Credit Risk Analysis Report

1 Problem Statement

The objective of this project is to classify customers of a German Bank as good or bad credit risk. The dataset used in this project consists of information of 1000 customers. Each customer in this dataset is flagged as good or bad credit risk. This dataset was originally prepared by Prof. Hofmann from University of Hamburg.

2 Data Wrangling

This step consists of inspection of collected dataset and to identify any potential problems (e.g. missing attributes, outliers, duplicate entries etc.). In this dataset, their are total 10 attributes available for every customer. These attributes are Age, Sex, number of jobs, Housing, Saving accounts, Checking account, Credit amount, Duration, Purpose, Risk. Few of the customers only had checking or saving accounts but not both. Absence of a account is indicated by NaN in the dataset. To prevent any further processing issues I replace NaN values by "No Account". After this we move to exploratory data analysis.

3 Exploratory Data Analysis

In this step, I explore correlations (Figure 1) between different features of the dataset. In figure 1, diagonals represent histogram of the attributes and non-diagonal figures represent correlation plots between each pair of attributes. We observe positive correlation between credit amount and duration attributes with correlation coefficient value being 0.62. Other attributes do not display any correlation. Therefore, we cannot drop any of the features. For the categorical features (e.g. Sex, Purpose etc.) we also plot barplot for counts of each unique value. We don't observe much any similarity between barplots of any two categorical features. Therefore, we cannot drop any of these features. Now in the next step we prepare our dataset for modeling.

4 Data Preprocessing

In the preprocessing step, I first convert all the categorical features except "Purpose" into numerical features using one hot encoding. I do not apply one hot encoding to "Purpose" because this is the response variable. Next, I split the explanatory variable, consisting of all data except the response variable, and response variable into training and testing data. I use 75% of data for training purpose and 25% for testing purpose.

5 Modeling

6 Model Evaluation

7 Conclusions and Future Work

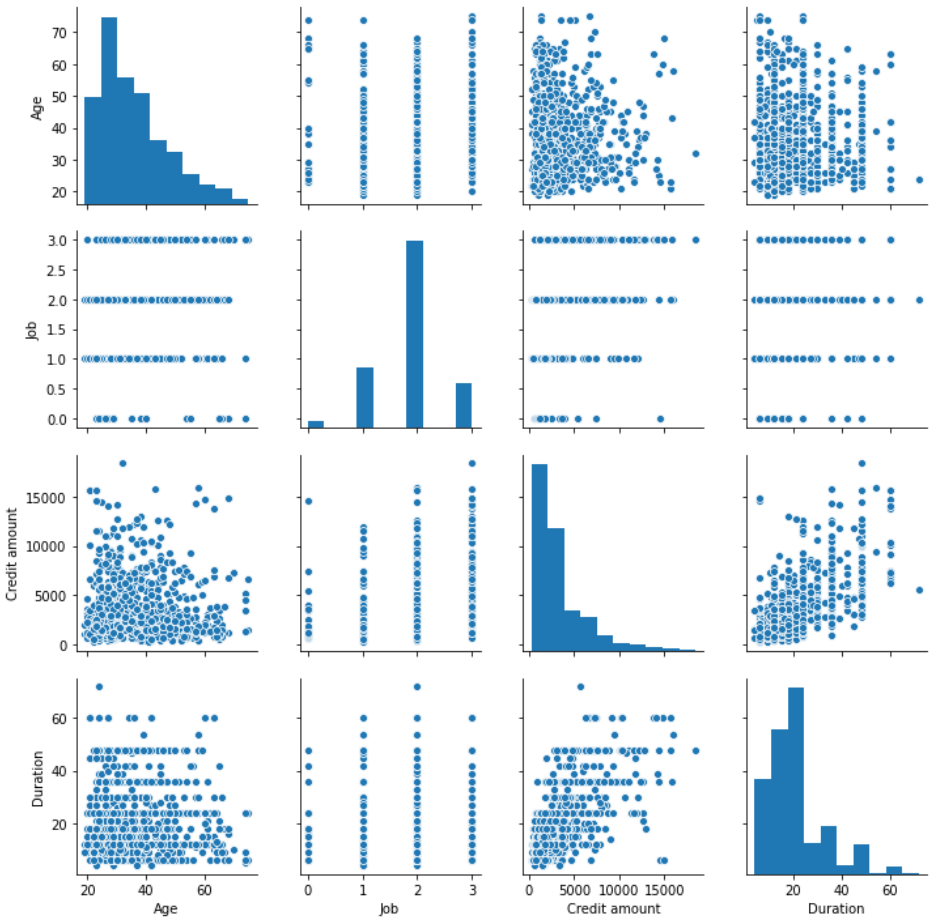

Figure 1: Correlation plot between different features of the dataset