National College of Ireland

Project Submission Sheet – 2022

Student Name: Aditya P. Shinde, Kalpesh J. Dhande, Rahul S. Vaydande, Tsai Shih Yang

Student ID: 20178883, 20185821, 20181663, 21101825

Programme: MSCDAD_B **Year:** 2022

Module: Domain Application of Predictive Analytics

Lecturer: Vikas Sahni

Submission Due Date: 29-04-2022

Project Title: Vehicle Loan Default Prediction using XGboost classifier

Word Count: 2598

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Aditya P. Shinde, Kalpesh J. Dhande, Rahul S. Vaydande, Tsai Shih Yang

Date: 29-04-2022

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Vehicle Loan Default Prediction using XGboost Classifier

1st Rahul S. Vaydande
Msc. In Data Analytics
National College of Ireland
Dublin, Ireland
x20181663@student.ncirl.ie

2nd Aditya P. Shinde
Msc. In Data Analytics
National College of Ireland
Dublin, Ireland
x20178883@student.ncirl.ie

3rd Kalpesh J. Dhande
Msc. In Data Analytics
National College of Ireland
Dublin, Ireland
x20185821@student.ncirl.ie

4th Tsai Shih Yang
Msc. In Data Analytics
National College of Ireland
Dublin, Ireland
x21101825@student.ncirl.ie

***Abstract*—Many people all around the world rely on banking institutions to lend them money to buy a vehicle. This is the continuation document for the proposed project design. We will design a model that will assist banks and financial lending institutions in determining whether or not the loanee will be able to repay the loan. To build the model, we'll have used the LT vehicle loan default prediction data, which is made up of past loanee's information. In order to predict loan default, we will use the XGboost model. On the basis of the confusion matrix, accuracy, f1score, precision, and recall score, we will evaluate the results of the built model.**

***Index Terms*—XGBoost, SMOTE, ROC, Vehicle Loan Defaulters .**

I. Research and investigation

In this research paper [1], the author has utilized machine learning models such as logistic regression, random forest and LightGBM to predict credit defaulter risk on Peer to Peer (P2P) platform which connects the lenders and borrowers. Out these implemented models, XGBoost proved to be the best performing model with highest accuracy. The proposed approach also uses resampling method such as SMOTE, NearMiss and manual 1:1 random sampling has the dataset which is available suffers a class imbalance problem. We will be taking inspiration from this research and try to implement the XGBoost machine learning model on our dataset.

Similarly, in this study [2] also the researchers have used logistic regression and XgBoost ML models to anticipate if a customer would take out credit or become debtors. The models trained on training data produced almost same results but when run on test data the logistic regression model had an accuracy of 88 percent whereas, for XGBoost it was 92 percent. This research has not utilized any kind of pre-processing techniques as the previous paper and still it has managed to produce a good output.

In this Next study [3], the Loan Repayment Behavior Prediction is done on the provident fund using different ML algorithms namely, Cart Decision tree, Random Forest, XGBoost, CatBoost, AdaBoost, and LightGBM. Out of these models, the best classification accuracy was achieved by XGBoost. In this research, the author has also pointed out the importance of feature engineering which will be adopted in our research also so to transform features into more useable formats.

In this comparative study [4], the researchers have used logistic regression and XGBoost model to test compare the performance on the credit risk prediction dataset. The XGBoost ML model outperformed the Logistic Regression model. The dataset contains different types of continuous variable which were transformed depending upon the business requirement for this research.

The author of this research [5] proposes the use of an eXtreme gradient boosting tree (XGBoost) classifier to build a credit risk evaluation model for banking institutions. To deal with unbalanced data, cluster-based under-sampling is used. Finally, in comparison to other widely used classifiers like logistic regression, self-organizing algorithms, and support vector machine (SVM), the area under the ROC and classification accuracy are the assessment indicators. The findings of this study show that XGBoost classifier employed in this research outperforms other 3 and could be a better tool for developing credit risk models for banking institutions.

The authors of this research study [6] developed a system that uses past banking data to forecast whether a loan application will be granted or rejected. They compared classification methods such as XGBoost, random forest, and decision tree in their study. Variables such as applicant income, loan amount, credit history, education, and so on were used to develop the model. After cleaning the data, the authors used classification models, and using the XGboost Classification, they were able to get the greatest accuracy of 78 percent.

The author of research paper [7] built a model to forecast personal loan evaluation based on historical data. In some cases, feature selection does not produce satisfactory results. Authors employed feature selection approaches such as logistic regression, AIC-Logistic regression, and BIC-Logistic regression to solve this problem. When it comes to feature selection, logistic regression performs well. After selecting features and cleaning the data, the author compared the performance of the three classification algorithms. Age, gender, land area loan credit history, and other characteristics were chosen. The authors used algorithms such as XGboost, decision trees, and KNN to get the maximum accuracy of roughly 88 percent utilizing the XGboost model.

In this study [8], the authors offer a system in which they predict whether or not the loanee would repay the loan. The estimate is based on loan data from the bank that is providing the loan as well as the demographics of the loanee. They used the XGboost Classification model for this. Data science Nigeria gave them with the information they needed. They employed features in three sections: Demographics, which contains personal information about the loanee, loan performance, which contains further information about the application and a default flag, and past loan data, which contains previous loan applications. The authors used the XGboost classification model to forecast loan default after cleaning the dataset, and they were able to reach an accuracy of roughly 80

The author of this study [9] created a way for predicting whether or not a person should be granted a loan. The author employed a hybrid model based on the random forest and XGboost models in this study. They employed random forest for feature selection and XGboost for prediction in this study. They used the Applicant score, which is derived using annual income, indebtedness, employment length, and other factors. They used four prediction models to compare the results, with the XGboost model achieving the greatest accuracy of 91 percent.

The authors of this study [10] built a model that predicts whether a person will repay a loan after it has been approved based on loanee data. The author used historical data for creating this model. The data they used in this investigation came from the loan data of lending clubs. They needed to clean the data before applying the model because there were many missing values and some of the columns were unnecessary. They employed statistical approaches to fill in the missing values. The feature selection process is broken down into three parts. First, they employed the standardization method, then numerical mapping, and finally extracting variables with model relevance from the original variables. They employed three distinct classification methods to develop the model: XGboost, decision tree, and logistic regression, with the XGboost doing the best, with an accuracy of roughly 98 percent.

II. Implementation

In this part of our project design, we are focusing on the potential vehicle loan default prediction which we are making use of Knowledge Discovery in Databases (KDD) methodology for Selection, Preprocessing, Transformation, Data Mining, and Interpretation / Evaluation. This method has been decided for the historical dataset of vehicle loan default and considers our target variable is categorical and literature reviews of loan default prediction model reference hence we have decided to use the XGBoost model in the task. The steps can see in Fig. 1

A. Data Injection

We have got the Train data and we injected the vehicle loan dataset to the Jupyter notebook can see Fig. 2.

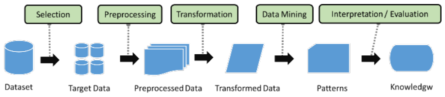
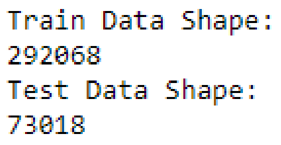


Fig. 1. KDD method



```
Train Data Shape:
292068
Test Data Shape:
73018
```

Fig. 2. Data shape

B. Data Cleaning

From the data amount there are 40 columns but not all the columns can help us to determine the loan. We have used the heatmap() function to check the correlation of data to clean some columns which have lack relevance with loan default can see Fig. 3. Hence we can dig into the data.

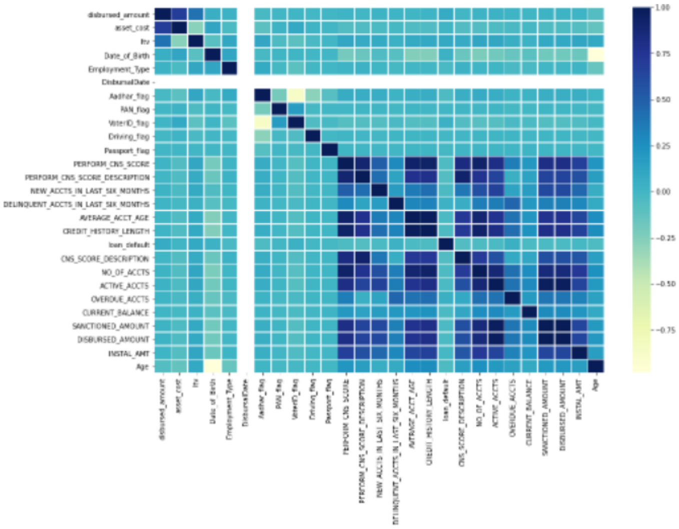


Fig. 3. Correlation heatmap

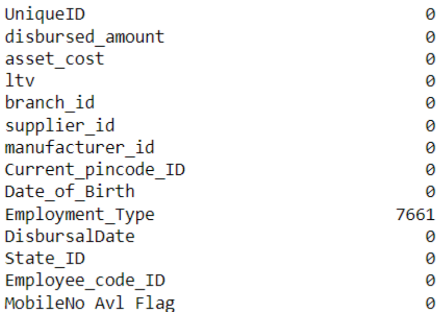
There are ID columns which we have decided to drop the unique value because these values do not help us to determine the loan and some columns are not relevant to Data Processing such as UniqueID, MobileNo_Avl_Flag, Current_pincode_ID, Employee_code_ID, manufacturer_id, supplier_id and so on.

•Missing Value

There are 7661 missing values in the column of Employment Type and the crosstab() function to check the frequency of columns employment type and loan default and based on this frequency we filled Salaried replace the missing value can see Fig. 4.

•Outliers

The Primary and Secondary Accounts have been combined and created new columns such as



```
UniqueID                    0
disbursed_amount            0
asset_cost                  0
ltv                         0
branch_id                   0
supplier_id                 0
manufacturer_id             0
Current_pincode_ID          0
Date_of_Birth               0
Employment_Type          7661
DisbursalDate               0
State_ID                    0
Employee_code_ID            0
MobileNo_Avl_Flag           0
```

Fig. 4. Missing value

ACTIVE_ACCTS NO_OF_ACCTS OVERDUE_ACCTS CURRENT_BALANCE and we have found outliers from these new columns and dropped the outliers by using use mode() function can see Fig. 5

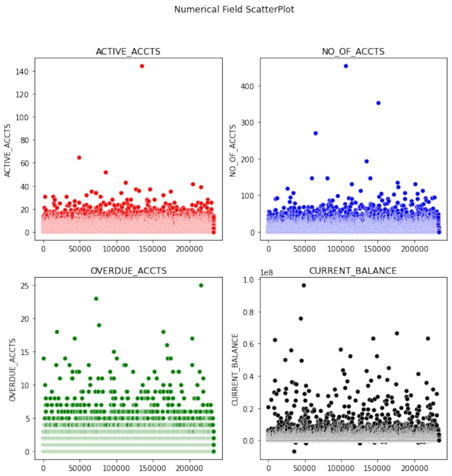
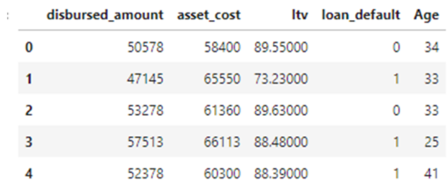


Fig. 5. Visualization on outlier

C. Data Pre-processing and Transformation

We got to the conclusion that data transformation is needed in our analysis and perhaps plays a major role. Data transformation in KDD refers to the process of converting raw data from unstructured to structured format. This phase is crucial in data management and integration because it brings the data to the point where it can be directly injected into the ML Model. The first transformation which we carried out was for the date to age conversion as the date field won't be of much use if we use it as it is. Changing all the date field to age field which can be helpful in our prediction. The fig below shows the transformed date fields. The CNS Score columns have been transformed into Common Group as Numeric Features which can show the risk scores of clients of loan.

In the next step, with the help of box_cox, we have transformed 16 continuous numerical variables such as assest_cost, disbursed_amount, age and so on into data that resembles a normal distribution as nearly as possible. Because modifications may remove white noise, as they can increase the prediction potential of our models.



	disbursed_amount	asset_cost	ltv	loan_default	Age
0	50578	58400	89.55000	0	34
1	47145	65550	73.23000	1	33
2	53278	61360	89.63000	0	33
3	57513	66113	88.48000	1	25
4	52378	60300	88.39000	1	41

Fig. 6. Date field transformed

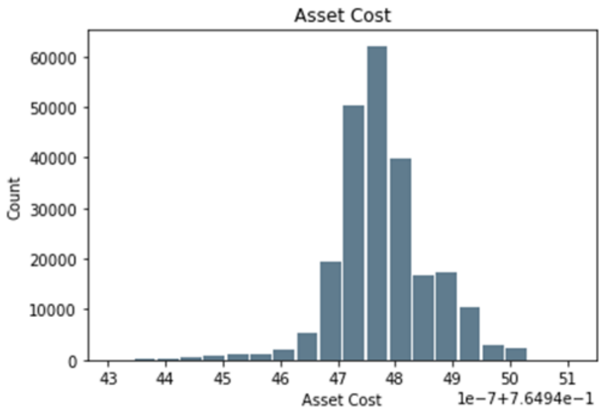


Fig. 7. Numerical field transformed into Normal Distribution

As discussed in the related work section, our dataset suffers from a problem of class imbalance i.e., the positive cases of defaulters are less as compared to the negative cases. Due to this, the model may fail to predict the positive class as the model has not been exposed to such cases. Resampling data is amongst the most often used methods for dealing with an unbalanced dataset. Resampling are of two types under-sampling and oversampling. Synthetic Minority Oversampling Technique (SMOTE) is just an oversampling approach that generates synthetic samples for such minority class. This approach aids in overcoming the overfitting issue caused by random oversampling. It concentrates on the feature space to produce new examples by interpolating between positive instances that are close together. After getting the balanced dataset the only part as shown in fig 8, which is pending would be to split the dataset into train and test sets. Hence, we have split the dataset in 80:20 ratio. The model will train on train data and will be evaluated on the test data

D. Model Implementation

Now we'll work on refining our business model. There are numerous machine learning algorithms available to solve classification problems. Many studies have shown that the XGboost model is effective for categorization. We will construct a model to forecast whether a loanee will default or

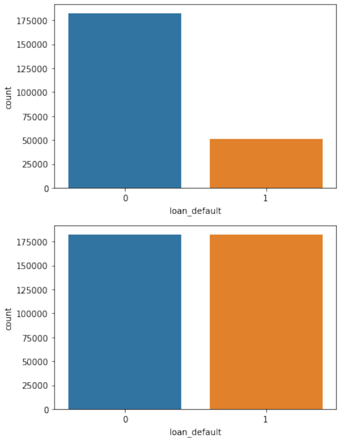


Fig. 8. SMOTE Transformation

not as part of our research. We can observe that the XGboost has always done better for dichotomous variable prediction and loan-related research, as evidenced by the research from research and investigation section. That's why we have used the XGboost Algorithm to create our business model.

III. Research Findings and Interpretation

We have used the Xgboost model for our loan defaulter's prediction and we were able to achieve an accuracy of 83.71 % and a precision of 90.76 %. The non -defaulters' prediction gave an accuracy of 92.41 % and 74.96 % of defaulters were correctly classified as shown in fig 9.

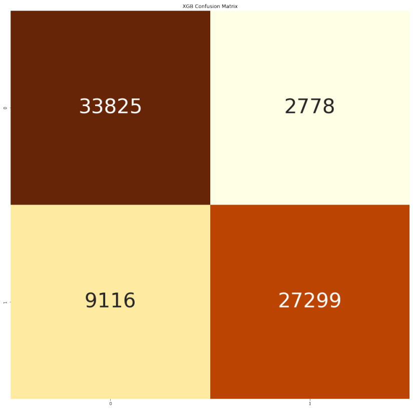


Fig. 9. Confusion Matrix

The main objective of our business is customer satisfaction and avoiding unnecessary inconvenience to our genuine customers, at the same time we want to reduce the losses caused by the defaulters. Even though our objective is defaulter prediction our main priority is to avoid inconvenience to our genuine customers and make our process smoother for them. Our model prediction is built in such a way that permits a few false cases for defaulters as a trade-off to avoid the classification of non-defaulters into the defaulter class and to make the process difficult for them unnecessarily. Thus, our prediction has high accuracy in classifying non-defaulters compared to a bit low but still very good accuracy in classifying defaulters. Our model can be used by agents to analyze the features and documents of any given user and if there is a likelihood of default predicted by our model, then the agent can employ more rigorous and in-depth background checks to make sure that the customer is eligible for repayment or not. This model can also be used to smoothen the process of users who are predicted to be non-defaulters to increase customer satisfaction and improve the quality of service. Hence our model Can be used as a tool to speed up and simplify the process of non-defaulters to increase customer satisfaction and at the same time can be used to reduce losses caused by defaulted by carrying out stringent and in-depth checks and background verification of predicted defaulters. Thus, our model will serve two purposes at the same time and can have a very positive impact if introduced in the customer review process

References

[1] T. Chen, ”Credit Default Risk Prediction of Lenders with Resampling Methods,” 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2021, pp. 123-127, doi: 10.1109/MLBDBI54094.2021.00032.

[2] R. Dwidarma, S. D. Permai and J. Harefa, ”Comparison of Logistic Regression and XGBoost for Predicting Potential Debtors,” 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 2021, pp. 1-6, doi: 10.1109/AiDAS53897.2021.9574350.

[3] L. Ke et al., ”Loan Repayment Behavior Prediction of Provident Fund Users Using a Stacking-Based Model,” 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), 2021, pp. 37-43, doi: 10.1109/ICCCBDA51879.2021.9442613.

[4] Y. Li, ”Credit Risk Prediction Based on Machine Learning Methods,” 2019 14th International Conference on Computer Science Education (ICCSE), 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

[5] Chang, Y.C., Chang, K.H. and Wu, G.J., 2018. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Applied Soft Computing, 73, pp.914-920.

[6] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, ”Prediction of Modernized Loan Approval System Based on Machine Learning Approach,” 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-4, doi: 10.1109/CONIT51480.2021.9498475.

[7] Wang, K., Li, M., Cheng, J., Zhou, X. and Li, G., 2022. Research on personal credit risk evaluation based on XGBoost. Procedia Computer Science, 199, pp.1128-1135.

[8] Odegua, R., 2020. Predicting Bank Loan Default with Extreme Gradient Boosting. arXiv preprint arXiv:2002.02011.

[9] Koduru, M., PranatiChunduri, M., Phanidhar, M. and Srinivas, D.K., 2020. RF-XGBoost Model for Loan Application Scoring in Non Banking Financial Institutions. International Journal of Engineering Research Technology (IJERT) ISSN, pp.2278-0181.

[10] Li, Z., Li, S., Li, Z., Hu, Y. and Gao, H., 2021, April. Application of XGBoost in P2P Default Prediction. In Journal of Physics: Conference Series (Vol. 1871, No. 1, p. 012115). IOP Publishing.