Georgia Institute of Technology

MGT 6203 FINAL PROJECT REPORT TEAM 63, SPRING 2023

Smart Lend: Analyzing Creditworthiness for Successful Loan Repayment

Sumanth Chundru

Yee Man Bergstrom

Julia Nordbakk

Devin LaCrosse

Natalie Mo

Table of Contents

1. Choice of Topic, Business Justification, and Problem Statement

1.1 Choice of Topic:

Getting a loan when you have an insufficient or non-existent credit history can be a challenge. This group is often denied credit or targeted by untrustworthy lenders with horrible interest rates. We want to see if we can include other alternative data points to predict if these clients can repay their loans.

1.2 Business Justification:

Limiting a client's predicted ability to repay a loan on one data point, could automatically reject a whole population of clients that would be good candidates for a loan. This puts that population at a huge disadvantage, but it also could be a missed business opportunity for lenders. Home Credit operates in a lot of countries where borrowing is still a novelty and there is a growing middle class. This includes China, India, Indonesia, the Philippines, Vietnam, and Russia. These countries have a combined population of almost four billion people. Home Credit approves about 200,000 loans every day and two-thirds are new customers.[1]This is a huge opportunity! Their ability to be successful very much depends on how quickly and reliably they can assess and predict the creditworthiness of potential customers.

1.3 Problem Statement:

We will determine a client's predicted ability to repay a loan using data other than a credit score. Using the predictors such as principal amount, loan term, and repayment schedule the highest probability for repayment will then be approximated.

1.4 Literature Survey:

In the article "Using data mining to improve assessment of credit worthiness via credit scoring models " Bee Wah Yap , Seng Huat Ong and Nor Huselina Mohamed Husain discuss how they used a credit scorecard model, logistic regression model and decision tree model to discriminate between defaulters and non- defaulters of monthly club subscription payment. All of their models were showing similar results. The conclusion was that these techniques are highly useful for the purpose of credit scoring.[2] In the article "Predicting creditworthiness in retail banking with limited scoring data" a team focuses on the banking sector in developing countries and applies logistic regression (LR), Classification and Regression Tree (CART) and Cascade Correlation Neural Network (CCNN).[3]

[1] https://www.raconteur.net/sponsored/home-credit-world-leading-consumer-lender-keeps-its-startup-spirit/

[2] https://www.sciencedirect.com/science/article/pii/S0950705116300156

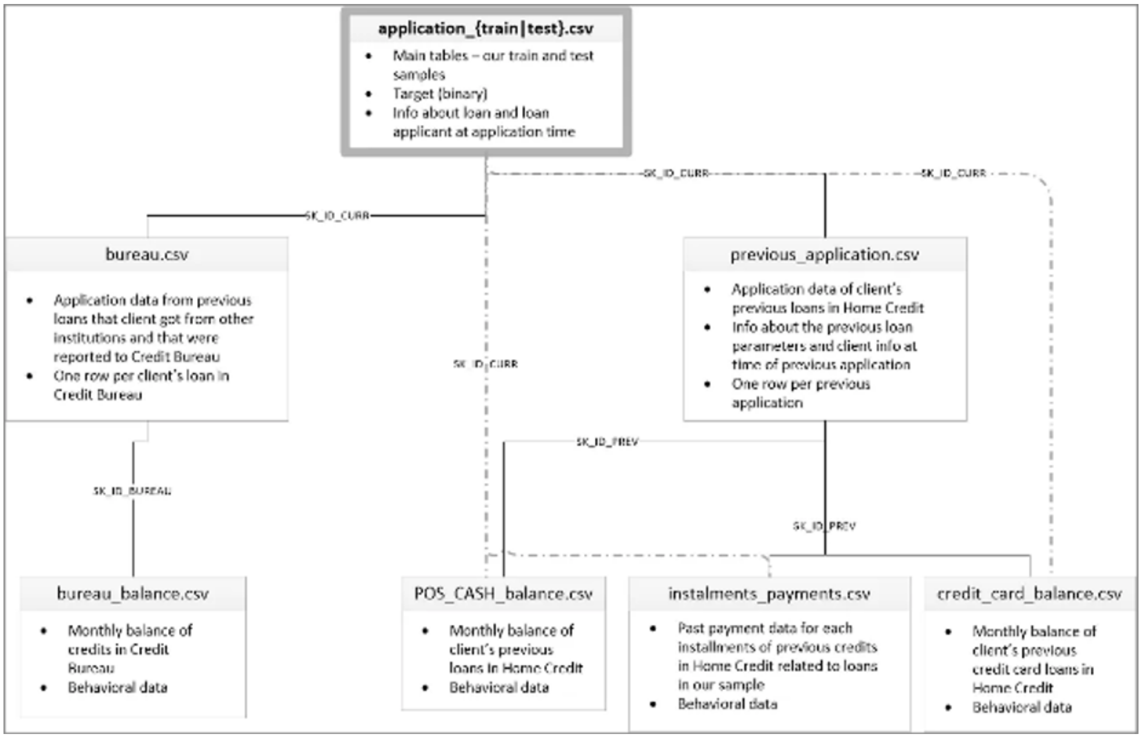
[3] https://www.sciencedirect.com/science/article/pii/S0950705116300156

2. Understanding of the Data and Exploratory Data Analysis

2.1 Understanding Data

We have a total of 7 data files altogether. The data is segregated into multiple text files that are related to each other such as in the case of a Relational Database. The datasets contain extensive features such as the type of loan, gender, occupation as well as income of the applicant, whether he/she owns a car or real estate, to name a few. It also consists of the past credit history of the applicant. A total of 221 columns are present in the 7 data files. One of the major challenges we faced is merging them into one single master train and test datasets. As we can see from the below figure, we have two main ID's that we can utilize to merge these dataframes. 'SK_ID_CURR'

We have a column called 'TARGET', which acts as the input that we take to train a classifier that makes the default predictions, and our problem at hand is a 'Binary Classification Problem', because given the Applicant's 'SK_ID_CURR' (present ID), our task is to predict 1 (if we think our applicant is a defaulter), and 0 (if we think our applicant is not a defaulter).


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
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
SK_ID_PREV
SK_ID_BUREAU
SK_ID_PREV
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

2.2 Data Cleaning & Merging

Given we had seven data files to merge together, we want our strategy to be simple to understand and retain as much information as possible. This is complicated by the many 1-to-N relationships that exist between the records in the multiple files. To achieve our goals, we added features that count how many of each type of record before merging just the latest record. This ensures our models are exposed to the applicant's history but is more focused on the applicant's most recent information.

After merging, we cleansed the data to prepare for model training and evaluation as follows

- We identified and dropped 68 features that contained over 50% missing values.
- We confirmed that we do not have any applications that are over 50% blank.
- We split our dataset into training and evaluation.
- We imputed our datasets with the mean, median, and mode for the continuous, discrete, and factor features respectively calculated from the training split. This step is necessary as logistic regression cannot handle missing values.
- We identified and dropped 3 unethical features to comply with the Equal Credit Opportunity Act.
- We identified and dropped 38 highly correlated features. This step is necessary as logistic regression cannot handle multicollinearity.

2.3 Exploratory Data Analysis

'Exploratory Data Analysis' is one of the most important and critical parts of the Machine Learning pipeline: Without an understanding of the data, we would not be able to make sense of the data, preprocess the data if needed, make a strategy to deal with missing values and outliers, and thus affecting our model prediction. Below are some observations we found out from our initial scoop through the data:

- We have only 8% of data labeled as loan defaulters. We need to counter this in modeling.
- Most of the loans given were less than a million USD. Around 30% took loans in the range of 25,000$. [Exhibit2]
- Single and Civil Marriage do show a higher default rate. Widow has at least minus two percentage points of default rate than the mean of our entire dataset (8%) [Exhibit1]
- The AMT_CREDIT is left-skewed and people who took large amounts of loans are likely to repay them as we have a very tiny proportion of defaulters above 2 million.
- Most of the loans are given to females 202K loans versus 105K loans for males. We can see some high default rates for men who are single and aged between 30-50 along with older men who are separated.
- Most people applying for loans are in the range of (35-40) years whereas this is followed by people in the range of (40-45) years. The number of applicants in people aged <25 or aged>65 is very low. People aged in the bucket (25-30) years and (30-35) years have a large chance of being deemed not capable of loan repayment. [Exhibit3]
- There was some considerate distinction between the two classes for these variables EXT_SORUCE_1, EXT_SORUCE_2, EXT_SORUCE_3. These three variables itself constitute to around 50% of correlation with the TARGET variable. [Exhibit4]

- Applicants with a Higher Value of Credit Amount across various income types have a Higher Likelihood of deemed capable of Loan Repayment, especially in the case of 'Unemployed', 'Student' and 'Businessmen'. [Exhibit5]
- Men & Women with Cash Loans have higher chances of being deemed capable of loan repayment based on their Credit Amount. [Exhibit6]

2.4 Feature Engineering

Apart from having the features from all the 8 datasets we did create additional features that might assist us in increasing the performance of the models. One can think of these ML-engineered features as combinations of the existing features (e.g., payment rate combined with geographic area for a mortgage loans portfolio) and highly predictive dummies signaling whether a risk driver is within an especially relevant domain. We did create lot of additional features for each dataset and aggregate them to ID level. To avoid the curse of dimensionality caused by too many features and identify the key factors in credit risk, we removed the irrelevant and redundant features by feature selection. In addition, there were 4 rows of the data where the gender was XNA and most of the data for these rows were missing so we removed these rows before performing any feature engineering. In addition, we also performed some special transformations for other columns such as the Date column. Few of the Date columns had value 365243 which is nearly equal to 1000 years and so we replaced them with Missing tag and with 0 in further stages of our analysis.

From the main dataset (i.e., application_train.csv) we can create different percentages and annual rates from the existing features already present.

- DAYS_EMPLOYED_PERC = ratio of DAYS_EMPLOYED and DAYS_BIRTH
- INCOME_CREDIT_PERC = ratio of AMT_INCOME_TOTAL and AMT_CREDIT which reflects the firm's past evaluation on applicant ability to pay back.
- INCOME_PER_PERSON = ratio of AMT_INCOME_TOTAL and CNT_FAM_MEMBERS which reflects past household ability to pay back the loans.
- ANNUITY_INCOME_PERC = ratio of AMT_ANNUITY and AMT_INCOME_TOTAL
- PAYMENT_RATE = ratio of AMT_ANNUITY and AMT_CREDIT

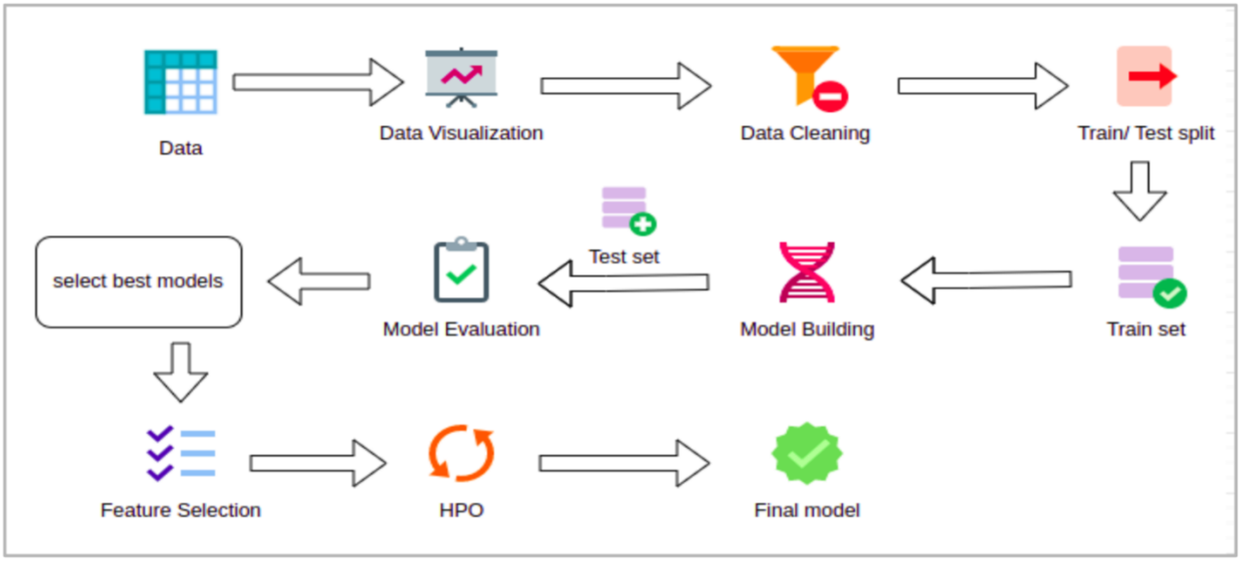
For other datasets we did also create some statistical features like Count, Max and Sum based on second level ID's - 'SK_ID_BUREAU' and 'SK_ID_PREV' as we wanted to check how many applications did the applicant have previously and what were their characteristics. 'Mean' was one of the measures we used extensively for all the numeric and continuous variables in our dataset as it conveys on average what was the amount the applicant has requested for in their previous applications and what amount was sanctioned as loan etc. This measure was quite helpful for our 'credit_card_balance.csv' dataset as we could get the average spend of an applicant and their main usage.

Even though R accepts factor variables we did use one-hot encoding for some categorical variables that have less no of categories within them. For others, we selected the categories that turned out to be useful from our model feature importance's. Null values can be handled by XGBoost so we left them as it is for the said model and replaced the numeric with mean and categorical null values with mode during the execution of other models.

3. Methodology and Modelling

3.1 Methodology

The final goal of this project is to automate the loan approval prediction process completely. The dataset needs to be analyzed comprehensively. The approach we followed for this project is shown by the diagram below. Model based feature selection will be used from the best models that were chosen from the initial model evaluation process.
The process of selecting a final machine learning model from among a group of candidate machine learning models for a particular training dataset of Loan customer is called model selection.



3.2 Modeling

One of the first studies to apply machine learning techniques in credit risk was from the article "Machine-learning algorithms for credit-card applications" [4]. In the article, the authors tested a series of algorithms for assessing credit default risk, integrating two models: (1) a general computational model based on a selection process and a pairing procedure, and (2) an artificial neural network (ANN) connective model. Although the results are limited by the small number of observations of the database and the characteristics of the techniques tested, the study supports the relevance of the

[4] https://calendar.app.google/yf38BUNDZhBRCLQp9

use of machine learning tools for credit analysis. Other studies focused on using SVM, KNN and RF (random forest) models for identifying credit risk applications.

Bagging (Bootstrap Aggregating), proposed by Leo Breiman is based on bootstrap samples that aggregate or combine individual predictors to establish a better final predictor[5]. The author verified the variance of the combined predictor is lesser or equal to the variance of any other individual predictor used. Another paper that showed the superiority of the ensemble classifiers was "A comparative assessment of ensemble learning for credit scoring". [6]

So, we decided to use Logistic Regression (LR), a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set as our baseline model. Decision Trees follow the structure of an upside-down tree, dividing data into branches. The model comprises a series of logical decisions, like a flowchart, with nodes indicating a decision to be made on an attribute. The branches reflect the choice of the decisions. XGBoost (XGB) as it is built using trees, handles sparse data, and learns from multiple weak learners. LightGBM (LGB), another variant of boosting tree, is also used. Support vector machines (SVM) which aims to create a hyperplane that could lead to partitions of data on groups reasonably homogeneous. Once the models are executed on the train data (215,258 samples) we compare the results of the test data (92,553 samples) and select the top k-features that contribute most to the predictability of the credit risk. The data split was done strategically meaning the ratio of neg to pos would be the same in both train and test sets.

4. Results

4.1 Performance metrics

We use standard metrics to analyze the performance of the credit classification models. Major metric we concentrated on was the AUC (Area under the ROC Curve -ROC [Receiver Operating Characteristic]) measurement which provides a precision criterion for the validation set, to compare results from the models. AUC is given by:

$$AUC = \frac{1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN}}{2}$$

4.2 Results

All models were developed in R and tested on the same sets of samples. We perform two types of experiments: (1) Used a cleaned version of the data to predict applicants' credit risk. In this experiment, no additional features apart from raw columns were explored. (2) Added a few new features that we developed based on domain knowledge. Based on Figure2 we can see how important feature engineering was in improving the model.

[5] https://academic.oup.com/imaman/article-abstract/4/1/43/656001

[6] https://www.sciencedirect.com/science/article/abs/pii/S095741741000552X

The performance of our models in the two cases indicated above is discussed in the following lines. When compared to the other models, XGB fared the best in the first scenario, with an AUC score of 0.756, [Figure1] reinforcing our belief in the ensemble method's superiority. Our baseline model LR performed better as well, with an AUC score of 0.731. [Exhibit7] shows the feature importance scores for the XGB model. It is apparent that EXT_SOURCE_3, EXT_SOURCE_2 contribute to more than half of the importance. Additional variables considered crucial in estimating overall prediction include AMT_GOODS_PRICE, DAYS_EMPLOYED, and INSTALLMENT_AMT_PAYMENT.

One intriguing characteristic is that, while EDUCATION_TYPE and OCCUPATION_TYPE are significant features, not all categories within these variables are significant. The logistic regression model's output assists in finding the essential categories from these variables. According to the LR model summary, Accountants, Core Staff, and medical category profession categories are adversely connected with credit risk, whereas Drivers and Low Skill Laborer's are positively correlated with credit risk. The same is true for those with a Lower Secondary education who pose a credit risk. Interestingly, having an automobile turned out to be a significant variable in model performance. Overall, the XGB outperformed our baseline LR by 3.5%.

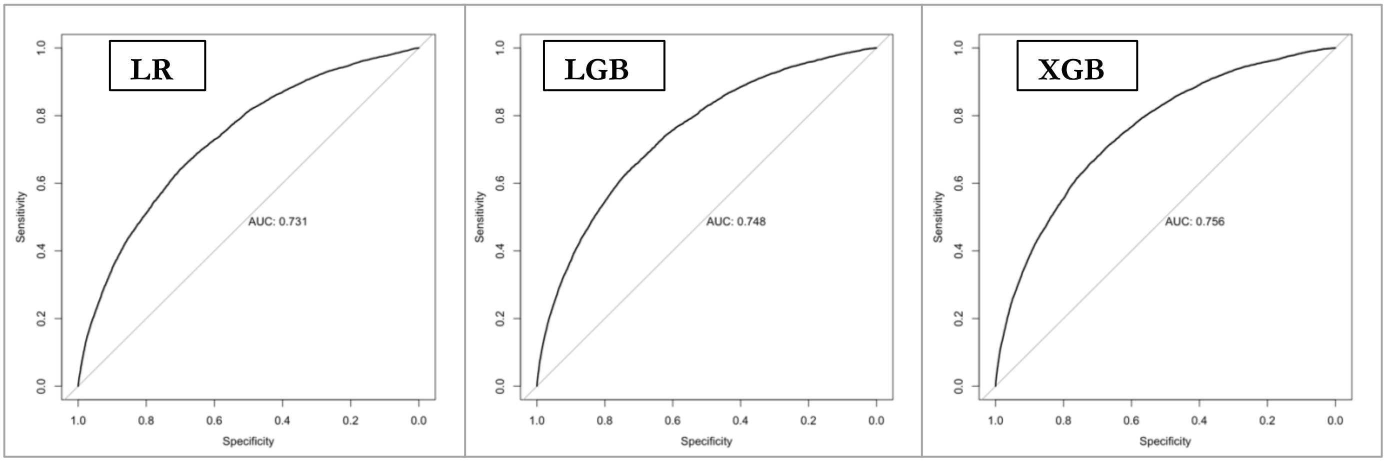


Figure1

In the second experiment, feature engineering improved our AUC score by 2.5%. XGB, as in the prior experiment, performed best with an AUC of 0.771 [Figure2]. PAYMNET_RATE proved to be a differentiator, reflecting how long the applicant will take to repay the loan. DPD_MEAN, CNT_DRAWING_ATM_CURRENT, CNT_PAYMENT_MEAN, are all variables that demonstrate how important it is to incorporate the applicant's past credit line history when modelling. These variables discuss the length of previous credit, late payments in past credit, and the number of draws from credit, all of which appear to be significant based on domain expertise. LR with AUC score of 0.756 and LGB (LightGBM) with AUC score of 0.764 also outperformed the prior experiment and indicates that tree-based models have good credit risk modeling abilities. The XGB feature importance plot is shown in [Exhibit8]

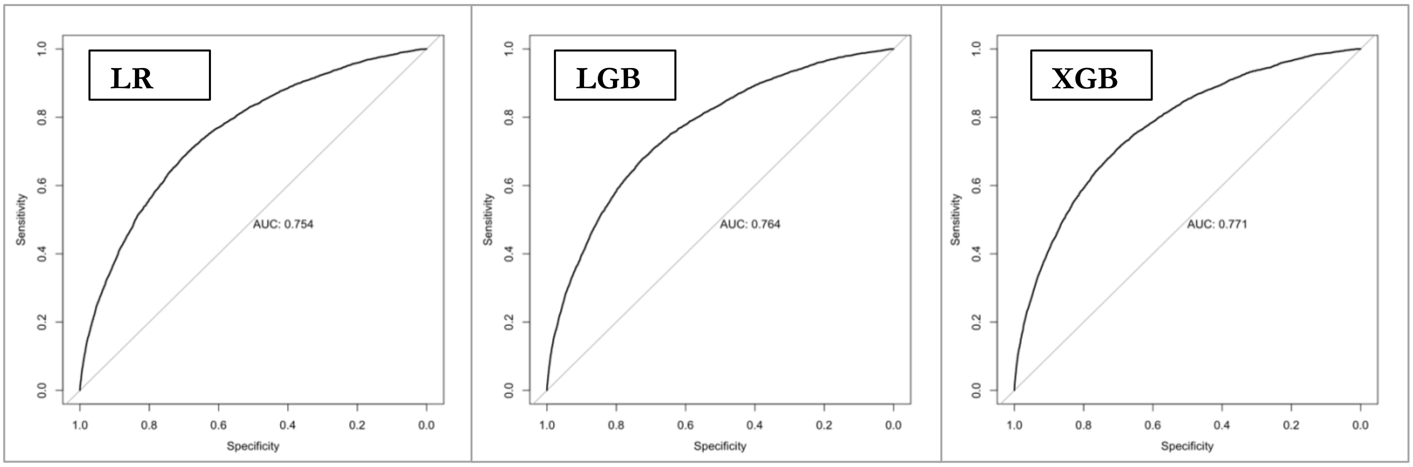


Figure2

4.3 Interpretability

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. We use SHAP (SHapley Additive exPlanations) which assigns each feature an importance value for a particular prediction. Rather than use a typical feature importance bar chart, we use a density scatter plot (summary plot in Exhibit9) of SHAP values for each feature to identify how much impact each feature has on the model output for individuals in the test dataset. Features are sorted by the sum of the SHAP value magnitudes across all samples. It is interesting to note that the EXT_SOURCE_1 feature has more total model impact than the PAYMENT_RATE feature, but feature importance from xgboost said otherwise. Clearly from the summary plot [Exhibit9] we can see how small values for EXT_* variables have high positive shap values indicating that applicants with low EXT_* values pose high risk.

SHAP dependence plots show the effect of a single feature across the whole dataset. They plot a feature's value vs. the SHAP value of that feature across many samples. SHAP dependence plots are like partial dependence plots, but account for the interaction effects present in the features and are only defined in regions of the input space supported by data. Dependency plots for the top four variables are shown in the visuals section [Exhibit10]

4.4 Graphics and Visuals:

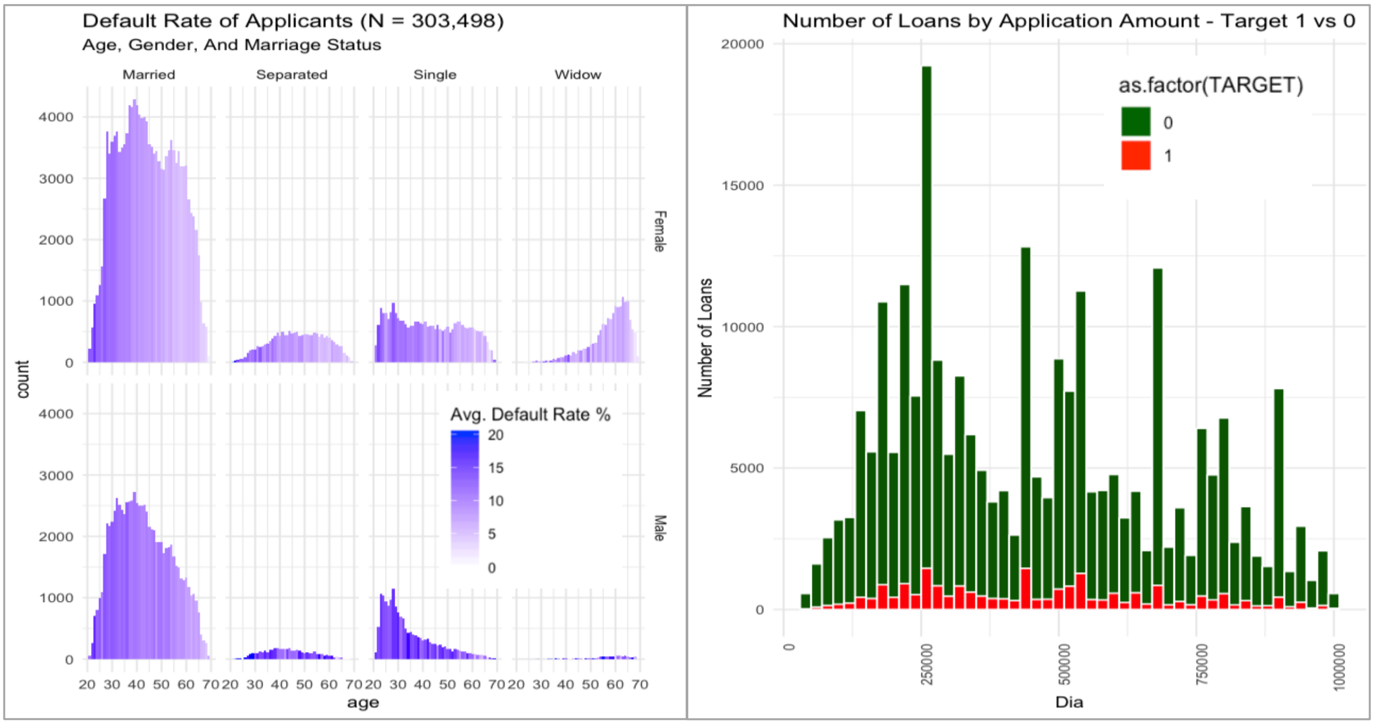


Exhibit1 & Exhibit2

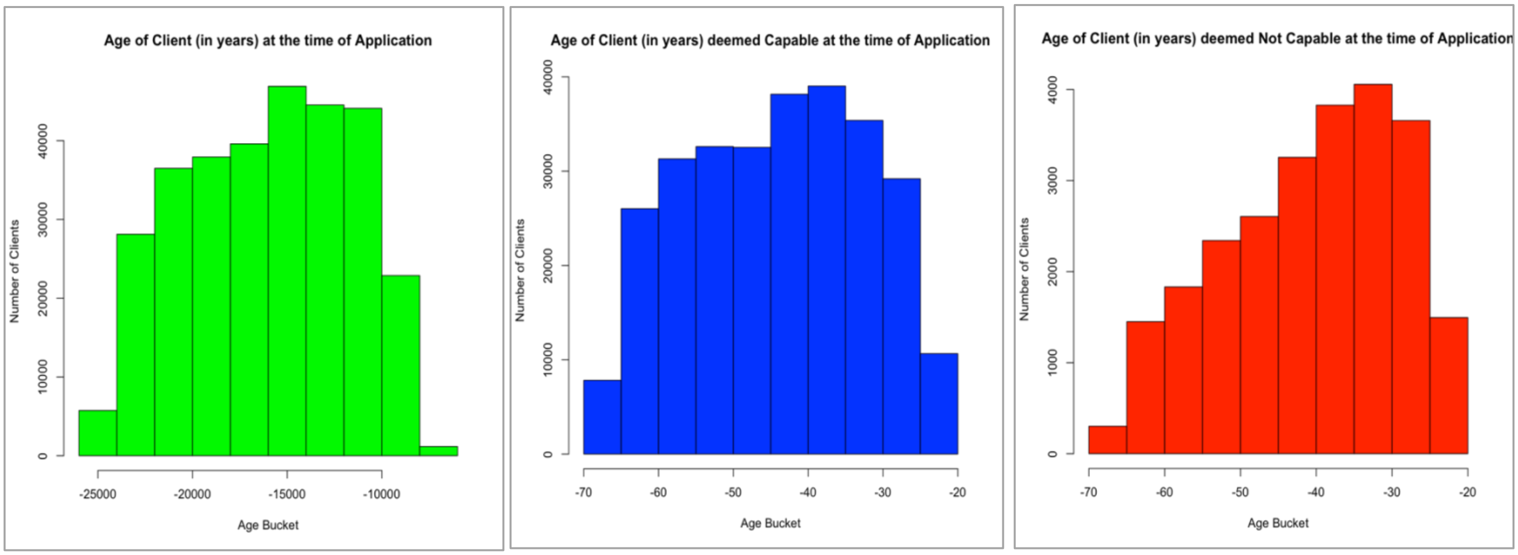


Exhibit 3

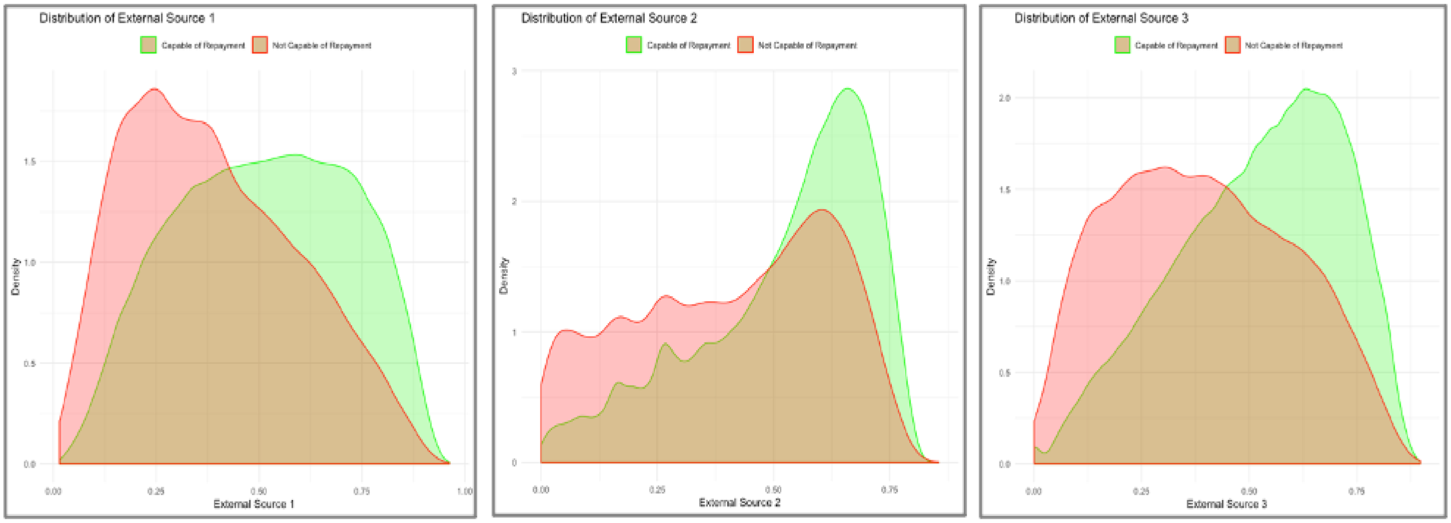


Exhibit 4

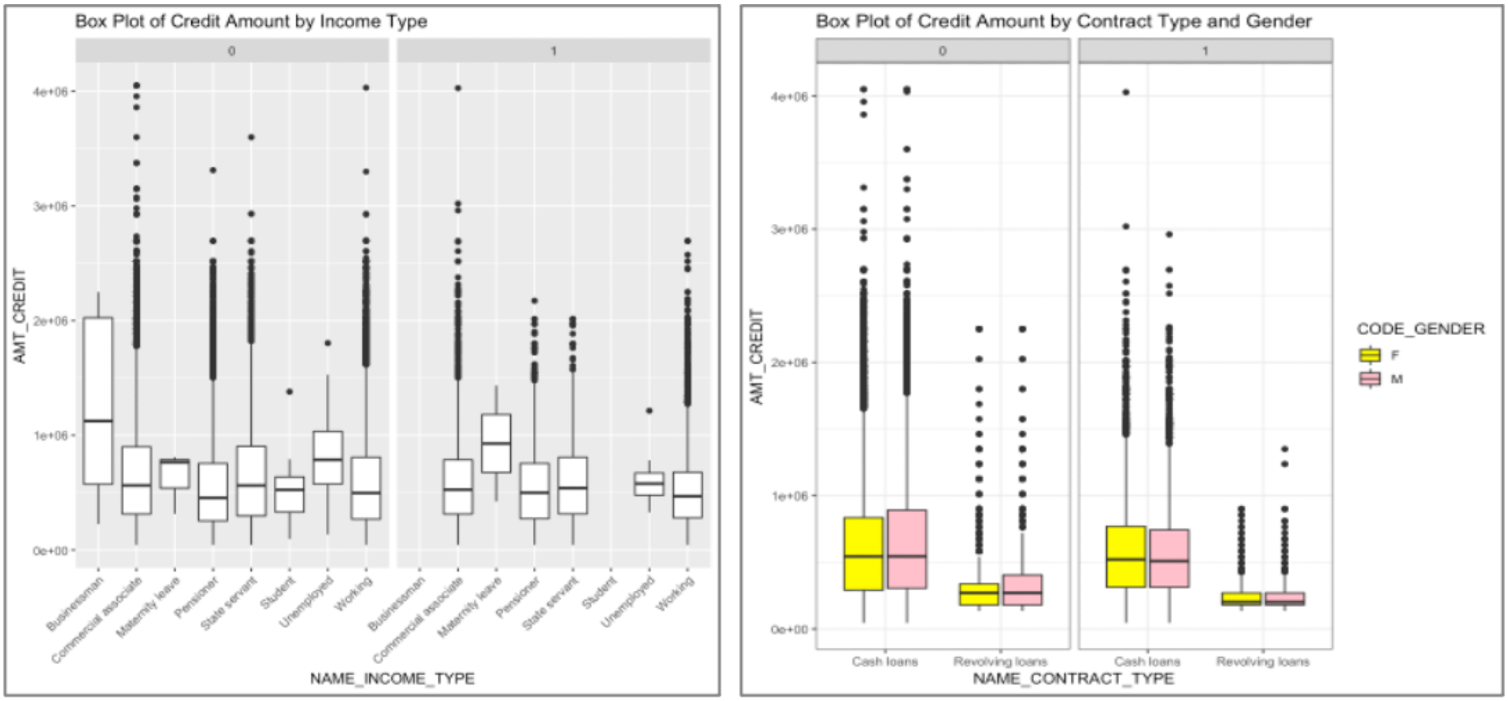


Exhibit5 & Exhibit6

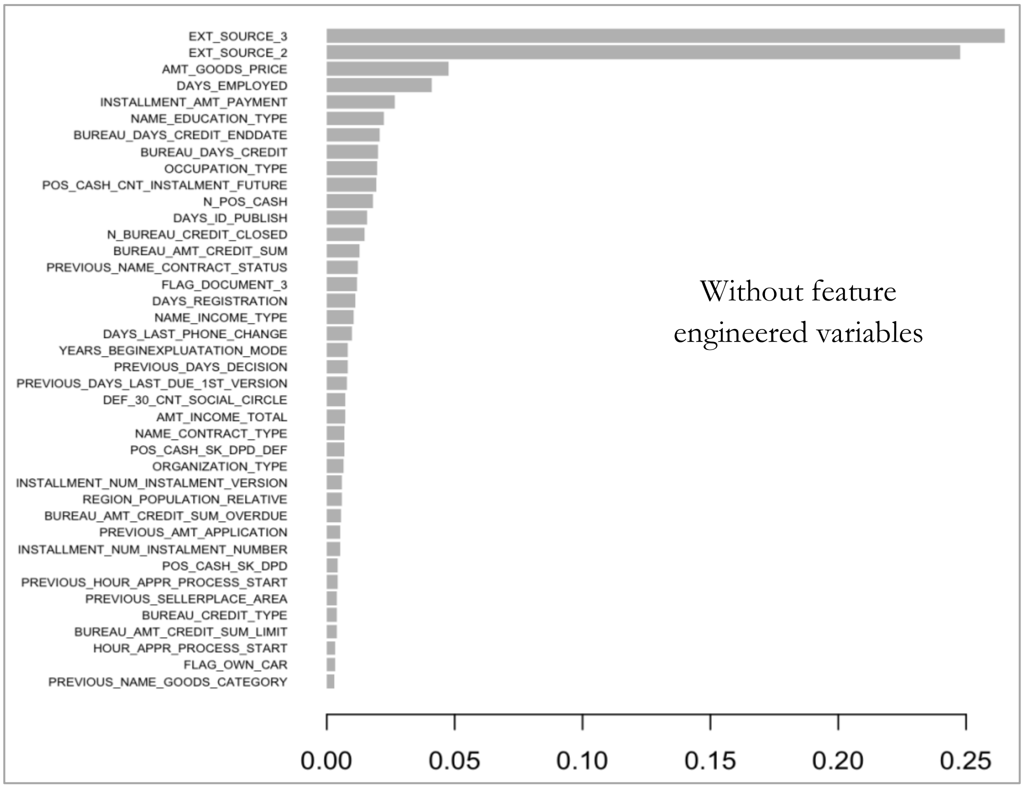


Exhibit7

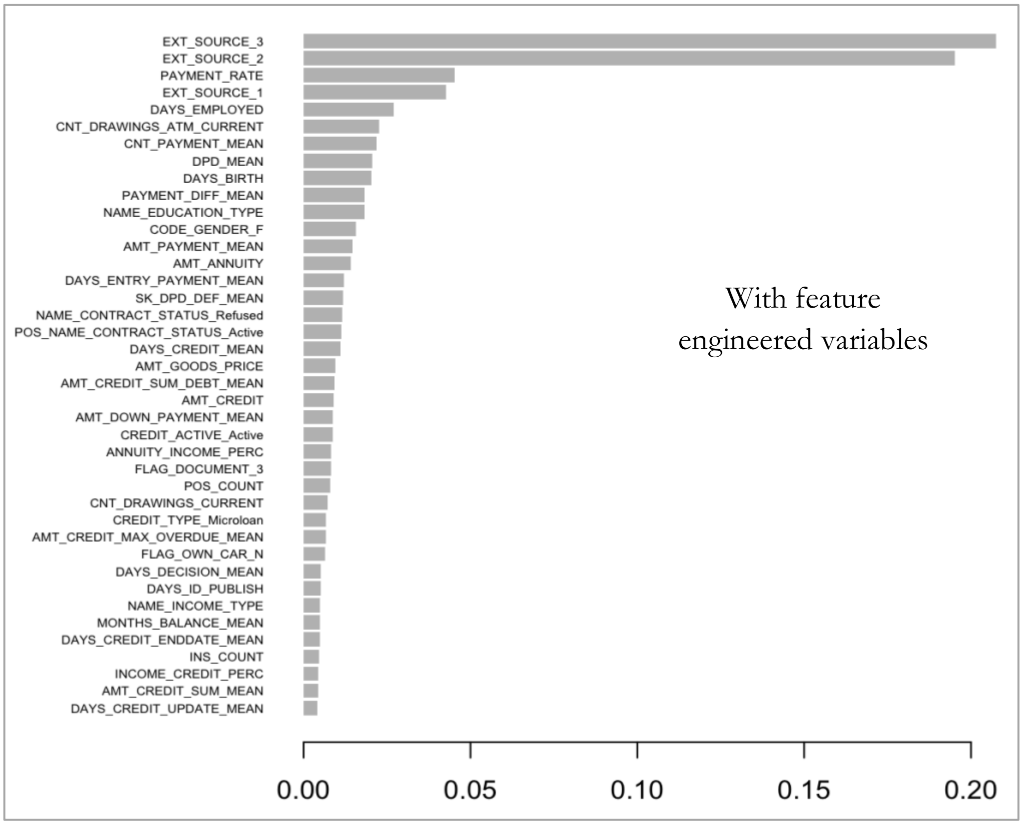
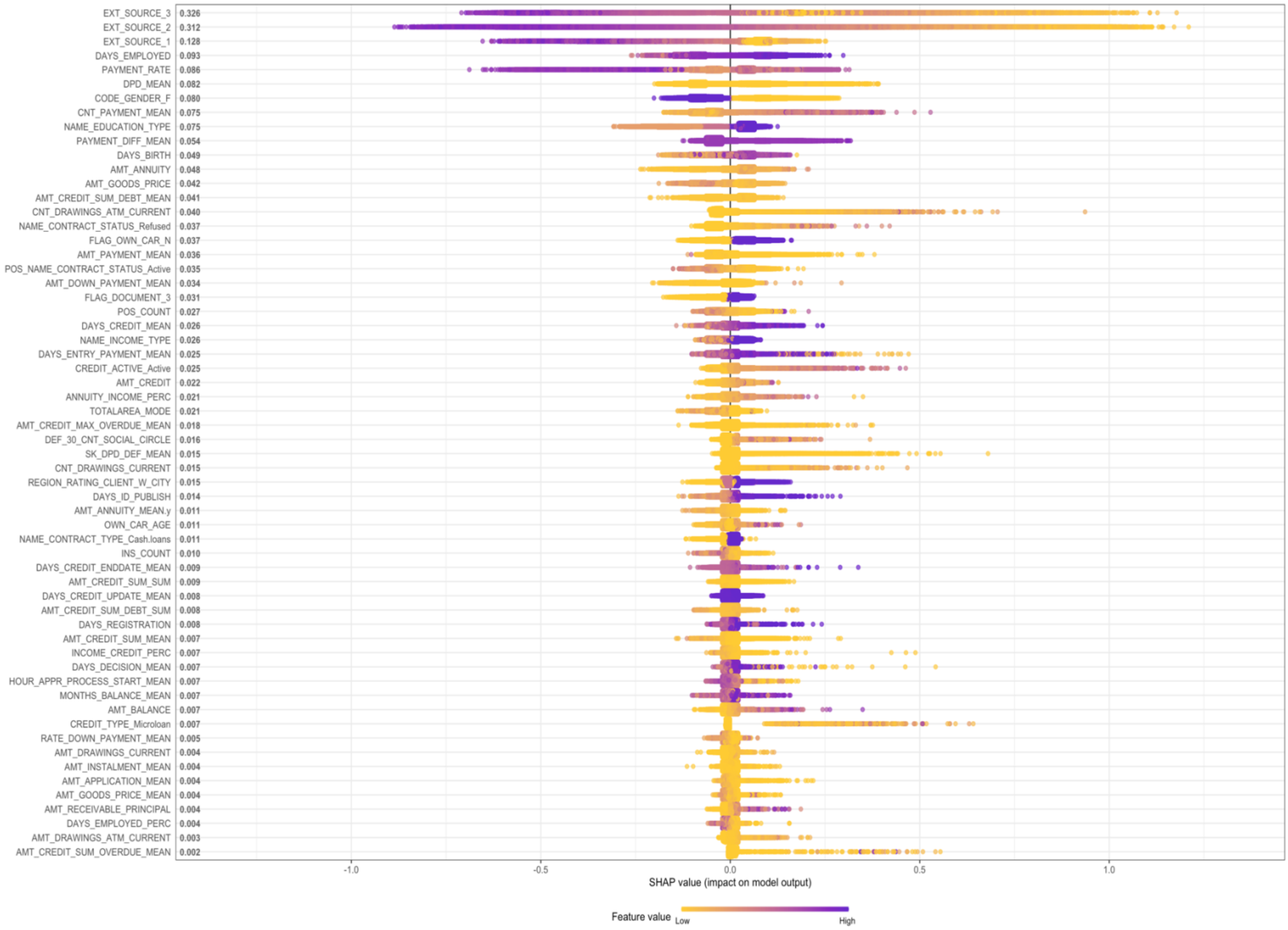
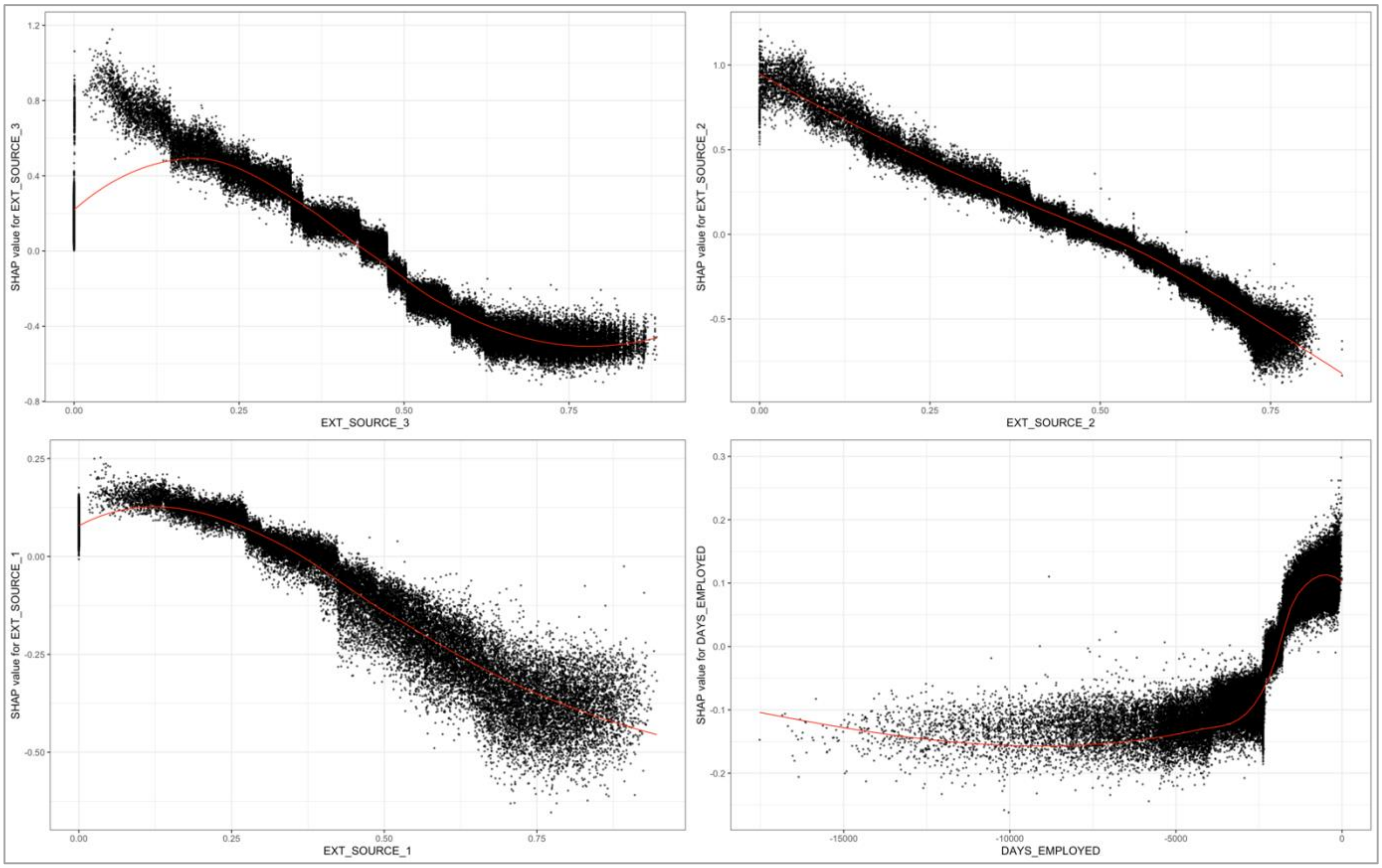


Exhibit8



Summary Plot – Exhibit9

Dependency Plots – Exhibit10


SHAP value for EXT_SOURCE_3
EXT_SOURCE_3
SHAP value for EXT_SOURCE_2
EXT_SOURCE_2
SHAP value for EXT_SOURCE_1
EXT_SOURCE_1
SHAP value for DAYS_EMPLOYED
DAYS_EMPLOYED