Applications of Machine Learning in Banking Services

Akshen Doke
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x18191592@student.ncirl.ie

Abstract- **Banking sector has always been generating a lot of data in its day to day operations, now with technology being integrated everywhere, the amount of transactions happening has also increased a lot in banking services. In this paper we would be following Knowledge Discovery in Database(KDD hereafter) methodology and see how machine learning techniques can be applied on three different datasets related to banking services to identify credit card based fraud transactions, classify whether the customer has taken personal loan based on certain parameters and check if a customer is willing to take loan or not based on the marketing campaign executed by the bank. Machine Learning models based on algorithms like Support Vector Machine, Naïve Bayes and Random Forest were used during this research and their performance was evaluated on metrics like confusion matrix, ml_test(package for bunch of metrics), accuracy on all 3 datasets, which also had some class imbalance which was taken care of by using down sampling, oversampling and SMOTE technique. The resulting models were accurate and reliable enough, future work includes deployment of these models by banking firms in a simulated environment to check and optimize their crackdown on detection of unwanted fraud credit card transactions and target customers who will fulfil specific criteria while lending loan thus minimizing loss and increasing overall profitability of the bank.**

Keywords—Data Mining, Machine Learning, Fraud and Risk Detection, class imbalance, loan defaults

I. Introduction

To any nation be it developed or developing, a healthy banking system is an indication of an healthy and reliable economy but sometimes due to fraud, loan defaults and less lending, banks suffer tremendous loss which not only affects that particular banking firm but also takes away the trust from banking industry as a whole thus impacting the growth of the country from where the bank is and effecting it's economy. As per [1] China had to blacklist accounts of more than 9 million customers for defaulting on their loan payments and freezing up to 28 billion USD deposits to secure the assets for the banks. Such issues have been happening worldwide as many security patches lies unfilled. Also, loans provided to customers without much financial background checks have resulted into global financial meltdown in the past. The 2008 US mortgage housing crisis which lead to recession is the prime example of situations that if things left uncheck and unevaluated might result into a disaster. Along with detecting loan default, we also need to identify potential customers who can apply for a loan and have the capacity to pay it in time. This can be done by analyzing the data gathered during the marketing of banking services which will help banks target new potential customers for their loan or mortgage services.

Another popular banking service called Credit Card is where probability of fraudulent transactions taking place is seen, even though the number of such transactions occurring in the dataset we have are very less compare to genuine ones but the impact such transactions leave due to the amount withdrawals from such a transaction is high and enough to blow a dent in the banks treasury. Hence detecting and pausing such transactions can save banks from incurring losses. In today's times, Modern Computers have enormous processing speed and relatively huge storage capacity to gather, analyze and derive outputs based on the data collected, hence statistical principles can be applied to get deep insights of the data and thus solve or detect problems and issues which were previously unthought of or went unseen due to limited data processing capabilities. For the datasets we have we will be using a computer principle called Machine Learning in which algorithms based on statistics and probability are used to make a computer predict or classify input data based on the historic behavior and performance of the dataset similar in nature to that of the input data. Machine Learning improves our ability to perform the task we desire as it's based on complex and reliable algorithms which are designed to do the task given to it in an optimized manner. It has countless applications in almost every field where in human redundant effort is required but the advantage of machine learning models is that it is more reliable as it's not prone to human errors like inconsistency and gives same output 24*7 provided a static nature. New machine learning models based on the dynamic principles of the nature i.e. change is being designed but they can't be trusted completely because it

[1] P. T. o. India, "Hindustantimes," Hindustan Times, 13 January 2018. [Online]. Available: https://www.hindustantimes.com/world-news/9-million-loan-defaulters-blacklisted-in-china-27-billion-frozen/story-Z9KBvxffasUwsD8LRLjpPJ.html. [Accessed 11 August 2020].

is not certain that whether they would act in the wellbeing of human society or not.

In this project we have made use of various machine learning algorithms like Decision Tree, Random Forest and clustering principles like K-Means to classify our data based on certain parameters present in our dataset. The resultant model based on the training of such data can be used to deploy on the backends of the bank to detect fraudulent credit card transaction, identify potential customers for mortgages and predict loan defaults of an existing customer.

II. RELATED WORKS

A lot of previous work has been done using machine learning in the domain of banking, in this section we would explore some of the papers referred while conducting the mentioned project. Credit card transaction fraud has been studied by many researchers, [1] proposed a novel model by combining two approaches of machine learning one which is based on neural networks called auto-encoders and the other based on support vector machine principle called one-class support vector machine (OSVM). Here the researchers first ran the dataset of credit card transactions separately on each model and then combined the two and performed the training again. The researchers used various evaluation methodologies to measure the performance of their proposed method and compared them with other methods as well, they claim that their proposed model performed better in comparison to other models which operated single handedly without any combination or tuning whatsoever but one thing worth mentioning here is that the researchers used a quite imbalanced dataset to begin with and also did not used any example of fraud transactions as a training parameter on their proposed model. The results obtained were satisfactory as per claims but since the approach used to make these claims are somewhat unconventional, it still has some gaps to fill.

[2] also worked on credit card fraud detection addressing the fact of class imbalance, initially they divided the data as per transaction amount ranging from high to medium and low. Later they made use of a technique called sliding window to extract some features to determine the behavioral patterns of each card-owner followed by calculating the mean and each time when a new transaction occurred, they popped the old one from the window. After this process they trained various classifiers on each group and extracted the features which indicated the fraudulent transaction. Due to class imbalance their classifiers didn't perform optimally hence they used a class balancing technique called Synthetic Minority Over-Sampling Technique aka SMOTE and then they plotted the performance of the models before applying SMOTE and after to prove that they got better performance compare to the previous ones. This result is a motivation for us in this project to use SMOTE as a technique to handle class imbalance. The novel approach they proposed made use of a feedback option called concept-drift where in the change of parameters in each transaction is taken into consideration. Finally, they also made use of conventional algorithms like Logistic regression, Decision Tree and got better results. Use of Artificial Neural Networks was mentioned in [3] to support loan decision for Jordanian commercial banks however their findings said that use of logistic regression based model was better compare to radial basis function used in ANN. [4] [5] [6] made use of Artificial Neural Network technology to determine whether to accept or reject credit applications, to evaluate credit giving risks and forecast earnings per share by comparing backward propagation a neural network architectural approach with genetic algorithms. [7] in there works on early fraud detection made use of a method called Support Vector Machine with Spark (SVM-S) that can create a model which can evaluate validity of upcoming transactions based on customer behavior, the result of their experiments showed that their technique of SVM-S had better performance overall in comparison to Back Propagation Networks. Some researchers also proposed a unique method for credit card fraud detection based on outlier detection which worked on the principle of distance sum in accordance to the infrequency and unconventionality of frauds in transactions [8].

For our next dataset which is dealing with loan defaults we can take inspiration from [9] who made use of machine learning models based on supervised learning approach to identify loan defaults. The dataset they used was obtained from a peer to peer lending platform based in the United States which has class imbalance to tackle this issue they made use of SMOTE technique but only on the training part of the dataset, the testing part was left untouched. They also made use of dimensionality reduction procedure on the parameters of the dataset to get better values which can be fed as input to Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection and Principle Component Analysis (PCA) for feature extraction were chosen. Feature scaling was also done where-ever needed. For classification, 4 supervised algorithms were used and their hyperparameters were tuned using GridSearchCV for each as per best fit for them. Also, a comparative study of models based on GridSearchCV and 5-fold cross-validation was performed which resulted into SVM based model performing better amongst all other models. Overall the authors did the required assessment of machine learning techniques and came out with their findings of SVM performing better in comparison to tree-based models. [10] worked on the data of Indian housing market which although was growing but also had many loan defaults, the author made use of an so-called HYBRID data mining model to predict the defaults using Classification and Regression Trees aka CART, five different training and testing models were built on these principles to ensure that they are reliable and reproducible. Another study based on principles of feature selection [11], MCDM [12] Independence component analysis [13] and combining feature selection methods with classifiers [14] was carried on by [15] on a China based commercial bank wherein the results they obtained showed K-NN is reliable in default prediction.

[16] conducted a study on Turkish Banks for detecting loan defaults by analyzing consumer behavior concerning bank loans. They ran 4 classification algorithms using 2 resampling techniques on their highly imbalanced dataset, as seen in other papers as well SMOTE was used for handling this issue however in their research they observed that even after using SMOTE the result they obtain was not up to the mark and they concluded their research by saying that the procedure of SMOTE misguided their classification algorithm and resulted in lower performance of when measured by metrics. Thus, the results from this study are interesting as it differs from the normal scenarios which we have seen so far.

Personal Bankruptcy has been an issue in lot of parts around the world, [17] tried to create a decision tree based model to predict personal bankruptcy for Malaysia based customers as they found that the number of bankruptcy rates in Malaysia from 2007 to 2014 were around 131, 282 which is an alarming number for a small country like Malaysia as such scenarios affect the overall countries economy and their ability to secure foreign investments and loans in there ventures. The results of their classification model based on decision tree were quite unsatisfactory earlier when they ran it over on an unbalanced dataset, later on they made use of random undersampling technique to correct this imbalanced and saw an improvement in specify rate. In future they expressed a will to implement SVM, LR and Naïve Bayes model on the same sampled dataset. In addition [18] made use of decision tree to make predictions of financial distress, [19] made use of both decision tree and logistic regression to build a prediction model for identifying corporate financial debts and non-performing assets. Meanwhile [20] reported that more than 20,000 cases of personal bankruptcy in Malaysians were declared by people aged between 25 to 34.

Conclusion

In the papers above we saw how researchers previously made use of machine learning models based on algorithms like Decision Tree, Random Forest, SVM and Neural Networks in banking services to assess and predict the risks involved for lending institutions to avoid loan defaults, frauds and identify potential customers correctly in order to minimize loss of a bank.

III. DATA MINING METHODOLOGY

For this study, we would be following the Knowledge Discovery in Databases (KDD hereafter) approach which is good for large datasets, pattern recognition and machine learning as per [2] the unifying objective of the whole process of KDD is to extract knowledge from the raw input data fed to it in form of large databases. Figure below gives a pictorial depiction of the same.

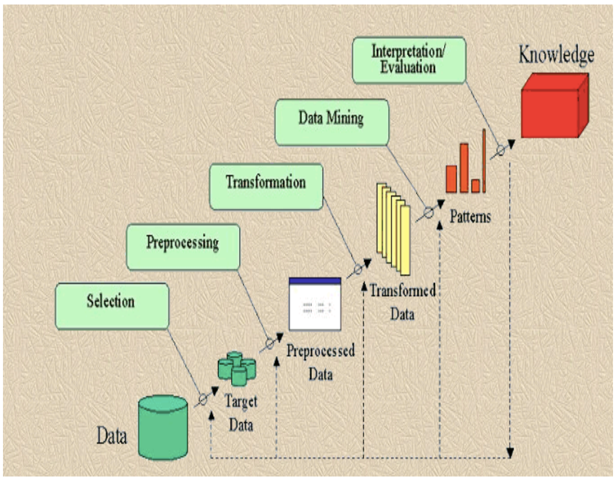

Figure 1: **KDD Process** Source: 2

Since the 3 datasets we used for this study are of banking sector but the nature of them is different as they are based on different banking services, we need to implement the above seen steps for all three of them.

Summary of Datasets

1. The data is related to direct marketing campaign majorly phone calls from a Portugal based banking firm.

   Number of Attributes: 17

   Number of Instances: 45,211

   Source: UCI Machine Learning Repository

[2] P.-S. S. U. Fayyad, "Overview of KDD Process," 1996. [Online]. Available: https://tinyurl.com/y4sr7bcn. [Accessed 12 August 2020]

2. This dataset is a collection of transactions made via credit card in September of 2013 by many Europeans cardholders.

   Number of Attributes: 31

   Number of Instances: 284,807

   Source: Kaggle

3. This dataset consists of responses of customers for personal loan selling queries, it contains information like customer demographics, customer's relationship with the lending institutes and their response to loan offers.

   Number of Attributes: 14

   Number of Instances: 5,000

   Source: Kaggle

First step in all 3 of our datasets is to load the data and check for any missing values, it's important to do this check as any inconsistency in our data might lead us to biased or unfair results. In this step of preprocessing if we find any missing values then we can replace it with either the mean of that column or depending upon the nature of values of that column or we can remove the complete row if it doesn't affect our model as a whole. We made use of a package called naniar which gives us a visual representation whether any of our value is missing or not also we plot the datatype of columns

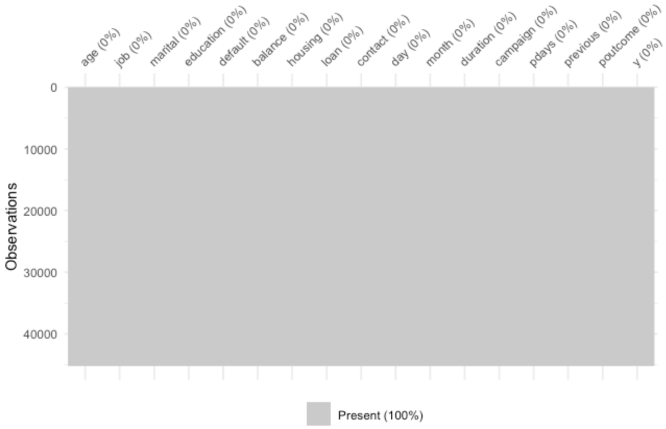

Figure 2: **Missing value visualization** Source: Rcode

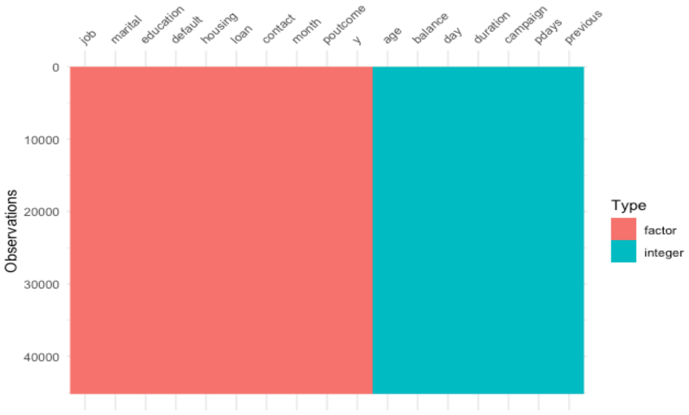

Figure 3: **Datatype plot** Source: Rcode

After this since we are doing a classification problem, we check whether if dependent variable is proper or not since we have to divide the complete dataset into two parts namely train and test set which are going to be feed to the model for training and evaluation, we have to make sure that the data isn't imbalanced. We do this by plotting the distribution of our dependent class variable and see the results.

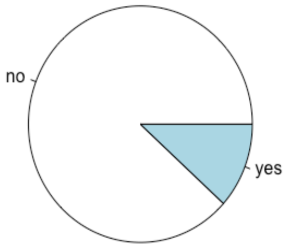

Figure 4: **Class Imbalance** Source: Rcode

To fix this imbalance issue firstly we have to convert the imbalanced class column to numeric and labelled them as '1' for no and '2' for yes, then to factor which is a R specific datatype after which we can either downSample or upSample our dataset. These functions are available in the caret library in R. In this case since we have frequency of 5,289 for the minority class from the total count of 45,211 we can down Sample the majority class and get to similar level and balance the distribution.

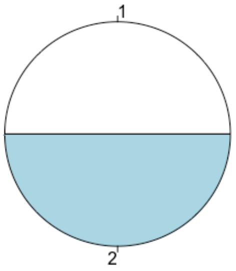

Figure 5: **Class balance** Source: Rcode

After performing the down Sampling process and fixing the class imbalance issue, our newly processed data is ready to be divided into train and test sets. We divide the data into 70 and 30 respectively. Next step is we train our model on the trainset and evaluate its performance on the test set of the divided data. To get better results from our models, we can only take in those inputs which actually affects the nature of our dependent variable.

A. *Feature Selection/Extraction*

In any machine learning model, the input parameters are one of the key factors after the modelling algorithms to be used, in case of our first dataset i.e. marketing related to Portuguese banking institutions it is clear from the names of the columns that which of them are going to affect the dependent variable. So, we structure the “formula” as it’s known in R and feed it to our model along with the train set from our original dataset. In the next dataset, which is about credit card transaction details, the data available from the source did not contain any column names to maintain privacy of the identity of the user and his/her data. In such scenarios in order to get the required features we can take the correlation of columns and see which ones are related the most to the dependent variable. As per Guilford, “A coefficient of correlation is a single number that tells us to what extent two things are related, to what extent variation in one go with variations in the other”[3]. Below given is the plot of correlation for credit card dataset, the density of color of the dot in the square indicates how much the column variable affects the row variable. Based on this plot we can only select the variables which affects our target variable the most and use them to build our model. This way the result which we get from our model would be reliable and can be trusted.

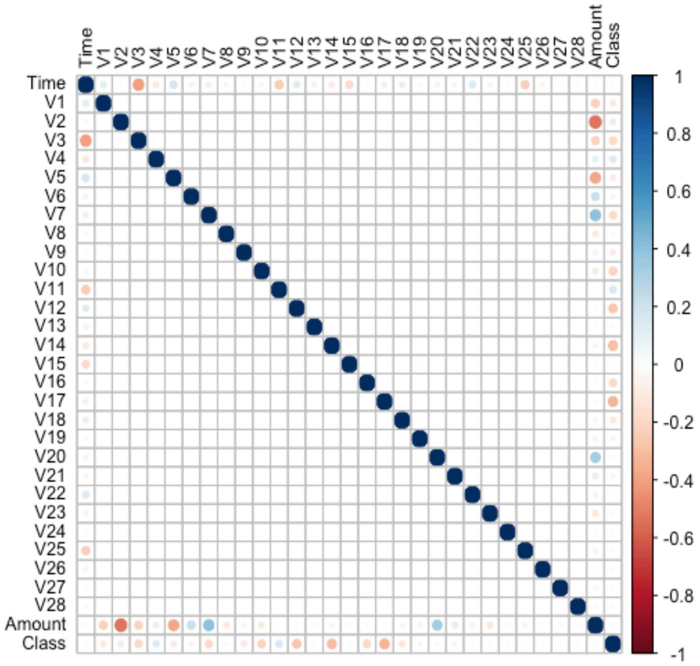

Figure 6: **Correlation Plot** Source: Rcode

B. *Modelling*

For all our datasets we are going to create a classification model which will help us determine based on the input parameters that whether the output would be yes or no, For this we would be making use of 5 modeling algorithms from which 4 are based on pure classification and 1 is based on clustering from which we can determine where it is a yes or no based on the clusters.

- **SVM**

  SVM stands for Support Vector Machine, it is a good classification and regression algorithm which was first proposed by Vapnik to tackle the pattern recognition issue [4].

- **Decision Tree**

  Decision Tree is one of the most popular predictive modelling approaches where in the data is split based on different decisions and conditions. The leaf node (bottom most node) contains the actual answer to the question for which we are trying to classify. They fall under supervised learning category [5].

[3] Shivangi, "Psychology Discussion Net," Psychology Discussion, [Online]. Available: https://www.psychologydiscussion.net/educational-psychology/statistics/correlation-definitions-types-and-importance-statistics/2788. [Accessed 15 August 2020].

[4] Subham, "Hacker Earth," Hacker Earth, [Online]. Available: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/. [Accessed 15 August 2020].

[5] A. C. Bahnsen and S. Villegas, "AppGate," 9 March 2018. [Online]. Available: https://blog.easysol.net/machine-learning-algorithms-4/. [Accessed 16 August 2020].

- **KNN**

KNN stands for K nearest neighbor which is a clustering algorithm, it has a centroid point around which points with similar features are located, based on this logic if any new point is introduced then it is classified based on the cluster of its nearest neighbor, if k neighbors lies next to the newly plotted point then based on the majority count, the newly plotted point is classified.

- **Naïve Bayes**

The Naïve Bayes Classifier algorithm is based on supervised machine learning that makes use of Bayes Theorem. Advantages of using Naïve Bayes Classifier is that they are quite simple yet fast also they are quite scalable and also works on small datasets.[6]

- **Random Forest**

This classifier is a model which consist of N number of decision trees, these are similar to ensembles approach in neural network where "it uses a bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree." [7]

The above-mentioned models will help us create robust machine learning models which will classify the data based on its features. We apply Decision Tree and SVM on our credit card dataset for fraud detection, Random Forest and KNN on our bank loan dataset and Naïve Bayes on our bank mortgage dataset respectively. In the next section we will discuss the performance of all these models and see whether they're reliable enough to be accepted in the real world.

IV. Evaluation

After we've created our models it's very important to see whether they are efficient and reliable hence we evaluate their performances based on certain evaluation metrics to measure the model's outcomes.

For this study we are making use of a package available in R called 'ml_test'[8] which gives us good metrics to evaluate classification models. The function ml_test returns various range of metrics which includes accuracy, F1, F2, Youden, OP, precision, recall, specificity, NPV, MK etc. Also, we make use of another very simple, yet powerful evaluation metrics called confusion Matrix which will help us measure the reliability of our classification models.

Results for Credit Card fraud Dataset:

For SVM Model when we evaluate the model on test-set we get the following confusion Matrix table

	0	1
0	2269	179
1	30	2181

Here 1 indicates fraud and 0 indicates normal, since our data was highly imbalanced, we made use SMOTE function to handle this issue. The resulting dataset was balanced and thus eliminating any chance of being biased towards majority class. Looking at the outcome of our confusion matrix we can calculate the accuracy by dividing the sum of true positives by total inputs values which gives us an **accuracy of 0.95** and we also performed Root Mean Square Error(RMSE) which resulted out to be **0.19** a RMSE value near to zero indicates positive outcome.

Results based on ML_TEST on SVM

Metrics	0	1
Diagnostic odds ratio (DOR)	628.4706	628.4706
F1	0.9515072	0.9524014
Negative Predicted Values (NPV)	0.9810071	0.9240560
Specificity	0.9254167	0.9806394

For Decision Tree Model the accuracy obtained was 0.9575 which is pretty good whereas other metrics also showed positive outcomes with an error rate of **0.042** as calculated from the ml_test function. Two models were applied on this dataset with a bunch of metrics for evaluation. Next for the dataset about loan acceptance,

[6] T. Yiu, "Towards Data Science," Medium, 12 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed 16 August 2020]

[7] T. Yiu, "Towards Data Science," Medium, 12 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed 16 August 2020]

[8] G. Dudnik, "Package 'mltest'," R projects, 16 November 2018. [Online]. Available: https://cran.r-project.org/web/packages/mltest/mltest.pdf. [Accessed 16 August 2020]

we implemented two models namely Random Forest and K Nearest Neighbor (KNN). Although our dataset was unbalanced and we had to upsample the minority class, our KNN based model gave an accuracy of **0.82** which is pretty reasonable. This model's accuracy can be improved if we only take 2 or 3 independent variables as opposed to 6 parameters which we took but then the reliability of the model in real world would be in question.

Results from Random Forest were similar to that of KNN with accuracy of 0.87 and other metrics listed below

Results based on ML_TEST for Random Forest

Metrics	NO	YES
F1	0.8754633	0.8724374
F2	0.8857057	0.8622242
Jaccard [calculated as TP/ (TP+FP+FN)]	0.7785102	0.7737374
MK	0.7489168	0.7489168

Overall performance of both the models was satisfying and the results from this model can be considered when deployed in real world situation although the final decision for any business call must be taken by a human.

For the final dataset of Bank Marketing to classify where the customer will subscribe for a term deposit or not, we made use of two machine learning classification algorithms which were Naïve Bayes and Random Forest.

The Accuracy obtained for Naïve Bayes was 0.63 while other metrics are stated below

Results based on ML_TEST for Naïve Bayes

Metrics	NO	YES
Precision	0.61722	0.66616
Recall	0.73128	0.54179
Geometrics Mean	0.6294466	0.6294466
Specificity	0.54179	0.73128

For Random Forest we got an accuracy of 0.64 then we plot a variable importance plot using the model obtained from random forest to see how much each variable affects the dependent variable

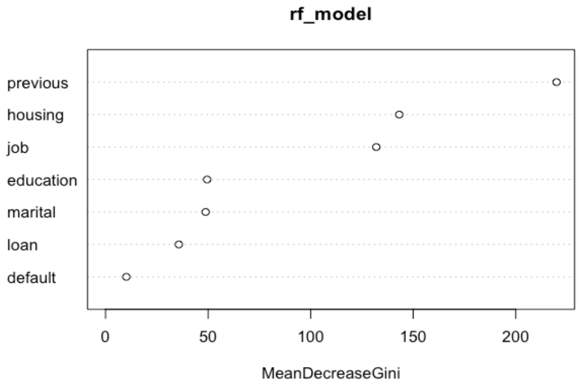

Figure 7: **varImpPlot** Source: Rcode

We can clearly see that the parameter default has very low impact on our model hence we remove that feature and then run our model again, this time we get an improvement of 1% in our accuracy, the overall classification still remains a challenge as another important factor for determining whether the customer will take the loan or not depends on its need who's data is still not available.

V. Conclusion and future work

To conclude the study, overall performance of all the models was satisfactory, out of the five models which were implemented and tested on the three datasets, Random Forest and Decision Tree gave solid outcomes as Random Forest was implemented on 2 models and Decision Tree showed low error rate, however as discussed in the Related Studies section, Artificial Neural Network is worth considering but the only issue is that it required many parameters and high computational power.

In case of the credit card fraud detection, the results obtained from the model were quite satisfactory however since we are using a publicly available dataset and also the number of fraud transactions are low, so the effectiveness of the model needs to be improved with more real world data in future. In case of loan default and loan sales marketing, even though the model classifies based on the input parameters and we can act upon the results, we have to keep one thing in mind that loan defaults can also be due to economic downfalls like recession, job loss of that customer or death which are surely a possibility but aren't included as a parameter in our datasets. Future work in this domain lies in building a robust model with all the mentioned parameters taken as input for training the model, also even if such models are deployed on the forefront we should always keep a human to take important decisions as the risk involves financial downfalls.

VI. REFERENCES

[1] M. Jeragh and M. AlSulaimi, "Combining Auto Encoders and One Class Support Vectors Machine for Fraudulant Credit Card Transactions Detection," *IEEE,* vol. 18, pp. 1-7, 2018.

[2] V. N. Dornadula and G. S, "Credit Card Fraud Detection using Machine Learning Algorithms," in *ICRTAC*, Chennai, 2019.

[3] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach," *Review of Development Finance,* vol. 4, no. 1, pp. 20-28, 2014.

[4] A. Khashman, "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes," *Elsevier,* vol. 37, no. 9, pp. 6233-6239, 2010.

[5] ElianaAngelinia, G. Tollob and AndreaRolic, "A neural network approach for credit risk evaluation," *The Quarterly Review of Economics and Finance,* vol. 48, no. 4, pp. 733-755, 2008.

[6] Q. Caoa and M. E.Parry, "Neural network earnings per share forecasting models: A comparison of backward propagation and the genetic algorithm," *Decision Support Systems,* vol. 47, no. 1, pp. 32-41, 2009.

[7] N. K. Gyamfi and D. J.-D. Abdulai, "Bank Fraud Detection Using Support Vector Machine," *IEEE Annual Information Technology,* pp. 37-41, 2018.

[8] W.-F. Yu and N. Wang, "Research on credit card fraud detection model based on distance sum.," in *International Joint Conference on Artificial Intelligence*, 2009.

[9] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossian, N. H. Ghani, H. Arif and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," in *IEEE*, Dhaka, 2019.

[10] V. Nagadevara, "APPLICATION OF HYBRID METHODOLOGY TO PREDICT HOUSING LOAN DEFAULTS IN INDIA," *Journal of International Management Studies,* vol. 15, no. 3, pp. 43-50, 2015.

[11] J. Antucheviciene, A. Zakarevicius and Z. E. K, "Measuring congruence of ranking results applying particular MCDM methods,," *Informatica,* vol. 22, no. 3, pp. 319-338, July 2011.

[12] H. Abdi and L. J. Williams, "Principal component analysis," *Interdisciplinary Reviews Computational statistics,* vol. 2, no. 4, pp. 433-459, 2015.

[13] ChristianJutten and JeannyHerault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing,* vol. 24, no. 1, pp. 1-10, 1991.

[14] B. Back, T. Laitinen and K. Sere, "Neural networks and genetic algorithms for bankruptcy predictions," *Expert Systems with Applications,* vol. 11, no. 4, pp. 407-413, 1996.

[15] G. Kou, Y. Peng and C. Lu, "MCDM APPROACH TO EVALUATING BANK LOAN DEFAULT MODELS," *Technological & Economic Development of Economy,* vol. 20, no. 2, pp. 292-311, 2014.

[16] U. KOC and T. Sevgili, "Consumer loans first payment default detection: a predictive model," *tubitak,* pp. 1-18, 2019.

[17] S. H. S. Nor, S. Ismail and B. W. Yap, "Personal bankruptcy prediction using decision tree model," *Emerald Insight,* vol. 24, pp. 1-15, 2018.

[18] A. Gepp and K. Kumar, "Predicting financial distress: a comparison of survival analysis and decision tree techniques," *Procedia Computer Science,* vol. 54, pp. 396-404, 2015.

[19] M. Chen, "Predicting corporate financial distress based on integration of decision tree classification and logistic regression," *Expert Systems with Applications,* vol. 38, no. 9, pp. 11261-11272, 2011.

[20] Bernama, "Alarming 22,581 M'sians aged 25 to 34 declared bankrupt in last four years," Straits Times, Kaula Lampur, 2016.