The Application of Data Mining Techniques for Financial Risk Management: A classification framework

Article · August 2020

CITATIONS
0

READS
145

1 author:

Tariq Saeed
Taibah University
9 PUBLICATIONS **13** CITATIONS

SEE PROFILE

The Application of Data Mining Techniques for Financial Risk Management: A classification framework

Tariq Saeed[†]
tmian@taibahu.edu.sa
[†]Department of IS, College of Computer Science and Engineering, Taibah University, Madinah Almunawarah, Kingdom of Saudi Arabia

Summary
Over the last few decades the world has witnessed a surge in the reliance on financial services (e.g. banking, credit cards, insurance), whilst the advent of the internet has led to a sharp rise in the number of online transactions. Both of these factors are driving an increase in the prevalence of financial fraud, precipitating the need for a novel approach to financial risk detection and management. One solution that has become feasible due to the availability of high amounts of storage spaces and computational power that has emerged over the past decade is data mining.
This paper sets out to examine the usage of data mining to detect and mitigate financial risks arising from financial frauds. The study used a Kaggle dataset and conducted experiments using several different classification metrics. The best performance for identifying creditworthy customers in banks was achieved by the Random Forest classifier..
Key words:
Random Forest, Data Mining, Bagging, Support Vector Machine, Financial Risk, Credit risk

1. Introduction

Technology is an essential aspect of financial risk analysis, as well as a tool for providing an alarm with regard to future trends. Data mining is used to gather and select valuable information [1]. It is applied in the big data models since it employs machine learning and artificial intelligence techniques to find relevant information in a scenario in a pile of data [2][3]. Therefore, it facilitates the application of relevant data sets to real-time financial risk management. Organizations are utilizing large data sets to establish patterns essential for operational, liquidity, market, legal, and credit risk mitigations [4] [5]. However, to employ data-driven conclusions and automation, it is necessary to develop models for addressing real-time challenges [6]. Data mining involves the integration of modeling techniques with statistical, artificial intelligence, and machine learning [7]. The combination of these tools ensures that firms can address varied situations based on the insights gained from unsupervised and supervised learning, as well as decision making. In this regard, learning and data processing algorithms have been developed to achieve the financial organization's goals, especially in the collection and gathering of information [8]. Data mining is efficient due to its ability to process information from several platforms, as well as databases while achieving the desired outcomes. Humans can take longer to process such information and achieve the efficiency gained from deep and machine learning [9]. Hence, data mining is now used in financial risk management to establish profiles, trends, and real-time insights.

2. Fundamental Techniques for Data Mining

Data mining is the process of extracting information or discovering hidden and valuable knowledge based on some patterns from large data. Data mining is favorable for different enterprises, for example, manufacturing, advertising, risk management and so on. Accordingly, there is a requirement for a standard data mining process. This data mining process must be reliable. Additionally, this procedure ought to be repeatable by businessmen with practically zero knowledge or expertise about data science.

Data Mining Process

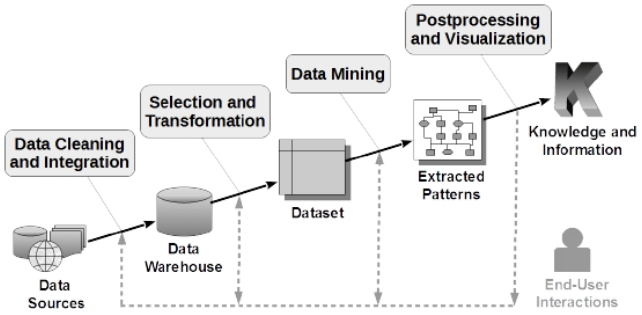


Fig. 1 Data Mining Process

The data mining process is classified in two stages: i) data preparation/data preprocessing, and ii) data mining. The data preparation process includes data cleaning, data integration, data selection, and data transformation. The second phase includes data mining, pattern evaluation, and knowledge representation.

Manuscript received August 5, 2020
Manuscript revised August 20, 2020

a. Data Cleaning

It has been observed that the real-world data, also known as raw data, is usually inconsistent or unreliable. First step is to clean the target data e.g. filling in the missing values and compute the required parameters/values.

b. Data Integration

The raw data i.e. data collected from several sources tends to be in different formats and locations, for example, spread over several different databases, spreadsheets or documents. The data from all these sources needed to be formatted using a common metadata or data dictionaries to reduce errors in the data integration process. Data integration also tries to reduce redundancy without losing any data.

c. Data Selection

At the data selection phase data relevant to the analysis is pulled from the data sources. This procedure sniff through the stored big data or historical data and creates a subset of data to achieve the set objectives of the analysis.

d. Data Transformation

Data transformation is a process to pull data from several different sources and glean it together and make it suitable for mining.

e. Data Mining

This is the process of to identify and extract patterns found in the target data set. Data mining is done in several steps such as classification, prediction, and clustering, among others.

f. Patterns Evaluation

Pattern Evaluation is the process of detecting strictly increasing patterns representing knowledge based on given measures. While, a pattern is interesting if it is potentially useful and easily understandable.

g. Knowledge Representation

Knowledge representation is presenting data to the target audience in a structured manner and appealing way.

Industry Standard Process for Data Mining

The Industry Standards Process consists of 4 phases that occur in a cyclical process:

i. Data Sources

Once the user requirements are identified, effort is focused to identify the available data and its sources.

ii. **Data Exploration and Preparation**

This is a multi-stage step where several different activities took place such as such as data load, data integration. Once the target data is collected the "surface" properties of the acquired data need to be examined and reported. Later by using querying, reporting, and visualization data is explored to tackle the data mining questions or objectives. Finally, the quality of the data is assessed making sure that the acquired data is complete and there are no missing values etc.

iii. Modeling

This is the stage where a suitable modelling technique is selected, in the context of the business goal, and a test scenario is generated to validate the quality and efficiency of chosen model. Then, by using modeling tools one or more models on the dataset need to be prepared. Finally, to make sure that the models meet business initiatives the selected models are presented to the project's participants/stakeholders.

iv. Deployment Model

The information gleaned from the above stages needed to be presented to the stakeholders so they can employee this for business decision making. This information must be presented to meet the business requirements. This stage could be as simple as creating a report or as complex as a repeatable data mining process across the organization. A deployment plan also contains a maintenance plan to protect the integrity of the base data. The final report, usually, contains a project insight, outcomes and a comprehensive project review to identify the future enhancements and improvements.

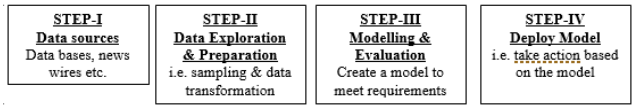


Fig. 2 Data Mining Process in Financial Industry

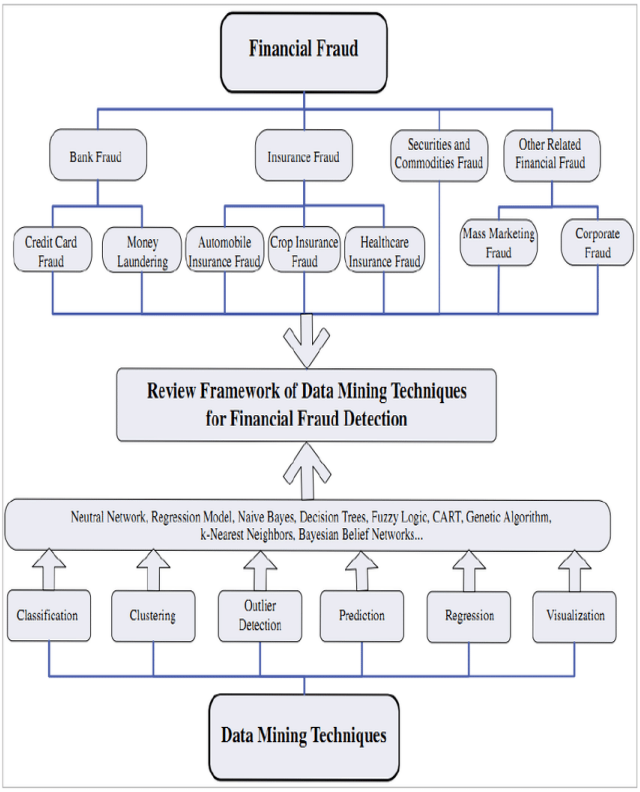
Data mining involves various techniques in gathering and preprocessing of data. One of the techniques utilized in data classification and prediction is the decision tree (Olson & Wu, 2017). This approach involves methodologies for the inductive learning algorithm. The learning in this methodology involves obtaining the appropriate rules for irregular and disorganized data. It is associated with simple comparative patterns, which can also be transferred to database query languages. This attribute is essential in ensuring that the data can be harmonized from different databases, given the necessity for big data [7]. Another essential attribute of the decision tree is the enhanced accuracy in terms of similarity of data sets.

Association Rules Technique is used to find relationships between datasets. Mostly, it is used as a mining technology to find relationships between data from multiple databases. This approach achieves precise, useful, and clear results [10]. Additionally, it can be used in indirect data mining. Furthermore, it can be used in dealing with lengthened data. The method employs algorithms, such as FP-growth

and Apriori algorithm. These approaches correspond to finding existing links between data sets.
Another crucial approach for data mining is the clustering analysis, which is also known as unsupervised classification. It is a critical tool for obtaining higher accuracy, as well as similarity, for data categories. It entails dividing data into categories, after which deliberations for different groups are made to minimize the gap [11]. This methodology entails four steps: i) features and selection; ii) similarity determination or calculation; iii) grouping; iii) and iv) clustering outcomes [1]. Algorithms, such as BIRCH, k-medoids, k-means, ROCK, and CURE, are used clustering analysis. This approach is suitable for organizing information into a classification mode, as well as grouping the data that does not have a description.

3. Literature Review

Financial fraud is becoming an everyday problem for the financial world. Figure 3 shows that there exist several different types of financial frauds prevalent in the industry (e.g. bank fraud, insurance fraud etc.). All these types of frauds could be detected by using data mining techniques, such as classification, clustering, outlier detection, prediction, regression, and visualization.



Figu. 3 Financial Fraud Detection

Data Mining for Credit Risk Management
The use of credit cards is becoming a norm in the financial world. The banks and credit card issuers use customer credit history to evaluate the customers' credit worthiness and create a credit rating.
The following figure explains how a financial institution uses stored data on the customers' financial history to rank the worthiness of the customer to issue or decline a request for a credit card.

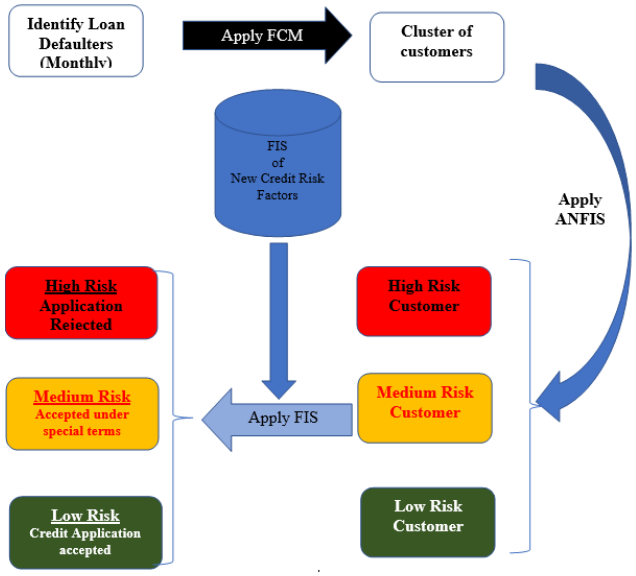


Fig. 4 Model for Customer Classification for Issuing Credit Cards

Data mining is used in financial risk management, whereby related models are widely studied and applied. Credit rating models and loan risk evaluations utilize data mining. Data mining technology related to intuitive quantitative grading is used to evaluate bank account credit [4]. Credit scores utilize data mining models to obtain indicators for ratings of a client. Consequently, the financial institution decides whether to provide the applicant with the desired limit or reject the application. Additionally, data mining employs other critical information for a customer, such as unusual usage of the credit card, illegal money loss, fraud, or extreme spending [4]. The insights and learning gained from extensive data mining processes provide the financial company with an appropriate overview of the client, which becomes the determinant for approval or disapproval of lending.
Credit worthiness is related to the losses caused by a client or party that fails to fulfill contractual payment or debt. Other risk metrics are considered with regard to the increased risk in relation to the transaction period. Before the development of machine learning and artificial intelligence tools, financial institutions relied on probit, logit, and classical linear regressions for credit risk modeling. Today, organizations have turned to artificial

intelligence, which involves several approaches for gathering and developing connections for data sets (Davenport, 2018). It is imperative to note that traditional mechanisms have not been complete in addressing the dynamic nature and volatile aspects of economic systems, which influence credit risk, as well as the clients' ability to fulfill contractual terms. Credit risk management capabilities are significantly enhanced through modern data mining strategies, especially neural networks and machine learning algorithms.

While both machine learning and artificial intelligence continue to gain popularity in the credit risk assessment models, these mechanisms combine modern tools with statistical methods to achieve accurate modeling mechanisms. It has been observed that the credit default swap (CDS) growth has contributed to the large-scale unpredictable elements used in determining the cost of default [12]. Data mining for machine learning has been proved to outperform traditional mechanisms, as well as benchmark models, in terms of accuracy and hedging measures establishments.

Data mining is applied in the development of Small and Medium Enterprise (SMEs) and consumer lending structures, given the availability of potential data. Application of modern mechanisms, such as machine learning, for data mining, have proved effective whereby they contribute to more savings as compared to traditional models. The application of machine learning with multivariate detection proves useful in the estimation of credit risk [13]. The corresponding outcomes influence decisions for lending [13].

Various models have been developed for evaluating creditworthiness to reduce non-payment risks. Financial institutions are required to develop their assessment systems based on context-based situations [14]. Some of the past approaches have failed to predict customer behavior accurately, hence leading to losses. Most of these models that financial organizations have utilized are static, but credit risk patterns are influenced by several factors, including political fluctuations [14]. As a result, such factors as sanctions have contributed to credit payment failures.

Essentially, political and social and factors affect financial risk patterns. Dynamic modeling can accommodate economic, social, and political trends. Data analysis in financial risk assessments eliminates the somewhat biased human judgment [15]. To effectively achieve data mining efficiency in financial risk management, it is imperative to outline the predictors. An adaptive network-based fuzzy inference system (ANFIS) can be effective in providing the customer profile [15]. The corresponding financial models should be flexible and adaptable to real-life situations, particularly political and economic factors. The proposed approach for data mining and processing by [15] examines the bad customers' profiles for a particular period. The resultant outcomes are used in the customer assessment, and changes in general customer profiles are replicated on the system. The most critical feature of the model is that customer assessment factors are dynamically established.

Designing appropriate techniques for data mining requires consideration of the required capabilities. Algorithms to be employed, including learning classifier systems, instance-based algorithms, Bayesian Networks (BNs), k-nearest neighbor, support vector machines, fuzzy modeling, neural networks, rule induction, and decision trees, are considered on the bases of their capabilities [15] [16]. These techniques are categorized as machine learning, artificial intelligence, or classical statistics.

The hybrid model is employed in the creation of algorithms for credit risk examination. It employs the feature selection and classification algorithm based on ensemble teaching [1]. The hybrid model consists of three stages: i) gathering and preprocessing of data; ii) feature selection; iii) and classification. Various algorithms are useful in the feature selection, including relief attribute evaluation, information gain ratio, genetic algorithms, and component analysis [15]. Afterwards, the appropriate model is used in the ensemble classification algorithm. The Feature Selection (FS) algorithm can also be integrated into this stage [15]. The third stage, which involves the classification of the dataset, reveals that adaptive boosting can achieve higher accuracy.

Criteria optimization is another essential factor considered in the improvement of the data mining algorithm. It uses kernel function for mapping input points and fuzzy function for multi-criteria optimization [15]. These functions use dimensional feature space with the latter being used at each data point, while the former is utilized to map high dimensional features. Unequal penalty attributes are used in reducing imbalanced classes, hence overcoming challenges associated with factorization incongruence. It is imperative to consider an accounting model integration to enhance predictive capacity [15]. It is essential to note that the techniques employed in the hybrid model do not indicate superiority since they capture different attributes of credit risk management. In addition, structural policies can also be integrated into the system design to improve bankers' confidence and reduce credit risks in asset-based financing. The hybrid data mining approach involves classification and clustering for credit scoring by identifying homogeneity and inconsistency, which are then used for exclusion or isolation [15]. One of the advantages of this model is that it can employ several data classifications rather than the distinct or bad categories. The diverse classification models are useful in ensuring that good and bad customers are classified accurately based on a well-deliberated cut-off point.

Since macroeconomic variables have a significant impact on credit risks, it is imperative to develop criteria for predicting external environment trends. As a result, the hybrid Support Vector Machines (SVM) model is used to evaluate credit scores using: the neighborhood data sets for input selection; kernel parameter search grid; optimal hybrid input; and determination of accuracy selection between methods [15]. The comparison between logistic regression, linear discriminant analysis, and SVM based classifier proved SVM to be more accurate [4]. One of the concerns is the ability to combine classifiers involved in machine learning. The credit prediction is more accurate in classifier ensembles, hence the preferred option. Therefore, the algorithms and models, are selected based on their overall accuracy [15].

Some models emphasize large collateral to secure loans, as well as increase the accuracy in measuring the customer repayment. Nevertheless, these models are likely to be inaccurate in unpredictable and unusual situations, such as sanctions [4]. The behavior of the customer also changes with time. Accounting for customer credit risk and behavior requires the development of a sophisticated model that addresses the unpredictable circumstances or crises [17]. In this regard, [18], proposed a risk assessment model covering some factors of concern. The model requires a batch of data produced by a generalized additive model whereby learning units and supervision about environmental training are considered static units. A Gini coefficient has been used in measuring the previous month's data [15]. While a full memory time window can be used to append new changes with regard to the customer profile, it has proved to be less adaptive to major trend variations. To counter this limitation, a fixed short memory can be implemented. Despite this sophisticated model outdoing static techniques with regard to preventing future losses, it has disadvantages because of the operational norms in the banking industry [19]. It is essential to note that the banking industry normally uses static models that utilize long-term records for credit scoring models.

Data Mining for Operational and Enterprise Risk Management

Organizations, whether small, medium, or large, can utilize data mining tools, big data, and business intelligence tools to gain insights, as well as improve business operations [7]. Data mining can be applied in management to establish patterns that can be used in analyzing behavior and trends [15]. Consequently, the resultant insights are used to develop a sustainable strategy. To further mitigate enterprise risk, data mining is utilized in capturing operational and industry-based trends, as well as concerns, hence improving business competitiveness and enterprise market share.

Both direct and indirect financial losses can result in operation breakdowns. Internal and external events can subject a financial institution to risks [20]. Natural disasters, procedural neglect, operation errors, frauds, and failed systems can subject an organization to numerous risks [19]. According to [21], Machine learning solutions to address some of these issues employ data mining tools to gather the information that helps to address operational risk exposures and complexity in explaining varied scenarios [22]. On the same note, artificial intelligence and machine learning is an indispensable method for identifying as well as measuring risk exposure to determine its effects on operations [22]. Shifting trading risk requires the establishment of an appropriate risk mitigation strategy. Additionally, data mining can be employed in data collection for repetitive processes and extensive document information. One of the advantages of the repetitive process with regard to providing operational data is that it can be accumulated into large data sets, hence the effectiveness of machine learning.

Organizations face numerous risks that require automated detection systems for referenced risk management. Financial institutions utilize algorithms for data mining, as well as machine learning and artificial intelligence, to mitigate the numerous risks that can happen without detection [15]. In this regard, data mining is used to design financial control mechanisms to protect data, systems, and clients. The ability for machine learning and artificial intelligence to enhance data mining models increase the pace for routine tasks, and unstructured process data to reveal risk [21]. These methodologies promote automation in many aspects, hence reducing the risks associated with human errors. In this regard, the evaluation of networks and clients becomes easier [15]. Employees and traders can also be monitored using this data analysis. Regarding behavior, clustering and classification are essential tools for developing profiles in which electronic, trade, and voice communication data is used to establish patterns, as well as latent risks [19]. Alerts are also established using information prioritized data to determine suspicious activities. Core artificial intelligence tools used for financial fraud detection employ data mining in selecting the appropriate data sets.

Data mining technology has been useful in detecting early warning signs in the banking sector. In this dimension, data mining is utilized in determining large scale financial risk amid the advent of the financial crisis. Primarily, it is used in the analysis that helps to reflect bankruptcy, payment crisis, and financial deterioration [17]. The warning models with respect to these factors, are used to provide control and early warning for enterprises. The management is then tasked with the responsibility of prudent financial management, as well as creditors, investors, and policymakers' insights. Various statistical, machine learning, and artificial intelligence methods,

including logistic regression, neural network, and linear regression, are used to develop a warning model for the financial crisis [19] [23]. These early signs are useful for financial institutions, which are bound to avoid unwarranted risks.

The implementation of financial risk analysis requires clarification of the aims and objectives to determine the appropriate tools, as well as relevant data sets. The data mining applications in financial firms follow six steps: i) content and objectives clarification; ii) data collection; iii) preprocessing; iv) data mining; v) evaluation and interpretation; vi) and knowledge assimilation [24]. The first step involves clarification of the relevance of the information or content to be collected to make the whole process accurate in addressing the impending financial risks. The second stage consists of the preparation of data acquisition from the varied sources, including the data warehouse and accounting systems. The third process involves rectifying the problems that may be present in the collected data sets [24]. This stage aims to address difficulties related to cumbersome data structures, non-standard content, and incomplete data sets [24]. The fourth step involves data mining, whereby the appropriate algorithm is selected and used in collecting the desired information. Appropriate algorithms, as well as business intelligence, are used in this phase for evaluation and interpretation of the outcomes [24]. Various models are presented for decision making in a financial enterprise. These models should be relevant to the body of finance. The final and sixth step for a data mining application in business processes is the assimilation and application of knowledge to the business information system. The corresponding insights are used in financial risk analysis as well as decision making.

Prediction of financial crises using data mining is a crucial aspect for businesses. It is useful in providing financial risk warnings through the analysis of data corresponding to an enterprise. The results are useful in providing technical support and decision-making (Abbas et al., 2019). Therefore, in the selection of data mining processes, relevant indicators with respect to enterprise-specific metrics are considered. In this regard, time-series characteristics are applied in data mining. The first idea is the identification and application of time series characteristics, which can be based on the lifecycle [25]. The second phase relates to finding relevant laws through the corresponding data. Finally, the laws are used to relate the data to financial processes and features, making relevant predictions.

Data Mining for Market Risk Management

Numerous data sets are available in the financial markets for trading and investments. Market risk management strategies attempt to reduce exposure to the negative implications of corresponding shifts in portfolios. Data mining and processing using machine learning are useful in market risk management [6]. It is pertinent to note that appropriate techniques and systematic methods are fundamental for validation and modeling, and it is important to establish the role of market models in trading. Each mechanism is suitable for a particular purpose. One of the fundamental aspects that require data mining and machine language mechanisms is stress testing for market models [14]. The role of data mining, in this case, is gathering as well as processing data to determine the emerging risk with respect to trading behavior.

Both machine learning and neural networks for financial market analysis, as well as stakeholders' behavior, corresponding to the improvement of the current modeling strategies, which are either incomplete, invalid, or false. Hence, data mining is a critical aspect of model risk management, as well as the improvement of the current models [19]. Machine learning for data mining has been useful for model validation in various firms due to the efficiency gained after conducting machine simulations. In this case, data mining processes differ in the implementation of unsupervised or supervised methods [8]. These models can establish new patterns. In essence, market risk management models are used in real-time monitoring, deviation testing, and validation [6]. Since machine learning techniques and artificial intelligence drive the corresponding methods, data mining approaches are necessary for gathering and preprocessing data tests.

Data mining also assists in risk management for large firms whose assets have a significant impact on market pricing. Data mining driven by machine learning and artificial intelligence and the consequent application of clustering techniques is useful for the implementation of countermeasures for illiquid markets trading [5]. It is essential to note that simulations, as well as real measures, require the appropriation of methodologies for data mining that accurately measure profits and losses in case a large firm enters the market [26]. Similarly, measures can be developed whenever significant equity or assets are released to the market. Data mining is an essential process in machine learning techniques used to identify the relationship between the release of assets to the market and their observable quantitative impact.

Data mining is also used in trading algorithms whereby learning is reinforced, so market reactions trigger changes that are embedded in future activities. The foreign exchange trading analysts use decision trees and neural networks to warn traders [6] [26].

Data Mining for Legal Risk Management

Compliance is a critical component for financial institutions, given the controls enacted, especially after the financial crisis. The risk management processes are deemed incompatible with bureaucratic functions for regulatory compliance, but they are linked in that they are

related to the systems for risk management [27]. Compliance is, therefore, linked to the varied implications for enterprise risk management. Similarly, operational, market, and credit risks are considered in compliance risk management [27]. Data mining, machine learning, as well as artificial intelligence are employed in ensuring that the bulky data sets are evaluated to guarantee that they are compliant with the legal provisions [2]. Additionally, data mining is critical in the determination of non-conventional data, which violates regulatory frameworks and provisions [27]. In most cases, data mining is continually employed to monitor the operation and credit aspects of a company. However, tracking on a real-time basis is crucial in avoiding breaches related to compliance. Financial firms also can utilize data mining for minimizing regulatory capital that is spent annually.

Individuals from all walks of life in the banking and financial services are affected by automated socially sensitive decisions, which at times exclude some groups based on the modeling and algorithms employed for sorting profiles [28]. Today, companies can utilize personal and collective data for evaluating socially sensitive decisions. Data mining tools can establish unfair use of automated decisions, also known as web-lining [28]. Web-lining outcomes can be unfair to the marketing and pricing of commodities. Data mining with regard to legal provisions analyzes the applied algorithms to determine whether their use, intentionally or unintentionally, harms, or discriminate against a particular group[28]. The focus of the corresponding algorithms for data mining is to eliminate biases with regard to race, gender, or other social biases. Therefore, experts seek to understand the implications of algorithms utilized by financial institutions to ascertain their social implications at the business level and avoid consequent effects regarding regulatory and legal frameworks.

Institutions can utilize discrimination-aware data mining (DADM) to address discriminatory practices and outcomes, some of which could also subject them to legal actions [28]. This approach focuses on the extraction and analysis based on social bias rules. It is imperative to deal with the legal aspect of the models used for processing data in financial institutions. Fundamentally, it is difficult for the current data mining models to address all the challenges related to automated discrimination since statistical observation is not sufficient [28]. Instead, these processes are multijurisdictional, spanning several legal areas, including but not limited to data protection and equality rules. For instance, in the US automated discrimination is purportedly addressed under the legal framework [28]. While legal discrimination may be dealt with at an institutional level, data mining algorithms and models may promote social discrimination. Similar, machine learning and artificial intelligence models should, therefore, be used to ascertain preprocessing fairness with regards to equality and legal provisions adherence.

4. Experimental Consideration

This section presents the performance and accuracy of the proposed algorithm. Based on the parameters of confusion matrix (Table 1) F-measure, Accuracy, MCC and ROC were calculated and used to evaluate the accuracy of the proposed algorithm.

		Actual Values: Defected	Actual Values: Non-defective
Predicted Values	Defective	True Positive (TP)	False Positive (FP)
	Non-defective	False Negative (FN)	True Negative (TN)

Table 1: Confusion Matrix

In the above confusion matrix:
TP = A positive instance detected positive.
FP = A negative instance detected positive.
FN = A positive instance negative by the algorithm.
TN = A negative instance by the algorithm.

The following is a brief description of the performance measures used to evaluate the accuracy of the algorithms:

F-Measure is calculated by evaluating the Precision (i.e. the ration of True Positive (TP) instances with respect to the total number of instances which were classified as positive) and Recall measure (i.e. the ration of True Positive (TP) instances with respected to the total number of positive instances).

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

Now

$$F - Measure = \frac{Precision * Recall * 2}{Precision + Recall}$$

Accuracy is the ratio of correctly classified instances to the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

AUC estimates how well an algorithm can distinguish between defective and non-defective classes.

$$AUC = \frac{1 + TP_\gamma - FP_\gamma}{2}$$

5. Experiment

For the experiment, the Kaggle dataset (https://www.kaggle.com/uciml/german-credit) was adopted [29]. The original dataset contains 1000 entries with 20 categorial/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person to whom a credit is extended by a bank. Each person is classified as good or bad credit risks according to the set of attributes. For classification, Gaussian Naïve Bayes, Random Forest, J48, and MLP as base learners were used for smooth dataset and bagging, respectively. The tests were performed utilizing Pycharm in a Python domain. The nature of the classifiers in this examination was estimated utilizing accuracy, precision, recall, ROC, and F-score of classification. It is essential to underscore that weighted normal was utilized to figure these measurements. The point behind the decision of the weighted normal was to evaluate measurements for each group label and to clarify the label uniqueness. The classifier execution was estimated with the training dataset dependent on 10-fold cross-validation. The resulting algorithm was used for performing experiments. First, a Kaggle dataset and a list of classifiers were given and then repeated over datasets, as shown in Line 7. As shown in Line 8, the program retains training and test sets based on 10-fold cross-validation with data rearranging before splitting. For each fold, the loop in Lines 9–20 focused on training the classifiers, obtaining predictions and assessing evaluation metrics. The average score was calculated. The method defined in Lines 7–28 was repeated through all over the dataset.

Input: *Data-set, Classifiers*
Results: *AvgAccuracy, AvgPrecision, AvgF-Score, and AvgAUC*
1: Data set: Kaggle Dataset for Fake News
2: Classifies ⟵ {RF,NB,MLP,J48, Bagging(RF,NB,MLP,J48),Smooteddataset(RF,NB,MLP,J48)}
3: All Accuracy Scores ⟵ {}.
4: All Recall Scores ⟵ {}.
5: All Precision Scores ⟵ {}.
6: All F– Scores ⟵ {}.
7: All AUC Scores ⟵ {}.
8: For DS $\in$ *Datasets* **do**
9: **for** $Xtrain, Xtest \in$ KFold (nsplits = 10, shuffle = $True$).$split(DS)$ **do**
10: $(Xtrain, Xtest) \leftarrow PerformStandardScaler(Xtrain, Xtest)$;
11: $ResampledXtrain,\ ResampledYtrain \leftarrow SMOTE(Xtrain, Ytrain)$;
12: **for** clf $\in Classifiers$ **do**
13: $clf \leftarrow TrainClassifier(clf, ResampledXtrain, XtrainLabels)$;
14: $predictions \leftarrow predict(cls, Xtest)$;
15: $Accuracy \leftarrow ComputeAccuracy(predictions, XtestLabels)$;
16: $Recall \leftarrow ComputeRecall(predictions, XtestLabels)$;
17: $Precision \leftarrow ComputePrecision(predictions, XtestLabels)$;
18: $F\text{-}score \leftarrow ComputeFmeasure(predictions, XtestLabels)$;
19: $AUC \leftarrow ComputeAUC(predictions, XtestLabels)$;
20: $AllAccuracyScores \leftarrow AllAccuracyScores \cup Accuracy$;
21: $AllRecallScores \leftarrow AllRecallScores \cup Recall$;
22: $AllPrecisionScores \leftarrow AllPrecisionScores \cup Precision$;
23: $AllFScores \leftarrow AllFscoreScores \cup F\text{-}score$;
24: $AllAUCScores \leftarrow AllAUCScores \cup AUC$;
25: **end**
26: **end**
27: $AvgAccuracy \leftarrow ComputeAvgAccuracy(AllAccuracyScores)$;
28: $AvgRecall \leftarrow ComputeAvgRecall(AllRecallScores)$;
29: $AvgPrecision \leftarrow ComputeAvgPrecision(AllPrecisionScores)$;
30: $AvgF\text{-}score \leftarrow ComputeAvgFmeasure(AllFScores)$;
31: $AvgAUC \leftarrow ComputeAvgAUC(AllAUCScores)$;
32 **end**
33 **return** AvgAccuracy,AvgRecall,AvgPrecision,AvgF-score,AvgAUC

Fig. 5 Experiment for classification

6. Results:

Experimental results are shown in Tables 2-5 and Figures 6−8. Tables shows various classification metrics. The best performance in terms of identifying good credit in banks achieved by Random Forest, Bagging Random Forest classifiers with 73.5%, and 75.3% accuracy and best performance on smoothed dataset was by Random Forest with 74.1% accuracy. Same for Recall and F-Measure the performance of Random Forest is best with values 72.2%, 74.1%, 75.4% for Recall and 72.6%, 74.1%, 75% for F-measure. In general, an AUC of 0.5 suggest no discrimination, 0.7 to 0.8 is considered excellent and more than 0.9 is outstanding Again Random Forest is an excellent performer with values of 74.0%, 74.8 and 75.8%.

Table 2: Recall

Base Learner				Smoothed Dataset				Bagging Smoothed dataset			
RF	NB	MLP	J48	RF	NB	MLP	J48	RF	NB	MLP	J48
72.2	69.8	68.1	70.4	74.1	69.7	69.1	72.9	75.4	71.3	70.8	73.5

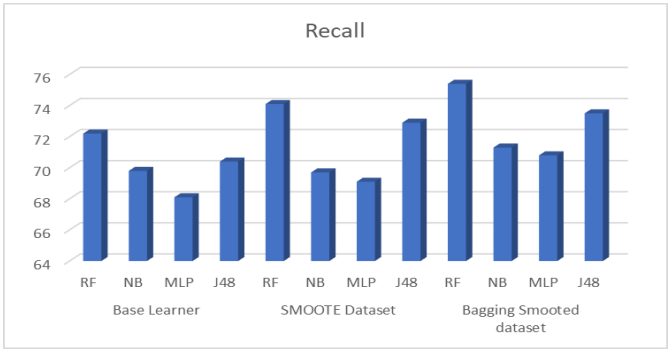


Fig. 6 Performance of classifiers (Recall) for bank credit

Table 3: F-measure

Base Learner				SMOOTE Dataset				Bagging Smoothed dataset			
RF	NB	MLP	J48	RF	NB	MLP	J48	RF	NB	MLP	J48
72.6	70	68.4	71.3	74.1	70.3	69.5	72.4	75.3	70.5	71	73.2

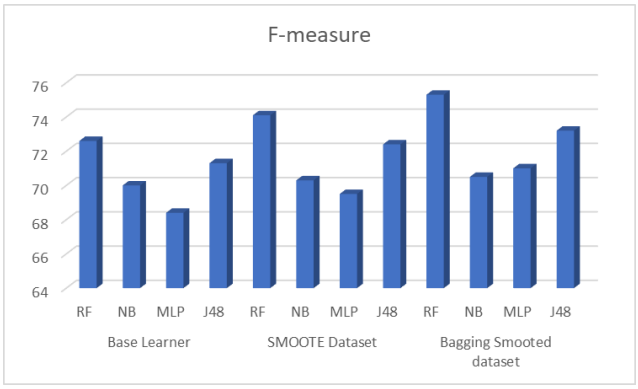


Fig. 7 Performance of classifiers (F-measure) for bank credit

Table 4: AUC

Base Learner				Smoothed Dataset				Bagging Smoothed dataset			
RF	NB	MLP	J48	RF	NB	MLP	J48	RF	NB	MLP	J48
74	74.6	66.1	69	74.8	73.4	70.3	69.2	75.8	73.6	72.1	74.6

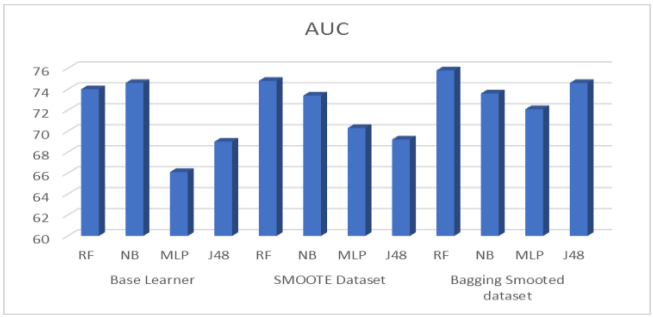


Fig. 8 Performance of classifiers (AUC) for bank credit

Table 5: ACCURACY

Base Learner				Soothed Dataset				Bagging Smoothed dataset			
Random forest	NB	MLP	J48	RF	NB	MLP	J48	RF	NB	MLP	J48
73.5	72	68.8	70.4	74.1	69.7	69.1	72	75.3	70	70.8	73

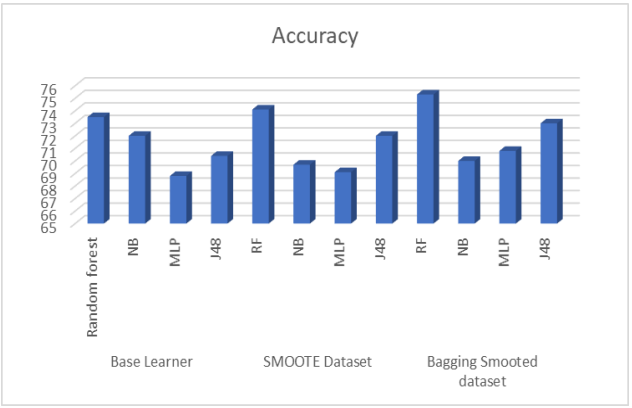


Fig. 9 Performance of classifiers (Accuracy) for bank credit

7. Conclusion

Data mining has been an essential technology utilized as a tool for developing modern models for financial risk management. This technology, together with statistical regression, machine learning and artificial intelligence is used in vast areas for financial risk assessment and management with cluster analysis, decision tree, and association rule being the fundamental approaches. Data mining has mainly been employed in credit risk assessment modeling and client profiling by determining credit ratings. Various techniques, as well as algorithms, including hybrid model and SVM, have been primarily used to improve the classification accuracy. In operational risk management, data mining has been utilized to develop financial controls for fraud and other critical issues, such as bankruptcy and financial crises. Market risk management has been applied in developing models that address real-time patterns in the financial markets. These models are also used to alert traders, as well as large enterprises, following quantitative implications of an asset injection into the market. In this research, with chosen dataset, the best performance was given by classifier random Forest in all type of metrics: Accuracy, AUC, Recall and F-Measure (given in Tables 2-5 and shown in figures 6-9). Similar modeling is used to create alerts for operational risk management. While organizations focus on the operational, market, and credit data mining models, legal-based modeling is essential for reducing risks and costs of breaching legal frameworks. Therefore, data mining can be used to eliminate algorithms and models for credit, market, and business operations that perpetuate social injustices by developing tools to decode the negative implications of the existing models as well as eliminating potential threats for legal breaches.

Acknowledgment

The author would like to express his cordial thanks to Prof. M. Z. Khan for his valuable advice.

References

[1] Hassani, H., Huang, X., & Silva, E. (2018). Digitalisation and big data mining in banking. Big Data and Cognitive Computing, 2(3), 18. doi: 10.3390/bdcc2030018

[2] Dicuonzo, G., Galeone, G., Zappimbulso, E., & Dell'Atti, V. (2019). Risk management 4.0: The role of big data analytics in the bank sector. International Journal of Economics and Financial Issues, 9(6), 40-47. doi: 10.32479/ijefi.8556.

[3] Wang, L., & Alexander, C. (2016). Machine learning in big data. International Journal of Mathematical, Engineering and Management Sciences, 1(2), 52-61. doi: 10.33889/ijmems.2016.1.2-006.

[4] Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. Risks, 7(1), 29. doi: 10.3390/risks7010029.

[5] İncekaraa , Ahmet and Çetinkayaa , Harun (2019). Liquidity risk management: A comparative analysis of panel data between Islamic and conventional banking in Turkey. Procedia Computer Science, 158, 955-963. doi: 10.1016/j.procs.2019.09.136.

[6] Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications, 67, 126-139. doi: 10.1016/j.eswa.2016.09.027.

[7] Davenport, T. (2018). From analytics to artificial intelligence. Journal of Business Analytics, 1(2), 73-80. doi: 10.1080/2573234x.2018.1543535.

[8] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. (2019). A systematic review on supervised and unsupervised machine learning algorithms for data science. Unsupervised and Semi-Supervised Learning, 3-21. doi: 10.1007/978-3-030-22475-2_1.

[9] Heaton, J., & Polson, N. (2016). Deep learning for finance: Deep portfolios. SSRN Electronic Journal. doi: 10.2139/ssrn.2838013.

[10] Zhan, F., Zhu, X., Zhang, L., Wang, X., Wang, L., & Liu, C. (2019). Summary of association rules. IOP Conference Series: Earth and Environmental Science, 252, 032219. doi: 10.1088/1755-1315/252/3/032219.

[11] Cavalcante, R., Brasileiro, R., Souza, V., Nobrega, J., & Oliveira, A. (2016). Computational intelligence and financial markets: A Survey and Future Directions. Expert Systems with Applications, 55, 194-211. doi: 10.1016/j.eswa.2016.02.006.

[12] Son, Y., Byun, H., & Lee, J. (2016). Nonparametric machine learning models for predicting the credit default swaps: An empirical study. Expert Systems with Applications, 58, 210-220. doi: 10.1016/j.eswa.2016.03.049.

[13] Figini, S., Bonelli, F., & Giovannini, E. (2017). Solvency prediction for small and medium enterprises in banking. Decision Support Systems, 102, 91-97. doi: 10.1016/j.dss.2017.08.001.

[14] Aziz, S., & Dowling, M. (2018). Machine learning and AI for risk management. Disrupting Finance, 33-50. doi: 10.1007/978-3-030-02330-0_3.

[15] Moradi, S., & Mokhatab Rafiei, F. (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financial Innovation, 5(1). doi: 10.1186/s40854-019-0121-9.

[16] Cerchiello, P., & Giudici, P. (2016). Big data analysis for financial risk management. Journal of Big Data, 3(1). doi: 10.1186/s40537-016-0053-4.

[17] Abbas, F., Iqbal, S., & Aziz, B. (2019). The impact of bank capital, bank liquidity, and credit risk on profitability in post-crisis period: A comparative study of US and Asia. Cogent Economics & Finance, 7(1). doi: 10.1080/23322039.2019.1605683.

[18] Sousa, M., Gama, J., & BrandÃ£o, E. (2016). A new dynamic modeling framework for credit risk assessment. Expert Systems with Applications, 45, 341-351. doi: 10.1016/j.eswa.2015.09.055.

[19] Olson, D., & Wu, D. (2017). Data mining models and enterprise risk management. Springer Texts in Business and Economics, 119-132. doi: 10.1007/978-3-662-53785-5_9.

[20] Weeserik, B., & Spruit, M. (2018). Improving operational risk management using business performance management technologies. Sustainability, 10(3), 640. doi: 10.3390/su10030640.

[21] Choi, T., Chan, H., & Yue, X. (2017). Recent development in big data analytics for business operations and risk management. IEEE Transactions on Cybernetics, 47(1), 81-92. doi: 10.1109/tcyb.2015.2507599.

[22] Kou, G., Chao, X., Peng, Y., Alsaadi, F., & Herrera-Viedma, E. (2019). Machine learning methods for systemic risk analysis in financial sectors. Technological and Economic Development of Economy, 25(5), 716-742. doi: 10.3846/tede.2019.8740.

[23] Alzeaideen, K. (2019). Credit risk management and business intelligence approach of the banking sector in Jordan. Cogent Business & Management, 6(1). doi: 10.1080/23311975.2019.1675455.

[24] Hou, Y., & Yuan, Z. (2019). Financial risk analysis and early warning research based on data mining technology. Journal of Physics: Conference Series, 1187(5), 052106. doi: 10.1088/1742-6596/1187/5/052106.

[25] Jin, M., Wang, Y., & Zeng, Y. (2018). Application of data mining technology in financial risk analysis. Wireless Personal Communications, 102(4), 3699-3713. doi: 10.1007/s11277-018-5402-5.

[26] Chandrinos, S., Sakkas, G., & Lagaros, N. (2018). AIRMS: A risk management tool using machine learning. Expert Systems with Applications, 105, 34-48. doi: 10.1016/j.eswa.2018.03.044.

[27] Arner, D., Barberis, J., & Buckley, R. (2016). The emergence of Regtech 2.0: from know your customer to know your data. SSRN Electronic Journal. doi: 10.2139/ssrn.3044280.

[28] Carmichael, L., Stalla-Bourdillon, S., & Staab, S. (2016). Data mining and automated discrimination: A mixed legal/technical perspective. IEEE Intelligent Systems, 31(6), 51-55. doi: 10.1109/mis.2016.96.

[29] Kaggle, https://www.kaggle.com/uciml/german-credit