National College of Ireland

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:	21-02-2022
Project Title:	Domain Application of Predictive Analytics Project Design
Word Count:	1926

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:	21-02-2022

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):	

Domain Application of Predictive Analytics Project Design

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*—In this research, we will use machine learning and business intelligence approaches to conduct analysis for Automobile loan defaulter prediction. We will be utilizing open source LT vehicle loan application dataset with a range of attributes in our research. We will conduct our explanatory analysis using Power BI visualisations while keeping numerous ethical issues in mind. We are offering several machine learning and statistical methodologies for loan default prediction, as well as the commercial value of this project.**

***Index Terms*—Vehicle loan, Defaulter, Machine Learning, Python, Power-BI, Banking**

I. Introduction

People throughout the whole world depend on financial institutions to loan them funds for a multitude of reasons., including overcoming financial limits and achieving personal ambitions. Banks and financial lending institutions perform critical roles in the credit system business. The demand for vehicles is rising every day, and with it, the need for financial financing, since there are many banks and financial lending organizations that offer loans for car purchases. Various bank or money lending institutions provide vehicle loans depending on the individual's credit history, value of the assets, and debt-to-income ratio which are provided to the company. There are lot firms who depends on the loans for their profit as the loan procedure includes the lot of fees and also the interest on the loan. There is, however, a major risk because the borrower's repayments are never guaranteed. In these circumstances, banks and financial institutions struggle and incur losses. If we can handle this loan in a planned manner, it will benefit both the banks and the borrower. We can use the predictive analytics in this. Understanding the borrower's credit score necessitates a thorough understanding of statistics. So, in order to grant a loan, banks rely on the expert who makes the decision. However, many banks and financial businesses now employ machine learning and deep learning to interpret the credit score and evaluate whether or not the borrower is capable of repaying the loan. The primary goal of this research is to build the machine learning model which helps in the loan lending process and determine the optimal way for a financial institution that properly identifies whom to lend to and assists banks in identifying loan defaulters for much-reduced credit risk [1]

Fig. 1. Word Cloud

II. Project Dataset

We are utilizing LT automobile loan data to do this project. We are forecasting the default loan borrower using a machine learning approach.. This Dataset is public dataset which we have got from the Kaggle. In this we have 2 data files, one is the training data which we will use for training the model and another data is test data which we will use for testing and forecasting the loan defaulters. The dataset consist of the historical data from the LT loan firm.
This dataset consist of 41 features. It have various information which is as follows:

1) Loanee information such as age, identification verification in the flag system, and so on.
2) Loan information such as disbursement data, loan to value ratio, and so on
3) Credit data and history which includes credit score, number of current accounts, status of other loans, credit history, and so on.

Using this data, we will get relevant insights using business analytics and we will be doing predictive analysis using loan default as our dependent variable through Machine learning approaches.

III. PROJECT GOAL

The implementation of this project will help us achieve the goals mentioned below.

- Finding features that affect the loan default.
- Developing visualisations that will provide insights into our data.
- Building machine learning models to provide a solution for vehicle loan default prediction.

IV. ETHICAL CONCERNS

While carrying out our research we had to consider a lot of ethics. We needed to make sure that we discriminate based on race, color, creed gender, or community while prediction of defaulters. We also needed to make sure the privacy and anonymity of the participants is maintained and no data is used without prior consent. We have taken care of the following ethical concerns in our project.

- Voluntary participation
- Informed consent
- Anonymity

We have not used name of participants in our dataset to avoid the identity of the participants being disclosed in any form at any stage. We have also not used any data that has information about race,caste, creed, gender, etc

V. BUSINESS IMPORTANCE OF PROJECT

Banks and financial organisation earn their biggest profit from the interest earned on loans. People take loans for various reasons ranging from starting a new business to very individual needs like buying a house or car. But handing out loans is useless if the banks are unable to collect their EMI's. hence it is very important to be assured that the borrower will be able to pay back the loan. A credit score is a good indicator for this purpose but not enough to guarantee that the loan amount will be repaid. Hence it becomes very important the separate out the "ideal borrower"(borrower with guaranteed repayment)from possible defaulters. Loan default prediction is crucial for banks and other financial organizations to avoid possible losses. By using prior data acquired at banks and financial institutions, we can make predictions about present borrowers who may become default clients. Using machine learning techniques and predictive analytics, we can gain a decent notion of the behavioral patterns of the borrower. This will assist banks and financial institutions in reconsidering their terms and loan approval for the borrower.

VI. VISUALISATION FOR PRELIMINARY ANALYSIS

From the figure 2 of default loan percentage, we have found there is 22.747 percent of loan cases are defaulters and the total default loan account is 3 billion. This number shows in the financial institutions the default of vehicle loans has a major loss. The prediction will help us to reduce this amount in the future . According to figure 3 he amount of Self-employed candidates defaulting is higher compared to a salaried person. This might be simply because of the fact that

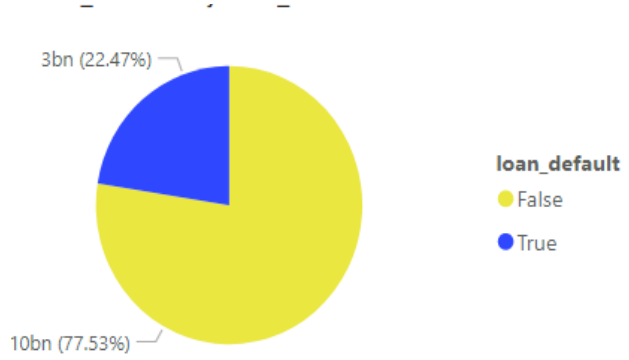

Fig. 2. Default loan cases percentage

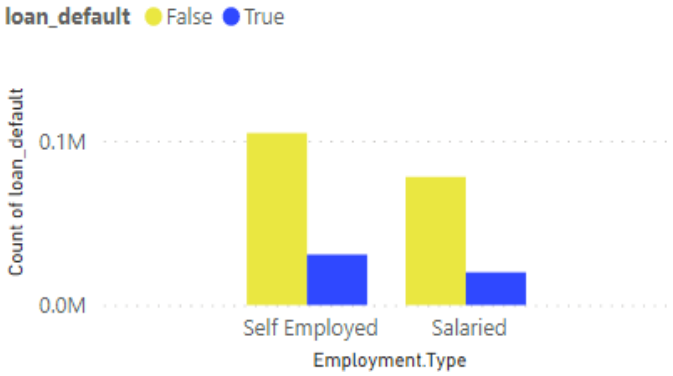

Fig. 3. Employment Type of candidates and default status

self-employed people have a high share in taking loans or due to the fact that they might not have a stable income. Take

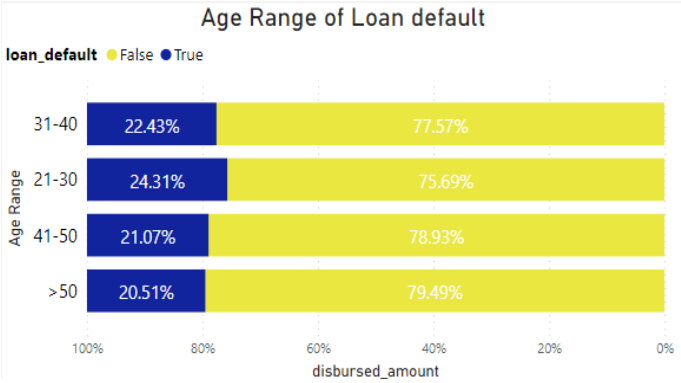

Fig. 4. Age range based classification of candidates

a look at figure 4 The age between 21-30 has the highest percentage of defaulters which is 24.31% compared to other age groups. This might be the case of inexperience and lack of financial education to handle debts. We can see the delinquent accounts of the last six months in figure 5 which shows us the number of times the candidate has defaulted in the last six months. The number of people taking the loans and defaulting is the number of people who have never defaulted before.

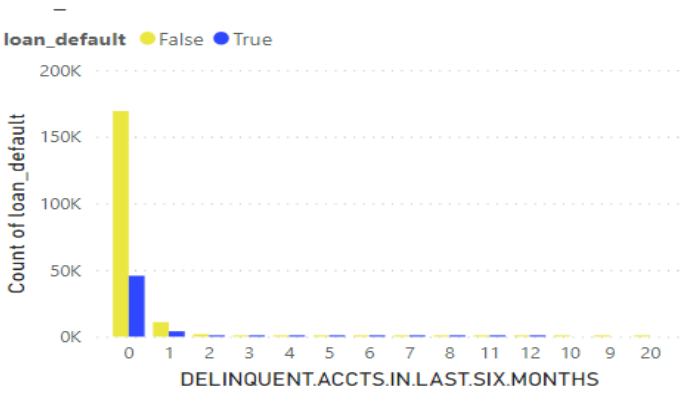

Fig. 5. Deliquent accounts in last six months

This might be because people who always pay their debt on time come in this group who do their due diligence to never miss a payment but this group can also be comprised of new inexperienced people who have taken debt for the first time and might have the most share in loan defaulters due to their lack of knowledge and proper planning From figure 6 we

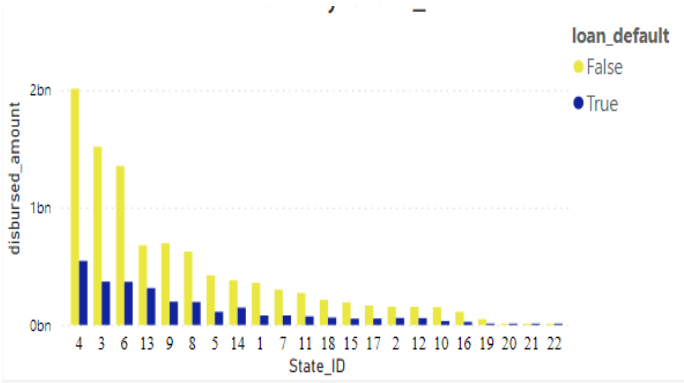

Fig. 6. State wise distribution of loan

could find that the car loan amount in State ID 4, 3, and 6 is much higher than that in other areas but the repayment ability is also quite strong in these areas. In addition, we have noticed that area 13 would have high default amount compared to the repayment amount.

VII. APPLICABLE PREDICTION TECHNIQUES

When compared to other business sectors, banking sector create massive amounts of data as compared to others, due to this they need to utilise the data for analysis and prediction for growth and planning. For these institutions, loan default prediction is a priority task as it helps them in better decision making. These institutes, in general, hold two kinds of information: soft data and hard data. Customers' transactions, statements, and other hard details are examples of hard data. This information is quantitative and is an important aspect in predicting credit risk for banks. Soft data, on the other hand, is less freely accessible and comprises the organization's policies as well as other hidden characteristics connected to the customer's application. In this study, we hope to emphasize the relevance of data in loan default prediction []

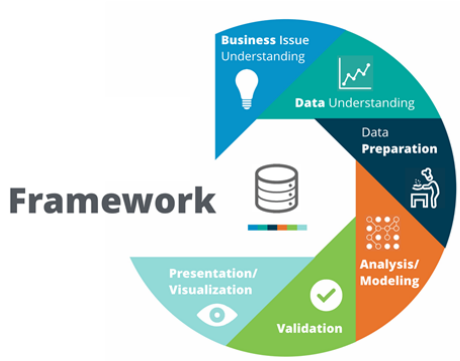

Fig. 7. Data Mining Methodology

We have a basic understanding of the patterns existing in the data by examining the infographics presented above. We will explore different machine learning algorithms depending on the features of the dataset and the nature of our target variable. The first step in building any machine learning model is the identification of features for further analysis. We would be able to determine the most relevant factors necessary for prediction by studying the relations between the independent variable and the dependent variable. Pearson's correlation matrix can be used for this. Because our dataset comprises a variety of categorical variables, this technique might be preferable to feature selection techniques such as Chi-square Test or Principal component analysis (PCA). Following the feature selection process, we will select the best fit machine learning model for our predictions.

Because we have statistical data, supervised machine learning approaches would be the ideal fit for our scenario. Furthermore, because our dependent variable is categorical, classification-based machine learning approaches would be suitable in our scenario.

The logistic regression is the most basic yet simplest type of classification technique in machine learning hence, for our model to be cost effective and easy to understand we will be using it as our first model. The next model which we plan to utilize for this project is the Random Forest, which is a type of decision tree regression, due to amount of accuracy it provides as it generates N number of trees (N is a user defined value for no. of trees). Both the decision tree and the logistic regression methods provide outcomes that are comparable to one another. When these models are trained on lesser datasets, there is a significant variation in their performance. The XGBoost classifier is another method that is well suited for classification issues. Unlike gradient boosting methods, this employs a multi-threading technique to make the best use of available hardware, resulting in more precise and quicker results. [2]

We will evaluate multiple supervised machine learning algorithms based on performance factors such as precision,

accuracy, and recall in order to attain greater prediction results. Based on the past research, we could prioritize the machine learning algorithms in order of importance. [3] [4]
Following the selection of an optimal machine learning algorithm based on our requirements, we will execute our research in python programming. To accomplish our objective in this project, we would take the following steps:

1) Data injection in Jupyter notebook
2) Data cleaning
3) Data Transformation
4) Splitting data in train and test datasets
5) Applying different ML models on Train data
6) Evaluation model performance on Test data
7) Creating insightful visualizations based on predictions

REFERENCES

[1] "ShieldSquare Captcha", Iopscience.iop.org, 2022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012042/pdf.

[2] Arxiv.org, 2022. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2002/2002.02011.pdf.

[3] P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277458.

[4] Scitepress.org, 2022. [Online]. Available: https://www.scitepress.org/Papers/2018/68724/68724.pdf.