KeAi
CHINESE ROOTS
GLOBAL IMPACT

Available online at www.sciencedirect.com

ScienceDirect

The Journal of Finance and Data Science 8 (2022) 145–161
http://www.keaipublishing.com/en/journals/jfds/

The use of predictive analytics in finance

Daniel Broby

Ulster University, Department of Accounting, Finance and Economics, Cathedral Quarter, Belfast, BT15 1ED, Northern Ireland, United Kingdom

Received 8 March 2022; revised 26 April 2022; accepted 16 May 2022
Available online 20 May 2022

Abstract

Statistical and computational methods are being increasingly integrated into Decision Support Systems to aid management and help with strategic decisions. Researchers need to fully understand the use of such techniques in order to make predictions when using financial data. This paper therefore presents a method based literature review focused on the predictive analytics domain. The study comprehensively covers classification, regression, clustering, association and time series models. It expands existing explanatory statistical modelling into the realm of computational modelling. The methods explored enable the prediction of the future through the analysis of financial time series and cross-sectional data that is collected, stored and processed in Information Systems. The output of such models allow financial managers and risk oversight professionals to achieve better outcomes. This review brings the various predictive analytic methods in finance together under one domain.

Keywords: Predictive analytics; Finance; Fintech; Regtech; Risk; Statistics; Machine learning; Decision support systems; Information systems

Contents

E-mail address: d.broby@ulster.ac.uk.
Peer review under responsibility of China Science Publishing & Media Ltd.

https://doi.org/10.1016/j.jfds.2022.05.003

1. Introduction

This paper reviews and explores the way that predictive analytic methods are used by scholars in finance, and in computational fields of study, to make forecasts. A systemic identification of methods and use cases is presented, specifically drawn from data sourced either internally or externally from financial Information Systems (IS). As a review, this paper does not explain the modelling process itself. This is adequately covered in many textbooks, such as that by.[1] It is instead a compendium of explanatory statistical and computational methods for the financial services industry, complied using bibliometric and keyword searches. It is relevant to the IS domain as predictive techniques are being increasingly incorporated into Decision Support Systems (DSS). The outputs are used in industry to achieve better commercial or product outcomes.

Finance is a broad term associated with activities in capital markets that encompasses intertemporal and portfolio decisions.[2] These activities frequently generate time series. These can be used to predict future values and or returns.[3] The foundation of statistical inference from these time series is the weak stationarity they exhibit in their means, variances, and covariances. It also lies in cross sectional series that are generated by financial transactions and markets. These provide a bedrock for statistical analysis and can in turn be imported into a DSS to predict customer, company or industry insights. An example is the prediction of a company's beta using cross sectional analysis. This is derived from the capital asset pricing model and represents the firms systematic risk relative to the market. The beta is the slope coefficient and is calculated simply by using a least squared regression.[a]

The term predictive analytics is more commonly applied when moving beyond explanatory statistical models to computational prediction. The latter are standard in DSS, and the former are increasingly being embedded into them. Table 1 highlights the key differences between these two approaches. The algorithmic techniques of the latter use test data sets to determine their validity and accuracy.[b] In contrast, statistical models use confidence intervals and significance tests. That said, as computational and statistical methods are used together in data analytics, both are covered in this review.

Within the field of finance, predictive analytics are used to examine historic IS data sets. The term covers both explanatory statistics and predictive computational methods. In column 2 of Table 1, explanatory statistics are listed as descriptive and backward looking. In column 3 of Table 1, predic-tive computational methods make forecasts and are listed as forward looking.[5] Unlike much applied forecasting in finance, the conclusions of a computational approach are drawn from the data rather than human interpretation. Indeed,[6] warn that researchers using computational predictive techniques must be comfortable with the idea that such research starts with data rather than with theory.[7]; however, warn that these same researchers should be wary of over-fitting, where the temptation might be to build a theory around the data and then modify the data to strengthen the theory.

In order for DSS to undertake predictive analytics itis necessary to have clean descriptive data from a corporate IS or a data warehouse. With the increasingly large, varied and complex nature of data, this becomes more of a challenge

[a] It can be calculated with greater degrees of sophistication to address time varying influences by using generalized autoregressive conditional heteroskedasticity (GARCH) and Kalman filter models.[4]

[b] The use of accuracy here refers to accuracy in the general sense. However, and for the avoidance of doubt, in machine learning applications, accuracy in the strict statistical sense is often less important than precision/recall.

Table 1
Differences between explanatory and predictive methods.

	Explanatory statistics	Predictive computational
Research problem	Testing hypothesis	Predicting possible outcomes
Systems output	Insight generated	Learning generated
Variables	Dependent and independent	System features/observables
Risks	Type I and type II errors	Overfitting
Optimization	Minimize model bias	Variance-bias trade-off
Evaluation	Statistical significance and Confidence intervals	Accuracy of out of sample performance

when faced with the heterogeneous, autonomous, complex and evolving "Big Data" commonly found in finance.[8,c] This raises a key point for IS. A distinction has to be made between internally and externally generated information and data. As such, section 2 of this paper reviews those datasets that have to be imported to an IS, and section 3 those that are generated by it. Addressing approaches to internally gathered raw data requires an interactive software-based system. Addressing approaches to external data requires suitable imported data-sets or a system for gathering distributed data.

Within predictive computational methods, the learning process is either unsupervised or supervised.[9] Unsupervised learning is a form of descriptive modelling that does not have an identified target variable or label. In finance, an example would be the reduction of the variables influencing a share price into number of factors. The maximum common variance of the variables is then assigned a score which gives the factor exposure of individual securities. In supervised learning, the label by contrast, both the variables and the nature of the desired predicted outcome are known. Using the example of a share price, this would be used where the past performance and attributes of a security are known but the analyst wishes to predict the future price.

Predictive analytics can be used in combination with IS data-sets to classify event outcomes, such as loan defaults, credit defaults and customer churn. It can be used to forecast numeric values, such as the price of a security or customer retention. It can further be used to identify anomalies, such as if a credit card transaction is fraudulent or if an insurance claim is falsified. It can also group data clusters found in IS, such as customer segmentation for sales targeting or identification of dissatisfied customers. Finally, as mentioned, it can be used for forecasting time series.

All predictive analytic models begin with a construction process. As mentioned, the data in the IS can be either internally or externally sourced. The Cross Industry Standard Process Model for Data Mining (CISPM - DM) outlines the steps for the internal data.[d] This includes gaining an understanding of the context and the data, as well as the preparation of that data. It details the modelling, evaluation and deployment stages. In the first step of the CISPM, IS domain experts are required. This is especially the case in finance, as there are extensive theories that explain asset pricing, the nature of banking and the behaviour of customers. These all require separate modelling. In the data preparation stage, variables need to be categorised and tabulated by finance domain experts into attributes, descriptors, variables, fields and features. Where external data and computational methods are applied, the IS need to be adapted accordingly.

The identification of outliers in the data is essential for good predictive analytics. Outliers can, for example, have a marked impact on the cross-sectional distributional properties of financial ratios.[10] These can be adjusted using winsorizing or trimming techniques. As a method for prediction, outlier models are also used. These are particularly useful for identifying fraudulent financial transactions. For example,[11] present a multivariate identification strategy that detects outliers in financial data. Outliers should be investigated thoroughly as they provide important information about the topic being investigated.

Once the data has been prepared, the predictive model can be built into the DSS. Such models can be divided into either parametric or nonparametric.[12] The former are where the distributions are known. Security prices, for example, are assumed to have a normal distribution by many capital asset pricing models. Machine learning models do not make such assumptions. They are therefore more flexible and iterative. They are also able to model non-linear data. For

[c] This type of data is increasing rapidly due to distributed computing.
[d] The CRISP-DM process model, methodology, reference model, and user guide can be found at: https://www.the-modeling-agency.com/crisp-dm.pdf.

example, they can be used to make share price trading models which include many time varying variables. In this way, nonparametric models distinguish predictive analytics from data mining and/or computational statistics, although there is a large overlap. As such, they are increasingly being applied in finance research and built into IS, as shall be shown.

In reviewing the methods used, this paper provides a contribution by extending the IS domain literature. It does this by including those scholars who have used data based predictive analytics to investigate questions of academic interest in finance. In gathering them together and reviewing them, it facilitates the adoption of the various techniques into DSS and research environments.

2. Methodology

To identify the methods, a broad literature review was conducted using multiple search engines in order to identify the most important uses of predictive analytics in finance. These are broken down by section later in the paper. This search followed the Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR) protocol proposed by.[13]

The use of the SPAR-4-SLR method is designed to consolidate knowledge in a domain. It ensures the review does not include replicative research and focuses on claims of novelty. This means the 53,500 results for the search term "Predictive Analytics in Finance" can be usefully processed. The numerous use cases and methods were assembled, arranged and assessed. The domain search identified over 187 major contributions, suggesting sufficient maturity to justify the subject for review. This is considered a good base, as[14] suggest that systematic reviews can be used at the domain level when more than 40 papers have been identified as suitable. The main techniques used are identified in Table 2. For the sake of brevity, not all of these papers are referenced in the synthesis of this review. The criteria for inclusion being usefulness to DSS and the significance of results produced by the method.

The identified key methods are addressed in each of the section headings. Care is taken to exclude papers that do not advance knowledge or add significant contribution to understanding of predictive analytic techniques. A more focused search was then undertaken on each of the theme headings, extracting those papers with the greatest citations and from the most learned journals. This excluded single country studies, as these typically build on the method of previous scholars.

Two broad categories were identified. Firstly, statistical methods that draw inferences, and secondly computational methods that find broad predictive patterns. Both can be used in DSS to undertake predictive analytics on a wide variety of finance variables. To ground the development of the predictive analytic domain, two research questions (RQ) were posed.

- RQ1 asked which top contributing authors in finance have utilised predictive analytics in their research.
- RQ2 asked what predictive analytic methodological choices were made and what the research context was.

Table 2
Main technique identified in literature search. This table shows the main method used in highly cited papers that claim unique insights. The count includes papers where the method may have been described with a different term. For example, the use of the GARCH approach is merged with time series.

Technique	Count	Merged	Technique	Count
Time series	42		Segmentation analysis	7
(Incl. described as GARCH)		(7)		
(Incl. described as ARIMA)		(4)		
Linear regression	32		Association rules	6
Neural networks	21		Optimization techniques	6
(Incl. described as ANN)		(8)	Quantile analysis	4
Logistic regression	13		Multivariate analysis	5
(Incl. described as LOGIT)		(4)		
Non linear regression	12		Classifiers	6
Clustering techniques	8		Random forests	5
SVM	8		Similarity approaches	4
Machine Learning	8			

The extant literature relevant to the search criteria is reviewed in section 3 and 4. These papers used the main statistical, classification, regression, clustering, association and time series models. These search terms were added to the key research areas in order to ensure all methods were captured. Search terms that also covered computational methods, such as machine learning and artificial intelligence, were also added in combination with the finance problem keywords. The insights drawn from this systematic approach are covered next.

2.1. Classification methods

Classification methods include Logistic Regression, Näıve Bayes, Stochastic Gradient Descent, K-Nearest Neighbours, Decision Tree, Random Forest, and Support Vector Machine (SVM). These are all common in Machine Learning.

Classification is a technique that helps predict discrete (generally binary) outcomes. Logistic regression is used to predict an outcome that either happens or does not. The most common classification method is the decision tree. This consists of nodes that form a rooted tree, meaning it is a directed tree that details how the classification is related, also with binary decisions. Random Forest is merely an extension of the decision tree approach to multiple trees. Naive Bayes is also popular, as it calculates the possibility of whether a data point belongs within a certain category.

Predictions using classification matrices impact many areas of finance, including portfolio construction, risk management, option pricing and strategic hedging. As such, there is a large body of literature devoted to such methods. Classification methods are used in financial DSS to make forecasting reports. For example,[15] predict stock market direction, which is either up or down, using a Support Vector Machine (SVM) and the classification method. In computational models such as these, classification is the process of partitioning a data-set into two classes. For example, a set of bank customers who never go overdrawn and a set of bank customers that includes those that do. Techniques that can be combined with classification in this way include predictive tables, co-variance matrices, similarity functions, Artificial Neural Networks (ANN) and the aforementioned SVM classifiers. Examples of how each are used in finance are presented next.

Predictive tables are used in analysing high frequency and ultra high frequency trading data where volatility clustering, heavy tails and time of day/week/month anomalies can be identified. For example,[16] analyse transaction batch frequency on 1.28 million buy transactions and 1.30 million sell transactions. With this form of analysis,[17] warn that the speed of transactions increases the errors. Further, that data has to be synthesised using linear interpolation for periods where there is no trading. That said, it is a new and growing part of the literature and shows what insights can be extracted from both internal and external IS.

Many finance problems are quadratic. These include univariate, where an unknown variable has to be predicted, and optimization problems. In the former, graphs can be used to predict points on a quadratic function. In the latter, covariance matrices can be used in the analysis of conditional variances, co-variances and correlations of financial returns.

Similarity functions play an import part in the financial prediction process. In particular when using certain parametric and nonparametric regressions neural nets, linear and nonlinear classifiers. The use of similarity functions in finance is normally where case based decisions are required. For example,[18] use it to predict bankruptcy by identifying similar past cases of financial distress. The main feature of this approach is the matching and retrieval of the prior cases. This takes the features of a prior bankruptcy and weights them. The matching process then retrieves these prior bankruptcies based on a weighted sum of their features compared to the company or companies being investigated.

The usefulness of computational methods in DSS is enhanced when addressing non-linear problems. Artificial Neural Networks and Support Vector classifiers are computational units that are particularly good at handling such non linear financial prediction problems. A review of their use in finance is produced by.[19] ANN's use a set of inputs in combination with additional inputs from other computational units to decide the value of an output. This allows them to compute bias adjustments and assign weights. By changing these weights, the ANN learns what the optimal outcome is. For example, ANN's have been used in fund management to make stock selection predictions. Several of these papers in science journals make use of technical indicators which are frowned upon by finance academics.[20] For example,[105] provide a Support Vector classifier approach based on weighted trading volume. They conclude that the classifier has limited accuracy and did not perform well with their proposed trading strategy. They suggest that future

research in this area could apply Random Forest methods to determine which features are likely to produce better results.

2.2. Regression methods

Regression techniques, used in explanatory modelling, are also used extensively in predictive finance. Indeed, they are the most common form of forecast models in DSS.

The most widely used model in finance is the capital asset pricing model (CAPM). This is essentially a single factor linear regression model: $E(Ri) = Rf + B\,(E(Rm)\,Rf)$. It is based on the expected return dependent variable $E(Ri)$ having a linear relationship to the market return $E(Rm)$ explanatory variable. As such, when individual security returns are regressed against the market, the resulting slope delivers the volatility of a security's return relative to the volatility of the returns of the market (Beta). This can then be used to describe the relationship between systematic risk, that element of returns that can't be diversified away, and expected return for assets. There are many papers that use the CAPM to make predictions. The reader is directed to 21 for an overview.

There are of course many types of regression, both linear and non linear, single or multiple. Many time series exhibit correlations but in finance it is difficult to identify whether the correlation between two or more variables is causal. Further, regressions suffer from a number of issues including omitted variables, reverse causality, mismeasurement and a limited focus. That said, this has not limited their use by finance academics who use there significance results to justify the acceptance or rejection of hypothesis.

When performing regression analysis to make predictions, the numerical output needs to be supplemented by a measure of how confident the researcher can be with the output prediction. These can be stated in confidence prediction or tolerance intervals.[22] warn of p-hacking, the manipulation of data to obtain desired *p*-values. They find this prevalent in financial papers that use of panel regressions to test for cross sectional equity risk factors.

Another way to incorporate a degree of confidence is to use quantile regression.[23] This method calculates the conditional median of the least squares rather than the mean (or any percentage of a particular value). It has the advantage of being able to order and sort samples. For example, there is dispersion and skew between the top paid employees in a company and the bottom paid ones.[24] therefore use this approach to identify if conditionally (predicted) high wage employees have a link between pay and performance than conditionally low-wage employees. Within finance, this is useful where the conditions of normal or linear relationships are not present, for example where the data exhibits homoscedasticity or independence.

The reason finance scholars use regressions is to reduce complex relation-ships found between many variables. As mentioned, the CAPM does this well. The normal distribution assumption, however, is just that, an assumption. Computational methods do not start with that as a basis and as such are well suited to DSS.

2.3. Clustering and association rule methods

Clustering and association rules are descriptive unsupervised classification approaches for data in IS when the various groupings, classifications and sub classifications are unknown. Within finance,[25] suggest they are important techniques for the understanding of time series. It also facilitates the visualization of the data relationships. For example, they can be used to classify securities into groups that share common characteristics and whose returns tend to vary and co-vary together without resort to a regression model such as the CAPM.

Association rules (AR) use Machine Learning (ML) models to check for patterns or related events in a database. For example, they are suggested by[26] as a way of detecting fraud in banks. Transactions or events that are usually associated with fraud, such as claiming frequent refunds, can be data mined and then flagged.

Both clustering techniques and association rules can be used in combination with other methods and in computational finance. For example,[27] used clustering in combination with a neural network based model to predict financial bankruptcy.

2.4. Time series models

Time series models, generated either internally or imported to an IS, are based on the assumption that a data series has a structure that exhibits itself over time. For example, the business cycle gives rise to trend and seasonal variation.

The internal structure of a time series may also exhibit auto correlation, such as when new information on a stock is gradually, rather than instantly reflected in its price. Because of this, regressions are less effective on such data series. As a result, predictive analytics can be applied using autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models.[28] Also, more sophisticated conditional modelling can be applied that utlitizes autoregressive conditional hetroskedasity (ARCH) and GARCH models.

ARMA, ARIMA, ARCH and GARCH are used according to ones as-sumption of the nature of the financial time series. These can be stationary (single period), integrated (seasonal), conditional (time varying) or lagged conditional (time varying with a delay). ARMA (stationary) and ARIMA (integrated) are both flexible predictive methods, however they are not appropriate to address nonlinearity and volatility, both common features of financial time series. ARCH (conditional) and GARCH (lagged) conditional models are better in this respect, and therefore more widely used in financial time series forecasting. This is also relevant because financial returns typically have an outlier-prone distribution. ARCH or GARCH models can be used to capture this effect in volatility.

GARCH and ARCH models are used in many conditional asset pricing applications.[29] find that conditional covariances of securities are variable over time and are an important determinant of time-varying risk premia. Within economics, time series models are often combined with Bayesian models in order to make forecasts of trends and cycles.[30] They can also, as with the other statistical techniques, be combined with computational methods.

3. Predictive analytics using external information and data

Organizational IS that source data from customers, vendors, regulators, and competitors are better able to make informed predictions than those that just rely on internally generated data. External data refers to syndicate data, which is data sourced from data suppliers, such as Bloomberg, Reuters, Datastream, and NielsenIQ. There are a number of external data sets that are particularly well suited to machine learning, including, Quandl, IMF, Simfin, Global Financial Development and Eurostat.

There is limited literature on acquisition of external data and its integration into a banks IS or data warehouses. Indeed, not all banks have an integrated data warehouse due to cost and complexity issues.[31] provide an interesting study on the subject. They conclude that the motive for banks to develop a data warehouse is to analyze and track customer behaviour and to increase operating efficiency.

3.1. Economic prediction

From the top down, predicting the future direction of the economy is important to almost all aspects of finance. As mentioned, time series models are popular autoregressive amongst scholars and these are almost exclusively from external databases. That said, they can be integrated with corporate IS and indeed are at the research departments of large investment banks.

There is a substantial amount of peer reviewed papers that use predictive methods in economics. Mostly, these are regression based. Such Prediction models include linear or logistic regressions. Nonlinear methods include many of those introduced such as classification models, random forests and penalized regression such as Least Absolute Shrinkage and Selection Operator (LASSO) and least-angle (LARS) regressions.

The most common predictive approach is to use a stochastic vector dif-ference equation with an event that requires policy action. The analyst then specifies a model based on assumptions related to the exogenous variables. Examples include the Wharton model (MARK III) and various ARIMA mod-els which forecast quarterly macroeconomic variables.[102]

Following the financial crisis of 2008, there was an increase in papers us-ing predictive analytics to forecast and model such events. For example,[32] present a Coupled Market State Analysis (CMSA), where a set of dynamic states represents a crisis induced by the dynamic market interactions. The predicted outcome outperformed a number of other approaches including ANN and logistic regressions. A survey of the ANN methods in finance is provided by.[33] They distinguish between uni-variate and multivariate forecasting.

The economy is, however, fundamentally uncertain. As such, there is a lack of observed accuracy in the predictions in economics such as inflation and GDP growth.[34] Despite the power of computational models, there isn't a great deal of successful research utilizing predictive analytics in the field of economics. This is because the primary drivers of model

performance are its complexity. The later is defined by the number of model parameters and their corresponding relationships.[35] propose a model confidence set (MCS) to help econometricians choose between models given a predetermined level of confidence.

3.2. Earnings prediction

Corporate earnings are easier to forecast than the economy, although they too have a high degree of idiosyncratic complexity. As with economic variables, the data has to be sourced from an external vendor. Unlike in economics, however, regressions are not good predictive techniques. Nonlinear models such as tree induction algorithms, neural networks, naive Bayesian learning, and genetic algorithms have had better success in forecasting earnings Neural networks, for example, can be used to forecast earnings based on goodness-of-fit and additional training factors in the error function including net profit, direction, and time horizon.[36]

Using computational methods to predict corporate earnings addresses the twin issues of objectivity and independence in human analysts and as such are well suited to DSS. The issue of accuracy is, however, very pertinent to valuation, a common objective of predictive models in finance. SVM therefore may have an edge when used in a systems context.[37] checked the results of their own support vector regression prediction on quarterly earnings and found it to be better than an ARIMA model.[38] also identify SVM models as appropriate for earnings prediction. In a similar vein,[39] compared the various techniques and found that genetic algorithms have a few advantages over the others in making predictions.

3.3. Stock price, returns and volatility prediction

The stock-market, because it is considered as important to the process of financial intermediation, is an area which has proved extremely challenging to predict. Despite this, there is an extensive literature on stock price prediction The strong form of the Efficient Market Hypothesis suggesting that prices listed in it reflect all available information.[40] The stock-market is therefore considered by[41] and other finance scholars to be efficient if prices would be unaffected by reveal-ing any available information to everyone.

There are a number of estimation models based on matrices in the litera-ture, the two most popular academic ones being the scalar Dynamic Conditional Correlation (DCC) model of[42] and the Dynamic Equicor-relation (DECO) model of.[43] The predictions made by such models are conditional on the Generalized Auto Regressive Con-ditional Heteroskedasticity (GARCH) estimates of previous stages. In an expansive review of covariance estimation methods[44] conclude that none of the methods work in all scenarios. The sample period, rebalancing criteria and portfolio constraints influence each of the models in different ways. Also, GARCH models have proved impractical for large datasets. Univariate GARCH models are therefore becoming popular in the literature. There is also quite a large and growing literature on estimation error in portfolio optimization.[45]

Many investors and financial service companies make fundamental forecasts and base investment strategies on them.[46] A smaller number make technical forecasts based on historic price movements. Although the latter is considered ineffective, many of the predictive models are based on this approach.[47–49]

There is a great deal of literature on forecasting returns, the methods used being relevant to DSS designed for this purpose.[50] suggest a predictive approach can based on a set of forecasting models, a search technology, a real time price feed, a risk premium model and an estimate of transaction costs. This leads to two types of approach, namely, statistical and computing. The most common statistical method is the estimation of the conditional mean using classical least-squares and maximum likelihood estimation. Other statistical techniques include decision trees, various multiple linear regressions, exponential smoothing, autoregressive integrated moving averages (ARIMA), and generalized autoregressive conditional heteroskedasticity (GARCH) volatility modelling.

The problem is that traditional statistical models such as these are not always good in capturing the complexity of stock price behaviour. The results that they deliver are based on restrictive assumptions. For example, although leverage affects all stock returns differently, such models would conclude that it will affect them all in the same way. Therefore, predictive analytic techniques that incorporate non linear techniques such as SVM, Support Vector Regression (SVR), and Neural Networks (NN) can prove more informative. For example,[15] used a support vector machine method to predict market direction.

There is also substantial literature on the predictability of stock returns based on the time horizon.[51]; for instance, found that stock return predictability is a function of the return horizon. They found that they could predict variation in aggregate returns over time. These were about three percent in their short term sample and 25 percent in their longer time horizon. They used a simple model for stock prices that combined stationary price and random walk elements.

3.4. Optimal portfolio prediction

Predictive models are also used to forecast optimal outcomes when faced with quadratic problems. Modern Portfolio Theory postulates that it possible to construct a meanvariance "efficient frontier" that consists of an optimal portfolio of securities which deliver the maximum possible expected return for the risk taken.[52] divide the role of predictive analytics into market analysis, financial analysis, and earnings estimates. They also argue the textual analysis of market research reports can provide predictive information.

There is a growing body of literature that makes use of ML techniques in such an approach.[53] Most scholars classify the optimization challenge as a supervised learning problem.[54,103] For example,[55] propose a Performance Based Regularization (PBR) ML method. This constrains the sample variances leading to a lower estimation error.

Machine leaning can also be used in combination with regression techniques to achieve optimal portfolios.[56] take a quantile regression approach arguing it captures the characteristics and stochastic relationship of the variables used to forecast the expected returns and hence how to incorporate stocks into an optimized portfolio.[57] derive a quantitative trading model based on linear regression and support vector regression models that predict stock movement.

Prediction and estimation are data mining tasks. Typically, they are done using a valid sample of the data. Although by using big data techniques, entire financial data-sets can be analysed. Financial markets, by their nature, are concerned about the accuracy of the estimate. As such, if the entire data-set is not utilized the sample error is important.[58] argue that to use the dynamic CAPM to make predictions one should use the Variance Risk Premium (VRP) due to its high predictive power.

3.5. Audit and compliance prediction

The audit and compliance functions have been slow to adopt computational techniques but[59] predict these will become more important in a third wave of systems that they predict professionals will embrace.[60] caution, however, that predictive analytics in this area involve increased model complexity and may lack explanatory insights. Despite this, there is a nascent research agenda focused on regulatory technology (regtech) and predictive data mining.

Predictive analytics can also be used for enforcement. For example,[104] documented the statistical methods for predicting money laundering. Typically, a profile of legitimate behavior is compared with a money laundering example. The two sets are then blended into a single nu-merical value representing suspected laundering. This is done using a Bayes ratio related metric.

The majority of audit and compliance checks are risk based. Predictive analytics based on tobit and logit regression models can, however, be applied to the outcomes of past audits to identify likely non-compliance. For example,[61] investigated the tax system. They used the tobit model to target to identify predicted tax evasion and the logit model for noncompliance with the tax code.

3.6. Credit score prediction

A credit history allows a lender to evaluate the risk of extending a loan.[62] provide an overview of the literature on statistical techniques and evaluation criteria used in credit score prediction. They con-clude that there is no one technique which is superior. They also claim that credit score models have proved successful in their application.[63] found similar results testing nonlinear classifiers, including neural net-works, support vector machines, generalised boosting, and random forests. They even found that simple linear logit and probit classifiers were also able to predict with reasonable accuracy.

Early credit scoring used a number of characteristics captured by a banks internal IS, such as occupation, length of employment, marital status, in-come, and rent/mortgage payments. Multivariate regression and/or discriminant analysis are then applied.[64] These methods, however, do not take into account the complex interaction between these factors.

With the advent of Open Banking, there is now the possibility to get more granularity and more immediate information using third party permissioned data.

There are a number of techniques that are used in retail banking score pre-diction including logistic regression (LR), discriminant analysis, Classification and Regression Tree (CART), artificial neural network (ANN) and Cascade Correlation Neural Network (CCNN). These are used to build scoring models based on information in IS. These can be divided between function-based methods and if-then induction methods.

Improvement in accuracy is important for credit scoring as it avoids bad loans being made. It is possible to measure the success of these using ROC curves and Gini coefficients. The robustness of these can be tested using Kolmogorov Smirnov curves. It can also be observed that ANN performs poorly in small samples when incorporating poorly defined attributes or small data sets.[65]

4. Predictive analytics using internal information and data

As explained, the bulk of predictive analytics within a financial institution are based on internally generated information and data, In order for these to be well decomposed,[66] suggest that IS be built with a representational model, a state-tracking model, and a good decomposition model, The deep structure of IS will therefore become more relevant. Similarly, as computational methods are embedded within DSS, they too will become more relevant.

General Systems Theory (GST) is used to support the goals and strategy of an organization. In this respect, within finance, the key goal of a DSS is to deliver optimal financial solutions. This means either driving profits or avoiding losses. The subsequent subsections cover the prediction of these drivers. In a review of the role of big data in predictive analytics[67] note that IS uses of predictive models can be divided into three phases. The first, early database systems with structured data in relational database management systems (RDBMS). The second, information gathered externally from the Internet. The third, a more recent wave based on the internet of things. The key methods that make use of such internal data are covered next.

4.1. Customer acquisition and attrition prediction

The most common DSS predictive analytics relate to customer behavior. Customer acquisition prediction methods used in the literature include Monte Carlo Simulation, Logistic Regression or Neural Network. Avoiding losing customers has also been investigated. A number of methods have been used to predict customer retention including Customer Life Time Value: Hybrid Data Mining, Markov Chain, Optimization, Analytical Hierarchy Process, SVM, and Quantile Regression.[68] investigated three other methods, logistic regression, decision tree and discriminant analysis for accuracy. They conclude the discriminant approach is the least accurate.[69] provide a good overview of the various methods.[70] suggest SVM is the best approach, They used a bank database to predict optimal levels of customer churn.

[71] produce a method for linking customer acquisition with customer retention. This combines a Tobit model with a binary Probit one. They argue this avoids misleading inferences. They conclude by critiquing firms that do not collect information on their customers in a data warehouse, supporting the view that this is an important aspect of DSS.

[72] proposed a way of improving customer attrition prediction by integrating emotions in emails and evaluating multiple interaction classifiers. They suggest this can be done in combination with logistic regression, SVM and random forests. They favour the latter.

Link analysis can be used to identify connections and records in a financial company's database. For example, a private client asset manager can identify family links amongst high net worth individuals. An approach to doing this on scale, the Knowledge Discovery in Databases (KDD) method is explained by.[73] They define it as a “nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. This can then be used in financial services to target new customers.

4.2. Customer segmentation and sales prediction

Also central to DSS for management is the forecasting of sales. Scholarly work on methods of sales prediction falls into two camps, evaluation-focused and prediction-focused. Either way, prediction models rely on classification and segmentation, a process done at the IS rather than DSS level. Attributes are typically classified into either “benefit” or

"cost" attributes. An example of a benefit attribute in finance is response-time to a customer enquiry. A cost attribute would be the fee level relative to competitors, These then need to be ordered, a method for which is explained by.[74]

Customer segmentation is best done with visualization techniques. Statistical techniques include link analysis, deviation detection, dependency modeling summarizing. Regression forests, random forests, and logistic regression. Customer segmentation allows for a more targeted marketing effort and even aids with sales prediction.[75] suggest that there are three methods of clustering that may be suitable to help with the segmentation of the diverse nature of a bank's customer base. They suggest it is important to have a knowledge of density based approaches. These help with data analysis and identify outliers and any closely related clusters from the extracted data.

Closely allied to the concept of sales prediction is customer satisfaction. This is because existing customers make up the bulk of a mature firm's revenue. In finance, satisfaction is linked to outcomes, experiences and value for money. All of these can be measured in terms of expectations, performance, disconfirmation and satisfaction. There are two groups of models which can be applied to predict satisfaction in financial services. The first grouping combines structural equation models (SEMs) and regression models. These have not proved very accurate as they are based on linear assumptions, and dissatisfaction tends to spread in a nonlinear way. For example, a banking data privacy leak might result in a spike in dissatisfaction. A second group of artificial neural network (ANN) models are used in practice. The literature on the use of these techniques in academic papers is, however, sparse due to sensitivity about the commercial nature of the data.

Data from the internet can be used to extend insights gained from internal IS data.[76] investigated customer satisfac-tion based on their disconfirmation between two products using an analysis of variance method on observed response. A binary approach was taken by.[77] They looked at social media posts and classified them as either satisfied or dissatisfied, identifying emotional language. They then used a SVM classifier with a linear kernel to use personality traits to predict outcomes.

There is also some literature focus on customer life time value and cross selling prediction. Methods used include sequence analysis, hybrid data mining, Markov chain, optimization, analytical Hierarchy Process, SVM, quantile regression techniques and segmentation. For example,[78] undertook a singular value decomposition which used keyword similarity to quantify customers online reviews and[79] use a kernel-based learning method to predict a variety of customer behaviours.

4.3. Credit default prediction

Unlike credit score prediction which is largely done using external data sets, credit default prediction is done using data from internal IS. The business cycle, varying interest rates and business conditions can all give rise to default on extended credit. Predictive methods include decision Tree Regression, Altman Z, ANN, RS, and SVM.[80] provide an overview of the various techniques and propose an ANN approach.

There are many other factors which can contribute to a default, thereby complicating its prediction. There is a substantive literature on these as well as creditsensitive instruments. The primary focus of the literature is on corporate bonds and their default. A large part of this consists of papers on empirical structural models based on[81] work on the pricing of corporate liabilities and predictions using option pricing, or reduced-form models that exogenously predict default probabilities.[82] Structural models are non-parametric. They make the assumption that the researcher has the same information as a company's insiders, which is often not the case. Reduced-form models are parametric. They only require that the researcher has the same information as the market. The definition of bankruptcy presents a data identification issue for predictive models. In the United States, there are six levels of bankruptcy under the Bankruptcy Code. Scholars have preferred to use the term failure as a result of this. Some view failure as a bankruptcy, rehabilitation or liquidations filing, whilst others others focus on financial stress or credit default.[83] used a datamining approach to develop bankruptcy prediction models for various economic conditions. Their accuracy rates were all better than random guesses. That said,[84] also tested various bankruptcy models and found that a human expert outperformed the predictions of the IS forecasting techniques.

The literature on bankruptcy prediction goes all the way back to the 1930's.[85] The early models are based on univariate single variables or financial ratios. Most studies are now multivariate, reflecting the many contributing factors to financial distress. The Z-Score model is one of the most well know financial distress model, giving a score that predicts bankruptcy[86] This multivariate model has been improved over time with[87] proposing a probit and[88] a logit approach. The former identified sample selection and over-fitting methodological issues.

4.4. Financial and credit fraud prediction

Financial markets attract fraudsters and DSS are used to identify these early.[89] classify the literature on financial fraud into three streams, namely false financial disclosures, fraudulent financial schemes, and financial misselling. The techniques used in the literature include a combination of statistical, data visualization, data mining, and filtering tools.
[90] point out that prediction is made difficult due to the rarity of fraud, the large numbers of explanatory variables and the large number of different types of fraud.[91] provide a classification and literature review on the latter. It is clear from their work, and the work of other scholars, that successful frauds are complex and hard to detect. For example, fraudulent activity at Madoff Investment Securities LLC was documented for ten years prior to being exposed by a fund manager who mathematically showed that the performance track record was mathematically impossible to achieve.

Credit card fraud is typically only identified with a time lag. The perpetrators therefore have time in which to commit multiple offences with the same account. Big data analysis can help in a preventative way.[101] Predictive analytics can be used on a large volume of data in combination with machine learning algorithms, data mining approaches such as support vector and random forest, together with more traditional logistic regressions. All these approaches are based on a benchmark data-set of normal spending patterns. In this way, it is possible to detect unusual spending patterns or frauds in real time.

Supervised and unsupervised predictive analytic approaches are used to detect credit card fraud. In supervised approaches, models are based on comparing two samples, one with clean legitimate transactions and one with known fraud. It is then possible to teach the machine to classify new transactions as either legitimate or fraudulent. In unsupervised approaches, models are also based on comparing two samples, one with clean legitimate transactions and one with suspected fraud. In this approach, the suspect trans-actions show up as outliers. The statistical output gives the probability of a fraudulent transaction.

There is also significant credit card application fraud.[92] propose using peer group analysis. This involves smaller classifications of groupings where the researcher calculates the standardized distance from the peer group behaviour. In this way, the slope of credit card spending over time can be compared on newly issued cards. Sudden and excessive usage versus the peer group would merit investigation.

4.5. Cyber crime prediction

As with other predictive methods, both statistical and computational methods are used in DSS to predict cyber crime. The most common include various forms of regression and time-series. The computational approaches include Artificial General Intelligence (AGI), ANN, ML, genetic algorithms, fuzzy logic, NLP and robotics. Often, the techniques are layered on top of each other. For example,[93] use data mining to get the data, apply association rules, then k-means clustering, followed by classifiers.
[94] provide an overview of computational techniques. They point out that there are a number of challenges faced by IS. These include inadequate information about the threat, the large amount of real time network data, and the changing nature of the treat, The literature on cyber crime is focused on narrow and specific issues. Theses include denial of service, corporate blackmail, cyberbullying, stalking, scams, robbery, identity theft, defamation and harassment.[95] produce a survey of computational methods used.

Much of the literature on prediction focuses on determining the magnitude of an event once it has started.[96]; for example, predict the severity of denial of service attacks. The challenge with these approaches is getting the data from the internet rather than an IS.

An interesting approach was taken by[97] who investigated spontaneous deception by analysing language used in interactive online games. They used logistic regressions on this data and found that deceptive players had certain language traits that could be predicted. These include cognitive load, latency, and wordiness.
[98] proposed a genre tree kernel method that uses fraud cues to predict phishing and thereby enhance anti-phishing capabilities. One technique to not just predict but to prevent is to maintain a Reputation Black-List (RBL). Combined with the aforementioned techniques, they can be combined to make predictions for future security incidents.

5. Discussion

The predictive analytics presented are increasingly being embedded in DSS. These systems source data from internal IS and external databases. Fig. 1 depicts how this process works in order to provide useful information to management in financial services. Obviously, each DSS is unique and the methods will have to be selected for individual tasks. As such, no one method or collection of methods can be prescriptive. Importantly, it is the learning element that represents an advance in DSS capabilities.

The literature has shown that explanatory methods that rely on causal relationships have proved less successful, in non-linear problems, than some computational methods. It is suggested that this supports the case that financial DSS should incorporate both techniques in order to provide management of financial services with a more diverse set of explanatory tools. This enables non computing financial staff to interact with data stored in an IS and retrieve it in an automated fashion to make predictions. These forecasts differentiate such DSS from Management Information Systems (MIS). The former are designed for analysts and senior management, the latter for performing set functions and aimed at middle management.

It has to be emphasised that there is a major limitation to the methods reviewed in the literature. This is because there is less coverage in the literature of the many failed attempts to utilize such methods. Within finance numerous investment strategies fall into this camp. There are also strategies that produce false positives. This occurs within finance because so called "successful" strategies are sometimes justified by over-fitting back-tests. The success is down to data-

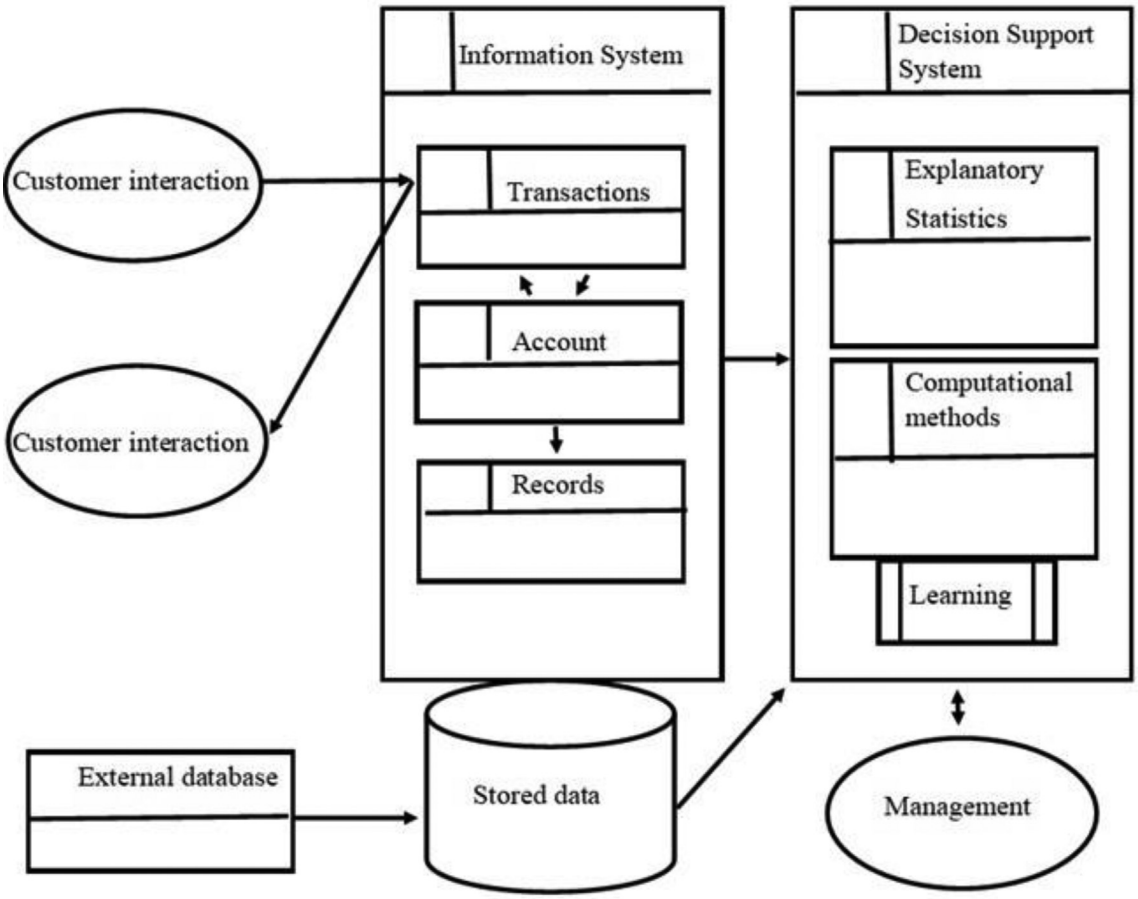

Fig. 1. The diagram illustrates a banking Information System (IS) which incorporates a Decision Support System (DSS) that has predictive analytics embedded in it. The DSS is able to perform explanatory statistic as well as computational methods. The IS enables management to make more informed decisions based on both internal and external databases.

snooping rather than skill. A similar issue is seen in those "successful" strategies that are "based on sound academic research", but which deliver poor results on a risk adjusted basis.

Another area which the literature does not cover well is that of unstructured data. There has been less coverage of methods that use such data despite the growth of social media predictive tools. This is expected to be a growing area of academic interest in the future. A whole new genre of methods is being developed around the use of language text and context. These are covered by.[99] They review the literature on predictive analytics in finance using text sources such as news reports, blog posts, internet posts, and corporate filings. For example,[100] use text regression from SEC annual report filings to predict risk. The evolving role of social media will mean that Natural Language Processing (NLP), Information Retrieval (IR), structured and unstructured Data Mining (DM) methods will increasingly be used.

Overall, the literature illustrates the rich and varied use of predictive methods for DSS. It is clear, however, that an understanding of their performance is important to the evaluation of their suitability for use. Various scholars suggest that descriptive, diagnostic, and predictive methods may well extend in the future to incorporate prescriptive analytics. This would prove an interesting future direction for DSS research and development.

6. Conclusion

This paper presented a rich variety of methods in its review of predictive analytics in finance. A keyword index search was made which covered the explanatory statistical techniques and computational models. This was done through a review of the top contributing authors from the disciplines of Information Systems, Finance, Computing and Management Science. The identified scholars all utilised predictive analytics in their research. Their analytic and methodological choices were identified in the context of their research. It was shown that scholars used such methods to deliver statistical results, descriptive outputs, decision forecasts and optimized outcomes. It is suggested that these can all be embedded in DSS to aid management decision making. This allows management to combine IS data with computational power, thereby enhancing the ability to make better financial, investment and lending decisions.

It was shown that statistical and computational techniques sourced from IS, either externally or internally, have been used by academics to solve a number of financial questions. These include the use of external data to forecast the economy and corporate earnings. In this respect, predictive models were shown to have been used in combination with external data to analyze historic time series and help with the forecasting of returns. Similarly, descriptive models were shown to have been used in finance to quantify relationships, and help in the prediction of risk. These include the use of internally generated data to predict and understand the customer, as well as modelling credit and risk.

The paper addresses important IS domain gaps, especially in the practical implementation of predictive analytics and modeldriven decision support. It was found that within the finance field, the term predictive analytics is synonymous with identification of anomalies and forecasting using computa-tional models that use data collected, stored and processed. Such techniques are increasingly being embedded in IS, thereby enhancing managements ability to forecast. It was also found that there are multiple methods can be used to address particular problems. From a systems design perspective, choices therefore need to be made. This is supportive of the need for IS professionals to have an overview of the literature. The identification of academic methods in this paper should therefore prove useful to IS and DSS systems designers. It is concluded that as DSS will evolve and incorporate and embed more computational methods. These will help with analysis of nonlinearity and complex financial questions. Within academia, it is expected that greater use of internally IS generated data will be used to provide scholarly insights. It is also expected that the methods used in academia will be increasingly used in industry.

Declaration of competing interest

The author has none to declare.

References

1. Kuhn M, Johnson K. *Applied Predictive Modelling Springer. New York Heidelberg Dordrecht London*. 2013.
2. Summers LH. On economics and finance. *J Finance*. 1985;40:633–635.
3. De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. *Int J Forecast*. 2006;22:443–473.

4. Choudhry T, Wu H. Forecasting the weekly time-varying beta of UK firms: garch models vs. kalman filter method. *Eur J Finance*. 2009;15:437–444.
5. Tsay RS. *Analysis of Financial Time Series*. vol. 543. John Wiley & Sons; 2005.
6. Müller O, Junglas I, Vom Brocke J, Debortoli S. Utilizing big data analytics for information systems research: challenges, promises and guidelines. *Eur J Inf Syst*. 2016;25:289–302.
7. de Prado MML. *Machine Learning for Asset Managers*. Cambridge University Press; 2020.
8. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. *IEEE Trans Knowl Data Eng*. 2013;26:97–107.
9. Dixon MF, Halperin I, Bilokon P. *Machine Learning in Finance*. Springer; 2020.
10. Frecka TJ, Hopwood WS. The effects of outliers on the crosssectional distributional properties of financial ratios. *Account Rev*. 1983:115–128.
11. Adams J, Hayunga D, Mansi S, Reeb D, Verardi V. Identifying and treating outliers in finance. *Financ Manag*. 2019;48:345–384.
12. Zhao Z. Parametric and nonparametric models and methods in financial econometrics. *Stat Surv*. 2008;2:1–42.
13. Paul J, Lim WM, O'Cass A, Hao AW, Bresciani S. Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *Int J Consum Stud*. 2021;45(4):1–16.
14. Paul J, Criado AR. The art of writing literature review: what do we know and what do we need to know? *Int Bus Rev*. 2020;29:101717.
15. Huang W, Nakamori Y, Wang SY. Forecasting stock market movement direction with support vector machine. *Comput Oper Res*. 2005;32:2513–2522.
16. Broby D, Basu D, Arulselvan A. The role of precision timing in stock market price discovery when trading through distributed ledgers. *J Business Thought*. 2019;10:1–8.
17. Fabozzi F, Focardi SM, Jonas C. High-frequency trading: methodologies and market impact. *Review of Futures Markets*. 2011;9:7–38.
18. Park CS, Han I. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Syst Appl*. 2002;23:255–264.
19. Coakley JR, Brown CE. Artificial neural networks in accounting and finance: modeling issues. *Intell Syst Account Finance Manag*. 2000;9:119–144.
20. Menkhoff L, Taylor MP. The obstinate passion of foreign exchange professionals: technical analysis. *J Econ Lit*. 2007;45:936–972.
21. Fama EF, French KR. The capital asset pricing model: theory and evidence. *J Econ Perspect*. 2004;18:25–46.
22. Harvey CR, Liu Y. Lucky factors. *J Financ Econ*. 2021;141(2):413–435.
23. Koenker R, Hallock KF. Quantile regression. *J Econ Perspect*. 2001;15:143–156.
24. Hallock KF, Madalozzo R, Reck CG. Ceo pay-for-performance heterogeneity using quantile regression. *Financ Rev*. 2010;45:1–19.
25. Aghabozorgi S, Shirkhorshidi AS, Wah TY. Timeseries clustering–a decade review. *Inf Syst*. 2015;53:16–38.
26. Tackett JA. Association rules for fraud detection. *J Corp Account Finance*. 2013;24:15–22.
27. Shah JR, Murtaza MB. A neural network based clustering procedure for bankruptcy prediction. *Am Bus Rev*. 2000;18:80.
28. Franses PH, Van Dijk D, et al. *Non-linear Time Series Models in Empirical Finance*. Cambridge University Press; 2000.
29. Bollerslev T, Engle RF, Wooldridge JM. A capital asset pricing model with time-varying covariances. *J Polit Econ*. 1988;96:116–131.
30. Pole A, West M, Harrison J. *Applied Bayesian Forecasting and Time Series Analysis*. Chapman and Hall/CRC; 2018.
31. Hwang HG, Ku CY, Yen DC, Cheng CC. Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in taiwan. *Decis Support Syst*. 2004;37:1–21.
32. Cao W, Cao L. Financial crisis forecasting via coupled market state analysis. *IEEE Intell Syst*. 2015;30:18–25.
33. Huang W, Lai KK, Nakamori Y, Wang S, Yu L. Neural networks in finance and economics forecasting. *Int J Inf Technol Decis Making*. 2007;6:113–140.
34. Boero G, Smith J, Wallis KF. Uncertainty and disagreement in economic prediction: the bank of england survey of external forecasters. *Econ J*. 2008;118:1107–1127.
35. Hansen PR, Lunde A, Nason JM. The model confidence set. *Econometrica*. 2011;79:453–497.
36. Yao J, Tan CL. A study on training criteria for financial time series forecasting. In: *Proceedings of International Conference on Neural Information Processing*. Citeseer; 2001.
37. Fischer JA, Pohl P, Ratz D. A machine learning approach to univariate time series forecasting of quarterly earnings. *Rev Quant Finance Account*. 2020;55:1163–1179.
38. Amani FA, Fadlalla AM. Data mining applications in accounting: a review of the literature and organizing framework. *Int J Account Inf Syst*. 2017;24:32–58.
39. Dhar V, Chou D. A comparison of nonlinear methods for predicting earnings surprises and returns. *IEEE Trans Neural Network*. 2001;12:907–921.
40. Degutis A, Novickytė L. The efficient market hypothesis: a critical review of literature and methodology. *Ekonom*. 2014;93:7–23.
41. Malkiel BG. Efficient market hypothesis. In: *Finance*. Springer; 1989:127–134.
42. Engle R. Dynamic conditional correlation: a simple class of mul- tivariate generalized autoregressive conditional heteroskedasticity models. *J Bus Econ Stat*. 2002;20:339–350.
43. Engle R, Kelly B. Dynamic equicorrelation. *J Bus Econ Stat*. 2012;30:212–228.
44. Trucíos C, Zevallos M, Hotta LK, Santos AA. Covariance prediction in large portfolio allocation. *Econometrics*. 2019;7:19.
45. Britten-Jones M. The sampling error in estimates of meanvariance efficient portfolio weights. *J Finance*. 1999;54:655–671.
46. Ou JA, Penman SH. Financial statement analysis and the prediction of stock returns. *J Account Econ*. 1989;11:295–329.
47. Fama EF, Blume ME. Filter rules and stockmarket trading. *J Bus*. 1966;39:226–241.
48. Van Horne JC, Parker GG. The randomwalk theory: an empirical test. *Financ Anal J*. 1967;23:87–92.
49. Jensen MC, Benington GA. Random walks and technical theories: some additional evidence. *J Finance*. 1970;25:469–482.
50. Timmermann A, Granger CW. Efficient market hypothesis and forecasting. *Int J Forecast*. 2004;20:15–27.

51. Fama EF, French KR. Permanent and temporary components of stock prices. *J Polit Econ*. 1988;96:246–273.
52. Pachamanova DA, Fabozzi FJ. Recent trends in equity portfolio construction analytics. *J Portfolio Manag*. 2014;40:137–151.
53. Ayodele TO. *Types of machine learning algorithms. New advances in machine learning*. 3. 2010:19–48.
54. Henrique BM, Sobreiro VA, Kimura H. Literature review: machine learning techniques applied to financial market prediction. *Expert Syst Appl*. 2019;124:226–251.
55. Ban GY, El Karoui N, Lim AE. Machine learning and portfolio optimization. *Manag Sci*. 2018;64:1136–1154.
56. Ma L, Pohlman L. Return forecasts and optimal portfolio construction: a quantile regression approach. *Eur J Finance*. 2008;14:409–425.
57. Ta VD, Liu CM, Addis D. Prediction and portfolio optimization in quantitative trading using machine learning techniques. In: *Proceedings of the Ninth International Symposium on Information and Communication Technology*. 2018:98–105.
58. Van Binsbergen JH, Koijen RS. Predictive regressions: a present- value approach. *J Finance*. 2010;65:1439–1471.
59. Elliott RK. The third wave breaks on the shores of accounting. *Account Horiz*. 1992;6:61.
60. Butler T, O'Brien L. Understanding regtech for digital regulatory compliance. In: *Disrupting Finance*. Cham: Palgrave Pivot; 2019:85–102.
61. Hashimzade N, Myles GD, Rablen MD. Predictive analytics and the targeting of audits. *J Econ Behav Organ*. 2016;124:130–145.
62. Abdou HA, Pointon J. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intell Syst Account Finance Manag*. 2011;18:59–88.
63. Jones S, Johnstone D, Wilson R. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *J Bank Finance*. 2015;56:72–85.
64. Chatterjee S, Barcun S. A nonparametric approach to credit screening. *J Am Stat Assoc*. 1970;65:150–154.
65. Féraud R, Clérot F. A methodology to explain neural network classification. *Neural Network*. 2002;15:237–246.
66. Wand Y, Weber R. On the deep structure of information systems. *Inf Syst J*. 1995;5:203–223.
67. Jeble S, Kumari S, Patil Y. Role of big data and predictive analytics. *International Journal of Automation and Logistics*. 2016;2:307–331.
68. Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH. Defection detection: measuring and understanding the predictive accuracy of customer churn models. *J Market Res*. 2006;43:204–211.
69. Lazarov V, Capota M. Churn prediction. *Bus. Anal. Course. TUM Comput. Sci*. 2007;33:34.
70. Kumar AS, Chandrakala D. An optimal churn prediction model using support vector machine with adaboost. *Int J Sci- entific Res Computer Sci, Eng and Information Tech- nology*. 2017;2:225–230.
71. Thomas JS. A methodology for linking customer acquisition to customer retention. *J Market Res*. 2001;38:262–268.
72. Coussement K, Van den Poel D. Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Syst Appl*. 2009;36:6127–6134.
73. Fayyad U, Piatetsky-Shapiro G, Smyth P. The kdd process for extracting useful knowledge from volumes of data. *Commun ACM*. 1996;39:27–34.
74. Cohen WW, Schapire RE, Singer Y. Learning to order things. *J Artif Intell Res*. 1999;10:243–270.
75. Zakrzewska D, Murlewski J. Clustering algorithms for bank customer segmentation. In: *5th International Conference on Intelligent Sys- Tems Design and Applications (ISDA'05)*. IEEE; 2005:197–202.
76. Churchill Jr GA, Surprenant C. An investigation into the determinants of customer satisfaction. *J Market Res*. 1982;19:491–504.
77. Herzig J, Feigenblat G, Shmueli-Scheuer M, Konopnicki D, Rafaeli A. Predicting customer satisfaction in customer support conversations in social media using affective features. In: *Proceedings of the 2016 Con- Ference on User Modeling Adaptation and Personalization*. 2016:115–119.
78. Zhou S, Qiao Z, Du Q, Wang GA, Fan W, Yan X. Measuring customer agility from online reviews using big data text analytics. *J Manag Inf Syst*. 2018;35:510–539.
79. Kitchens B, Dobolyi D, Li J, Abbasi A. Advanced customer analytics: strategic value through integration of relationship-oriented big data. *J Manag Inf Syst*. 2018;35:540–574.
80. Yeh IC, Lien Ch. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst Appl*. 2009;36:2473–2480.
81. Black F, Scholes M. The pricing of options and corporate liabilities. In: *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*. World Scientific; 2019:3–21.
82. Jarrow RA, Turnbull SM. Pricing derivatives on financial securities subject to credit risk. *J Finance*. 1995;50:53–85.
83. Sung TK, Chang N, Lee G. Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. *J Manag Inf Syst*. 1999;16:63–85.
84. Kim CN, McLeod Jr R. Expert, linear models, and nonlinear models of expert decision making in bankruptcy prediction: a lens model analysis. *J Manag Inf Syst*. 1999;16:189–206.
85. Bellovary JL, Giacomino DE, Akers MD. A review of bankruptcy prediction studies: 1930 to present. *J Financ Education*. 2007:1–42.
86. Altman EI. *A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy*. New York: Corporate Financial Distress; 1983.
87. Zmijewski ME. Methodological issues related to the estimation of financial distress prediction models. *J Account Res*. 1984;59:82.
88. Ohlson JA. Financial ratios and the probabilistic prediction of bankruptcy. *J Account Res*. 1980:109–131.
89. Reurink A. Financial fraud: a literature review. *J Econ Surv*. 2018;32:1292–1325.
90. Perols JL, Bowen RM, Zimmermann C, Samba B. Finding needles in a haystack: using data analytics to improve fraud prediction. *Account Rev*. 2017;92:221–245.
91. Ngai EW, Hu Y, Wong YH, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decis Support Syst*. 2011;50:559–569.
92. Bolton RJ, Hand DJ, et al. *Unsupervised Profiling Methods for Fraud Detection*. Credit scoring and credit control VII; 2001:235–255.

93. Lekha KC, Prakasam S. Data mining techniques in detecting and predicting cyber crimes in banking sector. In: *2017 International Con-Ference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. IEEE; 2017:1639–1643.
94. Gandotra E, Bansal D, Sofat S. *Computational techniques for pre- dicting cyber threats. Intelligent computing, communication and devices*. 2015:247–253.
95. Prabakaran S, Mitra S. Survey of analysis of crime detection techniques using data mining and machine learning. In: *Journal of Physics: Conference Series*. IOP Publishing.; 2018, 012046.
96. Fachkha C, Bou-Harb E, Debbabi M. Towards a forecasting model for distributed denial of service activities. In: *2013 IEEE 12th International Symposium on Network Computing and Applications*. IEEE; 2013:110–117.
97. Ho SM, Hancock JT, Booth C, Liu X. Computer-mediated deception: strategies revealed by language-action cues in spontaneous communication. *J Manag Inf Syst*. 2016;33:393–420.
98. Abbasi A, Zahedi F, Zeng D, Chen Y, Chen H, Nunamaker Jr JF. Enhancing predictive analytics for anti-phishing by exploiting website genre information. *J Manag Inf Syst*. 2015;31:109–157.
99. Das SR, et al. Text and context: language analytics in finance. *Found Trends® Finance*. 2014;8:145–261.
100. Kogan S, Levin D, Routledge BR, Sagi JS, Smith NA. Predicting risk from financial reports with regression. In: *Proceedings of Hu- Man Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2009:272–280.
101. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: a comparative study. *Decis Support Syst*. 2011;50:602–613.
102. Dhrymes PJ, Peristiani SC. A comparison of the forecasting performance of wefa and arima time series methods. *Int J Forecast*. 1988;4:81–101.
103. Krollner B, Vanstone BJ, Finnie GR. *Financial Time Series Forecasting with Machine Learning Techniques: A Survey*. ESANN; 2010.
104. Sudjianto A, Nair S, Yuan M, Zhang A, Kern D, Cela-Díaz F. Statistical methods for fighting financial crimes. *Technometrics*. 2010;52:5–19.
105. Z bikowski K. Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Syst Appl*. 2015;42:1797–1805.