPLUS money lending). Int J Comput Appl 94(8):8887. https://doi.org/10.5120/16362-5772
5. Zamula A, Kavun S (2017) Complex systems modeling with intelligent control elements. Int J Model Simul Sci Comput 08:1750009. https://doi.org/10.1142/S179396231750009X
6. Abiyev RH (2014) Credit rating using type-2 fuzzy neural networks. Math Probl Eng. https://doi.org/10.1155/2014/460916
7. Buikstra E, Fallon AB, Eley R (2007) A bi-level neural-based fuzzy classification approach for credit scoring problems. Rural Remote Health 7:543. https://doi.org/10.1002/cplx
8. Darwish NR, Abdelghany AS (2016) A fuzzy logic model for credit risk rating of Egyptian commercial banks. Int J Comput Sci Inf Secur 14:11–19
9. Louzada F, Ara A, Fernandes GB (2016) Classification methods applied to credit scoring: a systematic review and overall comparison. Surv Oper Res Manag Sci 21:117–134. https://doi.org/10.1016/j.sorms.2016.10.001
10. Mammadli S (2016) Fuzzy logic based loan evaluation system. Procedia Comput Sci 102:495–499. https://doi.org/10.1016/j.procs.2016.09.433
11. Rao TVN, Reddy K (2017) Application of fuzzy logic in financial markets for decision making. Int J Adv Res Comput Sci 8(3). https://doi.org/10.26483/ijarcs.v8i3.3020
12. Ye J (2017) Aggregation operators of trapezoidal intuitionistic fuzzy sets to multicriteria decision making. Int J Intell Inf Technol 13:1–22. https://doi.org/10.4018/IJIIT.2017100101
13. Ren S (2017) Multicriteria decision-making method under a single valued neutrosophic environment. Int J Intell Inf Technol 13:23–37. https://doi.org/10.4018/IJIIT.2017100102
14. Mohmad Hassim YM, Ghazali R (2013) Functional link neural network—artificial bee colony for time series temperature prediction. Springer, Berlin, pp 427–437
15. Li EY (1994) Artificial neural networks and their business applications. Inf Manag 27:303–313. https://doi.org/10.1016/0378-7206(94)90024-8
16. Liao SH, Chu PH, Hsiao PY (2012) Data mining techniques and applications—a decade review from 2000 to 2011. Expert Syst Appl 39:11303–11311
17. Goonatilake S, Treleaven PC, Philip C (1995) Intelligent systems for finance and business. Wiley, Hoboken
18. Bentley PJ, Kim J, Jung G-H, Choi J-U (2000) Fuzzy darwinian detection of credit card fraud. In: 14th Annual fall symposium of the Korean information processing society, vol 14
19. Hoffmann F, Baesens B, Martens J et al (2002) Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring. Int J Intell Syst 17:1067–1083. https://doi.org/10.1002/int.10052
20. Bojadziev G, Bojadziev M (2007) Fuzzy logic for business, finance, and management. World Scientific, Singapore
21. Laha A (2007) Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring. Adv Eng Inform 21:281–291. https://doi.org/10.1016/j.aei.2006.12.004
22. Hájek P, Olej V (2015) Intuitionistic fuzzy neural network: the case of credit scoring using text information. Springer, Cham, pp 337–346
23. Khashei M, Mirahmadi A (2015) A soft intelligent risk evaluation model for credit scoring classification. Int J Financ Stud 3:1–12
24. Malhotra R, Malhotra DK (2003) Evaluating consumer loans using neural networks. Omega 31:83–96. https://doi.org/10.1016/S0305-0483(03)00016-1
25. Kogut B, MacDuffie JP, Ragin C (2004) Prototypes and strategy: assigning causal credit using fuzzy sets. Eur Manag Rev 1:114–131. https://doi.org/10.1057/palgrave.emr.1500020
26. Ong C, Huang J, Tzeng G (2005) Building credit scoring models using genetic programming. Expert Syst Appl 29:41–47. https://doi.org/10.1016/j.eswa.2005.01.003
27. Dahal K, Hussain Z, Hossain MA (2005) Loan Risk analyzer based on fuzzy logic. In: 2005 IEEE international conference on e-technology, e-commerce and e-service. IEEE, pp 363–366
28. Jiao Y, Syau Y-R, Lee ES (2007) Modelling credit rating by fuzzy adaptive network. Math Comput Model 45:717–731. https://doi.org/10.1016/J.MCM.2005.11.016
29. Lahsasna A (2009) Evaluation of credit risk using evolutionary-fuzzy logic scheme. University of Malaya, Kuala Lumpur
30. Wang S (2010) A comprehensive survey of data mining-based accounting-fraud detection research. In: 2010 international

conference on intelligent computation technology and automation. IEEE, pp 50–53
31. Grace AM, Williams SO (2016) Comparative analysis of neural network and fuzzy logic techniques in credit risk evaluation. Int J Intell Inf Technol. https://doi.org/10.4018/IJIIT.2016010103
32. Sohn SY, Kim DH, Yoon JH (2016) Technology credit scoring model with fuzzy logistic regression. Appl Soft Comput 43:150–158. https://doi.org/10.1016/J.ASOC.2016.02.025
33. Paul U, Biswas A (2017) Consumer credit limit assignment using bayesian decision theory and fuzzy logic—a practical approach. J Manag 4:11–18
34. Zurada J (2010) Could decision trees improve the classification accuracy and interpretability of loan granting decisions? In: 2010 43rd Hawaii international conference on system sciences. IEEE, pp 1–9
35. Rescher N (1969) Many-valued logic. McGraw-Hill, New York
36. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353. https://doi.org/10.1016/S0019-9958(65)90241-X
37. Baghban A, Jalali A, Shafiee M et al (2019) Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids. Eng Appl Comput Fluid Mech 13:26–39. https://doi.org/10.1080/19942060.2018.1542345
38. Ohno-Machado L, Lacson R, Massad E (2000) Decision trees and fuzzy logic: a comparison of models for the selection of measles vaccination strategies in Brazil. In: Proceedings AMIA symposium, pp 625–629
39. Martens D, Baesens B, Van Gestel T, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. Eur J Oper Res 183:1466–1476. https://doi.org/10.1016/j.ejor.2006.04.051
40. Yaseen ZM, Sulaiman SO, Deo RC, Chau KW (2019) An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J Hydrol 569:387–408. https://doi.org/10.1016/j.jhydrol.2018.11.069
41. Wu DD, Olson DL (2014) A decision support approach for accounts receivable risk management. IEEE Trans Syst Man Cybern Syst 44:1624–1632. https://doi.org/10.1109/TSMC.2014.2318020
42. Massad E, Burattini MN, Ortega NR (1999) Fuzzy logic and measles vaccination: designing a control strategy. Int J Epidemiol 28:550–557
43. Zhang H, He H, Zhang W (2018) Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. Neurocomputing 316:210–221. https://doi.org/10.1016/J.NEUCOM.2018.07.070
44. Yu L, Wang S, Lai KK (2009) An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: the case of credit scoring. Eur J Oper Res 195:942–959. https://doi.org/10.1016/j.ejor.2007.11.025
45. Freeling ANS (1980) Fuzzy sets and decision analysis. IEEE Trans Syst Man Cybern 10:341–354. https://doi.org/10.1109/TSMC.1980.4308515
46. Malhotra R, Malhotra DK (2002) Differentiating between good credits and bad credits using neuro-fuzzy systems. Eur J Oper Res 136:190–211. https://doi.org/10.1016/S0377-2217(01)00052-2
47. Hoffmann F, Baesens B, Mues C et al (2007) Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. Eur J Oper Res 177:540–555. https://doi.org/10.1016/j.ejor.2005.09.044
48. Capotorti A, Barbanera E (2012) Credit scoring analysis using a fuzzy probabilistic rough set model. Comput Stat Data Anal 56:981–994. https://doi.org/10.1016/j.csda.2011.06.036
49. Muslim MA, Nurzahputra A, Prasetiyo B (2018) Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction. In: 2018 International conference on information and communications technology (ICOIACT). IEEE
50. Tang T-C, Chi L-C (2005) Predicting multilateral trade credit risks: comparisons of logit and fuzzy logic models using ROC curve analysis. Expert Syst Appl 28:547–556. https://doi.org/10.1016/J.ESWA.2004.12.016
51. Mierswa I, Wurst M, Klinkenberg R et al (2006) Yale: rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining—KDD'06. ACM Press, New York, p 935
52. Huang JJ, Tzeng GH, Ong CS (2006) Two-stage genetic programming (2SGP) for the credit scoring model. Appl Math Comput 174:1039–1053. https://doi.org/10.1016/j.amc.2005.05.027
53. Chen W, Ma C, Ma L (2009) Mining the customer credit using hybrid support vector machine technique. Expert Syst Appl 36:7611–7616. https://doi.org/10.1016/j.eswa.2008.09.054
54. Rahman N (2018) Data mining techniques and applications. Int J Strateg Inf Technol Appl 9:78–97. https://doi.org/10.4018/IJSITA.2018010104
55. Wang S, Li Y, Shao Y et al (2016) Detection of dendritic spines using wavelet packet entropy and fuzzy support vector machine. CNS Neurol Disord: Drug Targets 16:116–121. https://doi.org/10.2174/1871527315666161111123638
56. Zhang YD, Yang ZJ, Lu HM et al (2016) Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Access 4:8375–8385. https://doi.org/10.1109/ACCESS.2016.2628407

Chapter 4

Decision Support System on Credit Operation Using Linear and Logistic Regression

This chapter consists in the following paper:

Decision Support System on Credit Operation Using Linear and Logistic Regression

Germanno Teles, Joel J. P. C. Rodrigues, Sergei Kozlov, Ricardo A. L. Rabêlo and Victor Hugo C. Albuquerque

Expert Systems, Wiley, ISSN:1468-0394, pp. e12578, May 2020.

DOI: doi.org/10.1111/exsy.12578

According to Journal Citation Reports published by Thomson Reuters in 2019, this journal scored ISI journal performance metrics as follows:

ISI Impact Factor (2019): 1.546

Journal Ranking (2019): 93/136 (Computer Science, Artificial Intelligence)

Received: 21 June 2019 | Revised: 7 April 2020 | Accepted: 27 April 2020
DOI: 10.1111/exsy.12578

SPECIAL ISSUE PAPER

Expert Systems WILEY

Decision support system on credit operation using linear and logistic regression

Germanno Teles[1] | Joel J. P. C. Rodrigues[1,2,3] | Sergei A. Kozlov[2] | Ricardo A. L. Rabêlo[3] | Victor Hugo C. Albuquerque[4]

[1]Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, 6201-001, Portugal
[2]ITMO University, St. Petersburg, Russia
[3]Federal University of Piauí (UFPI), Teresina, PI, Brazil
[4]University of Fortaleza (UNIFOR), Fortaleza, Brazil

Correspondence
Germanno Teles Instituto de Telecomunicações, Universidade da Beira Interior Covilhã 6201-001, Portugal.
Email: germanno.teles@ubi.pt

Funding information
Brazilian National Council for Research and Development, Grant/Award Numbers: 200450/2015-8, 309335/2017-5; Government of Russian Federation, Grant/Award Number: Grant 08-08; Fundação para a Ciência e Tecnologia/Ministério da Ciência e Tecnologia e Ensino Superior: Project UIDB/EEA/50008/2020

Abstract

The act of lending is based on trust in the borrower to honour the obligation of paying back the lender. Greater spreads on credit operations may help predict the expected recovery of the credit, based on the sufficiency and liquidity of the guarantee. This study aims to understand how predictive models can provide different estimations of expected recovery based on the same data sets. It classifies credit by the formulation of a rule that describes the values of a categorical variable according to some specified definition. It finds that a simple logistic regression model can easily be extended to a multiple logistic regression model by integrating more than one prediction variable, which indicates increasing difficulty in obtaining multiple observations with an increasing number of independent variables. It compares the efficiency of the logistic regression with that of a linear regression in predicting whether recovery is due in a credit operation, and, thus, identifies the best model for this purpose.

KEYWORDS
credit scoring, finance, linear regression, logistic regression, machine learning

1 | INTRODUCTION

Technological advancements in machine learning can potentially contribute towards the creation of a powerful tool that allows for the prediction of events based on large unstructured data sets in contexts where classical research tools perform poorly (Witten, Frank, Hall, & Pal, 2016). Not only will such an invention be an end in itself, but it will also enable technical and scientific research in a broad continuum of sectors (Khondoker, Dobson, Skirrow, Simmons, & Stahl, 2016). Machine learning offers advantages which classical statistical tools could not provide by helping in the analysis of large unstructured data sets for highly accurate predictions (Bottou, Curtis, & Nocedal, 2018; Quinlan, 2014).

The main challenge in this process lies in formulating the perfect combination of old and new data, as well as the comparison between data and unstructured data, which relies on distributed strings, data mining, and machine learning techniques. Conventionally, textual communication such as e-mails, blogs, websites, and transcripts of phone conversations are the major sources of unstructured data; such data are difficult to categorize and require considerable pre-processing before they can be used in a model.

For example, the machine learning tools included in the software called Orange allow numerous industries to assess large sets of previously unstructured data to generate accurate estimations of significant industry-specific transformations (Figure 1).

Predictive analytics, which is the method of obtaining knowledge from existing data sets to decide guides and predict future outcomes and trends, include classification techniques and regression techniques. Classification techniques such as decision tree analysis, statistical analysis, neural networks, support vector machines, case-based reasoning, Bayesian classifiers, genetic algorithms, and rough sets help identify patterns in large unstructured data sets and generate cluster sets. Regression techniques include linear regression and logistic regression.

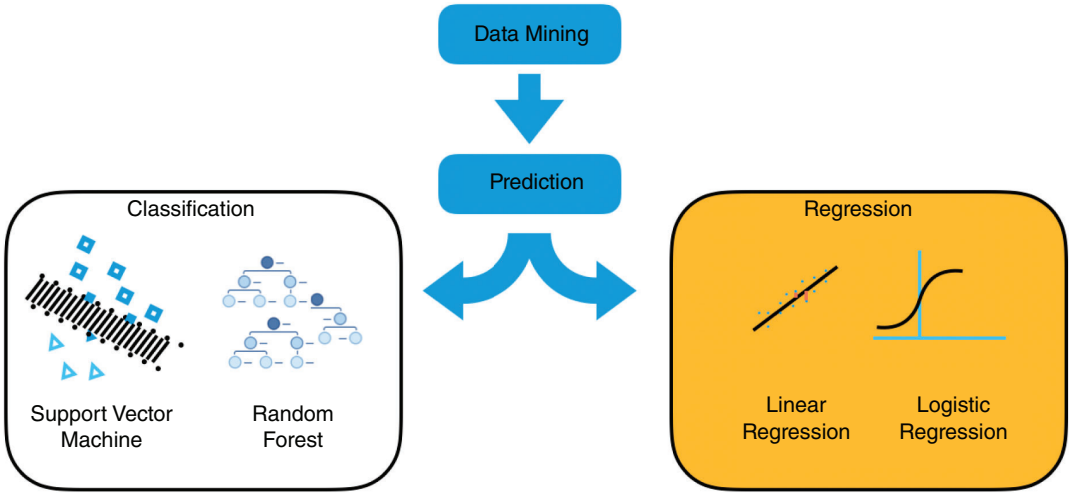


FIGURE 1 Supervised Learning Methods Algorithms used in Orange Software

1.1 | Classification and segmentation techniques for unstructured data sets

Machine learning techniques such as decision trees, logistic regression, Naïve Bayes, and random forest, which save time in classification of unstructured data sets, have a variety of uses: they can help find similarities in customer behaviour, identify targets of a campaign, and categorize documents. Brief descriptions of some of these techniques are given below.

Decision trees exemplify the "divide and conquer" formula: a training set is repeatedly divided until each division consists of examples from only one class. The general algorithm for decision tree building involves the creation of a root node and the assignment of all of the training set data to it. Next, the best attribute for splitting is selected, and a branch is added to the root node for each split value (i.e., the data is split into mutually exclusive subsets along the lines of the specific split). Lastly, Steps 2 and 3 for each and every leaf node until the conditions for stopping are reached (Figure 2).

1.2 | Naive Bayes

Here, probabilities of events are calculated from a decision tree by computing the relative frequency of each class in a leaf, and from a decision list by examining the instances that a particular rule covers.

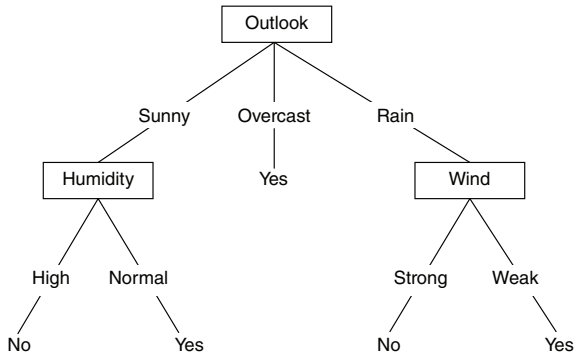
1.3 | Random forest

This technique involves the construction of decision trees using training sets to classify data and make predictions (regressions) based on the classified data of individual trees (Jadhav & Channe, 2016; Liaw & Wiener, 2002). Random decision forests can be used to rank the importance of variables in a regression or classification problem in a natural way (Figure 3).

1.3.1 | Regression techniques

Two types of regression techniques used frequently in predictive analytics include linear regression and logistic regression. Linear regression allows us to summarize and study relationships between two or more continuous (quantitative) variables. It consists of a set of dependent (predicted) and independent (predictor) variables, given in the form of ratios or intervals (Kleinbaum, Kupper, Nizam, & Muller, 2013). This technique explains if and how well a set of variables predicts some phenomenon. Linear regression (Figure 4) requires sound theoretical or conceptual reasons for the analysis (i.e., choice of dependent and independent variables). As a rule, independent variables must not be correlated (Jadhav & Channe, 2016).

FIGURE 2 Decision tree analysis example



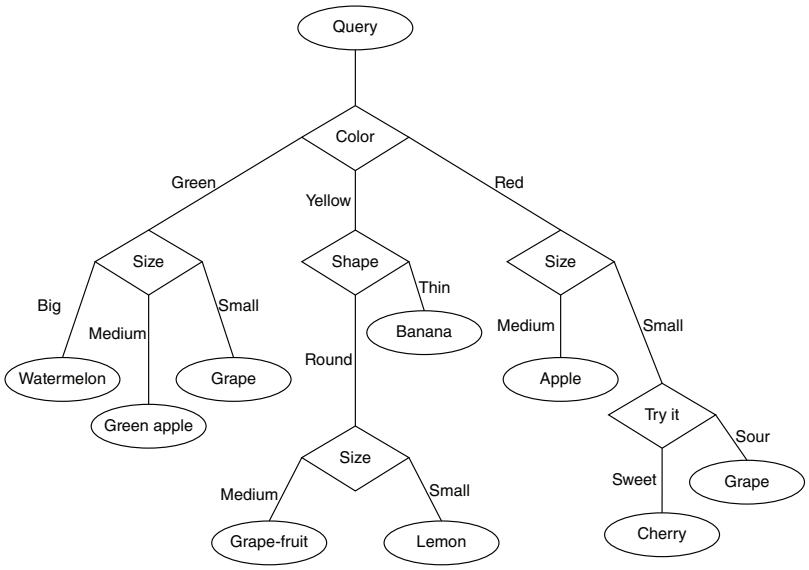


FIGURE 3 Random forest analysis example

Meanwhile, logistic (binomial) regression is used to estimate the probability of a binary response based on one or more predictors, and helps determine whether the presence of a certain condition increases the probability of a given outcome by a specific percentage (Figure 5) as functions that grow regularly at first, more quickly in the central growth period, and slowly at the end, levelling off at a maximum value after a period of time (Finlay, 2014; Gandomi & Haider, 2015; Gualtieri & Curran, 2015; Gunasekaran et al., 2017). In logistic regression, the dependent (predicted) variable is discrete (e.g., pass or fail, win or lose, alive or dead, positive or negative). The predictor can be either categorical or continuous or a mix of both (Kleinbaum et al., 2013).

Naive Bayes classifiers, logistic regression, and decision trees are, in fact, three alternative ways of representing a conditional probability distribution (Jadhav & Channe, 2016).

Further, this study aims to investigate the collateral as a variable in credit-scoring calculations applied to systems that use credit operations. The main contribution of this study includes a comparison of the performance of linear and logistic regression methods. The study highlights the fact that greater spreads on credit operations may help predict the expected recovery of credit. The credit scoring data are transformed into numbers representing one quality; the logistic regression gives binary outcome variable and the linear regression gives a continuous one in a real-world scenario.

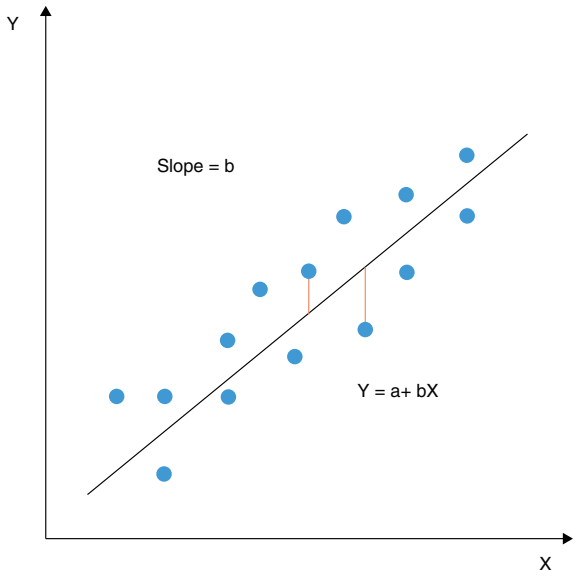


FIGURE 4 Sample Linear regression

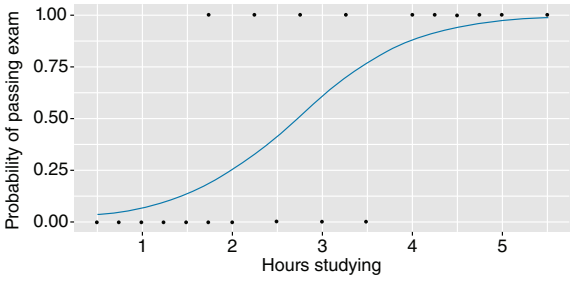


FIGURE 5 Sample Logistic regression

The remainder of the paper is organized as follows. Section 2 comprises the theoretical background of the study; it introduces the related works and discusses the linear and logistic regression methods used in the procedures for classifying the credit scoring approach. Section 3 discusses an approach to predicting credit recoverability using a data set comprising 1,890 samples. Section 4 presents the results of the analysis and identifies the superior model in terms of accuracy and closer fit with the data. The linear and logistic regression models are also compared and their issues discussed. Finally, Section 5 provides the conclusion and suggestions for future work.

2 | MATERIALS AND METHODS

According to Anderson (2007), credit scoring is the statistical analysis of borrowers' data using relevant models to represent the information in numerical values, which guide credit decisions. Scoring, in general, refers to the application of numerical tools to classify cases (such as of credit lending) according to some real or perceived quality (in this context, risk, desirability, or reliability) to distinguish between different cases in order to aid in objective and consistent decision making (for example, deciding whether to lend or not). For this, available data are transformed into uniform numerical values representing one single quality such as *reliability*. In credit scoring, the classification model is based on the behaviour of past borrowers whose repayment details are already known (Anderson, 2007). The attributes identified from the training data sets enable the model to evaluate the creditworthiness of new applicants. Based on the results of the statistical model combined with additional information about borrowers, lenders can then decide whether credit should be given.

In summary, the development of a predictive model requires the researcher to analyse available data of past borrowers using statistical models to identify the potential predictive or independent variables relating to creditworthiness. The outcome of the model, or the dependent variable, depends

on changes in the predictor variables. Essentially, the purpose of credit scoring is to provide lenders with concrete information about applicants' creditworthiness before a decision on granting of credit is taken, and to estimate the probability of repayment (Abdou & Pointon, 2011; Altman, 1968; John Banasik & Crook, 2007; Bandyopadhyay, 2006; Beaver, 1966; Bellotti & Crook, 2009; Berry & Linoff, 2004; Crook, Edelman, & Thomas, 2007; Gouvea, 2007; Kočenda & Vojtek, 2009; Larose & Larose, 2015; Leon, 2008; Mirzaei & Iyer, 2014; Mishra & Silakari, 2012; Nasrabadi, 2007).

2.1 | Credit scoring using linear regression

As mentioned earlier, linear regression provides a model for determining the relationships between dependent and independent variables. In credit scoring, the problem of categorization in this model can be represented using a dummy variable (Agresti, 2002; Akkoç, 2012; Baesens et al., 2003; Brown & Mues, 2012). In one of the earliest applications of linear regression in credit scoring, the model was limited to the evaluation of existing borrowers (Orgler, 1970). The author later used linear regression analysis for evaluating outstanding borrowers. Orgler observed that the information provided during the credit application process had lower predictability than existing data of previous repayment behaviour, which, unfortunately, were not reflected at the time of credit application. Similarly, the application of linear regression was extended to other informational aspects that could not be captured during the credit application process (Burr, 1988; Calvert, 2014; Chatterjee & Hadi, 2006; Chen, Chiang, & Storey, 2012; Daniel, 2015; Draper & Smith, 1998; Hand & Henley, 1997; Hosmer & Lemeshow, 2013; Thomas, 2000). A simple linear regression has the form of $y = b_0 + b_1x$. Thus, a multiple linear regression model that has multiple predictor variables can be represented as $y = b_0 + b_1x_{i1} + \cdots + b_nx_{in}$, where i is 1, 2, ..., m.

2.2 | Credit scoring using logistic regression

In logistic regression, the outcome variable is categorical, or binary (0 or 1). Wiginton (1980) reported that logistic regression provided superior classification in credit scoring than does discriminant analysis. Other studies also compared the performance of logistic regression to other classification models, such as random forests, in credit scoring. (Srinivisan & Kim, 1987).

The application of the maximum likelihood estimation technique helps overcome this limitation (Freund, Wilson, & Sa, 2006). Since issuing credit to new applicants is a two-case problem of “yes” or “no,” logistic regression is likely the ideal statistical model for credit scoring rather than linear regression (Hand & Henley, 1997).

$$\pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdots\cdots\cdots \text{where 1 is if it is overdue, and 0 otherwise.}$$

Thus, this statistical tool has been applied widely in credit scoring (Abdou & Pointon, 2011; Crook et al., 2007; Desai, Crook, & Overstreet, 1996). Other studies have evaluated additional scoring models, such as discriminant analysis and probit analysis, for building credit decision-making models (Boyes, Hoffman, & Low, 1989; Greene, 1998; Orgler, 1970; Steenackers & Goovaerts, 1989; Zekic-Susac, Sarlija, & Bensic, 2004). According to Anderson (2007), discriminant analysis is least used in credit scoring because of its likelihood of classification errors when predicting categories. Perhaps the oldest model used for credit scoring is probit regression, which was applied in 1934 by Chester Bliss (Abbott, 2014).

Probit regression uses the inverse cumulative distribution function associated with a standard normal distribution. In probit regression, normal distribution of the threshold values is assumed. In contrast, in discriminant analysis, the assumption is of multivariate normal distributions and equal variances. In probit regression and likelihood ratio test, estimates of coefficients can be tested individually for significance. Since discriminant analysis coefficients lack the uniqueness of the estimates of coefficients of probit regression, they cannot be individually tested. However, the probit regression model is more difficult to estimate than linear or logistic regression models.

Abdou (2009) and Banasik, Crook, and Thomas (2003) investigated and compared the performance of probit regression in credit scoring with other statistical scoring models. Some studies reported that classification results of probit regression were better than those of discriminant analysis, linear regression, and the Poisson model (Dionne, Artís, & Guillén, 1996; Greene, 1998). As such, the technique is widely accepted as a better alternative to the logistic regression.

3 | PROBLEM AND DATA CALCULATION

This study takes into account known information of a lender's borrowers from the lender's database, specifically their credit history. The population consists of existing borrowers. A sample of borrowers is chosen randomly based on their repayment details. The main aspects of borrower information are credit recovery details and whether recovery is overdue.

Regression techniques are used to analyse the data. Predictive analytics scoring models are generally based on more than one technique (Bonnes, 2014). Therefore, both linear and logistic regressions are used to predict credit recovery using a data set of 1890 records.

First, borrowers who meet the following criteria are shortlisted:

1. The value rate is not overdue (i.e., the individual is a good borrower who has not defaulted) in the first year, which, for purposes of this study, is considered the "learning period"; and
2. The selected individuals are observed to determine whether the value rates had become overdue.

An individual's survival time is given by T. Whether a default has occurred within 1 year is denoted by π (0, 1) π = 1 if credit is overdue, 0 otherwise.

3.1 | Simple linear regression model

In the following stage, a scatter plot of all points of the dataset is drawn to understand the nature of the correlation between the independent and the dependent variables. The graph may be used to observe the relationship between the variables x and y to examine the quality of the regression model. The data may be normalized to enhance the results of the data mining technique by scaling the values to a given range (−1, 1). The direction of and the strength of the variables is defined using the correlation coefficient and covariance.

For n observations, if the mean x and y are given by Equations 1 and 2, respectively, most of the points will lie in the first and third quadrants of the scatter plot if $(x_i-\bar{x})(y_i-\bar{y})$ is positive. However, most points will lie in the second and fourth quadrants if is negative (Figure 6).

$$\bar{x}=\frac{1}{n}\left(\sum_{i-1}^{n}x_i\right) \tag{1}$$

$$\bar{y}=\frac{1}{n}\left(\sum_{i-1}^{n}y_i\right) \tag{2}$$

The following equation is used to calculate the covariance of the independent and dependent variables:

$$Cov_{(x,y)}=\sum_{i-1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{3}$$

Here, n represents the number of observations. Two cases of covariance exist:

1. There is a positive relationship between x and y if $Cov_{(x,\ y)} > 0$; and
2. There is a negative relationship between x and y if $Cov_{(x,\ y)} < 0$.

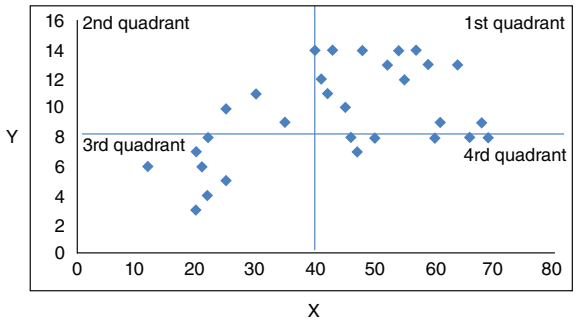


FIGURE 6 Scatter plot of the slope of the functions described above

Since adequate information may not be obtained by determining the covariance between the independent and the dependent variables, it is important to discuss the correlation between the two variables as follows:

$$Cor_{(x,y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=0}^{n}(y_i-\bar{y})^2}} \tag{4}$$

Next, all values are scaled to the range [−1,1] by normalizing the equation ($Cov_{(x,y)} = \sum_{i-1}^{n}(x_i-\bar{x})(y_i-\bar{y})$) Equation 3 (Kantardzic, 2011). Equation 4 gives the covariance of the normalized independent and dependent variables where $-1 \le Cov_{(x,\ y)} \le 1$.

Therefore:

1. There is a strong positive linear relationship between the independent and dependent variables if $Cov_{(x,\ y)}$ is around 1;
2. There is a strong negative linear relationship between the independent and dependent variables if $Cov_{(x,\ y)}$ is around −1; and
3. There is a non-linear relationship between the independent and dependent variables if $Cov_{(x,\ y)}$ is around 0.

Suppose the scatter plot of our dataset is linear. Then, the linear equation can be written as follows:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{5}$$

where α represent the *y*-intercept and β represents the slope of the linear function. The error is defined as ε and i = 1, 2, …., n.

The standard least squares method is applied to estimate the values of α and β in order to construct the linear function. The aim is to minimize the sum of the squares (*S*) of the regression line. Thus,

$$S = \epsilon_1^2 + \varepsilon_2^2 + ,\ldots,\varepsilon_n^2 = \sum_{i=1}^{n}\varepsilon_i^2. \tag{6}$$

The sum of the squares is minimized when S = 0 and a perfect regression line is produced.

$$\epsilon_i = y_i - \alpha - \beta x_i,\ i = 1,2,\ldots,n \tag{7}$$

Since $S = \epsilon_1^2 + \varepsilon_2^2 + ,\ldots,\varepsilon_n^2 = \sum_{i=1}^{n}\varepsilon_i^2$, therefore,

$$S = \epsilon_1^2 + \varepsilon_2^2 + ,\ldots,\varepsilon_n^2 = \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2. \tag{8}$$

The following equation gives the values of α and β:

$$\beta = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) / \sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{9}$$

and

$$\alpha = \bar{y} - \beta\bar{x} \tag{10}$$

The correlation between the independent variable *x* and the dependent variable *y* can be measured using the R-squared test, which gives the ratio of the sum of squares of regression (SSR) to the sum of squared errors (SSE). SSR, which explains the variance, is calculated using the following equation:

$$SSR = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{11}$$

The sum of squared deviations for y is given by

$$\text{SSD} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{12}$$

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y})^2. \tag{13}$$

Therefore,

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{SST - SSE}{SST} = 1 - \left(\frac{SSE}{SST}\right) = \frac{\sum_{i=1}^{n} (y_i - \widehat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{14}$$

where $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1, n = 1, 2, \ldots, n$

The R-squared test measured the goodness of fit of the regression model where $-1 \leq 0 \leq 1$. The regression model is perfect when R-squared equals 1. Therefore, the value of r is the square root of R-squared whose value is between 0 and 1.

3.2 | Multiple linear regression equation

The multiple linear regression model is derived from the simple linear regression equation above. Therefore,

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + \varepsilon_i, \; i = 1, \ldots, n \tag{15}$$

where m is the independent variable and y is the outcome variable. Thus, the model can be represented in matrix algebra as follows:

$$y_i = \begin{pmatrix} y_1 \\ yn \end{pmatrix} = \begin{pmatrix} \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{1m} \\ \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_n x_{nm} \end{pmatrix} \tag{16}$$

3.3 | Logistic regression equation

Suppose x is the predictor variable and y, which is binary (0, 1), is the outcome variable. Then, the relationship between the probability and the independent variable can be represented by the following logistic function:

$$\pi = \Pr(Y = 1 | X = x) \tag{17}$$

The graph of this equation is a non-linear sigmoid curve. Using this sigmoid curve, the equation can be written as:

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}. \tag{18}$$

This equation can be rewritten as follows:

$$1 - \pi = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \tag{19}$$

Dividing Equation 18 by Equation 19 gives

$$\frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 x}. \tag{20}$$

Taking logarithm of both sides to the base of e of the resulting equation gives

$$ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x, \tag{21}$$

where $ln\left[\frac{\pi}{1-\pi}\right]$ is the logit transformation used in logistic regression to check whether the model fits the dataset. The $\frac{\pi}{1-\pi}$ is the odds ratio where $\pi = \Pr(Y = 1|X = x)$ and $1-\pi = \Pr(Y = 0|X = x)$. The equation above is used to fit the data when the outcome variable is categorical with one or multiple predictor variables, which may be categorical or continuous. The maximum likelihood is used to determine the coefficients. Thus, because the independent variable is binary, for k independent variables, the equation can be written as follows:

$$ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, x_1, x_2, \ldots, x_k \tag{22}$$

Next, maximum likelihood estimation or least squares estimation is used to determine the significance of the model,

$$\ell(\beta|x) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1-\pi(x_i)]^{1-y_i}, \tag{23}$$

after which the coefficients of the logistic regression function are calculated as follows:

$$ln(\ell(\beta|x)) = \sum_{i=1}^{n} [y_i ln(\pi(x_i)) + (1-y_1) ln(1-\pi(x_1))] \tag{24}$$

The value of deviance D of the logistic model is calculated using

$$D = -2ln \sum_{i=1}^{n} \left[y_i ln\left(\frac{\pi_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\pi_i}{1-y_i}\right) \right]. \tag{25}$$

The following equation determines whether the predictor variable is significant:

$$G = 2\left\{ \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1-y_1)\ln(1-\pi_i)] - [\sum y_i \cdot \ln\left(\sum y_i\right) + \sum (1-y_1) \cdot \ln(\sum (1-y_1) - nln(n)] \right\} \tag{26}$$

Another way of testing the significance of the predictor variable is through the Wald test, as follows:

$$Z_{Wald} = \left(\frac{\beta_1}{Se(\beta_1)}\right), \tag{27}$$

where β_0 is assumed to be 0 and $Se(\beta_1)$ is the standard error.

The confidence level of 100(1-α)% is obtained from $\beta_0 \pm z.\ Se(\beta_0)$ and $\beta_1 \pm z.\ Se(\beta_1)$, where z is the critical score.

The corresponding coefficient is significant if $\beta_1 \neq 0$.

4 | RESULTS AND DISCUSSION

4.1 | Linear regression output

There were 14 independent variables in total. The "Recovered Value" was the dependent variable. Table 1 shows summary descriptive statistics of the variables, including the means and standard deviations of the sample (N = 1890). The standard deviation values show great variability,

TABLE 1 Descriptive statistics from linear regression output

	Mean	*SD*	Number of records
Recovered value	41,337.9328	569,948.66421	1890
Contract value	304,477.7335	1,523,440.18050	1890
Balance value	286,131.8563	1,384,559.50143	1890
Collateral value	19,936,961.3989	101,883,116.96788	1890
Value Tx rate	5.7750	3.44650	1890
Value Tx interest rate	5.1991	2.17972	1890
Value rate overdue	11.3492	2.65997	1890
Main value delay	31,285.5231	182,431.38267	1890

TABLE 2 Model summary relationship with the recovered value

					Change statistics				
Model	R	R Square	Adjusted R Square	*SE* of the estimate	R Square change	F change	df1	df2	Sig. F change
1	.445a	.198	.192	512,218.06582	.198	35.600	13	1876	.000

[a]Predictors: Constant, Delay in days, Main value delay, Percent used, Client size, Value rate overdue, Seniority level, Collateral value, Duration in years, Value tx interest rate, Value tx rate, Balance value, Contract value, Duration in days.

TABLE 3 ANOVA[a]

Model		Sum of squares	Df	Mean square	F	Sig.
1	Regression	121,424,412,519,181.830	13	9,340,339,424,552.450	35.600	.000[b]
	Residual	492,201,142,883,981.000	1876	262,367,346,953.082		
	Total	613,625,555,403,162.800	1889			

[a]Dependent variable: Recovered value.
[b]Predictors: Constant, delay in days, main value delay, percent used, client size, value rate overdue, seniority level, collateral value, duration in years, value tx interest rate, value tx rate, balance value, contract value, duration in days.

especially for the collateral value, contract value, balance value, main value delay, and recovered value. A multiple linear regression analysis was conducted to determine the significant predictors of the recovered values As shown in Tables 2 and 3, it was found that collectively, all the independent variables had a significant relationship with the recovered value F (13, 1876) = 35.60, p = .000 R^2 = .198 .

The regression coefficients (Table 4) show that only the contract value (t = 3.720, p = .000), collateral value (t = −7.660, p = .000), and main value delay (t = −3.589, p = .000) are significant predictors of the recovered value. However, the t-values of collateral value and main value delay are negative, implying strong negative relationships with the recovered value. The contract value has a strong positive relationship with the recovered value, indicating that the higher the contract values asked on the credit operation, the more difficult it is to find the recovered value. Effects of all the other variables are statistically insignificant.

Table 5 shows the correlation matrix between the different variables. The results indicate a strong and positive relationship between the contract value and recovered value (r = .402, p < .05), balance value and recovered value (r = .405, p < .05), and main value delay and recovered value (r = .283, p < .05). There is also a weak but statistically significant relation between collateral value and recovered value (r = .072, p < .05) as well as between tax rate and recovered value (r = .064, p < .05).

4.2 | Logistic regression output

4.2.1 | Explained variance

Table 6 shows a summary of the model which explains the variation in the dependent variable (the equivalent of R^2 in multiple regression) using the Cox & Snell R Square and Nagelkerke R Square values. Sometimes referred to as *pseudo* R^2 values because of their tendency to be lower than the R^2 values in multiple regression, they are explained in the same manner as the R^2. The explained variance, according to the model summary,

TABLE 4 Regression coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.	95.0% confidence interval for B		Collinearity statistics	
		B	SE	Beta			Lower bound	Upper bound	Tolerance	VIF
1	(constant)	−267,906.047	115,801.690		−2.313	.021	−495,019.718	−40,792.376		
	Contract value	.159	.043	.424	3.720	.000	.075	.242	.033	30.390
	Balance value	.069	.046	.169	1.517	.129	−.020	.159	.035	28.955
	Collateral value	−.001	.000	−.191	−7.660	.000	−.001	−.001	.684	1.462
	Value Tx rate	5,760.581	4,997.858	.035	1.153	.249	−4,041.365	15,562.528	.468	2.136
	Value Tx_Interest rate	9,198.842	6,970.850	.035	1.320	.187	−4,472.594	22,870.278	.602	1.662
	Value rate overdue	2,715.103	6,196.988	.013	.438	.661	−9,438.611	14,868.818	.511	1.956
	Client size	8,154.123	9,349.209	.022	.872	.383	−10,181.819	26,490.065	.652	1.533
	Main value delay	−.396	.110	−.127	−3.589	.000	−.612	−.179	.343	2.911
	Seniority level	30,521.282	18,111.066	.037	1.685	.092	−4,998.672	66,041.236	.868	1.152
	Percent used	637.888	452.979	.031	1.408	.159	−250.508	1,526.284	.876	1.142
	Duration in years	44,345.230	56,687.512	.221	.782	.434	−66,831.981	155,522.441	.005	185.852
	Duration in days	−114.391	155.882	−.207	−.734	.463	−420.111	191.329	.005	186.605
	Delay in days	18.316	21.802	.020	.840	.401	−24.442	61.073	.768	1.301

[a]Dependent variable: Recovered value.

TABLE 5 Correlation Matrix between the different variables

		Recovered value	Contract value	Balance value	Collateral value	Value Tx rate	Value Tx interest rate	Value rate overdue	Client size	Main value delay	Seniority level	Percent used	Duration in years	Duration in months	Duration in days	Delay in days
Recovered value	1		.402[a]	.405[a]	.072[a]	.064[a]	.108[a]	.008	−.030	.283[a]	.038	.031	.045	.044	.044	.011
			.000	.000	.002	.006	.000	.719	.196	.000	.095	.182	.052	.054	.054	.622
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Contract value		.402[a]	1	.981[a]	.522[a]	.139[a]	.232[a]	−.008	−.076[a]	.800[a]	.031	.030	.061[a]	.063[a]	.063[a]	−.024
		.000		0.000	.000	.000	.000	.731	.001	0.000	.178	.194	.008	.006	.006	.297
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Balance value		.405[a]	.981[a]	1	.480[a]	.139[a]	.238[a]	−.010	−.079[a]	.800[a]	.048[b]	.024	.061[a]	.062[a]	.063[a]	−.042
		.000	0.000		.000	.000	.000	.669	.001	0.000	.038	.307	.008	.007	.006	.065
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Collateral value		.072[a]	.522[a]	.480[a]	1	.116[a]	.208[a]	.005	−.076[a]	.395[a]	.048[b]	.000	.024	.024	.024	−.040
		.002	.000	.000		.000	.000	.836	.001	.000	.037	.994	.299	.298	.293	.081
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Value Tx rate		.064[a]	.139[a]	.139[a]	.116[a]	1	.232[a]	−.588[a]	−.388[a]	.111[a]	.051[b]	−.078[a]	−.334[a]	−.337[a]	−.336[a]	−.235[a]
		.006	.000	.000	.000		.000	.000	.000	.000	.026	.001	.000	.000	.000	.000
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Value Tx interest rate		.108[a]	.232[a]	.238[a]	.208[a]	.232[a]	1	.199[a]	−.486[a]	.208[a]	.060[a]	.001	−.039	−.037	−.037	−.018
		.000	.000	.000	.000	.000		.000	.000	.000	.009	.961	.092	.107	.112	.434
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Value rate overdue		.008	−.008	−.010	.005	−.588[a]	.199[a]	1	.080[a]	−.038	−.049[b]	.049[b]	.281[a]	.284[a]	.284[a]	.280[a]
		.719	.731	.669	.836	.000	.000		.000	.100	.035	.033	.000	.000	.000	.000
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Client size		−.030	−.076[a]	−.079[a]	−.076[a]	−.388[a]	−.486[a]	.080[a]	1	−.028	.034	−.047[b]	.151[a]	.152[a]	.152[a]	.175[a]
		.196	.001	.001	.001	.000	.000	.000		.223	.136	.042	.000	.000	.000	.000
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Main value delay		.283[a]	.800[a]	.800[a]	.395[a]	.111[a]	.208[a]	−.038	−.028	1	.008	.041	.015	.019	.019	−.013
		.000	0.000	0.000	.000	.000	.000	.100	.223		.744	.072	.516	.421	.410	.584
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Seniority level		.038	.031	.048[b]	.048[b]	.051[b]	.060[a]	−.049[b]	.034	.008	1	−.298[a]	−.186[a]	−.186[a]	−.187[a]	−.126[a]
		.095	.178	.038	.037	.026	.009	.035	.136	.744		.000	.000	.000	.000	.000
		1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890

TABLE 5 (Continued)

	Recovered value	Contract value	Balance value	Collateral value	Value Tx rate	Value Tx interest rate	Value rate overdue	Client size	Main value delay	Seniority level	Percent used	Duration in years	Duration in months	Duration in days	Delay in days
Percent used	.031	.030	.024	.000	−.078[a]	.001	.049[b]	−.047[b]	.041	−.298[a]	1	.195[a]	.197[a]	.197[a]	.018
	.182	.194	.307	.994	.001	.961	.033	.042	.072	.000		.000	.000	.000	.432
	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Duration in years	.045	.061[a]	.061[a]	.024	−.334[a]	−.039	.281[a]	.151[a]	.015	−.186[a]	.195[a]	1	.997[a]	.997[a]	.407[a]
	.052	.008	.008	.299	.000	.092	.000	.000	.516	.000	.000		0.000	0.000	.000
	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Duration In months	.044	.063[a]	.062[a]	.024	−.337[a]	−.037	.284[a]	.152[a]	.019	−.186[a]	.197[a]	.997[a]	1	1.000[a]	.408[a]
	.054	.006	.007	.298	.000	.107	.000	.000	.421	.000	.000	0.000		0.000	.000
	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Duration in days	.044	.063[a]	.063[a]	.024	−.336[a]	−.037	.284[a]	.152[a]	.019	−.187[a]	.197[a]	.997[a]	1.000[a]	1	.408[a]
	.054	.006	.006	.293	.000	.112	.000	.000	.410	.000	.000	0.000	0.000		.000
	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890
Delay in days	.011	−.024	−.042	−.040	−.235[a]	−.018	.280[a]	.175[a]	−.013	−.126[a]	.018	.407[a]	.408[a]	.408[a]	1
	.622	.297	.065	.081	.000	.434	.000	.000	.584	.000	.432	.000	.000	.000	
	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890	1890

[a]Correlation is significant at the 0.05 level (two-tailed).
[b]Correlation is significant at the 0.01 level (two-tailed).

TABLE 6 Model summary between variation in the dependent variable

Step	−2 log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.226	1.000

[a]Estimation terminated at iteration number 20 because maximum iterations was reached. The final solution cannot be found.

TABLE 7 Classification table[a]

			Predicted		
			Overdue		
Observed			.00	1.00	Percentage of correct classification
Step 1	OVERDUE	.00	53	0	0.0
		1.00	0	1837	100.0
	Overall percentage				97.2

[a]The cut value is .500.

TABLE 8 Variables in the equation

								95% C.I.for EXP(B)	
		B	S.E.	Wald	Df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Contract value	.000	.003	.000	1	.987	1.000	.993	1.007
	Balance value	.000	.004	.000	1	.986	1.000	.993	1.007
	Collateral value	.000	.000	.000	1	.993	1.000	1.000	1.000
	Recovered value	.000	.002	.000	1	.995	1.000	.996	1.004
	Value Tx rate	−2.332	252.094	.000	1	.993	.097	.000	3.714E+213
	Value Tx interest rate	6.566	576.560	.000	1	.991	.232	.000	
	Value rate overdue	2.649	352.889	.000	1	.994	14.142	.000	3.390E+301
	Client size	.922	961.971	.000	1	.999	2.515	.000	
	Main value delay	.000	.011	.000	1	.985	1.000	.980	1.021
	Seniority level	−.711	1,309.999	.000	1	1.000	.491	.000	
	Percent used	−.117	27.316	.000	1	.997	.890	.000	1.000
	Duration in years	2.652	4,499.162	.000	1	1.000	14.182	.000	
	Duration in months	−13.419	3,439.340	.000	1	.997	.000	.000	
	Duration in days	.446	120.289	.000	1	.997	1.563	.000	3.837E+102
	Delay in days	−.019	2.207	.000	1	.993	.981	.013	74.218
	Constant	−1.173	6,038.804	.000	1	1.000	.309		

[a]Variables: Contract value, balance value, collateral value, recovered value, value Tx rate, value Tx interest rate, value rate overdue, client size, main value delay, seniority level, percent used, duration in years, duration in months, duration in days, delay in days.

ranges from 22.6% (Cox & Snell R Square) to 100% (Nagelkerke R Square). Therefore, it is essential to report the Nagelkerke R Square because it provides the highest values.

4.2.2 | Category prediction

As mentioned earlier, logistic regression estimates the probability of an event occurring (in the present case, whether recovery is overdue). If the estimated probability of the event occurring is ≥0.5 (better than even chance), the event is classified as occurring, that is, recovery is overdue. Otherwise, the event is classified as not occurring, that is, no recovery is overdue. Therefore, logistic regression is essential in predicting whether events can be accurately predicted from independent variables. The effectiveness of the predicted classification against the actual classification is assessed using a classification table (Table 7). It is noted that the cut value is .500, implying that the probability of a case being classified into the

TABLE 9 Correlation matrix of the independent variables

	Constant	Contract value	Balance value	Collateral value	Recovered value	Value Tx rate	Value Tx interest Rate	Value rate overdue	Client size	Main value delay	Seniority level	Percent used	Duration in years	Duration in months	Duration in days	Delay in days
Constant	1.000	−.158	.159	−.098	.014	.422	−.167	.206	−.745	.032	−.604	−.495	−.223	−.754	.722	.425
Contract value	−.158	1.000	−.994	.851	−.313	.170	.773	−.695	−.283	.741	.166	.584	.243	.064	−.088	.490
Balance value	.159	−.994	1.000	−.858	.286	−.159	−.772	.679	.272	−.765	−.152	−.573	−.260	−.076	.100	−.474
Collateral value	−.098	.851	−.858	1.000	−.346	.286	.683	−.578	−.289	.415	.140	.510	.130	−.030	.012	.455
Recovered vlue	.014	−.313	.286	−.346	1.000	−.057	−.242	.219	.120	−.289	−.037	−.162	−.073	.001	.007	−.180
Value Tx rate	.422	.170	−.159	.286	−.057	1.000	.417	−.295	−.745	−.080	−.069	.179	−.464	−.533	.538	.714
Value Tx interest rate	−.167	.773	−.772	.683	−.242	.417	1.000	−.892	−.419	.449	.246	.773	.059	−.016	.005	.677
Value rate overdue	.206	−.695	.679	−.578	.219	−.295	−.892	1.000	.347	−.430	−.226	−.802	−.197	−.112	.129	−.729
Client size	−.745	−.283	.272	−.289	.120	−.745	−.419	.347	1.000	−.220	.133	−.131	.417	.789	−.773	−.802
Main value delay	.032	.741	−.765	.415	−.289	−.080	.449	−.430	−.220	1.000	−.089	.264	.403	.096	−.133	.331
Seniority level	−.604	.166	−.152	.140	−.037	−.069	.246	−.226	.133	−.089	1.000	.608	−.488	.011	.042	−.181
Percent used	−.495	.584	−.573	.510	−.162	.179	.773	−.802	−.131	.264	.608	1.000	−.019	.163	−.153	.384
duration in years	−.223	.243	−.260	.130	−.073	−.464	.059	−.197	.417	.403	−.488	−.019	1.000	.649	−.707	−.058
Duration in months	−.754	.064	−.076	−.030	.001	−.533	−.016	−.112	.789	.096	.011	.163	.649	1.000	−.997	−.368
Duration in days	.722	−.088	.100	.012	.007	.538	.005	.129	−.773	−.133	.042	−.153	−.707	−.997	1.000	.343
Delay in days	.425	.490	−.474	.455	−.180	.714	.677	−.729	−.802	.331	−.181	.384	−.058	−.368	.343	1.000

"recovery overdue" category is greater than .500. The percentage of accuracy in classification, which reflects the percentage of cases that can be correctly classified, is added to the independent variables.

4.2.3 | Variables in the equation

Table 8 shows the contribution of each variable to the model and its statistical significance. The Wald test estimates the statistical significance for each of the predictor variables. The "Sig." column provides the statistical significance of the test. These results indicate that none of the independent variables added significantly to the prediction model. However, keeping all other factors constant, the results show that the odds of recovery being overdue is 710.232 times higher for the value tax interest rate, 14.142 times higher for the value rate overdue, 2.515 times higher for client size, 1.563 times higher for duration in days, and 14.182 times higher for duration in years than the odds of no recovery being overdue.

Table 9 shows the correlation matrix of the independent variables. According to the results, only the balance value, value rate overdue, client size, duration in months, and duration in days have a positive correlation with recovered value. All the other independent variables are negatively correlated with the recovered value.

5 | CONCLUSION AND FUTURE WORK

Results reveal that while linear regression can be used for the prediction of a continuous variable in the credit process using collateral, it is not suitable for categorical data. Out of linear and logistic regression, linear regression is more applicable in determining groups of variables that can significantly predict an outcome. In contrast, logistic regression is better for predicting categorical variables, since it allows investigation of binary data. However, binary data must be available for logistic regression analysis, which might imply the need to transform available data into 0 and 1 to allow easier decision making. In summary, the two models can be used in predictive analytics to supplement each other.

The simple linear regression model does not fit our data set well because the value of R is .445, which tends towards 0 rather than 1. Comparatively, the R in the logistic regression is greater. Therefore, the logistic regression model better describes the functional relationship between the dependent and independent variables than the linear regression model does.

ACKNOWLEDGEMENTS

This work was supported by Brazilian National Council for Research and Development (CNPq) via Grants No. 200450/2015-8 and 309335/2017-5; by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020; by the Government of Russian Federation, Grant 08-08.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHOR CONTRIBUTION

Germanno Teles: Conceptualization, Methodology, Data curation, Writing, Original draft preparation; Joel Rodrigues: Supervision, Research Methodology, Revision, Validation; Sergei A. Kozlov: Visualization, Investigation, Validation; Ricardo A. L. Rabêlo and Victor Hugo C. Albuquerque: Writing, Reviewing and Editing.

ORCID

Germanno Teles https://orcid.org/0000-0001-5995-6603
Joel J. P. C. Rodrigues https://orcid.org/0000-0001-8657-3800
Ricardo A. L. Rabêlo https://orcid.org/0000-0003-1482-6404
Victor Hugo C. Albuquerque https://orcid.org/0000-0003-3886-4309

REFERENCES

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst* (Vol. 1, (675–704). Indianapolis, IN: Wiley. https://doi.org/10.1002/ejoc.201200111

Abdou, H. A. (2009). Genetic programming for credit scoring: The case of Egyptian public sector banks. *Expert Systems with Applications*, *36*(9), 11402–11417. https://doi.org/10.1016/j.eswa.2009.01.076

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, *176*(January), 161–176. https://doi.org/10.1002/isaf

Agresti, A. (2002). Categorical data analysis. *Statistical Methodology in the Pharmaceutical Sciences*, *13*, 389–470. https://doi.org/10.1002/0471249688

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*(1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2329297

Anderson, R. (2007). *The credit scoring toolkit - theory and practice for retail credit risk management and decision automation*. Oxford, NY: Oxford University Press.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*, 627–635.

Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, *183*(3), 1582–1594. https://doi.org/10.1016/J.EJOR.2006.06.072

Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *The Journal of the Operational Research Society*, *54*(8), 822–832.

Bandyopadhyay, A. (2006). Predicting probability of default of Indian corporate bonds: Logistic and Z-score model approaches. *Journal of Risk Finance*, *7*(3), 255–272. https://doi.org/10.1108/15265940610664942

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, *4*, 71–111. https://doi.org/10.2307/2490171

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 699–1707. https://doi.org/10.1057/jors.2008.130

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques for marketing, sales, and customer relationship management*. Indianapolis, IN: Wiley.

Bonnes, K. (2014). Predictive analytics for supply chains: A systematic literature review. Paper presented at: Proceedings of the BPM demo sessions 2014 co-located with the 12th international conference on business process management.

Bottou, L., Curtis, F., & Nocedal, J. (2018). Optimization methods for large-scale machine Leraning. *Society for Industrial and Applied Mathematics*, *60*(2), 223–311. https://doi.org/10.1137/16M1080173

Boyes, W. J., Hoffman, D. L., & Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, *40*(1), 3–14. https://doi.org/10.1016/0304-4076(89)90026-2

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Burr, D. (1988). On errors-in-variables in binary regression-Berkson case. *Journal of the American Statistical Association*, *83*(403), 739–743. https://doi.org/10.2307/2289299

Calvert, C. E. (2014). Developing a model and applications for probabilities of student success: A case study of predictive analytics. *Open Learning: The Journal of Open, Distance and e-Learning*, *29*(2), 160–173. https://doi.org/10.1080/02680513.2014.931805

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. Hoboken, NJ: Wiley-Interscience.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly*, *36*(4), 1165–1188. https://doi.org/10.1145/2463676.2463712

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447–1465. https://doi.org/10.1016/J.EJOR.2006.09.100

Daniel, B. (2015). Big data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37. https://doi.org/10.1016/0377-2217(95)00246-4

Dionne, G., Artís, M., & Guillén, M. (1996). Count data models for a credit scoring system. *Journal of Empirical Finance*, *3*(3), 303–325. https://doi.org/10.1016/0927-5398(96)00004-7

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. Canada: Wiley. https://doi.org/10.1002/9781118625590

Finlay, S. (2014). *Predictive analyteics, data mining and big data*. UK: Palgrave Macmillan. https://doi.org/10.1057/9781137379283

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis*. San Diego, CA: Elsevier.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/J.IJINFOMGT.2014.10.007

Gouvea, M. A. (2007). Credit risk analysis applying logistic regression, neural networks an genetic algotithms models. Paper presented at: POMS 18th Annual Conference. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.626.8493

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy*, *10*(3), 299–316. https://doi.org/10.1016/S0922-1425(98)00030-9

Gualtieri, M., & Curran, R. (2015). The Forrester wave™: Big data predictive analytics solutions, Q2 2015. *Forrester Research*, 1–18. Retrieved from https://www.predixionsoftware.com/Portals/0/Analyst reports/the Forrester Wave_big data predictive analytics Solutions_Q2 2015.Pdf.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70*, 308–317. https://doi.org/10.1016/J.JBUSRES.2016.08.004

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society Series A*, *160*, 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression*. Toronto, Canada: Wiley.

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research*, *14611*(1), 2319–7064.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*. Hoboken, NJ: Wiley-IEEE Press.

Khondoker, M., Dobson, R., Skirrow, C., Simmons, A., & Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical Methods in Medical Research*, *25*(5), 1804–1823. https://doi.org/10.1177/0962280213502437

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Muller, K. E. (2013). *Applied regression analysis and other multivariable methods*. New York: Duxbury Press.

Kočenda, E., & Vojtek, M. (2009). Default predictors and credit scoring models for retail banking. CESIFO working paper no. 2862C. Retrieved from https://ssrn.com/abstract=1519792

Larose, D. T., & Larose, C. D. (2015). *Data mining and predictive analytics*. Hoboken, NJ: John Wiley & Sons..

Leon, S. (2008). *Linear algebra with applications* (8th ed.). Dartmouth, MA: Pearson.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22. https://doi.org/10.1159/000323281

Mirzaei, T., & Iyer, L. (2014). Application of predictive analytics in customer relationship management: A literature review and classification. Paper presented at Proceedings of the Southern Association for information systems conference, 1–7. Retrieved from http://aisel.aisnet.org/sais2014/23/

Mishra, N., & Silakari, S. (2012). Predictive analytics: A survey, trends, applications, Oppurtunities & Challenges. *International Journal of Computer Science and Information Technologies*, *3*(3), 4424–5238.

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of Electronic Imaging*, *049901*, 049901. https://doi.org/10.1117/1.2819119

Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of Money, Credit and Banking*, *2*(4), 435. https://doi.org/10.2307/1991095

Quinlan, J. R. (2014). C4.5: Program for machine learning.

Srinivisan, V., & Kim, Y. H. (1987). Credit granting a comparative analysis of classificatory procedures. *Journal of Finance*, *42*(3), 655–683.

Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance Mathematics and Economics*, *8*(1), 31–34. https://doi.org/10.1016/0167-6687(89)90044-9

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *The Journal of Financial and Quantitative Analysis*, *15*(3), 757–770. https://doi.org/10.2307/2330408

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Cambridge, MA: Morgan Kaufmann.

Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: A comparison of logistic regression, neural network, and decision tree models. Paper presented at: 26th International Conference on Information Technology Interfaces; July 2004.

AUTHOR BIOGRAPHIES

Germanno Teles received the master's degree in applied computing from the State University of Ceará, Fortaleza, Brazil, in 2013. He is currently working toward the Ph.D. degree in informatics engineering with Instituto de Telecomunicações, University of Beira Interior, Covilha, Portugal. He is an IT Specialist with the Bank of Northeast, Fortaleza, Brazil. He is a member of the Next-Generation Networks and Applications Group supervised by Professor J. J. P. C. Rodrigues.

Joel J. P. C. Rodrigues [S'01, M'06, SM'06, F'20] is a professor at the Federal University of Piauí, Brazil; senior researcher at the Instituto de Telecomunicações, Portugal; and collaborator of the Post-Graduation Program on Teleinformatics Engineering at the Federal University of Ceará (UFC), Brazil. Prof. Rodrigues is the leader of the Next Generation Networks and Applications (NetGNA) research group (CNPq), an IEEE Distinguished Lecturer, Member Representative of the IEEE Communications Society on the IEEE Biometrics Council, and the President of the scientific council at ParkUrbis – Covilhã Science and Technology Park. He was Director for Conference Development - IEEE ComSoc Board of Governors, Technical Activities Committee Chair of the IEEE ComSoc Latin America Region Board, a Past-Chair of the IEEE ComSoc Technical Committee on eHealth, a Past-chair of the IEEE ComSoc Technical Committee on Communications Software, a Steering Committee member of the IEEE Life Sciences Technical Community and Publications co-Chair. He is the editor-in-chief of the International Journal on E-Health and Medical Communications and editorial board member of several high-reputed journals. He has been general chair and TPC Chair of many international conferences, including IEEE ICC, IEEE GLOBECOM, IEEE HEALTHCOM, and IEEE LatinCom. He has authored or coauthored over 850 papers in refereed international journals and conferences, 3 books, 2 patents, and 1 ITU-T Recommendation. He had been awarded several Outstanding Leadership and Outstanding Service Awards by IEEE Communications Society and several best papers awards. Prof. Rodrigues is a member of the Internet Society, a senior member ACM, and Fellow of IEEE.

Sergei A. Kozlov graduated with honors as engineer in quantum electronics from Leningrad Institute of Fine Mechanics and Optics (ITMO University now), Leningrad, USSR, in 1982. He received his PhD and Dr. Sci. Phys. and Maths. degrees from the Saint Petersburg State University, Saint Petersburg, Russia, in 1986 and in 1997 respectively. From 1986 to 2002, he worked for ITMO University, Saint Petersburg, Russia as engineer, assistant-, associate-, and full professor of the Physics Department at Natural Science Faculty. From2002 till now he has worked as a full professor, a Head of Photonics and Optoinformatics Department and a Dean of Photonics and Optoinformatics Faculty at ITMO University, Saint Petersburg, Russia. From 2013 till now he has worked as a Head of International Institute of Photonics and Optoinformatics at ITMO University, Saint Petersburg, Russia. He is the author of 250 articles. His research interests include femtosecond optics and femtotechnologies, nonlinear optics of few-cycle pulses and ultrafast data transmission, terahertz optics and biophotonics, quantum informatics. Prof. Kozlov is the member of SPIE and D.S. Rozdestvenskiy Optical Society.

Ricardo A. L. Rabelo received the B.Sc. degree in computer science from the Federal University of Piauí, Brazil, in 2005, and the Ph.D. degree in power systems from the São Carlos Engineering School, University of São Paulo, Brazil, in 2010. His research interests include smart grid, the Internet of Things, intelligent systems, and power quality.

Victor Hugo C. de Albuquerque [M'17, SM'19] is a professor and senior researcher at the University of Fortaleza, UNIFOR, Brazil, and Data Science Director at the Superintendency for Research and Public Safety Strategy of Ceará State, Brazil. He has a Ph.D in Mechanical Engineering from the Federal University of Paraíba, an MSc in Teleinformatics Engineering from the Federal University of Ceará, and he graduated in Mechatronics Engineering at the Federal Center of Technological Education of Ceará. He is currently a Full Professor of the Graduate Program in Applied Informatics of UNIFOR and leader of the Industrial Informatics, Electronics and Health Research Group (CNPq). He is a specialist, mainly, in IoT, Machine/Deep Learning, Patter Recognition, Robotic.

How to cite this article: Teles G, Rodrigues JJPC, Kozlov SA, Rabêlo RAL, Albuquerque VHC. Decision support system on credit operation using linear and logistic regression. *Expert Systems*. 2020;e12578. https://doi.org/10.1111/exsy.12578

Chapter 5

Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction

This chapter consists in the following paper:

Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction
Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabêlo and Sergei Kozlov

Journal of Artificial Intelligence and Systems, Institute of Electronics and Computer, ISSN: 2642-2859, vol.2, pp. 118-132, March 2020.

DOI: doi.org/10.33969/ais.2020.21008

IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2020, 2, 118–132
https://iecscience.org/journals/AIS
ISSN Online: 2642–2859

Artificial neural network and Bayesian network models for credit risk prediction

Germanno Teles[1], Joel J. P. C. Rodrigues[1,2,3,*], Ricardo A. L. Rabêlo[2], Sergei A. Kozlov[3]

[1]Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã 6201-001, Portugal
Email: germanno.teles@ubi.pt
[2]Centro de Tecnologia, Federal University of Piauí (UFPI), Teresina – PI 64049550, Brazil
Email: joeljr@ieee.org, ricardoalr@ufpi.edu.br
[3]Photonics and Optoinformatics Department, ITMO University, St. Petersburg 197101, Russia
Email: kozlov@mail.ifmo.ru
*Corresponding Author: Joel J. P. C. Rodrigues, Email: joeljr@ieee.org

How to cite this paper: Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabêlo, Sergei A. Kozlov (2020) Artificial neural network and Bayesian network models for credit risk prediction. Journal of Artificial Intelligence and Systems, 2, 118–132.
https://doi.org/10.33969/AIS.2020.21008

Received: January 5, 2020
Accepted: March 3, 2020
Published: March 9, 2020

Abstract

Credit risk threatens financial institutions and may result in irrecoverable consequences. Tools for risk prediction can be used to reduce bank insolvency. This study compares Bayesian networks with artificial neural networks (ANNs) for predicting recovered value in a credit operation. The credit scoring problem is typically been approached as a supervised classification problem in machine learning. The present study explores this problem and finds that ANNs are a more efficient tool for predicting credit risk than the aïve bayesian (NB) approach. The most crucial point is related to lending decisions, and a significant credit operation is associated with a set of factors to the degree that probabilities are used to classify new applicants based on their characteristics. The optimum achievement was obtained when the linear regression was equivalent to 0.2, with a mean accuracy of 85%. For the naïve Bayes approach, the algorithm was applied to four datasets in a single process before the entire dataset was used to create a confusion matrix.

Keywords

Artificial intelligence, Neural network, Bayesian network, Algorithms, Credit risk, Prediction

1. Introduction

Banks are exposed to a wide range of potential risks, ranging from those identified in the budgetary and technological structure to those related to brand reputation and those derived from the social and institutional environment. Because of the level of technology associated with Big Data, computing power, and data availability, most lending institutions have been compelled to renew their business models. Forecasting credit risk, active loan processing, and monitoring model reliability are vital for transparency and decision-making. From the viewpoint of machine learning, the problem has typically been approached as a problem of supervised classification. In this study, binary classifiers are built based on a machine model to predict the probability of loan default, and the results are compared

with those from artificial neural networks (ANNs) to examine their reliability and efficiency. A practical case is used to exhibit the applicability, efficiency, flexibility, and accuracy of data mining approaches to model ambiguous events related to measuring credit risk for financial institutions. However, aside from the technical questions necessary to understand it, confidentiality issues raised from the use of personal data are also important. The application of the ANN algorithm has long raised numerous ethical questions [1]. According to Williams, Sweeney, and Anderson (2007), such problems continue to be addressed through discussions on artificial intelligence. The underlying concern is the fear that an algorithm may take the decision power away from a human. Given that these debates and questions are legitimate, the present study focuses on the algorithms relevant to decision-making in the financial sector, especially those used to simplify/increase fluidity and speed processes [2][41].

Algorithms are sets of codes designed to attain predefined objectives. For example, in a recruitment process, they can discriminate among people based on their profiles. According to Yaseen (2019) the all those classifiers introduce on credit score approach are well-known base classifiers in this domain are used o his work they show results, analysis, and statistical tests demonstrate the ability of the proposed combination method to improve prediction performance against all base classifier. Baghban (2019) considered classification an example of supervised learning as training data associated with class labels a focus on study of various classification techniques, showing its advantages and disadvantages. The work of Wu et.al. (2019) present a study on BI (Business Intelligence) on a bank institution showing the potentially benefit business, increasing the visibility and recognition of research achievements. The behavior of a BN can be reinforced by the work of Abid et.al.(2017) that use this classifier on analyses of costumers loans default payment allowing providing an effective decision support system for banks[66] [5][38][64] . The same approach is used in providing loans to an enterprise, where a bank's lending decisions are based on the algorithm used [3]. Therefore, it is critical to comprehend the underlying problems and establish ways to regulate the use of algorithms [4].

1.1. Artificial neural networks

The study of ANNs can be traced to Frank Rosenblatt (1958) who focused on perceptron algorithms for the development of smart automated systems and software [6]. ANNs are a reliable and efficient approach for predicting outcomes. The method received tremendous support following the development of machine language. Odom and Sharda [43] applied neural networks to the evaluation of credit risk. Initially, the network was based on the Hebb system, which aimed to improve the input vector through perception and focused on increasing the accuracy of the model. The perception neural network was followed by backpropagation, developed by Rumelhart and McClelland [48]. Backpropagation refreshes its weight through the maintenance of history, commonly known as neural processing. However, more interest was focused on deep learning in 2008 when Angelini and colleagues performed the first credit risk analysis used by banking management to compute capital requirements. ANNs have also been used to calculate the variables necessary to evaluate credit risk [22].

ANNs were used to study the nervous system and the way the brain processes information. An ANN entails processing algorithm to model the brains of humans and comprises a large number of interconnected nodes (neurons) working as a system to solve pattern recognition or data classification problems, particularly in the field of biology [25],[28]. The purpose of ANN research is to develop a computational system with relatively low computational cost and time commitment. A range of tasks can be performed by ANNs, including classification, pattern-matching, approximation, function optimization, data clustering, and vector quantization. According to Rumelhart and McClelland, ANN properties initially include the following [22]:

- The cycle or speed of the implementation of ANNs is in nanoseconds.

- The processing time is rapid, and numerous operations can be performed simultaneously.
- he complexity and size of an ANN is subject to the network design and application in use.
- The data in ANNs are stored in contiguous memory sites that can become overloaded when the limit is exceeded.

Learning is the primary property of ANNs, and this occurs in two forms: structure and parameter learning. Parameter learning improves the weight linked to the network, whereas structure learning concentrates on network topology and confirms whether there are any changes within the framework [7]. Moreover, learning can either be supervised or unsupervised depending on expert knowledge [56]. Within supervised learning, reinforcement learning contains an activation function useful for calculating the exact output [8]. This feature is applied to the overall input to determine the total network production [9]c. Some types of activation functions include the binary step, hyperbolic tangent, bipolar sigmoid curve, and identity functions [12].

1.2. Naïve Bayesian (NB) approach

The NB was initially studied by applying Bayes' theorem. The approach comprises a supervised statistical classifier based on the assumption of conditional independence, where probability models are estimated using labeled data (i.e., each instance is assigned to a class) [10]. Mutual conditional probability distributions are used in the Bayesian classifier, allowing class-based conditional independencies to function between variables, with a graphical model used to depict the underlying relationships [13][63]. The random variables form either a continuous or discrete relationship while the attribute within the data may have a real Boolean variable to formulate the relationship [14]. Every arc in the acyclic graph represents the dependence probability, and all variables are independent of the non-descendant.

The formula of Bayes theorem is given as follows[15]:

$$P(A|B) = P(B|A)P(A)/P(B) \tag{1}$$

where in a sample referred to as A, the chances of all events occurring is h. Further, P (h—A) is consistent with Bayes' theorem, which can be stated mathematically as in Eq. 1. This classification is considered the optimal one [16] [59].

When the network topology and data are given in the multiple variables of a sample, data training is impartial. The variables are then used to determine the entries in the continuous probability table. This approach lowers computational costs and is suitable for problems where a strong relationship exists between the variables [61]. The approach is also highly advanced in comparison with support vector machines, and is also applicable to medical diagnosis [62]. Compared with other algorithms such as particle swarm optimization, neural networks, and machine language algorithms, studies based on support vector machines have been promising for assessing credit risk [11].

The present study demonstrates that various algorithms can be used in parallel to address the issue in question, which in the case presented here is loan provision [65]. Multiple strategies for identifying the choice of features (or variables), algorithm, and criteria can provide a solution. For instance, in the new Big Data and digital era, transparency is critical [22]. Strategies based on deep learning are also necessary to train data on the application, and machine learning algorithms and their use must be regulated to ensure accuracy.

The particular focus of this study is credit risk scoring and how distinct machine learning models can help lenders identify default [17]. Further, the stability of these models is examined based on the choice of variables or subsets. Although the methods used by banks in their decisions to award loans remain unclear, the application of classical linear models

in the banking sector is well known [29]. Finally, a transparent elastic approach is used as the benchmark, and its fit and decision rules are compared. To the best of the authors' knowledge, there are no solutions currently available in the literature for credit scoring based on ANNs and Bayesian networks. Further, this study aims to determine the best combination of parameters to use with ANNs and the BN approach to handle and precisely evaluate credit risk. The main contribution of this study is its proposed machine learning model or the combination of them, which is a rare approach to the problem of credit risk measurement. The study highlights the existing gap that prevents intelligence systems from addressing bank modeling concerns.

The paper is organized as follows. Section 2 presents the materials and methods, and describes the datasets and attributes used to forecast credit risk, the research data, and the prediction model. Section 3 reports the results and discusses the analysis when the training and test datasets were created using the Bayes and ANN approaches and presents the analysis defining a systematic method to handle and precisely evaluate credit risk. Finally, Section 4 summarizes the processes used to identify the best combinations and concludes the paper.

2. Data and methods

2.1. Research data

The dataset was collected from a financial institution, and a summary of 1,890 records was accessed and retrieved. As some attributes were missing, the dataset required preprocessing. The global mean approach was employed to replace the missing attributes (El-Shazly, 2002). Each record/instance was assigned to one of two classes: 1 (risky) or 2 (non-risky). The output of the processing represents the label as opposed to the value, where 1 indicates credit risk and 2 represents security (Demerjian, 2007). The predictor attributes include contract_value, balance_value, collateral_value, number_of_collateral, recovered_value, value_tx_rate, value_tx_interest_rate, value_rate_overdue, client_size, main_value_delay, seniority_level, duration_in_years, duration_in_months, duration_in_days, and delay_in_days. For accuracy, the attributes in the dataset were converted into classes, which is critical for preprocessing as it improves the accuracy of the algorithms [19]. Numerous banks have adopted these attributes to predict credit risk [67]. Moreover, to find the corresponding class, the data were normalized for ANNs, as the output ranges between 0 and 1. After normalization and conversion, the data were stored as a comma-separate values file and retrieved for further processing. The dataset was then classed into sets based on the age of the credit operation and trained. The weights were computed based on training and later tested.

2.2. Prediction model

The proposed model constitutes two complementary phases, the NB and ANN phases. The ANN phase is used to estimate the overall credit risk trend and establish the most significant factors, whereas the NB approach determines the probability of credit default when all variables have been measured [12][10]. Therefore, the two phases complement each other. From a technical perspective, although the same raw data were used to implement the two networks, no data flow exists between the systems [57]. Data implementation is independent because of the distinct rationales behind the two policies. Specifically, the output from one network cannot be used as an input of the other. The output was between 0 and 1 and the outcome was marked 2 when the result ranged between 0.75 and 0; otherwise, it was 1. The ANN was based on continuous raw data, most often coded in MATLAB, while the BN input data were converted into Boolean before being coded in MATLAB. The outcome of the two phases allows for validation and verification through the parallel and independent implementation of the dataset.

The proposed neural network is a multilayer perceptron (MLP) based on a feed forward architecture. The popularity of this architecture can be attributed to its link to the robust and powerful learning algorithm referred to as backpropagation learning. In the design of ANNs, the most critical element is accurate identification of the learning algorithm used in the training process [51][36]. In the present case, an active or supervised learning mechanism is used and an appropriate learning rule is applied. Gradient descent is used to adjust the relationship values and the Levenberg–Marquardt algorithm (LMA), one of the most common algorithms in computing and mathematics, was used to compute optimization problems that arise from generic curve fitting [24].

By selecting the chromosome with the lowest cost, the search process continued until the weights were turned into the target solution [37][40]. The learned credit risk function was configured using the autoregressive pattern to predict default risk via an MLP network [42]. The default condition of a given day is subject to the credit level of the previous days. Banks usually consider a time span of 30 days for credit strategies, and when confronting shortages, they invoke proxy funding resources [14][23].

NBs were used to identify the most critical risk indicators among those chosen as the model variables and their effect on each other and on the default risk measure was assessed. NBs are useful for graphically representing probabilistic relationships between variables [39]. They are crucial in data modeling, particularly when data are missing, because they characterize variables based on a combination of graphical models and statistical approaches. NBs can thus detect the possible relationships between variables and, thanks to causal inferences, help predict their trend using probability distribution functions regardless of the nature of the data [44]. Moreover, prior knowledge can be merged with existing data to provide accurate results, thereby leading to correct inferences. This method, together with Bayesian approaches, therefore prevents overfitting the data [49]. Hence, a Bayesian knowledge base allows researchers to draw conclusions and inferences about the relationships among the components of a system, making it the most suitable approach for achieving the second objective of this study.

2.2.1. Key parameters

The results generated by both the NB and ANN phases are the product of supervised learning. A network is produced starting from a random weight in the ANN classifier and the distributions in the NB classifier. The resulting system is trained using the dataset until the outcome is comparable to the distribution or pattern of the primary data[45]. The algorithms applied to the training data in the first and second phases were gradient descent and maximum likelihood estimation, respectively [13] [14] [20] [63]. Training based on algorithms is a standard procedure, and the parameters must be selected correctly. The primary concern is estimating the appropriate parameters for the trained function and probability distribution provided by ANNs and NBs to fit the data (Mileris, 2010). In the model, there are two sets of parameters, namely the sets of weights and binomial distribution parameters in ANNs and NBs, respectively.

During the first phase, the ANN characterizes a function of the datasets (input variables) and attempts to locate the most appropriate coefficients (weights) for the variables. Once the algorithm has learned the data, the target values can be estimated and hence default risk can be predicted[61] [18]. The learning process in the second phase occurs by applying the naïve Bayes rule. Nodes are identified using input variables and considered to contain the prior distribution. Similar parameters then define the previous characteristics of the network nodes.

2.2.2. Prediction and measurement of credit risk

Using the receiver operating characteristic (ROC) curve to evaluate a diagnostic test, an ANN interpolates between the Gauss–Newton algorithm and the gradient descent method [21]. Although the LMA is more robust than the Gauss–Newton algorithm, as with many fitting algorithms, it only finds the local minimum, which is not necessarily the global minimum [26]. To overcome this shortcoming, the genetic algorithm (GA) can be used to search the space of possible solutions [19]. GA first generates a random vector as the weight vector (chromosome), to which crossover and mutation are then applied. The output vector is calculated using inputs and weights, and the differences between the output and target values are introduced as the cost [36]. By selecting the lowest cost chromosome, the searching process continues until the masses evolve into a suitable final solution. Finally, to predict liquidity risk using an MLP network, the learned liquidity risk function is configured with an autoregressive pattern based on the type of risk under analysis[28]. The liquidity condition of a particular day strictly depends on the liquidity levels of the preceding days. Considering the nature of the problem, a powerful computational tool is needed to estimate and predict the credit risk function through the data provided. The ANN architecture with computationally intensive learning and massive parallelism through examples renders it suitable for the task at hand [34].

The ANN based on backpropagation and the naïve Bayes algorithm were implemented in MATLAB software. Initially, the data were divided into categories and the best combinations of the parameters were determined based on the learning rate, epoch, model checking rate, and number of neurons [27], as shown in Tables 1 and 2. The focus was not only on blending the parameters but also ensuring the accuracy of these combinations.

Table 1. Training dataset

Supervised Learning (Naïve Bayes Continuous)						
Parameters						
Lambda				0		
Homoscedasticity assumption				1		
Classifier Performances						
Error Rate			0.4037			
Value Prediction			Confusion Matrix			
	Recall	1-Precision		RISKY	NON- RISKY	Sum
RISKY	0.4721	59.63%	RISKY	642	446	1088
NON-RISKY	0.7227	63.85%	NON- RISKY	247	555	802
				889	1001	1890

Table 2. Test dataset

Supervised Learning (Naïve Bayes Continuous)						
Parameters						
Lambda				0		
Homoscedasticity assumption				1		
Classifier Performances						
Error Rate			0.3615			
Value Prediction			Confusion Matrix			
	Recall	1-Precision		RISKY	NON- RISKY	Sum
RISKY	0.6073	36.60%	RISKY	201	82	283
NON-RISKY	0.3735	42.63%	NON- RISKY	288	59	347
				489	141	630

3. Results and discussion

Twelve columns stood out in the data analysis: "Contract Value," "Balance Value," "Collateral Value," "Recovered Value," "Value Tax Rate," "Value Tax Interest Rate," "Main Value Delay," "Duration in Years," "Duration in Months," "Duration in Days," and "Delay in Days." These columns were eliminated immediately because the exercise focuses on categories or classes of data that operate as predictors of creditworthiness. Rather than hard figures and numbers, the true indicators of creditworthiness are scores. The columns that qualified as scores were "Value Rate Overdue," "Client Size," "Seniority Level," "Percent Used," and "Number of Collateral." The most important point is that these variables are related to lending decisions, and a significant credit operation is related to a set of factors to the degree that probabilities are used to classify new applicants based on their characteristics. Numeric data are removed, while specific classifications are preserved by creating an object that eliminates the identified columns. The training and test datasets were then created. The training dataset was used to train the model, whereas the test data were used to assess the model's accuracy. One-third and two-thirds of the data were allocated to the two sets, respectively. Next, the learning rate was iterated and plotted with different standards of accuracy to obtain a non-linear graph [31]. The optimum performance was obtained when the linear regression was equivalent to 0.2, with mean accuracy of 85%. For the naïve Bayes approach, the algorithm was applied to four datasets in a single process before the entire dataset was used to create a confusion matrix. The best outcomes of the iteration were sourced using the comparative approach [33].

3.1. Bayes approach

Table 1 describes the results of the two Bayesian models, showing that the classification rate improves when the indicators relating to "Value Rate Overdue," "Client Size," and "Seniority Level" are introduced. The best classification rates are 59.63% and 63.85% for the two classes. The criterion of the two types of errors (Type I and Type II) has been examined in numerous studies. The assignment of variables is as follows:

X1: Number_of_collateral
X2: Value_tx_rate
X3: Value_tx_interest_rate
X4: Value_rate_overdue
X5: Client_size
X6: Main_value_delay
X7: Percent Used
X8: Duration_in_years
X9: Seniority_Level

Type I error is also known as credit risk, which is the rate at which bad clients are classified as profitable. Therefore, when a bank has a significantly high rate, which implies that the rate of loan approval is too high, the potential for exposure to credit risk is considerable. Type II error is commercial risk, which is the rate at which applications of paying clients are rejected, with the bank experiencing an opportunity cost attributed to good customers. In the present study, the Type I error is exceedingly high at 52.79%. The introduction of seniority improves the outcomes and the classification rate rises. Further, the model based on the entire dataset indicator reduces the Type I and Type II errors to 32.97% and 39.27%, respectively. These findings show the correlation between value rate overdue and seniority and credit risk, concurring with previous results [35].

3.2. Artificial Neural Network approach

The objective of this approach was to estimate the credit risk function, and therefore continuous data were necessary. The only preprocessing of the dataset conducted was data normalization. As before, data were divided into two categories for training and testing at ratios of one-third and two-thirds [47][60]. The selected network comprised three layers: the MLP layer, the hidden layer, and an output layer [46]. The optimal structure was chosen through trial and error with the network assessed using the micro & small enterprises(MSE). The correlation between the output and target values, variance, mean residuals, learning process error, and root MSE were all used in the assessment of the network [46]. Because most of the hidden cases were sufficient for the network to perform optimally, several models were implemented using a single network. Figure. 1 exhibits the outcome of the assessment obtained from network training using the LMA.

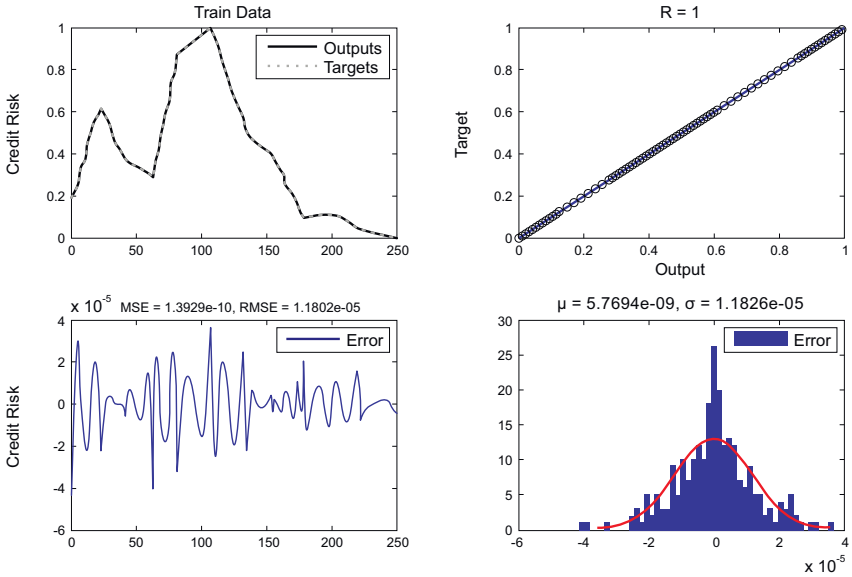


Figure 1. Analysis of the LMA-trained data.

The LMA was used to train the datasets; however, its inability to locate the global minimum influences the results from this approach[52]. Furthermore, the accuracy of the results was sufficient as long as the initial weights were a good approximation and the signal-to-noise ratio exceeded five [28]. For this reason, a meta-heuristic search algorithm, the GA, was used. Given that the GA has random behavior and is independent of its starting point, its application guaranteed that the LMA was functioning correctly. In addition, apart from overcoming the drawbacks of the LMA, the GA in figure 2 demonstrates that the dataset was sufficient to be modeled by any preferable algorithm. As shown in Table 3, the performance of the LMA was much better than that of the GA and it recognized data patterns accurately. As a result, credit risk was modeled using the LMA. Moreover, both the ANN and the Bayesian models provide reliable outcomes, but the former is more effective in the prediction of credit risk with an average score of 82% (Table 4).

The ROC curve in figure 3 aids the visualization and shows the trade-off between recall and precision. This allows the researcher to manipulate the false positive and true positive metrics [53][55]. The relationships among false positives, true positives, false negatives, and true negatives were further summarized using a simple confusion matrix. The probability of the above case occurring was 1, representing a 100% correlation, with 0 depicting no

relationship.

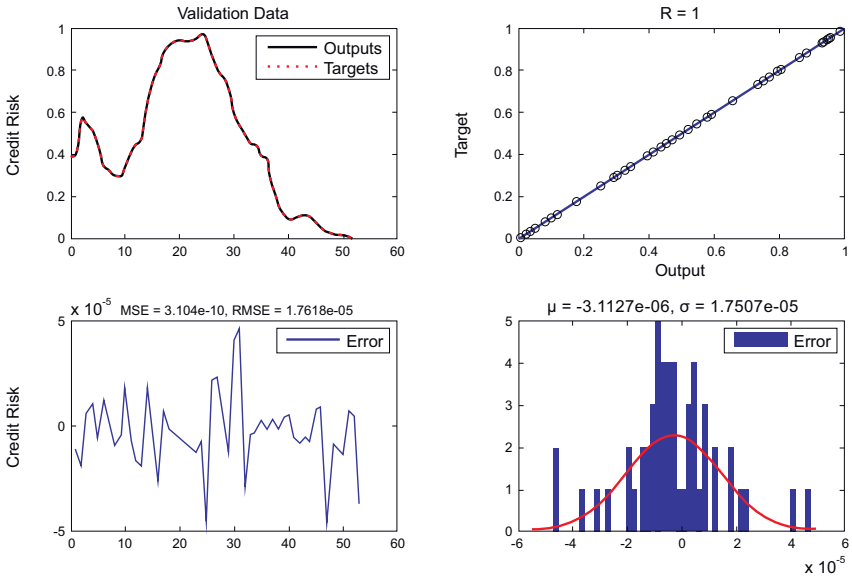


Figure 2. Assessment of the learning process: Validation based on the GA.

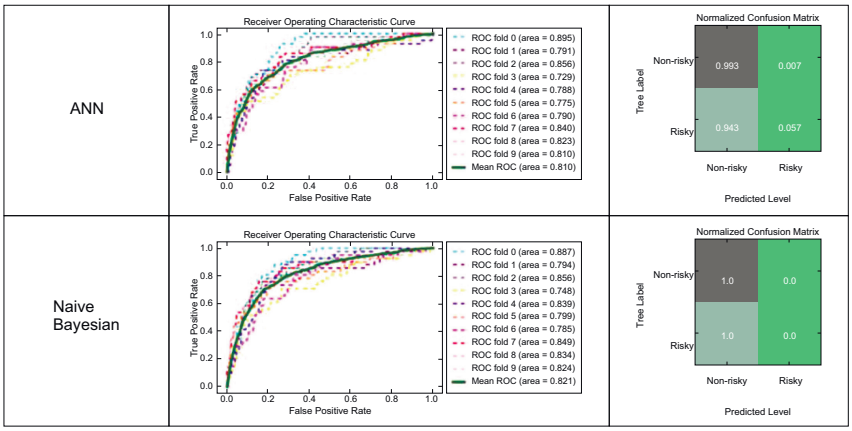


Figure 3. Classifier's ROC curve and confusion matrix of the ANN and NB.

Hence, compared with the NB, the ANN was more accurate and efficient as well as more promising for determining credit risk [30][58]. The Bayesian classification allocated the data classes into the tables where the values and attributes of an entity were predicted to be independent of others [31][32]. Because of the random nature of the ANN, the estimation accuracy was hugely dependent on the cases selected for training [34]. Therefore, the reported numbers could have changed slightly due to regular running. Similarly, the network structure becomes complex and quality is reduced because more time is needed to train the data[54].

This study addressed the issue of defining a systematic method to handle and precisely evaluate credit risk. The vagueness and ambiguity that characterize the credit risk concept complicate the formulation of an undisputable definition. Identifying the factors that

Table 3. Comparative analysis of BN and ANN performance by GA

Comparison Metric	Bayesian Networks	Artificial Neural Networks
Run Time	175s	6s
Training Data MSE	$9.1e^{-3}$	$13e^{-10}$
Validation Data MSE	$1.3e^{-2}$	$3.3e^{-10}$
Test Data MSE	$8.0e^{-3}$	$1.7e^{-10}$

Table 4. Comparison of classification accuracy

Algorithm used	Classification accuracy	Correctly classified cases	Incorrectly classified cases
Naïve Bayes	81.32%	1537/1890	353/1890
Neural Network	81.85%	1547/1890	343/1890

establish and influence credit risk to formulate a suitable functional form to estimate and predict its value is a difficult task [27]. Similarly, the spread and complexity of the credit risk phenomenon render traditional mathematical modeling techniques obsolete [50][37].

The present study proposed an approach that employs two of the most recent machine learning methods—Bayesian networks and ANNs—to address this concern. In the model, the variables were selected based on the data available from bank databases[52][55]. Despite the numerous capabilities of NBs and ANNs, machine learning methods, or a combination of them have been seldom used to approach the problem of credit risk measurement [54]. Therefore, this study bridged the existing gap that prevents intelligence systems from challenging bank modeling concerns. In particular, the focus was on the concept of insolvency as a characterization of credit risk [58]. As a result, inner factors were used to construct a model whose attributes permit prediction of credit risk issues. The case used bank data to demonstrate the accuracy, efficiency, flexibility, and rapidity of the data mining approaches when modeling events related to the measurement of credit risk. Implementation of NB and ANN can differentiate between the riskiest factors and estimate subject risk through the training and learning process. The results were highly consistent. Further, the binary outcomes gathered from the study depicted the ability of the NB–ANN approach to validate the findings through a parallel and independent implementation of the dataset.

4. Conclusion

The ANN algorithm based on backpropagation and the NB algorithm were implemented. Data were divided into categories to determine the best combination of the parameters. The best combinations were then observed, and the result was generated, gathered, and presented. The focus was not only on blending parameters but also on ensuring the accuracy of these combinations. As demonstrated, the optimum performance for ANNs was obtained when the linear regression was equivalent to 0.2, with mean accuracy of 85%. For the NB approach, the algorithm was first applied to four datasets in a single process and then the entire dataset was used to create a confusion matrix. The best outcomes of the iteration were sourced from the comparative approach. It was concluded that both the ANN and the NB models provide reliable outcomes, but the former is more effective for predicting credit risk with an average score of 82%. Future work may consider a comprehensive validation of the suggested method with other credit scoring databases identified by the high noise level sets method, with other methods such as forecasting routines used in the retail and consumer investment areas.

Acknowledgements

This work was supported by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020; by the Government of the Russian Federation, Grant No. 08-08; by National Council for Scientific and Technological Development (CNPq), Brazil, Grant No. 309335/2017-5]; by Ciência sem Fronteiras of CNPq, Brazil [Grant number 200450/2015-8].

Conflicts of Interest

There is no conflict of interest.

References

[1] Edward I Altman. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968.

[2] A. C. Antonakis and M. E. Sfakianakis. Assessing naïve Bayes as a method for screening credit applicants. *Journal of Applied Statistics*, 36(5):537–545, 2009.

[3] Amir F. Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks*, 12(4):929–935, 2001.

[4] B Baesens, T Van Gestel, S Viaene, M Stepanova, J Suykens, and J Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6):627–635, jun 2015.

[5] Alireza Baghban, Ali Jalali, Mojtaba Shafiee, Mohammad Hossein Ahmadi, and Kwok-wing Chau. Developing an ANFIS-based swarm concept model for estimating the relative viscosity of nanofluids. *Engineering Applications of Computational Fluid Mechanics*, 13(1):26–39, 2019.

[6] William H. Beaver. Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 4:71–111, 1966.

[7] Berk Bekiroglu, Hidayet Takci, and Utku Can Ekinci. Bank Credit Risk Analysis With Bayesian Network Decision. *IJAEST - INTERNATIONAL JOURNAL OF ADVANCED ENGINEERING SCIENCES AND TECHNOLOGIES*, 9(2):273–279, 2011.

[8] Leopold A. Bernstein. *Financial statement analysis : theory, application, and interpretation*. Irwin, 1993.

[9] Klaus Böcker. *Rethinking Risk Measurement and Reporting*, volume II. Risk Books, London, 2010.

[10] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, jul 1997.

[11] Xinying Zhang Chong Wu, Yingjian Guo and Han Xia. Study of Personal Credit Risk Assessment Based on Support Vector Machine Ensemble. *International Journal of Innovative Computing, Information and Control*, 6(5):2353–2360, 2010.

[12] R. H. DAVIS, D. B. EDELMAN, and A. J. GAMMERMAN. Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, 4(1):43–51, 1992.

[13] N Davutyan and S Özar. A credit scoring model for Turkey's micro & small enterprises (MSE's). In *13th Annual ERF Conference*, pages 16–18, Kuwait, 2006.

[14] Peter R. Demerjian. Financial Ratios and Credit Risk: The Selection of Financial Ratio Covenants in Debt Contracts. *AAA 2007 Financial Accounting & Reporting Section (FARS)*, 2007.

[15] Vijay S. Desai, Jonathan N. Crook, and George A. Overstreet. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95(1):24–37, nov 1996.

[16] Douglas Diamond. Financial intermediation and delegated monitoring. *Review of Economic Studies*, 51(3):393–414, 1984.

[17] Ilia D. Dichev and Douglas J. Skinner. Large-sample evidence on the debt covenant hypothesis. *Journal of Accounting Research*, 40(4):1091–1123, 2002.

[18] Alaa El-Shazly. Financial Distress and Early Warning Signals: A Non-Parametric Approach with Application to Egypt. In *9th Annual Conference of the Economic Research Forum*, number October, pages 1–25, Emirates, 2002.

[19] D J Hand and W E Henley. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Royal Statistical Society*, pages 523–541, 1997.

[20] James A Hanley and Barbara J McNeil. The Meaning and Use of the Area under a Receiver Operating (ROC) Curvel Characteristic. *Radiology*, 143(1):29–36, 1982.

[21] Martin F Hellwig. Risk aversion and incentive compatibility with ex post information asymmetry. *Journal of Economic Literature*, 438:1–25, 1998.

[22] Jih Jeng Huang, Gwo Hshiung Tzeng, and Chorng Shyong Ong. Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation*, 174(2):1039–1053, mar 2006.

[23] Huseyin Ince and Bora Aktan. A comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, 10(February 2015):233–240, 2009.

[24] Michael Jacobs and Nicholas M Kiefer. The Bayesian approach to default risk: A guide. *Center for Analytical Economics*, (March 2010), 2013.

[25] Vita Jagric, Davorin Kracun, and Timotej Jagric. Does non-linearity matter in retail credit risk modeling? *Finance a Uver - Czech Journal of Economics and Finance*, 61(4):384–402, 2011.

[26] Liu Jie and Song Bo. Naive Bayesian Classifier Based on Genetic Simulated Annealing Algorithm. *Procedia Engineering*, 23:504–509, jan 2011.

[27] J. W. Kay and D. M. Titterington. *Statistics and Neural Networks Advance at the Interface*. Oxford University Press, 2000.

[28] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, nov 2010.

[29] Adnan Khashman. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9):6233–6239, 2010.

[30] HC Koh, WC Tan, and CP Goh. A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business . . .* , 1(1):96–118, 2006.

[31] K Komorád. *On credit scoring estimation*. PhD thesis, Humboldt University, 2002.

[32] Adel Lahsasna, Raja Noor Ainon, and Teh Ying Wah. Credit scoring models using soft computing methods: A survey. *The International Arab Journal of Information Technology*, 7(2):115–123, 2010.

[33] Tian-Shyug Lee and I-Fei Chen. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, 28(4):743–752, may 2005.

[34] Pawel Lewicki and Thomas Hill. *Statistics : Methods and Applications*, volume 1st. 2006.

[35] Russell James Lundholm and Richard G. Sloan. *Equity valuation and analysis*.

[36] Dominik Maltritz and Alexander Molchanov. Economic Determinants of Country Credit Risk: A Bayesian Approach. In *12th New Zealand Finance Colloquium*, 2008.

[37] D. Martens, B.B. Baesens, and T. Van Gestel. Decompositional Rule Extraction from Support Vector Machines by Active Learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):178–191, 2009.

[38] Khalil Masmoudi, Lobna Abid, and Afif Masmoudi. Credit risk modeling using Bayesian network with a latent variable. *Expert Systems with Applications*, 127:157–166, aug 2019.

[39] Hamadi Matoussi and Aida Krichène Abdelmoula. Credit-risk evaluation of a Tunisian commercial bank: logistic regression vs neural network modelling. *International Journal of Accounting & Information Management*, 19(2):ijaim.2011.36619baa.005, jun 2011.

[40] Ricardas Mileris. Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian classifier. *Economics and management*, 15(1):1078–1084, 2010.

[41] Antonietta Mira and Paolo Tenconi. Bayesian estimate of credit risk via MCMC with delayed rejection. *Stochastic Analysis, Random Fields and Applications IV*, (January 2003):277–291, 2004.

[42] T M Mitchell. Generative and Discriminative Classifiers : Naive Bayes and Logistic Regression Learning Classifiers Based On Bayes Rule. In *Machine Learning*, page 432. McGraw-Hill Education, 2005.

[43] M.D. Odom and R. Sharda. A neural network model for bankruptcy prediction. *1990 IJCNN International Joint Conference on Neural Networks*, pages 163–168 vol.2, 1990.

[44] James A. Ohlson. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1):109, 1980.

[45] Krisna G. Palepu, Paul M. Healy, Victor L. Bernard, Sue Wright, Michael Bradbury, and Philip Lee. *Business Analysis and Valuation Using Financial Statements*. 2000.

[46] Belief Revision, Financial Distress, Sumit Sarkar, and Ram S. Sriram. Bayesian Models for Early Warning of Bank Failures. *Management Science*, 47(11):1457–1475, 2001.

[47] Bernard Rosner. *Fundamentals of biostatistics*. 2012.

[48] David E. Rumelhart and James L. McClelland. An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89(1):60–94, 1982.

[49] Okan Veli Safakli. Credit Risk Assessment for the Banking Sector of Northern Cyprus. *Banks and Bank System*, 2:21, 2007.

[50] Clifford W. Smith and Jerold B. Warner. On financial contracting. An analysis of bond covenants. *Journal of Financial Economics*, 7(2):117–161, 1979.

[51] Neda Soltani Halvaiee and Mohammad Kazem Akbari. A novel model for credit card fraud detection using Artificial Immune Systems. *Applied Soft Computing*, 24:40–49, 2014.

[52] Kent A. Spackman. Signal detection theory: valuable tools for evaluating inductive learning. *Proceeding Proceedings of the sixth international workshop on Machine learning*, pages 160–163, 1989.

[53] A. Steenackers and M. J. Goovaerts. A credit scoring model for personal loans. *Insurance Mathematics and Economics*, 8(1):31–34, mar 1989.

[54] Thomas Stibor. A study of detecting computer viruses in real-infected files in the n-gram representation with machine learning methods. *23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, 6096 LNAI(PART 1):509–519, 2010.

[55] Lili Sun and Prakash P. Shenoy. Using Bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, 180(2):738–753, jul 2007.

[56] Dennis Sweeney, David Anderson, and Thomas Williams. *Statistics for Business and Economics*. Thomson Learning EMEA, London, 7 edition, 2007.

[57] Anjan V Thakor. The Financial Crisis of 2007 – 2009 : Why Did It Happen and What Did We Learn ? *Review of Corporate Finance Studies Advance Access*, 4(2):1–51, 2015.

[58] Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, jan 2002.

[59] Chih-Fong Tsai. Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, 16:46–58, mar 2014.

[60] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1):223–230, jan 2011.

[61] David West, Scott Dellana, and Jingxia Qian. Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, 32(10):2543–2559, oct 2005.

[62] Dorota Witkowska, W Kaminski, Krzysztof Kompa, and Iwona Staniec. Neural networks as a supporting tool in credit granting procedure. *Information Technology for Economics & Management*, 2(1), 2004.

[63] Jiří Witzany. *Credit Risk Management Pricing, Measurement, and Modeling*. Springer, 2017.

[64] CL Wu and KW Chau. Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *Journal of Hydrology*, 399(3-4):394–409, 2011.

[65] Bee Wah Yap, Seng Huat Ong, and Nor Huselina Mohamed Husain. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, 38(10):13274–13283, sep 2011.

[66] Zaher Mundher Yaseen, Sadeq Oleiwi Sulaiman, Ravinesh C. Deo, and Kwok Wing Chau. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, 569(August 2018):387–408, 2019.

[67] Jozef Zurada, Niki Kunene, and Jian Guan. The Classification Performance of Multiple Methods and Datasets: Cases from the Loan Credit Scoring Domain. *Journal of International Technology and Information Management*, 23(1), 2014.

Chapter 6

Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation

This chapter consists in the following paper:

Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation

Germanno Teles, Joel J. P. C. Rodrigues, Ricardo A. L. Rabêlo and Sergei Kozlov

Software: Practice and Experience, Wiley, ISSN:1097-024X, pp. 1-9, May 2020.

DOI: doi.org/doi.org/10.1002/spe.2842

According to Journal Citation Reports published by Thomson Reuters in 2019, this journal scored ISI journal performance metrics as follows:

ISI Impact Factor (2019): 1.786

Journal Ranking (2019): 48/108 (Computer Science,Software Engineering)

Received: 13 December 2019 | Revised: 7 March 2020 | Accepted: 6 April 2020

DOI: 10.1002/spe.2842

SPECIAL ISSUE PAPER

WILEY

Comparative study of support vector machines and random forests machine learning algorithms on credit operation

Germanno Teles[1] | **Joel J. P. C. Rodrigues**[1,2,3] | **Ricardo A. L. Rabêlo**[2] | **Sergei A. Kozlov**[3]

[1]Instituto de Telecomunicações, Universidade da Beira Interior, Covilhã, Portugal

[2]Federal University of Piauí (UFPI), Teresina, Brazil

[3]ITMO University, St. Petersburg Russia

Correspondence
Joel J. P. C. Rodrigues, Federal University of Piauí, Campus Petrônio Portela, Av. Ministro Petrônio Portela, S/N - Bloco 8, Centro de Tecnologia 64049-550, Ininga, Teresina - PI, Brazil.
Email: joeljr@ieee.org

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: Grant No. 309335/2017-5, Process No. 200450/2015-8; Fundação para a Ciência e a Tecnologia, Grant/Award Number: Project UIDB/EEA/50008/2020; Government of Russian Federation, Grant/Award Number: Grant No. 08-08

Summary

Corporate insolvency has significant adverse effects on an economy. With the number of multinationals increasing rapidly, corporate bankruptcy can severely disrupt the global financial environment. However, multinationals do not fail instantaneously; objective strategies combined with a rigorous analysis of both qualitative and quantifiable data can go a long way in identifying an organization's financial risks. Recent advancements in information and communication technologies have made data collection and storage an easy task. The challenge becomes mining the appropriate data about a company's financial risks and implementing it in forecasting a company's insolvency probabilities. In recent years, machine learning has been incorporated into big data analytics owing to its massive success in learning complex models. Machine learning algorithms such as Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks, Gaussian Processes, and Adaptive Learning have been used in the analysis of Big Data to predict the financial risks of companies. In this paper, credit scoring is explored with regards to data processed using the collateral as an independent variable. The obtained results indicate that RF algorithm is promising for use in credit risk management. This research shows the advantages of the RF approach over the SVM algorithm are its speed and operational simplicity, and SVM has the benefit of higher classification accuracy than RF. The paper compares the SVM and RF algorithms to forecast the recovered value in a credit task. The execution of the projected intelligent systems uses tests and algorithms for authentication of the projected model.

KEYWORDS

Big Data, credit operation, machine learning, random forests, support vector machines

1 | INTRODUCTION

In recent years, the use of Machine Learning algorithms has been integrated in most contemporary credit risk prediction methods. There are many Machine Learning algorithms that have been incorporated into credit risk prediction models including Artificial Neural Networks (ANN), Naive Bayes, Dimensionality Reduction Algorithms, gradient Boosting Algorithms, and Decision Trees.[1] Different credit risk assessment models have been developed for the various models.

The use of Machine Learning algorithms in credit risk prediction for financial institutions raises numerous ethical debates. In 2017, a public forum from National Commission on Informatics and Liberty was held with the key topic of discussion being how personal data is used as well as the problems associated with Big Data and specifically the General Data Protection Regulation directive.[2] People are definitely afraid of how personal data is used and the fact that with Machine Learning the decisional power on this data will be placed on algorithms only serves to heighten these fears.

Choosing the appropriate machine learning algorithm depends on a couple of factors such as the size of data, the quality required as well as the nature of the data being analyzed.[3] Several questions also have to be answered before the machine learning technique is selected, such as what is the desired result? What is the translation technique to be implemented on the mathematical algorithm to generate commands in the subject computer? or How long does the algorithm have to operate to acquire sufficient learning ?[4-6]

Deplying an effective credit system to predict the recovery value presents many challenges. Some of these challenges can be recurrent with specific solutions; however, the general principles of credit risk solutions still apply. Machine learning has been widely used in many of these solutions by programming computers to learn without any special instructions.

Support Vector Machines (SVM) is usually described to produce better outcomes than other classifiers.[7] The Random Forests (RF) algorithm has demonstrated to manipulate huge dimensional data properly and is relatively resistant to overfitting.[8]

This study then aims to investigate two machine learning algorithms, SVM and RF, on their applicability, efficiency, flexibility, and accuracy in credit risk assessment. In this comparative study, two machine learning techniques are modeled, that is, RF and SVM,[9] focus on credit risk scoring and the impacts of distinct machine learning models to identify defaults by lenders. The main contribution of this study describes the use of SVM algorithm and the RF algorithm in addressing issues of recovery value loan provision.[10] A sample dataset from a bank was implemented using the algorithms with the objective of learning to classify and comparing a credit operation (recovery value), with different kernels and kernel parameters. Results for RF and SVM will be analyzed and compared for varying sets of data. The results from the various kernels are tuned with appropriate parameter settings. Besides the technical questions regarding understanding machine learning algorithms, referrals to the various discussions associated with confidentiality arising from the use of personal data in these algorithms are also discussed. It has been observed that multiple approaches can be implemented to address the fundamental objective of identifying the choice of features/variables, the algorithm and the corresponding criteria. In this digital era of Big Data, transparency is of utmost importance.[11] It is necessary that terms used in this field are ethical, clear, transparent, and appropriately defined. It is necessary to implement appropriate strategies to train data based on deep learning,[12] and machine learning algorithms and their implementation must be monitored to ensure accuracy.

The paper goes further to evaluate the stability of the two models based on the variables chosen. The method used by banks in making loans decisions is unclear; however, implementation of classical linear models in banking systems is adequately documented. For this paper, the transparent elastic approach was used as a benchmark.

The remainder of the paper is organized as follows. Section 2 presents a theoretical background, introduces related works, and discusses the problem of credit risk assessment on classifying the credit-scoring approach with SVM and RF. Section 3 analyzes the results, hence demonstrating which classifier is superior in terms of accuracy and discusses the issues of SVM, and RF models and draws a comparison between those two models. Finally, the conclusion and future scope are provided in Section 4.

2 | BACKGROUND AND RELATED WORK

Machine Learning is a subset of Artificial Intelligence that focuses on the development of strategies, methods, and algorithms. Therefore, the development of algorithms that enable a computer system to learn from the provided data and execute tasks and activities of design with sampling data together with performing tests on the new data. The field of machine learning has some strong links to statistics in various ways. There are multiple approaches and methods created for machine learning tasks. Neural Network techniques are widely used but have various limitations with regards to generalization, developing prototypes that usually get over fit with data.[12] This can be attributed to optimization procedures implemented for specific statistical approaches and parameter selection to determine the most appropriate model possible. This problem of credit risk assessment can best be solved with machine learning algorithms,[13] as follows:

1. Supervised learning: This type is composed of a target output variable (dependent variable) which is to be forecast from a set of predictors (independent variable). From these sets of attributes, a function mapping inputs to target outputs is generated. The training procedure remains in progress until the method attains the target level of precision. Examples include: RF, regression, decision tree and k-nearest neighbors.[14]
2. Unsupervised learning for this algorithm, there is no target output variable to forecast. This algorithm is mainly used to cluster a certain population into various groups to enable the segmentation of customers in different groups for a certain intervention. Examples of unsupervised learning include: K-means and a priori algorithms.[15]
3. Reinforcement Learning for this algorithm, the computers are trained to make particular decisions. The computer is exposed to an environment in which it can consistently train itself by implementing trial and error methods. The environment also provides an addition to rewards, unique numerical values that the agent attempts to maximize over time. The machine in the end learns from previous experiences and attempts to obtain the most suitable knowledge to make appropriate decisions. The best example of a reinforcement learning algorithm is the Markov Decision Process.[16]

The concept of empirical data modeling is appropriate for numerous applications in the field of computer science.[17,18] Empirical data modeling involves an induction procedure to create a model of the scheme from which it can derive responses of the system which are to be tried or observed. The observed data is finite and is considered a sample. This sampling is not uniform, and because of the great dimensional nature of the data, the input will be sparsely distributed. The problem is therefore often misrepresented.

SVMs were initially introduced to machine learning algorithms by Boser, Guyon, and Vapnik in 1992.[19] SVMs have since become a popular algorithm, especially with regard to handwritten digit recognition. Currently, SVMs are a crucial component of all research related to Machine Learning and are now regarded as a primary example of kernel methods.[20]

SVMs are a family of machine learning algorithms that discriminatively classify variables that were initially defined by a different hyperplane. In simpler terms, an input of labeled training data, the algorithm produces an output that consists of an optimal hyperplane categorizing the data based on a set of objective conditions. On a two-dimensional plane, this hyperplane is a line dividing the plane into two classes. For example, given two characteristics of an individual such as height and hair length, the two characteristics (variables) would first be plotted in two *n*-dimensional spaces with each point having two coordinates (support vectors) (Figure 1).

SVMs have been widely used in various machine learning applications such as facial recognition, target recognition, object identification, speaker identification, and handwritten digit recognition.[5,21]

RF is a machine learning algorithm proposed by Breiman.[8] It involves the construction of a forecast ensemble consisting of a set of decision trees growing in arbitrarily identified subspaces of data. Breimans concepts

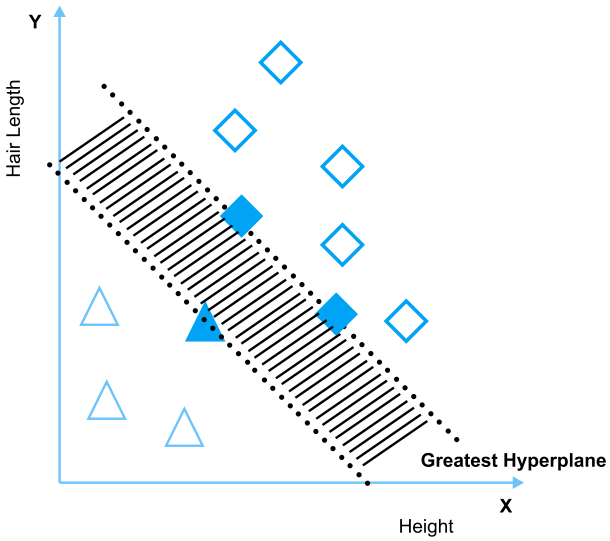


FIGURE 1 Illustration of support vector machine model
[Color figure can be viewed at wileyonlinelibrary.com]

were significantly subject to the prior scholarly works of Amit and Geman (1997)[19] on geometrical feature selection, the random subspace method developed by Ho in 1998.[22] Dietterich's (2000) casual split selection approach was also assimilated in the development of the RF approach.[23] Various empirical studies have shown that RF are increasingly becoming a strong competitor to some state-of-the-art approaches such as boosting (developed by Freund and Shapire in 1996) and Support Vector Machines.[24] The popularity of RF can be attributed to the fact that besides being fast and easy to implement, they also produce precise predictions and they can handle numerous input variables without overfitting.[25] They are considered among the most accurate machine learning techniques in the market.

In Breimans random trees method, each tree forming the collection is developed by first randomly choosing a small set of input coordinates (features or variables) to split on at each node and then calculating the best split with regards to these variables in the trained set. The Classification and Regression Tree methodology is then applied to grow the tree to its extreme size without trimming. This scheme of randomizing subsets is combined with bagging to resample with replacement, the training set of data whenever a new distinct tree grows.[26] In simple terms, the algorithm creates multiple decision trees and combines them to obtain a more precise forecast. Although the working of this approach appears simple, there are many driving forces involved in the mechanism making it challenge to analyze. In fact, the mathematical properties of RF remain mostly unknown to date, and most theoretical studies have focused on remote parts or stylized forms of the procedure.[27] Nonetheless, the arithmetic mechanism of true RF is yet to be entirely implicit and is still being explored.

Throughout this study, RF is assumed to comprise of a training model $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of independent and identically distributed (i.i.d.) $[0, 1]d \times R - valued$ random variables ($d \geq 2$) with a similar distribution to that of an independent simple pair (X, Y) satisfying $EY2 < \infty$. Space $[0, 1]^d$ has a standard Euclidean metric. For fixed $x \in [0, 1]d$, our objective is to approximate the regression function $r(x) = E[Y|X = x]$ by means of the data D_n. In other words, a regression function approximation rn is reliable if $E[r_n(X) - r(X)]^2 \Rightarrow 0$ as n $\Rightarrow \infty$.

2.1 | Study using the SVM

The objective of this approach is to implement the SVM learning technique to forecast the probability of loan defaults. The data needed training with each consisting of values for the set of input and output variables. The variables were chosen, and only those whose behavior was predictable were included. For the current comparative study, the data variables were considered the primary risk indicators and those borrowing were considered the subjects. The data collected for this study was from a bank, and it consisted solely of short-term loans since they make up the most significant share of loans. A dataset of 1890 credit files was sourced, and the subjects were classified as less risky or risky clients. The most constant variable is the probability of default together with a dummy variable, Y is equal to zero for a less risky client and one for risky clients. This means that $Y = 1$ when the repayment is delayed and $Y = 0$ when the payment is made in good time. SVM classifications were implemented to predict the potentials of class membership.

Before the SVM model was constructed, it was necessary to process the datasets in two standard deviation procedures. The data were split according to the average (mean) or assumed value, and all abnormal data was deleted. The final dataset consisted of 1890 samples with 1100 of them classified as good credits while 790 were classified as bad debts. For those in extremely bad credit positions, more than 3 years in default, they were classified as abnormal cases, and all their data was erased; this data represented only 50 samples. Since the quantity of the two sorts of samples is close, the SVM minimum requirements were met. The dataset is then divided into a training set and a test set. To show the learn and generalization capabilities of the SVM algorithm with regards to small samples, 30% (567 instances) of the sample data is chosen to build the SVM algorithm as a training sample. The remaining 70% (1323 instances) is randomly selected to test the generalization capability of the approach as a test dataset.

2.2 | Study using RF

The experiments for RF were conducted using Heuristic Lab and a modified RF Trees algorithm for classification. The key parameters to configure include: r, the ratio between 0 and 1; m, the number of variables; and nT, the number of trees. Appropriate selections of r and m impact the issue of noise tolerance in the training set meaning these parameters

need careful adjustment. Empirical tests for this research showed that r, m, and nT are the most appropriate parameters. In order to tune RF parameters, r and m were selected arbitrarily, and nine runs were analyzed for the various trees. The trees ranged from 50 to 500 with an increment of 50 for each run. It was determined that changing the parameters has no significant influence on the test performance and thus two tests were chosen for in-depth analysis.

3 | RESULTS ANALYSIS AND DISCUSSION

From the above analysis, a sample set (x, y) was constructed where $x = 4$ and the dimension y acts as an attribution sample. For good credits, $y = 1$ and $y = -1$ for bad credits. If the inner kernel function selects the polynomial kernel function or any other function, SVM is able to acquire the results of estimated performance and distribution of support vectors. The inner product function used for SVM in this paper is given by Equation (1) .

$$K(x_1, x_i) = exp\left\{-\frac{|x - x_i|^2}{\sigma^2}\right\}. \tag{1}$$

Figure 2 shows the results of the categorization model on the subject data over the 10-fold cross-validation with $R = 0.3$ and $M = 0.5$ (for the first) $R = 0.66$ and $M = 0.3$ (for the second).

The modified RDF algorithm makes the study of impacts of developed trees and variables easier. The relationship between nT and the performance of the model is not clear, but it becomes better with 500 trees. Table 1 below illustrates the mean, SD, median, maximum and minimum of the various measures used in this experiment.

Both SVM and RF models produced high evaluation results in the receiver operating characteristic (ROC) curve. However, because the classifications were not balanced, this indicator could not be used independently. The high frequency of hits from the majority class creates a bias in the results. Therefore, an alternative indicator is required, the Percentage True Correctly Classified (PTCC), to identify the level of accuracy in the class of interest. A proper assessment of this aspect shows a significant variation between the two projected models, in the range of 60% to 80%, with the best result realized by the SVM algorithm (Table 2). Table 2 presents the classifier model (ful l training set), 992 out of 1322 occurrences are correctly classified by the algorithm with 75.03% of accuracy. The table also gives four different statistical error measurements that measure the degree of relationship among the predicted and actual.

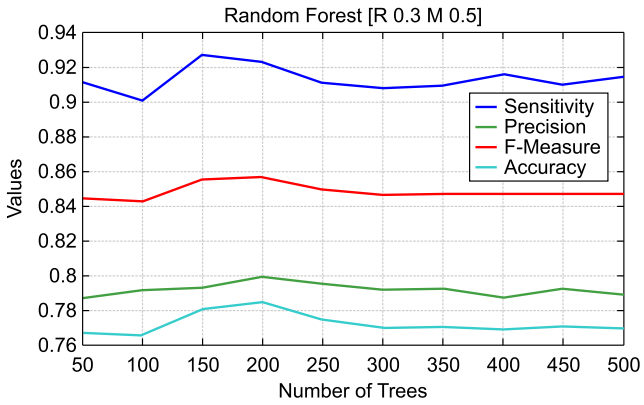


FIGURE 2 Results for the implantation of 10-fold cross validation for 10 runs ($r = 0.3$ and $M = 0.5$) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Statistics from the test data after implementation of the 10-fold cross-validation for 10 runs ($r = 0.3, M = 0.5$)

	Mean	SD	Median	Maximum	Minimum
Sensitivity	0.916	0.007	0.914	0.931	0.901
Precision	0.80	0.004	0.79	0.8	0.788
f-Measure	0.851	0.004	0.852	0.86	0.844
Accuracy	0.781	0.006	0.773	0.741	0.766

TABLE 2 Stratified cross-validation for support vector machines

		Training set	Total
Percentage true		992	75.03%
Incorrectly classified incidences		330	24.97%
Mean absolute error	0.3672		
Root mean squared error	0.5844		
Relative absolute error	63.22%		
Root relative squared error	96.4556%		
Total number of incidences		1322	100%

TABLE 3 Quality control of support vector machine (SVM) and random forest (RF) machine learning algorithms when used with various dimensionality reduction techniques

		Accuracy	Precision	Sensitivity	F1 Score
PCA	SVM	98.34%	100%	91%	96%
	RF	98.2%	100%	91.2%	93%
LDA	SVM	97.8%	100%	94.6%	97%
	RF	98%	100%	94.6%	97%
ISOMAP	SVM	98%	100%	94.6%	96.85%
	RF	98%	100%	94.3%	97%
Kernel PCA	SVM	98%	100%	94.6%	97%
	RF	98%	100%	91.2%	94%

The SVM and RF classifiers were analyzed using the error confusion matrix approach, which is representative of the whole thematic classification. The error confusion matrix can be used to determine the overall accuracy as well as specific endmember accuracy.[28]

The results of the error confusion matrices show that by using the RF classifier, a few of the endmembers are misclassified. However, the SVM algorithm produces more accurate classification results compared with RF(Table 3). The classifier accuracy procedure requires that the confusion matrix be representative of the entire mapped data area.[29] The results for accurately classified, unclassified, and incorrectly classified data can be obtained from the error confusion matrices. The overall accuracy of each algorithm is obtained by dividing the total number of accurate classifications by the total number of variables in the error confusion matrix. An error confusion matrix with all non-major diagonals having values of zero means that the classifier is 100% accurate. SVM showed the greater accuracy in this test having 53 (1.66%) unclassified incidences with the RF classifier having 58 (1.8%). For each classifier to construct its classification model, the RF classifier on average took 2.7 seconds while SVM took 27 seconds.

These data was sourced from a lending bank institution. The dataset included 1890 instances that were classified into two categories: 1100 good credit and 790 "bad/ defaulted credit." The initial dataset consisted of 16 variables that were classified into 10 qualitative and 6 numerical as shown in Table 4. The dataset used for this study was however processed to convert the original into 16 numerical variables, with the number 16 being an output variable.

All the classifications were conducted using a 10-fold cross-validation approach using three different tools, Heuristic Lab, Weka, and Keel. Default configurations were selected to create and test the models for comparison purposes. The algorithm used with Weka was the SVM using Linear Kernel.

The problems of credit risks have been consistently addressed in conferences on artificial intelligence.[30] The primary concern has been that personal data is used and there is also a growing fear that an algorithm could replace human beings in decision making. These questions are genuine, and this paper emphasizes the appropriate algorithms with regards to decision-making in lending institutions. Algorithms can be implemented to simplify a process while at the same time increasing its fluidity as well as increasing speed.[11] Algorithms include a set of code modelled to achieve set objectives. For example, an algorithm designed to perform a recruitment process introduces several discriminatory conditions based on individual profiles. A similar approach is adopted by lending institutions when making lending decisions to banks.[31] It is therefore of the essence to understand the underlying challenges and find ways to manage the use of algorithms.

TABLE 4 Original variables

Nž	Variable	Type
1	Contract value	Qualitative
2	Balance value	Qualitative
3	Collateral value	Qualitative
4	Number of collaterals	Qualitative
5	Recovery value	Qualitative
6	Value transformation rate	Qualitative
7	Value transformation interest	Qualitative
8	Value rate overdue	Qualitative
9	Client size	Qualitative
10	Main value delay	Qualitative
11	Seniority level	Qualitative
12	Percent used	Numerical
13	Duration in months	Numerical
14	Duration in years	Numerical
15	Duration in days	Numerical
16	Delay in days	Numerical

This study's primary objective was to compare the SVM and RF machine learning techniques based on performance in lending institutions and more precisely in the determination of recovery value. The credit risk concept is characterized by a vagueness that complicates the formulation of a limited definition for its identifying risk factors, and suitable functional forms to approximate and forecast its value is no easy task.[27] In similar fashion, the range and complexity of the credit risk concept render traditional mathematical models obsolete learning techniques.[28] This study compared the performance two approaches, the SVM and the RF approach in addressing the problem of credit risk assessment. In the study, the variables were chosen with regards to the data collected from the banks' records. Despite the numerous capabilities of support vector machines and RF prediction algorithms, the concern of credit risk estimation has barely been tackled with regards to machine learning algorithms let alone a combination of them. The current study, therefore, fills existing gaps that permeate intelligent systems from extremely puzzling bank modelling issues. The primary focus was on the idea of insolvency as a characterization technique of the credit risk.

Internal factors have been consequently implemented to construct models that permit the forecasting of credit risks. This case study that incorporated support vector machines on bank data exhibited high levels of precision, effectiveness and swiftness of the data mining techniques when modelling recovery value as a credit risk. The implementation of support vector machine and RF approach made it possible to establish the riskiest factors and estimation of the subject risk through the training and learning processes, and the outcomes were also very consistent. Additionally, the binary results from this study highlighted the capability of the support vector machines, and RF approaches to authenticate the findings via a parallel and uninfluenced application on the set of data.

The SVM algorithm was determined to be the most appropriate solution for determining recovery value. The ROC curves for SVM were generated and compared as in Figure 3. Both SVM and RF exhibited good performance in ROC curves; however, SVM was marginally better.

4 | CONCLUSION AND FUTURE WORK

Managing credit risks is important for the success of lending institutions. It is therefore of the essence to develop an effective aid for credit decision-making processes. The results of this study indicate that the RF algorithm is promising for use in credit risk management research. The main advantage of the RF approach over the SVM algorithm is its simplicity of operation. The fact that RFs takes much less time to construct its model means that it is more desirable in computerized

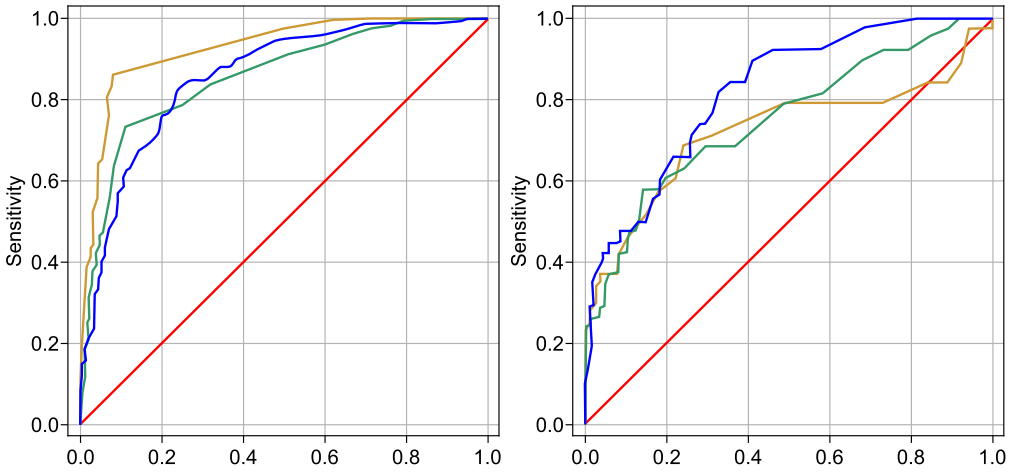


FIGURE 3 Receiver operating characteristic (ROC) curves for support vector machine (SVM) and random forests (RF). ROC curve for SVM (left) and ROC curve for RF (right) [Color figure can be viewed at wileyonlinelibrary.com]

environments. On the other hand, SVMs have the advantage of higher classification accuracy than RFs. A finding worth noting is the impact of injected randomness and the procedure of growing trees to produce optimal classification results. As has been experimentally proven, both algorithms are comparable, and each has a significant advantage over the other. It is, therefore, more advantageous to maximize on the advantages of each of these algorithms by using a hybrid model that will ensure the realization of highly accurate results that are also quite fast, rather than foregoing any of the advantages.

This being an initial study, it can further be enhanced to incorporate other predictive modelling approaches such as ANNs and Bayesian network model. Additionally, it is possible that the behaviors of loan default rates vary seasonally over a year due to circumstances that prevail over a year and the expenses tied to some specific dates throughout the year such as taxes and school fees, as well as seasonal variations in customer incomes. Future work will focus on developing more complex models accounting for seasonal variability.

More work can be done on how to grow the trees of RFs for optimal performance of decision trees, such as using multiple splitting ratios: 30/70, 40/60, 60/40, and 70/30. If the accuracies are fairly consistent then the algorithm is not sensitive to the number of samples. Hybrid models incorporating RF trees, SVMs, and other algorithms, also need to be thoroughly investigated and tested to develop more efficient systems. The future research opportunities that may be crucial for the prediction of credit risks may include the use of varying datasets, processing of the datasets to incorporate or remove various variables, research on the impact of each variable on the test performances, and modelling of varying problems.

ACKNOWLEDGEMENTS

This work was supported by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/EEA/50008/2020; by the Government of Russian Federation, Grant No. 08-08; by Brazilian National Council for Scientific and Technological Development (CNPq) via Grant No. 309335/2017-5; and by Ciência sem Fronteiras of CNPq, Brazil, through the Process No. 200450/2015-8.

ORCID

Joel J. P. C. Rodrigues https://orcid.org/0000-0001-8657-3800

REFERENCES

1. Lin WY, Hu YH, Tsai CF. Machine learning in financial crisis prediction: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev*. 2012;42(4):421-436. https://doi.org/10.1109/TSMCC.2011.2170420.
2. Stalla-Bourdillon S, Knight AE. Data Analytics and the GDPR: Friends or Foes? A Call for a Dynamic Approach to Data Protection Law. In: Leenes R, van Brakel R, Gutwirth S, de Hert P, eds. *Data Protection and Privacy: The Internet of Bodies*. Oxford: Hart Publishing; 2018;(March):1-22.

3. Wu X, Kumar V, Ross QJ, et al. Top 10 algorithms in data mining. *Knowl Inform Syst*. 2008;14(1):1-37. https://doi.org/10.1007/s10115-007-0114-2.
4. Crone SF. Finlay Steven. Instance sampling in credit scoring: an empirical study of sample size and balancing. *Int J Forecast*. 2012;28(1):224-238. https://doi.org/10.1016/J.IJFORECAST.2011.07.006.
5. Diederich J. *Rule Extraction from Support Vector Machines*. New York, NY: Springer; 2008.
6. Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Stat Methods Med Res*. 2016;25(5):1804-1823. https://doi.org/10.1177/0962280213502437.
7. Fassnacht FE, Latifi H, Stereńczak K, et al. Review of studies on tree species classification from remotely sensed data. *Remote Sens Env*. 2016;186:64-87. https://doi.org/10.1016/J.RSE.2016.08.013.
8. Breiman L. *Random Forests*. New York, NY: Kluwer Academic Publishers; 2001:5-32.
9. Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J Finance*. 1968;23(4):589-609. https://doi.org/10.2307/2329297.
10. Bekiroglu B, Takci H, Ekinci UC. Bank credit risk analysis with bayesian network decision. *IJAEST Int J Adv Eng Sci Technol*. 2011;9(2):273-279.
11. Antonakis AC, Sfakianakis ME. Assessing Naïve Bayes as a method for screening credit applicants. *J Appl Stat*. 2009;36(5):537-545. https://doi.org/10.1080/02664760802554263.
12. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn*. 1997;30(7):1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2.
13. Hachicha W, Ghorbel A. A survey of control-chart pattern-recognition literature (1991-2010) based on a new conceptual classification scheme. *Comput Ind Eng*. 2012;63(1):204-222. https://doi.org/10.1016/j.cie.2012.03.002.
14. Kabari LG, Nwachukwu EO. Credit risk evaluating system using decision tree – Neuro based model. *Int J Eng Res Tech*. 2013;2(6):2738-2745.
15. Pereira S, Portela F, Santos MF, Machado J, Abelha A. Predicting type of delivery by identification of obstetric risk factors through data mining. *Proc Comput Sci*. 2015;64:601-609. https://doi.org/10.1016/j.procs.2015.08.573.
16. Aven T. Risk assessment and risk management: review of recent advances on their foundation. *European J Operat Res*. 2016;253(1):1-13. https://doi.org/10.1016/j.ejor.2015.12.023.
17. Witzany J. *Credit Risk Management Pricing, Measurement, and Modeling*. New York, NY: Springer; 2017.
18. Fitch D. Structural equation modeling the use of a risk assessment instrument in child protective services. *Dec Supp Syst*. 2007;42(4):2137-2152. https://doi.org/10.1016/j.dss.2006.05.008.
19. Boser E, Vapnik N, Guyon IM. A training algorithm for optimal margin classifiers. Paper presented at: Proceedings of the 5th Annual Workshop on Computational Learning Theory; 1992:144-152; ACM.
20. Zhou H, Wang J, Wu J, Zhang L, Lei P, Chen X. Application of the hybrid SVM-KNN model for credit scoring. Paper presented at: Proceedings of the 2013 9th International Conference on Computational Intelligence and Security; 2013:174-177; IEEE.
21. Jiao L, Wang L, Gao X, Liu J, Wu F. Advances in natural computation. Paper presented at: Proceedings of the 2nd International Conference, ICNC 2006; September, 2006:24-28.
22. Davutyan N, Özar S. A credit scoring model for Turkey's micro & small enterprises (MSE's). Paper presented at: Proceedings of the 13th Annual ERF Conference; 2006:16-18; Kuwait.
23. Marqués AI, García V, Sánchez JS. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Syst Appl*. 2012;39(11):10244-10250. https://doi.org/10.1016/j.eswa.2012.02.092.
24. Biau G. Analysis of a random forests model. *J Mach Learn Res*. 2012;13:1063-1095.
25. Sweeney D, AD, Williams T. *Statistics for Business and Economics*. 7th ed. London, UK: Thomson Learning EMEA; 2007.
26. Lessmann S, Baesens B, Seow HV, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European J Operat Res*. 2015;247(1):124-136. https://doi.org/10.1016/j.ejor.2015.05.030.
27. El-Shazly A. Financial distress and early warning signals: a non-parametric approach with application to Egypt. Paper presented at: Proceedings of the 9th Annual Conference of the Economic Research Forum, Sharjah, UAE; October, 2002:1-25; Emirates.
28. Provost Foster, Fawcett Tom. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. Proceedings KDD-97; 1997:43-48.
29. Adams JB, Sabol DE, Kapos V, et al. Classification of multispectral images based on fractions of endmembers: application to land-cover change in the Brazilian Amazon. *Remote Sens Env*. 1995;52(2):137-154. https://doi.org/10.1016/0034-4257(94)00098-8.
30. Fawcett T. An introduction to ROC analysis. *Pattern Recog Lett*. 2006;27(8):861-874. https://doi.org/10.1016/j.patrec.2005.10.010.
31. Boardman, JW. Automating spectral unmixing of AVIRIS data using convex geometry concepts; 1993.

How to cite this article: Teles G, Rodrigues JJPC, Rabêlo RAL, Kozlov SA. Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Softw Pract Exper*. 2020;1–9. https://doi.org/10.1002/spe.2842

Chapter 7

Conclusion and Future Work

This chapter presents the main conclusions of the work performed in the course of this thesis. Besides, it shows important research topics to be considered as future works.

7.1 Final Remarks

This thesis project has described the importance of the use of smart DSS to help decision-makers in credit operations. This research is based on recent studies and identifies approaches and technologies used in several knowledge areas. The study about the use of these systems on credit scoring is vast, and it was divided into sub-topics, such as knowledge-based systems, data-based systems, and model-based systems. During the preparation of this document, several papers about these sub-topics are analyzed, enumerating the different methods.

Chapter 2 presented the survey paper entitled ”Classification Methods Applied to Credit Scoring with Collateral.” The paper provides an in-depth review of the state-of-the-art of the analysis includes 84 studies in this work to propose a using statistical methodology to conduct a meta-analysis comparing the results of classification methods. The result shows that SVM is the most commonly used classifier for credit scores, and while the system performs well, it does not apply approaches with collateral. This paper presents an in-depth review and evaluation of existing credit scoring classifier methods. An extensive review of literature focusing on classification methods applied in credit scoring is performed. It focuses mainly on methods for classifying an applicant for credit operations with the sufficiency of collateral, but we shall also briefly consider other associated problems in the credit industry, to their likely repayment behavior (e.g. ’default’ or ’not default’ with repayments). The analysis in this work proposes the use of statistical methodology to conduct a meta-analysis to compare the results of classification methods. It shows some cases that consider various probability distributions and also survival data. It also elaborates that collateral is not the first approach for credit scoring. There is a satisfactory statistic available for the collateral, the posterior probability distribution depends on the data only through this statistic, and thus, in many cases, we can reduce our data without loss of information. The general result shows that collateral is not the first approach for credit scoring, but when is used can be a high value on methods of model classification.

Chapter 3, entitled “Machine Learning and Decision Support System on Credit Scoring”, is compares the credit scoring performance of fuzzy sets and decision trees based on an artificial neural network to predict the recovered value. This paper is an initial study of collateral as a variable in the calculation of the credit score. Fuzzy logic makes some implicit assumptions that may make it even harder for credit grantors to follow a logical decision-making process. The study concludes that both models enable modeling uncertainty in the credit-scoring process. Therefore, mechanisms that help to characterize such complex scenarios are needed. The literature suggests that the application of multiple criteria for decision making can facilitate

the resolution of these issues. Other models, such as fuzzy logic, artificial neural networks, and decision tree models explicitly consider the underlying relationships and recognize the uncertainties (such as operational risks). In addition to the multi-criteria methods, additional complementary tools such as fuzzy sets or numerical simulations are increasingly being used in the credit-scoring process. Although fuzzy logic is more difficult to implement, it models more accurately the uncertanty.

Chapter 4, entitled "Support System on Credit Operation Using Linear and Logistic Regression," aims to understand how predictive models can provide different estimations of expected recovery based on the same data sets. Predictive analytics, which is the method of obtaining knowledge from existing data sets to decide guides and predict future outcomes and trends, including classification techniques and regression techniques. Classification techniques such as decision tree analysis, statistical analysis, neural networks, support vector machines, case-based reasoning, Bayesian classifiers, genetic algorithms, and rough sets help identify patterns in large unstructured data sets and generate cluster sets. Regression techniques include linear regression and logistic regression. A simple logistic regression model can easily be extended to a multiple logistic regression model by integrating more than one prediction variable, which indicates increasing difficulty in obtaining multiple observations with an increasing number of independent variables.

Chapter 5, entitled "Artificial Neural Network and Bayesian Network Models for Credit Risk Prediction", compares Bayesian networks with artificial neural networks for predicting recovered value in a credit operation. The study explores this problem and finds that ANNs is a more efficient tool for predicting credit risk than the naïve Bayesian (NB) approach. ANNs was used to study the nervous system and the way the brain processes information. An ANN entails processing algorithm to model the brains of humans and comprises a large number of interconnected nodes (neurons) working as a system to solve pattern recognition or data classification problems. The present study demonstrates that various algorithms can be used in parallel to address the issue in question, which in the case presented here is loan provision. Multiple strategies for identifying the choice of features (or variables), algorithm, and criteria can provide a solution. For instance, in the new Big Data and digital era, transparency is critical. Strategies based on deep learning are also necessary to train data on the application, and machine learning algorithms and their use must be regulated to ensure accuracy. The comparative approach was used to generate the best results of the iterations. The findings show that both ANN and NB models provide credible outcomes, but the ANN model is more valuable for predicting credit risk.

Chapter 6 "Comparative Study of Support Vector Machines and Random Forests Machine Learning Algorithms on Credit Operation", showed advantages of the RF approach over the SVM algorithm which are its speed and operational simplicity, and SVM has the benefit of higher classification accuracy than RF. The objective of this approach is to implement the Support Vector Machine learning technique to forecast the probability of loan defaults. The data needed training with each consisting of values for the set of input and output variables. The variables were chosen, and only those whose behavior was predictable were included. For the current comparative study, the data variables were considered the primary risk indicators and those borrowing were considered the subjects. The data collected for this study was from a bank, and it consisted solely of short-term loans since they make up the most significant share of loans. A dataset of 1890 credit files was sourced, and the subjects were classified as less

risky or risky clients. The most constant variable is the probability of default together with a dummy variable, Y is equal to zero for a less risky client and one for risky clients. This means that Y = 1 when the payment is delayed and Y = 0 when the payment is made in good time. SVM classifications were implemented to predict the potentials of class membership.This study's primary objective was to compare the SVM and RF machine learning techniques based on performance in lending institutions and more precisely in the determination of recovery value. The credit risk concept is characterized by a vagueness that complicates the formulation of a limited definition for its identifying risk factors, and functional forms suitable to approximate and forecast its value is no easy task. In similar fashion, the range and complexity of the credit risk concept render traditional mathematical models obsolete learning techniques. This study compared the performance of both approaches, the SVM and the RF approach in addressing the problem of credit risk assessment. In the study, the variables were chosen with regards to the data collected from the banks' records. Despite the numerous capabilities of support vector machines and random forests prediction algorithms, the concern of credit risk estimation has barely been tackled with regards to machine learning algorithms let alone a combination of them. The current study, therefore, fills existing gaps that permeate intelligent systems from extremely puzzling bank modelling issues. The primary focus was on the idea of insolvency as a characterization technique of the credit risk. The paper compares the SVM and RF algorithms to forecast the recovered value in a credit task. The execution of the projected intelligent systems uses tests and algorithms for authentication of the projected model.

The work presents the importance of qualitative and quantitative variables in the process of granting credit to banks and thus proposes an alternative methodology for financial institutions, based on sufficiency and especially their liquidity. The prediction of the recovery of collateral creates great expectations, especially when smart classifiers are used. This approach makes it possible to identify patterns with maturity to update asset data in the context of the guarantee, provided by financial institutions through the mapping of all guarantee related variables. Moreover, the creation of those mechanisms for recovery in Credit Score Systems using those parameters of the support and classifier support processes were applied to machine learning concept.

The main objective of this thesis was to create a model that uses contextual information to assess whether collaterals allow the recovery of credit granted in the decision making process that can help decision-makers in the credit operation. The impairment is related to the devaluation of assets or securities due to the lack of buyers or the excess of supply, as can occur in the real estate industry. In certain areas, prices for new apartments may not be "on par" with previous prices. Thus, we will say that the new prices are in a situation of impairment. The Client-Risk Assessment Model adopted by the banks has the function of classifying clients according to the level of risk, providing greater security for credit decisions and other financial and banking services, without becoming an element that inhibits the Institution's competitive power in the market. For this purpose, the research is based on previous studies, seeking to detect a viable approach. The systems investigated in each subtopic present us with data mining techniques as a necessary solution, and show the need to find the best classifier that will be used to meet the objectives of this research.

7.2 Future Work

To conclude this thesis, the following directions of future investigations that resulted from the developed work are suggested:

- Perform new literary research to analyze another way to measure a meta-analysis, such as from risk rate or risk difference, moreover, high-quality research is eventually published in scientific journals, other forms of publication may be included in this list in the future investigations. Notwithstanding these limitations, our systematic review provides important insights into the research literature on classification techniques applied to credit scoring and how this area has been moving over time;
- The particular strength of artificial-neural-network-based decision trees is their tendency to help comprehend sequential decisions and outcome dependencies. The model can play a complementary role to other scoring tools such as fuzzy assets, whereby the classes it creates can be used as fuzzy sets. However, a decision tree algorithm requires that the target attribute has only discrete values. Another disadvantage is that it performs poorly in terms of complex interactions in which the decision trees are redrawn every time new data are added to the model. In addition, decision trees are over-sensitive to the training set, irrelevant attributes, and noise. Create a model using fuzzy can be integrated, for example, neural networks, leading to higher prediction accuracy;
- Since numerous factors may have an impact on affordability and over-indebtedness, it is challenging to make predictions for the future. Affordability assessment is often based on application data, credit reports, and estimation of expenditure. Little information on the implemented affordability models is available in the public domain, except for the solutions offered by credit bureaus. There is even less information on models for credit operations. The existing literature on affordability and over-indebtedness models is also sparse. Nevertheless, a dynamic approach to affordability assessment may be preferred that takes into account possible changes in both income and expenditure and enables predicting for the future. Create comprehensive validation of suggested methods with another credit score database and be able to make comparisons;
- The proposed method is helpful for identifying a new way that has a strong possibility of being more advanced than similar previous classifiers. Second, in the proposed method, properties and functions were manually categorized into representative groups, but a future topic will automate this task using semantic technologies. Apply the concept of natural language processing (NLP) to analyze the sentiment added to the credit score.

Appendix A

Decision Support Systems to Predict a Sufficiency of Collateral for Credit Risk Operations

This appendix consists in the following poster:

Decision Support Systems to Predict a Sufficiency of Collateral for Credit Risk Operations

Germanno Teles, Joel J. P. C. Rodrigues

Poster Ciência 2018 - Science and Technology in Portugal Summit.

Decision Support Systems to Predict a Sufficiency of Collateral for Credit Risk Operations

Germanno Teles,
Joel J. P. C. Rodrigues
Instituto de Telecomunicações, University Beira Interior, Portugal;
National Institute of Telecommunications (Inatel), Brazil; Instituto de Telecomunicações, Portugal, ITMO University, Russia; University of Fortaleza (UNIFOR), Brazil

INTRODUCTION

- Basel Standards.
- The risk of non-payment.
- Probability of Default (PB) models.
- Kinds of Collateral
  - Fiduciary or Real

Problems

I. Which are the classifiers in the use of credit risk analysis and their performance on Decisions Support Systems (figure 1)?
II. Which classifiers use collateral as a parameter for calculating risk?
III. How to estimate the recoverability of a credit operation with a collateral as asset?

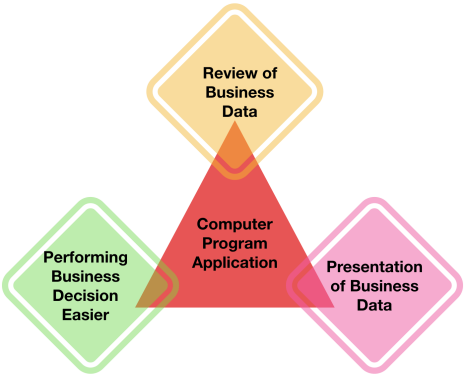


Figure 1: Decision Support Systems Proposal

METHODS

Models for measuring and estimate the probability of a customer becoming default.

- Credit Scoring
- Behavioural Scoring

An evaluation of existing credit scoring classifiers methods to choose the best one that will be used as a reference to evaluate and validate the proposed contributions. An extensive review of literature through the use of classification methods applied in credit scoring (Table 2)

CLASSIFIERS	Number of Papers
Bayesian Network	20
Neural Networks	19
Decision Tree Neuro-based	15
Fuzzy logic	16
Support Vector Machine	26
Logistic Regression	13
Linear Regression	13
Combined	17

Table 1 Summary of credit scoring classifiers

CLASSIFIERS USING COLLATERAL
Bayesian Network
Neural Networks
Decision Tree Neuro-based
Logistic Regression

Table 2 Summary of credit scoring classifiers using collateral

RESULTS

A review of the studies presented in Table 2 reveals that Bayesian Network, Neural Network, Decision Tree Neuro-based, and Logistic Regression are using collateral as a variable to improve the measurement of credit risk. Among those techniques, it can emphasize Support Vector Machine in Table 3 is a dominant classifier in credit scoring

CLASSIFIER DOMINANT
Support Vector Machine

Table 3 Summary of credit scoring classifiers dominant

CONCLUSIONS

Described the importance of the use DSS to help decision-makers in credit operations. This research is based on recent studies and identifies approaches and technologies used in several areas of the knowledge. In summary, we have performed a study of the current and essential condition for a decision support systems starting from the Basel Agreement until now, this work makes clear the importance of DSS system and the "intelligence" involved in them.

Appendix B

Intelligent Decision Support System on Credit Scoring

This appendix consists in the following poster:

Intelligent Decision Support System on Credit Scoring

Germanno Teles, Joel J. P. C. Rodrigues

Poster IEEE UBI Student Branch.

Intelligent Decision Support System on Credit Scoring

Germanno Teles, Joel J. P. C. Rodrigues
Instituto de Telecomunicações, University Beira Interior Covilhã, Portugal;
Instituto Nacional de Telecomunicações (Inatel), Brazil, Portugal

INTRODUCTION

Intelligent Decision support systems (IDSS) and Knowledge-based Expert Systems (KES) are two of the stronger general paths on Machine Learning (ML). The mechanization of auditing rules just implements true and high-speed purposes that take over part of the responsibility of the auditors, in particular, tasks that are usually routine and are prone to error if accounts lose concentration. Credit scoring is considered an important tool to pre-qualify borrowers and assist managers to make better risk decisions to the business. Between statistics and artificial intelligence (AI) there are several techniques used in credit scoring, such as Decision Trees, Neural Networks, Linear Regression, Logistic Regression, Bayesian Networks, Support Vector Machine (SVM), etc.

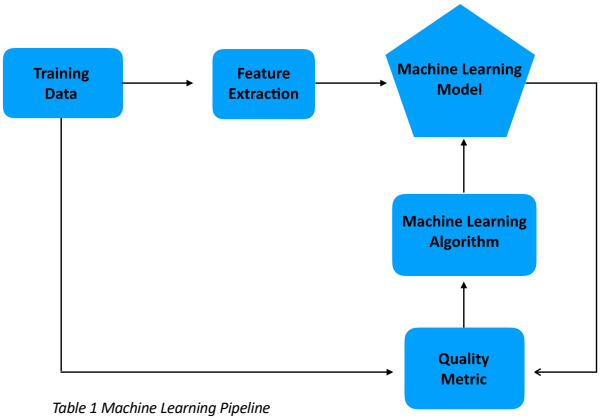


Table 1 Machine Learning Pipeline

Problems
I. Which are the classifiers in the use of credit risk analysis and their performance on Decisions Support Systems ?
II. Which classifiers use collateral as a parameter for calculating risk?
III. How to estimate the recoverability of a credit?

CLASSIFIERS	Number of Papers
Bayesian Network	20
Neural Networks	19
Decision Tree Neuro-based	15
Fuzzy logic	16
Support Vector Machine	26
Logistic Regression	13
Linear Regression	13
Combined	17

Table 2 Summary of credit scoring classifiers

METHODS

An extensive review of literature through the use of classification methods applied in credit scoring. Weighting of guarantees is the last aspect that indicates the quality (sufficiency and liquidity) that the guarantees must mitigating effect.

CONCLUSIONS

Described the importance of a new IDSS approach focusing in help the credit operations on financial institutions. Perform an evaluation and validation of the classifiers with a real dataset.

UBIsym in Healthcare Engineering, Covilhã, Portugal, July 24th 2018

Appendix C

Using Natural Language Processing with Sentiment Analysis to Improve Credit Score on Collateral Reports

This appendix consists in the following poster:

Using Natural Language Processing with Sentiment Analysis to Improve Credit Score on Collateral Reports

Germanno Teles, Joel J. P. C. Rodrigues

Poster LxMLS 2019, 9th Lisbon Machine Learning School

Using Natural Language Processing with Sentiment Analysis to Improve Credit Score on Collateral Reports

Germanno Teles[1],
Joel J. P. C. Rodrigues[2,3,4]
[1]Instituto de Telecomunicações, University Beira Interior, Portugal;
[2]National Institute of Telecommunications (Inatel), Brazil; [3]Instituto de Telecomunicações, Portugal,
[4]Federal University of Piauí, Teresina - PI, Brazil

- Sentiment Analysis is a field of NLP;
- Sentimental Analysis is helping Institutions enhance their policies;
- Can reach the goal of gaining better insight into different areas;
- Classifying the information expressed on reports using Long Short Term Memory networks(LSTM)

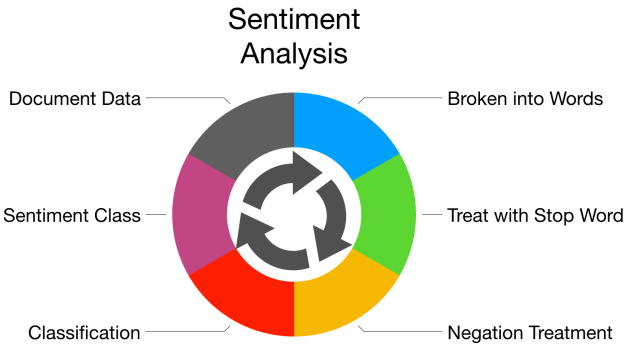


Figure 1: Sentiment Analysis Approach

Challenges
I. Avoid ambiguous words
II. Noise annotation
III. How to treat gap to learn to connect the information

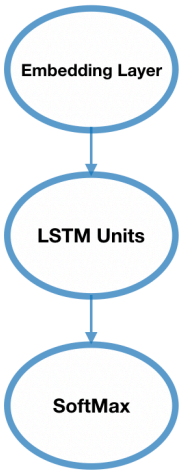


Figure 2: LSTM network

A conventional neural network is incapable of learning from previous experiences because the information does not pass from one step to the next. On the contrary, Recurrent Neural Networks(RNN) learns information from the immediately previous step. If was use Recurrent Neural Network (RNN). An evaluation of existing credit scoring classifiers methods to choose the best one that will be used as a reference to evaluate and validate the proposed contributions. Unfortunately, as that gap grows between the relevant information and the point where it is needed to become very large, RNNs become unable to learn to connect the information. LSTM can resolve this problem because it uses gates to control the memorizing process

LSTM beats the other models when we want our model to learn from long term dependencies. LSTM's ability to skip, retrieve and renew the information pushes it one step ahead of RNNs.

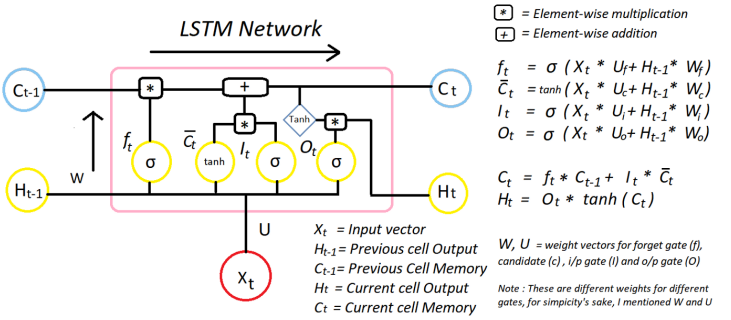


Figure 3: Diagram for a LSTM cell at T time step(Source : http://colah.github.io/posts/2015-08-Understanding-LSTMs)

In summary, we have performed a study of real data of collateral data for a decision support systems starting from the Basel Agreement, this work makes clear the importance of DSS system and the "intelligence" involved in them.