Home Credit Default Risk (Kaggle Competition) - Final Report

Misha Salkinder

Problem Statement:

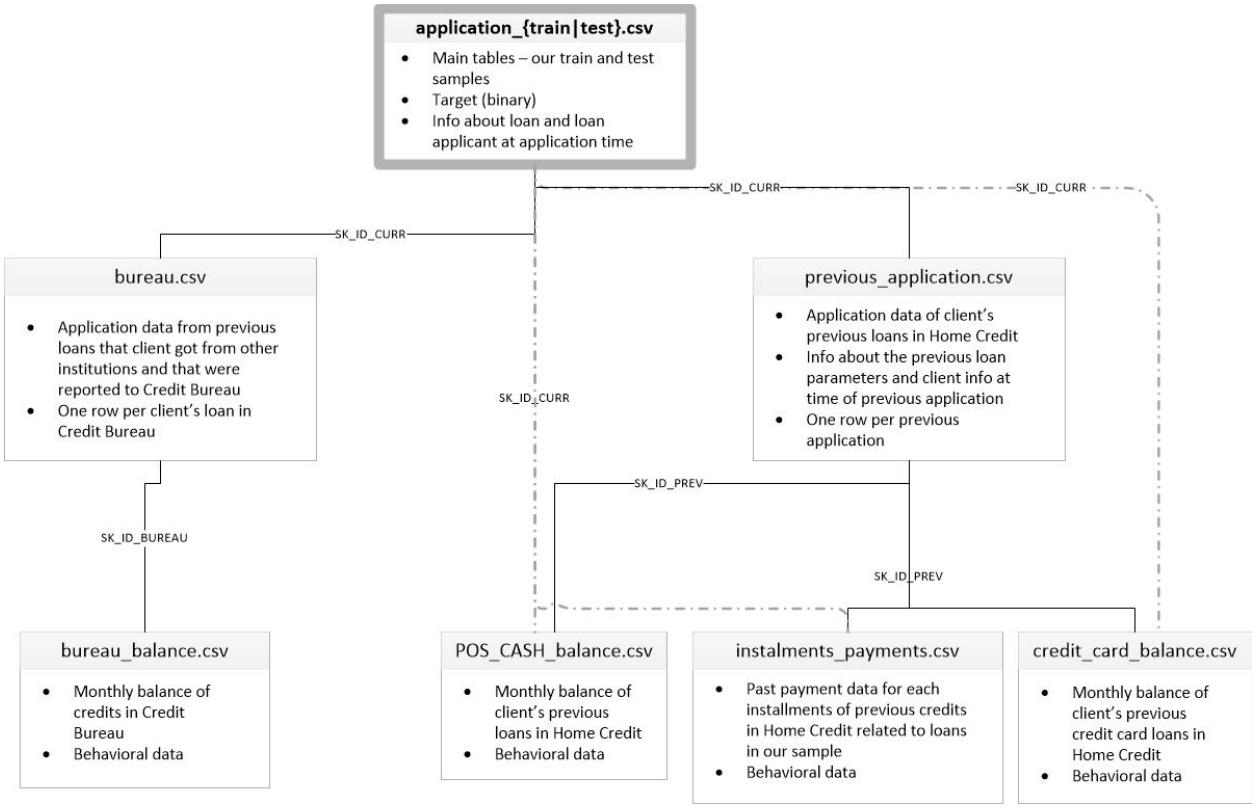
Home Credit is an organization that serves the unbanked population with access to loans. Such individuals that do not have a built-up credit score have a challenging time securing loans from financial institutions. Home Credit are trying to address this challenge by utilizing various other types of information sources to make their decision on offering a loan to a prospect. This though, comes with risk as a client's ability to repay is more difficult to discern. As such, Home Credit are seeking to understand whether a client is likely to be able to repay a loan and subsequently make a decision to offer or reject the loan or perhaps adjust some of the loan conditions. 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.

Dataset Description:

There are seven data sets that are at the disposal of Home Credit:

> **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 between the 7 data files can be seen here:



Data Wrangling and Cleaning:

There was significant preparatory work necessary in order to combine all of the data into a single workable file. Since the data in the application_train data set is on the application level, it was necessary to group data in the other data sets to the same level.

Initial data preparation involved mostly grouping variables by the mean value of the feature. While these averages were helpful in the building of the early models, it was necessary to group past transactions and subsequently information about the client by more than the mean values of the features. As such, two functions were created:

- *Numeric variable conversions* - This function took in continuous features and returned the aggregate information in the form of the mean, max, min, count and sum.

- *Categorical conversions* - As per its name, this function returned the mean and count for the categorical variables in the supplemental data sets.

The number of variables that were created through these functions resulted in an exponentially larger number of features. As opposed to grouping by the mean, which would have resulted in a single variable for each feature by client, this method extracted as many as 25 features from a

single numeric feature. This is because information on a past transaction would first be grouped by the statistics mentioned above and then go through the same process when grouping by the client. This process resulted in the growth of the data-set from 300 features to over 3000.

Once grouped, these data-sets were joined as per the above links graphic to be finally joined to the main application file by the Id of the current application. Because most applications that are present in the main file do not have previous available history, there were several empty values for each of the applications.

Subsequently, there were approximately 19% of the data that had the majority of data missing - this is because of the inability to match the current application with this previous application information. This data was discarded due to having a large enough data-set and the ratio of the defaulters to non-defaulters was similar after discarding these applications.

Some other important decisions were taken place before we moved to exploring the data:

- Nan occupation type was presumed to be unemployed and was marked as such
- 3 key external sources (which later proved to be critical for modeling) had a number of missing values, which were imputed through the mean. These were EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3.
- Values that were blank for noting who the applicant was accompanied by were marked as 'unaccompanied'.

New Variables:

Aside from the variables mentioned above, which were created systematically, a number of new variables were created, mostly through ratios of seemingly important financial characteristics:

- *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.

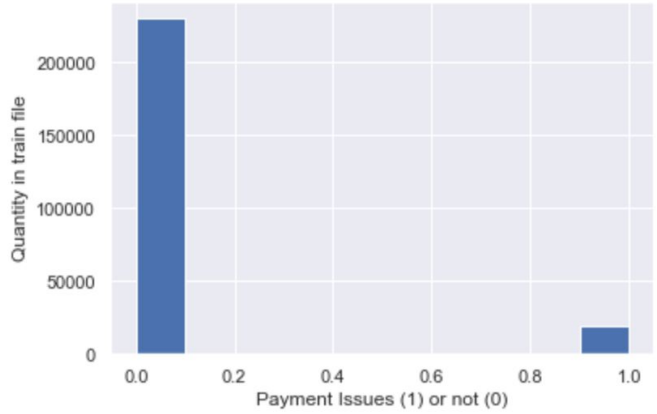
- *Income per person* - The income for the applicant divided by the number of people in the household.

- *Annuity Income percentage* - An attempt to show "repayment power" through the ratio of the annual annuity to the yearly income.

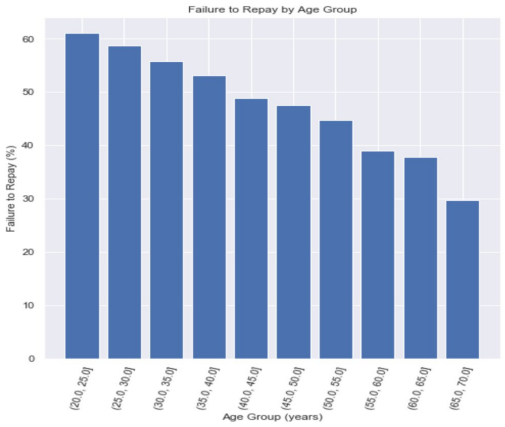
- *Payment rate* - The ratio of how much the individual will have to pay annually by the total credit of the loan.

Data Exploration:

The distribution of the target variable was important to review to decide whether upsampling or downsampling will be needed in order to balance the proportion of instances with payment issues to those without. Indeed, due to the ratio of the quantities in the train file, downsampling was later utilized.



Distribution plots as well as pair-plots helped identify any immediate correlations as well as outliers. Of the variables that were explored, most seemed reasonably distributed. For example, the age of the applicant variables did not seem to offer any unexpected observations, while it did later show a strong correlation to the target variable.



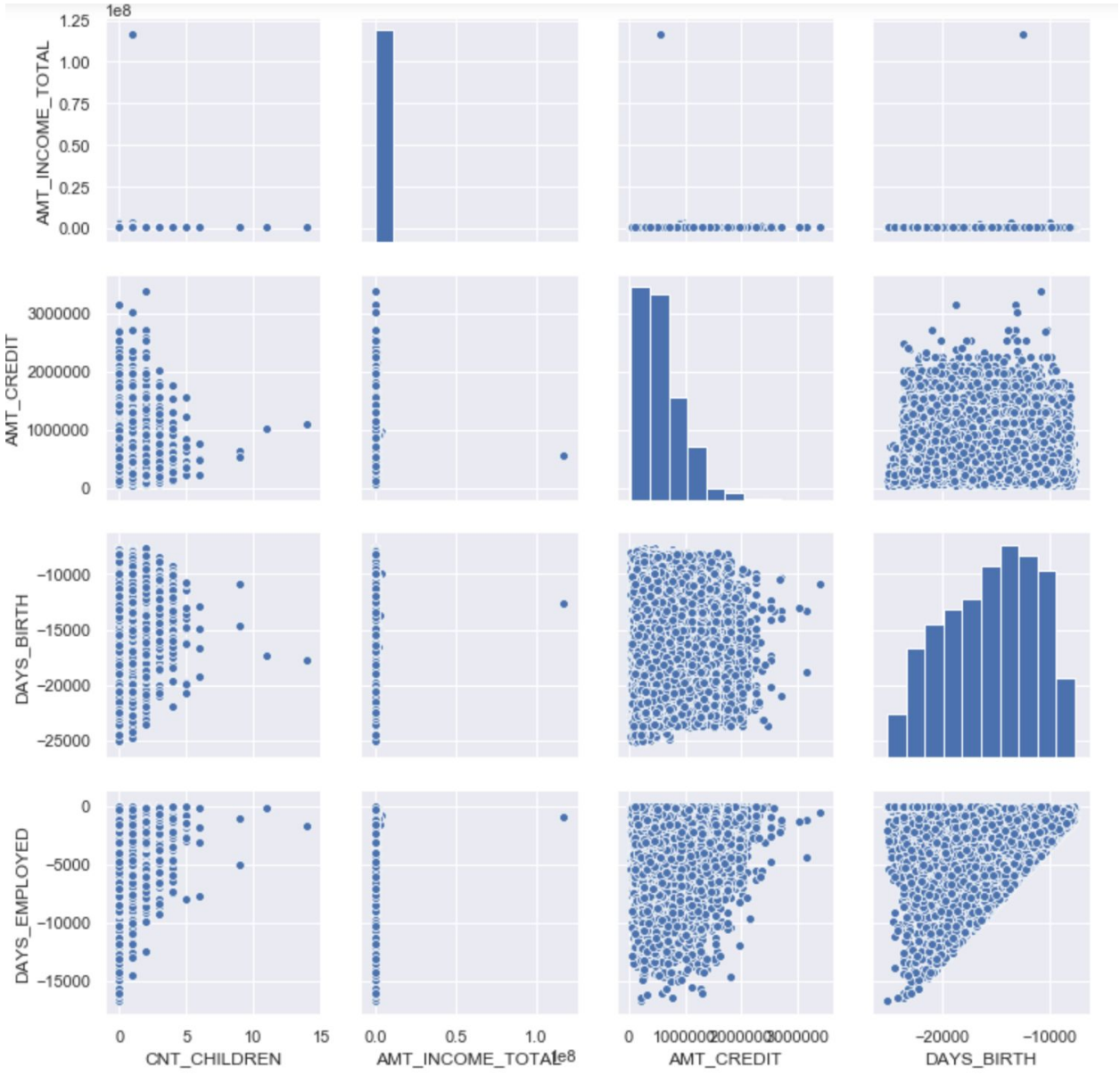
There were a few exceptions however that were adjusted:

- The ['DAYS_EMPLOYED'] variable, which denotes the number of days an applicant was employed prior to the application date had 18% of the records with -1000 years of employment. This must have been a coding issue, which I decided to address with the assumption that these were unemployed and changed those records to zeros.

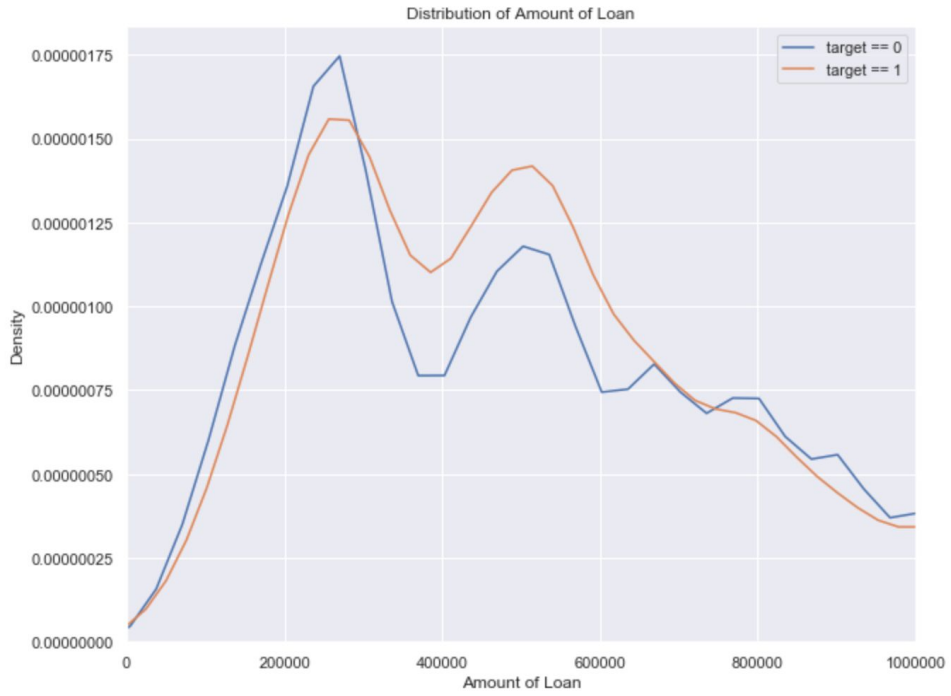
- The ['AMT_INCOME_TOTAL'] variable had an extremely long tail on the distribution with the maximum value of $117M. Given the extremely large an atypical salary, I decided to

cut down the data to within three standard deviations from the mean.

Pair-plots:



Subsequent exploration included looking at correlation heat maps between variables in the train file to identify potential interesting patterns. Similarly, 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.



Before moving on to modeling, the data was cut down to address collinearity through Variance Inflation Factoring. This helped eliminate some of the duplicate variables that resulted in the merging of the data as well as some highly collinear variables.
While the initial Linear Regression models excluded some variables, such as EXT_Source_1, EXT_Source_2 and EXT_Source_3, these later proved to be very significant in the Random Forest and GXBoost Models. As such, these variables were reintroduced into the data set for the Regression models as well.

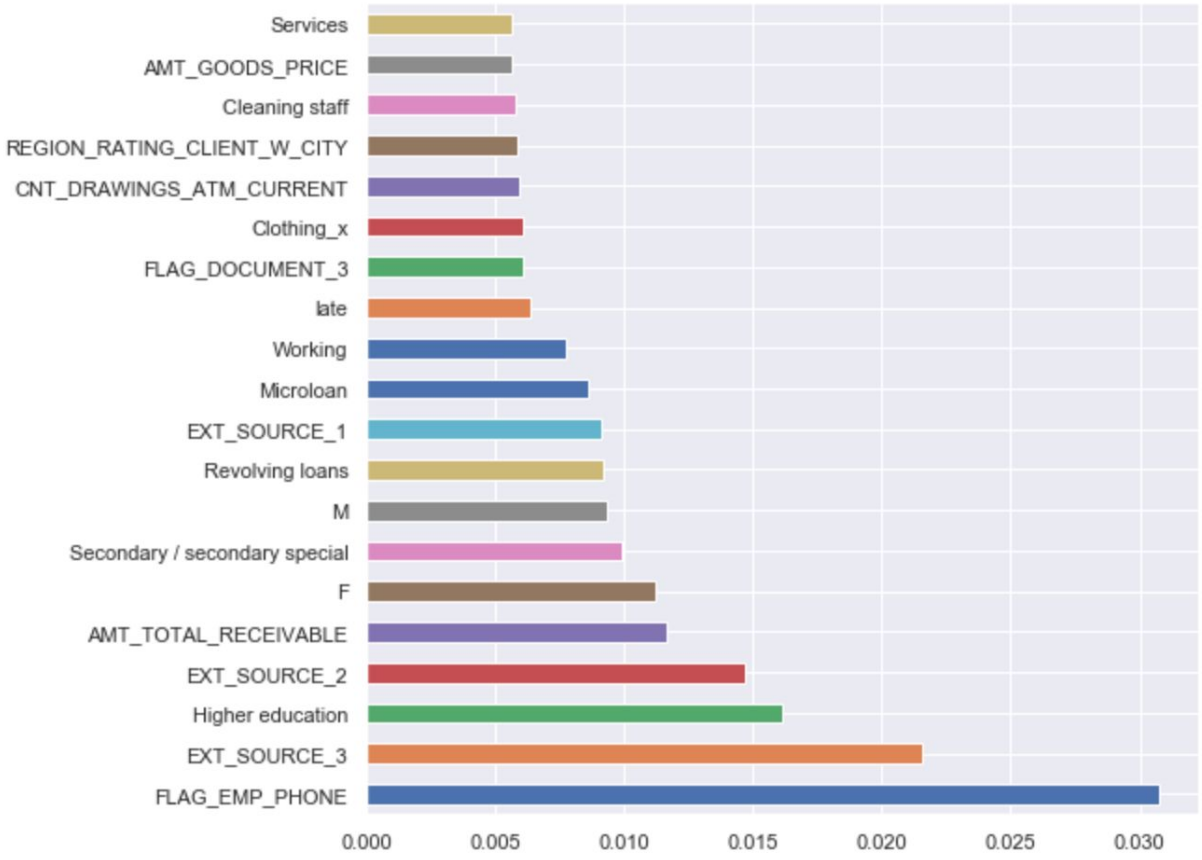
Early Models:

The measuring criteria for this competition is the AUC ROC, which is a measure of the ratio of the True Positive to the False positive observations in our test data.

The 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 hyperparameter tuning. Lastly, using the XGBClassifier, results as high as .704 were achieved - again after some hyperparameter tuning.

The image below shows the feature importances of that last model. It is interesting to see how the EXT_SOURCE variables proved to have a heavy weight in these models. Also, it is interesting to see that the above mentioned age variable (DAYS_BIRTH) was the fourth highest feature in terms of its importance for the model.

Feature Importances:



XGBoost Tuning and Final Feature Importances:

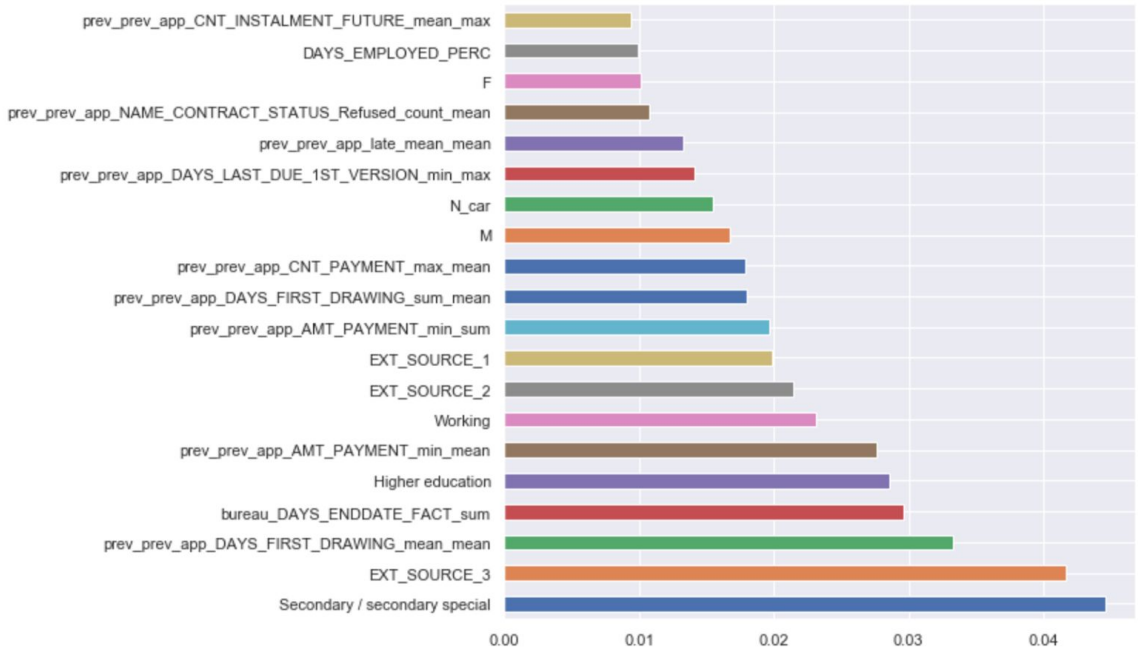
Bayesian Optimization was used to tune the XGBoost Classification models. Subsequently, a kfold cross validation with 10 splits was conducted to evaluate the validity of the models.

The averaged CV result was a roc_auc of 0.774 with a standard deviation of 0.05.

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.

It is interesting to review the feature importance to try and identify some of the key contributing characteristics to the challenge to repay a loan.

Feature Importances:



So what are some of the 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 one's 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:

While the results are close to the top of the Kaggle leaderboard, this exercise could use some further feature engineering to further improve the models. Here are a few ideas of what I would try if time allowed:

1. **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 dependants and more could be fruitful.

2. **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.