Capstone Project 1 - Milestone Report

Overview

1. Define the problem

How to predict the probability of default of credit card bill payments based on demographic data, past payment data, credit limit and bill amount.

1. Identify your client

Client for this project can be any bank or financial institution who has a credit card of lending business.

This problem is faced by banks and financial institutions for their credit card portfolios. This problem exists even for other credit portfolios with monthly payments.

Banks can take proactive steps to prevent the default. Prevention is much less expensive than recovery and collection after the default.

1. Describe your data set, and how you cleaned/wrangled it

I would like to use ‘default of credit card client’s data set’ from UCI data reporsitory.

It has 30000 records.

The dataset did not have any nulls. But, data in fields like marital status, education were not quite understable. I used pandas map function to give meaning names to the values of these fields. I also created a column called Age-Range.

Violin plots gave some idea about number of missed payment for different genders.

I tried to answer the following questions through data wrangling:

# Which group has the highest average credit limit?

# Which group has the lowest average credit limit?

# Which group is comprised of highest percentage of people who have a balance-to-limit rating less than or equal to 30%?

# Which group has the lowest utilization or balance-to-limit rating?

# Which group has the highest amount of debt, is the most likely to default, and is the most likely to miss a payment?

# Which group has the lowest amount of debt, is the least predicted to default, and is not likely to miss a payment?

1. List other potential data sets you could use

I had also considered credit card fraud detection dataset from Kaggle for my capstone project 1.

1. Explain your initial findings

The main goal of my investigation is to figure out, which variables among the 23 have higher impact in predicting probability of default next month. I took a series of steps and explored data to figure out the impacts:

1. Do age of the credit card holder have anything to do with default:
   1. The 60-69 age group was found to have higher percentage of defaulters that other age groups.
   2. The next two vulnerable groups are 70-79 and 50-59
2. Next I went on to check the last bill payment of the credit card holder. Does this amount say anything?
   1. The proportion of default is interestingly higher when the last payment amount is low.
   2. The default percentage keeps on going down as the last payment amount increases.
   3. The default percentage is higher if the last payment amount is below 10,000
3. Education impacts default:
   1. Those with just high school education tends to default more.
   2. The next two groups are University and Graduate school.
   3. The less the education, more the default percentage.
4. Does marital status have an impact:
   1. The marital status of ‘others’ have some interesting behavior of highest rate of default
   2. ‘Married’ people tend to default more than ‘Single’ folks.
5. What about gender?
   1. Clearly, the males tend to default more than the females
6. I was currious to know whether credit has any relationship with default
   1. The clear finding is: card holders with lower credit credit limits tend to default more
7. Is a large number of missed payments a precursor to default?
   1. It is evident that if there are two or more missed payment, the chance of default increases significantly
   2. But, there is a strange observation. I have observed that there are defaults when there has been zero missed payments.
8. I also tried to find out correlations between all variables
   1. As expected, the missed payments and credit balance tend to indicate to some extent, that a default is coming.

Findings from inferential statistics :

1. There is significant difference between male and female default rates . Male default rate is higher
2. With p value being zero we concluded that credit limit and default are independent of each other. But. there is little bit of negative correlation between the two. In other words, higher the credit limit, lower is the chance of default.
3. There are few more findings. Those have been included in the code.
4. Share the Capstone Project 1 code and milestone report related to Github repository

Data Wrangling

<https://github.com/dipanjan123/Dipanjan-Capstone-project1/blob/master/DataWrangling-Capstone.ipynb>

Data Visualization

<https://github.com/dipanjan123/Capsone_Data_Visualization/blob/master/Data%20Visualization%20Capstone%201.ipynb>

Inferential statistics

<https://github.com/dipanjan123/Dipanjan-Capstone-project1/blob/master/Inferential%20statistics%20Capstone1.ipynb>