1. Task Description

Hello,

Now that you have created a dedicated local programming environment, you're ready to begin your work by preparing and exploring the data. Before we dive in let's review some notes about the project:

Problem:

1. Increase in customer default rates - This is bad for Credit One since we approve the customers for loans in the first place.
2. Revenue and customer loss for clients and, eventually, loss of clients for Credit One

Investigative Questions:

1. How do you ensure that customers can/will pay their loans? Can we do this?

As you progress through the tasks at hand begin thinking about how to solve this problem. Here are some lessons we learned form a similar problem we addressed last year:

1. We cannot control customer spending habits
2. We cannot always go from what we find in our analysis to the underlying "why"
3. We must on the problem(s) we can solve: What attributes in the data can we deem to be statistically significant to the problem at hand?
4. What concrete information can we derive from the data we have?
5. What proven methods can we use to uncover more information and why?

I'll be expecting a report on your experience in a few days.

Thanks,

GR

Guido Rossum
Senior Data Scientist
Credit One
www.creditonellc.com

1. Task Solution

The solution for this task is divided in two documents, both documents will be loaded in the classroom.

- Module 5 - Task 2A – EDA Credit One.pdf
- Module 5 - Task 2B - M5T2_EDA_Credit_One.zip

2. Analysis

The EDA of the original dataset. We have learned some potential business values from this analysis.

1 - There are the total 30,000 observations in our credit card dataset. There are 6636 observations (22.12%) are defaulted credit card holders. There are 23364 observations (77.88%) are not defaulted credit card holders.

2 - In the defaulted credit card holder group, 43.29% are male, while 56.71% are female. In the not_ defaulted group, 38.585% are male, while 61.415% are female. It showed proportion of male in the defaulted group is higher than not defaulted group.

3 - The mean of age in the defaulted credit card holder group is 35.48 and 75% of this group is under 41. And mean of age in the not defaulted credit card holder group is 35.41 and 75% of this group is under 41. There are no significant differences in age distributions between these two groups.

4 - The mean of Amount of the given credit in the defaulted credit card holder group is 130,109.66, and 75% of this group is under 200,000. The mean of Amount of the given credit in the not defaulted credit card holder group is 178099.73, and 75% of this group is under 250,000.The results show that Amount of the given credit in the defaulted group is lower than not defaulted group.

5 - In the defaulted credit card holder group, 11.10% of this group is graduate level, 6.79% is university level and 4.12% are high school. For a total of 22.12% from the total of credits. In the not defaulted group, 28.50% of this group is graduate, 35.67% is university level and 12.27% are high school. The not defaulted group has a larger proportion of higher education level.

6 - In the defaulted credit card holder group, 11.14% of customers are singles, 10.69% of customers are married, and 0.28% of customers are divorce. In the not defaulted group, 42.08% of customers are singles, 34.84% of customers are married, and 0.8% of customers are divorce. The not defaulted group has a slight larger proportion of married customers.

3. Conclusions

Exploratory data analysis is a very effective tool to get deep insight into our data through visualization methods. We can get a lot of useful information about our dataset before building up our model. Through EDA, we can explore our data from different angles, visualize our results and get new ideas to analyze our model. Through EDA, we can easily communicate and present our results to stakeholders without technical background.

This analysis indicates gender, amount of the given credit, education and marriage status have some effect on default rate. Age is not a significant factor to affect default rate. Before building up our model, we gain some insight into factors that affect default rate, which can help us to solve high default rate problem.