PROJECT LLD

Project Title	Credit Card Default Prediction
Technologies	Machine Learning Technology
Domain	Banking
Project Difficulties level	Intermediate
Submitted By	Atyab Hakeem

1. INTRODUCTION

Banks provide loans and credit cards to their customers, allowing them to make purchases and pay later. However, an increasing number of credit card users are defaulting on their payments, which poses problems for banks in terms of profitability and trust from investors and stakeholders. One solution to this problem is to identify potential credit card defaulters ahead of time and implement measures to mitigate the risk of default.

This can be achieved by using machine learning algorithms to identify potential defaulters before they default. By analyzing the financial history and behavior of credit card users, banks can develop predictive models that can identify customers who are at high risk of defaulting on their payments. Once potential defaulters are identified, banks can take steps to mitigate the risk of default, such as by requiring these customers to provide additional collateral or by imposing stricter limits on their credit card usage. By taking these measures, banks can protect their profitability and maintain the trust of their investors and stakeholders.

2. PROBLEM STATEMENT

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. The goal is to predict the probability of credit default based on credit card owner's characteristics and payment history.

3. DATASET COLUMNS DESCRIPTION

Column 1 - ID: ID of each client

Column 2 - LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary = credit)

Column 3 - SEX: Gender (1=male, 2=female)

Column 4 - EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

Column 5 - MARRIAGE: Marital status (1=married, 2=single, 3=others)

Column 6 - AGE: Age in years

Column 7 - PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

Column 8 - PAY_2: Repayment status in August, 2005 (scale same as above)
Column 9 - PAY_3: Repayment status in July, 2005 (scale same as above)
Column 10 - PAY_4: Repayment status in June, 2005 (scale same as above)
Column 11 - PAY_5: Repayment status in May, 2005 (scale same as above)
Column 12 - PAY_6: Repayment status in April, 2005 (scale same as above)

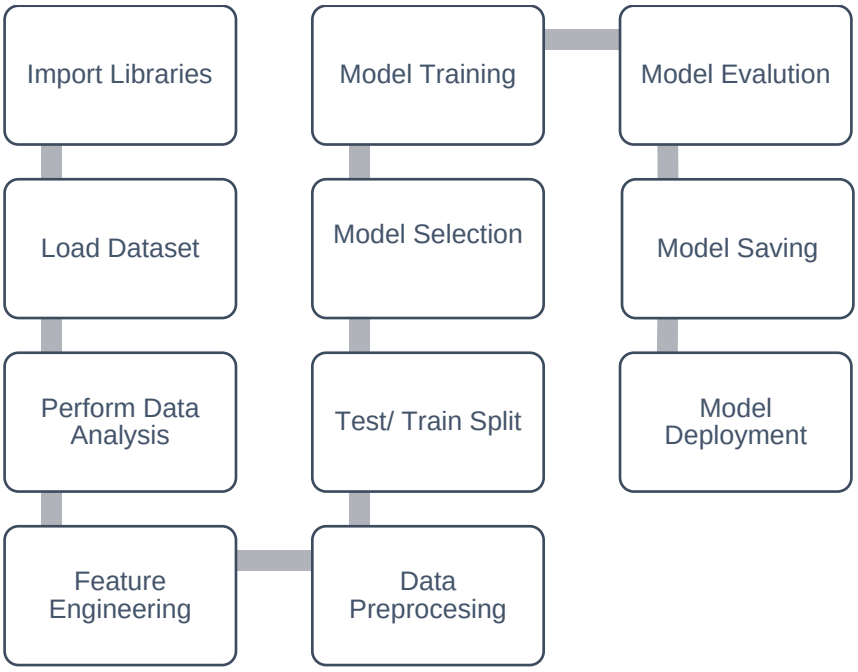
Column 13 - BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
Column 14 - BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
Column 15 - BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
Column 16 - BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
Column 17 - BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
Column 18 - BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

Column 19 - PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
Column 20 - PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
Column 21 - PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
Column 22 - PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
Column 23 - PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
Column 24 - PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

Column 25 - default.payment.next.month: Default payment (1=yes, 0=no)

4. ARCHITECTURE

4.1. Methodology:



4.2. Data Description:

The dataset was taken from Kaggle (URL: https://www.kaggle.com/uciml/default- of-credit-card-clients-dataset), This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

4.3. Data Analysis

In Data Analysis, the various trends in the dataset exhibited by the numerous features were visualized and examined. In addition, possible reasons were also hypothesized. For instance in the ‘SEX’ column, it was found that that there are more women than men in the dataset and men have a slightly higher chance of defaulting on payments. Also, in the ‘Education’ column there are few people on the 'unknown' categories (0, 5, 6) and, although their probabilities of default are not exactly close, all of them are lower than the probabilities found for the 'well defined' labels (1, 2 and 3). Furthermore, predominant level of education in the dataset is 'University', followed by 'Grad School', 'High School', 'Unknown' and 'Others'. Considering only the first three levels a higher education translates to a lower chance of default.

Pandas, Matplotlib, and Seaborn libraries were predominantly used for visualization of the data.

4.3. Data Pre-processing.

The inspection of the data revealed the characteristics of the numerous features in addition to the tendencies and distribution of the data. As such, it was found that there were no null values in any of the columns. Moreover, the feature columns were also not multi-collinear i.e. strong correlation didn’t exist between the any two features. However, due the ranging values of the feature columns, standard scaling was performed to bring the values to a single scale.

4.4. Feature Engineering

The ID column was dropped to the lack of significance and contribution to the default tendencies. Due to the lack of multi-collinearity and significant contribution by the rest of the columns, the rest of the columns were included as features and corresponding preprocessing was performed on the data.

4.5. Train/Test Split

This library was imported from Sklearn to divide the final dataset into the ratio of 67-33%, where 67% of the data was used to train the model and the latter 33% was used for validation.

4.6 Selecting Model

Naives Byes, XGBoost, and RandomForest algorithms were used to train the model.

4.7. Evaluation

Naïve Byes had the accuracy of 77.98%, XGBoost 82.14%, and RandomForest with an accuracy of 81.11%.

4.8 Save Model

As XGBoost model gave the highest accuracy it was saved using the pickle library which in a binary mode.

4.9 Deploy in Local Host

A simple frontend was created using HTML and a webapp was created and hosted using the Flask library.