Home Credit Default Risk

https://www.kaggle.com/c/home-credit-default-risk

Silvi Fitria
silvifitria1@gmail.com

Overview

Goal
Predict whether or not an applicant will be able to repay a loan using historical loan application data (probability of each applicant is repaying loan)

Data
Training : application_train.csv (307511 observations and 121 features)
Testing : application_test.csv (48744 observations and 121 features)
Data checking : around 69% most of the features have missing value

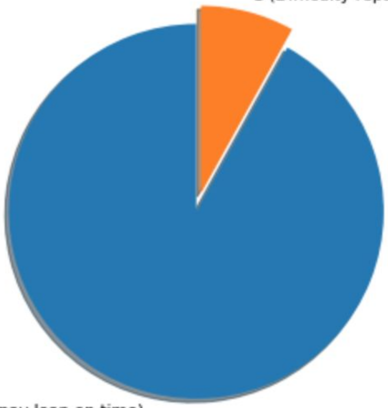
Methods
1. Logistic Regression
2. Random Forest
3. Light Gradient Boosting Machine (Light GBM)

01 Exploratory Data Analysis

...

Distribution of Target

```
TARGET
0    282686
1     24825
dtype: int64
```



There are far more loans that were repaid on time than loans that were not repaid. It’s about 8.78% applicant who had payment difficulties.

Detect Anomalies

Days of Birth

The numbers in the DAYS_BIRTH column are negative because they are recorded relative to the current loan application. To see these stats in years, mutliple by -1 and divide by the number of days in a year.

Days of Employed

The maximum value (besides being positive) is about 1000 years. One of the safest approaches is just to set the anomalies to a missing value and then have them filled in (using Imputation) before machine learning.

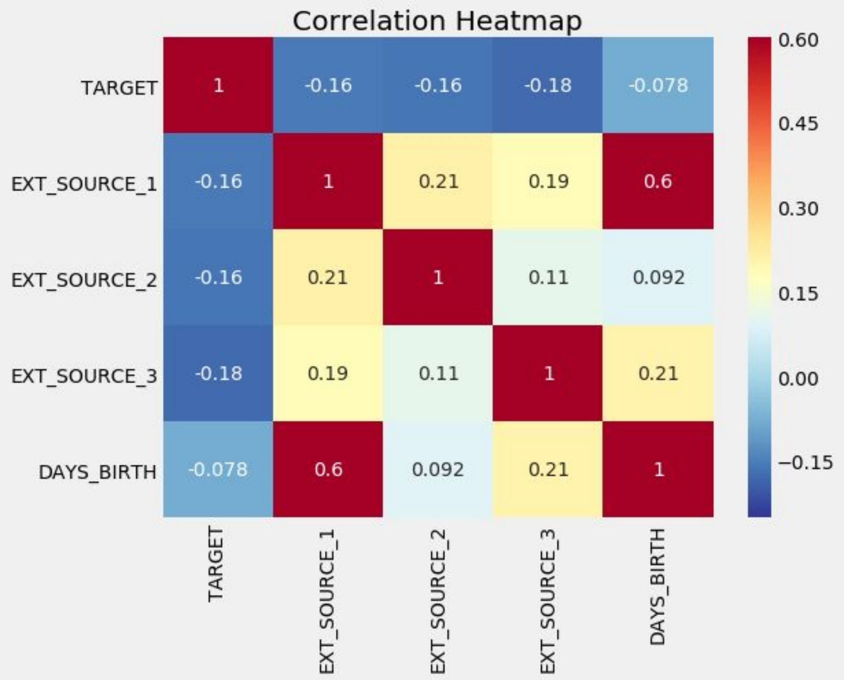
69% of features have missing value

Impute the missing value with median

Correlation

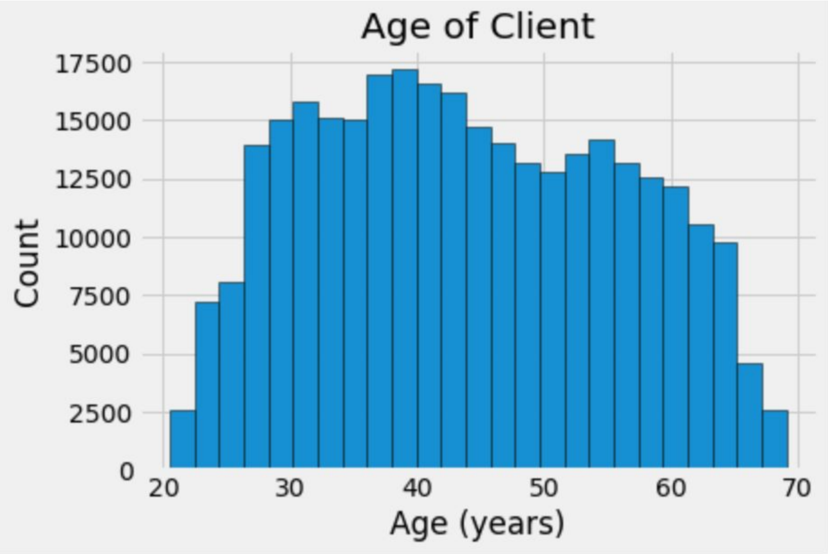
The DAYS_BIRTH is the most positive correlation. DAYS_BIRTH is the age in days of the client at the time of the loan in negative days (for whatever reason!). The correlation is positive, but the value of this feature is actually negative, meaning that as the client gets older, they are less likely to default on their loan.

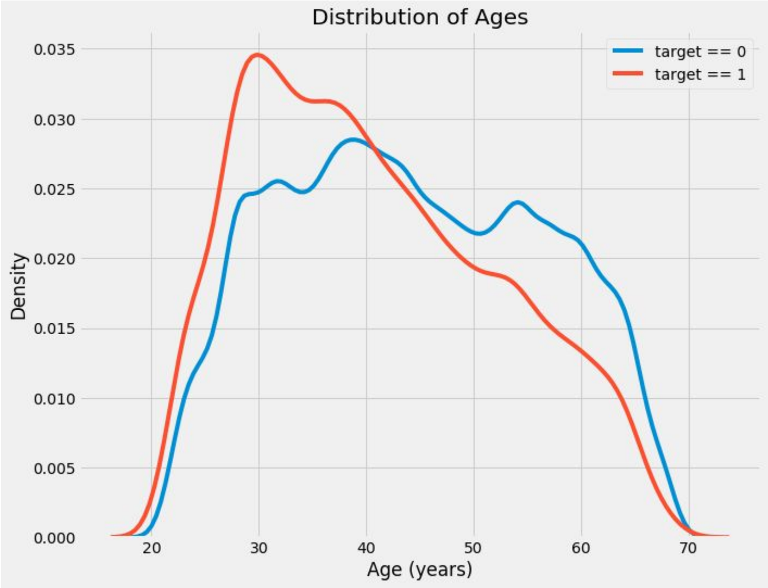
Correlation



All three EXT_SOURCE features have negative correlations with the target, indicating that as the value of the EXT_SOURCE increases, the client is more likely to repay the loan. DAYS_BIRTH is positively correlated with EXT_SOURCE_1 indicating that maybe one of the factors in this score is the client age.

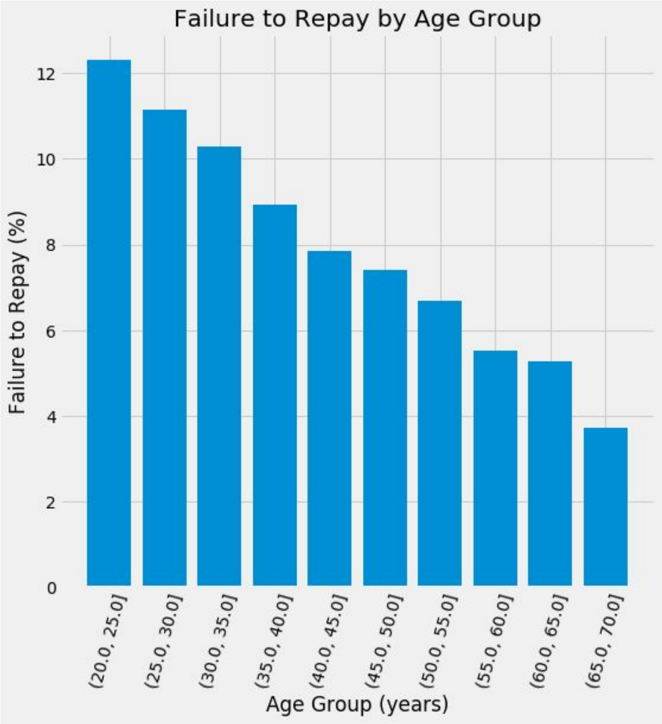
KDE Plot





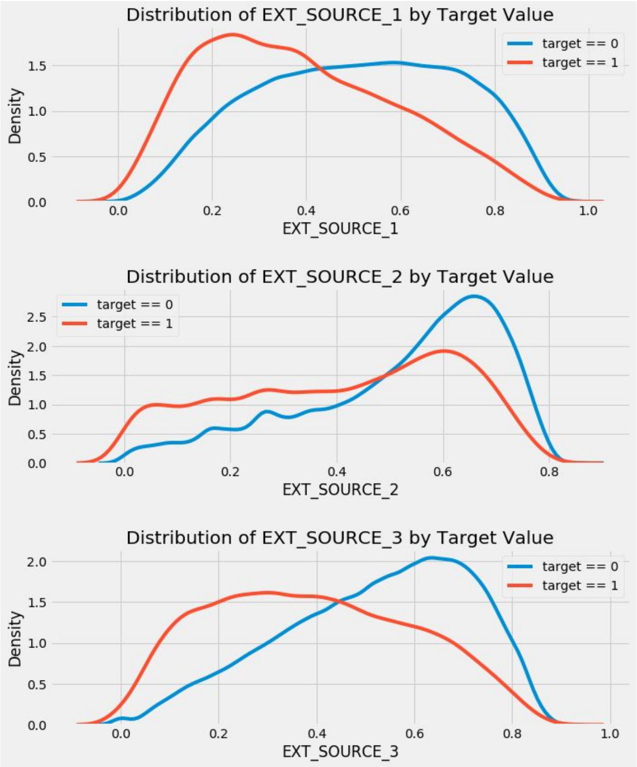
By itself, the distribution of age does not tell us much other than that there are no outliers as all the ages are reasonable.

Failure to Repay by Age



The younger applicants are more likely to not repay the loan. The rate of failure to repay is above 10% for the youngest three age groups and below 5% for the oldest age group. The 3 variables with the strongest negative correlations with the target are EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3.

Distribution of Features



EXT_SOURCE_3 displays the greatest difference between the values of the target. This feature has some relationship to the likelihood of an applicant to repay a loan. The relationship is not very strong (in fact they are all considered very weak, but these variables will still be useful for a machine learning model to predict whether or not an applicant will repay a loan on time.

02 Feature Engineering

...

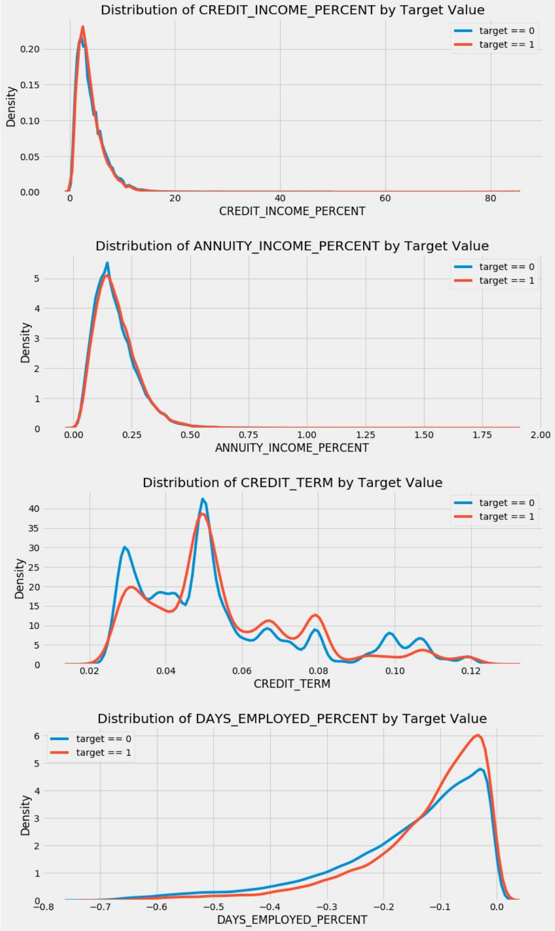
Polynomial Features

This features make powers of existing features as well as interaction terms between existing features. For example, make a new features EXT_SOURCE_1^2 and EXT_SOURCE_2^2 and also variables such as EXT_SOURCE_1x EXT_SOURCE_2, EXT_SOURCE_1 x EXT_SOURCE_2^2, EXT_SOURCE_1^2 x EXT_SOURCE_2^2, and so on. These features that are a combination of multiple individual variables are called interaction terms because they capture the interactions between variables. In other words, while two variables by themselves may not have a strong influence on the target, combining them together into a single interaction variable might show a relationship with the target.

New Features: 35 Features

Domain Knowledge Features

Exploring attempts at applying limited financial knowledge



03 Modeling

Logistic Regression

Before modelling, preprocessing the data by filling in the missing values (imputation) and normalizing the range of the features (feature scaling). This model predicted using probability between 0 and 1, if the predicted target near to 1, indicates that applicant will had a difficulty payment. Here is the head of predicted result.

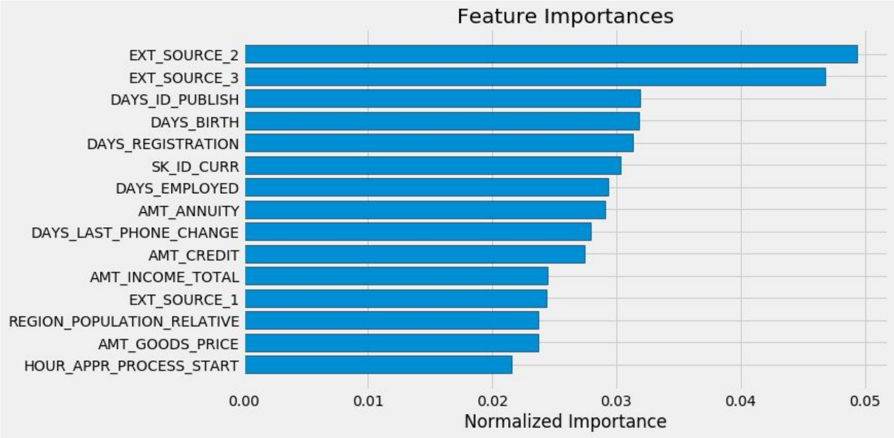
	SK_ID_CURR	TARGET
0	100001	0.087750
1	100005	0.163957
2	100013	0.110238
3	100028	0.076575
4	100038	0.154924

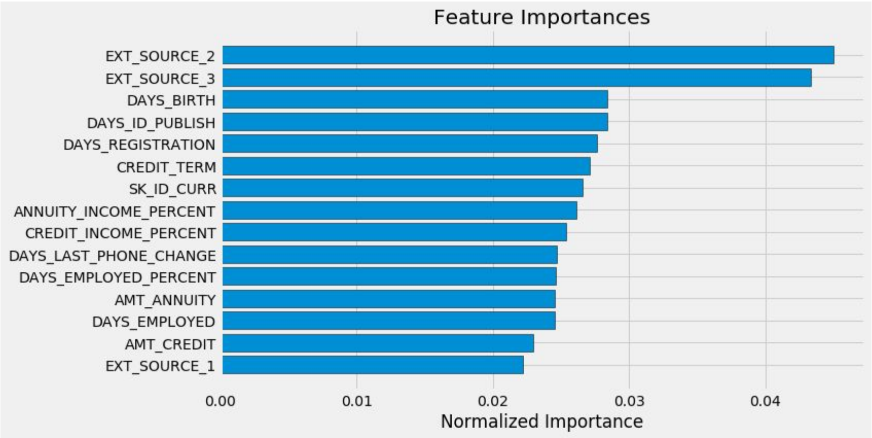
Random Forest

The Random Forest is a much more powerful model especially for more than hundreds of trees. This model uses 100 trees in random forest. This model is also predicted using probability between 0 and 1, if the predicted target near to 1, indicates that applicant will had a difficulty payment. Here is the head of predicted result.

	SK_ID_CURR	TARGET
0	100001	0.5
1	100005	0.5
2	100013	0.5
3	100028	0.5
4	100038	0.5

Light Gradient Boosting Machine (LGBM)





EXT_Source_2 ; EXT_Source_3, and DAYS_ID_PUBLISH are a significant features than other features.

04 Evaluation

...

Accuracy

To measure the accuracy, it should be use ROC AUC and also F1 Score. Because of the sample submission for testing target is univariate (only 0.5), it means not a binary categorical, that's why I use MSE and RMSE to measure the accuracy. Meanwhile, the target result of modeling is probability between 0 and 1.

Logistic Regression	
MSE	0.155
RMSE	0.394

Random Forest	
MSE	0.024
RMSE	0.154

Random Forest Domain Features	
MSE	0.104
RMSE	0.322

Based on three models deployed, random forest is the best model to predict the probability of loan repaying for each applicant. Random forest has a smallest MSE and RMSE than the other methods.

Result of probability each applicant can be seen in **random_forest_baseline.csv**

Result

Statistics descriptive of probability using the best models

Mean	0.357582
St Dev	0.057691
Min	0.13
Max	0.65
Quartile 25%	0.32
Quartile 50%	0.35
Quartile 75%	0.40

Mean of probability for each applicant is below 0.5 and probability of 75% applicants is 0.40. It means that 75% applicants will repay the loan.