

# DATA-2206-01 Capstone

Group #6 Final Project Plan

**Credit card payment default prediction**

Professor – Thomas Francescutti

Kanishka Saxena – 100815602

Mayurkumar Chavda - 100839938

Darshanbhai Avaiya – 100848457

Sainadh Chittajallu - 100843160

Date – 9th October, 2022

**Architecture of the Product**

In this capstone project, we are building and developing an algorithm which can predict whether a client is going default next month’s credit card payment or not. First step of our project is to collect the data from UCI machine learning repository. Second step will be exploratory data analysis of dataset, third step will be to train a machine learning algorithm and in final step, we will optimize the trained model and make it ready for use.

Optimized model will have higher prediction accuracy and it will be efficient but we will have only 30,000 observations in our dataset. So, to train a robust model which can be deployed into production, we require more data. Moreover, this data is from year 2005 credit card payments and that is why we will require more data over the time period after year 2005 to build highly accurate algorithm. These are the limitations of our project.

**Project Techniques and Applications**

In our capstone project we are working on dataset about “Credit card payment default prediction”. Most banks offer their clients credit cards and lines of credit. Along with banks, several financial organisations and certain retail chains have formed partnerships with banks to provide their consumers credit cards. Credit limit or lines of credit are determined by a customer's credit history and usage, but occasionally a customer will spend excessively from their credit limit, leaving them with insufficient funds when it comes time to pay their bills. This results in a loss for the credit card provider organisation, and the assets become non-performing assets of the business.

As we all are currently pursuing data analytics for business decision making from Durham college and while pursuing this course, we came across various topics and tasks related to data analysis which helped us in knowing the significance of "DATA," which is further divided into different categories of data. The project team started by taking a set of entirely raw data, which is where the first stage of analysis starts.

There are tens of thousands of data points available for analysis and study, allowing us to examine almost every aspect of business. By shifting your viewpoint, you can offer insightful commentary that can ignite genuine change. There is a large amount of data available in credit card payment processing. There are various uses for this information.

Our Team will be working on dataset “default of credit card client’s dataset” from UCI machine learning repository. This data was gathered as part of research aimed at the case of customer’s default payments in Taiwan. Data contains 24 variables and 30,000 records. We will use this data in our project to help solving a business problem in financial institution.

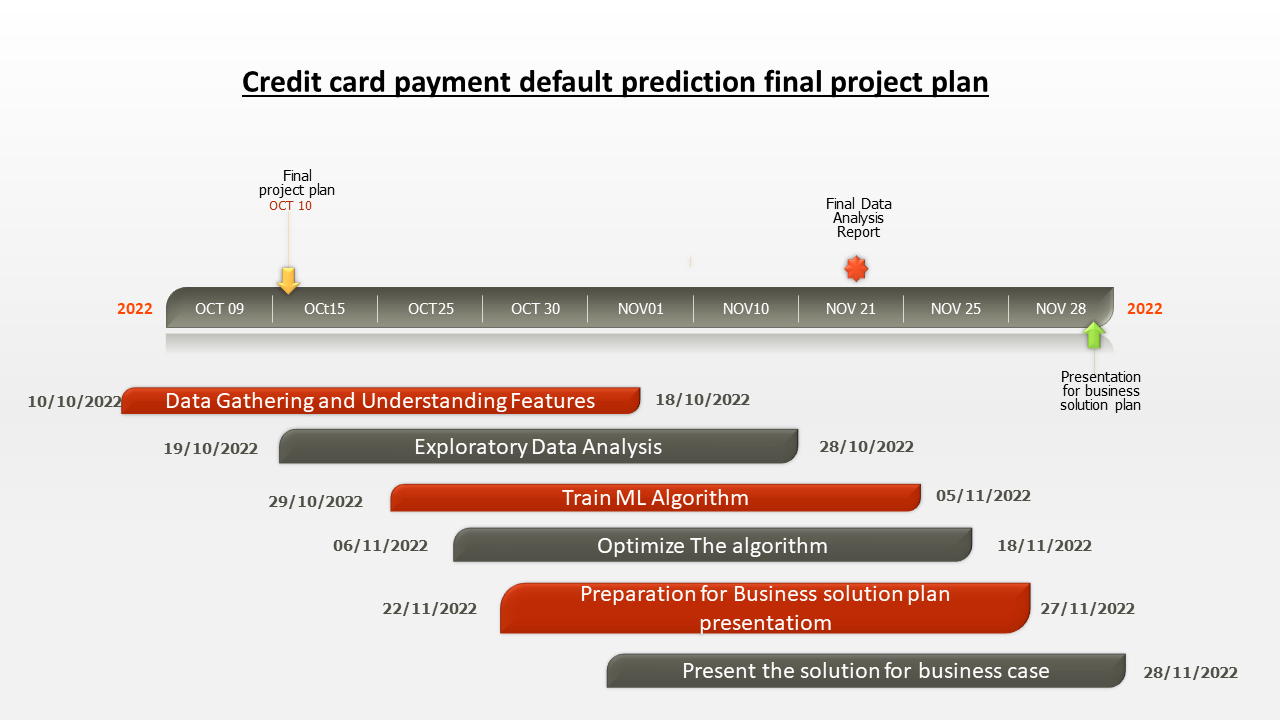
While pursuing the course we learned about various techniques and tools for data analysis and data visualization. We will use data analysis skills throughout the project, including a variety of aspects of raw data analysis, data cleansing, data distribution, datatype formatting, altering data category, removal of unwanted & null data, gaining insights into trends, removing invalid data entry, categorising the analysed data, and more.

We will try to apply analytics approach on this project which will be predictive. We will train a binary classification model which will predict whether a customer is going to default next month’s payment or not. Some of variables in dataset are categorical but we can use encoding techniques like one hot encoding to convert it into numerical columns and use it to train a model. We will use Python programming language and libraries like Pandas for data cleaning and manipulation, Scikit-learn to train machine learning algorithm and Seaborn to produce visualizations. If there is any missing data, then we will use imputation techniques to fill that data. Using visualizations like density curve and distribution plots, we will know that whether data is skewed or not. If data is skewed, then we can use normalization or transformation methods to normalize the dataset.

On this project we will try implement feature engineering because as a team we think Due to the 23 features that can be used in our dataset, feature engineering will be essential to this endeavour. Therefore, the team must determine the pertinent features that can be useful in training a binary classification model. By using feature engineering techniques, we can add new features and remove superfluous ones, helping to increase algorithm accuracy.

These tools assist in choosing one or more data sets for data trend analysis, which subsequently assists the team in developing insights that can be useful for any financial organization to improve efficiency and develop strategies for risk management.

**Project Timeline**

****