Spring 5-25-2017

CREDIT SCORING USING LOGISTIC REGRESSION

Ansen Mathew
San Jose State University

Computer Science

Ansen Mathew

has passed the defense for the project

CREDIT SCORING USING LOGISTIC REGRESSION

Digitally signed by Leonard Wesley (SJSU)
DN: cn=Leonard Wesley (SJSU), o=San Jose State University, ou, email=Leonard.Wesley@ssu.edu, c=US
Date: 2017.05.24 14:22:49 -07'00'

Advisor's Signature Dr. Leonard Wesley — Date: 05/24/2017

Robert Chun
Digitally signed by Robert Chun
DN: cn=Robert Chun, o=San Jose State University, ou=Computer Science, email=robert.chun@sjsu.edu, c=US
Date: 2017.05.18 18:07:45 -07'00'

Committee Member's Signature Dr. Robert Chun — Date: 05/18/2017

Raghavendra Keshavamurthy
Digitally signed by Raghavendra Keshavamurthy
DN: cn=Raghavendra Keshavamurthy, c=US, o=SAP, ou=SAP, email=raghavendra.keshavamurthy@sap.com
Date: 2017.05.18 18:30:11 -07'00'

Committee Member's Signature Mr. Raghavendra Keshavamurthy — Date: 05/18/2017

NOTE: The advisor should send the final report to the graduate coordinator so that the student can be cleared for graduation

San José State UNIVERSITY

Form Revised 06/14/11

CS 298 Final Project Report

CREDIT SCORING USING LOGISTIC REGRESSION

A Project Report

Presented to

The Department of Computer Science

San Jose State University

In Partial Fulfillment

of the Requirements for the

Computer Science Degree

by

Ansen Mathew

May, 2017

The Designated Project Report Committee Approves the Project Report Titled

Credit Scoring using Logistic Regression

by

Ansen Mathew

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

May 2017

Dr. Leonard Wesley Signature: _____
Department of Computer Science

Dr. Robert Chun Signature: _____
Department of Computer Science

Mr. Raghavendra Keshavamurthy Signature: _____
Project Leader, SAP

ABSTRACT

This report presents an approach to predict the credit scores of customers using the Logistic Regression machine learning algorithm. The research objective of this project is to perform a comparative study between feature selection and feature extraction, against the same dataset using the Logistic Regression machine learning algorithm. For feature selection, we have used Stepwise Logistic Regression. For feature extraction, we have used Singular Value Decomposition (SVD) and Weighted Singular Value Decomposition (SVD). In order to test the accuracy obtained using feature selection and feature extraction, we used a public credit dataset having 11 features and 150,000 records. After performing feature reduction, Logistic Regression algorithm was used for classification. In our results, we observed that Stepwise Logistic Regression gave a 14% increase in accuracy as compared to Singular Value Decomposition (SVD) and a 10% increase in accuracy as compared to Weighted Singular Value Decomposition (SVD). Thus, we can conclude that Stepwise Logistic Regression performed significantly better than both Singular Value Decomposition (SVD) and Weighted Singular Value Decomposition (SVD). The benefit of using feature selection was that it helped us in identifying important features, which improved the prediction accuracy of the classifier.

ACKNOWLEDGEMENTS

I am very grateful to my Project Advisor **Dr. Leonard Wesley** for his constant support and encouragement throughout the Master's project. His critical inputs helped me focus on the right path to complete this project.

I would also like to thank my committee members **Dr. Robert Chun** and **Mr. Raghavendra Keshavamurthy**, for their valuable time and suggestions during this project.

Last, but not least, I would like to thank my parents, my sister and friends for supporting and believing in me.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION AND MOTIVATION FOR CREDIT SCORING.

1.1 Credit Scoring, it's needs and benefits.

Credit is a very important product in banking and financial institutions. There is always a customer in need of a loan. Since Loans are always accompanied by risks, it is important to identify suitable applicants, and there have to be a means to determine and separate the good applicants from the bad. To solve this issue, financial institutions such as banks started developing credit scores. Using the customer's credit scores lenders can define the risk of loan applicants. By calculating the credit score, lenders can make a decision as to who gets credit, would the person be able to pay off the loan and what percentage of credit or loan they can get (Lyn, et al., 2002).

Lenders generally use "historical" data gathered from customers to build the scorecard for the applicants. They did this by gathering valuable information about candidates like the applicant's income, type of work, working current place, residual status, financial asset, time with the bank, credit history, if he/she had default or problem with payment. Credit scoring became widely used after the 1980s (Lyn, et al., 2002). In the past, only banks used credit scoring, but then it was extensively used for issuing credit cards, as another kind of loan. Currently, credit scoring is used in credit cards, club cards, mobile phone companies, insurance companies and government departments.

Credit scoring is beneficial from both the lenders and customers' point of view. From the bank's perspective, it helps them in evaluating potential clients and setting a credit limit based on their credit score. This helps the banks to avoid credit risk. Credit scoring is also a faster process in determining the credit worthiness of a customer, as compared to the traditional method which is time-consuming. From the

perspective of the client, they can keep on improving their credit score and extend their credit limit (Mester, 1997). Thus, credit scoring can help avoid unnecessary credit risk to both lender and customer.

As per (Mester, 1997), there are three main benefits of credit scoring. The main advantage of credit scoring is that each client is evaluated quickly. Also, since this system is automated, it results in a lot of cost savings to the lenders. As customers need to provide only the information used in the scoring system, applying for credit becomes easy to the customers. Also, this helps lenders to implement the same criteria in making credit decisions to all customers regardless of their gender, race, or other factors. Thus, this process is more objective for all customers and avoids discrimination in any form.

1.2 Types of credit scoring.

There are several credit score formulas in use, each having unique characteristics:

The FICO Score – The Fair Isaac Corporation has introduced the FICO score model which has now emerged as the most widely accepted credit scoring model in the industry.The FICO score scale runs between 300 to 850 points.

The FICO scores are not directly provided to the clients. Experian, TransUnion, and Equifax are the vendors who sell these scores to their customers. These credit agencies maintain the credit history and files of their clients. The credit score is determined based on the information present in the customer's file at that point in time.

The PLUS Score is another user-friendly credit score model which was developed by Experian with scores ranging from 330 to 830, to help customers understand how lenders view their creditworthiness. Higher scores represent a greater likelihood that the customers would pay back their debts and consequently be seen as being a

lower credit risk to lenders. During the time the client's information can change. Also, their credit score may be different from time to time.

(https://www.creditkarma.com/article/differentscores)

The Vantage Score- Vantage Score created by Experian, TransUnion, and Equifax is a new credit scoring model to support a consistent and accurate approach to credit scoring. This score provides lenders with nearly same risk assessment across all three credit reporting companies, and the Vantage scale ranges from 501 to 990.

No matter which scoring models banks use, it pays to have a good credit score as a customer with higher score gets approved with a lower rate of interest.

1.3 FICO Scoring Method

According to the FICO model analysis, most of the population has credit scores between 600 and 800. Also, a score of 720 or higher will enable a person to get the most favorable interest rates on a mortgage, as per the data from Fair Isaac Corporation. Two Percent of the total population has credit scores below 499 whereas, 5 percent have scores between 500-549 . 8 percent of the American people have scores between 550-599 , twelve percent have between 600-649 , fifteen percent have scores between 650-699 -15 percent , eighteen percent have credit scores in the range of 700-749 . Twenty-seven percent have excellent scores ranging from 750 to 799 whereas thirteen percent have a very good score range of 800 and above.

Statistical Models are used on the credit report of an applicant to determine their FICO score.The internal logic behind the FICO is kept confidential by the credit scoring agencies. However, five main factors are considered for developing FICO scores. They are the previous credit history, amount of loans, the amount of time credit has been in use and whether the person has applied for new credit, and the different types of credit held by the applicant.

2 LITERATURE REVIEW.

2.1 Credit Scoring Model based on Improved Tree augmentation Bayesian classification.

In this paper, (Fan, et al., 2013) have proposed a new Credit Scoring System based on Feature extraction and Bayesian Classification using improved tree augmentation. It first uses principal component analysis (PCA) to transform the features into a lower dimension and thereby simplify the network's inputs. After that, an improved Bayesian model is used for classification.

Building a Credit Scoring System

The following flowchart depicts the steps involved in building the model:

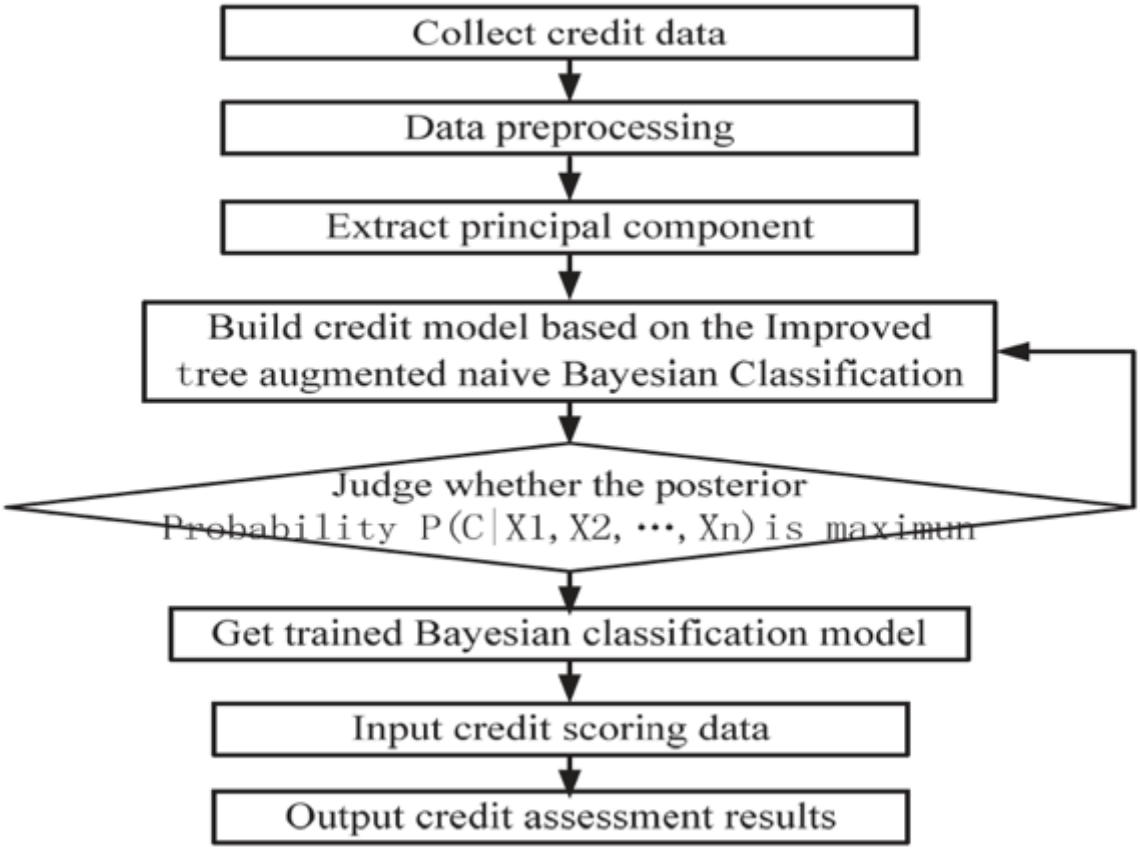

Figure 1:Steps to build credit scoring Model

Analysis and Results:

For conducting the experiments, they have used the German credit data, which has around 1000 records. The data is divided such that 700 records predict the target variable as '0', which means that that person has a good credit score. While 300 records predict the target varaiable as '1', which means that the person has a bad credit score. After pre-processing and removing the outliers, they have used principal Component Analysis (PCA) to extract the principal component from the original features. These principal components are then passed into the Bayesian classification model, which is then used for building the model. The dataset is split up into training and test sets and the model is then scored against the test set. They achieved an accuracy of 78 percent after the analysis.

Conclusion:

The authors observed that after applying principal component analysis to the model, there was a 2 percent increase in accuracy from 76 percent to 78 percent.

As part of the future work, the authors posit that different machine learning algorithms could be used to improve the accuracy of the model. Also, the above method could be used in several different datasets and a comparative study could be performed on them, to determine how effective this approach is on different datasets.

2.2 Credit Scoring Decision Support System.

In this paper, (Dukic, et al., 2011) have used Logistic Regression machine learning algorithm as a model for building its decision support system.

Model Formulation

After the model, has been constructed, i.e. following the determination of logistic regression parameters, it is relatively simple to calculate the probability that the

analyzed loan applicant may default on the loan. To be fairer when making the assessment and the decision whether to approve a loan, it is necessary to consider a range of socio-demographic characteristics and financial char of the loan applicant (if the relational features are included in the model). Socio-demographic characteristics include the loan applicant's gender, age, education level, marital status and members of household. Among other things, financial indicators comprise the salary, other income, expenditures, debts and account balance. This kind of data is frequently not available to the bank, or at least not in a sufficiently long time series. Even when the bank has access to such data, they are only of historical significance and cannot predict future behavior of the loan applicant. Given that future values of the loan applicant's financial indicators cannot be estimated with certainty at the time when credit worthiness is assessed, it is questionable to what extent the probability of default is valid.

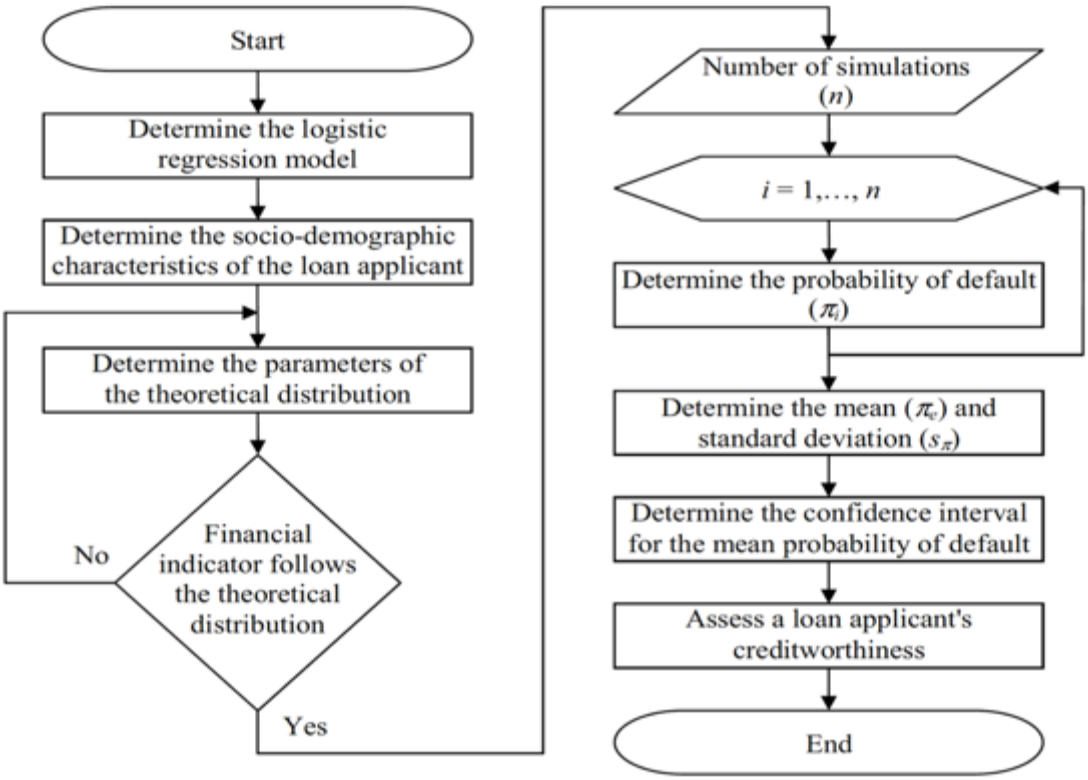

Figure 2: Main phases of the proposed decision support system

The proposed decision support system aims to improve the assessment of the loan applicant's credit worthiness. In this system, financial indicators are defined as arbitrary features with simulated values. It is the responsibility of the person making the decision to determine theoretical distributions for the financial indicators. In cases when historical data are available, the hypothesis that the financial indicators follow a certain distribution needs to be checked by an adequate statistical test. For this purpose the Kolmogorov-Smirnov test can be used.

The assessment of the loan applicant is made based on the determined confidence interval. If the threshold for the mean probability of default is within the boundaries of tolerance, the applicant will be granted a loan, and otherwise not.
In the credit scoring decision support system proposed in this paper, the authors assume that a larger number of simulations will be performed. The system then delivers the loan applicant assessment based on the threshold for the mean probability of default.

Conclusion

Adequate software applications need to be developed if the proposed decision support system is to be used for conducting quick and simple analysis of many loan applications. Decision making based on this system could be additionally improved by conducting sets of simulations sets.

According to the authors, socio economic factors like age, gender, marital status etc. are not taken into consideration while calculating the credit risk of a customer/borrower. Hence, if these factors into account, the credit worthiness of a customer could be measured more accurately.

2.3 An Empirical Study on Credit Scoring Model for Credit Card by using Data Mining Technology.

In this paper, (Li, et al., 2011) investigate the accuracy of the credit scoring model using 5 different machine leaning algorithms. They have used neural network, decision tree, logistic regression, regression tree and interaction detector for building the model. They first apply feature extraction to extract the principal component which denotes whether the customer has defaulted or not. Then a comparative study is done between the five different models, to check which model can classify the dataset more correctly.

Approach

Data Set: The data set was provided by one of the commercial banks in China. This dataset contained personal, family and credit/debit card information of the customers. It contained around 28 features and 80000 records.

Applying Principal Component Analysis to find the target variable: Among the 28 features in the data set, there was high correlation among the 8 features as shown in the table below:

Table 1: Correlation matrix between the 8 features

	Bad debt record	Returned check record	Dishonor of bill record	Compulsory card frozen record	Overdue record	Bad credit record	Large sum loan
Bad debt record	1	0.903	0.891	0.927	0.921	0.674	0.922
Returned check record	0.903	1	0.853	0.955	0.942	0.706	0.894
Dishonor of bill record	0.891	0.853	1	0.876	0.859	0.61	0.926
Compulsory card frozen record	0.927	0.955	0.876	1	0.952	0.789	0.914
Overdue record	0.921	0.942	0.859	0.952	1	0.716	0.865
Bad credit record	0.674	0.706	0.61	0.789	0.716	1	0.685
Large sum loan	0.922	0.894	0.926	0.914	0.865	0.685	1

Then, they have used PCA to extract the target variable to find whether the person defaulted or not. Hence, the dataset consisted of 20 features which were divided into ‘good credit’ set and ‘bad credit’ set.

Table 2: Cumulative variance of the features

Component	Initial Eignenvalues		
	Total	% of Variance	Cumulative%
1	6.042	86.31	86.31
2	0.513	7.33	93.64
3	0.200	2.86	96.50
4	0.102	1.46	97.96
5	0.061	0.87	98.83
6	0.051	0.73	99.56
7	0.031	0.44	100

Model Result and effect evaluation: Table 3 shows that decision tree performed the best as compared to the other prediction models, with a 100% accuracy for the

training set and the testing set. The Neural Network Model performed second best with an accuracy of 94 percent. The other models gave an average prediction accuracy between the range of 69 to 82 percent.

Table 3: prediction accuracy of five models

Model	*Training Set*			*Testing Set*			*Total Samples*
	Default	*Non-default*	*Total Set*	*Default*	*Non-default*	*Total Set*	
C5.0 Decision Tree	100%	100%	100%	100%	99.96%	99.97%	99.99%
Neural network	93.5%	96.7%	95.3%	97.9%	95.24%	96.44%	95.6%
Chi-squared automatic interaction detector	88.9%	74.6%	81.1%	87.4%	75%	80.6%	80.95%
Logistic model	86.3%	72.5%	78.7%	87.5%	71.4%	78.7%	78.7%
Classification and regression tree	96.6%	47.5%	69.7%	94.9%	45.2%	67.7%	69.1%

Conclusion

According to the authors, Credit scoring using different machine learning algorithms are used by many lending organizations, to control and mitigate the credit risks arising out of a default. In this data analysis, Decision Tree performed best for classification while the regression model was the least helpful among the five models to classify customers into default and non-default set.

Here, the authors have used Feature extraction technique like PCA to exact a dependent variable, and the outcome of the logistic regression is not very impressive and is not comparable to the C5.0 Decision Tree model. They have not considered a feature selection method to predict the outcome of the class. This is a technical gap that they have failed to address in this paper, which we would like to take up as our research topic, to conduct a comparative study on credit scoring by using feature

extraction methods like PCA against feature selection models like stepwise logistic regression.

2.4 Credit scoring model based on Bayesian Network and Mutual information.

In this paper, (Zhuang, et al., 2015) have looked at feature selection techniques like Bayesian Network Mutual Information (BNMI), to reduce the degree of uncertainty among empirical attributes. They then used the learned Bayesian Network to adaptively adjust according to the mutual information. They then conducted experiments to compare the BNMI model with three different baseline models.

The proposed Model

Overview of the BNMI Model

The BNMI model is divided into four phases which includes Data preprocessing, BN structure learning, Markov Blanket (MB) extraction, and parameter fitting and prediction. Data preprocessing consists of data cleansing and attribute ranking. In attribute ranking, the mutual information (MI) between each attribute and the target/class variable is calculated. BN structure learning consists of two steps. The first step learns a BN structure from data using Hill Climbing algorithm. In the second step, they propose a novel MI based algorithm to score and obtain the attributes MI list containing the most related attributes of the class variable. In the MB (Markov Blanket) extraction phase. First, the MB (Markov Blanket) of the class variable is obtained. Then, the MI list in phase two is used to re-examine MB of the class variable and further improve it by adding parents from the MI list not present in the current MB. Finally, the BN's parameters are fitted in the first phase, resulting in a full functional BN (Bayesian Network). Then the resulting BN can be used for classification and prediction tasks. The overview of the proposed BNMI model is as shown below:

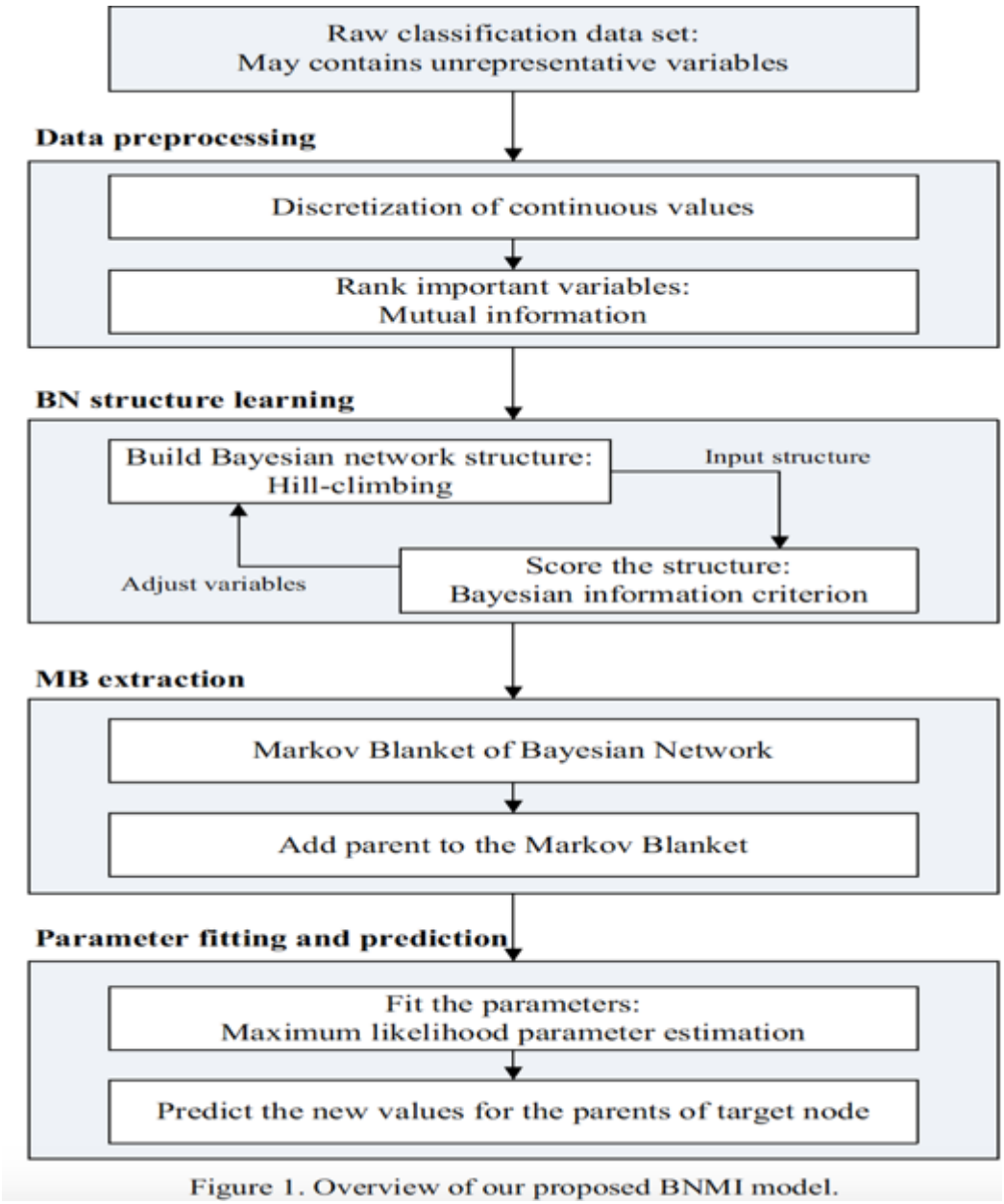

Figure 1. Overview of our proposed BNMI model.

Figure 3: BNMI model

Algorithm Design:

a. First the Mutual Information (MI) between the target variable are calculated.
b. Algorithm for building Bayesian network based on Mutual Information (The Build BN Algorithm).

c. Parents adding algorithm: It first obtains the attributes with largest MI with the class variable, and then it inserts one attribute into the MB of the class variable iteratively.

d. Parameters fitting and prediction: BN is used on testing data or new data to predict the customers' credit performance.

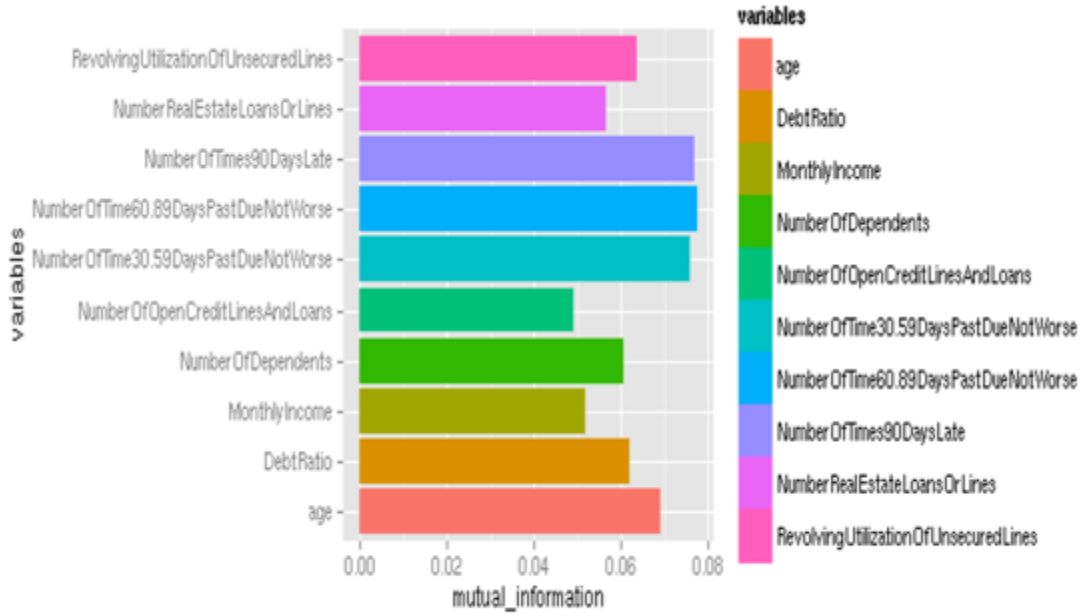

Figure 2. Mutual information of the data set.

Figure 4: Mutual Information

Experimental Results and discussion.

a. Dataset: The Dataset was obtained from "kaggle.com".In this study, the dataset is transformed into a form where the numerical variables "RevolvingUtilizationOfUnsecuredLines" and "DebtRatio" are discretized. The target variable "SeriousDlqin2yrs" is divided into two categories. Because the variables "MonthlyIncome" and "NumberOfDependents" contains missing values (NA), they transform the NA to categorical "unknown". The final data set used in this study consists of 11 columns and 150000 lines. Lastly, the data set is divided into 125,000 instances for "training data" and 25000 instances for "testing data".

b. Experimental Results: After computing the MI between target and other variables, they found that the features "NumberOfTimes90DaysLate", "NumberOfTime60.89DaysPastDueNotWorse" and "NumberOfTime30.59DaysPastDueNotWorse" have the top three MI values that are greater than 0.07. Also after applying the BNMI algorithm to improve BN leanining, it was observed that the features which had the greatest impact on the target class were "RevolvingUtilizationOfUnsecuredLines","NumberRealEstateLoansOrLines", “NumberOfTimes90DaysLate”, “NumberOfTime60.89DaysPastDueNotWorse”, and “NumberOfTime30.59DaysPastDueNotWorse”.

c. Comparison of Accuracy: The ROC plot in the figure below shows the accuracy of decision network, neural network, Bayesian network and BNMI. The AUC values of decision tree, neural network, Bayesian network and BNMI are 0.7792127, 0.8470511, 0.7814991 and 0.850851 respectively. The AUC of neural network and BNMI are higher, which are 0.8470511 and 0.850851, respectively. So, based on the data set, neural network and BNMI has high accuracy, and BNMI is slightly higher than the neural network model and achieves the best accuracy overall.

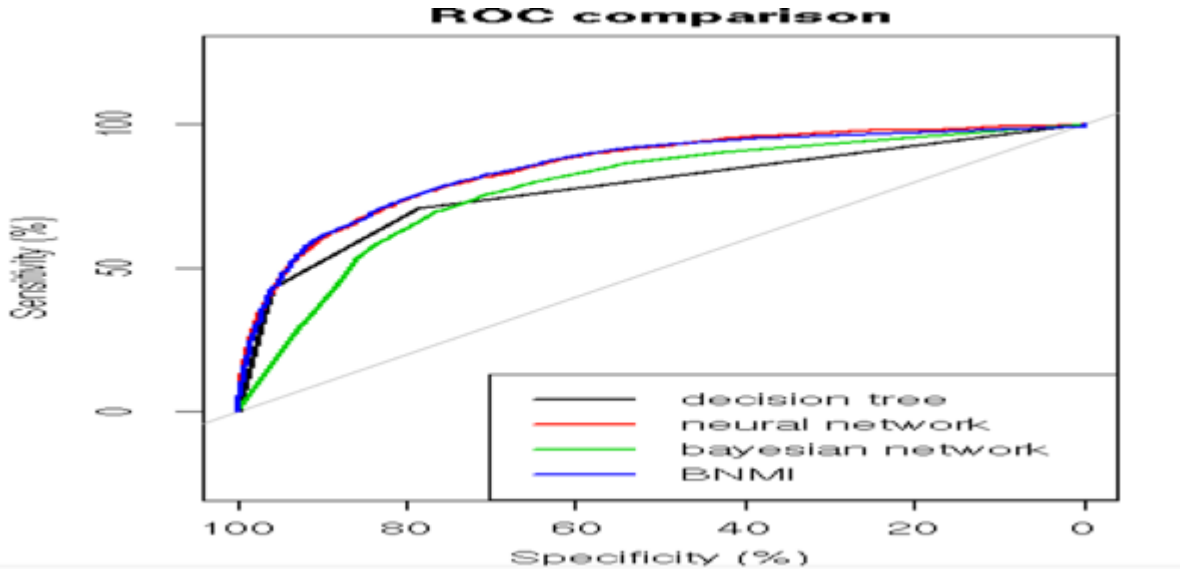

Figure 5: ROC comparison between BNMI and three baseline models.

Conclusion

In this paper, the authors have proposed a new scoring model called BNMI, which combines the advantages of both BN and MI, to build a better credit scoring model. The experiments conducted by them show that their BNMI model outperforms three existing baseline models (decision tree, neural network, and Bayesian network) in terms of receiver operating characteristic (ROC), indicating promising application of BNMI in credit scoring area. Here, they also conclude that performing using a feature selection technique like BNMI improved the accuracy of their model from 78 percent to 85 percent. As part of their future work, they plan to do a comparative study between other scoring algorithms to evaluate and build a Bayesian network.

2.5 Building classification models for customer credit scoring.

In this paper, (Benyacoub, et al., 2014) explore HMM(Hidden Markov Models) as a classification technique for credit scoring.

Background

Hidden Markov Models is a type of supervised machine learning algorithm. It could be used as a potential machine learning algorithm for predicting credit scores. Baum-Welch Algorithm provides HMM with the model parameters after a series of observations.

Classification Approach

As shown in the fig.6, the authors have followed three phases in their classification approach. They are Data preparation, Model building and Model validation.

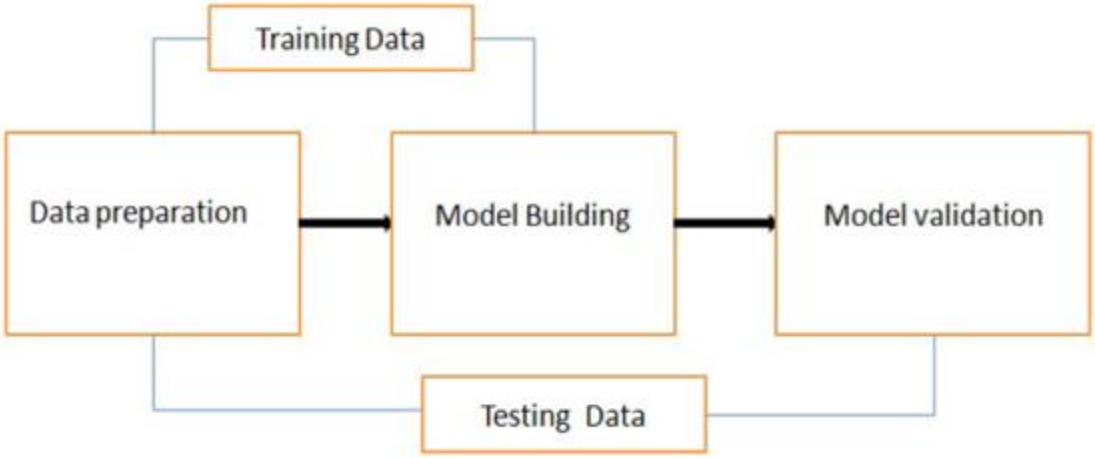

Figure 6: The classification approach for credit scoring

Experiments

a. Data: German credit dataset and Australian credit dataset were used to perform these experiments. Both the datasets were obtained from UCI machine learning repository.
b. Results and Analysis: They used the Matlab tool to compute the model results. With both the datasets they kept the number of iterations fixed i.e. 1000.

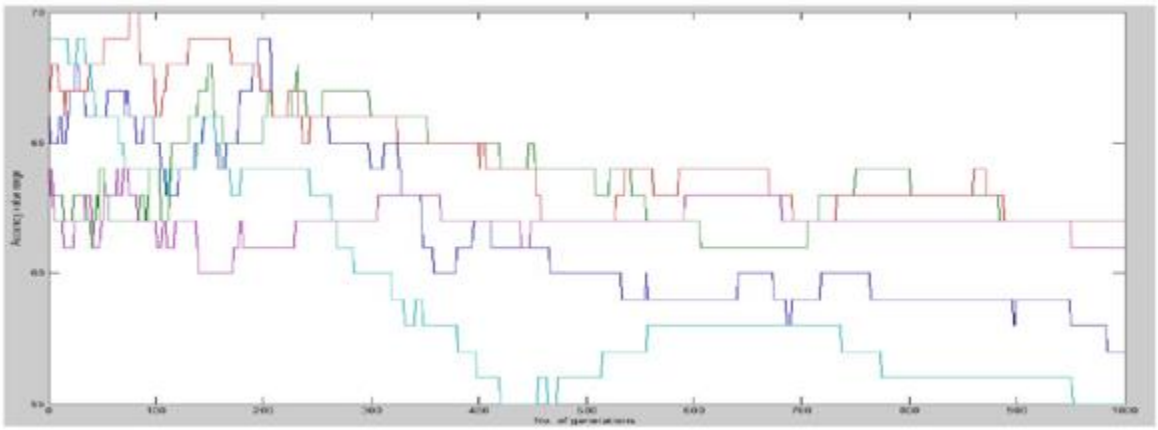

Figure 7: HMM prediction accuracy for German Credit Set.

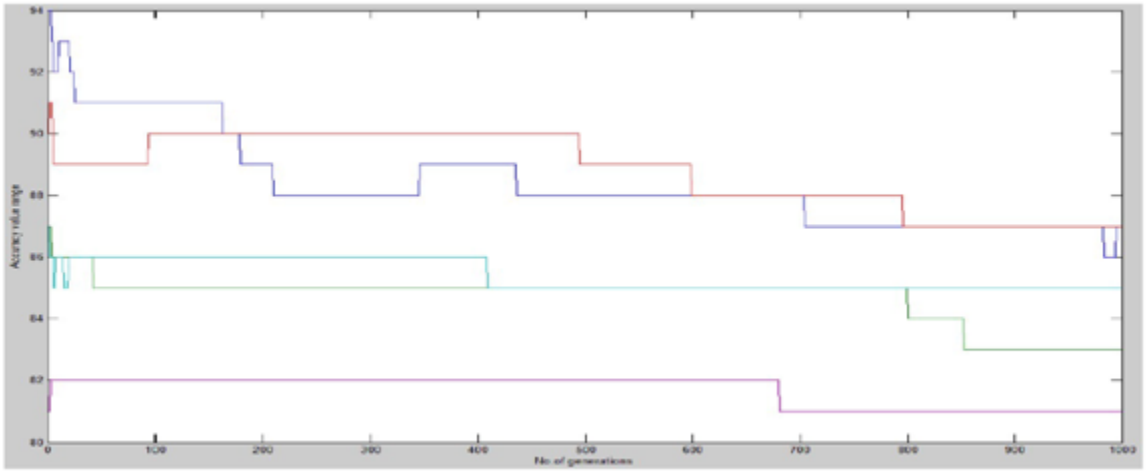

Figure 8: HMM prediction accuracy for Australian Credit Set.

Figure 7 and Figure 8 state the experimental results of the Hidden Markov Models and Baum-Welch model after 1000 iterations. As shown in both figures, after 200 iterations, the accuracy of the model starts increasing. When the model reaches the 1000 iteration, the accuracy decreases.

Conclusion:

In this paper, the authors have proposed a novel approach for detecting customers that may default in the future by making use of Hidden Markov Models (HMM). One of the major advantages of using such a supervised learninfg algorithm such as HMM is that it uses an iterative approach to do the prediction. As shown in the figures above, significant improvement in accuracy is observed using Hidden Markov Models and Baum Welch.

2.6 A comparative study of discrimination methods for credit scoring

In this paper, (Chen, et al., 2010) examine several sophisticated and highly effective machine learning algorithms, such as Skew-normal discriminant analysis (SNDA), Skew-t discriminant analysis (STDA), Stepwise discriminant analysis (SDA),

Sparse discriminant analysis (Sparse DA), Flexible discriminant analysis (FDA), and Mixture discriminant analysis (MDA) for screening credit card applicants.

Evaluation

The machine learning algorithms are evaluated by their ability to distinguish between defaulting customers and non-defaulting customers. Customers with good scores sually have good credit history while applicants with bad score usually have bad credit history. They are generally divided into three classes:

a. **The Total Percentage of Correctly Classified Cases (Total PCC)**

   The total percentage of correctly classified cases (total PCC) is the probability of correctly classifying a future observation by using 5-fold cross validation.

b. **The Bad Rate Among Accepts(BRA)**

   The bad rate among accepts is the number of customers who have a good credit score but eventually turn out to be non-creditworthy by defaulting on their credit.

c. **The ROC (Receiver Operating Characteristics) curve**

   An ROC plot is fraction of true positive rates (TPR) to the fraction of false positive rates (FPR). It is defined as the ratio of sensitivity vs. (1 – specificity).

Empirical Analysis

a. **Dataset:**

   They have used the German dataset to conduct their anlysis. This dataset consists of 20 features having 1000 records.

b. Results:

The results for the Total PCC are shown in table 4. Skew normal discriminant analysis and Skew-t discriminant analysis peforms better than all the other discrimination methods.

Table 4: Total PCC

TOTAL PCC OF THE AUSTRALIAN CREDIT DATASET FOR THE SNDA, STDA, SDA, SPARSE DA, FDA AND MDA.

Method	Total	Good	Bad
SNDA	.8425	.8869	.7060
STDA	.8375	.8860	.7344
SDA	.7330	.8173	.6736
SPARSE DA	.7450	.7461	.7142
FDA	.7750	.8054	.6863
MDA	.7984	.8107	.5859

The results for the BRA are shown in table 4. Skew normal discriminant analysis and Skew-t discriminant analysis peforms better than all the other discrimination methods because of the lower BRA values.

Table 5: BRA

BRA OF FOR THE GERMAN CREDIT DATASET FOR THE LDA, SNDA, STDA, SDA, SPARSE DA, FDA AND MDA.

	Accept Rate			
Method	75	80	85	90
SNDA	.1267	.1456	.1796	.2123
STDA	.1151	.1402	.1667	.2022
SDA	.1773	.1988	.2259	.2478
SPARSE DA	.2633	.2688	.2706	.2833
FDA	.2026	.2113	.2341	.2522
MDA	.2013	.2175	.2353	.2556

The ROC curves for Skew normal discriminant analysis and Skew-t discriminant analysis gives the best AUC values.

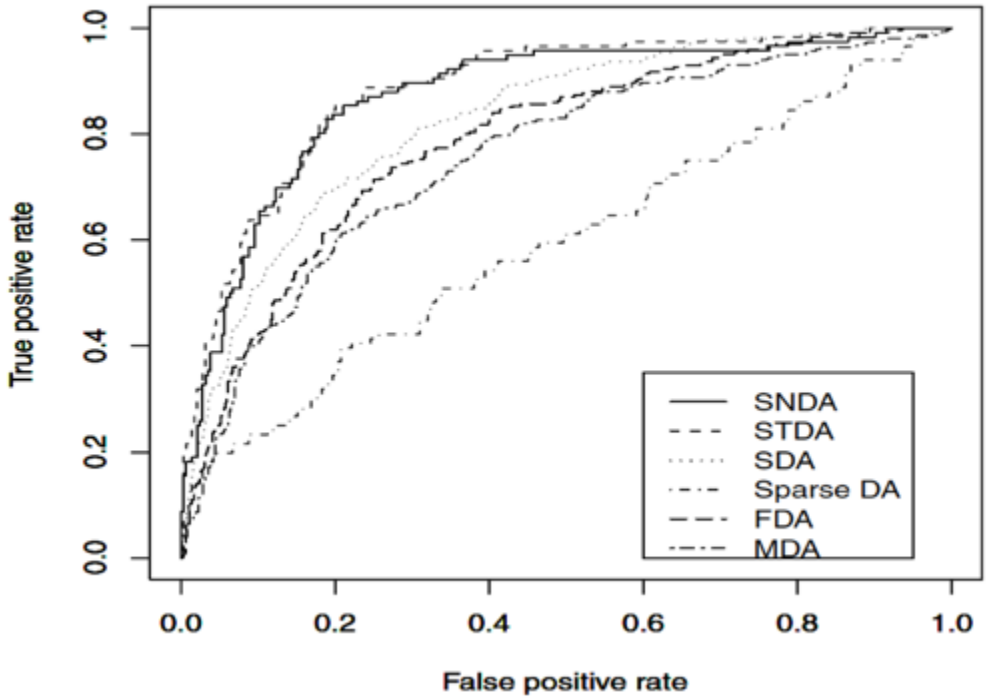

Figure 9: ROC curve

From the results, it can be observed that the Skew normal discriminant analysis and Skew-t discriminant analysis performed better than all others techniques. According to the authors, each of these methods discussed in this study would perform better for different datasets. Hence, as part of the future work, the authors would like to test these these methods on multiple datasets to ascertain whether the same results would be achieved.

2.7 Application of the Hybrid SVM-KNN Model for Credit Scoring

In this paper, (Zhou, et al., 2013) have used an ensemble model using Support Vector Machine and K-Nearest Neighbors algorithm to improve the performance of Support Vector Machine in terms of its prediction accuracy. This approach uses combines the salient features of both these machine learning algorithms.

Experiment

They have used the German Credit dataset and the Austrailan Credit dataset from the UCI machine learning repository to conduct their experiments. The German Credit dataset consists of 20 features with 1000 records. While, the Australian Credit dataset consists of 14 features with 690 records.

Results

They have used the MATLAB tool to conducts their experimental analysis. For the Support Vector Machines, they have used the Radial Basis Fuction as the kernel. The distance function for the K-Nearest Neighbors algorithm is as given below:

$$d = \left| \sum_{i=1}^{m} \omega_i e^{\frac{-\|x_i - x\|^2}{2\alpha^2}} + b \right|$$

Also, the parameters for the Support Vector Machine are taken as default. After conducting experiments, it can be observed that the hybrid ensemble Support Vector Machine and K-Nearest Neighbors model has a higher accuracy than both when individually using SVM and KNN when conducting experiments. The below table gives information regarding the accuracy, after the model has predicted the credit score.

Table 6: Accuracy rate for SVM-KNN, SVM and KNN respectively.

Names	Accuracy		
	$SVM + KNN$	SVM	KNN
Australian	**86.19%**	84.76%	84.29%
German	**75.98%**	73.27%	72.67%

The ensemble model using Support Vector Machine and K-Nearest Neighbors performs better than both the individual models. However, the distance function using KNN takes a lot of time in terms of computation.

As a future work, they would like to reduce the time taken to compute the distance and hence improve the efficiency of the algorithm.

2.8 Recombining Forecasts Used in Personal Credit Scoring.

In this paper, (Ming-hui, et al., 2006) present a new approach to personal credit scoring by using a combination of ensemble methods from three different Neural Networks and comaparing their performance with individual machine learning models like linear and logistic regression.

Dataset

They use the consumption loan data of a commercial bank, which had data for about 1057 customers. They used 529 records to train the model and 528 records to test the data.

Approach

In this paper, they chose RBF which is a forward neural network, Elman which is a feedback neural network and LVQ which is a competitive neural network to carry

out their prediction. The reason they chose these models was to determine the validity of the models in personal credit scoring by comparing their results to different combining models.

Results

After conducting experiments, it can be noted that the three combined prediction methods such as RBF, Elma and LVQ using Neural networks have a better precision of 94 percent when compared to individual methods such as linear regression, logistic regression etc.

Conclusion

Therefore, from the results it can be observed that using an ensemble method by combining the 3 neural networks gave a better prediction accuracy than individual machine learning models like linear regression.

3 RESEARCH HYPOTHESIS AND OBJECTIVES.

3.1 Research Objective

Based on all the technical gaps that are addressed in my literature review, my research interest would be to "**Perform a comparative study between Stepwise Logistic Regression which is a feature selection technique and Singular Value Decomposition (SVD), which is a feature extraction technique, to improve the accuracy and performance of credit scoring using the Logistic Regression Algorithm**".

3.2 Hypotheses

Alternate Hypothesis

Stepwise Logistic Regression as a feature selection algorithm should improve the accuracy and performance of credit score prediction model, as compared to a feature extraction algorithm like Singular Value Decomposition (SVD) by approximately 14% and Weighted Singular Value Decomposition (Weighted SVD) by approximately 10%.

Null Hypothesis

Stepwise Logistic Regression as a feature selection algorithm will not improve the accuracy and performance of credit score prediction model, as compared to a feature extraction algorithm like Singular Value Decomposition (SVD) by approximately 14% and Weighted Singular Value Decomposition (Weighted SVD) by approximately 10%.

Note:

As a part of my literature review, I found some information, based on which I am stating this hypothesis. In two of the papers (Fan, et al., 2013 and Zhuang, et al., 2015), who used a similar kind of dataset: In one, they have applied a model on the dataset after applying PCA (which is a feature extraction technique) and they achieved an accuracy of 78%. In the other, they have applied a model on the dataset after using a feature selection technique and they achieved an accuracy of 85%. This shows an increase for the feature selection technique by around 7%. The experiments I plan to perform are of a similar nature and hence, the above hypothesis of an increase in percentage of 10 percent for a feature selection technique is justified, and should result in a better model.

4 EXPERIMENTAL DESIGN

The experiments defined below are intended to test the hypothesis posited above. All experiments will measure the effect of carrying out the experiments by employing the metrics described below:

4.1 Calculate the accuracy of the credit score prediction model, using Stepwise Logistic Regression, a feature selection technique.

4.2 Calculate the accuracy of the credit score prediction model, using Logistic Regression after using Singular Value Decomposition (SVD), a feature extraction technique.

4.3 Compare the accuracy obtained using both the above models.

4.4 Apply weights to important features, before performing (Singular value Decomposition) SVD on the dataset.

4.5 Calculate the accuracy of the credit score prediction model, using Logistic Regression, after using Weighted Singular Value Decomposition (Weighted SVD).

4.6 Compare the accuracy obtained using Stepwise Logistic Regression, with the accuracy obtained using Weighted SVD (Singular Value Decomposition).

4.7 Select the Feature Reduction Technique which gives the best accuracy after performing the above experiments.

5 APPROACH AND METHOD

5.1 Data Exploration

5.1.1 Data Set Description.

For the conducting the experiments, as stated in the Experimental Design section, We would be using the dataset from "kaggle.com" called "Give me some credit". This dataset consists of 11 features and 150,000 records. The table below highlights the Features, their description and their corresponding datatype.

Table 7: Feature Name, Description, Datatype

Feature Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
Age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	Integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	Integer

1. Serious Delinquency in 2 years: This is the predictor/dependent variable. It has a binary value of either 1 or 0. A value of 1 means that the borrower is delinquent and has defaulted on his loans for the last 2 years, while a value of 1 means that the borrower is a good customer and repays his debts on time for the last two years.
2. Revolving Utilization of unsecured Lines: Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits, i.e. ((total non-secured debt)/ (total non-secured credit limit)).
3. Age: This represents the Age of borrower in years
4. NumberOfTime30-59DaysPastDueNotWorse: This feature represents the Number of times borrower has been 30-59 days past due but no worse in the last 2 years.
5. Debt Ratio: This feature represents monthly debt payments, alimony, living costs divided by the monthly gross income
6. Monthly Income: This feature represents the Monthly income of the individual
7. Number Of Open Credit Lines And Loans: This feature represents the number of open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)
8. Number of Times 90 Days Late: This feature denotes the number of times borrower has been 90 days or more past due.
9. Number of Real Estate Loans or Lines: This feature denotes the Number of mortgage and real estate loans including home equity lines of credit
10. NumberOfTime60-89DaysPastDueNotWorse: Number of times borrower has been 60-89 days past due but no worse in the last 2 years.
11. Number of Dependents: Number of dependents in family excluding themselves (spouse, children etc.).

5.1.2 Data Visualization using Scatter plot and Heat map of the Raw Data

5.1.2.1 Scatter Plots of the Independent variables with respect to the dependent variable.

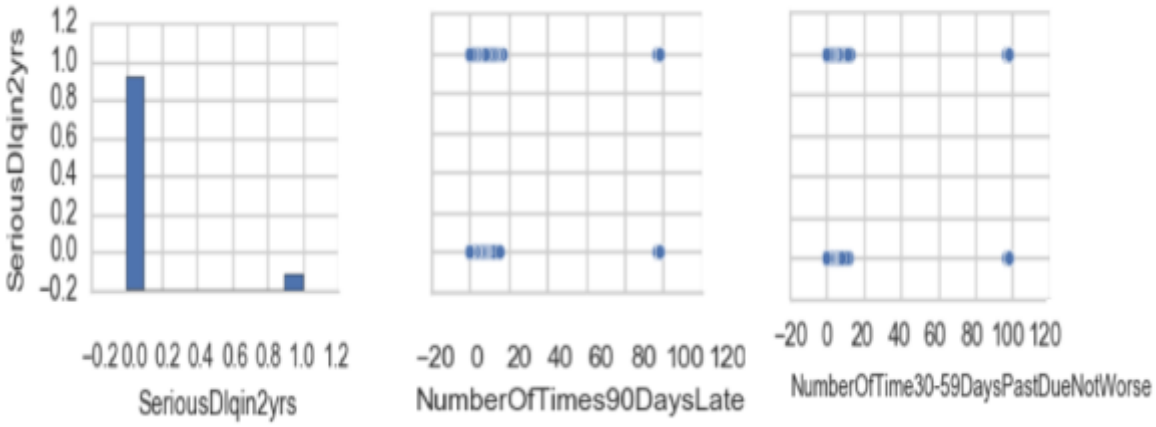

Figure 10: Scatter plot of Independent variables "NumberOfTimes90DaysLate", "NumberOfTimes30-59DaysPastDue" with the Dependent Variable

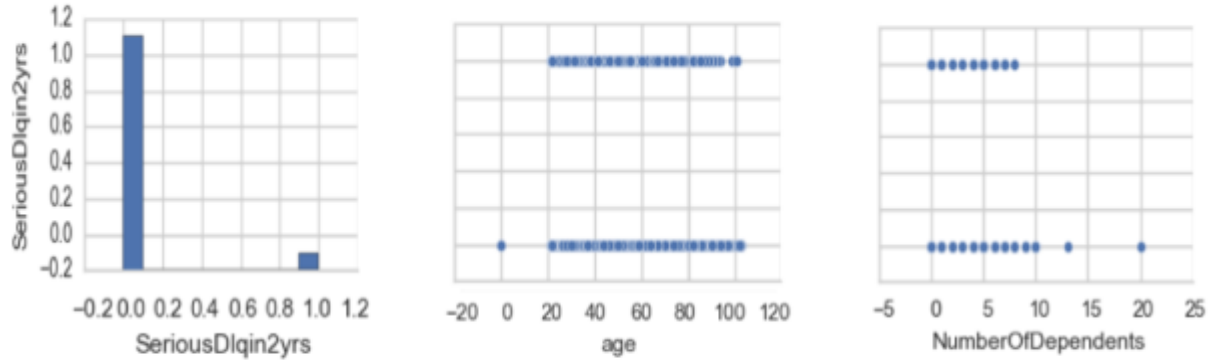

Figure 11: Scatter plot of Dependent variables "age", "NumberOfDependents" with the dependent variable.

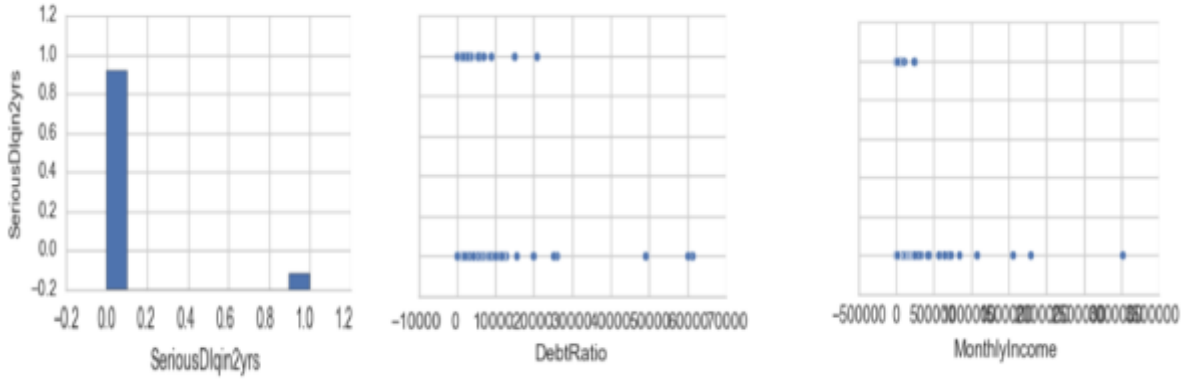

Figure 12: Scatter plot of Dependent variables “Debt ratio”, “Monthly Income” with the dependent variable.

As shown here, we can see the features have a lot of outliers and wrong data which would be handled in the Feature engineering section.

5.1.2.2 Heat Map which denotes the correlation between the independent features and the dependent feature.

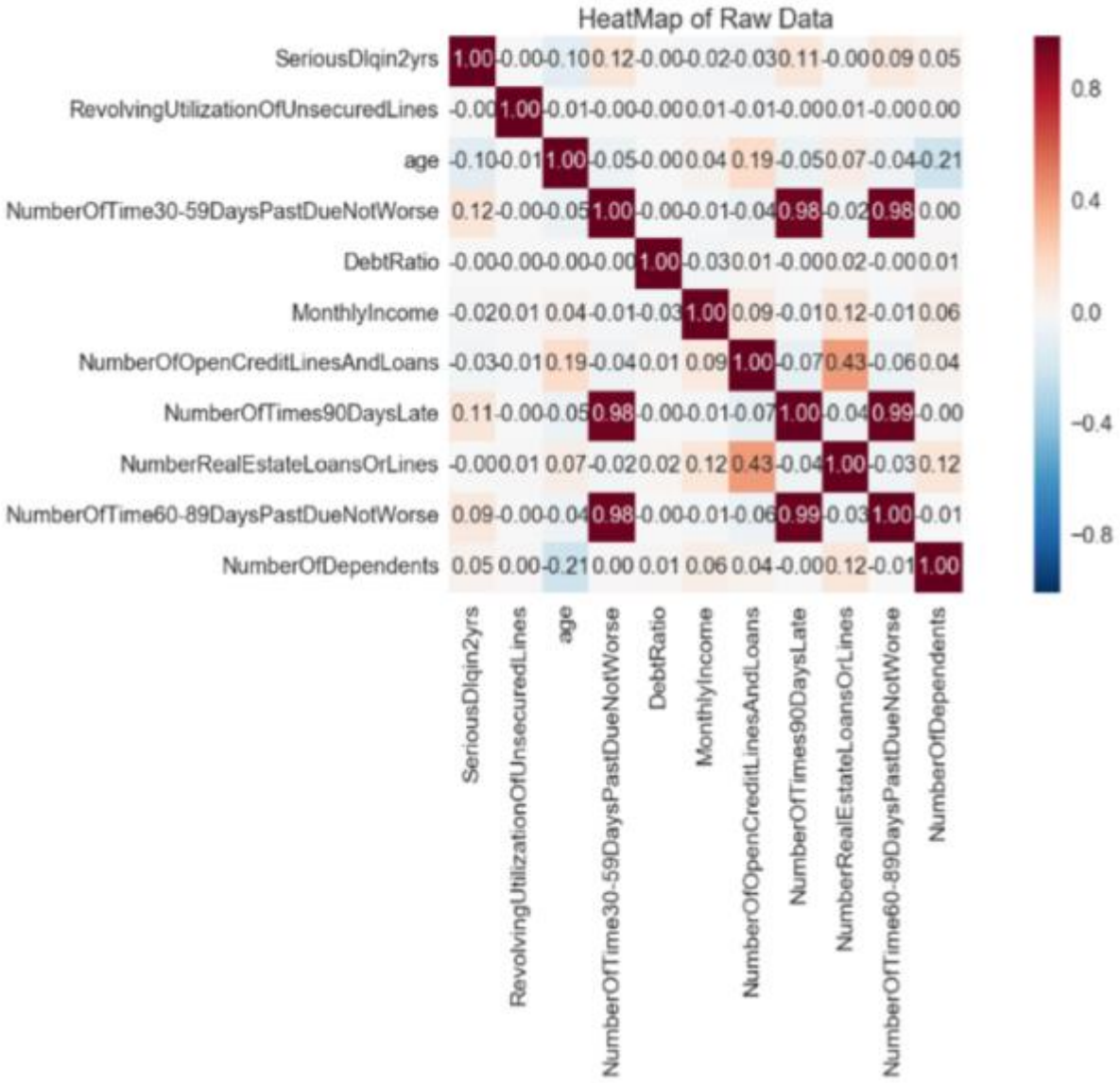

Figure 13: Heat Map of the Raw Data

The features have a very low correlation w.r.t to the independent variable, hence the data would have to be cleaned and processed so that the data becomes linear and correlated.

5.2 Feature Engineering

5.2.1 Removing missing values.

We first dropped the rows containing missing values or nan values. There were around 29,731 records which had missing values. After dropping those records, there were 120,269 rows remaining in the dataset.

5.2.2 Removing outliers/illogical values in the dataset.

- As shown in the Fig.1, the scatterplot shows the data points for the features "NumberOfTime30-59DaysPastDueNotWorse", "NumberOfTime60-89DaysPastDueNotWorse" and "NumberOfTimes90DaysLate". All these features have values ranging from "0 to 20" and have outliers in the form of values "96" and "98". Therefore. we used the "pandas" library of python to drop rows having these values.
- The "age" variable is a continuous variable from 0 to 100. But to be qualified as a borrower, the person must be an adult of 18 years. There were certain records, which had a value of "0", that did not make sense. Hence, dropped all those records which had the "age" variable having a value of 0.
- The "debt ratio" feature has values ranging from 0 to 168835. The data is spread across continuously from 0 to 15000. The values above this range look to be outliers as shown in the scatterplot. Therefore, values above this range would be dropped.
- The "Monthly Income" feature has values ranging from 0 to 107,2500. But most the records have values ranging from 0 to 100,000 in the data set, as shown in the scatterplot above. Hence, all the other

records having values greater than 100,000 were dropped from the data set.

- The “RevolvingUtilizationOfUnsecuredLines” feature is a ratio of the total amount of non-secured debt to the total non-secured credit limit. Hence, this feature should have values between 0 and 1, but some of the records have negative values and some of the records have values greater than 1, with the maximum value being 50,000. Therefore, we have kept the records which range from 0 to 1, and dropped the other records.
- The “NumberOfDependents” feature has values ranging from 0 to 20. As shown in the scatter plot, most of the records are clustered around the values from 0 to 10. Hence, we would be dropping all those records with values 15 and 20 which are outliers as shown in the scatter plot above.
- The “NumberOfRealEstateLoansOrLines” feature has values ranging from 0 to 54. As shown in the scatter plot, most of the records are clustered around the values ranging from 0 to 10. Hence, dropping all values above this range.
- The “NumberOfOpenCreditLinesAndLoans” feature has values ranging from 0 to 58. As shown in the scatter plot, most of the records are clustered around the values from 0 to 10. Hence, we would be dropping all those records with above 10 which are outliers as shown in the scatter plot above.

5.2.3 Scatter plot of the processed data.

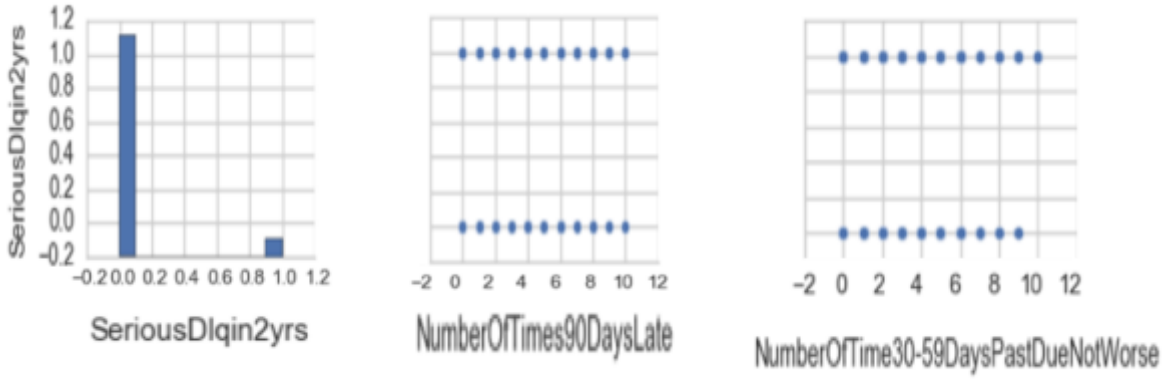

Figure 14: Scatter plot of Independent variables "NumberOfTimes90DaysLate", "NumberOfTimes30-59DaysPastDue" with the Dependent Variable

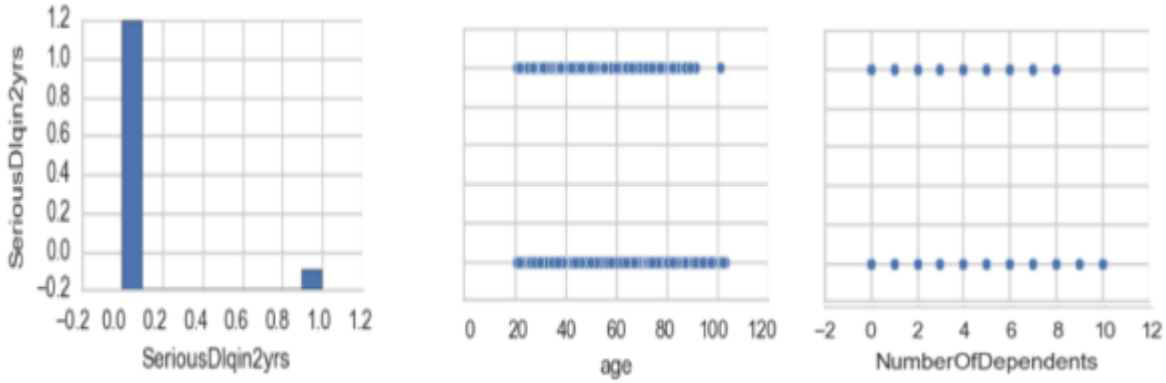

Figure 15: Scatter plot of Dependent variables "age", "NumberOfDependents" with the dependent variable.

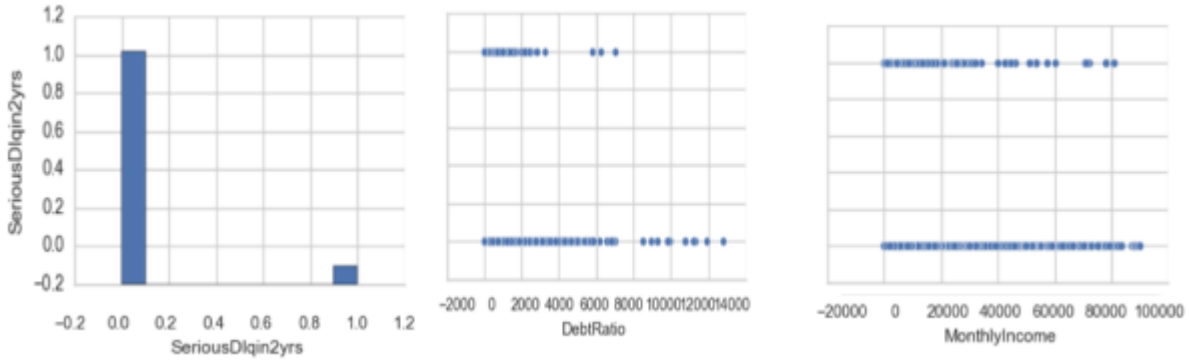

Figure 16: Scatter plot of Dependent variables “Debt ratio”, “Monthly Income” and “RevolvingUtilizationOfUnsecuredLines” with the dependent variable.

5.2.4 Heat Map after processing the data.

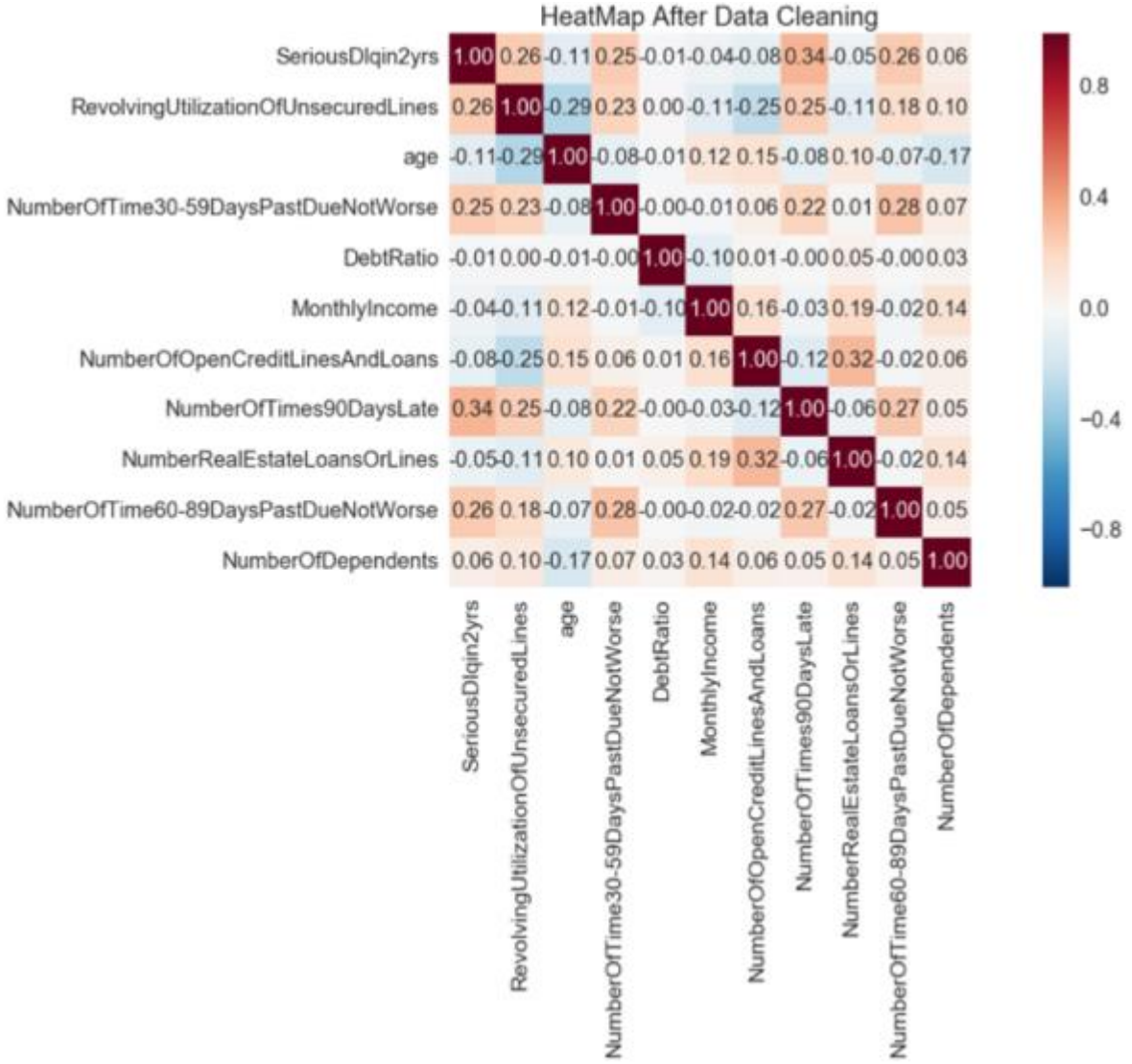

Figure 17: Heat Map after Feature Engineering

As shown in the figure above, we can see that the 4 variables "NumberOf90DaysLate", "NumberOfTime30-59DaysPastDueNotWorse", "NumberOfTimes60-89DaysPastDueNotWorse" and "RevolvingUtilizationOfUnsecuredLines" are having high correlation wr.t the independent variable.

5.2.5 Balancing the data.

The data is highly unbalanced with 111912 records having the predictor or target class as 0, and 8357 records having the predictor or target class as 1. Only 7 percent of the entire dataset has records with the target variable equal to 1.

Therefore, if the data is not balanced then it would result in a highly-skewed model, which would have the capability of predicting class 0 more than class 1. Hence, balancing the data is very important. Here, we take a random sample of records belonging to the target class 0 which is equal to the number of records belonging to target class 1.

This would help the classifier learn about each class equally and thus make a better prediction.

5.3 Feature Selection.

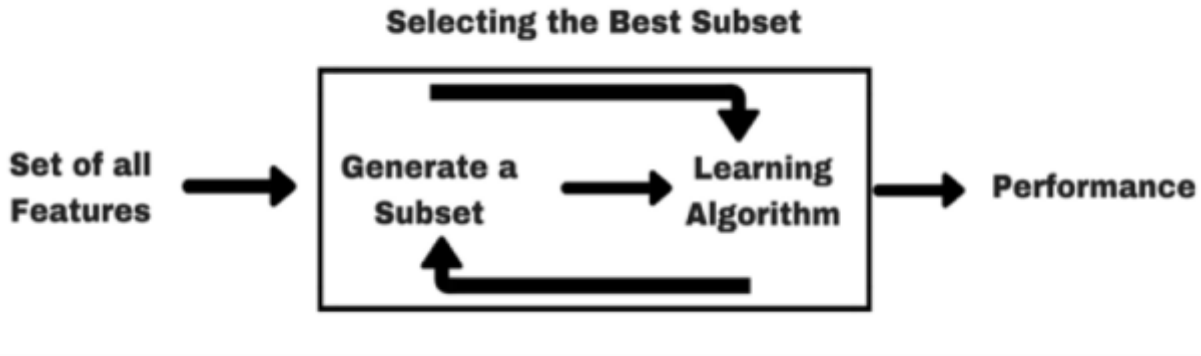

Figure 18: Feature selection approach

Feature selection is one of the two ways in which dimensionality reduction can be achieved. Given the entire number of features in the dataset, feature selection is the process of identifying the optimal subset of features based on an objective function. Feature selection helps in improving the prediction accuracy of the classifier, mining performance of the classifier.

5.3.1 Stepwise Logistic Regression using Recursive Feature Elimination (RFE).

Stepwise Logistic regression is a feature selection method which is used to add or remove features to the model, based solely on the importance of the features in terms of their statistical values. We will be using the Recursive Feature Elimination (RFE) procedure of "scikit-learn" package to perform feature selection. In Recursive Feature Elimination (RFE), an external estimator first assigns weights to all the features which are provided for training, and subsequently creates subsets or features based on the weight of each feature. We are using the forward approach, where it starts with no features and subsequently adds features based on their importance of their weights.

5.4 Feature Extraction.

Feature Extraction is another way in which dimensionality reduction can be achieved. In Feature Extraction, all the original values are transformed into principal components which are the linear combinations of the original features. Since, the dataset is not square, we would be using the Singular Value Decomposition (SVD) approach.

5.4.1 Singular Value Decomposition

We would be using "Truncated SVD" for feature extraction from the "scikit-learn" package. "Truncated SVD" performs feature extraction by setting the smallest singular values to 0.

5.4.2 Weighted Singular Value Decomposition.

Weighted Singular Value Decomposition (SVD), assigns weights to some of the important features, before applying Singular Value Decomposition (SVD). Standardizing the data is a pre-requisite for Weighted SVD. Standardizing the data, means rescaling the features to have a mean of 0 and variance of 1. After standardizing, weights are assigned to important features, by multiplying them with a scalar quantity greater than 1.

5.5 Classification

After dimensionality reduction, we use Logistic Regression Machine learning algorithm for training and testing the credit scoring model. We have partitioned the dataset such that 70 percent was used for training the model and 30 percent was used for testing the model.

6 RESULTS

6.1 Result of Stepwise Logistic Regression using Recursive Feature Elimination.

- Using 3 features ("NumberOf90DaysLate", "NumberOfTimes60-89DaysPastDueNotWorse" and "RevolvingUtilizationOfUnsecuredLines"), we get the following output:
  - Output:

Accuracy = 0.769764957265

AUC = 0.769615454878

Feature_rank = [2 1 1 4 3 5 8 1 7 6]

Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

Table 8: Classification Report for 3 features

Class	Precison	Recall	F1-score
0	0.77	0.78	0.77
1	0.77	0.76	0.76
Avg/Total	0.77	0.77	0.77

As shown above, the ‘feature_rank’ array corresponds to the rank assigned to each feature in the features array by the Recursive feature elimination (RFE) estimator. A rank of 1 means that the corresponding feature has been selected for performing classification task.

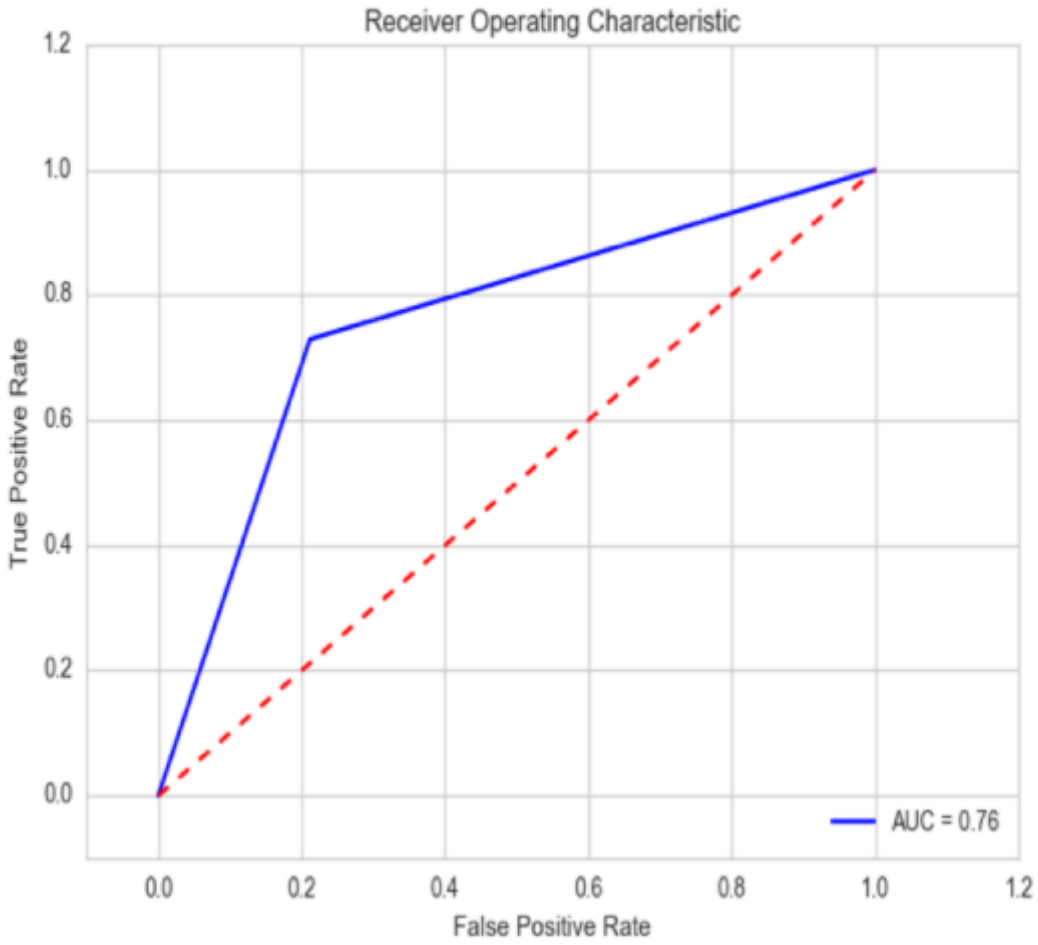

Figure 19: ROC curve for the 3 features

- Using 4 features ("NumberOf90DaysLate", "NumberOfTimes60-89DaysPastDueNotWorse", "RevolvingUtilizationOfUnsecuredLines" and "NumberOfTime30-59DaysPastDueNotWorse"), we get the following output:
    - Output:

      Accuracy = 0.782051282051

      AUC = 0.781969309463

      Feature_rank = [1 1 1 3 2 4 7 1 6 5]

      Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

      Table 9: Classification Report for 4 features

Class	Precison	Recall	F1-score
0	0.79	0.79	0.79
1	0.78	0.78	0.78
Avg/Total	0.78	0.78	0.78

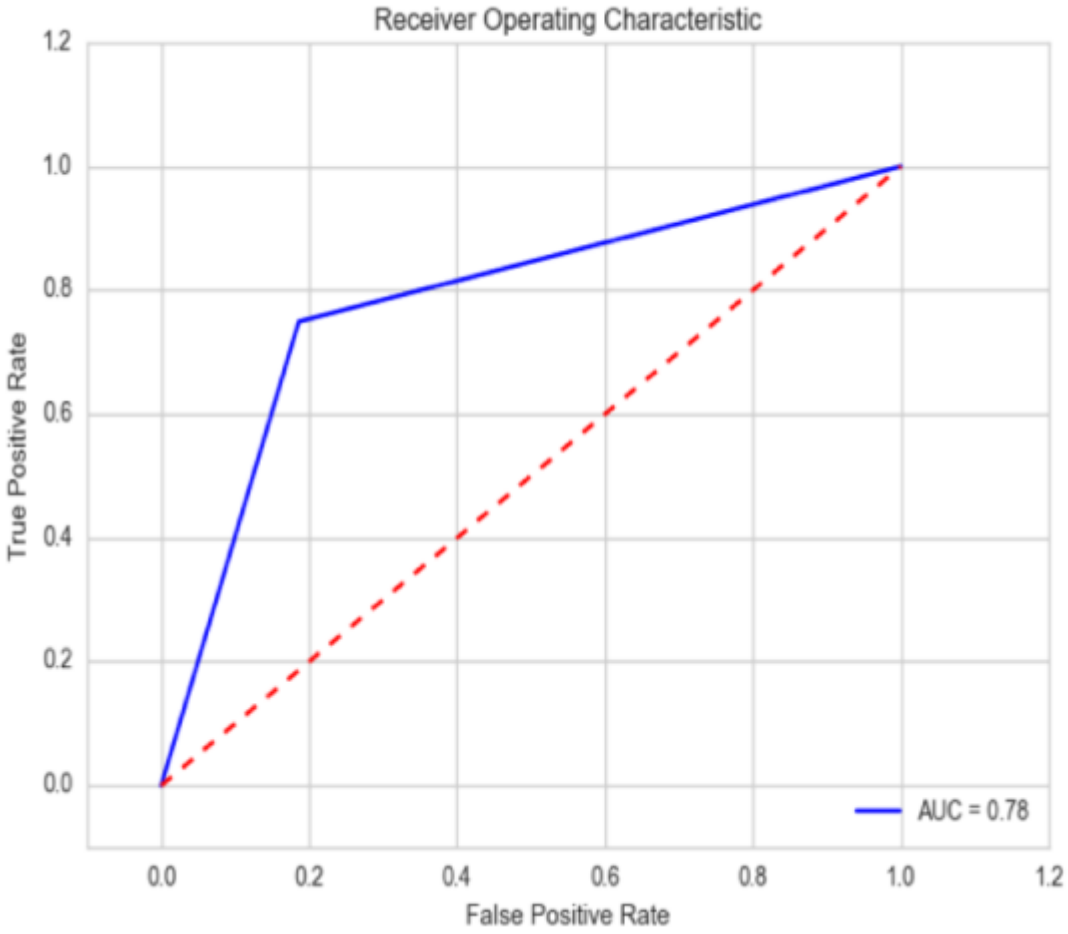

Figure 20: ROC curve for 4 features

- Using 5 features ("NumberOf90DaysLate", "NumberOfTimes60-89DaysPastDueNotWorse", "RevolvingUtilizationOfUnsecuredLines" and ""NumberRealEstateLoansOrLines"), we get the following result:
    - Output:

      Accuracy = 0.778846153846

      AUC = 0.778708439898

      Feature_rank = [1 1 1 2 1 3 6 1 5 4]

Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

Table 10: Classification Report for 5 features

Class	Precison	Recall	F1-score
0	0.78	0.79	0.78
1	0.78	0.77	0.77
Avg/Total	0.78	0.78	0.78

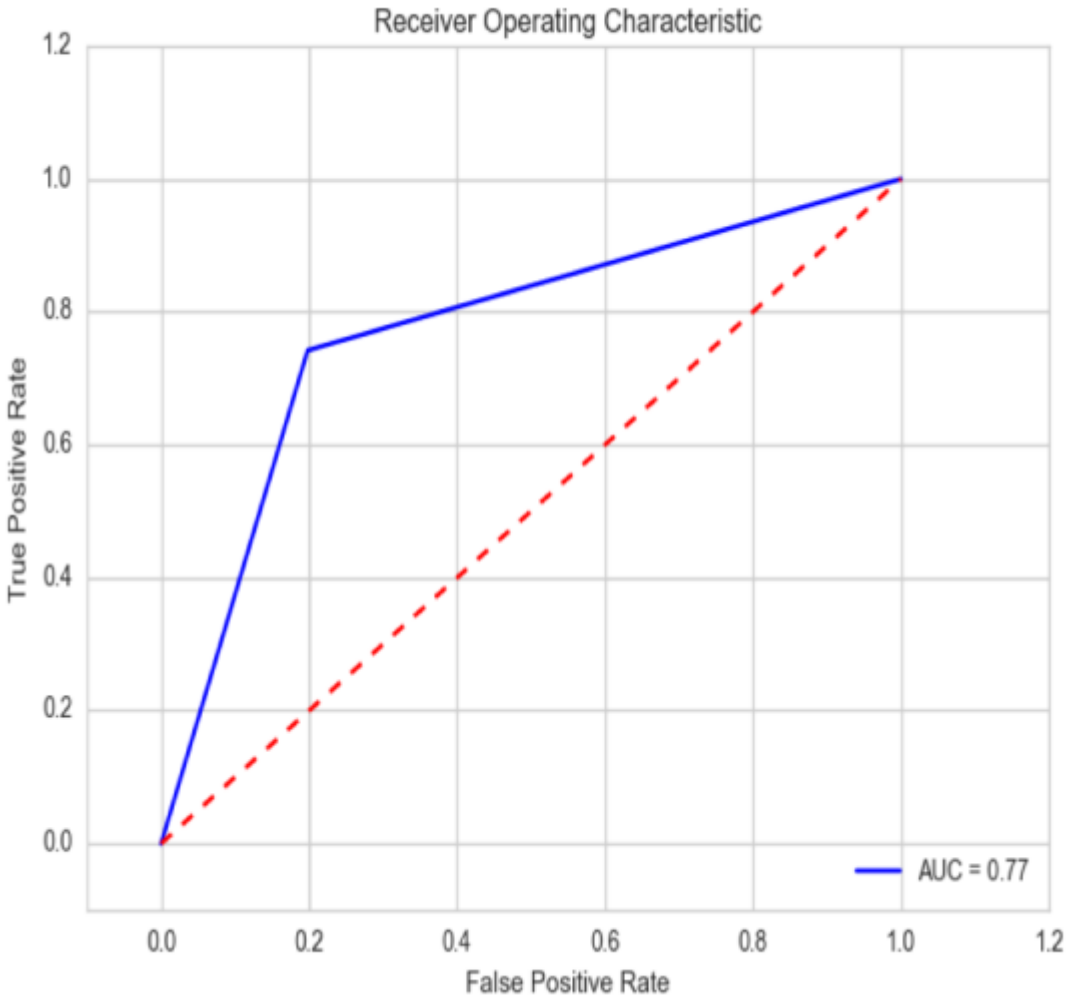

Figure 21: ROC curve for 5 features

As we can see from the results above, the classifier gives an accuracy of 76 percent with 3 features, increases to 78 percent with 4 features and then it again decreases to 77 percent with 5 features. Hence, we can see that Feature selection performs best with 4 features. Other results with different combination of features are shown in the appendix section of this report.

6.2 The Result of Feature Extraction using Singular Value Decomposition (SVD).

Accuracy of the Classifier is: 0.641320293399

AUC = 0.641871527895

Table 11: Classification Report for SVD

Class	Precison	Recall	F1-score
0	0.63	0.68	0.65
1	0.66	0.61	0.63
Avg/Total	0.64	0.64	0.64

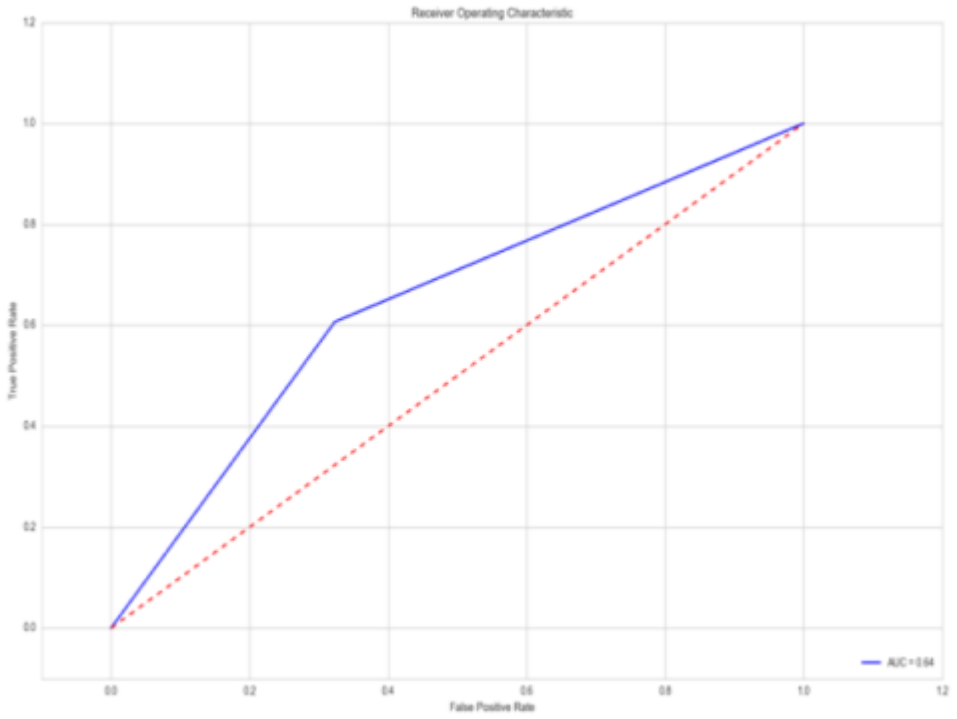

Figure 22: ROC curve for SVD

6.3 The Result of Feature Extraction using Weighted SVD (Singular Value Decomposition)

Accuracy of the Classifier is: 0.68141809291

AUC = 0.683525189303

Table 12: Classification Report for Weighted SVD

Class	Precison	Recall	F1-score
0	0.63	0.68	0.65
1	0.66	0.61	0.63
Avg/Total	0.64	0.64	0.64

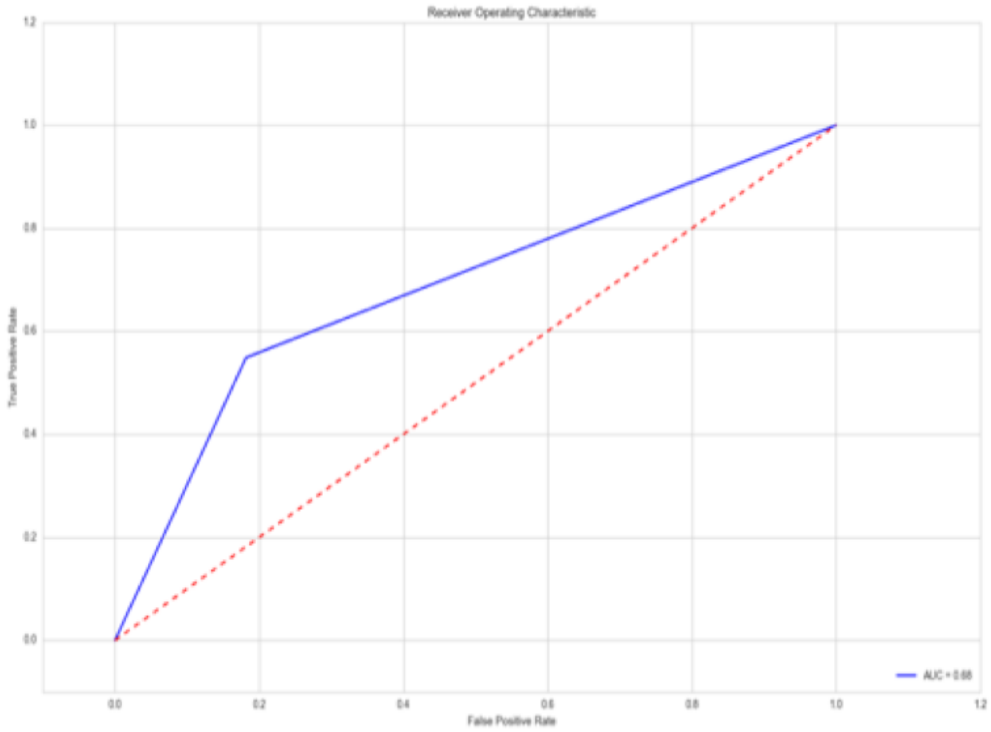

Figure 23: ROC curve for Weighted SVD

The table below provides the summary of the results,

Table 13 : Comparison of Results

Feature Reduction Technique	Accuracy	AUC	Precision	Recall	F1-score
Stepwise Logistic Regression	0.78	0.78	0.78	0.78	0.78
Singular Value Decomposition (SVD)	0.64	0.64	0.64	0.64	0.64
Weighted Singular Value Decomposition (SVD)	0.68	0.68	0.70	0.68	0.68

7 DISCUSSION

From the results we can see that, Feature Selection using Stepwise Logistic Regression performs significantly better than Feature Extraction using Singular Value Decomposition (SVD) and Weighted Singular Value Decomposition (SVD). Using Stepwise Logistic Regression, we see that selecting the 4 features, "NumberOf90DaysLate", "NumberOfTimes30-59DaysPastDueNotWorse", "NumberOfTimes60-89DaysPastDueNotWorse" and "RevolvingUtilizationOfUnsecuredLines" gave us the optimal accuracy in our credit scoring analysis.

Basically, in feature selection, we use distinct features for our model and we know that which features contributed towards the prediction. But in feature extraction, the features are transformed into a new reduced set of features, which might not be meaningful to us. In our problem of credit scoring, feature selection makes more sense

to the lenders or companies as they would want to know what kind of data to check before deciding a user's credit score.

In our model, these 4 features ("NumberOf90DaysLate", "NumberOfTime30-59DaysPastDueNotWorse", "NumberOfTimes60-89DaysPastDueNotWorse" and "RevolvingUtilizationOfUnsecuredLines") together determines whether the user is a defaulter or not. If we think logically also, these 4 features give us information about the defaulting of the users in different ways. "NumberOf90DaysLate" denotes the number of times the customer has defaulted more than 90 days in the past, which is highly correlated with the target variable "Serious Delinquents in 2 years". "NumberOfTimes60-89DaysPastNotWorse" denotes the number of times the customer has defaulted in his payments in the past 89 days. "NumberOfTimes30-59DaysPastNotWorse" denotes the number of times the customer has defaulted in the pass 60 days. Both these features are moderately correlated with the target variable. "RevolvingUtilizatiionOfUnsecuredLines" is defined as the ratio of the total non-secured debt to the total non-secured credit limit, which is a very good indicator of the type of borrower the customer is. Hence, these 4 features together give us the best accuracy in our credit scoring analysis.

Whereas in feature extraction, as the model combines all features to give a reduced set of features, some other features which might not be that relevant tend to deviate the model from these specific distinct features which are relevant. And that is why the accuracy seems to be less in this case.

8 CONCLUSION AND FUTURE WORK

From the experiments that we have carried out, we can observe that Feature selection using Stepwise Logistic Regression performed significantly better than Feature Extraction using Singular Value Decomposition and Weighted

Singular Weighted Decomposition (SVD). Feature Selection gave us the optimal accuracy using the 4 important features in our dataset, which was enough to predict the output of the target variable. Therefore, we can see that identifying important features in the dataset that mainly affect the accuracy of the credit scoring models can improve decision performance of the classifier, and improve the predictive accuracy while reducing overfitting risks in the model.

We have used Logistic Regression as the main machine learning algorithm for this project, but as future work other machine learning algorithms like XG Boost, Random Forests or ensemble models which use a combination of individual machine learning algorithms can be used to perform these experiments. As part of the future work other feature selection techniques like Tree based feature selection could be used to perform these experiments. Furthermore, the p value can be calculated to assess the degree to which observations are due to random events as part of the future work.

9 PROJECT SCHEDULE

The implementation of the project work took around 3-4 months. Within this time frame all the tasks mentioned in the method and approach section of the report were carried out. A basic credit scoring model was generated using Logistic Regression by the end of fifth week. By the end of 10 weeks all the tasks listed in the design of experiments were completed. The rest two weeks were utilized for writing the report. A more detailed schedule is elaborated in the table below.

Table 14: Project Schedule

EXPERIMENTS	WEEK
DATA SET EXPLORATION	0-1
LOADING DATA SET	1-2
FEATURE ENGINEERING	3-4
APPLYING LOGISTIC REGRESSION ON THIS MODEL	4-5
LOGISTIC REGRESSION USING STEPWISE LOGISTIC REGRESSION	5-6
LOGISTIC REGRESSION USING SINGULAR VALUE DECOMPOSITION (SVD)	6-7
FINDING MORE ABOUT THE NATURE OF THE PROBLEM	7-8
APPLYING WEIGHTS TO FEATURES BASED ON THEIR IMPORTANCE	8-9
LOGISTIC REGRESSION USING WEIGHTED SINGULAR VALUE DECOMPOSITION (SVD)	9-10

REPORT OF PROJECT WORK	10-12

10 REFERENCES

B. Benyacoub, S. El Bernoussi and A. Zoglat, "Building classification models for customer credit scoring," *2014 International Conference on Logistics Operations Management*, Rabat, 2014, pp. 107-111.

H. c. Chen and Y. c. Chen, "A comparative study of discrimination methods for credit scoring," The 40th International Conference on Computers & Industrial Engineering, Awaji, 2010, pp. 1-5.

D. Dukić, G. Dukić and L. Kvesić, "A credit scoring decision support system," *Proceedings of the ITI 2011, 33rd International Conference on Information Technology Interfaces*, Dubrovnik, 2011, pp. 391-396.

Y. q. Fan, Y. l. Yang and Y. s. Qin, "Credit scoring model based on PCA and improved tree augmented Bayesian Classification," *IET International Conference on Information and Communications Technologies (IETICT 2013)*, Beijing, 2013, pp. 169-175.

W. Li and J. Liao, "An Empirical Study on Credit Scoring Model for Credit Card by Using Data Mining Technology," *2011 Seventh International Conference on Computational Intelligence and Security*, Hainan, 2011, pp. 1279-1282.

Thomas C. Lyn, Edelman B. David, and Cook N. Jonathan (2002) Credit scoring and its application (America, 2002)

Mester J. Loretta ,(1997) ’’what’s the point of credit scoring?”

J. Ming-hui and C. Yu-fang, "Recombining Forecasts Used in Personal Credit Scoring," *2006 International Conference on Management Science and Engineering*, Lille, 2006, pp. 1719-1722.

Salome Tabagari, “Credit Scoring using Logistic Regression”, Tartu 2015

Y. Zhuang, Z. Xu and Y. Tang, "A Credit Scoring Model Based on Bayesian Network and Mutual Information," *2015 12th Web Information System and Application Conference (WISA)*, Jinan, 2015, pp. 281-286.

H. Zhou, J. Wang, J. Wu, L. Zhang, P. Lei and X. Chen, "Application of the Hybrid SVM-KNN Model for Credit Scoring," *2013 Ninth International Conference on Computational Intelligence and Security*, Leshan, 2013, pp. 174-177.

11 APPENDICES

11.1 Stepwise Logistic Regression results that were not included in the Results section.

- Using 1 feature, we get the following output:

  Accuracy = 0.742521367521

  AUC = 0.742848008769

  Feature_rank = [4 3 2 6 5 7 10 1 9 8]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans',' MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.76	0.72	0.74	952
1	0.73	0.76	0.74	920
avg / total	0.74	0.74	0.74	1872

- Using 2 features, we get the following output:

  Accuracy = 0.761217948718

  AUC = 0.761504384362

  Feature rank = [3 2 1 5 4 6 9 1 8 7]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans',' MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.78	0.74	0.76	952
1	0.75	0.78	0.76	920
avg / total	0.76	0.76	0.76	1872

- Using 6 features, we get the following output:

  Accuracy = 0.77938034188

  AUC = 0.779270186335

  Feature rank = [1 1 1 1 1 2 5 1 4 3]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans',' MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.78	0.78	0.78	952
1	0.77	0.77	0.77	920
avg / total	0.77	0.77	0.77	1872

- Using 7 features, we get the following output:

  Accuracy = 0.77938034188

  AUC = 0.779288454512

  Feature rank = [1 1 1 1 1 1 4 1 3 2]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents', 'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans',' MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.78	0.78	0.78	952
1	0.77	0.77	0.77	920
avg / total	0.77	0.77	0.77	1872

- Using 8 features, we get the following output:

  Accuracy = 0.775256410256

  AUC = 0.775211910851

  Feature rank = [1 1 1 1 1 1 3 1 2 1]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse', 'NumberOfTimes90DaysLate','NumberOfTime60-89DaysPastDueNotWorse','NumberOfDependents',

'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','
MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.78	0.78	0.78	952
1	0.77	0.77	0.77	920
avg / total	0.77	0.77	0.77	1872

- Using 9 features, we get the following output:

  Accuracy = 0.773653846154

  AUC = 0.773618012422

  Feature rank = [1 1 1 1 1 1 2 1 1 1]

  Features = ['NumberOfTime30-59DaysPastDueNotWorse',
  'NumberOfTimes90DaysLate','NumberOfTime60-
  89DaysPastDueNotWorse','NumberOfDependents',
  'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','
  MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.77	0.77	0.77	952
1	0.77	0.77	0.77	920
avg / total	0.77	0.77	0.77	1872

- Using 10 features, we get the following output:

  Accuracy = 0.77311965812

  AUC = 0.773037997808

  Feature rank = [1 1 1 1 1 1 1 1 1 1]

Features = ['NumberOfTime30-59DaysPastDueNotWorse',
'NumberOfTimes90DaysLate','NumberOfTime60-
89DaysPastDueNotWorse','NumberOfDependents',
'NumberRealEstateLoansOrLines','NumberOfOpenCreditLinesAndLoans','
MonthlyIncome', 'RevolvingUtilizationOfUnsecuredLines', 'DebtRatio','age']

	precision	recall	f1-score	support
0	0.77	0.77	0.77	952
1	0.77	0.77	0.77	920
avg / total	0.77	0.77	0.77	1872