Project Proposal

Jeuty Swargiary

March 2, 2018

## R Markdown Project proposal

This is an R Markdown document for Capstone Project Proposal.

## 1. What is the problem you want to solve?

Credit card defaults is a cost to card-issuing banks. I would like to find out the correlations between the demographic and financial variables contributing to the probability of a default payment.

## 2. 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 otherwise?

Creating a predictive model to predict a default payment will help card-issuing banks assess credit card debt risks and take appropriate risk management actions. Based on the predictive model outcome, the bank can take preventive actions and know when to withhold/cancel the credit card.

## 3. What data are you going to use for this? How will you acquire this data?

The dataset for Default of credit card clients is atken from <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>.

The dataset to be used is available in <https://archive.ics.uci.edu/ml/machine-learning-databases/00350/>. This dataset contains customer payment data from a bank and contains observations of card holders with default payments. It has demographic factors, and financial factors as defined below. The dataset contains ~30K observations and 23 variables and the last column contains information on default payments(1=yes, 0=no).

|  |  |
| --- | --- |
| **Demographic variables** | **Financial variables** |
| Gender | Limit Balance |
| Education | Payment History (April-September 2005) |
| Marital Status | Bill Statement (April-September 2005) |
| Age | Repayment status(April-September 2005) |

## 4. In brief, outline your approach to solving this problem (knowing that this might change later).

For Data wrangling, I would like to address the below points:

1. There are some categorical values which are not listed in the attribute information e.g. Education 0 (not listed), 5 (unknown), 6 (unknown) Marital Status 0 (not listed) I may decide to safely put them under “Other” category or exclude them since they are few in number.
2. Bill Statements: Some amounts of the monthly bill statements are negative. It needs to be investigated if the negative values can be interpreted as credit.
3. The Repayment status has -2 and 0 which is not listed in the payment delay description. It needs to be investigated how to interpret these values. According to the documentation, the measurement scale for the repayment status is: -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.
4. Renaming column name “default payment next month” to “default”

After Data wrangling, the next step would be to determine:

1. Correlation between each demographic variable (gender, education, marital status, age) contributing to the probability of a default payment
2. Correlation between each financial variable (limit balance, repayment status, bill statement and previous payment) contributing to the probability of a default payment

Finally the dataset will be used to create a predictive model to predict a default payment.

## 5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The deliverables will include the code used for data wrangling, exploratory data analysis and will include the key findings. It will also include the accuracy of the built model in predciting a default payment.