Milestone 2: Draft of White Paper

Credit Card Approval Prediction

DSC 680

Saurabh Shrestha

- **Business Problem**

The decision of approving a credit card or loan is majorly dependent on the personal and financial background of the applicant. Factors like, age, gender, income, employment status, credit history and other attributes all carry weight in the approval decision. Credit analysis involves the measure to investigate the probability of a third-party to pay back the loan to the bank on time and predict its default characteristic. Analysis focus on recognizing, assessing, and reducing the financial or other risks that could lead to loss involved in the transaction.

There are two basic risks: one is a business loss that results from not approving the good candidate, and the other is the financial loss that results from by approving the candidate who is at bad risk. It is very important to manage credit risk and handle challenges efficiently for credit decision as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision.

- **Background/History**

Credit card issuing institutions are becoming meticulous in approving credit cards for customers. In addition, the downturn of financial institutions during the US subprime mortgage and the European sovereign crisis has raised concerns about risk management properly. Hence, these challenges have attracted significant attention from researchers and practitioners. A wide range of statistical and machine learning techniques have been developed to solve credit card related problems. It is found that machine learning

techniques are superior to other traditional statistical techniques in dealing with credit scoring.

The decision of approving a credit card or loan is majorly dependent on the personal and financial background of the applicant. There are two basic risks: one is a business loss that results from not approving the good candidate, and the other is the financial loss that results from approving the candidate who is at bad risk. It is very important to manage credit risk and handle challenges efficiently for credit decisions as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision.

- **Data Explanation**

The dataset contains application record for credit card with the shape the dataset being 438557 rows and 18 columns. The variables provided are gender, have children, education type (higher education/secondary special), Income, days employed, housing type, family status to name a few.

Dataset can be found at: https://www.kaggle.com/rikdifos/credit-card-approval-prediction

Below is the explanation of the datasets:

application_record.csv		
Feature name	Explanation	Remarks
`ID`	Client number	
`CODE_GENDER`	Gender	

FLAG_OWN_CAR	Is there a car	
FLAG_OWN_REALTY	Is there a property	
CNT_CHILDREN	Number of children	
AMT_INCOME_TOTAL	Annual income	
NAME_INCOME_TYPE	Income category	
NAME_EDUCATION_TYPE	Education level	
NAME_FAMILY_STATUS	Marital status	
NAME_HOUSING_TYPE	Way of living	
DAYS_BIRTH	Birthday	Count backwards from current day (0), -1 means yesterday
DAYS_EMPLOYED	Start date of employment	Count backwards from current day(0). If positive, it means the person currently unemployed.
FLAG_MOBIL	Is there a mobile phone	
FLAG_WORK_PHONE	Is there a work phone	
FLAG_PHONE	Is there a phone	
FLAG_EMAIL	Is there an email	
OCCUPATION_TYPE	Occupation	
CNT_FAM_MEMBERS	Family size	

- **Method**

With the support of the data, the plan is to analyze and apply necessary data preparation techniques which could be useful for creating an effective model. I followed the process as follows:

1. Load the dataset.
2. Converted column names for readability.
3. Dropped unnecessary columns.
4. Handled missing and null values.
5. Converted non-numeric to numeric value. Such as Education column had Lower secondary, Secondary / secondary special, Incomplete higher, Higher education and Academic degree as unique value which were converted to numerical value 0, 1, 2, 3, 4, 5 respectively.
6. Exploratory Data analysis – I created multiple graphs and analyzed the dataset. Such as for marital status, I had higher applicants who were married vs lowest applicants from widowed. For housing type; highest applicants came from housing/apartment vs lowest with co-op apartment housing type, for gender: I had higher female applicant vs Male
7. Train and test the machine learning model – I used sklearn's model selection for splitting dataset into train and test set.

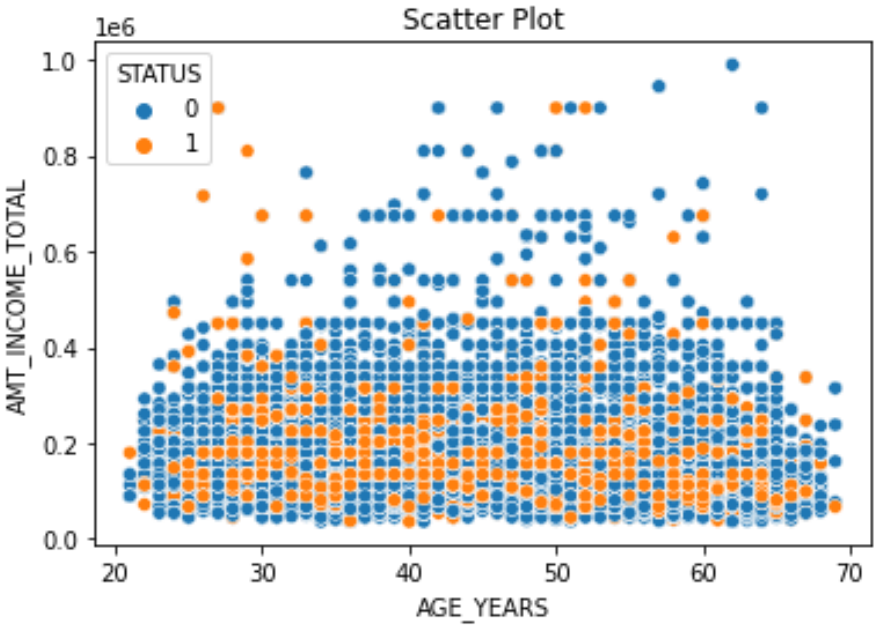
I trained a supervised learning model where status is our target variable. Before going through the modeling process, I dropped columns that were irrelevant for our models such as numerical columns (has: cell phone, phone, work phone, months balance) and categorical columns such as Income category, Marital status, Housing type, Occupation. These columns do not have any correlation and are deemed unnecessary.

- **Analysis (with Illustrations)**

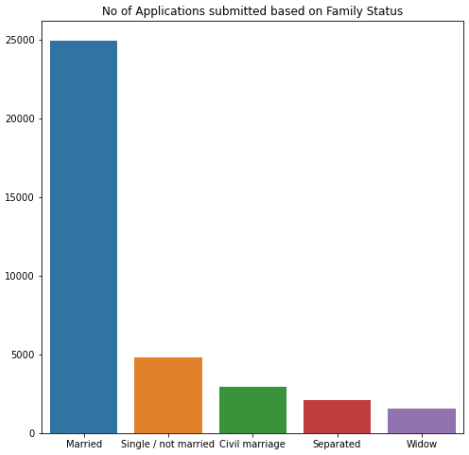
EDA Analysis: Here are examples of few graphs I created for EDA. The Status shows the binary values of either 1 or 0. 0 indicates that the applicant has paid their credit due

on time or has no loan remaining. Whereas 1 indicates that they are behind on their payments.

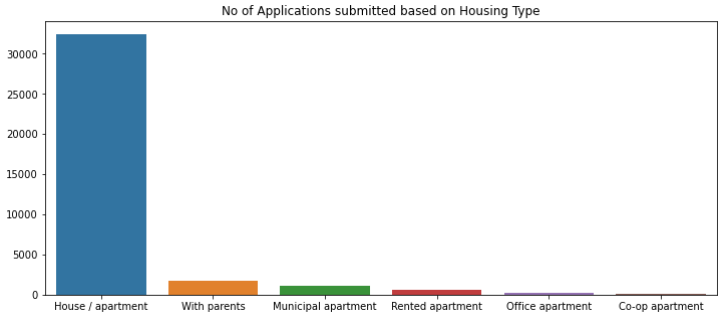
The above graph shows that the applicants are not good candidates if Total income & years of Employment is less.



The above graph shows that, majority of applicants who have higher income are more likely to pay their due on time. There is no correlation with age with their payments. I also analyzed the applicant's distribution data, here are some results that I found:

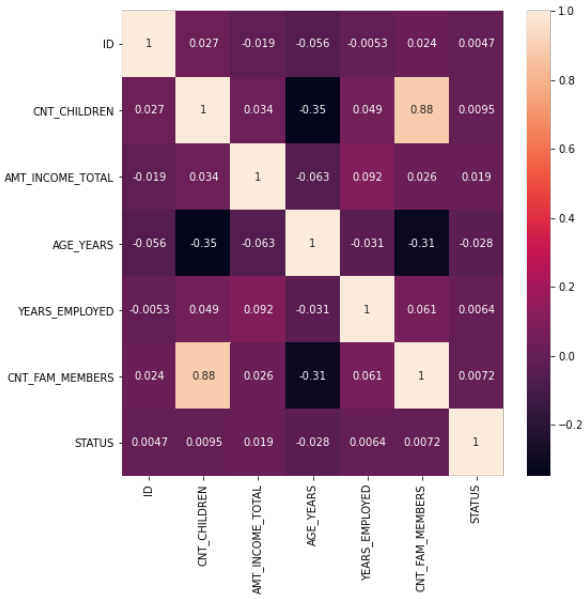


Majority of applicant's are married



Majority of applicant's lives in House / Apartment

Correlation: Below we have the seaborn correlation heatmap which shows that the features are not highly correlated to each other. In addition to that, the features are also evenly split between positive and negative correlation between two variables. This graph also shows that there is no column (Feature) which is highly co-related with 'Status'



- **Conclusion**

1. As the dataset is highly imbalanced, I have used SMOTE (Synthetic Minority Oversampling Technique) technique to understand which model performs better.
2. I took 2 passes at the Machine learning models, one with initial data and other with balanced data after performing SMOTE technique and the two results differed significantly.

After applying all the Machine learning models on the balanced dataset, I got that XGBoost Model is giving the highest accuracy of 84.14 %. SMOTE Sampling methods provided much better results compared to raw data.

- **Challenges**

The challenges I faced when working with this project was working with the dataset as a whole. In the dataset I found out that there were many features which were unnecessary for the model creation, to pick which variables were at first a challenge. Additionally, there were many columns which were not numerical in nature. I had to use label encoder to convert all non-numerical columns to numerical such as gender column had male and female categorization, owns cars were Y and N, education types were high school, doctorate, secondary/special education and so on. Another challenge was when modeling observation in one class was higher that the observation in another class. This created an imbalance. Majority class had over 35,000 observations whereas minority class had 768 observations. This resulted in higher accuracy in majority class and low or 0 in minority class. Hence, failed to capture the minority class.

- **Assumptions and Limitations**

I didn't know that there we maximum null values in one particular feature or that I had to work with so many categorical features. Another hypothesis I had was that the applicants who applied had higher age were going to be paid in linear progression. I was wrong. At one point in the age there comes a decline in salary, i.e. as the age of the applicant increased the salary also increase until certain age. Other assumption I had was that dataset will have balanced classes which I was wrong on.

For this project the limitation is that it is generated from a special location. This creates restriction on homogeneity of users. Reason being in a specific region we won't have all types of population. Example being from the analysis we found that maximum of the application is married and live in a house, comprising of above 70%.

- **Recommendations**

I would suggest this project would be helpful primarily to banks or credit union who provides a credit card directly to the end user. Additionally, this project could be useful to marketing companies as well. With the help of analysis, banks can determine what type of customers are likely to default on their credit card payment. This way they can avoid those types of customers and reject their application. Marketing companies can use the analysis where they can understand the credit card applicants structure of information such as age with salary, what type of application have salaries, which education type are more likely to earn higher salary and so on.

- **Implementation Plan**

I will be following the Predictive analytics processing steps with CRISP-DM which is explained in chapter 2 of *Applied Predictive Analytics Principles and Techniques* book. There are six stages as CRISP-DM namely:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

- **Ethical Assessment**

The reason for ethical assessment is to figure out what to do next. Ethics means the consideration of behaviors that will do good and assessment is what you do to figure out what you or someone else knows about doing a particular thing—and then figuring out what the next step is. For this project, I researched when working with the variables, followed the rules and regulations of ethics such as how could my system negatively affect impact individuals? who is the most vulnerable and why? how much error in prediction can the business accept for this use case? Explanation on which input factors had the greatest influence on the outputs?