**Introduction**

Tool used: Python

Used data: KGB.sas7bdat (known good/bad), REJECTS.sas7bdat as a reference

The assignment done by Daniil Golub, [daniil.golub2812@gmail.com](mailto:daniil.golub2812@gmail.com)

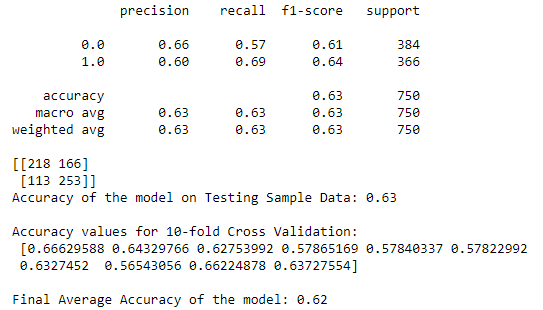
Resources used: scikit.learn documentation, stackoverflow forum, the attached book in pdf format

**General observations:**

* The number of loans (NMBLOAN) is dependent upon salary. People with income lower than 20000 tend to take loans more and vice versa.
* Number of children in a family correlates with the number of loans (more-less).
* The vast majority of people have no existing loans.
* There is no correlation between people’s income and paid loan (FINLOAN)
* Good score and bad score do not depend on number of loans while there is no strong connection between the score and income. However, good score tends to appear within people who have salary bigger than median in our observation.
* Person between 35-60 (adult) has the highest salary. Though there is not big difference between age and salary.
* People between 25-35 have the biggest number of children.
* People with the high salaries do not use any credit cards.
* Number of loans is dependent upon number of households (lower-bigger). People with 10 or more households do not take loans.
* There is no strong correlation between GB and income/number of households.
* T-Test shows that here is no correlation between student/young, student/adult in number of loans. However, there is some correlation between adult/senior in number of loans

**Logistic Regression Model and Scorecard**

In order to build logistics regression, I used independent variables – age, income, number of loans, finished loans, time at job. As a dependent variable I used column GB.



Our F-score is not high though. I may assume it is due to data which has no some strong connections between variables.

This supervised learning is a good for start since we achieved model accuracy is around 2/3.

Future improvements can be achieved if we collect more data that corresponds closer to our hypothesises and calculations.

**Raised issues**

Looking at credit modelling from SAS, it seems quite complicated to implement same techniques into Python e.g precise scorecard.

Also, as it is written above, according to my analysis, the provided data does not show a strong connection between variables which may seem complicated as well.

**Used libraries**

pyreadstat

numpy

pandas

seaborn

matplotlib

scipy.stats

sklearn.linear

sklearn.metrics

sklearn.model\_selection

**Some attached pictures from Python**

