Capstone Project 1: Project Proposal

Predict Credit card Default
Lakpa Sherpa

- **What is the problem you want to solve?**

  A Bank's main service is to provide loans, and credit cards to people who are in need of money to either invest on a business, purchase assets, pay school tuitions or other reasons. It is a great source of income for a bank until the loan or credit is set as a default. Then, the bank has to spend time and money to take legal actions against the debtor. The money may or may not be collected depending on the circumstances. However, Banks can take precautionary steps to pin out applicants who are at risk to default.

- **Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?**

  This project is useful for bank's risk management officers to find out high risk level credit card applicants and to recommend new loans and credit card to low risk loan/credit applicants who are a potential client for the bank. This project will predict if a client will default a credit card loan, so the bank can take precautionary measures to minimize the possibility. The bank can find out the variables which affects the client's behaviors or possibility to default. With the model we create from this project and the necessary client's data and informations feed into our model, we can predict if the client will or will not default. This will help the bank to make decision to issue credit card.

- **What data are you using? How will you acquire the data?**

  I am going to use a data set of Taiwanese credit card holders which was donated to UCI Machine Learning Repository for study and research purposes. The data set consists of 30,000 rows of records with 25 columns. You can download the excel file to acquire the data from the website.

  The Data Set includes following data from the client which will help us to create different machine learning models:-

➢ Client ID: Unique number used as identity of the clients.
➢ X1: Amount of the given credit (NT dollar): Limit_Balance
➢ it includes both the individual consumer credit and his/her family (supplementary) credit.
➢ X2: Gender (1 = male; 2 = female).
➢ X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
➢ X4: Marital status (1 = married; 2 = single; 3 = others).

- X5: Age (year).
- X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
- X6 = the repayment status in September, 2005
- X7 = the repayment status in August, 2005
- X8 = the repayment status in July, 2005
- X9 = the repayment status in June, 2005
- X10 = the repayment status in May, 2005
- X11 = the repayment status in April, 2005.
- The measurement scale for the repayment status is:-
- -1 = pay duly
- 1 = payment delay for one month
- 2 = payment delay for two months
- 3 = payment delay for three months
- 4 = payment delay for four months
- 5 = payment delay for five months
- 6 = payment delay for six months
- 7 = payment delay for seven months
- 8 = payment delay for eight months
- 9 = payment delay for nine months and above.
- X12-X17: Amount of bill statement (NT dollar).
- X12 = amount of bill statement in September, 2005
- X13 = amount of bill statement in August, 2005
- X14 = amount of bill statement in July, 2005
- X15 = amount of bill statement in June, 2005
- X16 = amount of bill statement in May, 2005
- X17 = amount of bill statement in April, 2005
- X18-X23: Amount of previous payment (NT dollar).
- X18 = amount paid in September, 2005
- X19 = amount paid in August, 2005
- X20 = amount paid in July, 2005
- X21 = amount paid in June, 2005
- X22 = amount paid in May, 2005
- X23 = amount paid in April, 2005
- Y(default Payment) 1 for yes and 0 for no

- **Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

❖ **Data Wrangling:**

- Import the file to the database as a pandas DataFrame
- Clean the data using different manipulating techniques to tidy and rearrange the data
- Change column names.

➢ Make separate columns for different gender, education level, and marital status.

❖ **Data Visualization:**

➢ Use Seaborn to visualize the data with different kinds of plot
➢ Plot age vs repayment status to see any pattern
➢ Plot marital status vs repayment status for different months
➢ Plot gender vs repayment status for different months
➢ Plot education vs repayment status for different months

❖ **Model:**

➢ Logistic Regression
➢ Decision Tree
➢ Random Forest
➢ Extra Trees
➢ Gradient Boosting
➢ Support Vector Machine
➢ Neural Networks(Optional)

- **What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

  The deliverables for this project would be a complete presentation of the project with all the documentations, visualizations and python codes on notebook. All the model created to predict default on credit card loan. The final submission will include the introduction, the steps and methods involved in this project from the beginning to the end of the project. Finally, present how we evaluated the accuracy of different models created, and recommend a model.