**Credit Card Default Prediction using Machine Learning Techniques Proposal**

Kevin Harmer, Chaithra K C

Due: 10/22/2022

**Objective**

Credit makes up a very significant portion of the banking industry. While interest payments from credit cards could likely fund the sector alone, credit remains a riskier investment due to the abundant amount of customers who fail to pay off their debt. Consequently, understanding who will default and when is crucial. Predicting this phenomenon not only controls the financial damage caused by payment failures but can also be reimplemented into the system for new customer approval and their associated credit limits. This report will address possible avenues of exploration and predictions to be made considering credit card defaults, while also providing possible insight into what factors indicate default payments.

**Data Set Description**

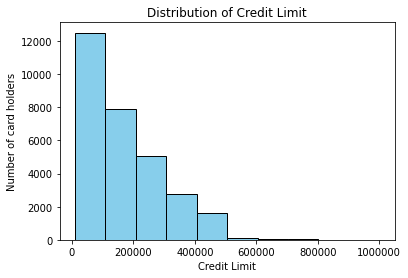
To address the problem around defaulted credit card payments, the case of customersâ€™ in Taiwan will be examined. This data set contains several clients with individual descriptive factors (sex, age, etc.) and information about their bills and payments. There are 40000 rows (each representing a client) and 25 columns. We break the columns into 4 groups, which include all the data except for ID:

* Customer Demographics: LIMIT\_BAL (credit limit), SEX (gender; 1 = male, 2 = female), EDUCATION (education level; 1 = Graduate School, 2 = College, 3 = High School, 4 = Other), MARRIAGE (Martial Status; 1 = married, 2 = single, 3 = other), and AGE (age of client in years)
* Historical Past Payments: PAY\_0-PAY\_6. PAY\_0 = 1 month previously, PAY\_2 = 2 months ago, and so forth (there is no PAY\_1 variable). Data measured in these columns are -1 = pay correctly, 1 = payment delay for 1 month, 2 = delay for 2 months, and a consistent pattern for higher values.
* Billing Statements over the past 6 months: BILL\_AMT1-BILL\_AMT6
* Payment Amounts over the past 6 months: PAY\_AMT1-PAY\_AMT6
* Boolean Default Variable: default payment next month

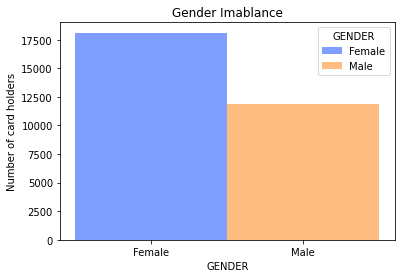
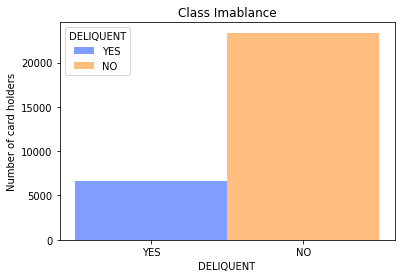
The data can be found at the link: <https://archive.ics.uci.edu/ml/machine-learning-databases/00350/>

**Preliminary Data Exploration**

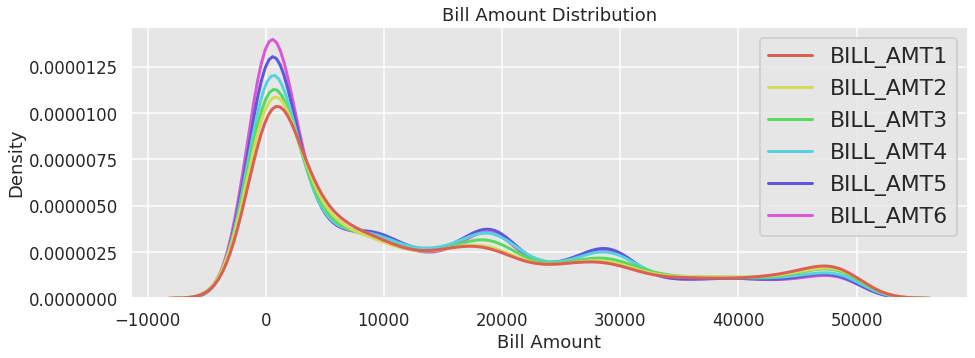
* The average age of credit card users is 35.5 years, and the median age is 34 years. 50% of users are between the ages of 28 and 41. It appears quite normal. However, the large difference between the 75th percentile age of 41 years and the maximum age of 79 suggests potential outliers.
* Some variables do not have documentation to explain their factors. Specifically, the pay codes have a -2 value and a 0 value, while EDUCATION (= 4, 5, and 6) and MARRIAGE (= 3) variables do not address the “other” inputs. For pay code, -2 will be interpreted as paying the bill in full so there is no following debt and 0 will be paying an amount that is satisfactory but not enough to decrease the debt due to interest. As for the other two, variable transformation will be used to get dummy variables for the known factors and the other inputs will be discarded.
* The average credit limit offered to the customers is about 167,500 NT$ and the median credit limit is 140,000 NT$. This suggests that the credit limit is not symmetrically distributed because the mean credit limit is not very close to the median credit limit. In fact, the distribution will most likely be right-skewed because the mean is greater than the median (seen below).



* There are 6,224 more women with credit cards than men in the dataset. Similarly, 23,364 customers did not default on their payments, compared to 6,636 customers who did (both shown by the visualizations below). Data balancing will need to be done before modeling.

* The fact that some of the bill amounts are negative suggests overpayment.



**Proposed Data Exploration Insights**

* Determine which characteristics have the highest influence on default loan prediction
* Compare each month’s payment and bill statement to past default differences
* Analyze different age groups and how they influence loan defaults
* Identify any difference between education
* Review relationships between sex, education, marriage, and age with limit\_balance to check for any discriminatory measures

**Proposed Predictions**

* We intend to develop machine learning models like Logistic Regression, Random Forest, Support Vector Machine, XGBoost, and Artificial Neural Network that take into account the six months (April 2005 - September 2005) credit card bill amount, payment amount, repayment status, and demographics to predict delinquency in credit card payments.
* Perform a time series analysis over a six-month period of payments, to determine the trend in bill amounts, payment amounts, and repayment status.
* Identify the variable(s) that most strongly affect the default payment.

**Model Inference Insights**

* Identify which demographic is most likely to default on a payment, then advertise payment plans and other promotions to the individuals. This can be done by analyzing the standardized parameters of the factored data (potentially including age group factors)
* Using the standardized parameters of data from the previous 6 months, results could identify which month is most likely to indicate a present default. This means that if someone defaults on a loan, payment plans and/or payment relief are the exact amount of months in advance.
* Understanding overall credit limit utilization and spending habits of customers. This could help the re-evaluation of customer credit limits
* Recognizing the importance of regular and irregular credit card payments in sensing delinquency. We can suggest customers who are falling behind in making required monthly payments to set up autopay, discontinue using the credit card, etc., to avoid defaulting the payment.

**Non-Spark Packages**

Matplotlib, Seaborn, Pandas, Numpy