Problem Statement

To predict whether a client, based on his/her work and education history, will be able to repay his/her loans on time

Findings from EDA

1. Most of the clients are Female.

2. Most of the clients who can pay off their home credit already have a home/apartment.

3. Most of the clients that can pay off their home credit have a job/are working.

4. People who have completed Secondary school studies/higher education are most likely to afford a home (taking into account that they have been working since they graduate from secondary school)

5. The type of jobs that allows the highest probability of repayment for home credit is/are working as businessmen.

6. Most of the clients that has income are already working.

7. Correlation matrix --> Shows us the most correlated variables such as AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.

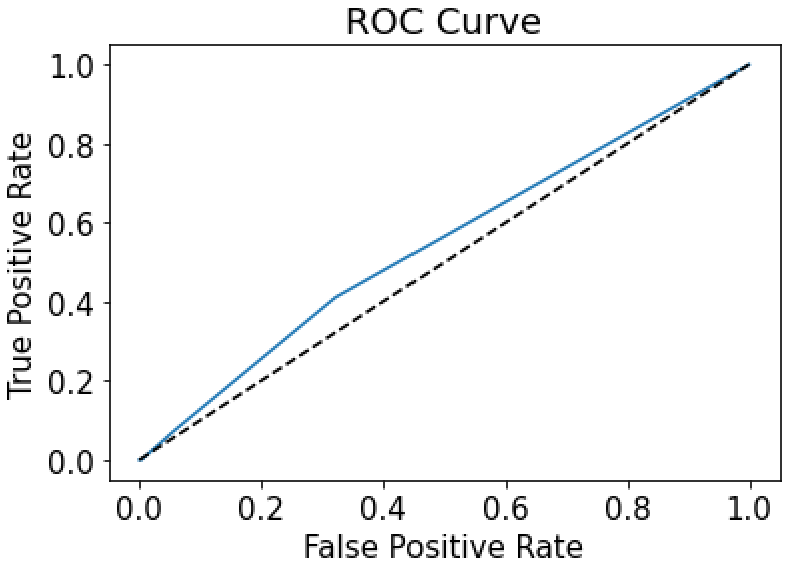
8. Income and Credit are positively correlated.

Classification Model used: K-Nearest Neighbours

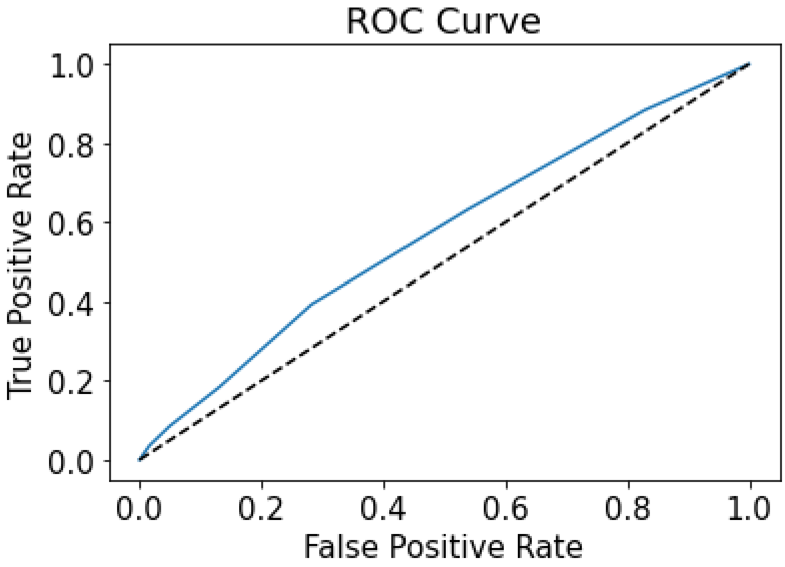
1. Calculates the distance/probability between a query and all the examples that are given in the data, selecting the specified number examples (K) closest to the query
2. Then, it will run predictions and check for validation within the model.

Model Performance:

Initial: 0.9155 accuracy with 0.4983 AUC score



After tuning: 0.9185 accuracy with 0.5 AUC score



Limitations of Dataset:

1. The data had too many unknown values and hence not able to produce better/more accurate values.
2. Dataset given is very heavy yet there is no explanation or definition of what the headers mean or at least they should describe what do the different headers do for the dataset.
3. Date of application is not given and therefore we are unable to distinguish new and old data.

Suggestions and Future Improvements:

Since the AUC score is low, a better machine learning language can be adapted to have better predictions.

Bayesian networks can be used for better prediction, detecting of any anomaly and time series prediction. It can also allow us to better interpret a joint probability distribution for this model.