Home Credit Default Risk Prediction

Dublin City University, Module - CA683 - Data Analytics and Data Mining

Sagar Ramachandra Murthy *Student ID - 20210320*
Chinmaya Laxmikant Kaundanya *Student ID - 20210168*
Rohith Raghuprakash Menon *Student ID - 20210989*
Sumit Sunil Khopkar *Student ID - 21261643*

***Abstract*—Estimation or assessment of debt default is a critical procedure that financial institutions should do to assist them in determining whether or not a loan applicant may become a defaulter at a later stage, so that they can process the application and decide whether or not to grant the loan. This research project focuses on predicting if a person will default the home loan or not. For this, we have explored various features in our dataset and identified key features which helps in determining if a person will default. Currently though there are Machine Learning and Deep Learning algorithms which have proven to highlight key features in this process, our approach focuses more on making this simpler by exploring various Data Mining techniques. With our approach, we could observe a formidable difference between the F1 scores in comparison between the models, before and after our research.**

***Index Terms*—Loan Default, Classification, Default Risk, Default Risk Prediction, Machine Learning**

I. Introduction

Having a home of your own is a dream to many people in all walks of life. Buying a home is easier said than done as it requires us to have a good credit history to avail a loan. Many people struggle to get loans as they have insufficient credit histories or sometimes a non-existing credit history. Unfortunately this sector of population is often taken advantage of and pressured by untrustworthy loan lenders. Home Credit is an organization that is striving to bring this financial inclusion for the unbanked and unorganized sector who are either unaware or use the banking lesser and thereby enabling them to have a positive banking or borrowing experience. This is done with the help of a variety of data sources and features which includes data about the loan, loan applicant, his/her previous loan transactions or history from current and other financial institutions, monthly credit balance that applicant had and also details of installment payment availed through the POS (Point of Sale) loans. The dataset for this research project is collected from kaggle.com. This dataset is of Home Credit group and was featured in one of the greatest prediction competitions on kaggle. The goal of the competition was to predict how capable each applicant is of repaying the loan. Similarly along with the prediction, the goal of our project is to identify key features which influence loan default. The further sections of this report will throw light on the existing literature, followed by a novel approach used by us, followed by results and conclusion.

II. Related Work

Domain expertise or Business Understanding plays a pivotal role for any data analysis or machine learning task. In order to gain that knowledge, we have referred to the existing literature in the domain of loan default risk prediction and also have referred to several base papers to understand it bottom-up.

To start with, Talha Mahboob Alam et al. [1] began their research by employing exploratory data analysis approaches, such as data standardization, to analyze financial datasets. Initially, for the predictive analysis they applied the GBDT model and then compared the results to those of typical machine learning algorithms. The GBDT model was found to be a greater prediction accuracy rate than typical machine learning-based models. Moses Lusinga et al. [2] proposed the use of Shapley Additive Explanations (SHAP) tool which is better for demonstrating details in ML models. SHAP can increase openness in financial models and build trust between the financier and the lander. ML techniques outperform statistical analytic techniques in most cases, with the exception of ANN in a few cases due to a lack of data and difficult parameter optimisation. They used SHAP to find the feature representation that has the greatest impact on the model's performance. Applicants and credit analysts can use the combination of XGBoost and SHAP to correctly anticipate a loan based on alternative data. At last they mentioned that the results would be better if the other dimensionality reduction techniques such as Autoencoder are used. Wang et al. [3] proposed a data mining model for insurance companies to predict loan applicants, and their data mining model included a rule generator and a recommendation mechanism. Because a policyholder with a lower interest rate is more likely to apply for a loan, they were able to get better outcomes. To classify loan default and non-loan default consumers, Xia et al. [4] suggested a credit score methodology. They built a model using the P2P lending dataset and preprocessed the columns due to noisy data. The results were put to the test using advanced gradient boosting models and keyword clustering-based approaches. To improve the performance of classifiers, they extracted prominent features. Their tests revealed that the Catboost model, which is based on gradient boosting, outperformed other standard models. In P2P lending, Zhou et al. [5] developed a decision tree-based model for client default prediction. For modelling, they used a variety of ensemble-based machine learning models. Missing values and excessive scarcity were also dealt with using data preparation techniques. They also graded the features, and fewer features that were related to each other were removed. Different Hyper-parameters were also optimized to boost the performance of classifiers. Their experiments demonstrated that leveraging high-dimensional data to make desirable predictions yielded promising outcomes. Lee et al. [6] used Deep Convolution Neural Network (DCNN) and other learning algorithms to recommend financial instruments, and the final findings showed that DCNN performed marginally better than Decision Tree, Support Vector Machine, and Logistic Regression in predicting accurate data. Bhoomi Patel et al. [7] used multiple machine learning models for predictive analysis for the forecasting of loan defaults. Logistic Regression, Random Forest, Gradient Boosting, and CatBoost Classifier were used to achieve the best results. In comparison to Logistic Regression, the Gradient Boosting technique produces better or equal results. CatBoost converts categorical values to numerical values using a range of statistics based on categorical features and numerical and categorical attributes. With respect to the given dataset, the CatBoost classifier and Gradient Boost provide nearly comparable accuracy. Furthermore, these models can be used to make better judgements about loan applications, potentially saving a financial institution from massive losses. in this work, three algorithms - the j48, BayesNet, and Naive-Bayes algorithms were utilized to create prediction models that may be used to predict and classify customer-introduced loans as good or bad by looking at customer behavior and prior repayment credit. The Weka application was used to implement the model. We discovered that the best algorithm for loan classification is the j48 algorithm after applying classification data mining approaches algorithms such as j48, BayesNet, and NaiveBayes. Because of its great accuracy and low mean absolute error, the J48 algorithm is the best. Hsu et al. [8] used a support vector machine (SVM) to classify a bank credit dataset and found that SVM accuracy increases as the number of data samples grow or other selection factors are used, making it more effective in credit rating. Turkson et al. [9] used supervised and unsupervised machine learning algorithms to predict creditworthiness using a bank credit dataset, with some systems achieving an accuracy rate of up to 80%. Moro et al. [10] looked at a telemarketing dataset from a Portuguese retail bank and used neural network models to predict success. Using the Iranian bank's dataset, Jafarpour et al. [11] targeted customer relationship management. Through multiple channels, the customer relationship management model assesses the relationship between consumers' requirements and banks, devising an equation that banks and lending corporations can use to anticipate loan customers. The book "Multiple Factor Analysis by Example Using R" [12] talks about the ineffectiveness of the standard procedure where the quantitative variables have to be converted to qualitative variables, break down their variation interval into classes, and run multiple correspondence analysis on the resulting homogeneous table. The more advanced method is to generate a principle components by the factor analysis approach that has enough positive qualities to be classified as a separate method.

The above literature review gave us a solid understanding of end-to-end analysis and different models currently being employed or implemented in financial domain. It also paved way for us to think through and come up with a novel approach to answer our current business problem or the problem statement.

III. Data Mining Methodology

Based on the above literature review, business and data understanding, let us look at each of the steps involved and have been followed by us to pre-process and prepare the data, model and evaluate the data and suggest ways for deployment. These steps were implemented in an iterable fashion so as to arrive at a better model and an approach which is faster, replicable and easily implementable. Let us look at each of the steps in detail below.

A. Data Understanding

The final dataset used for our project is a combination of data with a total data of about 2.68 GB (zipped) containing 307,511 rows and 122 columns. Every step followed by us have been deep-dived and explained in detail below however lets first get a quick view of the entire methodology used by us end-to-end here. The first step in our structured approach was to understand each feature/column of our dataset. Based on our feature understanding, we leaned towards the domain knowledge from the above literature review to fill in those gaps and only then we started with our data analysis. Next we proceeded with a thorough sanity check of the data followed by analyzing outliers. Though we identified outliers in few features, based on our research and business knowledge the data wasn't imputed or removed because these outliers can in-turn be the true gems. Next was to identify features which exhibit Multicollinearity. This was identified and handled effectively using VIF. Next, due to the sheer number of features present in our dataset, it was a wise choice for us to perform dimensionality reduction. Next, we also noticed that our dataset is imbalanced, which is we have data with less defaulters and more non-defaulters, hence we also used Synthetic Minority Oversampling Technique (SMOTE) to balance the data. Do note that the modelling stage covers and explains the model being run with and without these imputations and SMOTE to see how good or bad the models perform.

B. Data Preparation

1) Handling Missing Values: The first step in any Data Pre-Processing is to perform the basic sanity checks. We first start by checking for missing (null/nan) and erroneous values (special characters, prefix and suffix spaces) present in our dataset, followed by checking for duplicates. We observed that all the files used in our analysis have missing values and had to be treated before doing any further analysis. Missing values can occur for a number of reasons such as operator error, faulty device, respondent refuses to enter the data or They may be values, attributes entire records or entire sections which are missing. Handling such missing values in numerical variables by just inserting an average of the column or mode value in categorical variables could possibly draw wrong conclusions.

Initially we calculated the ratio of missing values of a feature to the total number of rows. We removed those features that had at least 50% of the values missing in that feature. No intuiton can be drawn to impute these values and hence it was better to leave them out. The threshold of 50% lead to the removal of 41 features. The next step was to determine if the values are Missing Conditionally at Random (MAR) or not. MAR implies that a relationship exists between the values of these variables and the other variables. Statistical tests can be preformed to confirm the existence of a relationship. If a relationship exists then the missing values can be imputed using regression or a propensity score model.

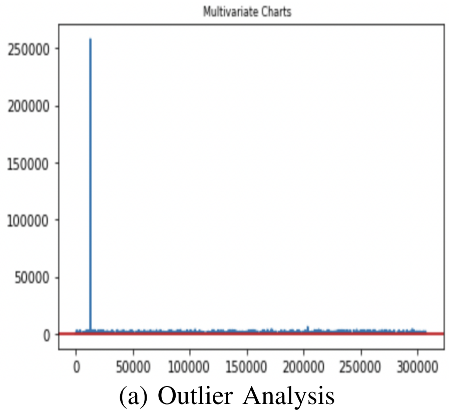
Two different approaches need to be followed in our case to perform statistical tests since the missing value data is continuous as well as categorical. In case of the categorical data, we have performed the "Chi Square" test to determine if the relationship is statistically significant. The categorical data was first separated in another dataframe and then encoded before carrying out the test. What we found out was that every variable containing a missing value has a significant relationship with the other variables for a significance level of 0.05, thus, confirming the nature of the missing categorical values to be MAR. Similarly, we performed the "Pearson's Correlation Coefficient" test for the continuous variables.

Two observations were laid out from the test. First one being that we found out the missing variables CNT_FAM_MEMBERS (Number of family members) and AMT_GOODS_PRICE (Goods price of good that client asked for) to have a significant correlation (Coefficient greater than a threshold of 0.8) with non-missing variables CNT_CHILDREN (Number of children) and AMT_CREDIT (Credit amount of the loan) respectively. And the second observation was that the other missing variables had a correlation with each other. Hence, imputation would lead to an incorrect estimation of the missing value. Based on this intuition we removed these features from our dataset.

After this, the other continuous variables which are either MCAR or NMAR were plotted to have a deeper look at them. The variable DAYS_LAST_PHONE_CHANGE (How many days before application did client change phone) had both an imbalanced set of values (lot of 0s in comparison to other values) and it also is difficult to determine whether the null values should have been 0 or are they deliberately left out. There were also some other variables which had a highly imbalance dataset in this same category. We decided to not use these variables. The remaining variables can be considered to be MCAR. These variables do not form a normal distribution and hence we decided to impute them using their Median. Now we handled the MAR values imputation explicitly.

The previously mentioned continuous MAR values were imputed using Linear regression. The approach that we followed to split the dataset for training and testing was that we kept the rows not having missing values as the training data and the rows having missing values were split to form the testing data. We tried to follow the same approach for the categorical variables (by using Logistic Regression instead of Linear), but the model failed to converge. This meant that we would have to find a way to handle each variable in its own way. Two of the three MAR variables were dropped since the number of missing rows was less. But the variable "OCCUPATION_TYPE" had one-third of the entire data missing. Hence, instead of dropping we replaced the missing values to a self-defined category of "Other Occupation". Also, looking at it with a business perspective, the type of occupation of the applicant will be of importance for the problem statement. Another vital information that we confirmed was that our dependent variable which is the "Target" variable does not contain any missing value.

2) Analyzing Outliers: A significant number of variables are recorded or sampled in many data analysis procedures. The discovery of outlaying observations is one of the first stages in obtaining a coherent analysis. Although outliers are frequently seen as errors or noise, they might contain vital information [13]. Outlier detection approaches are classified as either univariate or multivariate. In our dataset, we used multivariate outlier detection. The use of more than one variable to identify an outlier is known as multivariate outlier identification. Individual outliers in a single variable are considered in univariate outlier identification. If there are odd combinations of two or more variables, these univariate outliers may not be detected. We concentrated on depth-based control charts since they are a reasonably robust method. The Hotelling's T-Squared from n rows with k columns was used to compute critical values. The Mahalanobis distance was then computed for each row using the overall multivariate mean X with the help of scipy.spatial package.(xbar). The Mahalanobis distance is used in conjunction with the covariance matrix to compute individual Hotelling T Squared statistics, which is then compared to the crucial Hotelling's T Squared values. The chart shown below is the outcome of the method mentioned above and we observed that the customer with id '114967' is the most extreme outlier. Since these outliers may contain vital information, we have refrained ourselves from imputing them.



(a) Outlier Analysis

We also implemented the concept of interquartile range (IQR) on 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE' to observe the outliers in these columns. IQR is a statistical concept that divides a dataset into quartiles to measure statistical dispersion and data variability. We observed that 'AMT_INCOME_TOTAL' has the max value of 492 standard deviations from the mean and user '114967' has the highest income of 117,000k. It could be true that the user might have such a large salary, hence we decided not to remove this record.

3) Treating Multicollinearity: When a multiple linear regression analysis contains numerous variables that are substantially correlated not only with the dependent variable but also with each other, this is referred to as multicollinearity. Because of multicollinearity, some of the significant variables under investigation are statistically insignificant. In this study, the variance inflation factor is utilized to calculate how much the variance of the predicted regression coefficient is inflated when the independent variables are correlated [2]. VIF is calculated by regressing an independent variable against all other variables and applying the formula:

$$X1 = \alpha + X2 + ... + Xn$$

We then get the R2 from each variables regression model and then calculate the VIF using the following equation:

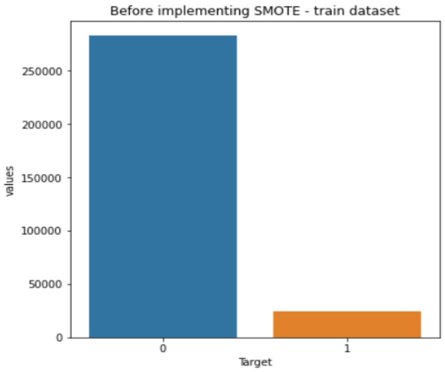
$$V.I.F = \frac{1}{1 - R^2}$$

If the value of VIF is 1 to 5, it indicates that the variables are moderately connected. VIF's difficult value ranges from 5 to 10, indicating strongly connected variables. If VIF is between 5 and 10, there will be multicollinearity among the predictors in the regression model, and if VIF is greater than 10, the regression coefficients will be calculated inaccurately due to the existence of multicollinearity so we may have to drop the variables with VIF ¿ 10 [14]. In our dataset we observed there are 10 variables with VIF value greater than 5 out of which 'AMT_INCOME_TOTAL', 'AMT_ANNUITY', 'DAYS_BIRTH', 'FLAG_MOBIL', 'CNT_FAM_MEMBERS', 'REGION_RATING_CLIENT' have VIF values greater than 10. So, we must drop these variables to reduces the multicollinearity. To handle these variables, we had different approaches. 'DAYS_BIRTH' was converted to the age by dividing the value by 365 and then converted it into category based on age range of 0-20, 21-40 and so on. This reduced the multicollinearity of the variable and still retained the information before we dropped the variable. To handle other variables, we implemented dimensionality reduction to remove these variables along with other variables thus reducing the number of variables. Detailed explanation of dimensionality reduction is given below.

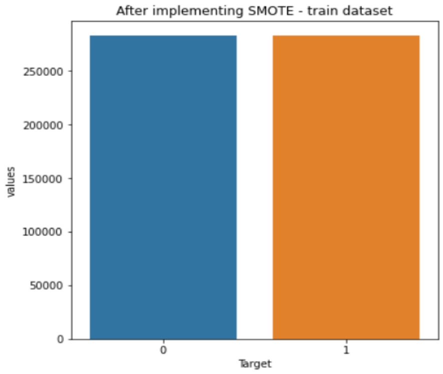
4) Dimensionality Reduction: When dealing with such high dimensional data, statistical and machine reasoning approaches confront a significant challenge, and the number of input variables is often decreased before a data mining strategy can be successfully implemented. Dimensionality reduction can be accomplished in two ways: by retaining only the most relevant variables from the original dataset (this technique is known as feature selection) or by utilizing the redundancy of the input data and identifying a smaller set of new variables, each of which is a combination of the input variables and contains essentially the same information as the input variables (this technique is known as dimensionality reduction) [15]. FAMD is a data exploration approach that works with both continuous and categorical variables. It may be thought of as a cross between PCA and MCA. More specifically, the continuous variables are scaled to unit variance, and the categorical variables are converted into a collection of dummy variables before being scaled using MCA's particular scaling. This guarantees that the effect of continuous and categorical variables in the analysis is balanced. It signifies that both variables are on equal footing when it comes to determining the dimensions of variability. This technique enables one to investigate the similarities between people while accounting for mixed factors, as well as the correlations between all the variables. From the previous analysis of our dataset, we observed few variables that have high multicollinearity with VIF value greater than and many redundant variables. Thus, to reduce this high dimensional dataset with multicollinear variable we have implemented FAMD using 'prince' package in our dataset to reduce the high dimensional categorical and continuous data into smaller set of new variables but retaining the information of the variables. In this research, we applied this technique on a set of continuous and categorical variables thus reducing the shape of dataset from (307511, 65) to (307511, 35). This reduced the high dimensional dataset along with removing the multicollinearity that we discussed in the previous section.

5) Handling Imbalanced Data: Imbalanced data-set classification has emerged as a major issue in data mining. Because the fundamental assumption of typical classification algorithms is that the distribution of classes is balanced, the techniques utilized in Imbalanced data-set Classification cannot reach an ideal impact. Considering the imbalance dateset classification, we offer an oversampling approach based on support degree to aid individuals in selecting minority class samples and generating new minority class samples [16]. We have implemented Synthetic Minority Over-sampling Technique (SMOTE) in our dataset. The SMOTE method is a traditional oversampling algorithm. The main principle behind SMOTE is that new positive class samples are created by linear interpolating between two near positive class samples and then added to the original dataset. By increasing the number of fresh minority class samples, the two classes might be balanced [16]. The graph highlighted below shows the data spread before implementing SMOTE.

We applied SMOTE technique from 'imblearn.over_sampling' package where we apply minority sampling strategy thus synthetically generating values to balance the dataset. After SMOTE implementation we get a balanced dataset. The graph highlighted below shows the data spread before and after implementing SMOTE to get a better view and understanding.



(b) Before SMOTE

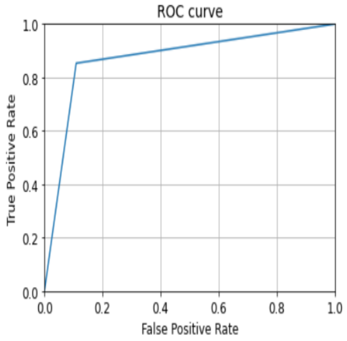


(c) After SMOTE

C. Modeling and Deployment

This section defines all of the data mining classification models that were employed, as well as the results that we obtained after applying each of them. For each model, the accuracy value obtained is also shown. The following is a list of the algorithms that were used:

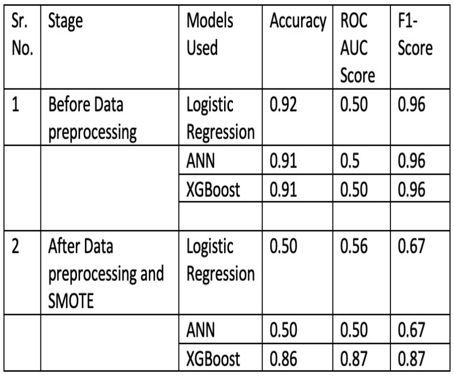
- Logistic Regression: The target column is a dichotomous variable having 0 for the people who will not default on the loan and 1 for the people who can default. By distributing observations to a different set of classes, Logistic Regression conducts categorical categorization. It can be used to classify data points into binary categories. A categorical classification is one in which an output is assigned to one of two classes (1 or 0). By altering its output with the logistic sigmoid function, logistic regression outputs a probability value.
- Artificial Neural Networks: Credit default concerns are rooted in neural networks, particularly Multi-layer Perceptron neural networks (MLP). MLPs are a type of feed-forward artificial neural network that has at least three layers of nodes. The input layer, hidden layers, and one output layer are among the layers. A basic Neural Network with two hidden layers and an output layer with Sigmoid activation is used.
- Extreme Gradient Boosting: Gradient Boosting is a machine learning strategy for classification and regression issues. In the form of an ensemble of mediocre prediction models, it creates a model for prediction. The weak classification algorithm is implemented on updated versions of the data in a sequential way in the boosting algorithm, resulting in a sequence of weak classifiers. Gradient Boosting is a type of ensemble learning that combines numerous weak decision trees to produce a powerful classifier. These decision trees are combined to produce a powerful gradient boosting model.



(d) Evaluation Metric

The above Machine Learning model can be deployed on cloud as a website or as an app. Model can be continuously improved by incremental re-training approach and deployed as a need basis as per the business requirement.

IV. Evaluation/Results



Sr. No.	Stage	Models Used	Accuracy	ROC AUC Score	F1-Score
1	Before Data preprocessing	Logistic Regression	0.92	0.50	0.96
		ANN	0.91	0.5	0.96
		XGBoost	0.91	0.50	0.96
2	After Data preprocessing and SMOTE	Logistic Regression	0.50	0.56	0.67
		ANN	0.50	0.50	0.67
		XGBoost	0.86	0.87	0.87

(e) Evaluation Results

Because of the over-representation of the majority class and under-representation of the minority class, the performance of different classifiers is generally high on imbalanced datasets. Even if the model's accuracy is higher, Loyola-González et al. [17] points out that the model's accuracy is biased towards the majority class, with minority classes impacting less to decide the accuracy. To solve this challenge, we tried to use a balanced dataset generated using SMOTE analysis to eliminate the dominant class's bias. Also accuracy is not the right metric for our business problem or problem statement. This points out towards the trade-off between precision and recall. In our case as identifying all the users who are susceptible to default and also identifying the user who will default is important, f1-score is a better metric to handle the trade-off. Also as a note, recall would be a bit more important than precision and we can have some leverage on the precision score. From the above table we can see that the initial F1 Score is higher, which is due to the data bias of more non-defaulters vs defaulters. Post SMOTE though the overall F1 Score is lesser, the data now is more balanced and without bias. We can also see that Extreme Gradient Boosting (XGB) is a better model compared to Artificial Neural Network (ANN) and Logistic Regression.

V. Conclusion and Future Work

Various algorithms were used to identify loan defaulters in this paper. To start with, we handled missing values by identifying their appropriate types such as NMAR/MCAR/MAR and eliminated only those variables which were found to be highly correlated with each other, using an appropriate statistical test. We then analysed the outliers and treated multicollinearity among the variables by calculating their VIF scores. Further to reduce the dimensionality of the data we applied the FAMD technique as there was a combination of multiple quantitative and qualitative variables. At last, before predictive analysis, we handled the imbalanced target column using SMOTE analysis and we evaluated multiple Machine Learning models in two different stages as mentioned in the evaluation section of this paper. Logistic Regression, Neural Networks and Extreme Gradient Boosting (XGB) were used to achieve the best results. In comparison to Logistic Regression, the Gradient Boosting technique produces better results. Further, from the overall research, these models can be used to make better judgements about loan applications, potentially saving a financial institution from massive losses.

Future work: Based on the business requirement and on how flexible the business wants to stretch the Recall and F1 Score, the model can be tuned accordingly. The model could be improved in the future by attempting to build a better dataset by selecting just important features and/or experimenting with thresholds. We've just looked at home loan defaulters so far, but a system may be developed to anticipate defaulters on all other types of loan as well.

References

[1] T. M. Alam et al., "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," in IEEE Access, vol. 8, pp. 201173-201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[2] M. Lusinga, T. Mokoena, A. Modupe and V. Mariate, "Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries," 2021 IEEE AFRICON, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570906.

[3] Williams, Graham J. "Rattle: a data mining GUI for R." The R Journal 1.2 (2009): 45-55.

[4] Y. Xia, L. He, Y. Li, N. Liu, and Y. Ding, "Predicting loan default in peer-to-peer lending using narrative data," J. Forecasting, vol. 39, no. 2, pp. 260–280, Mar. 2020.

[5] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," Phys. A, Stat. Mech. Appl., vol. 534, Nov. 2019, Art. no. 122370.

[6] K. H. Kim, C. S. Lee, S. M. Jo and S. B. Cho, "Predicting the success of bank telemarketing using deep convolutional neural network," 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Fukuoka, 2015, pp. 314-317.

[7] B. Patel, H. Patil, J. Hembram and S. Jaswal, "Loan Default Forecasting using Data Mining," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154100.

[8] C. F. Hsu and H. F. Hung, "Classification Methods of Credit Rating - A Comparative Analysis on SVM, MDA and RST," 2009 International Conference on Computational Intelligence and Software Engineering, pp. 1–4, Dec. 2009.

[9] R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, 2016, pp. 1-7.doi: 10.1109/ICAIPR.2016.7585216

[10] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, vol. 62, pp. 22–31, 2014.

[11] H.Jafarpour and H. Sheikholeslami Garvandani, "New Model of Customer Relationship Management in Iranian Banks," icbme.yasar.edu.tr, pp. 1–12, 2012.

[12] Multiple Factor Analysis by Example Using R-book

[13] I. Ben-Gal, "Outlier Detection," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2005, pp. 131–146. doi: 10.1007/0.387.25465.X.7.

[14] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," Am. J. Appl. Math. Stat., vol. 8, pp. 39–42, Jun. 2020, doi: 10.12691/ajams.8.2.1.

[15] C. O. S. Sorzano, J. Vargas, and A. P. Montano, "A survey of dimensionality reduction techniques," ArXiv14032877 Cs Q-Bio Stat, Mar. 2014, Accessed: Apr. 06, 2022. [Online]. Available: http://arxiv.org/abs/1403.2877

[16] K. Li, W. Zhang, Q. Lu, and X. Fang, "An Improved SMOTE Imbalanced Data Classification Method Based on Support Degree," in 2014 International Conference on Identification, Information and Knowledge in the Internet of Things, Oct. 2014, pp. 34–38. doi: 10.1109/IIKI.2014.14.

[17] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," Neurocomputing, vol. 175, pp. 935–947, Jan. 2016.

[18] M. Lusinga, T. Mokoena, A. Modupe and V. Mariate, "Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries," 2021 IEEE AFRICON, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570906.

[19] B. Patel, H. Patil, J. Hembram and S. Jaswal, "Loan Default Forecasting using Data Mining," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154100.