**Home Credit Default Risk Prediction Report**

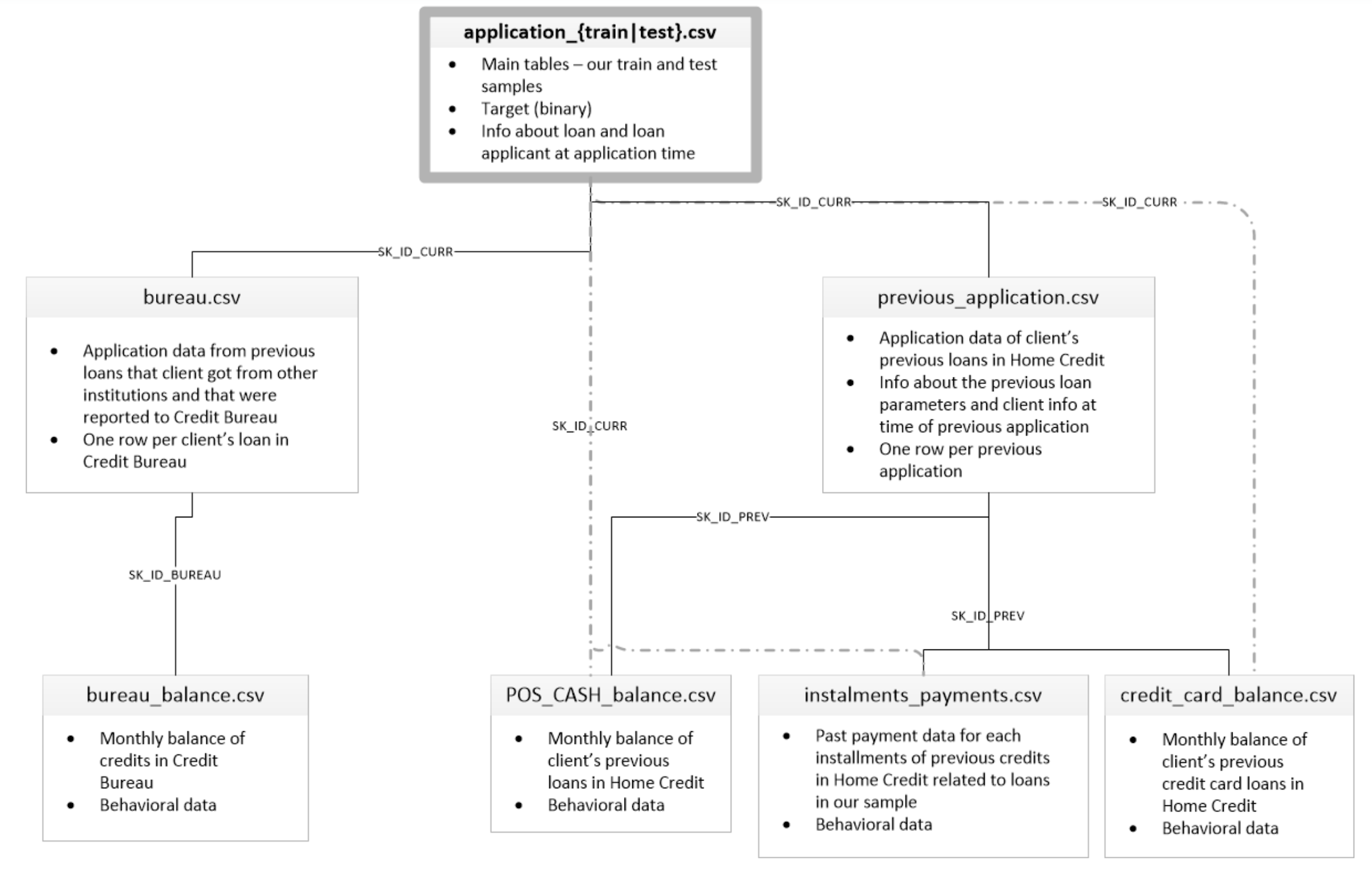
Long Fang, Xia Fu, Feiman Li

1. **Executive Summary**

To ensure that underserved people have a good loan experience, Home Credit uses various data, including telecommunications and transaction information, to predict the repayment ability of its customers.Forecasting whether or not a borrower will repay a loan is a critical business need. In this project, our group wants to predict the probability of repaying a loan for each borrower. The data is provided by Home Credit, a service dedicated to providing lines of credit (loans) to the unbanked population. The dependent variable is the probability of default and independent variables are features related to the performance of repayment. Through optimization, our group would compare different models and features to secure a relatively good performance to predict the result.

1. **Data Description**

There are 122 variables and 307,511 observations in the train dataset. As for the type of the 122 variables, 67 of which are float, 43 are integer, 13 are strings. As for data scope, we have a total of 218 variables including bureau balance data, installments payment data, credit card balance data, POS cash balance data and personal information as train dataset. All the datasets are recorded data and are concatenated with key variables. As for Data Structure, most of the variables are numeric and some of them, such as gender, whether their own realty and income type are categorical variables and need to be label encoded or set as dummies. As for granularity, there are some variables such as the number of requests for loans in a week, month and a year. For these kinds of granularity, we could just include one of the time scales after visualization. As for quality, the new dataset has lots of missing values since we used left join to merge our datasets. Fortunately, time did not impact all datasets. Therefore, we do not need to worry about the influence of time for data split and whether it is stationary or not.



1. **Evaluation Methods**

Since the dependent variable is a factor, we decide to use Logistic Regression and Advanced Tree Model such as Bagging Trees and Boosted Trees. To evaluate the accuracy, we first create a null model as a baseline, that is assuming that everyone will not default. This null model is used to compare with our model and select useful models. As the accuracy of the null model is pretty high, to evaluate our model performance, we choose to use the ROC curve, a two dimensional graph in which the false positive rate is plotted on the X axis and the true positive rate is plotted on the Y axis. A good model should have an ROC curve near the left top corner, which means TPR is always equal to FPR. Also, when we measure a classifier according to the ROC AUC, we do not generate 0 or 1 predictions, but rather a probability between 0 and 1. Thus, when we get into problems with unbalanced classes, accuracy is not the best metric.

1. **EDA**

**4.1 Data examination**

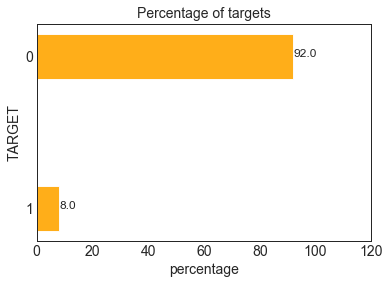
All the datasets connected with the primary key are easy to merge. Firstly, we concatenate the train and test dataset provided by the website, which has a total 356,475 observations. Then, we merged all 8 datasets as a new dataset, which has 356,255 observations and 218 columns. After merging our datasets, we did keep almost all observations.

**4.2 Data visualization**

To make sure how many features we can put into use in our final dataset, we checked for the missing values. For the application train dataset, 68 variables have missing values. We filled those missing values with either the mean or directly deleted, which will be explicit in the concrete analysis below. We begin the EDA with the application dataset.

**Train dataset:**

The target in our analysis is the “TARGET” column in the train dataset, which represents the result that the loan will be repaid (0) or not (1). 92% of people Target=0 repaid for their loan, the other 8% are those who are not able to pay for their loan.



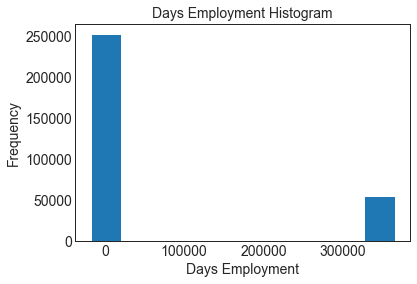
We wanted to see the distribution of the lender's age. Because in the dataset, it gives the days from birth rather than age, so that we divided the days by 365 to estimate their age and plot the figure below. The average age for our observations is 43.9. The minimum age is 20, and the maximum is 69.



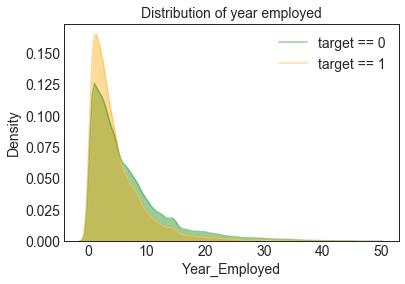
We wanted to know whether the age of lenders’ contribute to their final result. From the plot below, we found that young people more tend not to pay back than older people, which makes sense. Thus, age is an important factor that we should include in our model.



We then went deep into the days employed of our observations, and we found that there are many outliners who have worked 365243 days, which is like errors because that about 1000 years.

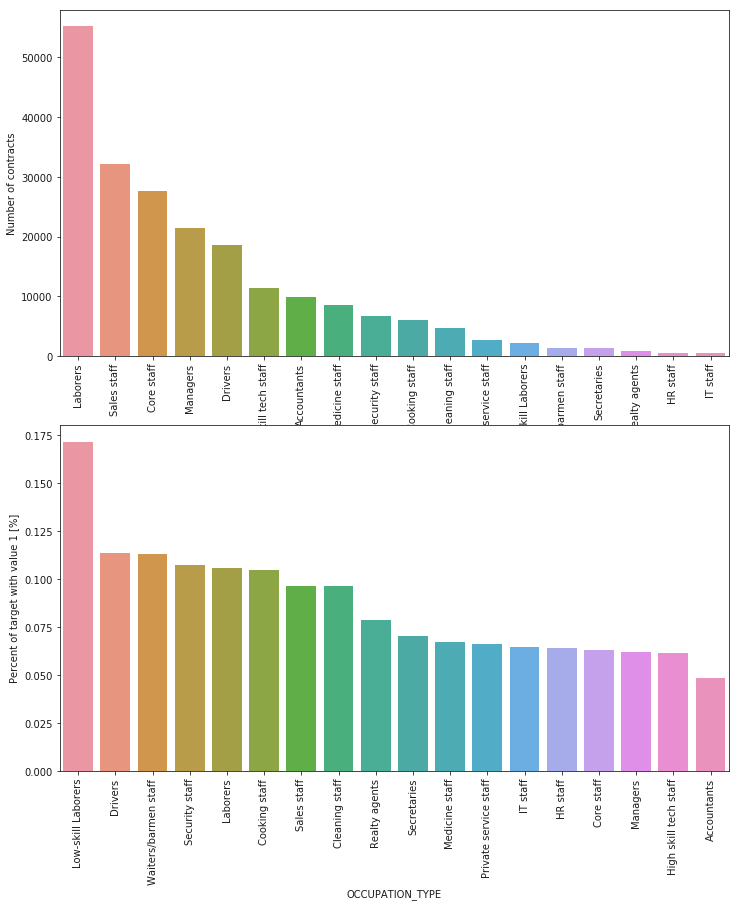


So I plot only those with employment days less than 365243 for the two target classes and divided the days by 365. Some difference between class 0 and class 1 could be found in this figure, that the distribution for people who tend to default is more skewed than those who are employed for a longer period of time.

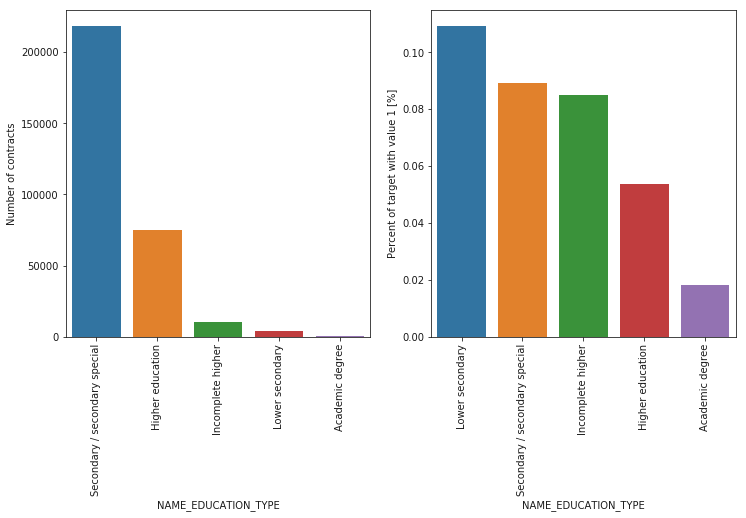


**Other Datasets:**

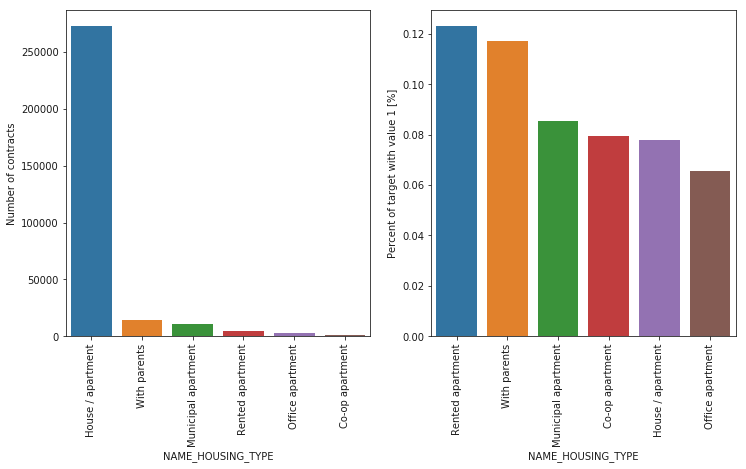
Most loans are borne by low-skilled workers, followed by salespeople. IT staff have the least number of loans. The category with the highest percentage of outstanding loans was low-skilled labor (above 17%, which is double the average), followed by drivers and waiters/barmen staff. It shows that different occupations will affect the outcome.



For Education, the customers mainly take up Secondary/Secondary Special, and followed by higher education. For those who default, lower the education level, higher the default rate.

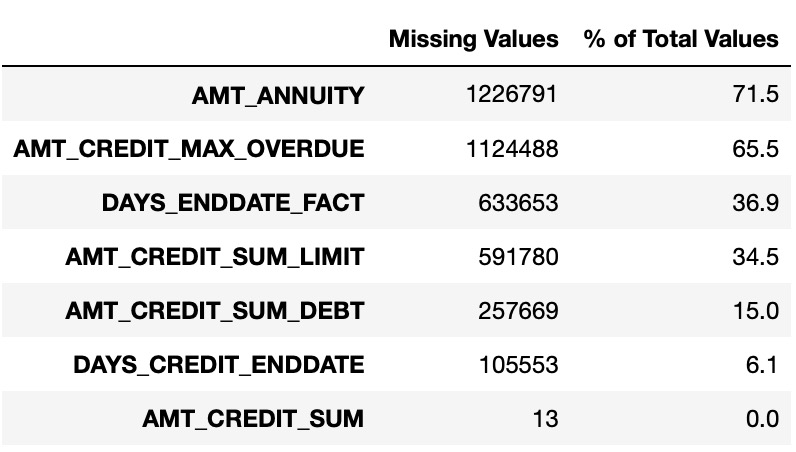


House/ Apartment contracts are the major but the default rate of house/ apartment contract type is lower than the average default rate. Rental apartment contract type, which is less than 10,000, has the highest default rate. Thus, it may not have a major effect on the target.



1. **Feature Engineering & Selection**

As we have many datasets, we need to deal with the missing value in each dataset first. For example, in the bureau dataset, there are a lot missing values in AMT\_ANNUITY, AMT\_CREDIT\_MAX\_OVERDUE, DAYS\_ENDDATE\_FACT and AMT\_CREDIT\_SUM\_LIMIT. As in EDA and for common knowledge, some variables may not contribute a lot in the model, we choose to delete those variables. Some of them, such as AMT\_ANNUITY, even if they have a large amount of missing value, are important to us. We choose to keep them and will deal with them later.



We merge train dataset and test dataset and then merge with another six datasets. There are many categorical variables such as AMT\_REQ\_CREDIT\_BUREAU and NAME\_YIELD\_GROUP. Thus, we get dummies and right now we have 367 variables and our target. Considering we still have many missing values, we try to input missing values. For categorical variables, the missing values are not considered anymore after we get dummies. According to different values, we use different methods to deal with. For those variables which values between 0 to 1, we set them as 0 or -1 to avoid them to have effects on the model. For those variables having large values such as income amount or credit amount, we choose to use mean or median value to fill the missing value, depending on their distributions. For example, AMT\_ANNUITY has some very large outliers that may affect mean value. Thus, we choose the median value to fill the missing value.

Because we have too many variables after merging and one hot coding (up to 367), we choose to use PCA to reduce dimensions. We set to keep 99% of the variation, and we retained 16 components. Then we projected our whole dataset to the 16 eigenvectors and split it into training (0.7) and validation (0.3) sets. Later on, we found that the more features we keep, the better the model. So we change to maintain 99.99999% variations, that is 36 dimensions retained.

1. **Modeling :**

Different algorithms have been used for the prediction.

**Null model:**

Null model is used as a baseline for any other models. We assume all of our validation sets will not default, which is 0. This model has the accuracy of 0.919 and the ROC AUC of 0.5 when we randomly applied the possibilities. Our other advanced model should have ROC AUC that are larger than 0.5.

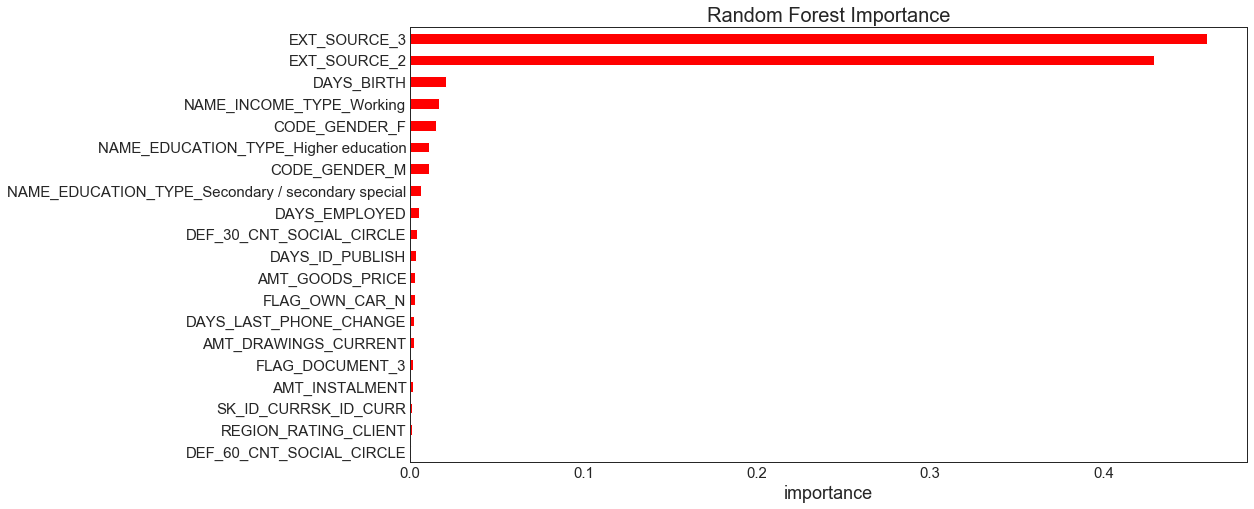
**Logistic regression model:**

We used the binary logistic regression model to fit and produce the probability for each record that will default. The ROC AUC is 0.61, which is improved compared to the baseline. The parameters are set default.

**Decision tree model:**

After adjusting the hyperparameters (max\_depth=6, max\_leaf\_nodes=100, min\_samples\_leaf=20), our model gives the ROC AUC of 0.624, which seems better than our logistic model.

In addition, considering that PCA is the dimension reduction method that the dependent variables are not taken into account. So we tried to use the non-transformed dataset to fit the same model and predict the possibility of default. The model gives us the importance for each feature and we plot the top 20 most important features and their importance:



This model has ROC AUC 0.71, which is a lot higher than the transferred dataset. This tells us that using the dimension not reduced dataset to train models is the priority.

1. **Future Improvements**

**7.1 Address outliers**

We could avoid outliers leveraging the process of fitting models and improve model performance. To deal with outliers and anomalies, we could just delete outliers or create a class named other’s to avoid extreme values. For numeric variables, we may take AMT\_ANNUITY as an example, if there are too low annuity or too high which are far away from the average values. We need to check the distribution and concentrate on addressing the heavy right or left tail. For categorical variables, to create a new category for infrequent categories is a must. If we do not do so, some categories may only exist in the train or test dataset that cannot be detected by models.

**7.2 Better data imputation**

For some of the missing values we just assigned mean and median value for them. We need a more comprehensive EDA so that we can deal with those missing values more carefully by using possible methods. Apart from missing values, we also need to address zero values. Some interval variables such prices have no meaning when they are equal to zero.

**7.3 Use Cross Validation**

To reduce the variability, we may use cross validation, performing multiple rounds of cross-validation with different subsets, to improve the accuracy for further improvements.