AN ACADEMIC REVIEW: APPLICATIONS OF DATA MINING TECHNIQUES IN FINANCE INDUSTRY

[1]SWATI JADHAV, [2]HONGMEI HE, [3]KARL JENKINS

School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, United Kingdom
E-mail: [1]s.jadhav@cranfield.ac.uk, [2]h.he@cranfield.ac.uk, [3]k.w.jenkins@cranfield.ac.uk

Abstract- With the development of Internet techniques, data volumes are doubling every two years, faster than predicted by Moore's Law. Big Data Analytics becomes particularly important for enterprise business. Modern computational technologies will provide effective tools to help understand hugely accumulated data and leverage this information to get insights into the finance industry. In order to get actionable insights into the business, data has become most valuable asset of financial organisations, as there are no physical products in finance industry to manufacture. This is where data mining techniques come to their rescue by allowing access to the right information at the right time. These techniques are used by the finance industry in various areas such as fraud detection, intelligent forecasting, credit rating, loan management, customer profiling, money laundering, marketing and prediction of price movements to name a few. This work aims to survey the research on data mining techniques applied to the finance industry from 2010 to 2015.The review finds that Stock prediction and Credit rating have received most attention of researchers, compared to Loan prediction, Money Laundering and Time Series prediction. Due to the dynamics, uncertainty and variety of data, nonlinear mapping techniques have been deeply studied than linear techniques. Also it has been proved that hybrid methods are more accurate in prediction, closely followed by Neural Network technique.
This survey could provide a clue of applications of data mining techniques for finance industry, and a summary of methodologies for researchers in this area. Especially, it could provide a good vision of Data Mining Techniques in computational finance for beginners who want to work in the field of computational finance.

Keywords- Data mining, Computational finance, Credit rating, Loan prediction, Money laundering, Stocks prediction.

I. INTRODUCTION

Huge amounts of raw, quantifiable data have been generated by organisations or enterprises. In 2011, 1.8 zettabytes of information were created globally, and that is expected to double every year [1]. This huge amount of data, if properly managed, modelled, shared and transformed, can help extract information that can be used to help understand, power and improve the business processes. The backbones of financial industry are data and quantitative business. Organisations can use these data to gain advantages over their competitors. It means data-driven decision making could provide new opportunities to boost the business strategies.

Data mining, also known as Knowledge Discovery, is to extract interesting nontrivial, implicit, previously unknown and potentially useful information or patterns from data in large databases [2]. The importance of data mining has been demonstrated for more than two decades. Forward-looking companies are placing data at the centre of strategic decision making.

Data analysis is usually carried out in two phases: Discovery and search. The patterns extracted in a discovery phase can be used in a search phase. Data mining is not data reporting. True data mining has its purpose, and it is a statistical process with some business goals, to enable business analysts to uncover useful patterns in available data [3]. Also, data preprocessing is the key step in data mining, as the quality and completeness of data is of utmost importance for data mining algorithms. Effective data mining technologies, corresponding to available data for different purposes, are helpful in predicting outcomes for new cases using the patterns recognised from known cases.

Computational Finance is the application of computational techniques to finance, and it is also referred to as financial engineering or quantitative finance. Financial mathematics, stochastic, numerical mathematics and scientific computing are combined to solve the real problems in finance industry. Computational Finance is important for the business world when it comes to corporate strategic planning by giving insights into what could happen in the future if a strategy is implemented, and predicting the risks associated with financial instruments.

The financial industry is a data-driven business, since the data generated is reliable and of high quality [4]. Big Data has become the asset of banks, as they are resources for analysing credit quality, monitor fraud and reduce customer churn. With passing time, various prediction techniques have been developed, and many of them have been employed in finance industry, especially after 1960s, following a triggered corporate financial distress [5]-[7]. Many computational finance problems can be mapped to data mining problems, thus corresponding data mining technologies can be applied to solve these problems. The techniques for association, classification, clustering, regression problems in data mining, which have been investigated extensively in the area of computational finance, included Support Vector

Machines (SVMs), Artificial Neural Networks (ANNs), Bayesian Classifier, Decision Trees (DTs), and Genetic Algorithms(GAs).

In this paper, we survey totally about 200 papers, in the research on data mining technologies for computational finance, published in the years from 2010 to 2015, in respects of classic business issues in finance industry, such as Credit rating (Section 2), Load prediction (Section 3), Money laundering (Section 4), Stocks prediction (Section 5) and Time series (Section 6). Finally, Section 7 concludes the survey.

Many researchers investigate multiple techniques from Data Mining, Machine Learning, Statistics, and Econometrics to solve problems in finance. For example, a study investigated Neural Networks, Support Vector Regression and Regression techniques to solve the Credit Rating problem, and the winning technique is Neural Network. In this review work, when observing the distribution of techniques for Credit Rating, all the 3 techniques are accounted, and the Neural Network is counted as a winner.

II. CREDIT RATING

The word ‘credit’ means ‘buy now and pay later’; the word ‘scoring’ refers to ‘the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them, and ensure objective and consistent decisions [8]. The process (by financial institutions) of modelling creditworthiness is referred to as credit scoring’.

Credit evaluation is one of the vital processes in banks’ credit management decisions. The process performs collection, analysis and classification of different credit elements and variables, which determines the credit decisions.

In finance industry, credit rating is used to check credit worthiness of a person, an authority, corporations, non-profit organizations or even governments. Credit Rating is expressed as a letter grade (such as 'AAA' and 'AA' for high credit quality and 'A' and 'BBB' for medium credit quality etc.). Whereas Credit Rating tells about creditworthiness of a business, corporation or government, Credit Score is expressed in numerical form and often used for individuals. Credit rating carries out credit evaluation, and assigns a score to credit report which represents credit worth of that entity. A credit score between 700 and 850 is considered as good. Figure 1 shows the credit scoring distribution for the general American public [9]. Nevertheless, both credit rating and score are used by creditors to assess a borrower's prospect of repaying a debt.

Data mining techniques are used to build credit scoring models, to help banks to make decision of accepting or rejecting a client’s credit. Business success of banking industry depends strongly on the evaluation of credit risk of potential debtors.

Credit risk analysis is an important part of financial risk management. Credit rating indicates a relative level of credit risk and it is a main analytical approach for credit risk assessment.

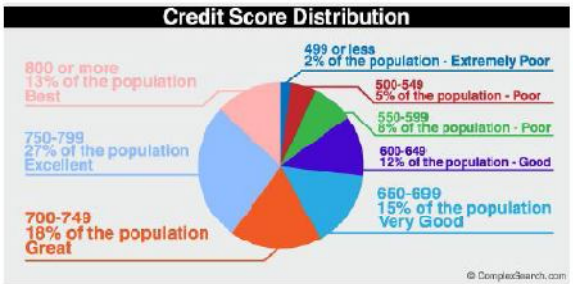

Figure 1. Credit Scoring Distribution.

A number of statistical based Data Mining techniques, such as SVMs, decision trees, neural networks, and k-nearest neighbours are used to construct credit scoring models. SVMs have been used widely for credit scoring [10]-[17]. Predictive models for credit card fraud detection are in active use in practice [18]. Research has shown that SVM is one of the most effective tools in credit risk evaluation. However, the performance of SVM is sensitive to the algorithms for the quadratic programming, to the parameters setting in its learning machines, and to the importance of different classes. Danenas et al. [19] used SVM based classifiers, and [16] claimed the weighted least squares SVM classifier achieved promising results. Similarly, to help assessment of the creditworthiness of loan applicants, Trustorff et al. [20] showed that the theoretical advantages of SVM classifiers can be used to improve the accuracy and the reliability of prediction of probabilities-of-default classification. Recently, for credit risk evaluation on larger databases, Danenas et al. [21] presented a technique for optimal linear SVM classifier selection based on particle swarm optimization technique. Also comparison with Logistic Regression and RBF networks was carried out which showed the proposed technique giving comparable results to the classic techniques.

Zhou et al. [14] used several SVM ensemble models to reduce inductive bias towards samples and parameter settings, shown by single SVM machines. In a similar study, Ghodselahi [22] used ensemble model of ten Support Vector Machine classifiers to improve the accuracy of classification for credit granting decisions. A hybrid ensemble model for credit risk combining both clustering and classification was designed. Ten SVM classifiers were the members of the ensemble model.

Hybrid methods, which combine different models to provide better performance than that can be gained by individual models, are becoming focus of lot of research work in finance market. An integrated approach using sampling to calculate F-score and SVM was proposed in [23]. A novel hybrid approach

for rectifying the data imbalance problem was proposed by employing k-Reverse Nearest Neighborhood and One Class support vector machine (OCSVM) in tandem in [24]. Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), Probabilistic Neural Network (PNN), Group Method of Data Handling (GMDH), Multi-Layer Perceptron (MLP) were used for testing. In a similar study for credit scoring [25], four approaches, including conventional statistical LDA and Decision tree, were combined with support vector machine classifier for feature selection to retain information. It was concluded that the hybrid credit scoring approach is mostly robust and effective in finding optimal subsets and is a promising method to the fields of data mining. The use of linear and quadratic discriminant analysis, logistic regression, multilayer perceptron, SVM, classification trees and ensemble methods [26] reduced misclassification up to 13.7%, compared with other classic models such as Linear and quadratic discriminant analysis, Logistic regression and Bagging. He et al. [27] used domain-driven multiple constraint-level programming-based classification outperforming the SVM, decision tree, and neural network in terms of sensitivity and Kolmogorov-Smirnov value while maintaining the tradeoff between time efficiency and acceptable accuracy and specificity. Whereas Kim et al. [28] used ordinal pairwise partitioning, cooperating with multi-class support vector machine for the credit rating, and it outperformed the multi-class support vector machine and other data mining techniques, such as Multiple Discriminant Analysis(MDA), Multinomial Logistic Regression(LOGIT), Case-based Reasoning(CBR) and ANN. Similarly, Huang [29] proposed a Gaussian-process-based multi-class classifier (GPC), which outperformed conventional multi-class classifiers and SVMs. Li et al. [30] proposed a vector machine based infinite decision agent ensemble learning system, in which, soft margin boosting was used to overcome overfitting, and the perceptron kernel was used to simulate infinite subagents. Other research in SVM-based hybrid methods can be found in [31]-[34]. Koutanaei et al. [35] developed a hybrid data mining model of feature selection and ensemble learning classification algorithms on the basis of three stages namely: Data gathering and pre-processing, followed by Employment of four Feature Selection algorithms of principal component analysis (PCA), genetic algorithm (GA), information gain ratio and relief attribute evaluation function. In the third stage, the classification results showed that the artificial neural network (ANN) adaptive boosting (AdaBoost) method had higher classification accuracy in credit scoring. Number of classifiers were compared in a study [36] such as logit/probit and LDA to fully nonlinear classifiers, including NNs, SVMs and more recent statistical learning techniques such as generalised boosting, AdaBoost and Random Forests (RFs). Study concluded that simpler classifiers such as Regression and LDA can be viable alternatives to more sophisticated approaches, particularly if interpretability is an important objective of the modelling exercise.

ANNs have been criticised for their 'black box' approach and interpretative difficulties but they are a very flexible family of models and are another well-known technology for credit scoring. A lot of efforts have been made onto the applications of neural networks in credit risk evaluation, and the back propagation learning algorithm is an efficient learning algorithm for training ANNs in the automatic processing of credit applications. Wah et al. [37] investigated three credit scoring models, logistic regression (LR) model, classification and regression tree (CART) model and neural network (ANN) model, to discriminate rejected and accepted credit card applicants of a bank. Results showed that the Neural Network model had a slightly higher validation predictive accuracy rate. Later, [38] proposed a hybrid system with genetic algorithm and artificial neural networks to find optimum feature subset to enhance the classification accuracy for retail credit risk assessment. Other research on ANN can be found in [39] and [40]. ANN and CART decision trees have shown that the forecast accuracy of credit rating process could be increased up to 96.5% [41].

For credit scoring, Marcano-Cedeno et al. [42] developed an ANN training algorithm, inspired by the neurons' biological property of meta-plasticity, which can be efficient when few patterns of a class are available, or when information inherent to low probability events is crucial. In situations where poor financial information was available, Falavigna [43] proposed a simulation model for assigning rating judgements to financial firms. It was the first tool able to forecast the default event two years before the bankruptcy.

Artificial neural networks, especially, MLPs were used [44] for credit scoring in microfinance industry, and it is shown that MLPs credit scoring can get higher accuracy in performance and lower misclassification costs than the classic Linear Discriminant analysis, Quadratic Discriminant analysis and Logistic Regression models. Self-Organizing Maps (SOM), another variant of ANN which is a clustering and unsupervised method is a common way for labeling the clusters and is called as Voted method. This method labels each cluster based on the majority class in it. The study [45] compared the capabilities of SOM that is labeled by Voted method and SOM that is labeled by a feedforward Neural Network in forecasting of credit classes. The comparison performed well in a commonly used benchmark, the Australian dataset.

In an effort to develop credit risk estimation models and to evaluate an influence of input data reduction on

credit risk models accuracy, Mileris and Boguslauskas [46] used ANNs and Logistic Regression(LR). The highest classification accuracy was shown by LR model followed by ANN. The Discriminant Analysis technique was the third accurate in classifying companies with credit risk.

The challenges of constructing credit scoring models lie in the availability of data and sample selection issues, and classification methods such as scorecards and decision trees are relatively easier to deploy in practical applications [47]. Also dual strategy ensemble trees can reduce the influence of the noise data and the redundant attributes of data to obtain the relative higher classification accuracy [48]. In cases of large class imbalance, the C4.5 decision tree algorithm, quadratic discriminant analysis and k-nearest neighbours perform significantly worse than Linear Discriminant Analysis and Gradient Boosting [49]. Bahnsen et al. [50] proposed an example-dependent cost-sensitive decision tree algorithm, by incorporating the different example-dependent costs into a new cost-based impurity measure and a new cost-based pruning criterion. Further the proposed method built significantly smaller trees in only a fifth of the time with a superior performance measured by cost savings. This could lead to a method with more business-oriented results creating simpler models that are easier to analyze. Zakrzewska [51] investigated a possibility of connecting unsupervised and supervised techniques for credit risk evaluation by building different rules for different groups of customers. Each credit applicant was assigned to the most similar group of clients from the training data set and credit risk is evaluated by applying the rules proper for this group. Results obtained with this technique of clustering and decision tree on the real credit risk data sets showed higher precisions and simplicity of rules obtained for each cluster than for rules connected with the whole data set. DT was again the subject of study in [52]. An ensemble approach based on merged decision trees, the correlated-adjusted decision forest (CADF) was introduced to produce both accurate and comprehensible credit risk models.

For credit rating, Tsai and Chen [53] developed four different hybrid models based on classification and clustering techniques providing highest prediction accuracy and lowest error rates in terms of credit rating, where a new technique, Grey Data Mining was introduced, based on Grey System's concept, Analytic Hierarchy Process technology and classical Data Mining technologies.

Rough sets have received less research attention in the area of credit scoring. [13] used rough sets with SVM to create a credit scoring classifier, outperforming linear discriminant analysis, logistic regression and neural networks. Integration of rough set, fuzzy set and probability theories was proposed in [54] for classifying credit risks. A basic parameter, representing the likelihood that a loan will not be repaid and will fall into default, was the inspiration behind this study. Chuang et al [55] developed a two stage hybrid model based on artificial neural networks and rough set theory for credit scoring. Some other research work can be found in [56]. Recently Shen and Tzeng [57] proposed an integrated hybrid soft computing model to resolve the financial prediction problem by adopting a dominance-based rough set approach to solve the financial performance prediction problem. Multiple criteria decision making and the influential weights of DANP (DEMATEL-based ANP) were used for further processing of core attributes along with the data from 2008 to 2011 from central bank of Taiwan for obtaining decision rules and forming an evaluation model. In the results, the proposed model showed that the top-ranking bank outperformed the other four banks.

Integrated approaches have been the favourite of researchers because they take advantage of different data mining techniques. Also in the cases where the training data and expert rules are insufficient and/or corrupted, mixed or hybrid approaches are required [58]. ANNs, SVMs and rough sets also have contributed well for the problem of credit rating. Bayesian networks and Fuzzy Apriori Genetic algorithms were explored in [59] and [36] respectively.

Ubiquitous Data Mining(UDM) classifier, which works in three steps such as: (1) constructing different models using datasets, (2) inducing rules from these models, and (3) consolidating these rules was used to predict credit ratings [60]. This study combined a number of classification models into one such that the performance of the consolidated model is better that of the original individual classification models in their classification accuracy and efficiency. Also the model was benchmarked against logistic regression (LR), Bayesian style frequency matrix (BFM), multilayer perceptron (MLP), classification tree methods (C5.0), and neural network rule extraction (NR) algorithms. Empirical results indicated that UDM outperformed all these single classifier models.

There exist numerous techniques and methods, but most of them depend on a particular data set or attribute set in question. In order to better apply these techniques and methods, getting insight of problems will help improve the performance of decision making or classification.

Figure 2 shows the distribution of surveyed techniques in the field of Credit Rating. Hybrid methods where usually multiple techniques are combined in stages, are most widely used techniques for predicting credit scores. The hybrid methods included combining SVM, ANN, DT, GA, Clustering, Regression, Rough sets and KNN. Many classic techniques were studied

in parallel to compare the performance. Such studies have been distributed in all the categories that were used. Modern classification techniques, despite being criticized as 'Black-box' techniques and being computationally expensive, often imply SVM and ANN. Many variants of SVMs, such as Least Square SVM, Multi-agent ensemble and Kernel-based, have been developed. MLP and RBF Neural Networks were often used. Obviously, complex non-linear techniques of ANNs and SVMs play significant roles for building credit scoring models in this period, totally accounting for 33% of studies. A traditional classification technique of regression is studied equally alongside SVM and ANN. Hybrid approaches account for 30%. Moreover, most hybrid approaches are the extension of SVMs and ANNs. Decision trees were studied in conjunction with SVM, Regression, ANN and Genetic Algorithms. A number of other techniques such as Gaussian process, Multiple Criteria Linear Programming-Optimization technique, Bayesian Network and Fuzzy logic have seen fewer studies.

Figure 3 depicts the number of wins in the usage of machine learning techniques for the purpose of prediction of credit ratings. Many studies reviewed in this paper have used multiple techniques for comparison purposes. Such studies are classified under all the techniques used. The winning technique is usually one, in some cases two giving comparable performance. Hence the number of wins is lesser that the number of techniques found in Distribution. In Figure 3, the hybrid methods have emerged as winning techniques more often compared to other techniques.

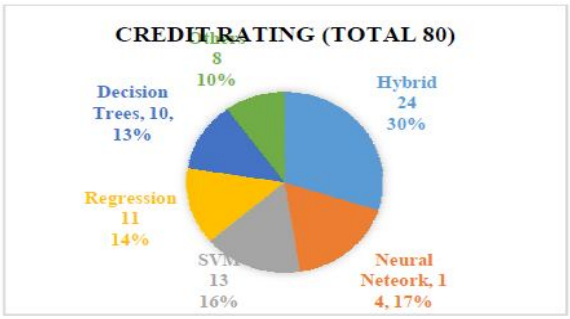

Figure 2. Distribution of surveyed techniques applied in Credit Rating.

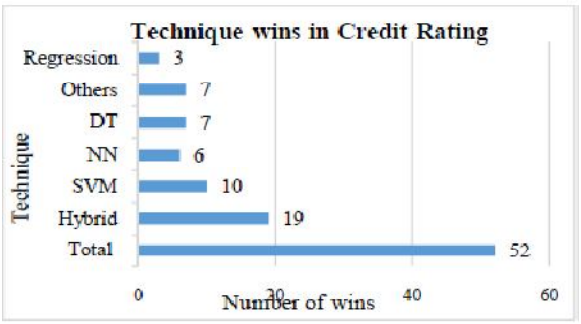

Figure 3. The numbers of wins for different techniques applied in Credit Rating.

In the area of Credit Rating, hybrid computational intelligence techniques have shown their superiorities over single techniques as can be seen from the number of wins above. By integrating the standard basic techniques into hybrid machine learning solutions, various intelligent searches and reasoning techniques have been developed to solve advanced and complex problems based on different domain knowledge.

III. LOAN PREDICTION

Banks and financial institutions rely on loan default prediction in credit risk management (Kou et al. 2014). They use data mining models to predict loan default before they decide to grant a loan with the goal of reducing defaults.

Chen and Chen [61] used the homogeneity between different city districts, the magnitude of the heterogeneity, and a prior distribution for the heterogeneity to formulate Loss Given a Default (LGD), which indicates the credit risk for a given default. Prediction of loan default has been studied widely by researchers. Reddy et al. [62] used attribute relevance analysis to predict loan default. This method excluded irrelevant attributes, thus reducing number of units for neural network model. The evaluation method of financing credit capacity was proposed in [63] combined with Kirkpatrick model and fuzzy neural network algorithm.

To predict loan risk well ahead of time using an imbalanced and large dataset, Srinivasan et al. [64] coupled Partial Least Squares Regression model and Variable Influence on Projection (VIP) scores to select the most important variables. This made the model less complex and computationally efficient, particularly for high risk loan records. RFs were used on large imbalanced data to predict loan defaults in [65]. The original RF algorithm was improved by allocating weights to decision trees in the forest during tree aggregation for prediction. The weights were easily calculated based on out-of-bag errors in training. To predict the performance of online peer-to-peer lending and classify the risk of loan into four categories, DT, ANNs and SVM were used in [66]. This study used RF for feature selection in the modeling phase and showed that the term of loan, annual income, the amount of loan, debt-to-income ratio, credit grade and revolving line utilization play an important role in loan defaults. But the prediction performance of SVM, Classification and Regression Tree (CART) and MLP were almost equal. But the study [67] found that DT enhanced with resampling techniques like AdaBoost performed better at enhancing the capabilities in classification than in prediction. Decision Trees are part of study in many studies in 2015 such as [68] for comparison between C4.5 and ID3 where C4.5 reached highest performance with data partition of 90%-10%; DT along with Fuzzy set theory where no preprocessing

or sampling was done for imbalanced datasets problem, outperforming C4.5 DT [69]; DT along with ANN and SVM with ANNs showing more accuracy than other classifiers [70]; comparative study of RF, SVM, LR and k-NN for identifying good borrowers in social lending with winner being RFs [71]; RFs giving better results than ID3 and C4.5 [72]; Fuzzy apriori combined with PCA creating a compact rule base and better results than the single fuzzy apriori model and other combined feature selection methods [73]; DT to build up a model to predict prospective business sectors in retail banking [74].

Cao et al. [75] developed a new model, particle swarm optimization combined with cost sensitive SVM to deal with the problem of unbalanced data classification and asymmetry misclassification cost in loan default discrimination problem. An extended tuning method for cost-sensitive regression and forecasting was suggested in [76] which was applied to loan charge-off forecasting on a real-world banking data. Logistic Regression used in [77] to estimate the probability of default for the customer credit in Vietnamese bank. Lasso Logistic Regression Ensemble was used in [78].

Many techniques from machine learning have received less attention such as KNN, Hidden Markov model, Loss Given Default, Association Rule mining, Wavelet ANN. In a study of dynamically monitoring loan service [79], the authors applied Hidden Markov Model to show that more accurate monitoring can be achieved by segmenting the defaulted data and training them separately. Chandra et al [80] presented a novel, hybrid soft computing system, SVWNN, based on integration of the sample-weighting SVM and wavelet neural network, to predict failure of banks. Support vectors along with their corresponding actual output labels were used to train the wavelet neural network (WNN). Further, Garson's algorithm for feature selection is adapted using WNN. Thus, the new hybrid, WNN-SVWNN, accomplishes horizontal and vertical reduction in the dataset as support vectors reduce the pattern space dimension and the WNN-based feature selection reduces the feature space dimension. To identify characteristic patterns of prospective lenders, Aribowo and Cahyana [81] used Association Rule Mining Classifier using Weighted Itemset Tidset tree. Another hybrid model [82] using ANN and SOMs outperformed the traditional Discriminant analysis and Logistic Regression, SVM and Random Forest classification models to predict bankruptcy. Hybrid models [83] made use of Association rule mining and process mining for fraud detection and the study in [84] used Particle Swarm Optimization and Support Vector Machines for bankruptcy prediction.

Loan prediction forms part of Credit risk analysis for the business of a financial institution. Data mining increases understanding by showing which factors should be included and which factors most affect specific outcomes [85].

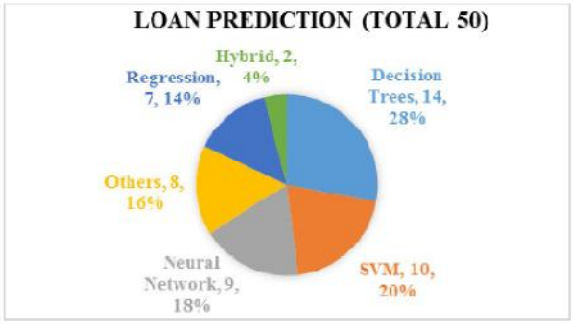

Figure 4. Distribution of surveyed techniques applied in Loan Prediction.

Figure 4 represents the use of data mining techniques applied for the problem of loan prediction from 2010 to 2015. Researchers have explored the effectiveness of Decision trees such as ID3, CART and C4.5 extensively. Hybrid techniques are still an emerging trend. The techniques like hybrid methods, KNN and Markov model have been investigated more. There is a lot of emphasis on complex, nonlinear supervised algorithms, such as SVM, ANNs as well as Regression. If we look at the winners among these studies, as shown in Figure 5, Decision trees and Neural Networks give better performance than other methods of SVM, Regression and Hybrid models, Markov model, Fuzzy set theory, KNN and Association Rule Mining.

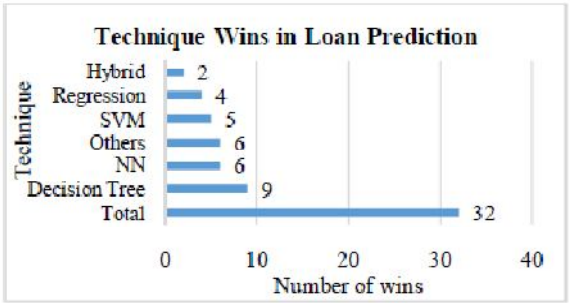

Figure 5. Numbers of wins for different techniques applied in Loan Prediction.

IV. MONEY LAUNDERING

Money laundering is the process whereby "dirty money" produced through criminal activity is transformed into "clean money", the criminal origin of which is difficult to trace. Financial Institutions are the most affected services for money laundering purposes [86]. Fraud is an extremely serious problem for credit-card companies. Visa and MasterCard lost over $700 million in 1995 from fraud. The cost of global payment card fraud grew by 19% last year to reach $14 billion. The cost of U.S. payment card fraud grew by 29% to $7.1 billion. In the rest of the world, card fraud grew by 11% to $6.8 billion [87]. Some of the well-known examples to support the importance of

data mining technology in financial institutions are: U. S. Treasury Department, Mellon Bank USA, Capital One Financial Group, American Express, MetLife Inc., Bank of America (USA) [88]. A system developed by the Financial Crimes Enforcement Network (FINCEN) of the U.S. Treasury Department called 'FAIS' detects potential money laundering activities from a large number of big cash transactions [88]. Mellon Bank has used the data on existing credit card customers to characterize their behavior to predict what they will do next, to predict which customers will stop using credit card in the next few months. Using data mining techniques, Capital One Financial Group tries to help market and sell the most appropriate financial product to 150 million potential prospects residing in its data warehouse. American Express uses data warehousing and data mining to cut spending and loan application screening. Metlife Inc. uses the "information - extraction" approach in which the input text is skimmed for specific information relevant to the particular application. Bank of America identified savings of $4.8 million in 2 years (a 400% return on investment) from use of a credit risk management system based on statistical and data mining analytics. They have also developed profiles of most valuable accounts in order to identify opportunities to sell them additional services. Recently, to retain deposits, they have used KXEN Analytic Framework in identifying clients likely to move assets and then creating offers conducive to retention [88]. As banking and payments have moved onto mobile and online channels, the opportunities for fraud have expanded. It is found that 50% of fraud is undetected until after the money has been lost. Patterns in data can be examined to identify chances of fraud occurrence, and prevention is possible.

Every financial institution is taking the responsibility of developing policies and procedures to fight money laundering. Money laundering regulation is seeing several transformations, thus financial institutions require establishing a well-defined plan against money laundering within their organizations.

In a seminal study [89] focused on financial fraud and money laundering, included reviews of classification methods such as SVM, Classification Trees and Ensemble Learning, Classification Rules and Rule Ensembles, Neural Networks, Bayesian Belief Networks, Hidden Markov Models.

In an interesting study [90] using Neural Networks in fraud detection, the auditors could use ANNs as complementary to other techniques at the planning stage of their audit to predict if a particular audit client was likely to have been victimized by a fraudster. Lot of studies have combined use of two or more classic techniques. Data mining techniques, such as clustering, neural networks, genetic algorithms, were applied for the cause of anti-money laundering detection in [91], and heuristics. Also Liu et al. [92] built a core decision tree with clustering algorithm to identify abnormal transaction. Dreżewski et al. [93] developed a system along with data importer and analysing algorithms, such as clustering and frequent-patterns-mining algorithms. In a study to evaluate different clustering algorithms, Cai et al. [94] showed that density-based clustering does not suit financial dataset. Normalised centroid-based clustering with higher DI or lower DBI gives the best number of clusters to help understanding financial data classification. K- means clustering method with multi-level feed forward network in [95] stressed that published financial statement data contains falsification indicators and ANNs provided highest accuracy. Clustering based anti-money-laundering system in [96] showed that the definition of client profiles that are more tailored to the system's goal, with a database of a greater time span (1 year) and a more thorough exploration of the types of attributes available for the used algorithms, produced better results.

Neural networks and regression models have been used for fraud detection since the dot com bubble burst that caused the 2000 stock market crash. Ravisankar et al. [97] used Neural Network, SVM and Logistic regression to detect fraud in the financial statement of big companies. Whereas Perols [98] used Neural Networks, SVMs, logistic regression and C4.5 to compare the performance of popular statistical and machine learning models, Zhou and Kapoor [99] used Neural Networks, Bayesian Networks, regression, decision tree to detect financial statement fraud. Wei et al. [100] presented an algorithm to mine contrast patterns along with neural network and decision forest to distinguish fraudulent behaviour from genuine behaviour.

Among the ANN and SVM models, kernel-based SVM methods are found to be most robust and accurate. To address the non-stationary problem in financial forecasting, Qin et al. [101] presented a novel non-linear combination of multiple kernel learning (MKL) model, called Gompertz model, for time series. This model showed good results compared to original MKL and single SVM, but with heavy computation burden. In the early warning system model for financial risk detection [102], a plenty of quantitative dependent variables, including 31 risk profiles, 15 risk indicators, 2 early warning signals and 4 financial road maps were considered in decision tree, using the CHAID algorithm to recognize those companies that need improvement. To meet the demands of the dynamic nature of business operations, Sun et al. [103] constructed a new dynamic financial distress prediction model by integrating financial indicator selection (a sequential floating forward selection method), principal component analysis and back-propagation neural network, optimized by genetic algorithm. This model

performed better than static models. In a study of credit loan fraud detection, Choi et al. [104] used individual level utility of each customer instead of the mean-level utility for classification using Decision Tree, Bayesian network and their Bagging to predict the probability of each customer being a fraud. Grammar-based Multi-objective Genetic Programming with Statistical Selection Learning (GBMGP-SSL) which applies the concepts of multi-objective optimization, token competition, and ensemble learning for evolving classification rules, in [105] to identify fraudulent information was able to obtain better performance in classifying fraudulent firms than LR, ANN, SVM, Bayesian networks and DTs.

Since money laundering is a non-linear problem and is a noisy process, because no distinct boundaries between legitimate and fraud accounts exist, neural networks offer the best dynamic solutions which are capable of evolution over time in this problem domain [106]. Figure 6 also shows the fact that the common classification technique of ANN remains the most common learning models in the area of Money laundering and fraud detection. The category shown as 'Other' covers techniques such as Genetic Programming, Bayesian networks and kernel learning and these techniques are studied comparably with the individual baseline models of ANN and DT.

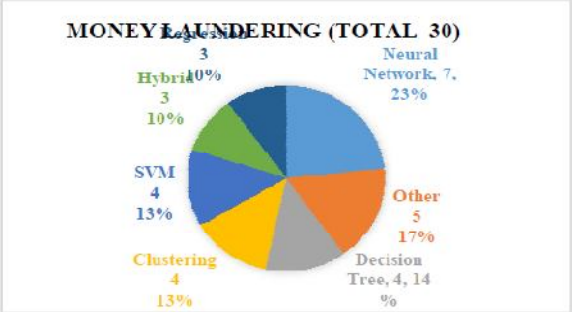

Figure 6. Distribution of surveyed techniques applied for Money Laundering.

Figure 7 below iterates the fact that ANNs and Other techniques discussed above have proved their accuracy in the field of detection of money laundering.

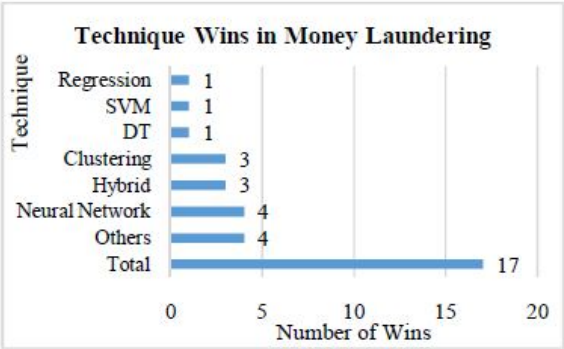

Figure 7. The Numbers of Wins for different Techniques applied in Money Laundering.

V. STOCKS PREDICTION

Stock prediction with the help of data mining techniques is vastly investigated by researchers in past decade. This application can be used by financial markets to make qualitative decisions. Prediction of stock market is not an easy task because of the fluctuations of the stock market. Many methodologies and models have been developed to predict the probability of profit-making in the stock market. This area has received the most attention from researchers.

In the past, stock price prediction models have been developed based on statistical or regression analysis tools. Some classic nonlinear models, such as Auto-regressive (AR), Vector Autoregressive (VAR) [107], Auto-regressive moving average (ARMA), auto-regressive integrated moving average (ARIMA), auto-regressive conditional heteroskedasticity (ARCH), and generalized ARCH (GARCH) [108], have been used to predict the values and trends of the stock market.

ANNs are a popular tool for financial forecasting because they are capable of dealing with very complex patterns. ANNs with high number of hidden layers and units can learn regularities having arbitrary complexity. ANNs have been investigated for studying stock markets in [109] and [110]. As fuzzy technology is a good approach to representing uncertainty of objects or concepts, recently, there were many studies focused on fuzzy techniques, combining with other data mining techniques to improve the robustness of system for stocks prediction. For example, Fuzzy ANN was explored in [111] and [112] to improve the stock market forecasting capability of the system. Recently a four layer BNNMAS(A bat-neural network multi-agent system) architecture for dealing with the distributed nature of stock prediction problem was proposed [113]. This multi-agent approach to create autonomous and independent subtasks to design an accurate prediction model used preprocessing methods in a parallel way such as data normalization, time lag selection and feature selection. Comparison with genetic algorithm neural network (GANN) and generalized regression neural network (GRNN) showed best MAPE statistics for the proposed model.

More recently, hybrid data mining technologies have been widely used to develop prediction models for stock price/index forecasting. Genetic Algorithm with ANN in wrapper is investigated by [114]. Araújo and Ferreira [115] worked on Morphological-Rank-Linear (MRL) filter combined with a Modified Genetic Algorithm (MGA). Cheng et al [116] used rough sets theory and GA, Huang [117] used GA and support vector regression; [118] used GA, decision tree, SVM; Qiu et al. [119] used C-fuzzy decision trees and k-nearest neighbors.

Evolutionary or Genetic Algorithms (GA) have been used routinely to generate useful solutions in a variety of search and optimisation problems. GAs are inspired by natural evolution, such as inheritance, mutation,

selection, and crossover. Since GAs can rapidly locate good solutions even for difficult search spaces, they have been applied in Stock market and other finance fields widely. Huang et al [120] used fuzzy-based GAs. Hsu [121] used self-organizing map and genetic programming. Associative classification is more accurate than a traditional classification approach but it is not good at handling numerical data and its relationships; which leads to an ongoing research problem of how to build associative classifiers from numerical data. Chien and Chen [122] proposed a highly competitive GA-based algorithm to build an associative classifier able to discover trading rules from these numerical indicators. Multi-gene Symbolic Regression genetic programming [123] evolving linear combinations of non-linear functions of the input variables was compared with traditional multiple linear regression model. This prediction model for the S&P 500 showed more robust results especially in the validation/testing case. Decision Forests again emerged as winners in the study in [124] followed by Support Vector Machines, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression. This study stressed the usage of ensemble methods in the field of stock price prediction.

Multivariate adaptive regression splines (MARS) is a multivariate, nonlinear, nonparametric regression approach. Because MARS itself has excellent variable selection capabilities, the difference between the degrees of significance for different variables can be analyzed, thus providing users with convenient data interpretation and higher user value [125]. Zarandi et al. [126] used MARS for stock price forecasting along with SVR and Adaptive Neural Fuzzy Inference Systems (ANFIS) on four different datasets. This technique was found to be more accurate than the other techniques in predicting all datasets. Recently, MARS along with fuzzy C-means has been researched more in other fields such as bankruptcy prediction [127] and [128].

Recently, SVM-based approaches have been investigated for the problems of stock price prediction. The following work can be found in this area:

Ding [129] applied SVR and compared with Ordinary Least Squares Regression, Back Propagation ANN, Radial Basis Function Networks (RBFN); Wen et al. [130] used SVM with box theory; Luo and Chen [131] integrated piecewise linear representation and weighted SVM. Kazem et al. [132] used SVR with chaos-based firefly algorithm. Yeh et al. [133] used multiple-kernel SVR approach. SVMs and traditional technical trading rules, such as Relative Strength Index (RSI) and the Moving Average Convergence Divergence (MACD) were studied in [134]. These rules were inputs to SVMs to determine the best situations to buy or sell the market. SVM itself can be modified with Fishers feature selection, Volume Weighted-SVM, input vector delays and technical indicators in combination with walk-forward optimization procedure successfully for the purpose of predicting short-term trends on the stock market [135]. But still, Least Squares Support Vector Machines (LS-SVMs) is a popular choice for classification in this area [136].

The sole use of a statistical model or a machine-learning method cannot adequately model all situations. Generally, it is more effective to combine different models that use different sources of information. Hybrid prediction models that combine some of these methods such as self-organizing maps (SOM), hidden Markov models (HMM), SVR, particle swarm optimization (PSO), Regression and simulated annealing (SA), have also been investigated for improving prediction accuracy and can be found in [137] and [138]. Liao and Chou [139] applied association rules and clustering to investigate the co-movement in the Taiwan and China (Hong Kong) stock markets. Chitrakar and Chuanhe [140] combined clustering algorithm *k*-Medoids with SVM and produced better performance in terms of Accuracy, Detection Rate and False Alarm Rate, compared to the combination of *k*-Medoids algorithm with Naïve Bayes classification. A two stage fusion approach involving SVR in the first stage and ANN, Random Forest (RF) and SVR in the second stage was proposed in [141] to address the problem of predicting future values of stock market indices. Experiments with single stage and two stage fusion prediction models showed that two stage hybrid models performed better than the single stage prediction models. The performance improvement is significant in case when ANN and RF are hybridized with SVR and moderate when SVR was hybridized with itself. The benefits of two stage prediction models over single stage prediction models become evident as the predictions are made for more number of days in advance. The best overall prediction performance was achieved by SVR–ANN model.

Hajek [142] applied several Prototype Generation Classifiers to predict the trend of the NASDAQ Composite index, and demonstrated that prototype generation classifiers were more accurate than SVMs and neural networks considering the buy or sell hit ratio of correctly predicted trend directions. Xiong et al. [143] proposed a swarm-based intelligent metaheuristic called firefly algorithm and multi-output SVR as a promising alternative for interval-valued financial time series forecasting problems and statistically outperforming in terms of the forecast accuracy.

Many application-oriented studies are being conducted. Tsai and Hsiao [144] applied multiple feature selection methods such as union, intersection and multi-intersection approaches to identify more representative variables for better prediction. Some researchers mined textual and contextual information from financial reports to predict stock price movement [145], [146]. Researches [147], [148] used sentiment classification and similarities in Candlestick charts for the task of stock prediction. Bollen et al. [149]

analyzed the text content of daily Twitter feeds by mood tracking tools along with use of Self-Organizing Fuzzy neural network and showed good results of predictions by considering specific public mood dimensions.

Use of statistics and other forecasting methods to predict stock prices was the focus of some research work. For example, [150] proposed a stock price prediction model able to extract data from time series data, news and comments on the news and to predict the stock price. They tested their model on numerical data only, yet missing text contents. Their results showed they were able to outperform other prediction methods such as SVR, Technical Analysis, Sentiment Analysis and Numerical Dynamics. The tone of the financial documents is significantly correlated with historical financial ratios such as profitability, liquidity, debt ratios, and stock price return. Hajek et al. [151] used sentiment information hidden in corporate annual reports successfully to predict short-run stock price returns with the application of several neural networks and ε-support vector regression models which performed better than linear regression models.

The nature of the stock market prediction problem requires the intelligent combination of several computing techniques rather than using them exclusively. More efforts have gone into developing hybrid methods, involving algorithms from basic data mining functions such as Association, Classification, Clustering and Regression along with methods from Statistics. These are closely followed by SVMs, ANNs and Regression techniques. Modern finance requires efficient ways to summarise and visualise the stock market data, and extract information from sentiments of market reports. The overall applicability of these methodologies and models still remains to be improved. Figure 8 below shows the contribution of different techniques in the field of stocks prediction.

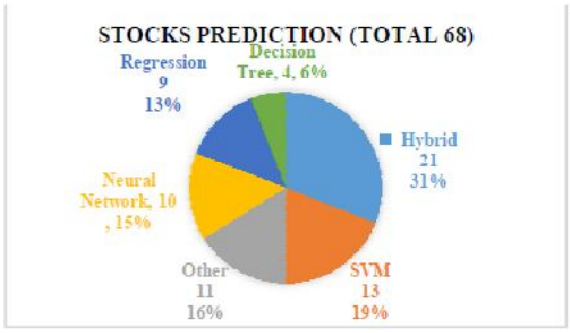

Figure 8. Distribution of surveyed techniques applied for Stocks Prediction.

The two stage methods of supervised learning, such as ANN, SVM, Regression and Decision Tree, which consist of first training and then predicting, have been used extensively in the area of stock market prediction. But the methods of hybrid machine learning have shown to be more accurate in prediction than single baseline methods. SVM has been receiving increasing interest in the areas ranging from Pattern recognition where it was originally applied, to Stocks prediction due to its remarkable generalization performance. This is proved from Figure 8 and Figure 9. The techniques of Genetic Algorithm, the econometrics model of Random Walk, Bayesian network, Naïve Bayes, Multiple Kernel learning, and Clustering are accounted as 'Other' techniques here, which follow the usage of SVM.

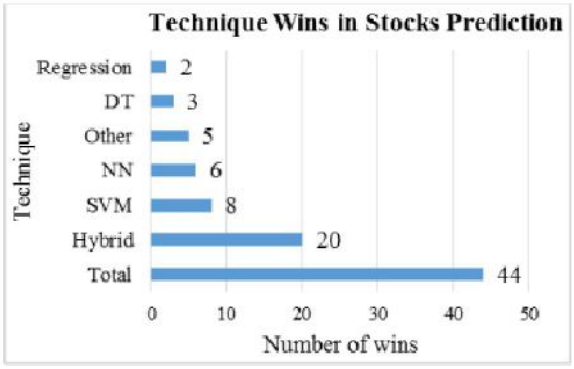

Figure 9. The number of wins for different methods in Stocks Prediction.

Irrespective of what combinations were used in hybrid learning studies, they have shown superiority over single models in the area of stock prediction.

VI. TIME SERIES

Time Series is a collection of organized data obtained from sequential measurements over a period of time. Time series data mining tries to extract all meaningful patterns from the shape of data. The goal of time series analysis is to predict the near future data based on past data. The relationship between the future value and previously observed values of variables is modeled in order to forecast the future values of a time series. Prediction of time series values is a distinct and extensive research area.

Modern time series forecasting involves many factors that are complexly correlated with each other, and hence it involves working with noisy, random, nonstationary and chaotic data. Time series forecasting is significant in the field of stock market investment since knowledge of the dynamic relationships among economic variables is essential for the investors to make right decisions at right times to maximize their financial profit and there is a strong dependency between future and past in finance market. According to [152], time series analysis may have one or more of the following objectives.

- Analysis and interpretation- find and interpret model to describe the time dependence in the data.
- Forecasting or prediction- given a sample from the series, forecast the next value, or the next few values.

- Control- adjust various control parameters to make the series fit closer to a target.
- Adjustment- in a linear model the errors could form a time series of correlated observations; adjust estimated variances to allow for this.
- Queries- moving averages, aggregates over time, year to year comparisons [153].
- Forecasting- forecasting of stock prices, sales and financial risk, estimate of credit worthiness, estimates of future financial outcomes for a company or country.

The research on time series prediction has been carried out by researchers from various communities such as machine learning, data mining, computer science, artificial intelligence, econometrics and statistics. This review work concentrates on the work carried out in Financial time series.

One of the most important static models for prediction, SVR has been applied in the prediction of financial time series with many characteristics of large sample sizes, noise, non-stationary, non-linearity, associated risk. Jiang and He [154] introduced local grey support vector regression (LG-SVR) and the use of grey relational grade as weighting function to adapt every test point in the time series locally and flexibly improved the performance of SVR. Classification is one of the main tasks in data mining. In the time series problem domain, special consideration is usually given due to the nature of the data. Sugimura and Matsumoto [155] proposed a system that acquires feature patterns and developed a classifier for time series data without using background knowledge given by a user. SVMs, which are good at better generalization of the training data, are studied in forecasting of financial time series aided by preprocessing and can be found in [156]. The key issue of deciding the parameters of the predicting model when using SVM was handled in this study by selecting the particle swarm optimization (PSO) method as the optimal tool to build a classifier, namely PSOSVM.

Hajek [142] demonstrated that the behaviour of stock price's movement can be effectively predicted using prototype generation classifiers. These methods are based on building new artificial prototypes from the training data set and improved the performance of nearest neighbour based classification.

Xiong et al. [157] proposed a swarm intelligence based FCRBFNN (fully complex-valued radial basis function neural networks) by using DPSO (discrete particle swarm optimization) and PSO (particle swarm optimization) for joint optimization of the structure and parameters of the model for interval time series forecasting. The proposed method improved prediction performance and statistically outperformed some well-established contenders like ARIMA and interval valued methods like Holt's linear trend method and MLP (Multi-Layer Perceptron applied to Interval-Valued Data) in terms of accuracy measure. Work in neural networks has concentrated on forecasting future values of the time series using current values. Some recent work on ANN can be found in [158], [159] for stochastic time effective ANNs, [160] for ANN compared with Auto-regression as well as in [161]-[165].

Heuristic approaches, based on ANN or Evolutionary Computation, have been shown to obtain some really good results in time series modelling and forecasting. Neural networks, offering flexible nonlinear modelling capability, can be adaptively generated through training with the features extracted from the data. Recently, ANNs have been used extensively in time series forecasting. Sometimes, classification problems can be transferred to optimisation problems to select best set of features to achieve the best classification performance. For example, feature selection was studied by [166] to present a novel CARTMAP neural network based on Adaptive Resonance Theory that incorporated automatic, intuitive, transparent and parsimonious feature selection with fast learning. Systematic ANN modeling processes and strategies for TSF were developed in [167]. Other ANN studies are: [168] and [169].

Saigal and Mehrotra [170] applied data mining techniques to financial time series data for calculating currency exchange rates of US dollars to Indian Rupees. It comprised of performance comparison of regression, vector autoregressive model and ANN on time series data in terms of the forecasting errors in accuracy generated by the models while predicting the currency exchange rates. The analysis was done using four Models: multiple regression in excel, multiple linear regression of dedicated time series analysis in Weka, vector autoregressive model in R and neural network model using Neural Works Predict.

Chen and Chen [171] proposed the fuzzy time series model based on the granular computing approach. It regulated the interval lengths during the iteration process using the entropy-based discretization method. The proposed model also used the binning-based partition to determine reasonable interval lengths by partitioning the universe of discourse and related linguistic values of each datum to change through repeated iterations. Another hybrid model Multiple Kernel Learning and Genetic Algorithm for Forecasting Short-Term Foreign Exchange Rates in [172] combined multiple kernel learning (MKL) for regression (MKR) and a genetic algorithm (GA) to construct the trading rules. More study of hybrid models is done in [173]. Some other techniques used in this field are Fuzzy set theory [174]; Regression, Fuzzy set theory, Ant Colony Optimization [175]; Fuzzy set theory, Clustering [176]; Regression [177], [178].

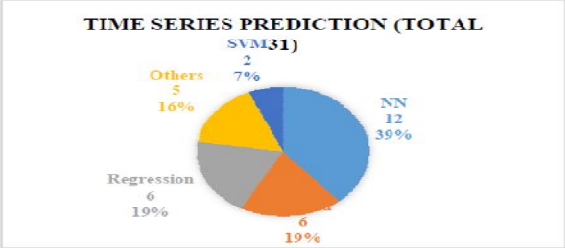

Figure 10. Distribution of surveyed techniques applied for Time series Prediction.

Time series data has such features that require more involvement by the data miner for preparation process than for non-series data [179]. Hence existing research in the field of time series prediction is inadequate and the problem of mining of sequence and time series data is considered as one of the challenging problems in the field of data mining. Dimensionality, representation, lack of well-established approaches in time series, handling of multivariate time series are some of the issues that still need to put much attention.

Figure 10 above shows that Regression and its variants having good explanation ability are widely used by researchers for time series prediction. Despite being time consuming in setting up, ANNs provide great classification and forecasting functionality. They are powerful with non-linear data, and hence are popular technique for time series forecasting. Hybrid machine learning models combine strengths of multiple knowledge representation model types and are researched quite often along with Regression. The category of 'Other' techniques, consisting of Clustering and Fuzzy set theory, has been investigated in a few studies but SVM has received less attention in financial time series prediction in the surveyed period since they are better suited for classification tasks.

Similar trend is seen in Figure 11 for the techniques emerging as winners in the area of time series prediction. ANNs prove to be the most winning techniques in predicting Time Series. Although ANNs can model both linear and non-linear structures of a time series, to handle both equally well, hybrid methods have proved successful and they closely follow usage of ANNs. Due to its simplicity and interpretability, Regression is still popular in time series forecasting and has emerged winner in five studies. Techniques such as Clustering and Fuzzy set theory have proved better at four studies in this period.

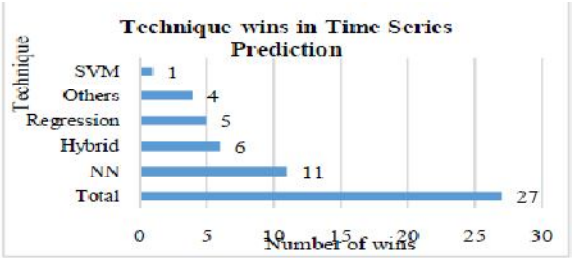

Figure 11. The number of wins for different methods in Time series Prediction.

SUMMARY

Financial data are being generated rapidly along with technological advances. Data mining techniques have been used to discover unknown patterns and predict future in financial markets and they have been investigated over a long period. Availability of vast data and potential significant benefits of solving finance problems have motivated extensive research in this field. Econometric modeling has always aimed to give empirical content to economic relations that deal only with real-life examples and data. But there is still room for us to develop more mature and efficient models to help the finance industry.

This work categorised the applications of data mining applications in finance industry in respects of Credit Rating, Loan Prediction, Money Laundering, Stocks Prediction and Time Series Prediction from 2010 to 2015. Many data mining techniques, including Regression, Neural Networks, Support Vector Machines, Decision Trees, Hybrid methods and some 'Other' techniques, e.g. Genetic Algorithms, have been applied in these areas. Different technologies have different performance for different finance problems. There is no doubt that data mining techniques play significant roles in finance industry.

The applications of data mining techniques in the fields of forecasting, pricing and prediction are related to various fields, such as Statistics, Econometrics, Artificial Intelligence and Machine learning. Regression, Clustering, Neural Networks, SVMs, Genetic Algorithms are the advanced techniques being popularly investigated in Computational Finance.

A lot of investment decisions in the area of finance industry are supported by data mining techniques, which help finance industry to understand the data and gain competitive advantage from the data. Hybrid models tend to evolve more and more. The synergy derived by machine learning, fuzzy logic, clustering and genetic algorithms have proved successful for hybrid methods. Nevertheless, the process should never ignore the importance of data quality, as the uncertainty of data always requires the robustness of the mining technology. .

The key findings of this study are:

(i) No single algorithm or technique works best across all types of datasets, problems. The choice is governed by the important aspects of dataset being used, the problem area, research objective, data preprocessing techniques involved, performance evaluation criteria, security, privacy and data integrity issues.

(ii) Hybrid models provide better and more accurate results, and hence are used more in the area of Credit Rating and Stock-market prediction. This proves the fact that unless subject to sufficiently rigorous tests entailed by hybrid techniques, a lot of the research could prove to be false.

(iii) Most investigation is for Credit Rating, followed by Stocks Prediction, Loan Prediction, Time Series Prediction and lastly Money Laundering.

(iv)What differentiates finance market is the world of unstructured data that is emerging as large source of actionable insights.

This survey could provide a clue of applications of data mining techniques for finance industry, and a summary of methodologies for researchers in this area. Especially, it could provide a good vision of Data Mining Techniques in computational finance for beginners who want to work in the field of computational finance.

REFERENCES

[1] J. Gantz and D. Reinsel, "Digital Universe," 2011. [Online]. Available: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[2] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[3] T. Khabaza, "Nine Laws of Data Mining," 2010. [Online]. Available: http://khabaza.codimension.net/index_files/9laws.htm.

[4] S. Sumathi and S. N. Sivanandam, *Introduction to data mining and its applications*, vol. 29. Springer, 2006.

[5] M. A. Aziz and H. A. Dar, "Predicting Corporate Financial Distress: Whither do We Stand?," *Dep. Econ. Loughbrgh. Univ.*, 2004.

[6] M. Hilston Keener, "Predicting The Financial Failure Of Retail Companies In The United States," *J. Bus. Econ. Res.*, vol. 11, no. 8, pp. 373–380, 2013.

[7] N. V Rao, G. Atmanathan, M. Shankar, and S. Ramesh, "Analysis of bankruptcy prediction models and their effectiveness: An Indian perspective," *Gt. Lakes Her.*, vol. 7, no. 2, 2013.

[8] R. Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, 2007.

[9] Complexsearch, "What is a Good Credit Score: 2016 Range, Credit Score Scale & Chart [Complete Guide]," 2016. [Online]. Available: http://www.complexsearch.com/what-is-a-good-credit-score/.

[10] F. L. Chen and F. C. Li, "Combination of feature selection approaches with SVM in credit scoring," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4902–4909, 2010.

[11] H. S. Kim and S. Y. Sohn, "Support vector machines for default prediction of SMEs based on technology credit," *Eur. J. Oper. Res.*, vol. 201, no. 3, pp. 838–846, 2010.

[12] A. B. Hens and M. K. Tiwari, "Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method," *Expert Syst. Appl.*, vol. 39, no. 8, p. 6774, 2012.

[13] Y. Ping and L. Yongheng, "Neighborhood rough set and SVM based hybrid credit scoring classifier," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11300–11304, 2011.

[14] L. Zhou, K. K. Lai, and L. Yu, "Least squares support vector machines ensemble models for credit scoring," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 127–133, 2010.

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

[16] L. Yu, X. Yao, S. Wang, and K. K. Lai, "Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 15392–15399, 2011.

[17] G. Wang and J. Ma, "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5325–5331, 2012.

[18] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, no. 3, p. 602, 2011.

[19] P. Danenas, G. Garsva, and S. Gudas, "Credit risk evaluation model development using support vector based classifiers," *Procedia Comput. Sci.*, vol. 4, pp. 1699–1707, 2011.

[20] J.-H. Trustorff, P. M. Konrad, and J. Leker, "Credit risk prediction using support vector machines," *Rev. Quant. Financ. Account.*, vol. 36, no. 4, pp. 565–581, 2011.

[21] P. Danenas and G. Garsva, "Selection of support vector machines based classifiers for credit risk domain," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3194–3204, 2015.

[22] A. Ghodselahi, "A hybrid support vector machine ensemble model for credit scoring," *Int. J. Comput. Appl.*, vol. 17, no. 5, pp. 1–5, 2011.

[23] A. B. Hens and M. K. Tiwari, "Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6774–6781, 2012.

[24] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 368–377, 2015.

[25] F.-L. Chen and F.-C. Li, "Combination of feature selection approaches with SVM in credit scoring," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4902–4909, 2010.

[26] M.-D. Cubiles-De-La-Vega, A. Blanco-Oliver, R. Pino-Mejías, and J. Lara-Rubio, "Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 6910–6917, 2013.

[27] J. He, Y. Zhang, Y. Shi, and G. Huang, "Domain-Driven Classification Based on Multiple Criteria and Multiple Constraint-Level Programming for Intelligent Credit Scoring," *Knowl. Data Eng. IEEE Trans.*, vol. 22, no. 6, pp. 826–838, 2010.

[28] K. J. Kim and H. Ahn, "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach," *Comput. Oper. Res.*, vol. 39, no. 8, pp. 1800–1811, 2012.

[29] S. C. Huang, "Using Gaussian process based kernel classifiers for credit rating forecasting," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8607–8611, 2011.

[30] S. Li, I. W. Tsang, and N. S. Chaudhari, "Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4947–4953, 2012.

[31] P. Hájek and V. Olej, "Credit rating modelling by kernel-based approaches with supervised and semi-supervised learning," *Neural Comput. Appl.*, vol. 20, no. 6, pp. 761–773, 2011.

[32] G. Wang and J. Ma, "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5325–5331, 2012.

[33] L. Yu, W. Yue, S. Wang, and K. K. Lai, "Support vector machine based multiagent ensemble learning for credit risk evaluation," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1351–1360, 2010.

[34] X. Zhou, W. Jiang, Y. Shi, and Y. Tian, "Credit risk evaluation with kernel-based affine subspace nearest points learning method," *Expert Syst. Appl.*, vol. 38, no. 4, p. 4272, 2011.

[35] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *J. Retail. Consum. Serv.*, vol. 27, pp. 11–23, 2015.

[36] S. Jones, D. Johnstone, and R. Wilson, "An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes," *J. Bank. Financ.*, vol.

56, pp. 72–85, 2015.
[37] Y. B. Wah and I. R. Ibrahim, "Using data mining predictive models to classify credit card applicants," in *Proc.- 6th Intl.Conference on Advanced Information Management and Service, IMS2010, with ICMIA2010 - 2nd International Conference on Data Mining and Intelligent Information Technology Applications*, 2010, pp. 394–398.
[38] S. Oreski, D. Oreski, and G. Oreski, "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment," *Expert Syst. Appl.*, vol. 39, no. 16, p. 12605, 2012.
[39] S. H. Ha and R. Krishnan, "Predicting repayment of the credit card debt," *Comput. Oper. Res.*, vol. 39, no. 4, pp. 765–773, 2012.
[40] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," *Expert Syst. Appl.*, vol. 37, no. 1, p. 534, 2010.
[41] S. C. Chen and M. Y. Huang, "Constructing credit auditing and control & management model with data mining technique," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5359–5365, 2011.
[42] A. Marcano-Cedeno, A. Marin-De-La-Barcena, J. Jimenez-Trillo, J. A. Pinuela, and D. Andina, "Artificial metaplasticity neural network applied to credit scoring," *Int. J. Neural Syst.*, vol. 21, no. 04, pp. 311–317, 2011.
[43] G. Falavigna, "Financial ratings with scarce information: A neural network approach," *Expert Syst. Appl.*, vol. 39, no. 2, p. 1784, 2012.
[44] A. Blanco, R. Pino-MejÃ-as, J. Lara, and S. Rayo, "Credit scoring models for the microfinance industry using neural networks: Evidence from Peru," *Expert Syst. Appl.*, vol. 40, no. 1, p. 356, 2013.
[45] A. AghaeiRad and B. Ribeiro, "Credit Prediction Using Transfer of Learning via Self-Organizing Maps to Neural Networks," in *Engineering Applications of Neural Networks*, Springer, 2015, pp. 358–365.
[46] R. Mileris and V. Boguslauskas, "Data Reduction Influence on the Accuracy of Credit Risk Estimation Models," *Eng. Econ.*, vol. 66, no. 1, 2015.
[47] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13274–13283, 2011.
[48] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," *Knowledge-Based Syst.*, vol. 26, pp. 61–68, 2012.
[49] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012.
[50] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6609–6619, 2015.
[51] D. Zakrzewska, "On integrating unsupervised and supervised classification for credit risk evaluation," *Inf. Technol. Control*, vol. 36, no. 1, 2015.
[52] R. Florez-Lopez and J. M. Ramon-Jeronimo, "Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5737–5753, 2015.
[53] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," *Appl. Soft Comput.*, vol. 10, no. 2, pp. 374–380, 2010.
[54] A. Capotorti and E. Barbanera, "Credit scoring analysis using a fuzzy probabilistic rough set model," *Comput. Stat. Data Anal.*, vol. 56, no. 4, p. 981, 2012.
[55] C.-L. Chuang and S.-T. Huang, "A hybrid neural network approach for credit scoring," *Expert Syst.*, vol. 28, no. 2, pp. 185–196, 2011.
[56] C.-C. Yeh, F. Lin, and C.-Y. Hsu, "A hybrid KMV model, random forests and rough set theory approach for credit rating," *Knowledge-Based Syst.*, vol. 33, pp. 166–172, 2012.
[57] K.-Y. Shen and G.-H. Tzeng, "A decision rule-based soft computing model for supporting financial performance improvement of the banking industry," *Soft Comput.*, vol. 19, no. 4, pp. 859–874, 2015.
[58] B. Kovalerchuk and E. Vityaev, *Data mining in finance: advances in relational and hybrid methods*. Springer Science & Business Media, 2000.
[59] C. K. Leong, "Credit risk scoring with bayesian network models," *Comput. Econ.*, pp. 1–24, 2015.
[60] J. K. Bae and J. Kim, "A Personal Credit Rating Prediction Model Using Data Mining in Smart Ubiquitous Environments," *Int. J. Distrib. Sens. Networks*, vol. 2015, 2015.
[61] T.-H. Chen and C.-W. Chen, "Application of data mining to the spatial heterogeneity of foreclosed mortgages," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 993–997, 2010.
[62] M. V Jagannatha Reddy and B. Kavitha, "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis," in *Signal Acquisition and Processing, 2010. ICSAP '10. International Conference on*, 2010, pp. 274–277.
[63] Y. Zhang, "A Novel FNN Algorithm and Its Application in FCC Evaluation Based on Kirkpatrick Model," in *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010, pp. 447–450.
[64] B. V Srinivasan, N. Gnanasambandam, S. Zhao, and R. Minhas, "Domain-Specific Adaptation of a Partial Least Squares Regression Model for Loan Defaults Prediction," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 474–479.
[65] L. Zhou and H. Wang, "Loan default prediction on large imbalanced data using random forests," *TELKOMNIKA Indones. J. Electr. Eng.*, vol. 10, no. 6, pp. 1519–1525, 2012.
[66] Y. Jin and Y. Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," in *Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference on*, 2015, pp. 609–613.
[67] R. S. Oetama, "Enhancing Decision Tree Performance in Credit Risk Classification and Prediction," *Ultim. J. Tek. Inform.*, vol. 7, no. 1, p. 50, 2015.
[68] R. K. Amin, Y. Sibaroni, and others, "Implementation of decision tree using C4. 5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)," in *Information and Communication Technology (ICoICT), 2015 3rd International Conference on*, 2015, pp. 75–80.
[69] J. A. Sanz, D. Bernardo, F. Herrera, H. Bustince, and H. Hagras, "A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data," *Fuzzy Syst. IEEE Trans.*, vol. 23, no. 4, pp. 973–990, 2015.
[70] R. Geng, I. Bose, and X. Chen, "Prediction of financial distress: An empirical study of listed Chinese companies using data mining," *Eur. J. Oper. Res.*, vol. 241, no. 1, pp. 236–247, 2015.
[71] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Syst. Appl.*, vol. 42, no. 10, pp. 4621–4631, 2015.
[72] S. Sathyadevan and R. R. Nair, "Comparative Analysis of Decision Tree Algorithms: ID3, C4. 5 and Random Forest," in *Computational Intelligence in Data Mining-Volume 1*, Springer, 2015, pp. 549–562.
[73] S. Sadatrasoul, M. Gholamian, and K. Shahanaghi, "Combination of feature selection and optimized fuzzy apriori rules: The case of credit scoring," *Int. Arab J. Inf. Technol.*, vol. 12, no. 2, pp. 138–145, 2015.
[74] M. Islam, M. Habib, and others, "A data mining approach to predict prospective business sectors for lending in retail banking using decision tree," *arXiv Prepr. arXiv1504.02018*, 2015.
[75] J. Cao, H. Lu, W. Wang, and J. Wang, "A loan default

discrimination model using cost-sensitive support vector machine improved by PSO," *Inf. Technol. Manag.*, vol. 14, no. 3, pp. 193–204, 2013.
[76] H. Zhao, A. P. Sinha, and G. Bansal, "An extended tuning method for cost-sensitive regression and forecasting," *Decis. Support Syst.*, vol. 51, no. 3, pp. 372–383, 2011.
[77] T. Duong, V. Tran, and Q. Ho, "A Proposed Credit Scoring Model for Loan Default Arobability: a Vietnamese bank case," in *International Conference on Qualitative and Quantitative Economics Research (QQE). Proceedings*, 2015, p. 52.
[78] H. Wang, Q. Xu, and L. Zhou, "Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble," *PLoS One*, vol. 10, no. 2, p. e0117844, 2015.
[79] H. Lee, N. Gnanasambandam, R. Minhas, and S. Zhao, "Dynamic Loan Service Monitoring Using Segmented Hidden Markov Models," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, 2011, pp. 749–754.
[80] D. K. Chandra, V. Ravi, and P. Ravisankar, "Support vector machine and wavelet neural network hybrid: application to bankruptcy prediction in banks," *Int. J. Data Mining, Model. Manag.*, vol. 2, no. 1, pp. 1–21, 2010.
[81] A. S. Aribowo and N. H. Cahyana, "Feasibility study for banking loan using association rule mining classifier," *Int. J. Adv. Intell. Informatics*, vol. 1, no. 1, pp. 41–47, 2015.
[82] F. J. L. Iturriaga and I. P. Sanz, "Bankruptcy visualization and prediction using neural networks: A study of US commercial banks," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2857–2869, 2015.
[83] R. Sarno, R. D. Dewandono, T. Ahmad, M. F. Naufal, and F. Sinaga, "Hybrid association rule learning and process mining for fraud detection," *IAENG Int. J. Comput. Sci.*, vol. 42, no. 2, pp. 59–72, 2015.
[84] Y. Lu, N. Zeng, X. Liu, and S. Yi, "A new hybrid algorithm for bankruptcy prediction using switching particle swarm optimization and support vector machines," *Discret. Dyn. Nat. Soc.*, vol. 501, p. 294930, 2015.
[85] R. Gerritsen, "Assessing loan risks: a data mining case study," *IT Prof.*, vol. 1, no. 6, pp. 16–21, 1999.
[86] F. Typologies and T. Reports, "Money Laundering and Terrorist Financing Trends in FINTRAC Cases Disclosed between 2007 and 2011 FINTRAC Typologies and Trends Reports – April 2012." 2012.
[87] J. Heggestuen, "The US Sees More Money Lost To Credit Card Fraud Than The Rest Of The World Combined," 2014. [Online]. Available: http://www.techinsider.io/the-us-accounts-for-over-half-of-global-payment-card-fraud-sai-2014-3.
[88] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.
[89] A. Sudjianto, S. Nair, M. Yuan, A. Zhang, D. Kern, and F. Cela-Díaz, "Statistical Methods for Fighting Financial Crimes," *Technometrics*, vol. 52, no. 1, pp. 5–19, 2010.
[90] M. Krambia-Kapardis, C. Christodoulou, and M. Agathocleous, "Neural networks: the panacea in fraud detection?," *Manag. Audit. J.*, vol. 25, pp. 659–678, 2010.
[91] N. A. Le Khac and M. Kechadi, "Application of Data Mining for Anti-money Laundering Detection: A Case Study," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010, pp. 577–584.
[92] R. Liu, X. Qian, S. Mao, and S. Zhu, "Research on anti-money laundering based on core decision tree algorithm," in *Control and Decision Conference (CCDC), 2011 Chinese*, 2011, pp. 4322–4325.
[93] R. Dreżewski, J. Sepielak, and W. Filipkowski, "System supporting money laundering detection," *Digit. Investig.*, vol. 9, no. 1, pp. 8–21, 2012.
[94] F. Cai, N.-A. Le-Khac, and M.-T. Kechadi, "Clustering approaches for financial data analysis: a survey," in *Proceedings of the International Conference on Data Mining (DMIN)*, 2012, p. 1.
[95] K. K. Tangod and G. H. Kulkarni, "Detection of Financial Statement Fraud using Data Mining Technique and Performance Analysis," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 549–555, 2015.
[96] C. Alexandre and J. Balsa, "Client Profiling for an Anti-Money Laundering System," *arXiv Prepr. arXiv1510.00878*, 2015.
[97] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *Decis. Support Syst.*, vol. 50, no. 2, pp. 491–500, 2011.
[98] J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms," *Audit. A J. Pract. Theory*, vol. 30, no. 2, pp. 19–50, 2011.
[99] W. Zhou and G. Kapoor, "Detecting evolutionary financial statement fraud," *Decis. Support Syst.*, vol. 50, no. 3, pp. 570–575, 2011.
[100] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, no. 4, pp. 449–475, 2013.
[101] H. Qin, D. Dou, and Y. Fang, "Financial Forecasting with Gompertz Multiple Kernel Learning," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 2010, pp. 983–988.
[102] A. S. Koyuncugil and N. Ozgulbas, "Financial early warning system model and data mining application for risk detection," *Expert Syst. Appl.*, vol. 39, no. 6, pp. 6238–6253, 2012.
[103] J. Sun, K.-Y. He, and H. Li, "SFFS-PC-NN optimized by genetic algorithm for dynamic prediction of financial distress with longitudinal data streams," *Knowledge-Based Syst.*, vol. 24, no. 7, pp. 1013–1023, 2011.
[104] K. Choi, G. Kim, and Y. Suh, "Classification model for detecting and managing credit loan fraud based on individual-level utility concept," *ACM SIGMIS Database*, vol. 44, no. 3, pp. 49–67, 2013.
[105] H. Li and M.-L. Wong, "Financial fraud detection by using Grammar-based multi-objective genetic programming with ensemble learning," in *Evolutionary Computation (CEC), 2015 IEEE Congress on*, 2015, pp. 1113–1120.
[106] P. J. G. Lisboa, A. Vellido, and B. Edisbury, *Business Applications of Neural Networks: The State-of-the-Art of Real-World Applications*, vol. 13. World scientific, 2000.
[107] S. Kumar, S. Managi, and A. Matsuda, "Stock prices of clean energy firms, oil and carbon markets: A vector autoregressive analysis," *Energy Econ.*, vol. 34, no. 1, pp. 215–226, 2012.
[108] H. Zhao, "Dynamic relationship between exchange rate and stock price: Evidence from China," *Res. Int. Bus. Financ.*, vol. 24, no. 2, pp. 103–112, 2010.
[109] A. A. Adebiyi, C. K. Ayo, M. O. Adebiyi, and S. O. Otokiti, "Stock Price Prediction using Neural Network with Hybridized Market Indicators," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 3, no. 1, pp. 1–9, 2012.
[110] M. M. Mostafa, "Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6302–6309, 2010.
[111] A. Esfahanipour and W. Aghamiri, "Adapted Neuro-Fuzzy Inference System on indirect approach TSK fuzzy rule base for stock market analysis," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4742–4748, 2010.
[112] C.-F. Liu, C.-Y. Yeh, and S.-J. Lee, "Application of type-2 neuro-fuzzy modeling in stock price prediction," *Appl. Soft Comput.*, vol. 12, no. 4, pp. 1348–1358, 2012.
[113] R. Hafezi, J. Shahrabi, and E. Hadavandi, "A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price," *Appl. Soft Comput.*, vol. 29, pp. 196–210, 2015.
[114] B. B. Nair, S. G. Sai, A. N. Naveen, A. Lakshmi, G. S. Venkatesh, and V. P. Mohandas, "A GA-artificial neural network hybrid system for financial time series forecasting," in *Information Technology and Mobile Communication*, Springer, 2011, pp. 499–506.
[115] R. de A. Araújo and T. A. E. Ferreira, "A morphological-rank-linear evolutionary method for stock market prediction," *Inf. Sci. (Ny).*, vol. 237, pp. 3–17, 2013.
[116] C.-H. Cheng, T.-L. Chen, and L.-Y. Wei, "A hybrid model

based on rough sets theory and genetic algorithms for stock price forecasting," *Inf. Sci. (Ny).*, vol. 180, no. 9, pp. 1610–1629, 2010.

[117] C.-F. Huang, "A hybrid stock selection model using genetic algorithms and support vector regression," *Appl. Soft Comput.*, vol. 12, no. 2, p. 807, 2012.

[118] B. B. Nair, V. P. Mohandas, and N. R. Sakthivel, "A genetic algorithm optimized decision tree-SVM based stock market trend prediction system," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 9, pp. 2981–2988, 2010.

[119] W. Qiu, X. Liu, and L. Wang, "Forecasting shanghai composite index based on fuzzy time series and improved C-fuzzy decision trees," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7680–7689, 2012.

[120] C.-F. Huang, B. R. Chang, D.-W. Cheng, and C.-H. Chang, "Feature selection and parameter optimization of a fuzzy-based stock selection model using genetic algorithms," *Int. J. Fuzzy Syst.*, vol. 14, no. 1, pp. 65–75, 2012.

[121] C.-M. Hsu, "A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming," *Expert Syst. Appl.*, vol. 38, no. 11, pp. 14026–14036, 2011.

[122] Y.-W. Chang Chien and Y.-L. Chen, "Mining associative classification rules with stock trading data–A GA-based method," *Knowledge-Based Syst.*, vol. 23, no. 6, pp. 605–614, 2010.

[123] A. Sheta, S. E. M. Ahmed, and H. Faris, "Evolving stock market prediction models using multi-gene symbolic regression genetic programming," *Artif. Intell. Mach. Learn. AIML*, vol. 15, pp. 11–20, 2015.

[124] M. Ballings, D. den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, 2015.

[125] C.-J. Lu, "Sales forecasting of computer products based on variable selection scheme and support vector regression," *Neurocomputing*, vol. 128, pp. 491–499, 2014.

[126] M. H. Zarandi, M. Zarinbal, N. Ghanbari, and I. B. Turksen, "A new fuzzy functions model tuned by hybridizing imperialist competitive algorithm and simulated annealing. Application: Stock price prediction," *Inf. Sci. (Ny).*, vol. 222, pp. 213–228, 2013.

[127] J. De Andrés, P. Lorca, F. J. de Cos Juez, and F. Sánchez-Lasheras, "Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS)," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1866–1875, 2011.

[128] A. Martin, V. Gayathri, G. Saranya, P. Gayathri, and P. Venkatesan, "A hybrid model for bankruptcy Prediction using genetic Algorithm, fuzzy c-means and mars," *arXiv Prepr. arXiv1103.2110*, 2011.

[129] Z. Ding, "Application of support vector machine regression in stock price forecasting," in *Business, Economics, Financial Sciences, and Management*, Springer, 2012, pp. 359–365.

[130] Q. Wen, Z. Yang, Y. Song, and P. Jia, "Automatic stock decision support system based on box theory and SVM algorithm," *Expert Syst. Appl.*, vol. 37, no. 2, pp. 1015–1022, 2010.

[131] L. Luo and X. Chen, "Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction," *Appl. Soft Comput.*, vol. 13, no. 2, pp. 806–816, 2013.

[132] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting," *Appl. Soft Comput.*, vol. 13, no. 2, pp. 947–958, 2013.

[133] C.-Y. Yeh, C.-W. Huang, and S.-J. Lee, "A multiple-kernel support vector regression approach for stock market price forecasting," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2177–2186, 2011.

[134] C. L. Dunis, R. Rosillo, D. de la Fuente, and R. Pino, "Forecasting IBEX-35 moves using support vector machines," *Neural Comput. Appl.*, vol. 23, no. 1, pp. 229–236, 2013.

[135] K. Zbikowski, "Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1797–1805, 2015.

[136] I. Marković, M. Stojanović, M. Božić, and J. Stanković, "Stock market trend prediction based on the LS-SVM model update algorithm," in *ICT Innovations 2014*, Springer, 2015, pp. 105–114.

[137] J.-J. Wang, J.-Z. Wang, Z.-G. Zhang, and S.-P. Guo, "Stock index forecasting based on a hybrid model," *Omega*, vol. 40, no. 6, pp. 758–766, 2012.

[138] C.-F. Huang, C.-H. Chang, B. R. Chang, and D.-W. Cheng, "A study of a hybrid evolutionary fuzzy model for stock selection," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, 2011, pp. 210–217.

[139] S.-H. Liao and S.-Y. Chou, "Data mining investigation of co-movements on the Taiwan and China stock markets for future investment portfolio," *Expert Syst. Appl.*, vol. 40, no. 5, p. 1542, 2013.

[140] R. Chitrakar and H. Chuanhe, "Anomaly detection using Support Vector Machine classification with k-Medoids clustering," in *Internet (AH-ICI), 2012 Third Asian Himalayas International Conference on*, 2012, pp. 1–5.

[141] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2162–2172, 2015.

[142] P. Hajek, "Forecasting Stock Market Trend using Prototype Generation Classifiers," *WSEAS Trans. Syst.*, vol. 11, no. 12, pp. 671–680, 2012.

[143] T. Xiong, Y. Bao, and Z. Hu, "Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting," *Knowledge-Based Syst.*, vol. 55, pp. 87–100, 2014.

[144] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches," *Decis. Support Syst.*, vol. 50, no. 1, pp. 258–269, 2010.

[145] S. W. K. Chan and J. Franklin, "A text-based decision support system for financial sequence prediction," *Decis. Support Syst.*, vol. 52, no. 1, pp. 189–198, 2011.

[146] M. Hagenau, M. Liebmann, M. Hedwig, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-specific features," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2012, pp. 1040–1049.

[147] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Syst.*, vol. 41, pp. 89–97, 2013.

[148] C.-F. Tsai and Z.-Y. Quan, "Stock Prediction by Searching for Similarities in Candlestick Charts," *ACM Trans. Manag. Inf. Syst.*, vol. 5, no. 2, p. 9, 2014.

[149] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[150] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," in *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*, 2011, pp. 800–807.

[151] P. Hajek, V. Olej, and R. Myskova, "Forecasting Stock Prices using Sentiment Information in Annual Reports-A Neural Network and Support Vector Regression Approach," *WSEAS Trans. Syst. (in Press. 2013)*, 2013.

[152] G. REINERT, "Time Series," 2002. [Online]. Available: http://www.stats.ox.ac.uk/~reinert/time/notesht10short.pdf.

[153] D. Shasha, "Time series in finance: the array database approach," *ACM SIGM0D, Abril*, 1999.

[154] H. Jiang and W. He, "Grey relational grade in local support vector regression for financial time series prediction," *Expert Syst. Appl.*, vol. 39, no. 3, p. 2256, 2012.

[155]H. Sugimura and K. Matsumoto, "Classification system for time series data based on feature pattern extraction," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, 2011, pp. 1340–1345.
[156]G. Zhiqiang, W. Huaiqing, and L. Quan, "Financial time series forecasting using LPP and SVM optimized by PSO," *Soft Comput.*, vol. 17, no. 5, pp. 805–818, 2013.
[157]T. Xiong, Y. Bao, Z. Hu, and R. Chiong, "Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms," *Inf. Sci. (Ny).*, vol. 305, pp. 77–92, 2015.
[158]L. A. Laboissiere, R. A. S. Fernandes, and G. G. Lage, "Maximum and minimum stock price forecasting of Brazilian power distribution companies based on artificial neural networks," *Appl. Soft Comput.*, vol. 35, pp. 66–74, 2015.
[159]J. Wang and J. Wang, "Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks," *Neurocomputing*, vol. 156, pp. 68–78, 2015.
[160]A. B. Kock and T. Teräsvirta, "Forecasting macroeconomic variables using neural network models and three automated model selection techniques," *Econom. Rev.*, pp. 1–27, 2015.
[161]T. D. Chaudhuri and I. Ghosh, "Forecasting Volatility in Indian Stock Market using Artificial Neural Network with Multiple Inputs and Outputs," *Int. J. Comput. Appl.*, vol. 120, no. 8, 2015.
[162]Y. Xiao, J. J. Liu, S. Wang, Y. Hu, and J. Xiao, "Multiple dimensioned mining of financial fluctuation through radial basis function networks," *Neural Comput. Appl.*, vol. 26, no. 2, pp. 363–371, 2015.
[163]H. Al--askar, D. Lamb, A. J. Hussain, D. Al--Jumeily, M. Randles, and P. Fergus, "Predicting financial time series data using artificial immune system--inspired neural networks," *Int. J. Artif. Intell. Soft Comput.*, vol. 5, no. 1, pp. 45–68, 2015.
[164]D. I. Vortelinos, "Forecasting realized volatility: HAR against Principal Components Combining, neural networks and GARCH," *Res. Int. Bus. Financ.*, 2015.
[165]Z. Chengzhao, P. Heiping, and Z. Ke, "Comparison of Back Propagation Neural Networks and EMD-Based Neural Networks in Forecasting the Three Major Asian Stock Markets," *J. Appl. Sci.*, vol. 15, no. 1, p. 90, 2015.
[166]C. Wong and M. Versace, "CARTMAP: a neural network method for automated feature selection in financial time series forecasting," *Neural Comput. Appl.*, vol. 21, no. 5, pp. 969–977, 2012.
[167]W. Yan, "Toward Automatic Time-Series Forecasting Using Neural Networks," *Neural Networks Learn. Syst. IEEE Trans.*, vol. 23, no. 7, pp. 1028–1039, 2012.
[168]M. Khashei and M. Bijari, "Fuzzy artificial neural network (p, d, q) model for incomplete financial time series forecasting," *J. Intell. Fuzzy Syst.*, vol. 26, no. 2, pp. 831–845, 2014.
[169]H. Niu and J. Wang, "Financial time series prediction by a random data-time effective RBF neural network," *Soft Comput.*, vol. 18, no. 3, pp. 497–508, 2014.
[170]S. Saigal and D. Mehrotra, "Performance comparison of time series data using predictive data mining techniques," *Adv. Inf. Min.*, vol. 4, no. 1, pp. 57–66, 2012.
[171]M.-Y. Chen and B.-T. Chen, "A hybrid fuzzy time series model based on granular computing for stock price forecasting," *Inf. Sci. (Ny).*, vol. 294, pp. 227–241, 2015.
[172]S. Deng, K. Yoshiyama, T. Mitsubuchi, and A. Sakurai, "Hybrid method of multiple kernel learning and genetic algorithm for forecasting short-term foreign exchange rates," *Comput. Econ.*, vol. 45, no. 1, pp. 49–89, 2015.
[173]D. Huang, X. Wang, J. Fang, S. Liu, and R. Dou, "A hybrid model based on neural networks for financial time series," in *Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on*, 2013, pp. 97–102.
[174]B. Sun, H. Guo, H. R. Karimi, Y. Ge, and S. Xiong, "Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series," *Neurocomputing*, vol. 151, pp. 1528–1536, 2015.
[175]Q. Cai, D. Zhang, W. Zheng, and S. C. H. Leung, "A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression," *Knowledge-Based Syst.*, vol. 74, pp. 61–68, 2015.
[176]W. Wang and X. Liu, "Fuzzy forecasting based on automatic clustering and axiomatic fuzzy set classification," *Inf. Sci. (Ny).*, vol. 294, pp. 78–94, 2015.
[177]A. Dutta, "PREDICTION OF STOCK PERFORMANCE IN INDIAN STOCK MARKET USING LOGISTIC REGRESSION," *Int. J. Bus. Inf.*, vol. 7, no. 1, 2015.
[178]Y. Bai, B. Wan, X. Zong, and W. Rao, "A Modified ARIMA Model Based on Extreme Value for Time Series Modelling," 2015.
[179]S. Bhattacharyya, P. Dutta, and S. Chakraborty, "Hybrid Soft Computing Approaches," 2016.

★★★

An academic review: applications of data mining techniques in finance industry

Jadhav, Swati

2017-05-31

Jadhav S, He H, Jenkins K, An academic review: applications of data mining techniques in finance industry, International Journal of Soft Computing and Artificial Intelligence, Volume 4, Issue 1, Pages 79 – 95.

http://ijscai.iraj.in/volume.php?volume_id=258