Identifying Optimal Parameters And Their Impact For Predicting Credit Card Defaulters Using Machine-Learning Algorithms

Muhammad Qasim Idrees[1], Humaira Naeem[2], Muhammad Imran[3], Asma Batool[4], Dr. Nadia Tabassum[5]
Department of Computer Science & Information Technology, Virtual University of Pakistan
[1]qasim.ch102@gmail.com, [2]humairanaeem@vu.edu.pk, [3]mimran@vu.edu.pk, [4]asmabatool@vu.edu.pk,
[5]nadiatabassum@vu.edu.pk

ABSTRACT

Data mining and Machine learning are the emerging technologies that are rapidly spreading in every field of life due to their beneficial aspects. The financial sector also makes use of these technologies. Many research studies regarding banking data analysis have been performed using machine learning techniques. These research studies also have many Problems as the main focus of these studies was to achieve high accuracy and some of them only perform comparative analysis of different classifier's performance. Another major drawback of these studies was that they do not identify any optimal parameters and their impact. In this research, we have identified optimal parameters. These parameters are valuable for performing the credit scoring process and might also be used to predict credit card defaulters. We also find their impact on the results. We have used feature selection and classification techniques to identify optimal parameters and their impact on credit card defaulters identification. We have introduced three classifiers which are Kstar, SMO and Multilayer perceptron and repeat the process of classification and feature selection for every classifier. First, we apply feature selection techniques to our dataset with each classifier to find out possible optimal parameters and In the next phase, we use classification to find the impact of possible optimal parameters and proved our findings. In each round of classification, we have used different parameters available in the dataset every time we include and exclude some parameters and noted the results of each run of classification with each classifier and in this way, we identify the optimal parameters and their impact on the results Whereas we also analyze the performance of classifiers. To perform this research study, we use the "credit card defaults" dataset which we obtained from UCI Machine learning online repository. We use two feature selection techniques that include ranker approach and evolutionary search method and after that, we also apply classification techniques on the dataset. This research can help to reduce the complexities of the credit scoring process. Through this study, we identify up to six optimal parameters and also find their impact on the performance of classifiers. Further We also identify that multilayer perceptron was the best performing classifier out of three. This research work can also be extended to other fields in the future where we use this mechanism to find out optimal parameters and their impact can help us to predict the results.

KEYWORDS: Data Mining, Machine Learning, Credit Scoring, Classification, Feature Selection, Gain Ratio, Optimal Parameters.

1. INTRODUCTION

Financial institutions, always care about the customer transactions. The reputations of these companies depend upon the safety of customer credit and their satisfaction; it also helps them to increase their profit as well. For example, the use of credit cards is spreading rapidly everywhere, and it benefits customers as they become free from the hassle of carrying the cash every time and everywhere. This aspect of credit cards looks very beneficial, but credit cards also have problematic aspects that become a worrying point for the customers and banking institutes as well. It includes credit card frauds and loan defaults,etc. The main object that was at risk in the whole situation was credit. The number of credit card frauds and the ratio of defaults is increasing with every single passing day. Financial pressure and other uneven circumstances affect banks' performance and cause banks to default. For that, purpose many supervising regulations are enforced that keep checks on the banks. [3] proposed a bank failure forecasting technique.

Banks use credit-scoring mechanisms to avoid Credit risk but the risk remains there. Accurate and precise credit scoring procedure is very important for the success of financial organizations [11]. The increasing numbers of credit card users create a competitive environment between different financial organizations. The use of technology becomes necessary as multiple techniques exist and implemented to curb credit risk best of them were data mining methodologies that stand above all others. Applications of data mining (DM) spread across different fields [14]. Its functionalities can be categorized as clustering, feature selection, association rule mining, and classification.

Although many solutions for credit card defaulters identifications have been utilized however, every method falls short of expectation. Data mining approaches based on machine learning techniques proved very helpful in the detection of these defaulters. It helps in doing so based on the previous behavior of customers because data mining techniques have the capacity to discover the unseen information from a large set of data [6]. Data mining based approaches are used to accurately foresee the defaulters based on their previous records and with the help of their personal information correctly with high accuracy, so one can easily rely upon their results.

1.1 CREDIT SCORING

It is also known as statistical analysis that was performed before providing loans to customers. Loan providing organizations and financial institutes perform this analysis. The faith of the loan application is decided based on the value of the credit score. It determines the ability of individuals to return the loan amount within the given period. The value of this score relies upon the previous payment record of the customer. A customer can have multiple credit score values, as there are many methods to calculate credit score value. An accurate and precise assessment system for credit risk is must and vital for financial institution. In such an unpredictable and changing economy as the rate of loan defaults are increasing, authorities of financial institutions are finding it more and more difficult to correctly assess loan requests and tackle the risks of loan defaulters [15]. Banking sector main risk factor involve giving loans and issuing credit cards it include the risks of non-payment. According to the Basel 2 guidelines, banks need to develop their own credit risk assessment systems [12]. Research community presents large numbers of studies in the past twenty years, which relates with the use of data mining techniques to detect frauds, score credits and manage risks, but issues such as data selection, algorithm design, and hyperparameter optimization affect the perceived ability of the proposed solutions which result under performance of these methods [20]

The growing popularity of cryptocurrencies was changing the dynamics of the world. However, there also exist some fear in the use of these currencies that was their rate of exchange. Recent research [9] presents a model for the prediction of the exchange rate for bitcoins one of the cryptocurrency. Everyday lots of applications are submitted in banks related to loan, but Banks have limited funds. In this situation, the right decision would be very helpful and the role of ML based prediction model becomes very critical. Most of ML based Prediction models use the logistic regression, random forest classifier, support vector machine classifier etc. A Bank’s profit and loss depend on the amount of the loans that is whether the Client or customer is paying back the loan. Recovery of loans is the most important for the banking sector [16].

1.2 DATA MINING

Data mining (DM) or knowledge discovery is a term used interchangeably. It is the process of finding information from a large amount of data. DM provides its benefits to every field such as in business it is used to find out data patterns and relationships, which can be used to make decisions. By using data mining, we can predict future trends. Data Mining (DM) techniques are consider vital for the banking sector, with help of this we can get valuable information from the huge volume of data and develop better strategy for management and customer [5].

Different techniques are present to perform data mining, which includes Classification, Clustering, Prediction, Regression and many more. The process of DM consists of three steps that include data preparation, removing unwanted data and finding valuable information. However, DM approaches may vary under different conditions mostly DM techniques classified on the base of the discovery of knowledge, used methods and on the base of the database [8].

1.3 MACHINE LEARNING

Machine learning (ML) is a data analysis method that is used to automate analytical model building.

Simulation of ML based models is highly linked to Computational Statistics and its main aim was to focus on prediction making via computers. Its concepts also co-related to Mathematical Optimization which relates models, applications and frameworks to the field of statistics [1]. ML is a branch of artificial intelligence. It, is based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It can further be divided into subtypes that include supervised learning, unsupervised learning and reinforcement learning.

In supervised learning algorithm takes a known set of input data and known responses and trains a model to generate future predictions and response of new data. Whereas when we train an artificial intelligence (AI) algorithm using information that is unclassified and not labeled and allow the algorithm to act on that information without guidance is known as unsupervised learning. Although ML concepts are applied in many fields most prominent of them are in DM. Supervised ML algorithms or techniques can be classified as Logic-based algorithms, which include decision trees. Next, we have perceptron-based techniques that consist of single-layered, multi-layer perceptron and Radial basis functions networks further we have Statistical learning algorithms which include Naïve Bayes classifiers and Bayesian Networks next two types are Instance-based Learning and Support Vector Machines [7]. Feature selection is a technique that provides us the most useful attributes out of all. Feature selection is considered one of the core concepts of machine learning and the performance of the model hugely depends on it. Filter methods, wrapper methods and the embedded methods are three form of feature selection [10].

1.4 CREDIT CARD FRAUDS

The term credit card frauds are widely linked with theft or fraud that happens with the involvement of credit or debit cards. It is defined as the unauthorized or illegal use of someone's card or its information without the permission of that person for any purpose that includes access funds and makes purchases or any other use of these funds. In [18] Credit card fraud defined as the physical loss of credit card or loss of sensitive credit card information.The instances of credit card frauds are increasing alongside that criminals are also getting smarter that makes it harder for authorities to detect them. Credit card frauds can be of many types. Online transactions numbers has grown in large quantities and share of online credit card transactions are in huge numbers [17].

Application Fraud is when someone applies for a credit card with a someone else identity, then this type of fraud takes place. Electronic credit card imprints a type of fraud that happens when your card information that is hidden in the magnetic strip was stolen by fraudsters. Card-not-present fraud is another type of fraud that happens if someone gets the detail of your credit cards such as card expiry date, account number and other details that are printed on the card and use this information without the need of the physical presence of the card. Another case of fraud can happen if someone steals the card from you or it gets lost by you accidentally. A fraudster who gets this card can use it to make transactions this type of fraud is called as stolen card fraud and if criminals use the card information of someone else and use this information to get a new card for themselves then this fraud is termed as Card ID fraud. All these types of frauds are named as card related frauds. While some fraud types are termed as Merchant related frauds. Merchants related frauds occur when a merchant or their workers use customer accounts or personal information and pass it to fraudsters. While triangulation frauds occur in which fraudsters use the website, on which they sell products on discounted prices and dispatch the sale item before processing payment. The customer considers these sites as normal legitimate eCommerce sites and passes their card detail to these sites. Once card information is given to the site, they use this information to buy products for themselves [2].

Some credit card frauds are related to the internet which includes site cloning in which the whole website or some web pages are cloned by fraudsters through which users place their orders to buy products. In another type of fraud, criminals use false websites to sell cheap products to customers after getting complete detailed information regarding their credit card, which they use for many other illegal purposes. Some time fraudsters also use credit card number generator software, which generates valid card numbers along with their expiry dates.

2. MOTIVATION

The dynamics of the world were changing and now our living in a digital world. Advancement of technology provides us multiple benefits in every aspect of life. In banking sector users now avail the services at their ease. Now we have options of internet banking. Banking sector becomes digital and provide many benefits, but on the other hands threat and risk factor are also increased and number of crimes related to banking sector was also increased. Quick actions are required to curb these criminal activities. Cases of Loan and credit card defaults are rising with every passing day, which result huge loss of money to companies and customers. The main reason of this research work was to identify such parameters that can be helpful for forecasting credit card defaulters and also helpful to design any system that identify credit card defaulters, research community can utilize this study in order to develop any predicting model for any field.

3. METHODOLOGY

In this research, we have analyzed banking data using ML algorithms. In our analysis, we identify such parameters that are helpful in credit card defaulters prediction and decision making. We also find the role and implications of these parameters in predicting defaulters and identify which algorithm (SVM, Kstar, Multilayer perceptron) shows promising results out of all that was used in this analysis. We use feature selection and classification methods to fulfill these tasks. Feature selection is a technique that provides us the most useful attributes out of all. It reduces the inputs and makes the analysis process to work faster and also increase the accuracy of models and reduce the complexity. In short, we define it as the method to find out important data attributes from a large datasets. Feature selection techniques are divided into a wrapper and Filter methods. We perform feature selection by using two techniques that include filter method based ranker approach and after that we use wrapper methods based technique named as an Evolutionary search method. Ranker method rank attributes based on evaluator in this case we use Gain ratio as an evaluator. Whereas evolutionary search method uses the Evolutionary algorithm that is a generic population-based algorithm. As evolutionary algorithm is easily customizable that make it the right option for any problem. After feature selection we move to the next phase, which is a classification. Classification is a supervised learning approach which classifies things based upon previous data inputs.

Our selected dataset consists of multiple attributes which are numbered from X1 to X23. These attributes include Limit balance which specifies the individual or the collective credit amount then comes the Gender (1= male; 2 = female) Education (1 = graduate school; 2 = university; 3 = high school; 4 = others) Marital Status (1 = married; 2 = single; 3 = others) and Age (year). After that from X6 – X11 shows the past history of payment months from April to September of the year 2005 where -1 means pay duly and 1 represents payment delay for one month. 2 means a payment delay for two months and so on. Whereas attributes X12-X17 represent the amount of the bill for each month. Further attribute X18-X23 signify the previous month paid payments. And the last attribute is a default payment in a coming month where 1 means yes and 0 mean no.

As feature selection just provides us best combination of attributes, but it does not provide the detailed information regarding parameters such as the impact of each parameter on the result and their role in identification of defaulters which we will find through classification We will use supervised machine learning algorithms (SMO, Kstar and Multilayer perceptron) due to their long-range of application in different fields. SMO is good in a situation when we have no advance knowledge of the dataset. Similarly, Kstar is a good choice to use if we have multiple problems and Multilayer perceptron is another good classifier because it has fast computation time. We individually test each optimal parameters that we have found through feature selection technique one by one with every classifier to check the influence of each parameter on the result. First, we select the classifier and then we run this classifier on the dataset with different combination of parameters (by including and excluding parameters). In this way we find how much impact each parameter have upon the result and how helpful they are for defaulters identification.

Classification results also provide us figures like Accuracy, Precision, Recall, Kappa Statistics, and F measure and also identify the effect of these parameters on these figures. The detailed working of this system is shown in the diagram below. As shown in the Fig. 1, first we perform data preprocessing to convert the dataset in the required format which is arff format which has its own syntax to write data. After the dataset is ready for testing, then we perform next step which is Feature Selection. After this step is completed, then we perform classification in the next step and we repeat the process of classification for every classifier with each parameter that we identify through feature selection technique. As we have to find the effect of each parameter on the results.Then we analyze the results in order to find the impact of parameters on the performance of the classifier.

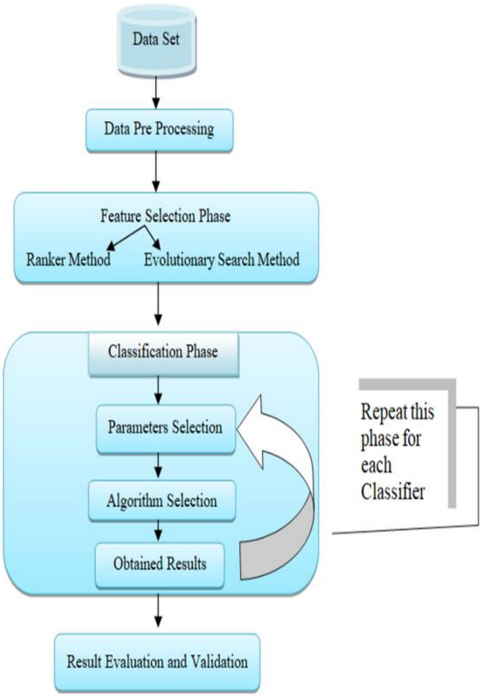

Fig.:1 Block diagram of proposed Research methodology

4. RESULTS AND DISCUSSION

4.1 FEATURE SELECTION

In this step, first, we use the Ranker approach to perform feature selection. In this approach, all attributes are evaluated separately and these attributes are evaluated based upon the evaluator. In our case, we select the gain ratio as an evaluator. Outcomes of the ranker approach method are described below. Six Attributes that represent the PREVIOUS PAYING HISTORY of customers is placed at the first six positions; these six attributes are PAY _0, PAY _2, PAY _3, PAY _4, PAY _5 and PAY _6. Whereas positions no 7,8 along with positions 10-13 are held by attributes which represents AMOUNT OF PREVIOUS PAYMENT. The names of these attributes are PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5 and PAY_AMT6. Attribute LIMIT BALANCE holds 9th position. Moving down further at 14th place, we have the attribute EDUCATION after which at 15th spot we have attribute AGE and attribute SEX was placed at the 16th spot. Whereas MARRIAGE attribute was ranked at 19th position. The last five positions are occupied by attribute BILL AMOUNT IN DIFFERENT MONTHS because they have zero gain ratios. The name of these attributes is BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5 and BILL_AMT6.

The highest gain ratio is achieved by attribute PAY_5 (PREVIOUS PAYING HISTORY) which is 0.090076 and the lowest one was 0.000759 for the attribute MARRIAGE. The graph in Fig.2 below shows the complete detail of the gain ratio achieved by different attributes.

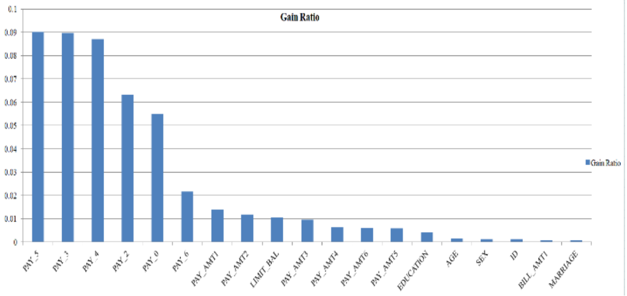

Fig.2 Percentage of the gain ratio of different attributes using Ranker Method

From the above graph in Fig. 2, we see that this approach ranks the “PREVIOUS PAYING HISTORY” as the most important attribute. However, the “AMOUNT OF PREVIOUS PAYMENT” attribute was ranked as a second important attribute. The third most important attribute was “LIMIT BALANCE.” Whereas attributes EDUCATION, AGE, SEX and MARRIAGE come after the attribute “LIMIT BALANCE.” After the Ranker approach, we move towards another approach of feature selection which was an evolutionary search method. We use this feature selection approach for each classifier differently and their results are as follows. First, we perform an evolutionary search using the "SMO" classifier, which points out 13 important attributes that include "SEX, EDUCATION, AGE, PAY_0, PAY_2, PAY_4, PAY_6, BILL_AMT2, BILL_AMT5, PAY_AMT1, PAY_AMT2, PAY_AMT3, and PAY_AMT4 ". Other details of this result include current mean fitness value that is 0.8052 whereas the maximum fitness value is 0.8096 and the minimum fitness value found was 0.7788. The term Fitness is the measure of the degree of adaptation of an attribute to its environment.

The attribute that holds bigger fitness value is more likely to be selected for recombination and these attributes have more chances that they will get adapted to the environment.

Next, we use the "Multilayer Perceptron" classifier and perform the evolutionary search again and this time we get 16 attributes. These attributes are "ID, LIMIT_BALANCE, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, BILL_AMT2, BILL_AMT6, PAY_AMT4, PAY_AMT5 and PAY_AMT6". Whereas the fitness values in this case, we receive, are as, follows mean fitness value stands at 0.8161 while the maximum fitness value stands at figure 0.8229 and the minimum fitness value is 0.7788. Moving further with evolutionary search method in next turn, we use the "KStar" Classifier and the results we get from this was quite different from the results which we get from previous two classifiers. This time we get only three attributes which are "LIMIT_BALANCE, BILL_AMT2 and BILL_AMT6." However, if we note down the fitness values,we find that mean fitness value, in this case, is -0.0004 while the maximum fitness value is -0 and minimum fitness value is -0.4197. This means all three fitness values are negative. The results of this classifier are much below than other two.

Fitness value plays an important role in the case of evolutionary search method, and selection of attributes depends upon these fitness values. Moving further if analyze the results of three classifiers in the case of evolutionary search method of feature selection we find that "Multilayer Perceptron" provides the best combination of attributes as it shows better mean and maximum fitness values which are 0.8161 for mean fitness and 0.8229 for maximum fitness.

Whereas "SMO" is second best classifier fitness values in this case are as follow. Mean fitness value was 0.8052 and maximum fitness value was 0.8096. At the last spot we have "KStar" classifier because all the fitness values are negative or zero in this case. Therefore, after this analysis, we find that best combination of attributes from all approaches, include following attributes "AGE, MARRIAGE, SEX and EDUCATION" along with "PREVIOUS PAYING HISTORY" and "AMOUNT OF PREVIOUS PAYMENT" as these attributes are common in the results of each approach. Therefore, in the classification phase, we will have more focus on these attributes.

4.2 CLASSIFICATION

After finishing the feature selection phase, we move toward the classification phase. In this phase, we mainly focus on those attributes that we find most useful through the feature selection process and use all three classifiers on the dataset to obtain the results of this phase we discuss the results of all three classifiers one by one as given below.

4.2.1 MULTILAYER PERCEPTRON

We start this process of classification by using a multilayer perceptron first as we find it the most successful classifier in the feature selection. First, we perform the classification of the complete dataset that includes 25 attributes in it and uses "DEFAULT PAYMENT NEXT MONTH" as a class attribute. The class attribute is a type of the attribute whose value we want to predict by using other attributes in the classification process. Through this first run of classification, in which we use 10-fold cross-validation as test mode, which remains the same throughout the process of classification. We find the percentage of 81.7933 % for correctly classified instances whereas for incorrectly classified instances we get the 18.2067 %. We continue this process of classification with the same classifier and other settings also remain the same in the next run, but this time we skip those attribute that represents the "PREVIOUS PAYING HISTORY" of the customer using same classifier and same test mode now percentage of correctly classified instances drop to 77.88 %. Whereas the percentage of incorrectly classified instances moves upward and reach 22.12 %. The clearly visible difference of results in two cases shows that "PREVIOUS PAYING HISTORY" holds an important role in predicting defaulters as it causes a decline of 3.9133 % in results so we marked it as an important attribute. Moving further in the next run, we only skip the attribute "AGE" and again run the model using the same setting. This time percentage of correctly classified instances were 81.7867% and for incorrectly classified instances, it is 18.2133% and through this, we find that without attribute "AGE" percentage of correctly classified instances are decreased by the margin of 0.0066 %, which signifies the impact of the attribute "AGE" on the result. On the next run now, we remove the attribute "EDUCATION" and obtain 81.81 % for correctly classified instances and 18.19 % for incorrectly classified instances. That means "EDUCATION" also influences the result, as the percentage of correctly classified instances is decreased by 0.0167%. Moving further this time, we remove the attribute "MARRIAGE" from the dataset and obtain results, are as, follows. The percentage of correctly classified instances were 81.7267%. While the percentage of incorrectly classified instances was 18.2733 % from these figures we clearly observe a decline of 0.0666 % in the percentage of correctly classified instances. Whereas when we use the dataset without the "SEX" attribute, then we get 81.7067 % for correctly classified instances and 18.2933% for incorrectly classified instances and in this case, the percentage of decline was 0.0866 %. On the last run, we check the "AMOUNT OF PREVIOUS PAYMENT" attribute we perform classification after removing it from the data set. This time percentage of correctly classified instances were 81.7333 % and for incorrectly classified instances, it was 18.2667 %. Now in this case percentage of correctly classified instances reduces by 0.06 % from the percentage that we get when we use a dataset with all attributes.

Table 1: Classification result using multilayer perceptron classifier

Name of Attribute	Impact on Result
Previous Paying History	3.9133%
Sex	0.0866%
Marriage	0.0666%
Amount of Previous Payment	0.06%
Education	0.0167%
Age	0.0066%

From the table 1 above, we find that “PREVIOUS PAYING HISTORY” attribute is the most impact-full attribute, which is followed by attribute “SEX” and at third spot attribute “MARRIAGE” and 4th and 5th spots is held by “AMOUNT OF PREVIOUS PAYMENT” and “EDUCATION” and attribute “AGE” is at 6th spot. In last 2 final runs of classification using “Multilayer Perceptron” classifier, in first case we only include these six attributes and we get a figure of 81.96 % for correctly classified instances and 18.04 % for incorrectly classified instances this is the best classification figure we get with this classifier and when we perform classification without these six attributes, the percentage for correctly classified instances decline reach to 77.88% and the percentage for incorrectly classified instances move up at 22.12%. In these last 2 final runs, we achieve the best and worst values of classification with this classifier and these figures show the importance of these six attributes.

After finding the effect of each parameter on the percentage of accuracy, next we move forward to find further details of these parameters on other important figures that we obtained from the results.

Kappa Statistics

From the definition of Kappa statistics kappa value between 0.01 – 0.20 is considered as none to slight while from 0.21 – 0.40 we consider fair and if this value ranges from 0.41 – 0.60 we call it moderate and if it’s between 0.61 – 0.80 the agreement is called as substantial and perfect agreement exist if value is between 0.81 – 1.00.

In our research, we obtain the value of kappa statistics in the best case of classification when we use only six optimal parameters was 0.3584 which is a fair agreement while in the worst case, when we exclude six optimal parameters from the dataset it was 0 that means no agreement. From this result, our finding of optimal attributes again proves right as we see kappa value move up to 0.3584 from zero, which shows that six attributes which we find optimal showing their effect on the kappa value as well.

Precision, Recall And F Measure

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0	0	?	0	?	?	0.624	0.31	1
	1	1	0.779	1	0.876	?	0.624	0.845	0
Weighted Avg.	0.779	0.779	?	0.779	?	?	0.624	0.727	
Confusion Matrix									
	a	b	<--Classfied as						
	0	6636	a = 1						

	0	23364	b = 0						

Table 2: Detailed result of classification without six optimal attributes using multilayer perceptron

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.335	0.043	0.69	0.335	0.451	0.392	0.731	0.515	1
	0.957	0.665	0.835	0.957	0.892	0.392	0.731	0.879	0
Weighted Avg.	0.82	0.527	0.803	0.82	0.795	0.392	0.731	0.798	
Confusion Matrix									
	a	b	<--Classfied as						
	2225	4411	a = 1						
	1001	22363	b = 0						

Table 3: Detailed results of classification with only six optimal attributes using multilayer perceptron

The above, Table 2 and 3 show the detailed result of classification in two different cases. The detailed in Table 2 was in the case when we do not include six optimal attributes in the data set and the next detailed shown in Table 3 was in the case when we include only these six optimal attributes in classification. The difference in results in the two cases is clearly visible. In the first case, the weighted average precision value is unknown whereas the recall value is 0.799 and the value of F-measure is not calculated. Whereas when we use only six optimal attributes, then these values get improved. Where the weighted average precision value stands at 0.803 and the value of recall is 0.820 while the value of F-measure is 0.795. Investigating further, we also see visible improvement in the values of TP Rate, which increase from 0.779 in the first case to 0.820 in the second case. The value of FP Rate declines from 0.799 to 0.527. TP Rate means the true positive rate or sensitivity and FP Rate means the false positive rate or fall out. The result of these two cases clearly showing that six identified optimal parameters show their impact on the results.

4.2.2 SEQUENTIAL MINIMAL OPTIMIZATION (SMO)

After completing the classification task with a multilayer perceptron, we move to the next classifier that “SMO” which we find as second best through the process of feature selection. Further, in this section, we discuss the results that we find in the classification phase using “SMO.”First, we perform classification using “SMO” with the complete dataset and in this case, the results are as follows. The percentage of correctly classified instances were 80.9267 whereas the percentage of incorrectly classified instances were 19.0733. In the next step, we start testing the effect of different optimal attributes on the result of classification using the “SMO” classifier. We start this process from the attributes that represent the “PREVIOUS PAYING HISTORY”. We remove these attributes from the dataset, and then perform the classification phase. The results that we obtained, in this case, are as follows. The percentage of correctly classified instances was 77.88 and the percentage of incorrectly classified instances was 22.12. From these results, we conclude that "PREVIOUS PAYING HISTORY" has a significant effect on classification results, and we measure a decline of 3.0467% in the percentage of correctly classified instances whereas we also observe a rise of 3.0467 in the percentage of incorrectly classified instances. Moving further, next, we test the attributes that represent "AMOUNT OF PREVIOUS PAYMENT" when we perform classification without these attributes the percentage of correctly classified instances was 80.9233% and the percentage of incorrectly classified instance was 19.0767%. Therefore, attributes that represent "AMOUNT OF PREVIOUS PAYMENT" do have little impact on the result and the percentage of correctly classified instances declined by 0.0034%. While moving forward now we perform classification without attribute "EDUCATION". This time percentage of correctly classified instances were 80.92%, whereas the percentage of incorrectly classified instances were 19.08% and from this, we figure out that there is a slight decline of 0.0067% in the percentage of the correctly classified instance as compared to the classification result, which we obtained while using complete dataset. We continue this process of classification in the next turn we exclude the "MARRIAGE" attribute from the dataset and run the classifier again. This time percentage of correctly classified instances were 80.9267%, while the percentage of incorrectly classified instances were 19.0733%, which means that attribute "MARRIAGE" left no effect on the result of classification. In order to find the effect of the attribute "AGE" we perform classification by excluding the attribute "AGE" through this run we find that percentage of correctly classified instances was again at 80.9267 %, while for incorrectly classified instances, it stands at 19.0733 % which mean attribute "AGE" has no effect on the result of classification. In the last round of classification, we exclude the attribute "SEX" from the dataset and this time percentage of correctly classified instances was 80.93% while the percentage of incorrectly classified instances was 19.07%. Through this run of classification, we find an unexpected result. Attribute "SEX" has affected the result negatively instead of decreasing the percentage of correctly classified instances increase by the margin of 0.0033%, while in all previous cases we observe that when we exclude an attribute from the data set percentage of correctly classified instances get decrease but not in this case.

Table 4: Classification result using SMO classifier

Name of Attribute	Impact on Results
Previous Paying History	3.0467%
Education	0.0067%
Amount of Previous Payment	0.0034%
Age	0.0%
Marriage	0.0%
Sex	- 0.0033%

From the table 4 above, we find that the "PREVIOUS PAYING HISTORY" attribute is the most impact-full attribute, which is fallow by attribute "EDUCATION" whereas attribute "AMOUNT OF PREVIOUS PAYMENT" was third most impactful attributes. While at 4th and

5th spots we find attributes “AGE” and “MARRIAGE” and at 6th spot we have the attribute “SEX”. However, we also perform 2 more runs of classification using the “SMO” classifier, in the first case we only include these six attributes in the dataset while in the second case we exclude these six attributes from the dataset. When we exclude these six impactful attributes from dataset percentage for correctly classified instances declines and reach 77.88%, which is a worst-case result. Whereas when we include only these six attributes in the dataset, then the percentage of correctly classified instances were 80.9233%.

Kappa Statistics

Kappa value is an important figure to find out the performance of any classifier. When we exclude six important attributes from the dataset, kappa value, in this case, stands at 0 which points out no agreement state while in another case when we only include six important attributes in the dataset then kappa value was 0.2731 which shows that there is a fair agreement. Therefore, the difference of kappa value in these two cases also proves the importance and impact of these six attributes on kappa value and overall results as well.

Precision, Recall and F measure

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0	0	?	0	?	?	0.5	0.221	1
	1	1	0.779	1	0.876	?	0.5	0.779	0
Weighted Avg.	0.779	0.779	?	0.779	?	?	0.5	0.655	
Confusion Matrix									
	a	b	<--Classfied as						
	0	6636	a = 1						
	0	23364	b = 0						

Table 5: Detailed Result of classification without six optimal attributes using SMO

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.236	0.028	0.706	0.236	0.354	0.33	0.604	0.336	1
	0.972	0.764	0.818	0.972	0.888	0.33	0.604	0.816	0
Weighted Avg.	0.809	0.601	0.793	0.809	0.77	0.33	0.604	0.71	
Confusion Matrix									
	a	b	<--Classfied as						
	1566	5070	a = 1						
	653	22711	b = 0						

Table 6: Detailed Results of classification with six optimal attributes using SMO

The above, Table 5 and 6 shows the detailed result of classification in two different cases. The detailed in Table 5 was in the case when we do not include six optimal attributes in the data set and the next detailed shown in Table 6 was in the case when we include only these six optimal attributes in classification. The difference in results in the two cases is clearly visible. In the first case, the weighted average precision value is unknown whereas the recall value is 0.779 and the value of F-measure is not calculated. Whereas when we use only six optimal attributes, then these values get improved. Where the weighted average precision value stands at 0.793 and the value of recall is 0.809 while the value of F-measure is 0.770. Investigating further, we also see visible improvement in the values of TP Rate, which increase from 0.779 in the first case to 0.809 in the second case. The value of FP Rate declines from 0.799 to 0.601. TP Rate means the true positive rate or sensitivity and FP Rate means the false positive rate or fall out. From the analysis of the results of these two cases, we clearly observe that six identified optimal parameters improve results when they are used for classification.

5 Kstar

While repeating the process of classification now in this turn, we select Kstar Classifier to perform classification. First, we perform the classification of the complete dataset that includes 25 attributes in it. Through this, in the first run of classification, we find the percentage of 57.98 % for correctly classified instances whereas for incorrectly classified instances we get the 42.02 % with these figures we find that Kstar is the worst performer out of three classifiers that we used. Moving forward, we continue this process of classification with Kstar classifier and other settings also remain the same in the next run, we skip those attribute that represents the “PREVIOUS PAYING HISTORY” of the customer using same classifier and same test mode now percentage of correctly classified instances drop to 57.3733%. However, the percentage of incorrectly classified instances moves upward and reaches 42.6267 %. The difference of results in two cases shows that “PREVIOUS PAYING HISTORY” holds an important role in predicting defaulters as it causes a decline of 0.6067 % in results. Moving further in the next run of classification, this time we skip the attribute “AGE” and again run the model using the same setting. This time percentage of correctly classified instances was 57.9867% and for incorrectly classified instances, it was 42.0133 %. This means that attribute “AGE” affects the process of prediction negatively as a percentage of correctly classified instances is increased by the margin of 0.0057 %. In the next run of classification, we remove the attribute “EDUCATION” and obtain 57.9133 % for correctly classified instances and 42.0867 % for incorrectly classified instances. That means “EDUCATION” influences the result as the percentage of correctly classified instances is decreased by 0.0667 %. Moving further and we continue the process of classification and this time we remove the attribute “MARRIAGE” from the dataset and obtain results as follows. The percentage of correctly classified instances was 57.9167%, While the percentage of incorrectly classified instances was 42.0833 %.

From these results we clearly observe a decline of 0.0633 in the percentage of correctly classified instances. Whereas when we test the dataset without the “SEX” attribute, then we get 57.9733 % for correctly classified instances and 42.0267 % for incorrectly classified instances and in this case, there was a decline of 0.0067 %. On the last run, we check the “AMOUNT OF PREVIOUS PAYMENT” attribute we perform classification after removing it from the data set. This time percentage of correctly classified instances were 63.7033 % and for incorrectly classified instances, it was 36.2967 %. Now in this case percentage of correctly classified instances was increased by 5.723 percent from the figure that we get when we use a dataset with all attributes.

Table 7: Classification result using Kstar classifier

Name of Attribute	Impact on Result
Previous Paying History	0.6067 %
Education	0.0667 %
Marriage	0.0633 %
Sex	0.0067 %
Age	-0.0057 %
Amount of Previous Payment	-5.723 %

From the table 7 above, we find that “PREVIOUS PAYING HISTORY” attribute is the most impact-full attribute, which is fallow by attribute “EDUCATION” and at third spot attribute “MARRIAGE” and 4th and 5th spots is held by “SEX” and “AGE” and at the last spot we have attribute “AMOUNT OF PREVIOUS PAYMENT”. One notable point here is that attribute “AGE and AMOUNT OF PREVIOUS PAYMENT” shows a negative impact on the result which means that when we excluded these two attributes from the dataset the percentage of classification increases. After this we perform 2 final runs of classification using “Kstar” classifier, first we only include these six attributes in the dataset and we get a figure of 72.35 % for correctly classified instances and 27,65 % of incorrectly classified instances and when we perform classification without these six attributes, the percentage of correctly classified instances decline reached to 62.5433% and the percentage of incorrectly classified instances move up at 37.4567%. After finding the effect of each parameter on the percentage of accuracy, next we move forward to find further details of these parameters on other important figures that we obtained from the results.

Kappa Statistics

In case of Kstar, we obtain the value of kappa statistics as follows when we use only six optimal parameter’s value of kappa statistics was 0.1612 which is considered as none or slight agreement while when we exclude six optimal parameters from dataset kappa statistics value decreases further and stand at 0.0442 which again shows none or slight agreement. From this result, our finding of optimal attributes again proves right as we see kappa value was better when we use only six optimal attributes and the kappa value decreases further when we exclude these six attributes from the dataset.

Precision, Recall and F measure

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class

	0.345	0.295	0.249	0.345	0.29	0.045	0.577	0.259	1
	0.705	0.655	0.791	0.705	0.746	0.045	0.569	0.814	0
Weighted Avg.	0.625	0.575	0.671	0.625	0.645	0.045	0.57	0.691	
Confusion Matrix									
	a	b	<--Classfied as						
	2291	4345	a = 1						
	6892	16472	b = 0						

Table 8: Detailed result of classification without six optimal attributes using Kstar

Detailed Accuracy By Class									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.315	0.16	0.358	0.315	0.335	0.162	0.634	0.314	1
	0.84	0.685	0.812	0.84	0.825	0.162	0.634	0.846	0
Weighted Avg.	0.724	0.569	0.711	0.724	0.717	0.162	0.634	0.728	
Confusion Matrix									
	a	b	<--Classfied as						
	2088	4548	a = 1						
	3747	19617	b = 0						

Table 9: Detailed result of classification with six optimal attributes using Kstar

The above, Table 8 and 9 shows the detailed result of classification in two different cases. The detailed in Table 8 was in the case when we do not include six optimal attributes in the data set and the next detailed shown in Table 9 was in the case when we include only these six optimal attributes in classification. The difference in results in the two cases is clearly visible. In the first case, the weighted average precision value is 0.671 whereas the recall value is 0.625 and the value of F-measure is 0.645. However,in another case when we use only six optimal attributes,then these values get improved. Where the weighted average precision value stands at 0.711 and the value of recall is 0.724 while the value of F-measure is 0.717. Investigating further, we also see visible improvement in the values of TP Rate, which increases from 0.625 in the first case to 0.724 in the second case. The value of FP Rate declines from 0.575 to 0.569. TP Rate means the true positive rate or sensitivity and FP Rate means the false positive rate or fall out. From the analysis of the results of these two cases, we clearly observe that six identified optimal parameters improve results when they are part of the dataset used for classification.

5. DISCUSSION

After completing the data-preprocessing phase, we formally started the research work and move towards first step that was feature selection for which we use two different methods in the first method we perform feature selection using a ranker approach in which we use the gain ratio as an evaluator. Through this, we find that attribute “PREVIOUS PAYING HISTORY” has the highest gain ratio and placed on top of the list while attribute “AMOUNT OF BILL STATEMENT” was at the bottom of this list with zero gain ratio. While continuing the process of feature selection this time we use an evolutionary search method, which was used with each classifier differently. We find multilayer perceptron as the most successful classifier as this classifier shows better fitness values (Max-Min Mean) whereas Kstar was the least performer of this method. After achieving the results of feature selection methods, we analyze these results and point out six attributes that are likely to be the optimal attributes (PREVIOUS PAYING HISTORY, AMOUNT OF PREVIOUS PAYMENT, AGE, SEX, MARRIAGE, and EDUCATION). Now we move towards classification phase in which we perform 8-10 runs of classification for each classifier in these runs of classification each time we make changes to the dataset. First, we use the complete dataset in classification and then in remaining runs we individually exclude each identified attribute (through feature selection) one by one to find out the impact of each attribute on the results. We also check the combined effect of these six identified attributes on the results and what impact they left if these attributes are not part of the dataset. During these tasks of classification, we also obtained some unexpected results as well. Attribute “SEX” in case of SMO classifier shows a negative impact on the result, which mean when we exclude this attribute classifier shows better results. A similar situation happens with attributes “AGE and AMOUNT OF PREVIOUS PAYMENTS” in the case of Kstar classifier, which shows a negative impact on, results as well. Whereas from table no 4-1, 4-2, 4-3 we observed that attribute “PREVIOUS PAYING HISTORY” was the most impact-full attribute under all three classifiers whereas other attributes shows randomly varying performance under these three classifiers.

Overall, these six attributes individually and in combination affect correctly classified instances percentage. We do check the validity of these six attributes through the values of other key indicators that we obtained along these results which include Kappa Statistics, Precision, and Recall and results of these figures proved that these six attributes are optimal attributes and they not affect the percentage of correctly classified instances but also shows their effect on the values of these key indicators and ultimately affect the results of predicting credit card defaulters. The detailed output of this research work along with the numerical values of each step described above in the results section.

We also analyze the performance of classifiers. We use three classifiers out of them multilayer perceptron was ranked as the best performer based on the results we achieved through this classifier in feature selection multilayer perceptron provide the best fitness values. After multilayer perceptron, we have SMO and Kstar shows the least performance in feature selection. The multilayer perceptron also holds the top spot in the classification phase as it provides the highest possible percentage of correctly classified instances.

SMO remains the second-best and Kstar again shows the least performance. From this analysis, we can say that the Multilayer perceptron was the most efficient and accurate classifier as compared to the other two that we used in this analysis. Table 4 below shows the detailed analysis of three classifiers along with the values of key figures.

Table 10: Performance analysis of Classifiers

Classifiers	**Max Fitness Value**	**Min Fitness Value**	**Highest Classification Accuracy**	**Lowest Classification Accuracy**
Multilayer Preceptron	0.8229	0.7788	81.96%	77.88%
SMO	0.8096	0.7788	80.9267%	77.88%
Kstar	-0	-0.4197	72.35%	57.98%

Table 10 below shows the comparison of our research methodology with the techniques used in previous research studies. It is clearly visible from the table that Multilayer Preceptron classifier shows better results as compared to classifiers used in this research and it also shows better results from other classifiers that are used in previous research studies.

Table 11: Performance Comparison of Proposed Methodology with previous Techniques

Proposed Research Methodology				
Classification Algorithm	**Accuracy**	**Recall**	**Precision**	**TP**
Multilayer Preceptron (MLP)	**81.96**	**0.820**	**0.803**	**0.820**
Sequential Minimal Optimization (SMO)	80.9233	0.793	0.809	0.809
Kstar	72.35	0.711	0.724	0.724
[4]				
Simple Cart	68.93	0.689	0.689	
J48	72.82	0.728	0.733	
Random Tree	67.96	0.68	0.668	
PART	73.79	0.738	0.762	
NB Tree	69.9	0.699	0.697	
Fuzzy	78.64	0.786	0.783	
[19]				
Bayes Net		0.796	0.781	0.796
Stacking		0.779	0.607	0.779
Navies Bayes		0.805	0.786	0.805
Random Forest		0.816	0.798	0.816
Random Tree		0.816	0.797	0.816
ZeroR		0.779	0.607	0.779
IBK		0.816	0.798	0.816
SMO		0.809	0.793	0.809
[13]				
FLDA	72.4	0.724	0.769	
J48	80.3	0.803	0.782	
Logistic Regression	81	0.81	0.795	
Naive Bayes	69.4	0.694	0.77	
Multilayer Preceptron (MLP)	81.7	0.817	0.799	
IBK	72.9	0.729	0.73	

6. CONCLUSIONS

In this research work, we have performed bank data analysis. The focus of this analysis is to identify the optimal parameters, which are helpful for the prediction of credit card defaulters. We have used a credit cards defaults dataset. In this research work, we use machine learning techniques and classifiers. First, we perform feature selection with two different methods and analyze these results and from this analysis, we point out some important parameters and focus on them in the classification phase, as these parameters are likely to be optimal. In the next phase, we perform classification using three different classifiers. In the classification phase, we find out the impact of each parameter, on overall results. We also validate the impact of each parameter through the result of some key figures. These figures include Kappa Statistics, precision, recall and F measure. In classification, we test each parameter individually and do check the combined impact of these parameters on the results as well. We do not put a full stop only at these results in this research we also analyze the performance of each classifier and find out which classifier performs best and validate these findings using key figures. Through this research, we find multilayer perceptron as the best classifier.

7. FUTUREWORK

In this research work, we utilized machine learning approaches to make a prediction about the identification of credit card defaulters by identifying optimal parameters. Machine learning is spread in every life field, it was used to make future predictions. Therefore, this study was really helpful for other fields as well. Future predictions are based on previous records and therefore identification of optimal parameters is important. Because with the help of these parameters we can improve prediction accuracy and we achieve these results in less time. Therefore, we can apply the methodology of this research work on datasets of other fields like the medical field where we can use the previous history of patients to make future predictions. Therefore, if we identify optimal parameters than we improve the results of future prediction, which can, prove life-saving for many patients. Similarly, We can also apply this methodology with other machine learning classifiers and feature selection techniques as well.

REFERRENCES

1. Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1), 0–15. https://doi.org/10.1088/1742-6596/1142/1/012012

2. Bhatla, T. P., & Prabhu, V. (2003). Understanding Credit Card Frauds, (June).
3. Gogas, P., Papadimitriou, T., & Agrapetidou, A. (2018). Forecasting bank failures and stress testing : A machine learning approach. *International Journal of Forecasting*, *34*(3), 440–455. https://doi.org/10.1016/j.ijforecast.2018.01.009
4. Gulsoy, N., & Kulluk, S. (2019). A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers, (November 2018), 1–12. https://doi.org/10.1002/widm.1299
5. Hassani, H., Huang, X., & Silva, E. (2018). Digitalisation and big data mining in banking. *Big Data and Cognitive Computing*, *2*(3), 1–13. https://doi.org/10.3390/bdcc2030018
6. Keramati, A., & Yousefi, N. (2011). A Proposed Classification of Data Mining Techniques in Credit Scoring, 416–424.
7. Kotsiantis, S. B. (2007). Supervised

Machine Learning : A Review of Classification Techniques, *31*, 249–268.

8. Lee, S. J., & Siau, K. (2001). A review of data mining techniques. *Industrial Management and Data Systems*, *101*(1), 41–46. https://doi.org/10.1108/02635570110365989

9. Mallqui, D. C. A., & Fernandes, R. A. S. (2019). Predicting the direction , maximum , minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. *Applied Soft Computing Journal*, *75*, 596–606. https://doi.org/10.1016/j.asoc.2018.11.038

10. Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia - Procedia Computer Science*, *91*(Itqm), 919–926. https://doi.org/10.1016/j.procs.2016.07.111

11. Mohammed EL Hassan, E. (2014). *College of Graduate Studies Credit Scoring Using Data Mining Classification : Application on Sudanese Banks Supervisor : Prof . Izzeldin Mohammed Osman*.

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

13. Pasha, M., Fatima, M., Dogar, A. M., & Shahzad, F. (2017). Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters, *17*(3), 178–183.

14. Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *6*(2).

15. Shoumo, S. Z. H., Dhruba, M. I. M., Hossain, S., Ghani, N. H., Arif, H., & Islam, S. (2019). Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, *2019-October*, 2023–2028. https://doi.org/10.1109/TENCON.2019.8929527

16. Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021). Prediction of Modernized Loan Approval System Based on Machine Learning Approach. *2021 International Conference on Intelligent Technologies, CONIT 2021*, 21–24. https://doi.org/10.1109/CONIT51480.2021.9498475

17. Tanouz, D., Subramanian, R. R., Eswar, D., Reddy, G. V. P., Kumar, A. R., & Praneeth, C. H. V. N. M. (2021). Credit card fraud detection using machine learning. *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, 967–972. https://doi.org/10.1109/ICICCS51141.2021.9432308

18. Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) : proceedings : March 20-21, 2019, Jahorina, East Sarajevo, Republic of Srpska, Bosnia and Herzegovina. *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, (March), 1–5.

19. Venkatesh, A., & Jacob, S. G. (2016). Prediction of Credit-Card Defaulters : A Comparative Study on Performance of Classifiers, *145*(7), 36–41.

20. Zhou, X., Cheng, S., Zhu, M., Guo, C., Zhou, S., Xu, P., … Zhang, W. (2018). A state of the art survey of data mining-based fraud detection and credit scoring. *MATEC Web of Conferences*, *189*. https://doi.org/10.1051/matecconf/201818903002