Fundamentals of Data Mining [IT3051]

Mini Project –Statement of Work Document

2024

Group Details

Name-Mining Masters

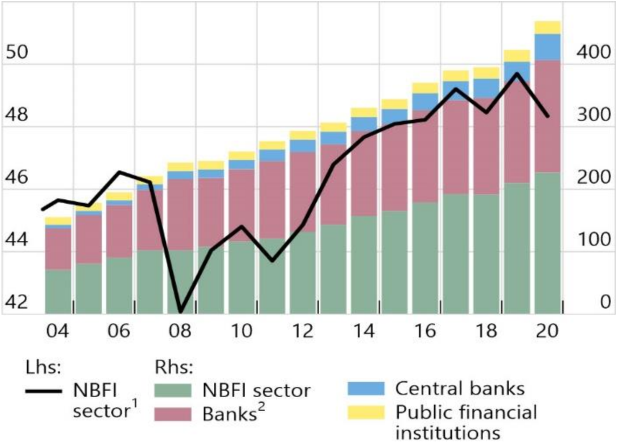
Name with Initials		Registration Number
1.	Deemantha P.H.H.C	IT22560162
2.	Jayarathna J.V.D	IT22577610
3.	Thrimanna A	IT22585530
4.	Wimalarathna B.P.K	IT22059604
5.	Pallewela S. M	IT22594136

Table of Contents

Background

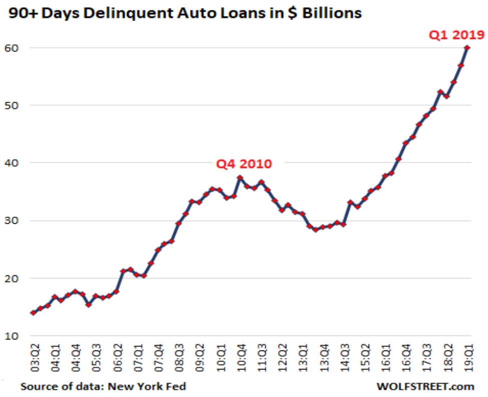
In today's context due to the current financial crisis and with the increased inflation rate and the increase in the interest rates, consumers have been stretched thin with regard to financial strength. This has created issues within the financial system.

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.


90+ Days Delinquent Auto Loans in $ Billions

Source of data: New York Fed

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**

  The current real-world business problem was identified in the financial sector. That would be facing challenges in assessing loan applications for approval.

  We used a publicly available dataset from Kaggle, which focuses on loan application data.

  The solution aims to predict whether the loan application should be approved or rejected based on the applicant's historical data, thereby minimizing the risk of approving loans that could result in defaults.

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

  The dataset we obtained was not ready for direct model implementation.

  As part of the preprocessing, data set would be cleaned (null values handled), handled missing values, removed outliers, normalized, reduced (with dimensionality reduction).

  After preprocessing, several models will be chosen based on their applicability to this problem, such as Decision Tree, GaussianNB, Logistic Regression, Random Forest, and SVM.

  Then we will build the models and train using the cleaned training dataset.

- **Evaluate the model**

  From that multiple models we must evaluated to find the optimal model.

  This evaluation was based on key performance metrics like the least error, most accurate, precision, recall, F1-score. Based on these metrics, will be select for the final implementation.

- **Make predictions**

Using the best-optimal model, we will make predictions on the test dataset to determine whether a loan should be approved or not.

- **Front-end development**

  As the final step, to make the solution to the client, the user-friendly front-end application would be built.

  It will give a better user experience and removes the technical complexity of the solution as well as it will ensure that non-technical users can easily interact with the system without needing a deep understanding of the underlying machine learning models.

Approach

We aim to build a predictive model from scratch to determine whether a loan can be approved or not.

So first we choose a data set. Then we will identify the ways to clean and prepare it. We plan to build five models using five different techniques used for binary classification.

By comparing the accuracy of the models, we will proceed to build the user interface.

Dataset: Automobile Loan Default Dataset

Data Preprocessing

- Dimensionality Reduction: We will remove columns that have no predictive power and keep only those that contribute significantly to predicting loan approval.
- Handling Missing Values: Rows with missing or null values will be removed using suitable techniques.
- Discretizing Continuous Values
- Normalization
- Data Splitting**:** The dataset will be split into training and testing sets

Building the models

- The dataset will be split into training and testing sets using Python language.

Analyzing and verifying the models

- After building the models, the testing dataset will be used for validation. We will evaluate the best model based on various metrics, including accuracy, precision, recall, F1-score.

Building the interface and server

To provide a user-friendly solution, we will develop a front-end interface where users can input loan application details and receive a prediction.

- Front-end: Developed using HTML and CSS
- Back-end: Implemented using Streamlit integrated with Python.

Deliverables

This system's main goal is to tell the lender whether the borrower can afford to return the loan amount that was taken out. The organization would gain from this in order to avoid financial failures due to loan defaults.

The ultimate goal for the model and system that our team devised, developed, and put into place is to be able to anticipate if the customer who is applying for a loan would be able to repay it without going into default. By refusing to lend money to customers who are likely to default on their auto loans, NBFIs will be able to address the issue of profit loss from defaulted loans.

Assumptions

- **Data Quality**: The dataset is clean, and any missing or erroneous data will be minimal and manageable through preprocessing steps such as imputation and normalization.

- Balanced Dataset: The dataset might have imbalanced classes (i.e., fewer default cases), requiring techniques such as SMOTE (Synthetic Minority Over-sampling Technique).

- Resource Availability: Computational resources (such as cloud or local machines) will be sufficient to handle the dataset and train models within the project timeline.

- Team Availability: All team members will contribute according to the planned timeline and will be available for meetings and collaborative work.

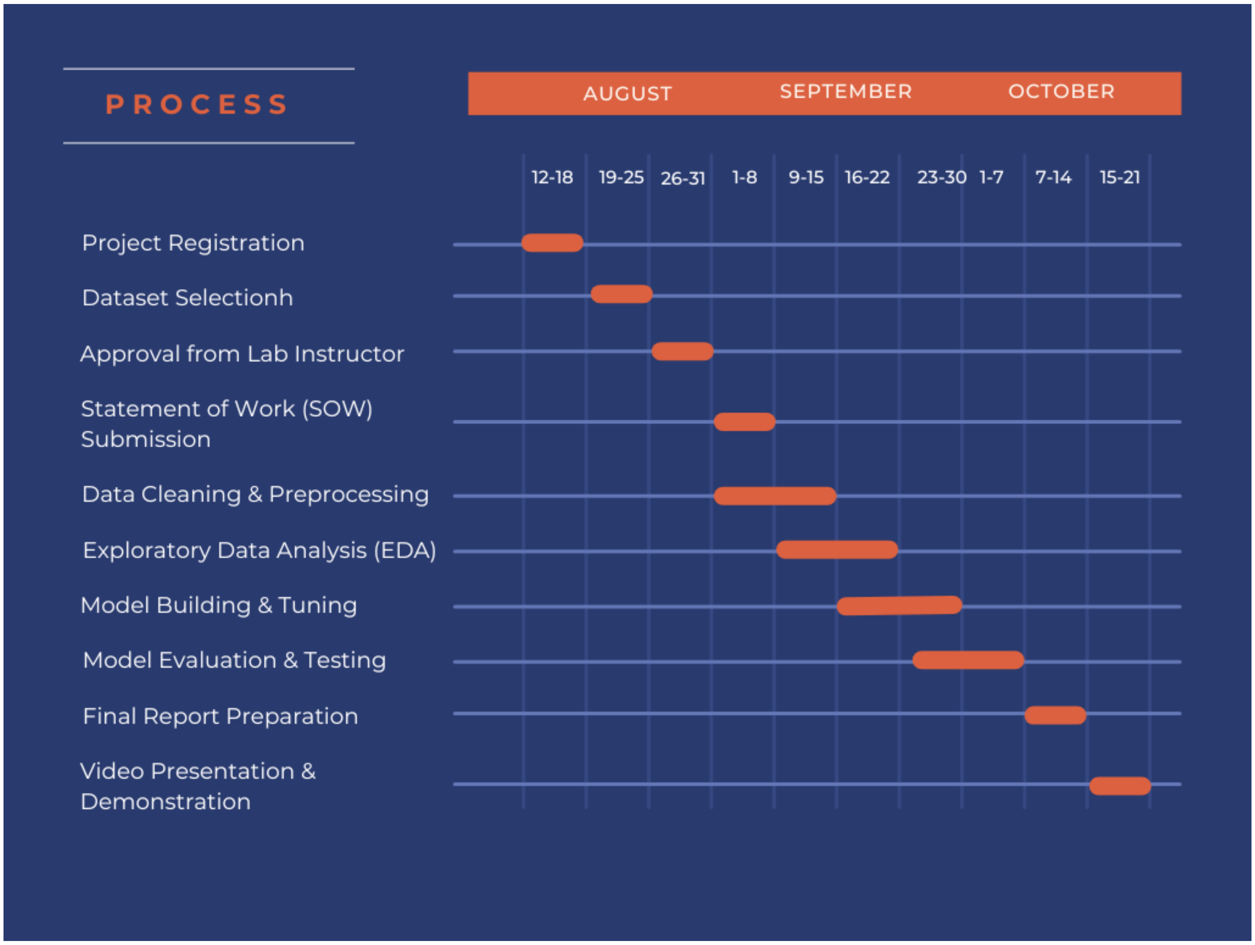
- Dataset Size: The dataset size will not be too large to require advanced distributed computing techniques like Spark.

Project Plan & Timeline

A thorough schedule for the project can be seen in the project management timeline that follows. It lays out every work that needs to be done along with a deadline so that everyone on the team knows when each step has to be completed and when the project will be finished as a whole.

The project timeline is essentially a summary of the deliverables for the project arranged chronologically. It keeps things moving along smoothly by outlining what must be done before starting a new task.

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	IT22560162	Deemantha P.H.H.C	Team Leader Solution Developer Business Analyst	Implement model Handle documentation Test alternate model Data analysis and process UI development Integration
2	IT22577610	Jayarathna J.V.D	Solution Developer Solution Tester	Implement model Test alternate model Handle documentation Data analysis and process Head UI development
3	IT22585530	Thrimanna A	Solution Developer Business Analyst	Implement alternate model Test model 1 Handle documentation Data visualization UI development
4	IT22059604	Wimalarathna B.P.K	Solution Developer Business Analyst	Implement alternate model Handle documentation Test model 1 Data analysis and process UI development
5	IT22594136	Pallewela S. M	Solution Developer Solution Tester	Implement alternate model Test model 1 Handle documentation Integration UI development