Predicting Loan Default Risk

# Any surprises from your domain from these data?

Dataset that I got seem to have the features that I’m looking for. I analyzed the datasets (demographics, performance and previous loans data) briefly. I checked if the joining key among these datasets is unique or not. It is as expected, and I don’t see any surprises so far.

# The dataset is what you thought it was?

Yes, I read the codebook and type of features in the dataset. I also did initial analysis on the dataset. I used the datasets from Zindi data challenge site, I have the information I’m looking for in the dataset. In my preliminary analysis, I feel like I would have to transform few fields before building machine learning models.

# Have you had to adjust your approach or research questions?

Yes, There are no changes to the research questions at this point.

# Is your method working?

Its working as expected so far. I loaded the data in Python using pandas library. I built a dataframe with the observations and started looking at the descriptive statistics to get a feel of the data. I joined the different datasets to get a wholistic view of the customer data. I also drew the correlation matrix among the features. I would have to convert some non-numeric data to numeric for few features. I will use Feature selection method to select only the most relevant and impacting features rather than all. I am also working on data visualizations to view the patterns in the data.

In order to predict load default risk, I would first need to understand all the features that could influence the loan default risk rate. I would have to create a derived dataframe with demographic, performance and previous loans data. I will then model on this derived dataframe with loan default risk as my target variable. I will first start with decision tree modeling, followed by random forest.

I have not started with modeling yet, but I will start with feature scaling (brining the features to uniform range using StandardScaler) and decision tree model. Will talk more about modeling after I’m done coding on it.

# What challenges are you having?

There are few challenges with the dataset I got. First challenge being the kind of values in the some of the fields. For example, there are some non-numeric values in few fields which I want to transform prior to modeling. I first need to get a derived dataframe before begin modeling.

There are no major challenges with the dataset or exploration data analysis so far. I also don’t see any major road blocks. It is a real-time problem and I think the dataset has enough features to cover all aspects of customer behavior.