KAGGLE COMPETITION

HOME CREDIT DEFAULT RISK

PROBLEM STATEMENT

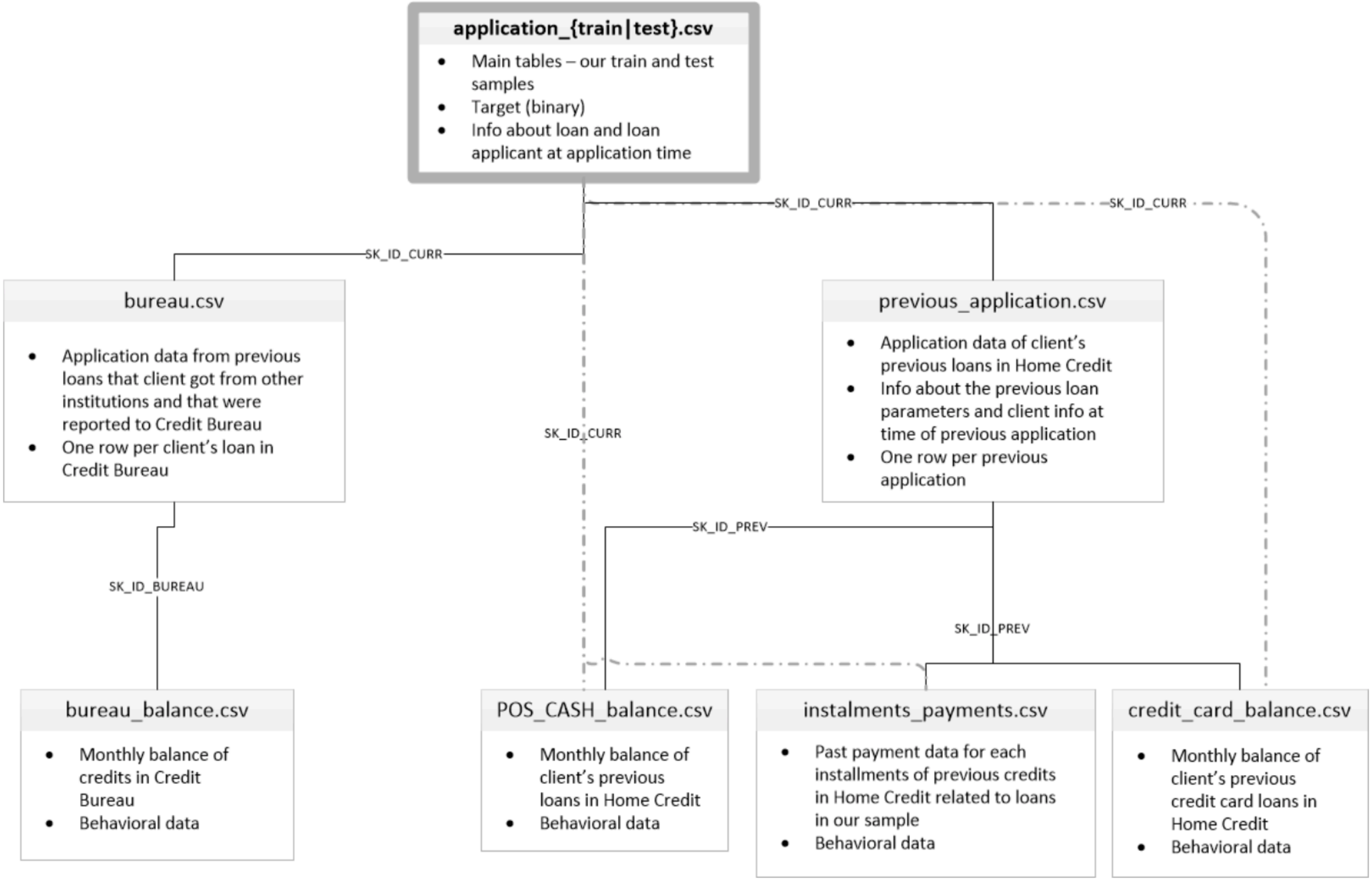
- Home Credit is an organization that serves the unbanked population with access to loans.
- This organization is trying to address the risk of such loans by utilizing various sources to make their decision on offering a loan to a prospect.
- Being able to better predict the repayment outcomes offers confidence to the financial institutions and empowers it to be able to offer much needed opportunities to their clients.

- **Objective**: Predicting the likelihood of an individual defaulting on a loan, given financial information from Home Credit and other sources.

SEVEN DATA SETS:

- **Application_train.csv** - this is the principal table and presents all of the application information. There is a single row per application, which has a unique identifier.
- **previous_application.csv** - this file presents previous applications for people in the sample through Home Credit. There is a row per each application.
- **installments_payments.csv** - this is the repayment history on loans given out through Home Credit for people in the sample. Each row is a made or missed payment.
- **bureau.csv** - credit information from other financial institutions that were reported to Home Credit. Each row represents a credit that was given to an individual in the sample.
- **bureau_balance.csv** - monthly information per credit, per loan for users in the sample. This is a long data set as it has (number of loans * credits associated for those loans * month duration for each of those credits) rows.
- **POS_CASH_balance.csv** - similar to bureau_balance.csv, this data set is the internal version of the previous monthly breakdown of balances for consumer credit and cash loans that were taken out through Home Credit.
- **credit_card_calance.csv** - each row in this data set represents a monthly balance of credit cards that were issued to applicants in the sample through Home Credit.

LINKS ACROSS THE DATA SETS



application_{train|test}.csv

- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv

- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv

- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv

- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv

- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv

- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv

- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

PROBLEM STATEMENT

EXTENSIVE DATA PREPARATION

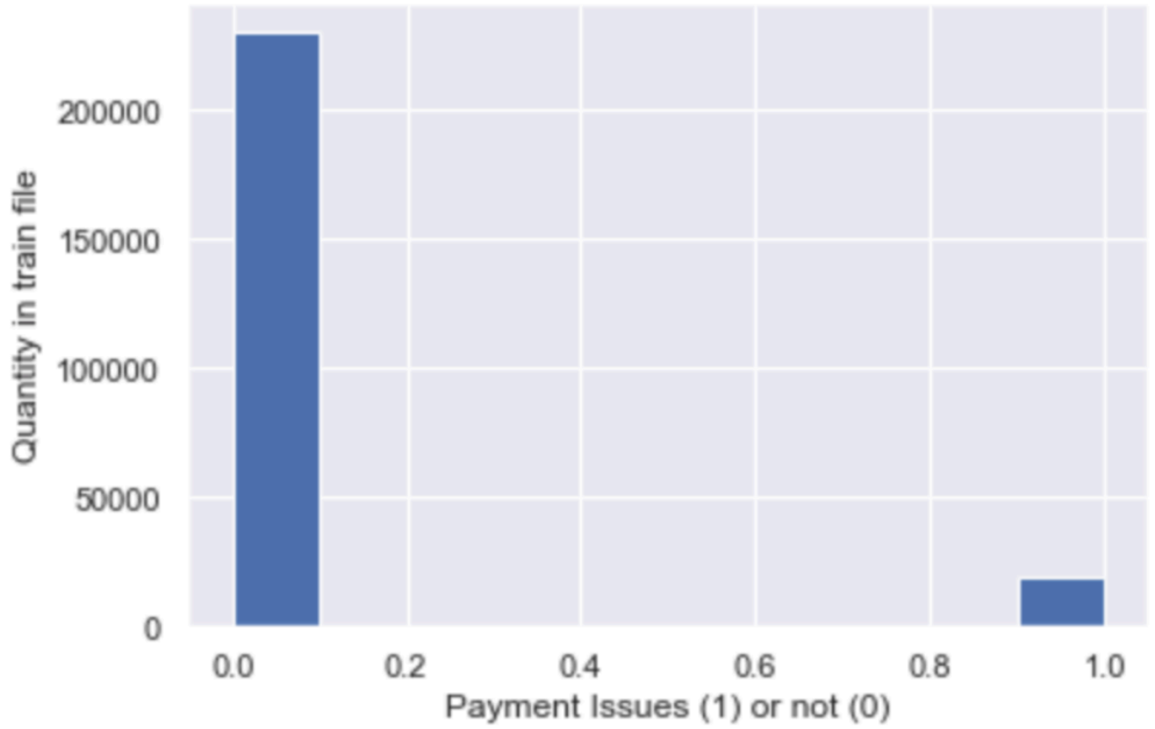
- *Numeric variable conversions* - function took in continuous features and returned the aggregate information in the form of the mean, max, min, count and sum.
- *Categorical conversions* - function returned the mean and count for the categorical variables in the supplemental data sets.

- All supplemental data sets (aside from the application_train_ were grouped using the above two functions. In the case of data sets that had previous transactions, two levels of grouping were necessary.
- This resulted is over 3000 variables that were the statistical metrics of past clients transactions.

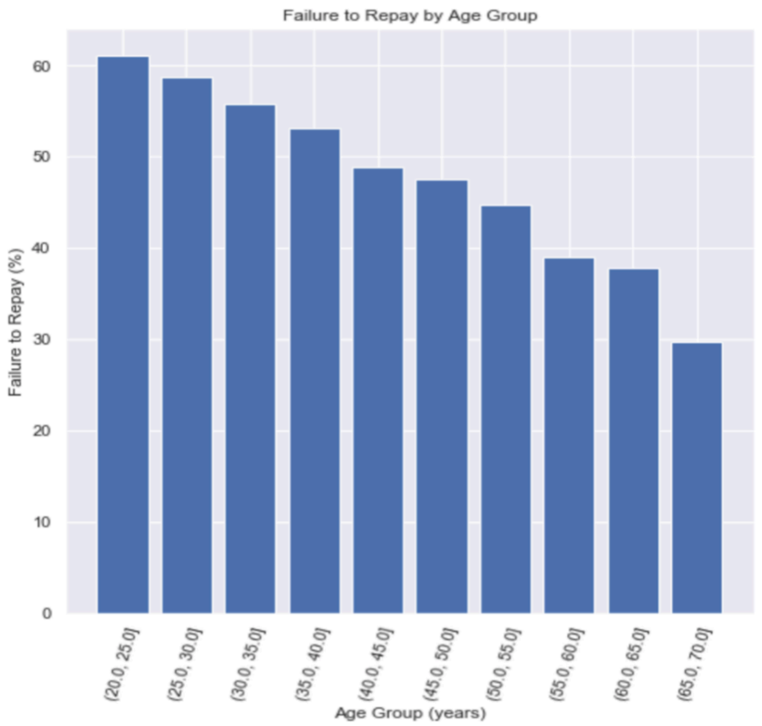
NEW SUBJECT MATTER VARIABLES

- *Days employed percentage* - a ratio of the days employed divided by the days since birth.
- *Income Credit percentage* - The ratio of the credit to income. Theory would suggest that the higher this ratio is, the more difficult it would be to repay the loan.
- *Income per person* - The income for the applicant divided by the number of people in the household.
- *Annuity Income percentage* - An attempt to show "repayment power" through the ratio of the annual annuity to the yearly income.
- *Payment rate* - The ratio of how much the individual will have to pay annually by the total credit of the loan.

INTERESTING PATTERNS IN THE DATA



A small percentage of the train data featured individuals who had challenges repaying their loans in some way.



When binning the age variable of the applicant, there appeared a very interesting pattern in relation to the failure to repay the loan.

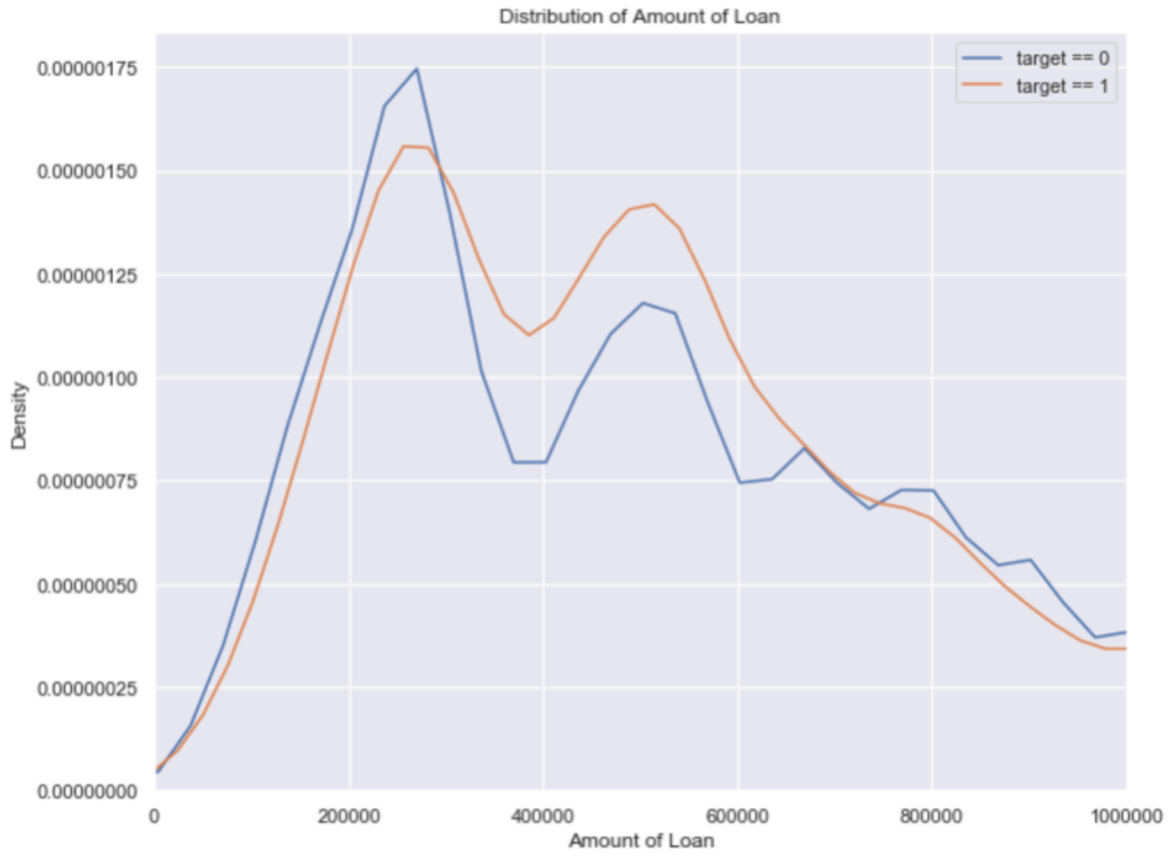
Further work: *It could be interesting to look at some binned features for further model improvement.*

OUTLIERS AND DISTRIBUTIONS

A number of outliers were identified and were corrected. An example of one such, illogical, outlier had to do with the days employed variable, which had a negative 1000 years inputted in a significant number of observations.

Additional, perhaps more logical but nonetheless out-of-the-norm, outliers, included annual income, which had a a value of 117M on the edge of its distribution. These were adjusted to within three standard deviations of the mean.

Distribution differences in relation to the target variable helped give an idea of the types of variables that could be important for distinguishing the difference in our subsequent models. Once such visible difference can be seen in the Amount of Loan variable.



PERFORMANCE CRITERIA

Due to the nature of the problem - identifying likely individual with challenges in repaying the loan - the AUC ROC measure was used to evaluate the models.

Accuracy would not be an appropriate measure in a case in which guessing that all individuals would be able to repay, as a 90%+ accuracy would not be meaningful.

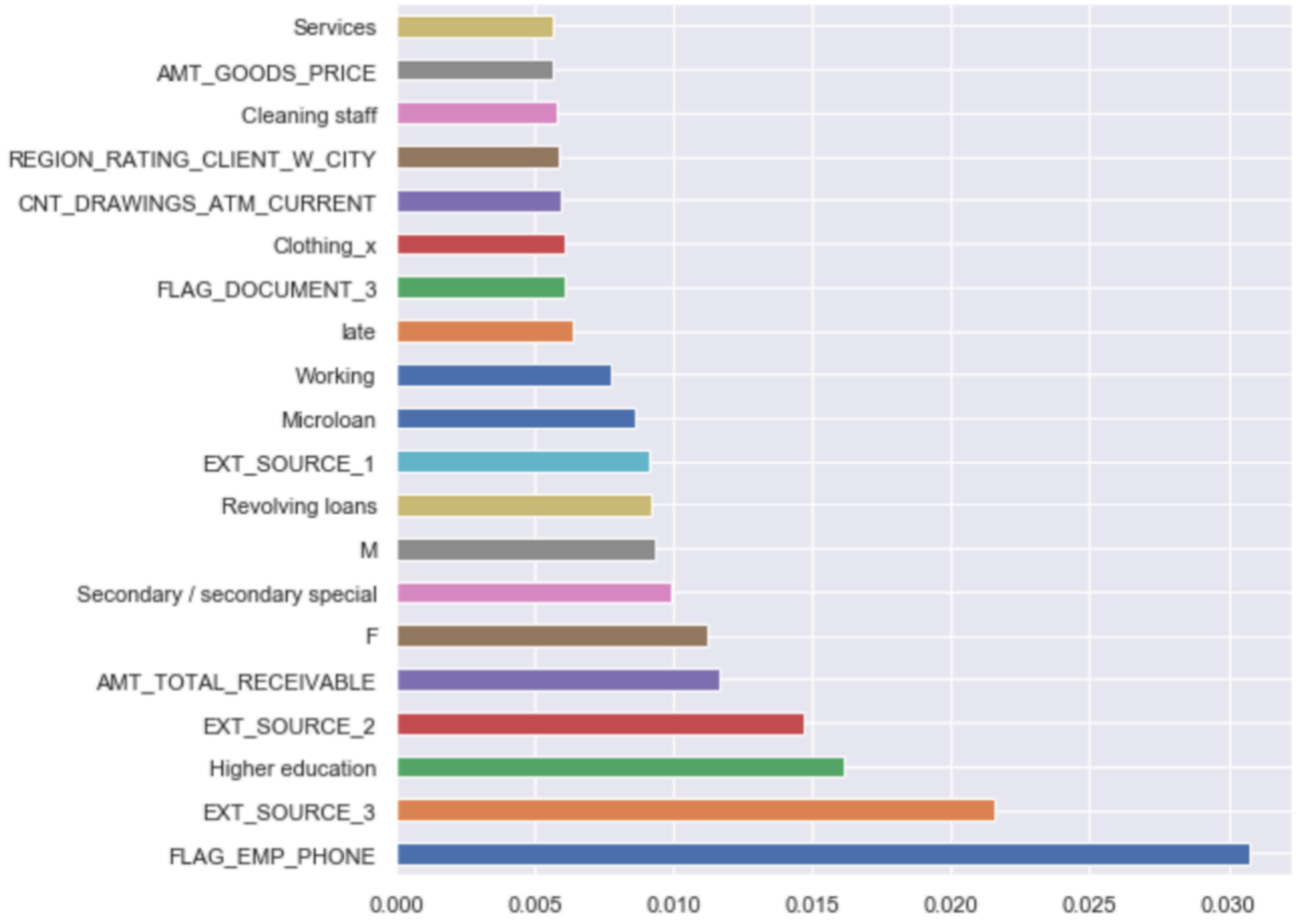
As such, the use of the AUC ROC, the ratio of the True Positives to the False Positive observations as well the Confusion Matrix, which showed a breakdown across all four types of identifications of the prediction, were utilized for performance measurement.

EARLY MODELS AND RESULTS

The first models to be run included a smaller and simpler data manipulation. All past aggregates were done through mean GroupBy. Additionally, these did not include any special subject matter variables.

Early results in terms of the accuracy of logistic regression classifier on test set was an underwhelming 0.57. Subsequent models using a Random Forest classifier gave AUC scores as high as 0.697 after some iterations of hyper-parameter tuning. Lastly, using the XGBClassifier, results as high as .704 were achieved - again after some hyper-parameter tuning.

Interestingly, three key variables stood out from the feature importance breakdown of the last model - the EXT_* features, which were unidentified metrics from an external source.



GXBOOST AND A MORE ROBUST DATA SET

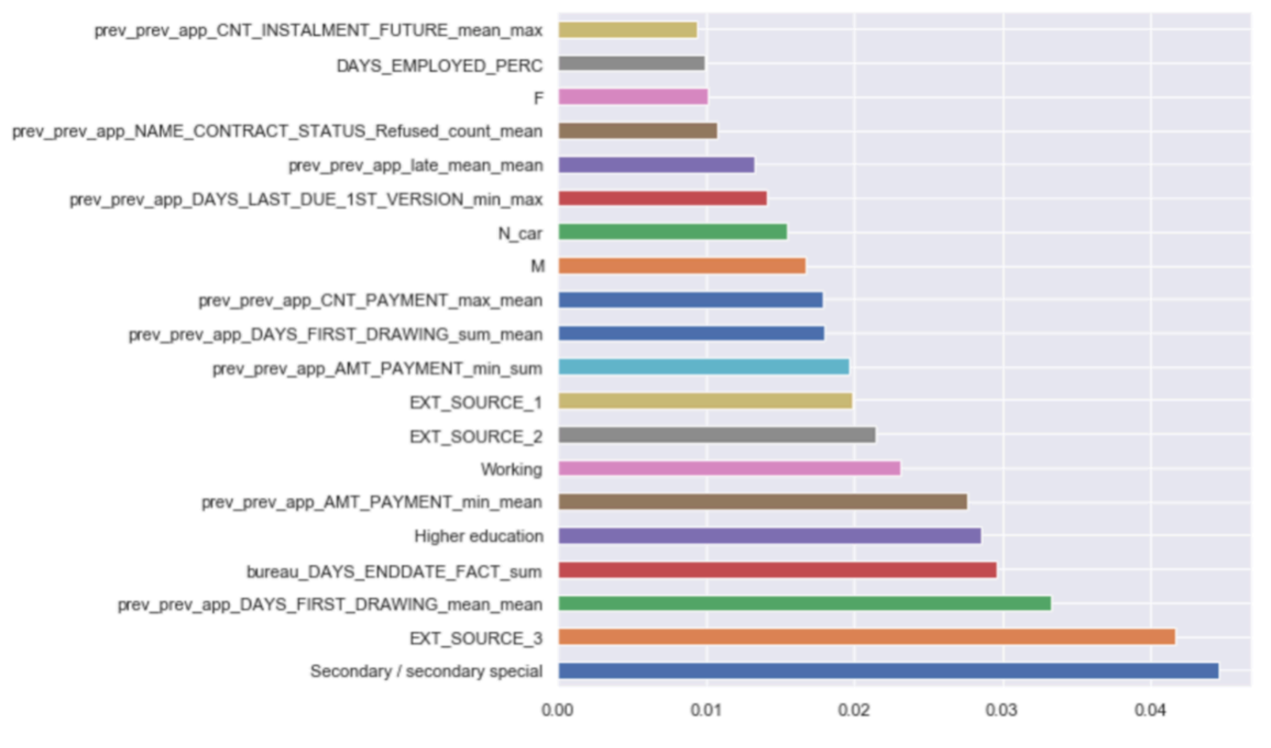
In order to attempt to better the performance of the early models, the before-mentioned data manipulation was conducted. This brought up the size of the data set from 400 variables to over 3000 and included the subject matter variables.

Bayesian Optimization was used to tune the XGBoost Classification models. Subsequently, a kfold cross validation with 10 splits was conducted to evaluate the validity of the models.

The averaged CV result was a roc_auc of 0.774.

This is a significant improvement on the early models that did not include as many variables and only relied on the mean groupings of past financial transactions. It can therefore be concluded that the new variables significantly help in the identification of features that might make someone more likely to have challenges in repaying a loan.

FEATURE IMPORTANCES



Interesting takeaways:

- **Education** is a key contributor - whether it is linked with other opportunities or financial hardships, it is interesting to see that having only secondary education (no higher education) is the leading for contributions to this model.
- **External resources** - while these variables were described in much detail, it is clear that they group some key elements that describe ones financial capacity. All three of these variables were important for all previous models and were in the top 10 features in this last model.
- **Percentage variables** - One that stands out is the DAYS_EMPLOYED_PERC as it represents how long someone has been in their current role. Interesting to see that longevity in a role, on average, negatively contributes to the ability to repay a Home Credit Loan.

Lastly, while it is difficult to understand the exact meaning of a grouped variables when a number of different stats are combined (particularly without specific subject matter expertise), it is evident that the major grouping exercise with several statistics was helpful in creating some important variables that positively contributed to the model.

FUTURE SCOPE OF WORK

Here are a couple of ideas of what I would try if time allowed:

- **Additional subject matter variables** - While some such variables were created, I am in no way an expert on financial loans. Bankers, or other in charge loans, could help identify key contributing factors to their decisions and what rations are considered. It is well possible that some ratios of income, age, number of defendants and more could be fruitful.
- **Rigorous imputing of data** - The key EXT_* variables were imputed by the mean of the observations, while it is likely that a better method exists. Given their importance to the models, a small change in the identification of the proper imputing could have positive results. One idea would be to run regression models in order to predict the EXT_ variable values for an individual given other data points that are available.