Home Credit Default Risk

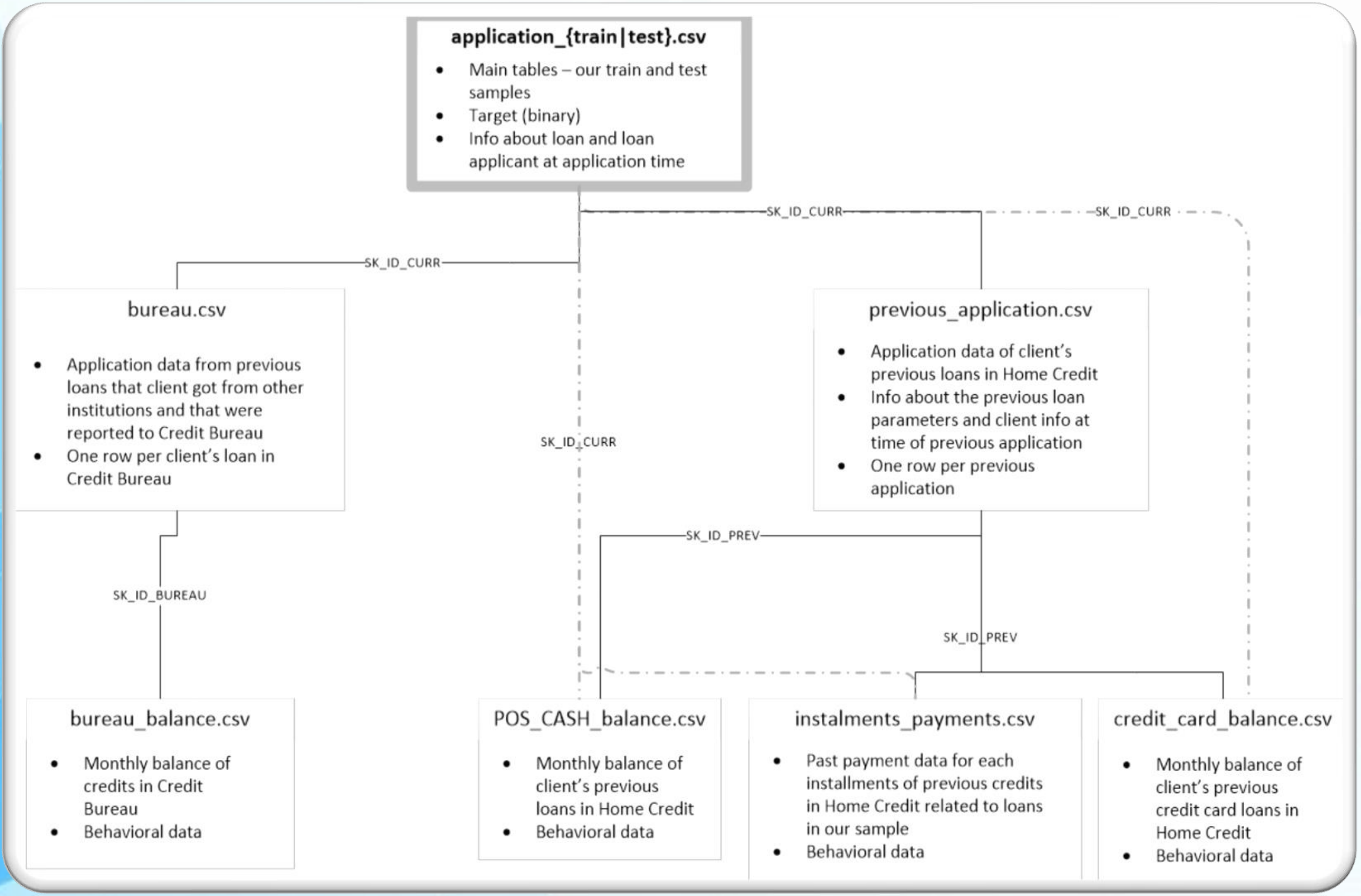
By Amir Helmy

Problem Description

- Many people struggle to get loans due to insufficient or non-existent credit histories.
- Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.
- Our aim is to make use of alternative data to predict the probability of default of a loan application.
- A typical performance measure for classification problems is the Area under the ROC Curve (AUC/ROC).

Data Description

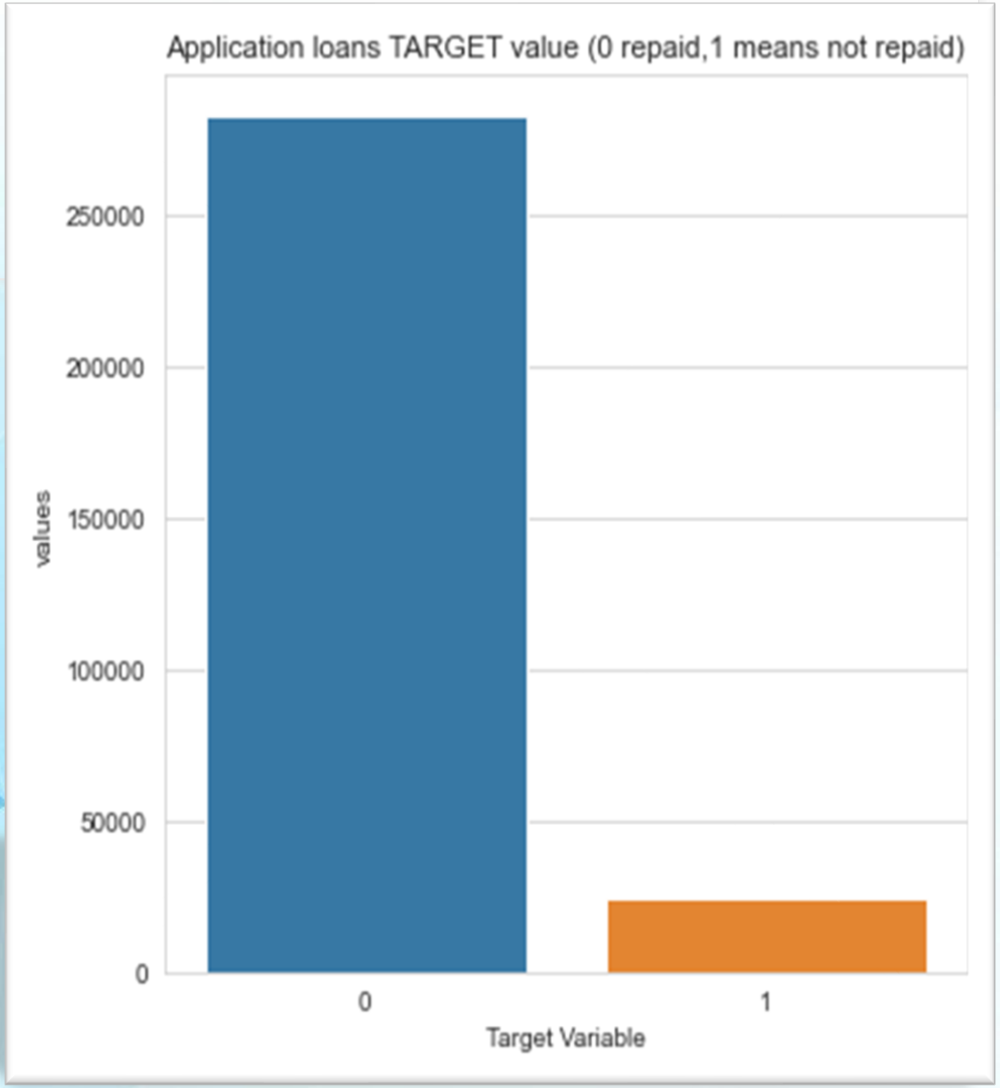
- We have different info about clients and credit applications.
- The main application data file contains 122 col 307511 rows.
- And0 total of 99 columns in 7 different historical data files related by foreign key.



Data explorations & key findings

Target feature

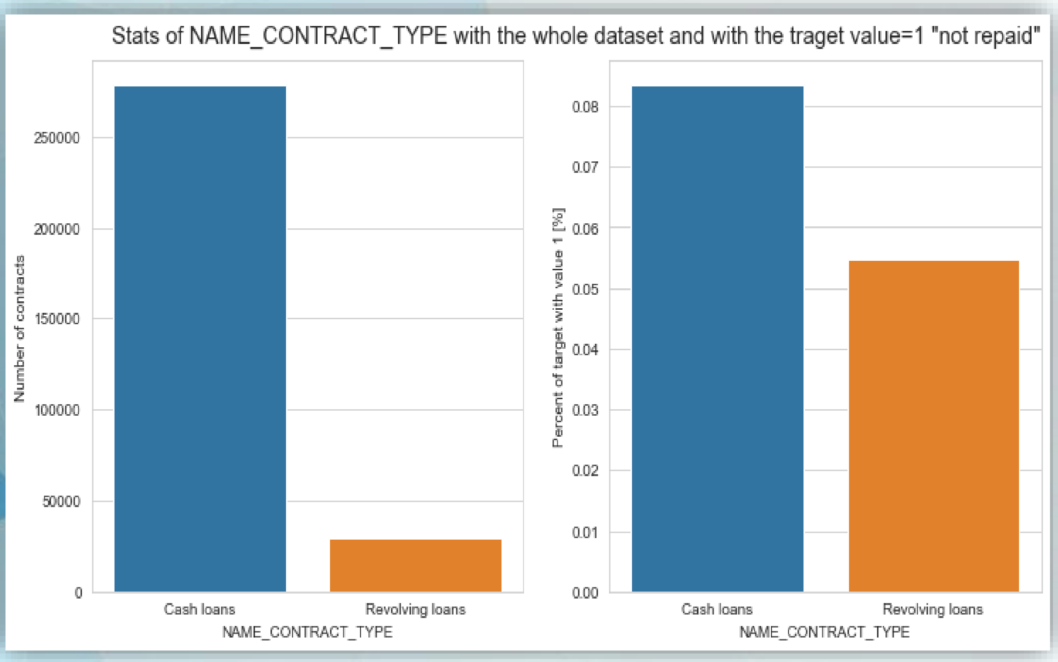
- 92% of the loans are repaid, 8% are not repaid
- Highly Imbalanced Data.
- Leading to a Biased Model.
- Under sampling to prevent majority class from dominating.



Data explorations & key findings

NAME_CONTRACT_TYPE

- Contract type Revolving loans are just a small fraction (10%) from the total number of loans.
- Larger amount of Revolving loans, comparing with their frequency, are not repaid.



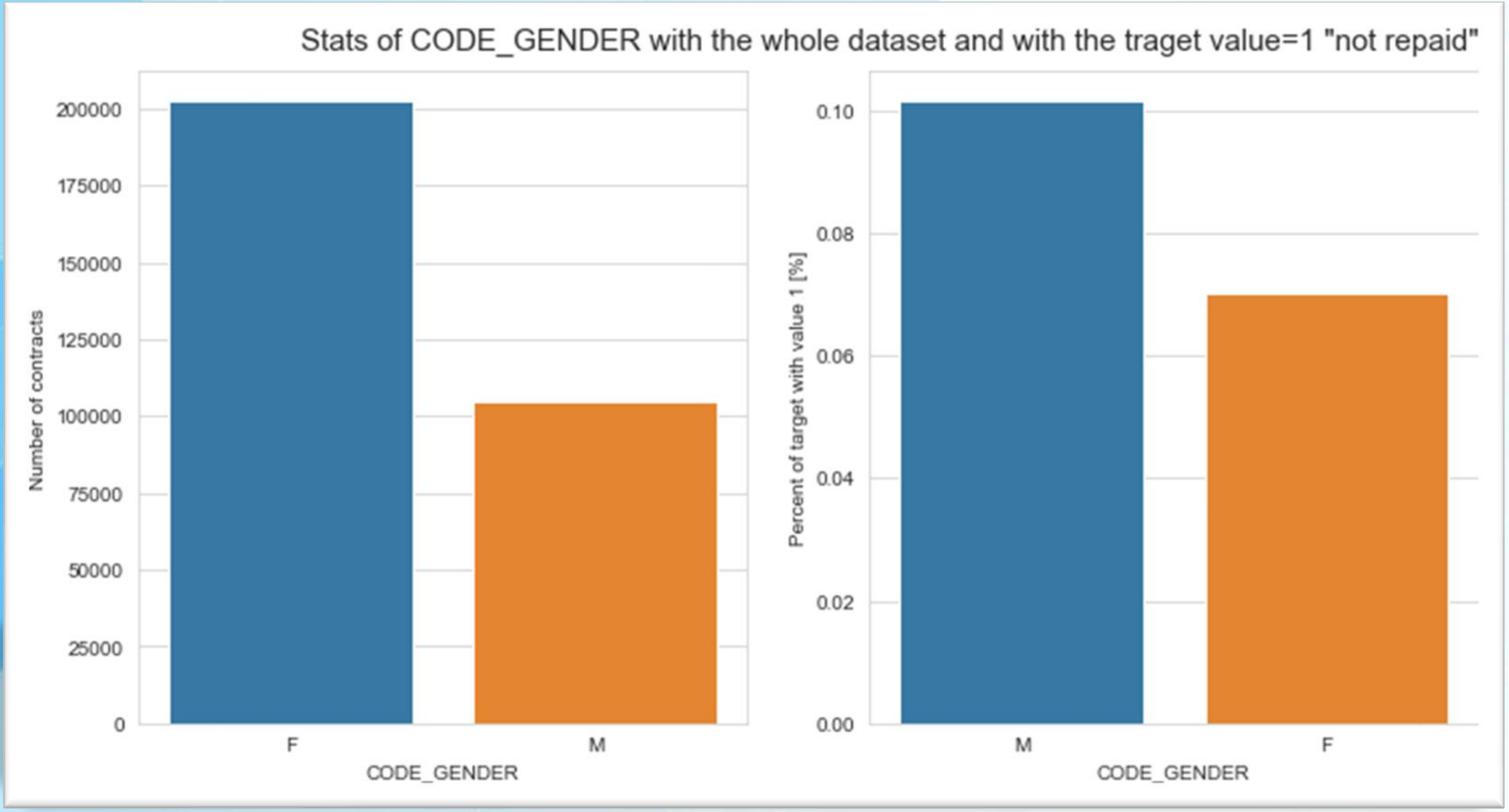
Conclusion

Beware when contract type is revolving as a larger amount of Revolving loans, comparing with their frequency, are not repaid.

Data explorations & key findings

CODE_GENDER

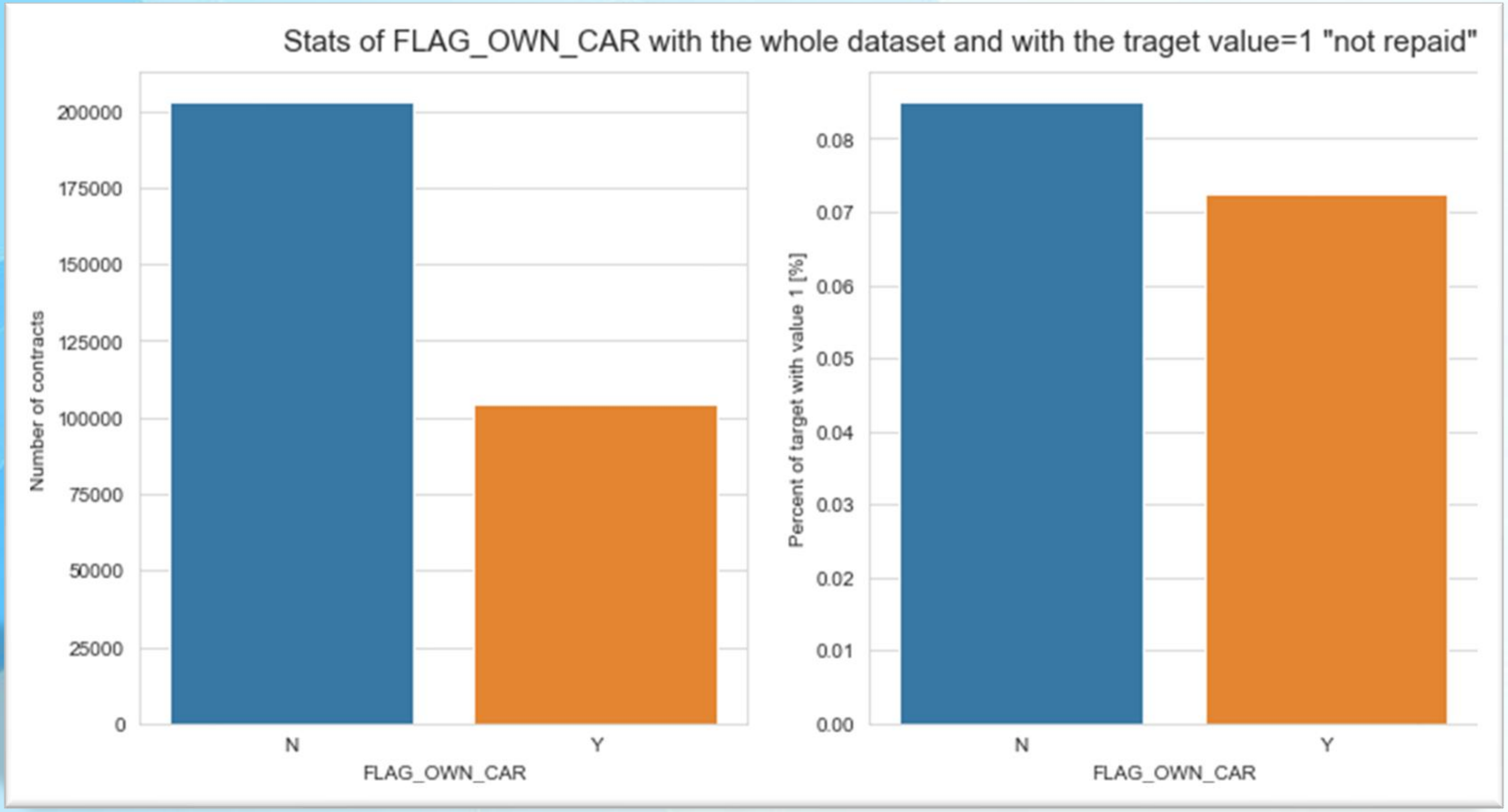
- The number of female clients is almost double the number of male clients
- Looking to the percent of unpaid credits, males have a higher chance of not returning their loans, comparing with women.



Data explorations & key findings

FLAG_OWN_CAR

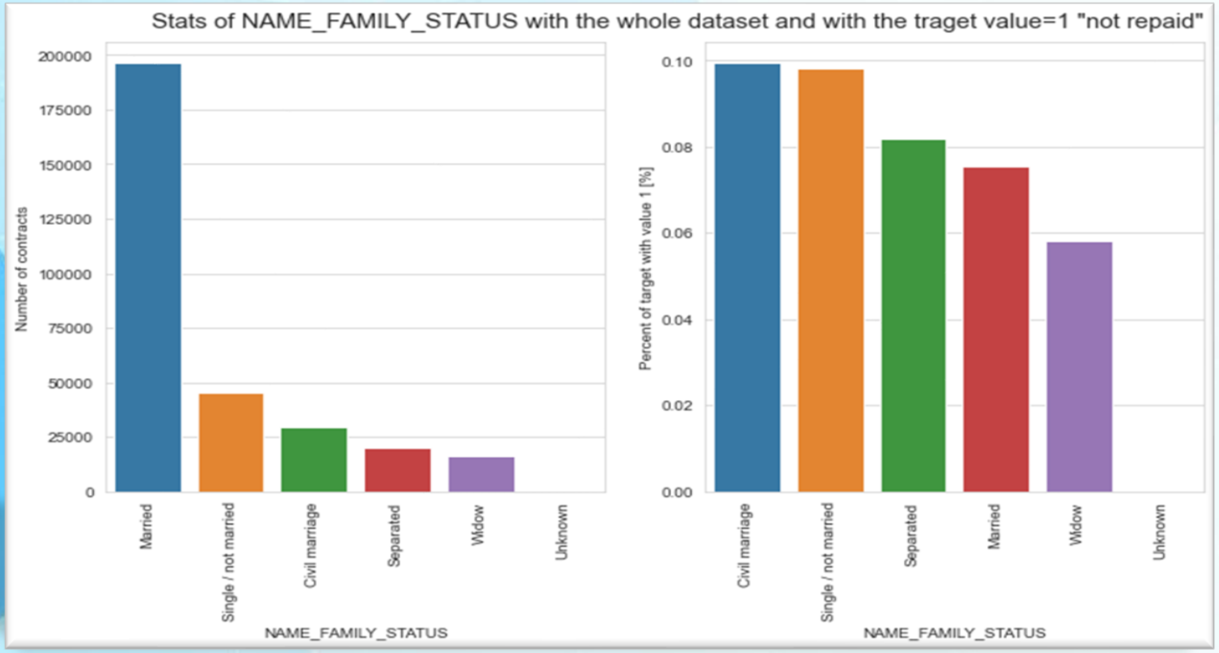
- The clients that owns a car are almost a half of the ones that doesn't own one.
- In terms of percentage of not repayment of loan ,The clients that doesn’t own a car are more likely not to repay than who does.



Data explorations & key findings

NAME_FAMILY_STATUS

- Most of clients are married, followed by Single/not married and civil marriage.
- In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment.
- Widow being the most likely to pay.



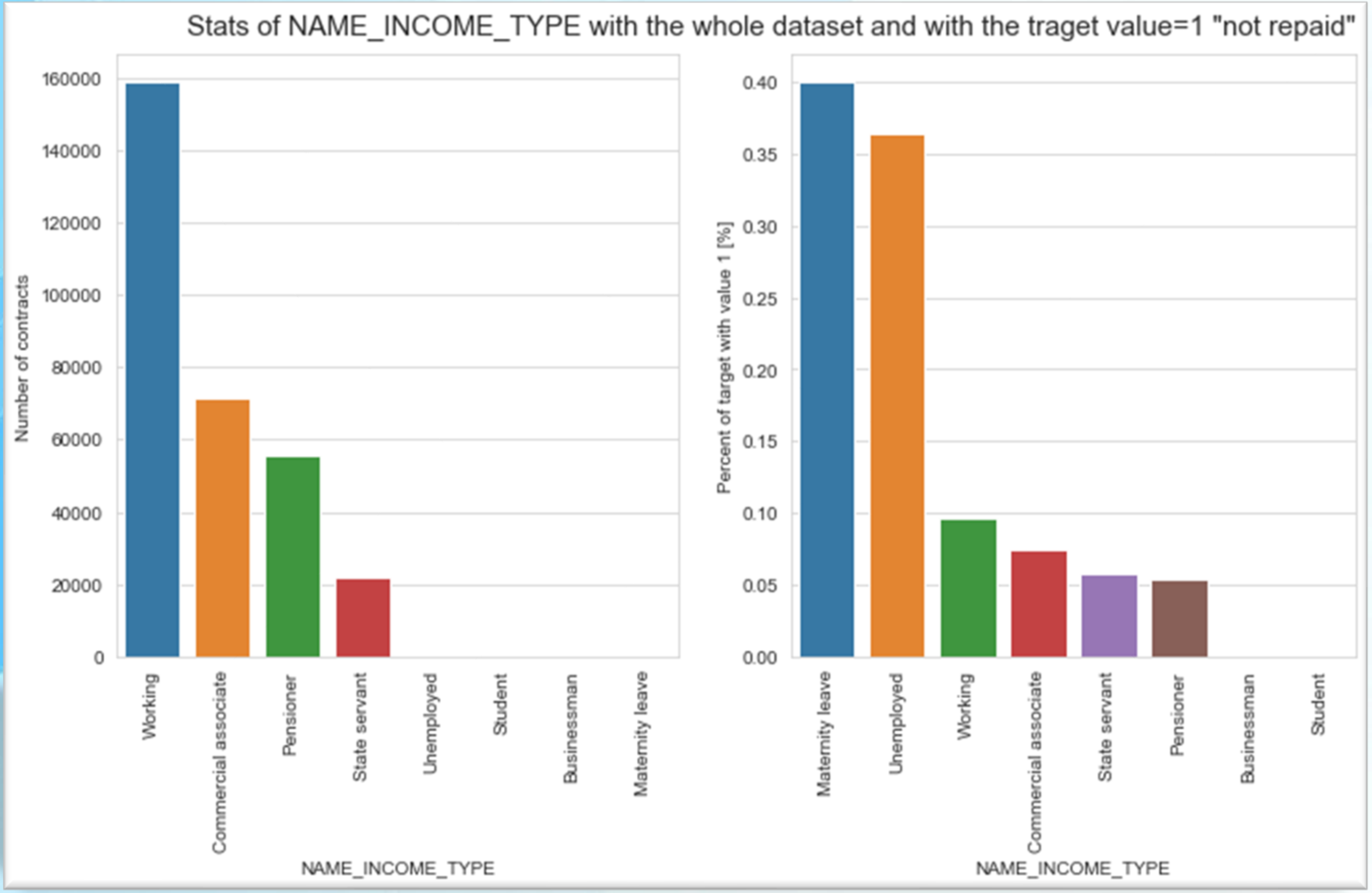
Conclusion

Widowed have the most repayment rate so they should be more assurance when lending them.

Data explorations & key findings

NAME_INCOME_TYPE

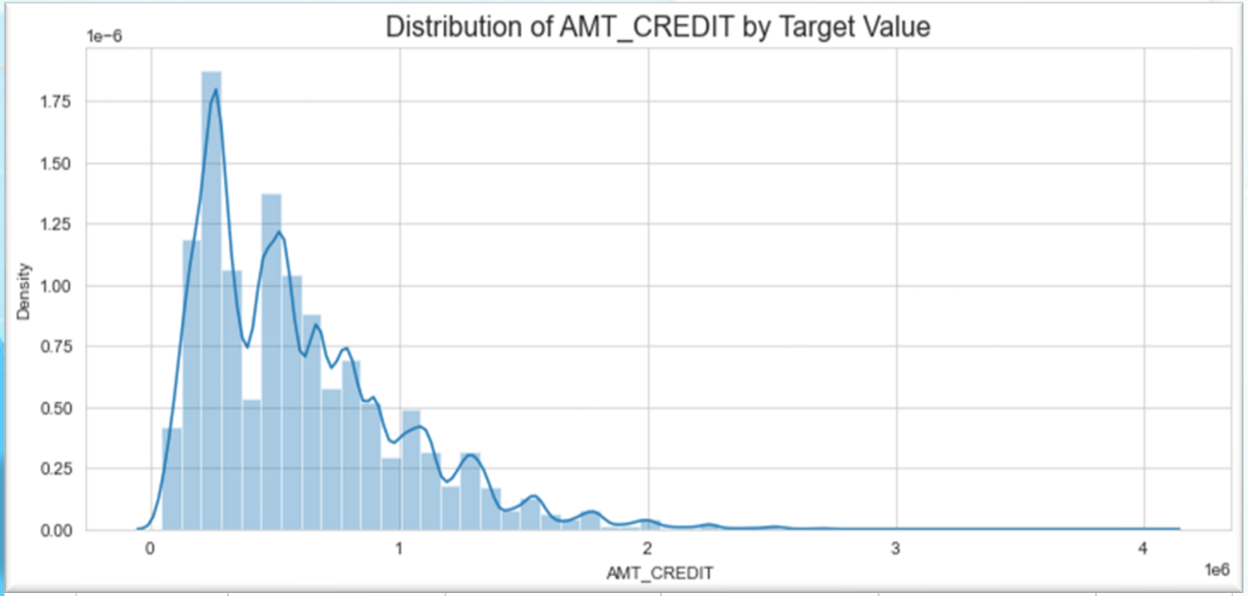
- Most income is from Working, followed by Commercial associate, Pensioner,etc..
- Applicants with Maternity leave income have almost 40% ratio of not returning loans, followed by Unemployed (37%).
- The rest of types of incomes are under 10% for not returning loans

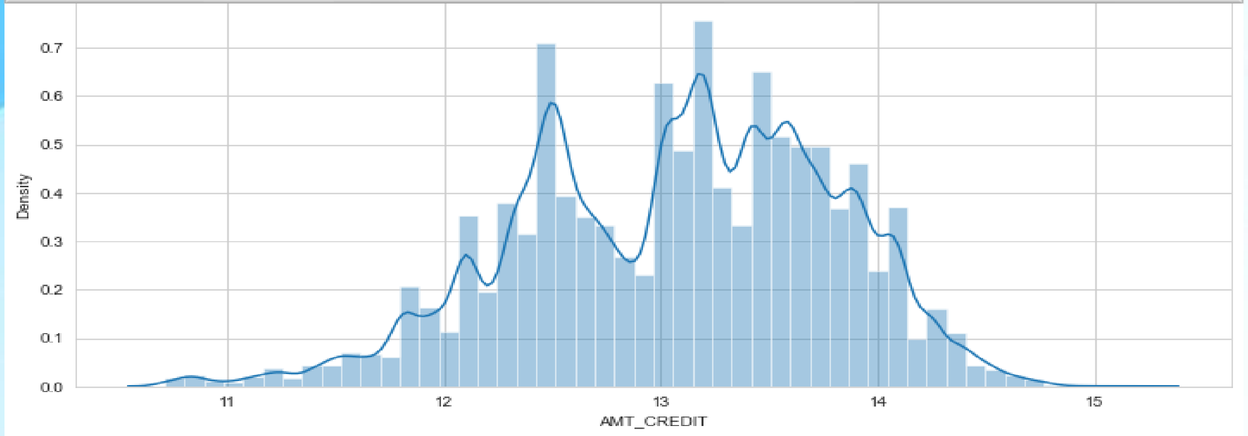


Data explorations & key findings

AMT_CREDIT

- Data points are mostly centered in the left side of the plot, **'Right skewed’** (under about 1.5 million)
- Certain models are sensitive to skewed data, so I applied **log transform** to get data close to normal distribution and correct It’s skewness.





Data Preprocessing

Missing Values

- Most of the datasets have missing values between 60% to 70%
  - Tried removing and Imputing with the median.
  - Imputing preformed better.
- Some columns needed special treatment
  - Example: Previous application Dataframe
  - Missing values means there is no previous loan to this client i.e the current application is the first.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4

Data Preprocessing

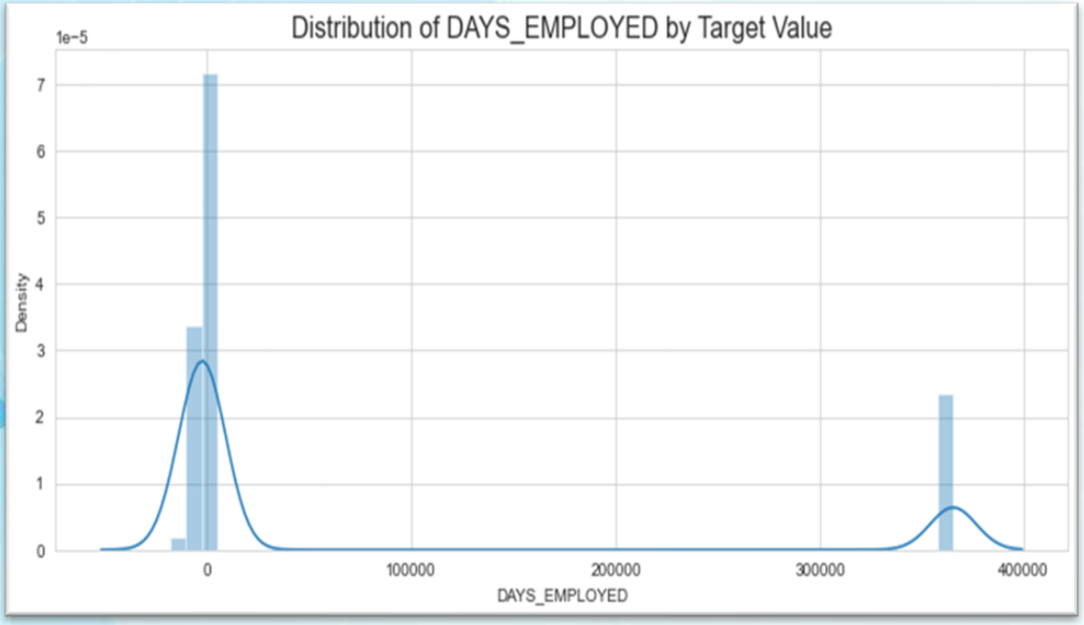
Treatment of the days of employment

Remove anomalies (Outliers)

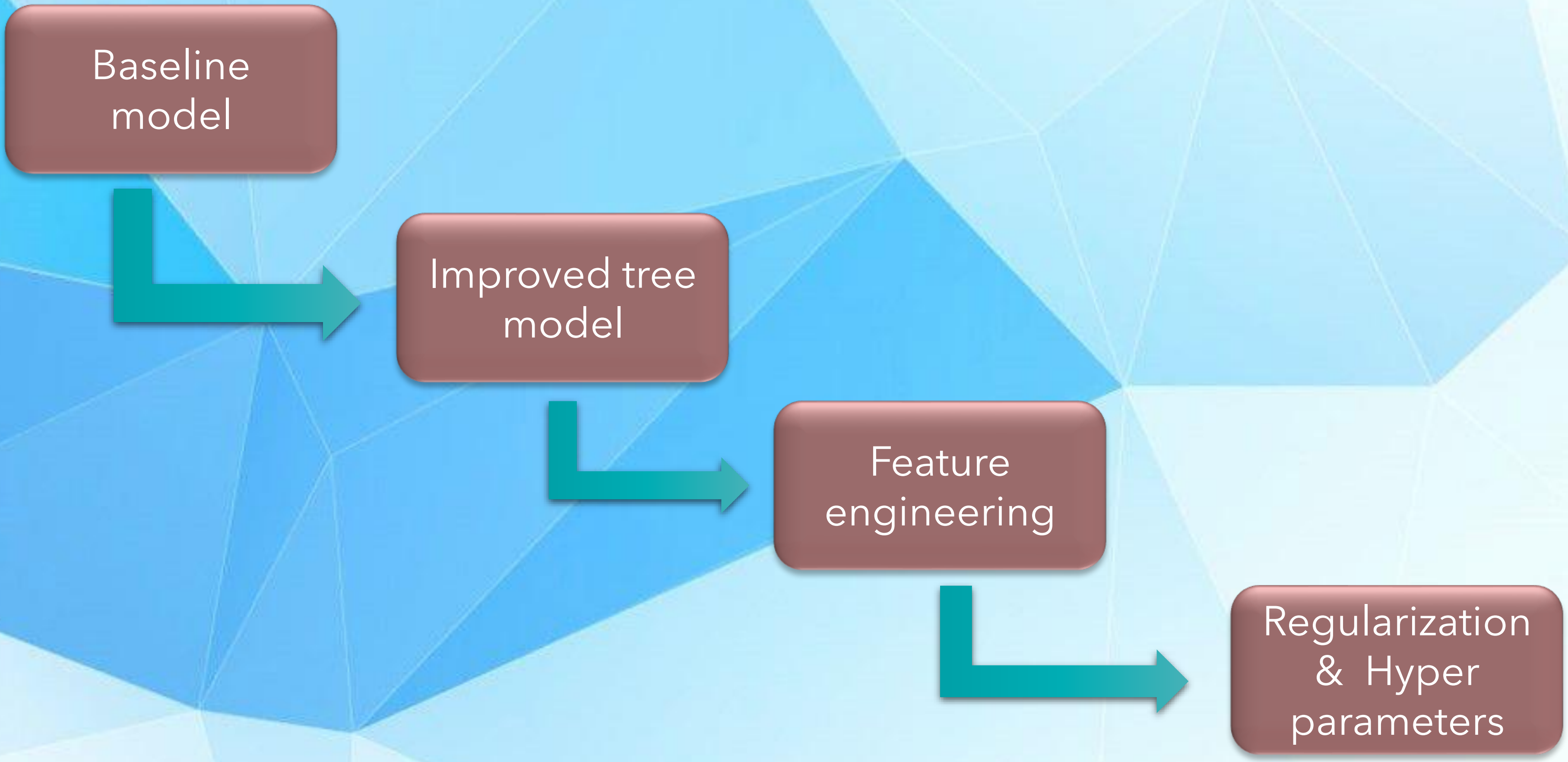
There are 55374 (5.4%) of applicants are about 1000 years old.

- Imputing.
- Added feature.

- I ended up using Random Forest.
  - Robust to outliers.



Approach in Data Modeling



Pipeline

Encode categorical variables

One Hot Encoder.

Normalize features

MinMaxScaller

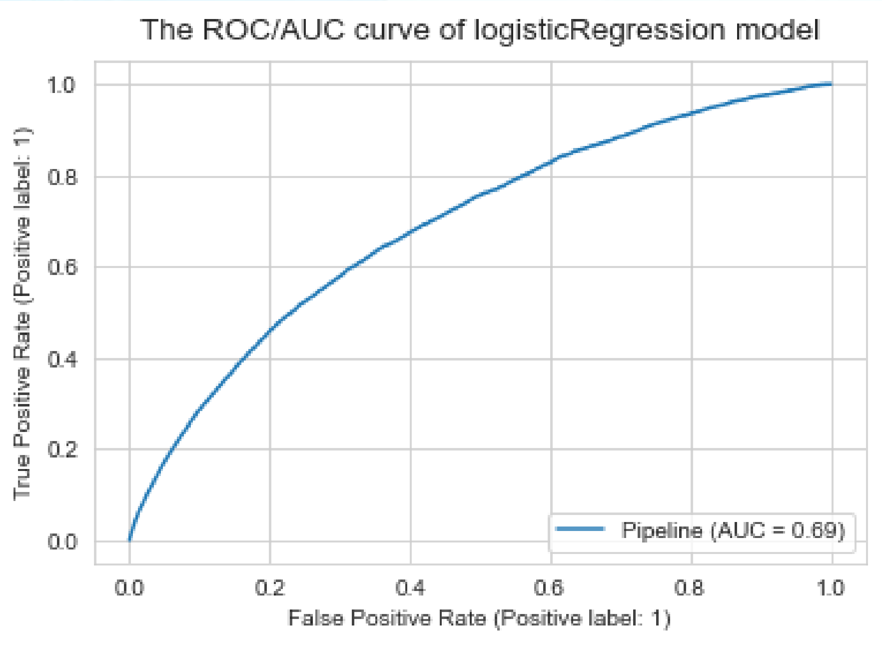
Imputing : median

SimpleImputer

Baseline Model

Logistic Regression

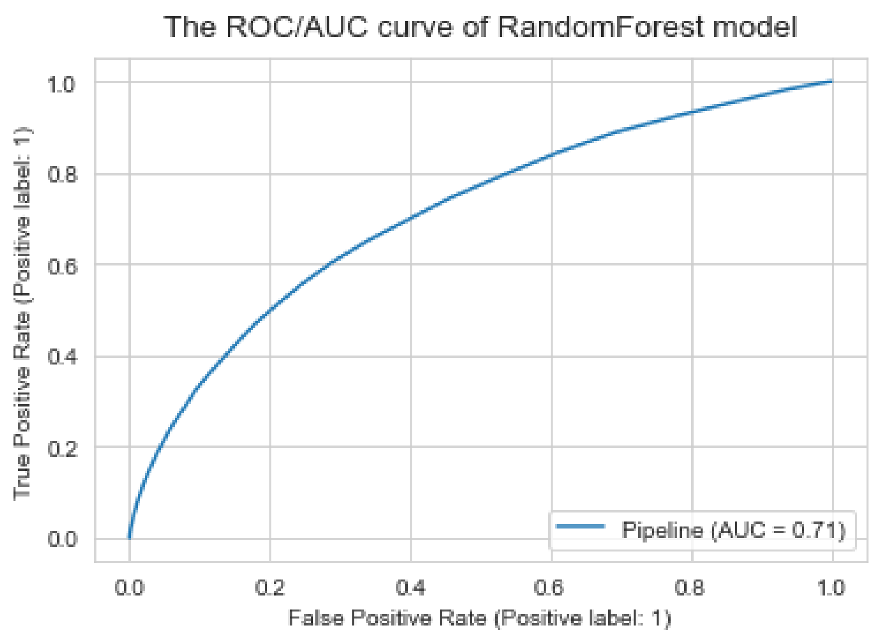
- Tried Logistic Regression without feature engineering
  - Scored around 0.691134
- Not bad



Improved Model

Random Forest

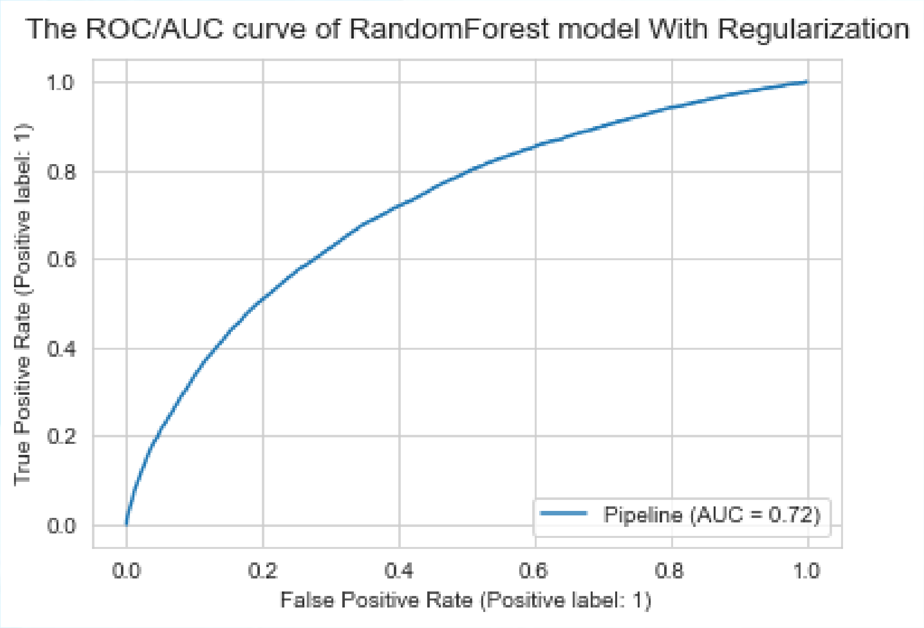
- Tried Random Forest without feature engineering
  - scored around 0.709793
- Tried Random Forest with feature engineering
  - Test data scored around: 0.7051048
  - Train data scored : 1.0
- Over Fitting problem.



Dealing with Overfitting.

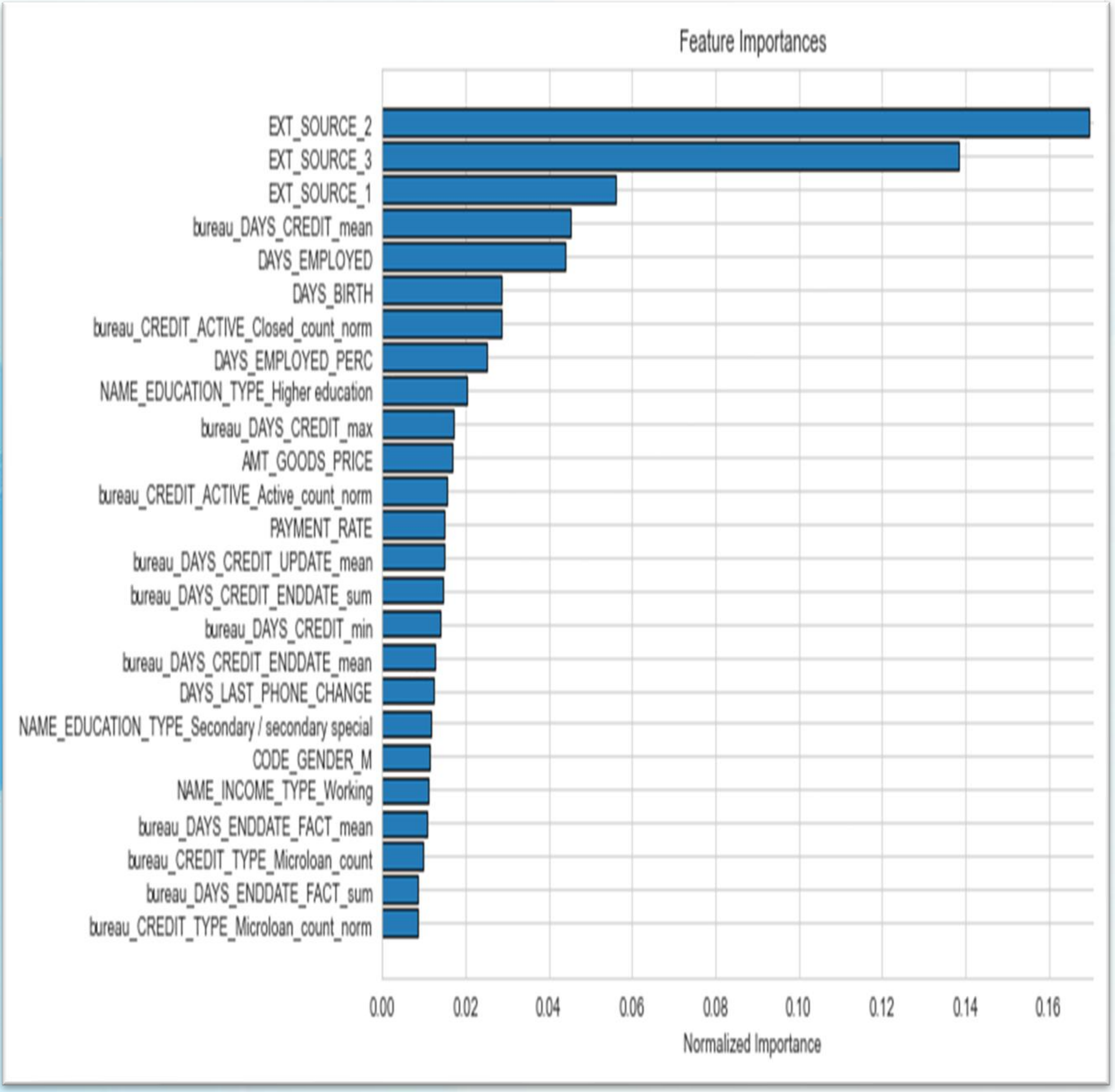
Regularization & Hyper parameter Tunning

- Tried Random Forest with feature engineering
  - Test data scored around: 0.72
  - Train data scored : 0.73
- Score Improved and No Over Fitting.
- Generalize to unseen data.



Feature Importance

- The most important features are Normalized credit score from external data source.
- Most of Important features are the ones I engineered.
  - Statistics of supporting table aggregation



Conclusion

- Home credit can quite rely on this model as a **secondary option** for now as it needs further enhancements.
- Home credit should consider these advices before lending money to applicants.
  - ✓ Beware when contract type is **revolving** as a larger amount of Revolving loans, comparing with their frequency, are **not repaid**.
  - ✓ Widowed have the most repayment rate so there should be more assurance when lending them.
  - ✓ Should focus on applicants with no children & (1: 2 children) as they are most frequent and with highest repayment rate.
  - ✓ Shouldn’t lend to applicants with 9 and 12 children as 100% of them don't repay.
  - ✓ Pay attention when someone isn't working or on maturity leave as 40% of them doesn’t pay back.

Thank you

Next Steps

- For EDA .
  - Explore the bereau dataset.
  - Merge the application_train and bureau on ID column
  - Explore these features (Credit status, Credit currency, Credit type, Duration of credit, Credit overdue 'CREDIT_DAY_OVERDUE', Credit sum 'AMT_CREDIT_SUM')
  - Remove the outliers from AMT_CREDIT_SUM and better plot the distribution.
  - Explore the Previous application data
  - Plot these features (Contract type, Cash loan purpose, Contract status, Payment type,Client type)
- For Feature engineering.
  - Develop new features from the categorical and numeric column from
 (Previous_application, POS_CASH_BALANCE, Installments, Credit_card_balance) datasets.
  - Then merge the to application data (Train & test)
  - Evaluate the model on these new features.

Next Steps

- For Dimensionality reduction.
  - Get important features from Randomforest
  - Select top features and evaluate.
  - If It preforms poorly I could add more feature until it gets better.
- For Data Modeling
  - I want to try some Gradient boosting algorithms like XGBoost, and LightGBM.
  - I think it will preform even better than Randomforest.