A Study on Credit Default Risk

Bo Liu
Springboard Data Science
March 2020

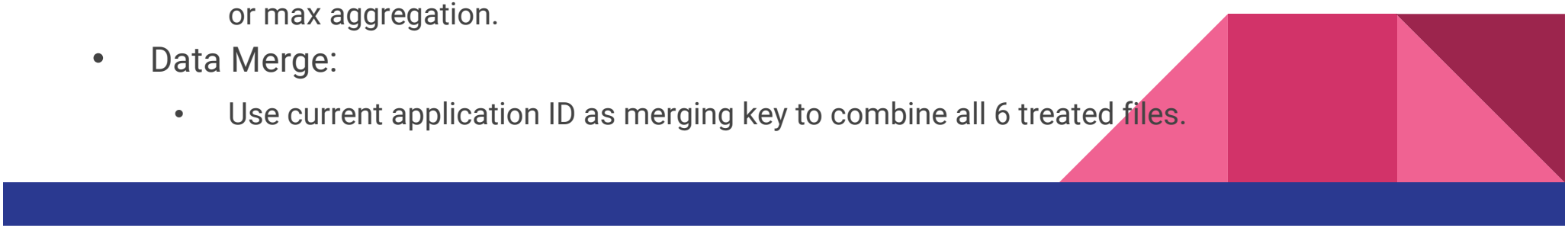
Table of Contents

Problem Statement

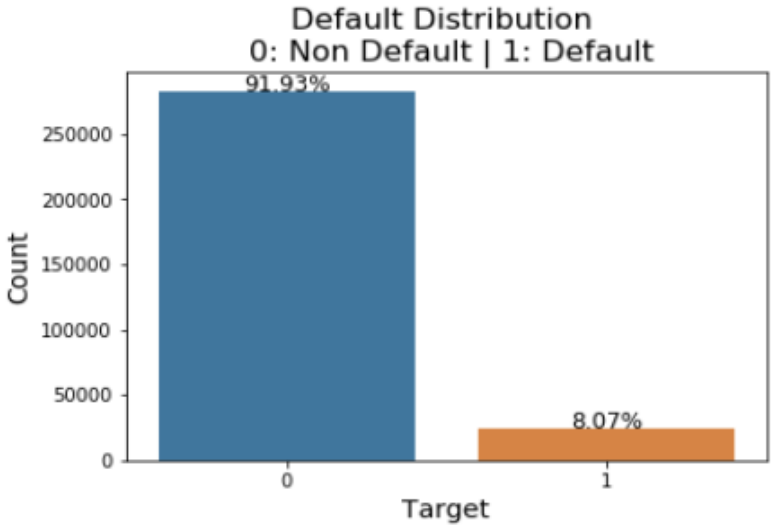
- Home Credit is a lending company that strives to provide positive and safe borrowing experience for its customers.
- Like all other lending companies, Home Credit leverages a wide variety of data sources from its customers such as bureau information, payment history of previous loans, and income information to predict the customer's credit worthiness and ability to repay the loans.
- Given the collected attributes and default / non-default labels for each customer, we are able to build classification models to predict the probability of loan default and obtain the top features that separate default from non-default loans. The model output can be used as guidance to approve or decline a new loan application.

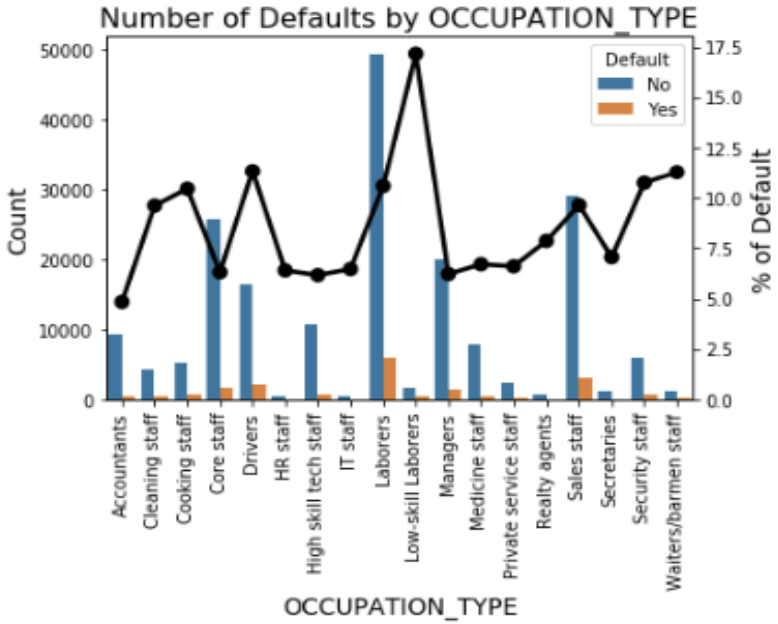
Data Acquisition & Wrangling

- Source: https://www.kaggle.com/c/home-credit-default-risk/data
- Raw Data:
  - 7 individual csv files – one is a lead current application file, the rest are supporting datasets with bureau or previous loan information.
  - Due to CPU processing power limitations, we only aggregated 6 csv files.
- Data Aggregation:
  - Numerical variables – keep one or two of the summary statistics such as average, sum, max or min within each sub ID, in this case could be previous application ID or bureau ID.
  - Categorical variables – if with high cardinality, some categories with low frequencies are grouped before one hot encoding, and aggregation is done after one hot encoding using sum or max aggregation.
- Data Merge:
  - Use current application ID as merging key to combine all 6 treated files.



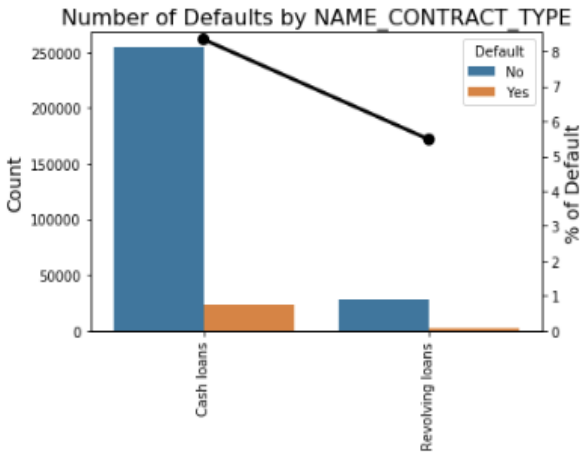
EDA – Default Rate & Categorical Variables

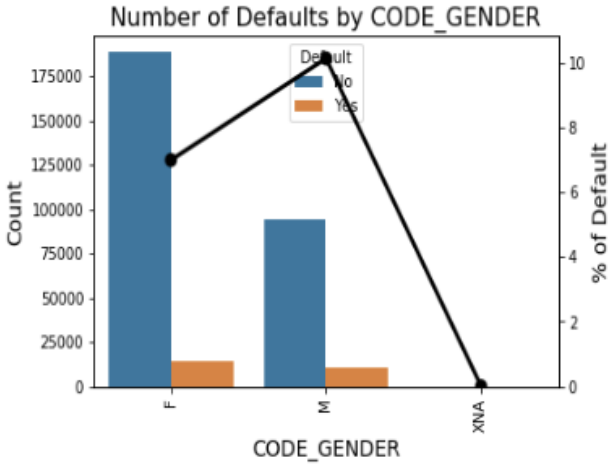


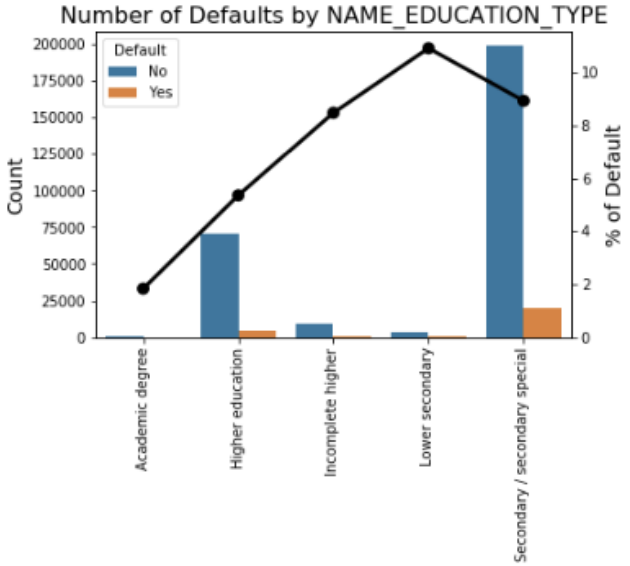


- Overall there are over 300,000 current applications with more than 300 variables after combining all 6 datasets.
- The default rate is only 8.07%, indicating highly imbalanced dataset.
- Occupation type has over 15 categories, those categories having <= 10,000 counts will be grouped into one group.

EDA – Categorical Variables

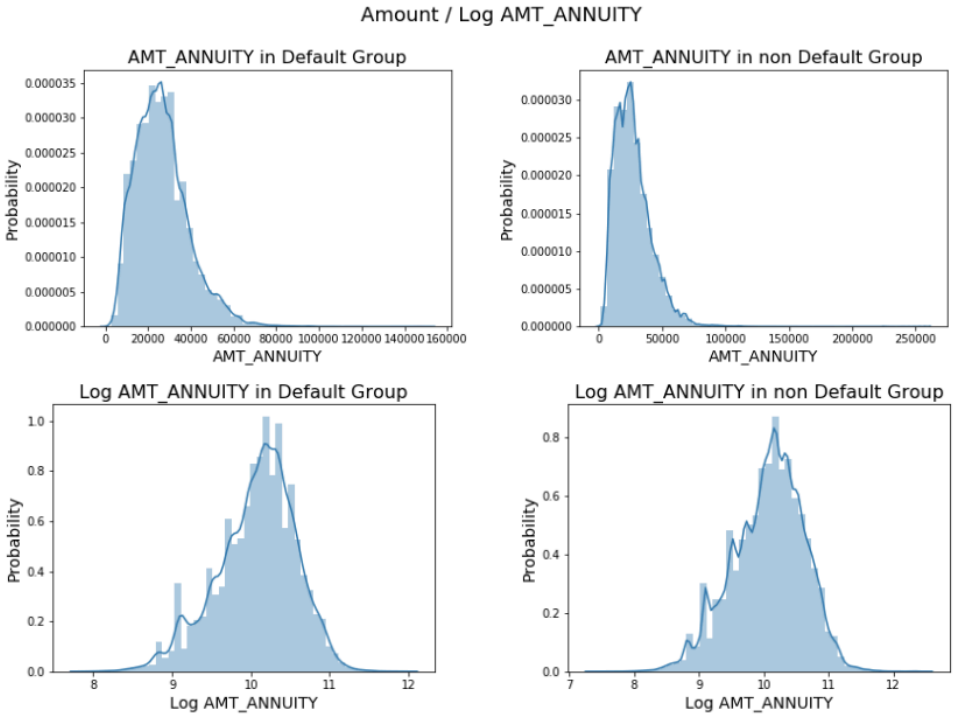


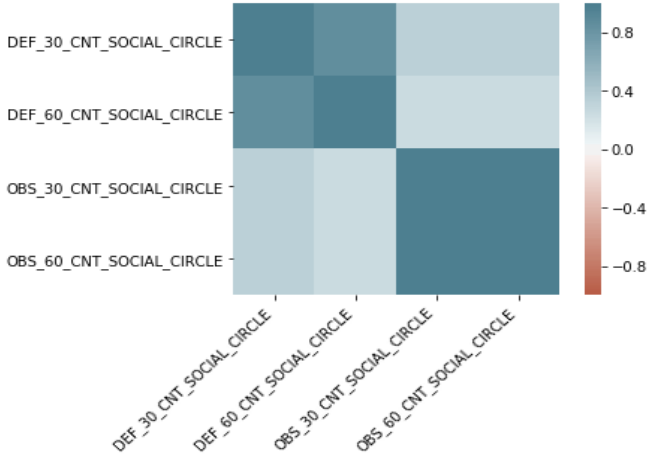




- Cash loans are the major type of loans with ~ 8% default rate. Default rate in Revolving loans is slightly lower.
- In addition to Female and Male, there a few records with XNA gender type. These non typical values are removed from the dataset.
- Majority of the customers are in secondary/special education type, which dominates the overall ~8% default rate. Lower secondary education segment has over 10% default rate.

EDA – Numerical Variables

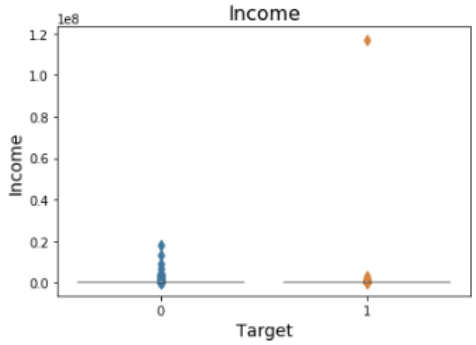


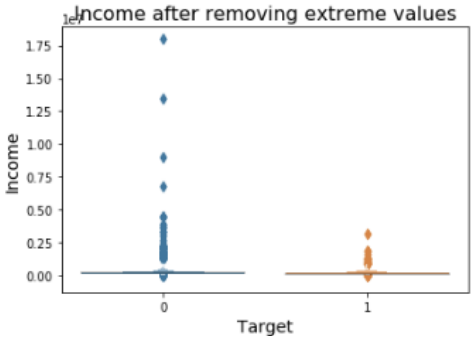


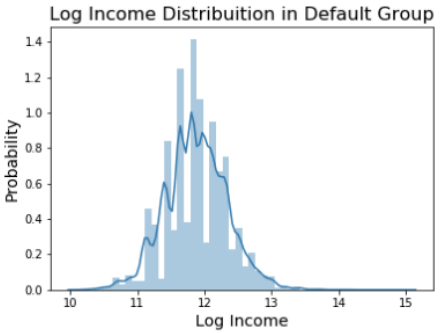
- Correlation heat map indicates there is high positive linear correlation between observations of client's social surroundings with observable 30 days past due and 60 days past due. Due to collinearity considerations, one of the 2 variables is removed.

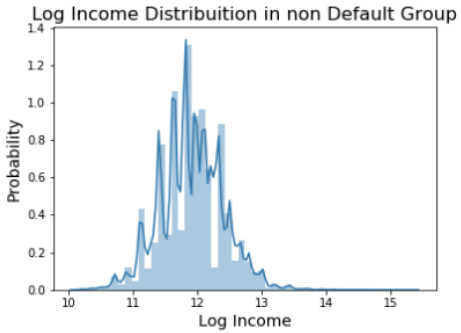
- The KDE plot shows the distribution of some continuous variables such as amount of annuity are quite skewed. A log transformation is applied to normalize the variable. Visualization of the default and non-default groups show no significant differences.

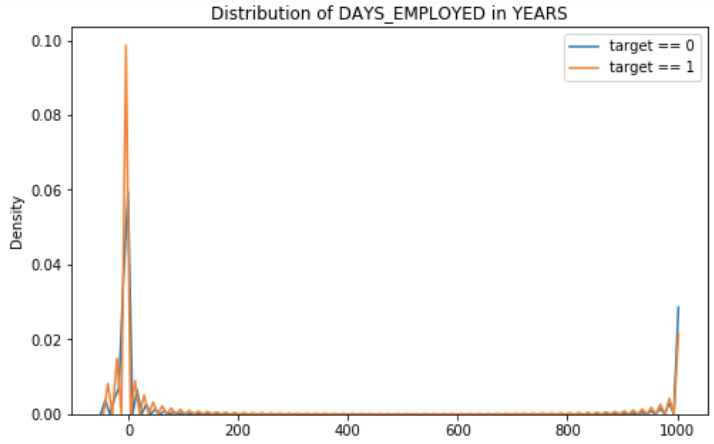
EDA – Numerical Variables











- Some employment length are over 100 years, which are apparently data errors. Further investigation indicates many age-type of variables are counting backwards from the time point of application, so the maximum should be 0. We therefore cap the values at 0.

- Boxplot shows that there is one extremely large income value in the default group ($120 million), and 4 very large values (> $7.5 million) in the non-default group. Recommend to remove the outliers and use log transformation to normalize the income data.

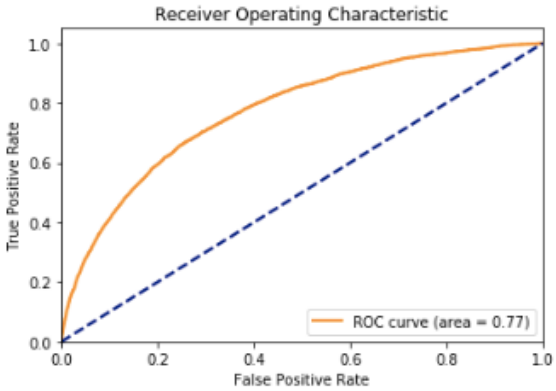
Modeling

Model	Average AUC	STD AUC
Random Forest	0.7500	0.0076
XGboost	**0.7722**	0.0074
LightGBM	0.7704	0.0065
Logistic Regression	0.7652	0.0040
KNN	0.5860	0.0053
Kernel SVM	0.7565	0.0013

- Due to the imbalanced nature of the data, we down sampled the non-default counts to be approximately the same as the default counts in the training data, while the test data is still a randomly held out 20% imbalanced dataset.
- We applied 6 classification models to the training set. For each model, a 5-fold CV with RandomizedSearchCV is run on the down-sampled balanced training set to select the best combination of hyperparameters.
- Table above shows the highest average AUC scores among the validation sets along with its standard deviation, for each of the model trained.
- XGboost beats LightGBM slightly with the highest AUC among all models.

Model Evaluation

	precision	recall	f1-score	support
0	0.96	0.71	0.82	56533
1	0.18	0.69	0.28	4968
micro avg	0.71	0.71	0.71	61501
macro avg	0.57	0.70	0.55	61501
weighted avg	0.90	0.71	0.78	61501



- Xgboost model has the highest average AUC score on the validation dataset, and hence selected to be the final model recommendation.
- We evaluated Xgboost model on the independent testing set, which is a random 20% of the original imbalanced dataset. AUC on the test set is 0.7738.
- Using 0.5 as cutoff threshold, precision and recall for each target group are shown in the report. Adjusting the cutoff threshold will lead to different precision and recall values and this will be left as a judgmental call from the credit analyst.

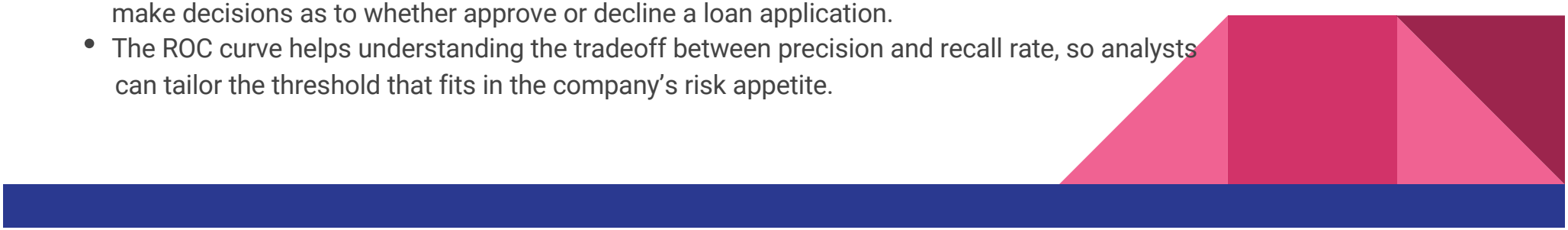
Summary

Data & Modeling:

- Data aggregation are done at sub ID level for each of the 5 supporting datasets and then merged with the major current loan application file.
- EDA and data wrangling are performed to treat missing values, data errors and outliers. One-hot encoding is used to convert categorical variables and log transformation is used to treat some numerical variables.
- The model with the highest AUC on the validation data is XGboost. LightGBM comes second and Logistic regression ranks the 3rd place, followed by SVM, random forest and KNN.
- In terms of computation times, tree-based methods are generally faster than other methods. Logistic regression and KNN take slightly longer than tree-based methods. Kernel SVM is the slowest to train and predict.

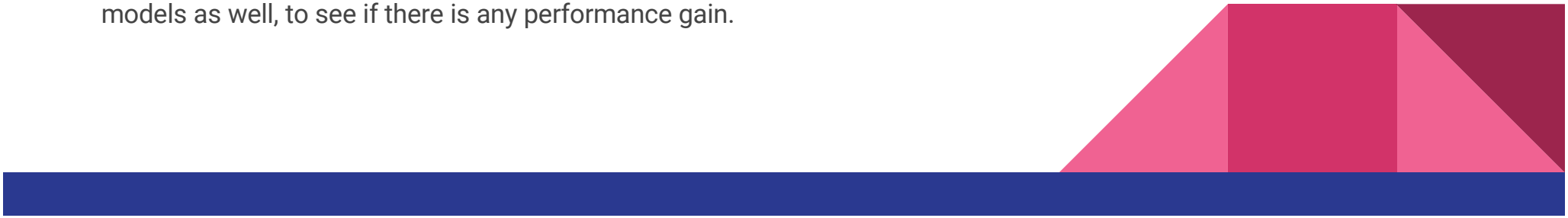
Business decisions:

- The model outputs probability of default for each application and using these numbers, analysts in Home Credit can make decisions as to whether approve or decline a loan application.
- The ROC curve helps understanding the tradeoff between precision and recall rate, so analysts can tailor the threshold that fits in the company's risk appetite.



Future Work

- Feature engineering can be further explored by looking into combinations of various columns, quadratic terms or using different aggregation methods at sub-ID levels.
- In the final dataset used in modeling, Bureau balance dataset was not included in the features due to limited time and computation power. In future work, it can be aggregated and incorporated into the training data.
- More missing value imputation techniques can be used to better impute the missing values based on overall shape of the distribution, rather than using the median for every numeric variable.
- More sampling techniques such as up sampling of the minority class, or SMOTE can be used to compare with the performances with down sampling.
- PCA or other dimension reduction methods can be leveraged to reduce the dimension of the feature space, however, model interpretability may be lost.
- In this project we only tried to use scaled features in Kernel SVM model, to speed up the convergence of the algorithm. For all other 5 models we used the original scale of the variables, as tree-based methods can deal pretty well with different scales of variables. In future work we can try using scaled / normalized features in Logistic Regression and tree models as well, to see if there is any performance gain.



Future Work

- In all of the models we trained, we throw in all of the over 300 features without picking any subset of features. In Logistic regression, L1 penalty was selected, so we automatically got some feature selection benefits. Tree-based models naturally produce the rankings of features, so we can try to use the top 100 or so to train the model again to see if there are any performance improvements. In future work, more variable selection techniques can be considered to fit a more parsimonious model rather than using the full set of features.
- In the SVM model cross validation step to select the optimal hyperparameters, due to the prohibitive training time of SVM algorithms, we only tried random search 2-fold cross validation with 2 random parameter combinations. More combinations can be tested to achieve possibly higher AUC provided with GPU or a high computation power CPU. Similar with tree-based models, more combinations can be tried when searching for best hyperparameters.
- Other classification models can be tested such as neural networks, linear and quadratic discriminant analysis, other tree-based methods such as Adaptive Boosting, or other bagging and ensemble methods.

Acknowledgement

- Thyago Porpino (mentorship)
- Kaggle (open data source)
- Springboard team (curriculum & administrative support)

Reference

Data Source:
https://www.kaggle.com/c/home-credit-default-risk/data
Data Wrangling & EDA Notebook:
https://github.com/lisalb168/Credit_Risk_Default_Prediction/tree/master/notebook
Final Report:
https://github.com/lisalb168/Credit_Risk_Default_Prediction/tree/master/reports