Sri Lanka Institute of Information Technology

- Fundamentals of Data Mining -

[IT3051]

Mini Project – Statement of Work Document 2022

Group – G07

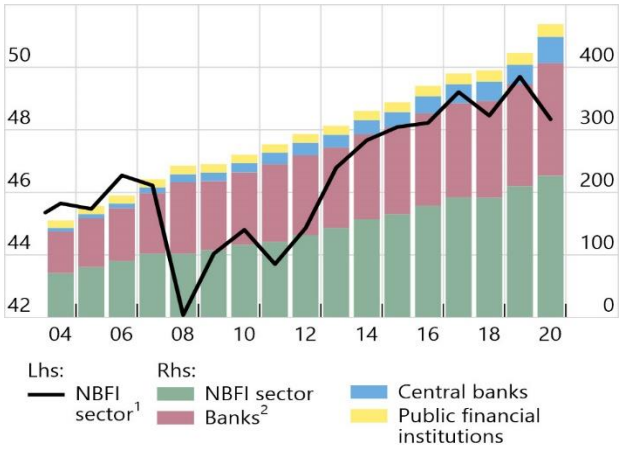
Jayasooriya C.A	**-**	**IT20250942**
Amanullath M.U	**-**	**IT20155520**
Gavindya N.A.C	**-**	**IT20409982**
Bandara T.M.Y.M	**-**	**IT20492052**
Rathnaweera R.P.W.G	**-**	**IT20237554**

Table of Contents

Background

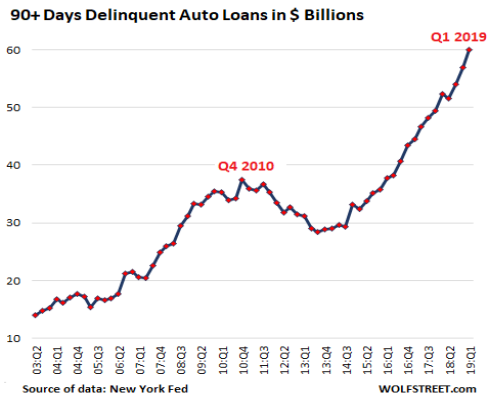
In today's context due to the current financial crisis and with the increased inflation rate and the increase in the interest rates, consumers have been stretched thin with regard to financial strength. This has created issues within the financial system.

Banks and Non-banking Financial Institutions (NBFI) are the pillars of a country's economy, if these institutions are to fail, the entire economy will face the threat of collapse. With the above-said reasons, one of the main threats to the current financial system is customers defaulting on the obtained loans. This has mainly affected the NBFIs as they do not have the strongest support from the Central Banks like Banking institutions.



Effect of financial institutions due to loan default

With the problems and inadequacy in public transport and every household needing a private vehicle, out of the overall loans, a high volume of loans belongs to the Automobile / Auto loan category, and they are the most vulnerable for customers defaulting, as the majority do not provide any financial return. Thus, defaulting on automobile loans has increased significantly in recent times and NBFI profit margins have been hit largely due to this as they are the prominent lenders of auto loans.



After a careful analysis of the loan defaulting problem, it could be understood that the most suitable and most practical solution for this issue would be to find a way to identify or predict the customer's ability to pay a loan which makes them eligible or not for loans, specifically automobile loans.

As group G07 we have identified this current problem and have decided to use our expertise to find a solution for the aforementioned problem. It was decided to address this issue from the perspective of the lender and more accurately, from the perspective of an NBFI as they are the most affected party in this issue.

The following presents the real-world business problem that we have identified, the method that we have planned to use and solve the problem, and our goal for the solution.

- **Problem** - The surge in defaults in the category of auto loans is making it difficult for NBFIs to report profits. The company's goal is to detect a client's loan repayment capacity and comprehend the relative weighting of each factor that affects a borrower's capacity to repay a loan.

- **Client** – A Non-Banking Financial Institution (NBFI). An NBFI is a financial institution that does not have a full banking license or that is not overseen by a national or international banking regulatory agency. Financial services, such as lending and investing, are made more accessible through NBFI.

- **Solution** – Predict whether a client can pay the requested loan. For each customer loan request, you must predict the default.

- **Goal** - Making a model which enables the loan approver (the NBFI) to predict whether the said customer can pay the requested loan, and then with the results of the prediction, they can approve or decline the loan request which would prevent the loss of profits due to defaulting of the loan.

- **Dataset selected** – The following is the publicly available dataset that we have selected.

    - Kaggle | Automobile Loan Default Dataset

Scope of Work

This project consists of 5 main layers namely,

1. User Interface Layer.
2. Data wrangling and data cleansing layer.
3. Data mining layer.
4. Model building and analysis layer.
5. Data visualizing layer.

A brief explanation of the above-mentioned layers is given below.

1. **User interface layer**
   The user interface layer is the layer which is the front end where users can interact with the system, users can select data or input relevant data, that are required for analytics. This layer mainly focuses on user-friendliness for the end-users where they can interact with the backend model of the system.
   When implementing the interface layer, the goal is to use a simple questionnaire that contains a user-friendly interface. This interface plays a vital role since it interacts with end-users.

2. **Data wrangling and data cleansing layer**
   This layer performs the data cleaning and preprocessing part for the chosen data, which helps to detect and correct corrupted inaccurate records. It identifies the incomplete or irrelevant parts of data and then it replaces, modifies, or deletes those parts using relevant preprocessing techniques which would help the model give more reliable results.
   Mainly this layer helps to process data by transforming and mapping them from the raw form into other formats which are more appropriate, accurate, and valuable for the process.

3. **Data Mining layer.**
   This layer mainly focuses on the process of analyzing the datasets and gathering the data using algorithms and transforming those gathered values. Mainly it extracts information from the dataset and transforms that extracted information into a comprehensible structure that is suitable for further analysis

4. **Model building and analysis layer.**

   This layer helps model the data. Which was transformed using the data mining layer, we use this layer to build predictive models that use the selected dataset to build a mathematical solution to predict the desired outcomes from the newly gathered data.

5. **Data visualizing layer**

   This is the layer that helps to graphically represent the outcome. Using this layer users can graphically view the predicted outcomes and get a clear understanding of the results obtained by using the model we have created.

Activities

- **Finding a real-world problem and defining a solution**

Using publicly available datasets a real-world problem was found which is both current and relevant.

Kaggle was used to find the dataset for our business problem, which is NBFIs losing profits due to clients defaulting on obtained Automobile Loans.

For the above real-world problem, the team was able to come up with a solution to predict the ability of a client to pay the loan which eliminates the root cause of the problem.

- **Data preparation, model construction, and training**

As the obtained dataset is dirty preprocessing of the dataset would occur, and the chosen data set would be cleaned (null values handled), normalized, reduced (with dimensionality reduction), and prepared to suit the implementation of the solution.

After extensive research on the problem and the solution, several models were chosen which could enable the implementation of the solution.

Then as the next step, the building of the chosen models would happen and after the models have been built, they would be trained with the prepared training set of data.

- **Evaluate the model**

As multiple models have been prepared for the solution, the most suitable one will be chosen. The evaluation of the models would happen and the best model out of the candidates would be chosen for characteristics like most accurate, least error, etc.

- **Make predictions**

Using the chosen optimal model predictions would be done to solve the business problem.

- **Front-end development**

As the final step, to release the solution to the client, a front-end application would be built. This gives a better user experience and removes the technical complexity of the solution and presents the solution in a user-friendly acceptable manner.

Approach

We will start building the model from scratch. So first we chose a dataset. Then after going through the dataset, we decided on how the dataset can be cleansed before using it to build the model. We plan to build two 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: Automobile Loan Default Dataset

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 (dimensionality reduction).
- Remove rows with null values (null value handling).
- Discretize the columns with continuous values. For example, the customer income.
- 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 two models for binary classification will be built.
- The following will be used to build the model
    - Algorithms – Decision Tree and Random Forest
    - 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

- React JS will be used for the front end.
- Streamlit with python will be used for the backend.

Deliverables

The prime objective of this system is to provide an understanding to the lender, of whether the customer can repay the loan amount taken by him/her. This would benefit the organization so that it will not face any financial failures with loan defaults.

As the ultimate goal, the model and the system designed, created, and implemented by our team should be able to predict whether the client requesting the loan would be able to pay the loan without defaulting. This would enable the NBFIs to solve the problem of profit loss due to customer defaulting by denying the clients who are predicted to be unable to pay requested Automobile Loans.

Assumptions

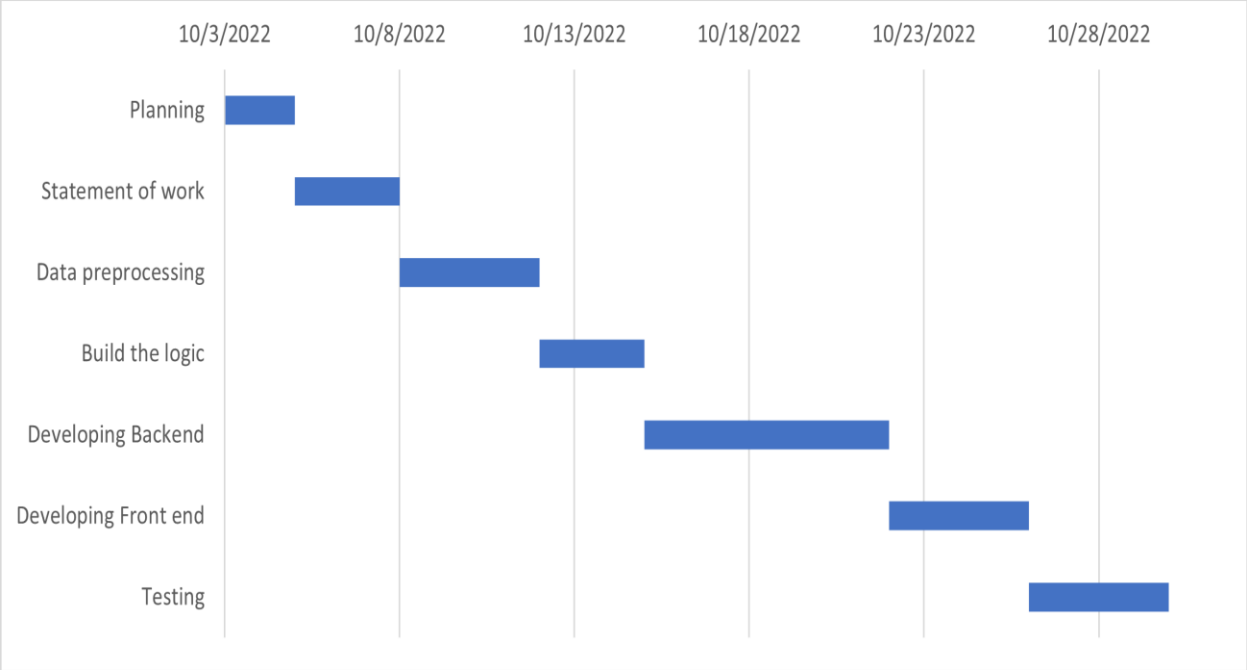
- All the source data are accurate.
- Independent and identically distributed data - true but unknown data distribution in which each of the training and test points is drawn independently.
- Initially, the whole training data is considered as root.
- Records are distributed recursively based on the attribute value.
- Assumption of no formal distributions. Being a non-parametric model can handle skewed and multi-modal data.

Project Plan & Timeline

The following project management timeline is a detailed schedule for the project. It spells out all the tasks involved and a deadline for each so that the entire team can see when individual steps will take place and when the whole project will be wrapped up.

At its core, the project timeline is an overview of the project's deliverables laid out in chronological order. It maps out what needs to be completed before a new task can commence and keeps everything ticking along nicely.

The Gantt Chart below presents the timeline in a visual format, which means stakeholders and team members can get a quick overview immediately.



Gantt Chart

Project Team, Roles, and Responsibilities

	Member IT Number	Member Name	Member Role	Member Responsibilities
1	IT20250942	Jayasooriya C.A	Team Leader Solution Developer Business Analyst	Implement model Handle documentation Test alternate model Data analysis and process UI development
2	IT20155520	Amanullath M.U	Solution Developer Solution Tester	Implement model Test alternate model Handle documentation Integration Head UI development
3	IT20409982	Gavindya N.A.C	Solution Developer Business Analyst	Implement alternate model Test model 1 Handle documentation Data visualization UI development
4	IT20492052	Bandara T.M.Y.M	Solution Developer Business Analyst	Implement alternate model Handle documentation Test model 1 Data analysis and process UI development
5	IT20237554	Rathnaweera R.P.W.G	Solution Developer Solution Tester	Implement alternate model Test model 1 Handle documentation Integration UI development