BSc (Hons) in Information Technology Data Science
3rd year 1st Semester
Faculty of Computing
SLIIT

Fundamentals of Data Mining
IT3051

Mini Project – Statement of Work Document
Group 4
(Mining Minds)

Name	IT Number	Group
Fernando W.D.A	IT22223708	Y3.S1.WD.DS.01.01
Gunathilaka.P.A.S.R	IT22136824	Y3.S1.WD.DS.01.02
A.D.Oshadhi Vibodha	IT22363770	Y3.S1.WD.DS.01.01
Daraniyagala G.K	IT22360182	Y3.S1.WD.DS.01.01
Gamage D.M.G.P.K	IT22188472	Y3.S1.WD.DS.01.02

Table of Contents

1. Project Overview

NBFIs or NBFCs which are the nonbanking financial companies that do not own a full banking license and do not come under the scope of different banking regulations for banking entities. However, these institutions perform functions that are crucial for economic accessibility of resources such as investment consultancy, risk pooling, saving and marketing brokering. One of its advantages is that it is able to accept clients with less restrictive credit requirements as those of the traditional banks. There are many types of NBFI or NBFC. Some of the most familiar are

1. Security and commodity entities as Example brokers/dealers, investment advisors, mutual funds, hedge funds or commodity traders.
2. Money services businesses (MSB).
3. Insurance companies.
4. Loan or finance companies.
5. Managers of credit card systems.

Automobile loan is a crucial necessity in everyone's life, whether it's for transport, commuting or even carrying out business activities, and for this reason customers approach companies to provide them with automobile loans. NBFIs act as a source for these clients wherein they get loans for their different financial needs. However, the opportunities of NBFIs attract more applicants due to lower credit standards and shorter credit term contracts which, in turn, leads to higher default opportunities. defaults can occur for various reasons

1. Sometimes, number of factors like loss of a job, sickness, or any other financial hardship, the client will be unable to make his/her loan payments.
2. Some clients take additional loans they cannot afford to service ultimately becoming delinquent with repaying the loans.
3. Poor management of finances or having unachievable goals as to income and expenditure can also result to a client's default.
4. Any background economic indicators like such like recession or inflation the clients will be forced to pay less on their loans due to less disposable income.

Stakeholders and Beneficiaries

The primary stakeholders in this study are the NBFIs who offer automobile loans to the consumers. Through the implementation of the predictive model applied in this project, these institutions will be in a much better position to minimize their risk of loan defaults hence increase operational profitability and stability. Clients who are less likely to default can benefit from better loan terms and increased access to credit. Also, the regulators and other policymakers will benefit from the stabilization of the lending platform along with better health of the sector.

Business Need for the Model

A specific vehicle loan defaulting client model has to be developed by the NBFI. This model will assist the NBFI in managing the risk of defaulting loans by enhancing the loan evaluation process, determination of credit limits, and initiating default risk reduction plans. Thus, the NBFI lacks a big-picture understanding of what makes clients repay the loans and therefore optimizes its lending practices and profitability.

Problems in Selecting the Dataset -

- Data Quality: The dataset might contain missing or erroneous values, which could lead to inaccurate modelling.
- Data Imbalance: If there are more non-default cases as compared to default cases in the dataset, the model could be biased.
- Feature Relevance: The data may have variables that do not have relevance in predicting a default, thus adding noise.
- Information Hazard: Special consideration of the laws on data protection is necessary since sensitive personal and financial data is involved.

Goal

A vehicle financing quant will develop a predictive model that helps the loan originator to determine to a very high degree of accuracy whether the client would default in settling a loan for a vehicle. The model concentrates on providing the NBFI with valuable information by analyzing various aspects which discourage loan defaulters and assist the institution in maintaining its financial health.

Solutions

- Data Preprocessing: Appropriately manage your data for any missing data values and check for uniformity.
- Feature Engineering: Focus on those characteristics that are more likely to lead to defaults on loans and perhaps develop additional ones.
- Model Development: Implement machine learning techniques such as logistic regression, decision trees or random forests to construct the prediction model for validation.
- Evaluation and Tuning: Assess the model against the F1 Score and adjust where necessary to enhance performance further.

Limitations

- Data limitations: The given dataset might not be sufficient in all the relevant aspects existing in other datasets causing deviation from the targeted model.
- Model generalizability: The model will not work satisfactorily with another dataset because it has been too well trained/overfitted the model.
- Economic Changes: The model is unable to take into consideration any such events that happen swiftly that can impair the ability of clients to repay.
- Assumptions: This model is based on the premise that one's past conduct is a good predictor of how one would behave in the future.

- **Data set –** **Automobile Loan Default Dataset**

2. Scope of work

This process focuses on 5 main stages

1. A suitable dataset selection stage
2. Data preprocessing stage
3. Model building stage
4. Model evaluation stage (Testing)
5. Data visualizing stage
6. User interface design stage

A brief explanation of the above stages is given below.

1. A suitable dataset selection stage

When choosing a dataset, you should be careful about its structure, attributes, features, and data quality. It is important to further explore the dataset selected according to these criteria and identify the problem implied by it and focus on ways to solve it. That is, we must have a basic understanding of the problem that can be solved by the dataset we choose, the way to solve it and the predictions that can be given to the end users.

2. Data preprocessing stage

Data preprocessing is an important stage in this process. It begins with **data cleaning**, where errors, inconsistencies, and irrelevant information are addressed by handling missing values, correcting outliers, and removing duplicates. Also, **data transformation** converts the data into a suitable format, involving tasks like normalization and encoding categorical variables. **Feature engineering** Create new features that can improve model performance by Combine existing features. **Feature selection** identifies and retains the most relevant features for the model by removing features with low variance. Finally, we can **Split data** into subsets for training, validation, and testing.

3. Model building stage

Model building is another important stage. We should choose an appropriate algorithm based on our problem such as regression for predicting continuous values, classification for categorizing data into classes, or clustering for grouping similar data points. After that we should train the model by using training data to minimize the error between predictions and actual outcomes. This stage directly contributes to accuracy, reliability and generates solutions to predict the desired outcomes.

4. Model evaluation stage (Testing)

This phase focuses on the testing part of the process. It involves testing the trained model to evaluate how well the model performs. The goal is to ensure that the model performs effectively and reliably in realworld scenarios.

5. Data visualizing stage

In this data visualizing stage users can graphically view the predicted outcomes by using charts, graphs, and plots, illustrate data patterns and the results of model predictions. It helps to get a clear understanding of the data more intuitively and communicate findings effectively.

6. User interface design stage

This is the stage in creating an intuitive and visually appealing interface that allows users to interact with a system or application easily. The main purpose of using an interface is to provide the required data to the user in a simple and userfriendly manner. First, we should integrate the model into the user interface and then deploy the system for realtime use, ensuring it's scalable and reliable. simply, this is a platform where users can input data, run predictions, and view results.

3. Activities

- **Find a real-world problem and define a solution**

The problem that we found is the struggling of NBFI's to mark the profits due to the increase in defaults in their vehicle loan category. It would be a great advantage for them if they were able to identify whether a customer would be able to repay the loan before lending them the money. So as a solution, we will be developing an application to predict whether a customer will be able to repay the loan without defaulting.

- **Check data availability and data preparation**

We were able to find a dataset for the above problem in Kaggle. Firstly, as we observed the dataset, we found it contained over 100,000 rows and 40 columns and it was not cleaned. Data preprocessing and Exploratory Data Analysis (EDA) techniques will be applied to prepare the data in a way that suits machine learning models and find the right features that affect our target variable respectively.

- **Model development and Evaluation**

We will split the preprocessed data into training and testing datasets. Each member of our team will select a classification model and train it using the training dataset. After that, we will evaluate the model's performance on new data by using the testing dataset. To measure the performance of our models, we will use performance metrices such as F1 score and accuracy. Furthermore, we will be fine tuning, feature engineering etc. as per the need to increase the model's accuracy on data. Finally, from the trained models, we will choose the optimal model as our final model that we will be using.

- **Frontend development and Deployment**

As the last step in our project, we will build user interfaces to give a better user experience and to reduce the technical complexity and deploy the application so that the client could use the solution (Automobile Loan Default prediction).

4. Approach

First, we choose a dataset. Then we go through the dataset and after that we decided to clean the data so that we can made a model to it. We plan to build five models using two different techniques used for binary classification. Then compare the accuracy of the models and proceed to build a UI to enter the properties used for the prediction and get the predicted value for the given data using the best model.

Dataset: https://www.kaggle.com/datasets/saurabhbagchi/dishnetworkhackathon

Data Preprocessing:

- Remove the columns without prediction power and only keep the columns that contribute to predicting whether the customer can afford to repay the loan or not
- Remove rows with null values
- Perform data normalization, reduction, and integration operations on the dataset and divide the dataset into two a training dataset and a testing dataset.

Building the models:

- Using the training dataset five models for binary classification will be built.
- The following will be used to build the model
    - Algorithms
    - Language – Python

Analyzing and verifying the models:

- Using the testing dataset, the models will be validated, and the best model will be chosen based on model accuracy and other metrics.

Building the interface and server:

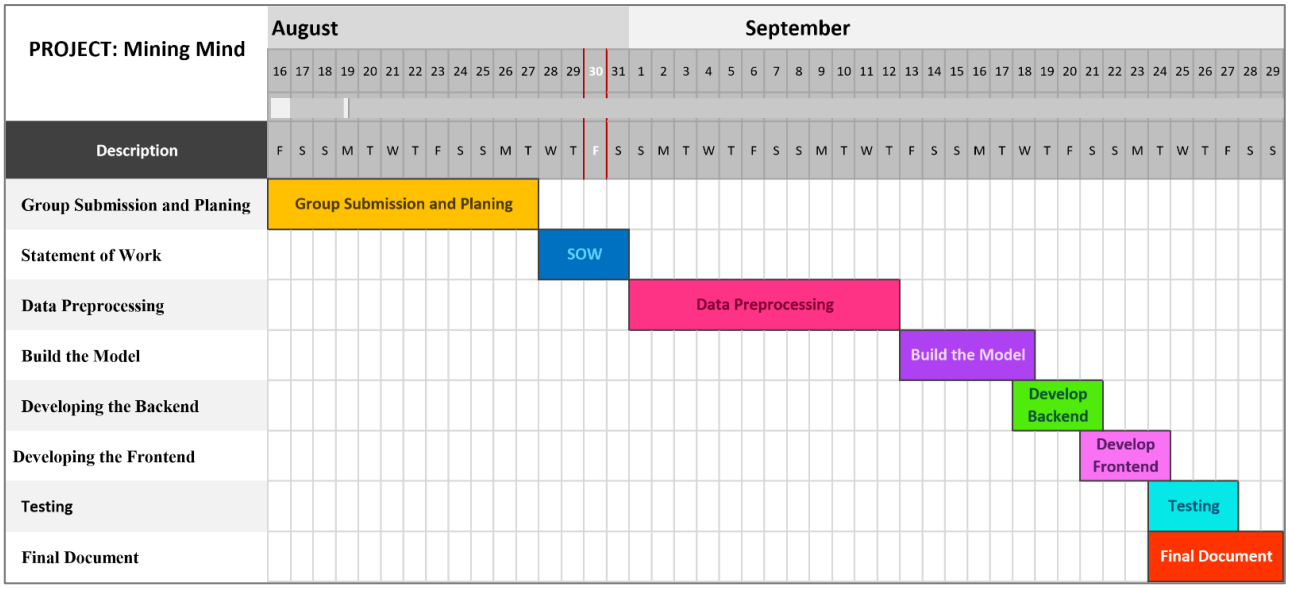
- Streamlit with python will be used for the backend.

5. Deliverables

Our primary objective of this system is to classify whether a customer will be able to repay the loan taken by him/her. This will give NBFI a better understanding of the customer's background and deny a loan if he/she won't be able to repay or accept the request if he/she is able to repay it back. Moreover, this will benefit both the organization by decreasing the losses made due to defaults and the customer by avoiding him/her being bankrupt.

Using the friendly user interface, lenders can enter the relevant customer details, and the system will output indicating whether the customer would default or not. This will help NBFI take the best decisions.

6. Project Plan & Timeline



7. Assumptions

- Assumes that the dataset is complete, meaning there are no missing or unrecorded values.
- Assumes that observations are independent of each other unless the project specifically deals with time series data or another form of dependent data.
- Assumes that the dataset is large enough to represent the underlying population and that the patterns discovered are statistically significant.
- Assumes that the data mining project complies with legal and ethical guidelines, ensuring that privacy and confidentiality are maintained.

8. Project Team, Roles, and Responsibilities.

	IT Number	Member Name	Role	Responsibilities
1	IT22223708	Fernando W.D.A	• Team Leader • Developer • Business Analyst	• Model implementation and Testing • Data analysis and process • Handle documentation
2	IT22136824	Gunathilaka.P.A.S.R	• Developer • Business Analyst	• Model implementation and Testing • Data analysis and process • Handle documentation
3	IT22363770	A.D.Oshadhi Vibodha	• Developer • Business Analyst	• Model implementation and Testing • Data analysis and process • Handle documentation
4	IT22360182	Daraniyagala G.K	• Developer • Solution Tester	• Test the model • UI development • Handle documentation
5	IT22188472	Gamage D.M.G.P.K	• Developer • Solution Tester	• Test the model • UI development • Handle documentation

***