GROUP 8

HOME CREDIT DEFAULT RISK

PRESENTATION

Jade Gosar, Karson Eilers, Paula Stefani

AUGUST 2023

Business Problem

- Insufficient or non-existent credit histories and variance among people's financial historical data.
- Users have difficulty in verifying and providing proof of credit accountability when getting loans
- Lenders have issues in identifying and predict which customers are at risk of default, possibly leading to mis-identifications and reduction of possible future customers that would be in fact reliable borrowers.

Purpose

The purpose of the proposed project is to create a supervised analytical model to help Home Credit predict how capable each applicant is of repaying a possible loan by identifying possible defaulters and non-defaulters.

Approach

01 DETERMINATION OF VARIABLES THAT ARE MORE IMPORTANT WHEN DETERMINING THE PREDICTION OF THE REPAYMENT ABILITY OF ENTERPRISE LOANS.

02 USE THE DIFFERENT FEATURES FOR THE CURRENT APPLICATION VS. PREVIOUS APPLICATIONS TO CREATE SUPERVISED ANALYSIS

03 TESTING OF MULTIPLE MODELS THAT CAN BE USED TO IDENTIFY DEFAULTERS AND NON-DEFAULTERS AND SUPPORT FUTURE ANALYSIS

04 FOCUS ON CAPTURING THE CLIENTS THAT TRADITIONALLY WOULD NOT HAVE ACCESS TO CREDIT

Data

DATASETS AND PROCESSES FOR EDA AND MODELING

Data Selection

Selection of Datasets:
clean_training_data.csv
clean_testing_data.csv

Data Preparation

Performing Data cleaning and removing NA's

Data Partition

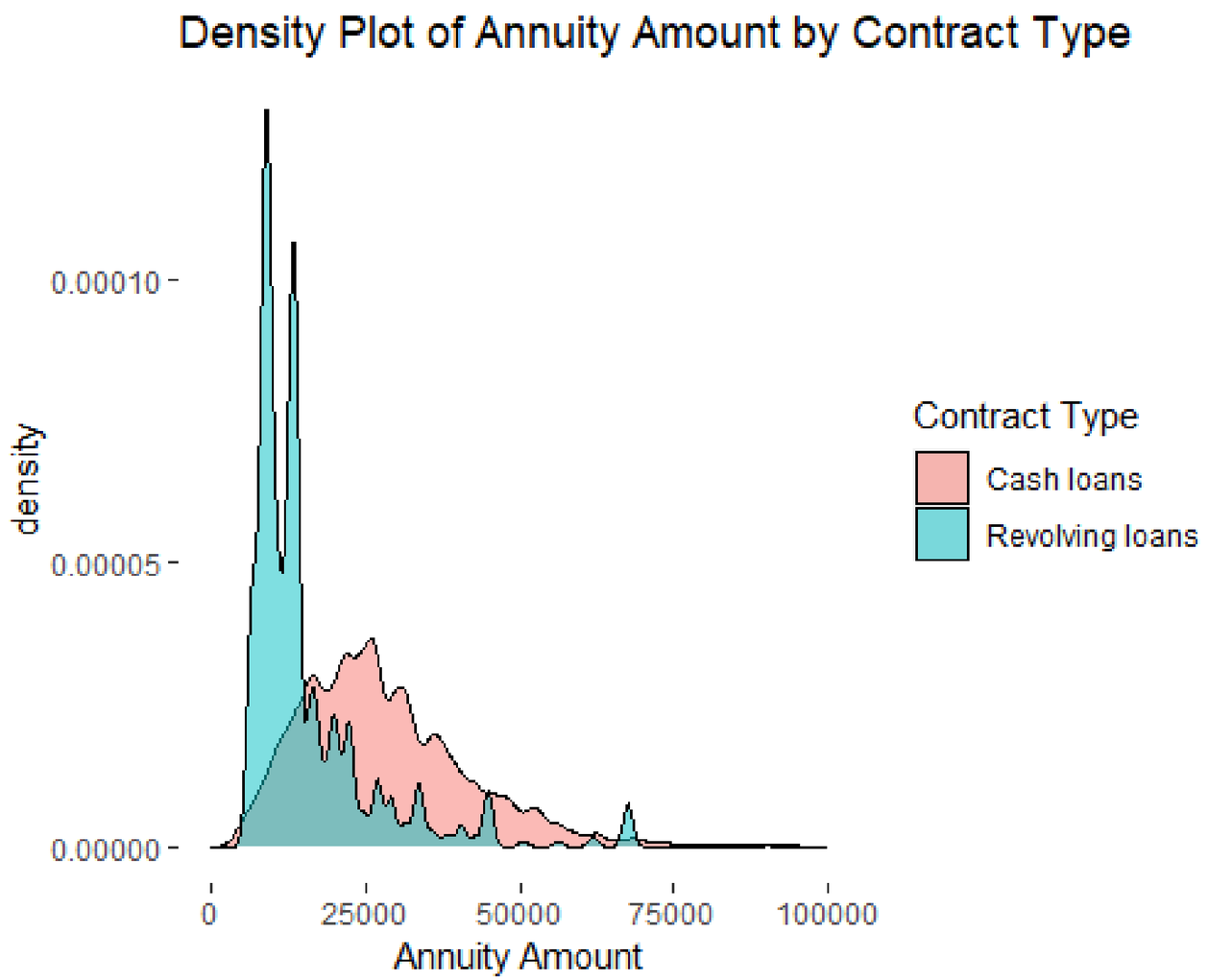
Creation of training subset of the training data with 70% of the data

GROUP 8

EDA

The group set a tolerance threshold of a 10% change in the TARGET variable

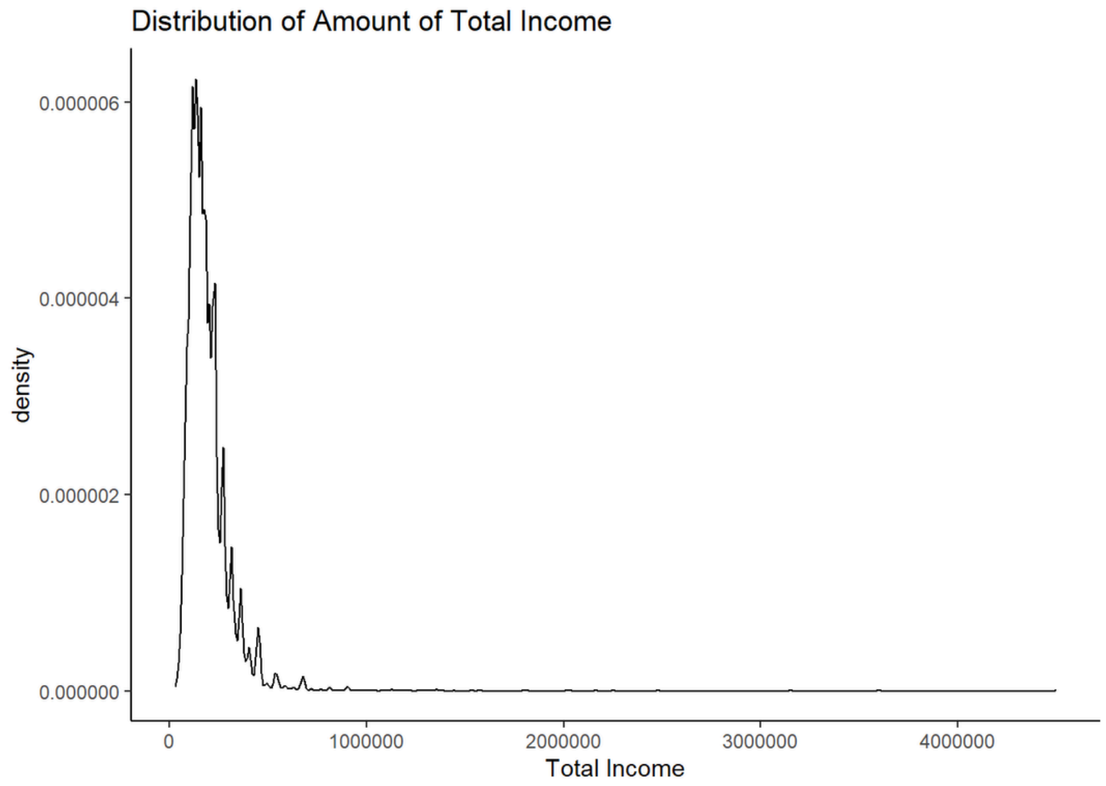
The cleaning methods used resulted in a 4.24% reduction in the TARGET variable



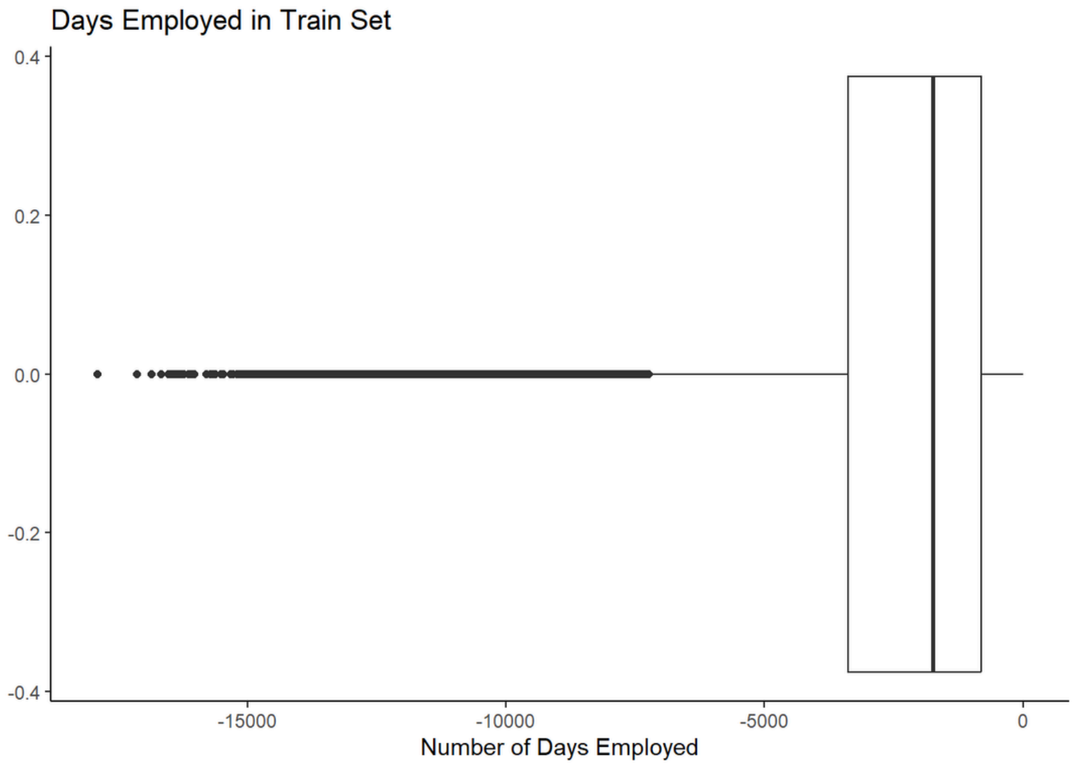
GROUP 8

EDA

INCOME



DAYS EMPLOYED



Modeling

Logistic Regression

The approach for the Logistic Regression Model was to increase model complexity and amount of variables while checking model accuracy and predictive efficiency.

Naive Bayes Classification

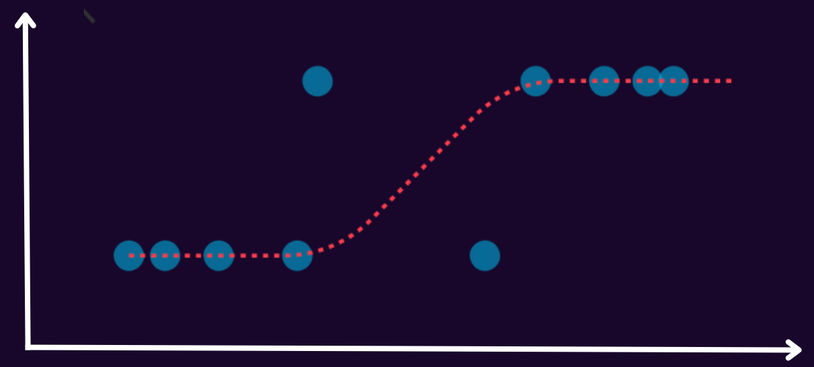
Three Naive Bayes models were created for better model comparisons. First two models use the existing class imbalance and the third will up-sample the target = 1 class.

Decision Tree & Random Forest

Our last models are a creation of a simple tree model to see if we could gain any initial insight, followed by random forest models with increased complexity and cross-validation.

GROUP 8

Logistic Regression



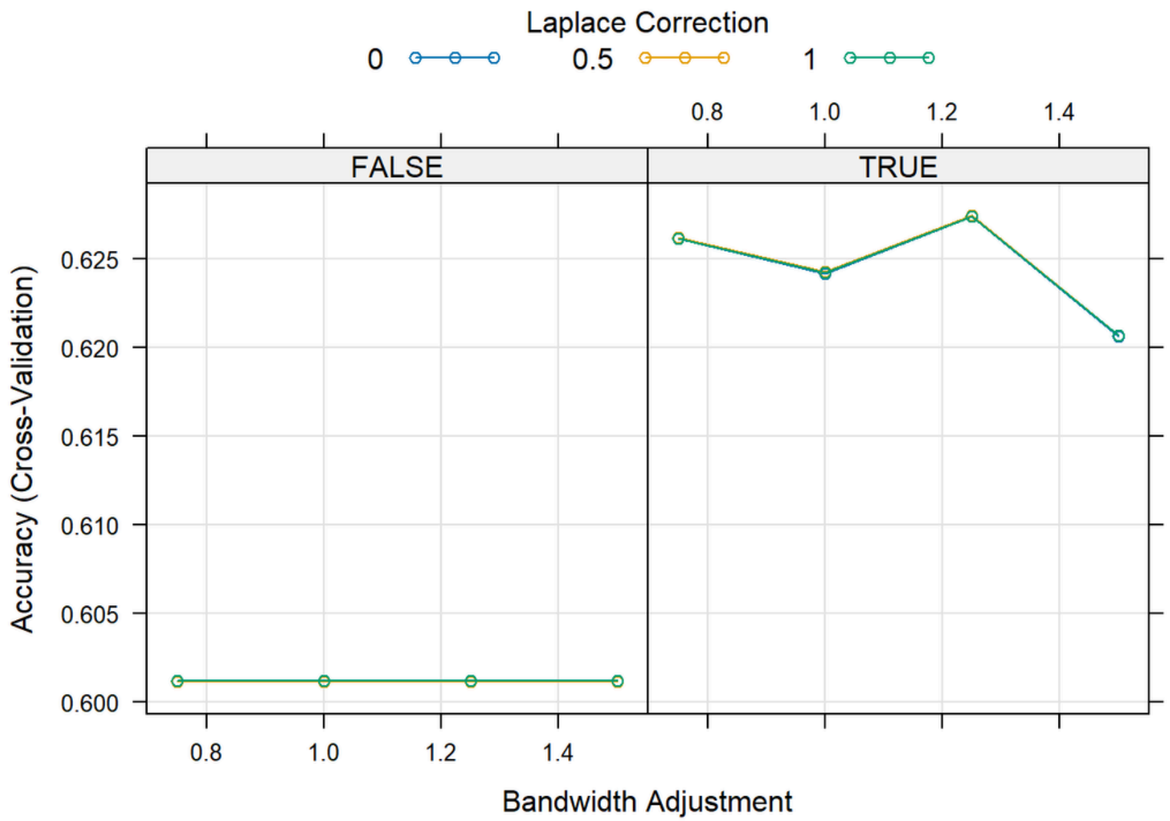
Strong predictors of Target:

AMT_INCOME_TOTAL + AMT_CREDIT + AMT_ANNUITY + DAYS_EMPLOYED + DAYS_CREDIT + NAME_INCOME_TYPE + NAME_EDUCATION_TYPE + NAME_CONTRACT_TYPE + CODE_GENDER + AMT_GOODS_PRICE + DAYS_ID_PUBLISH + AMT_REQ_CREDIT_BUREAU_YEAR + DAYS_LAST_PHONE_CHANGE + FLAG_WORK_PHONE + DAYS_BIRTH.

Results: We used AIC and Deviance as our main performance metrics. Both AIC and Deviance have improved from Model 1 to Model 2 and from Model 2 to Model 3, with a total change of 12700.6 less AIC and 834.5 less Deviance from model 1 to 3.

GROUP 8

Naive Bayes Classification



NB Model 1

01 Creation of baseline NB Model Employed a Laplace estimator of 1.

NB Model 2

02 Use the naive Bayes method in caret with cross-validation and several different tuning approaches.

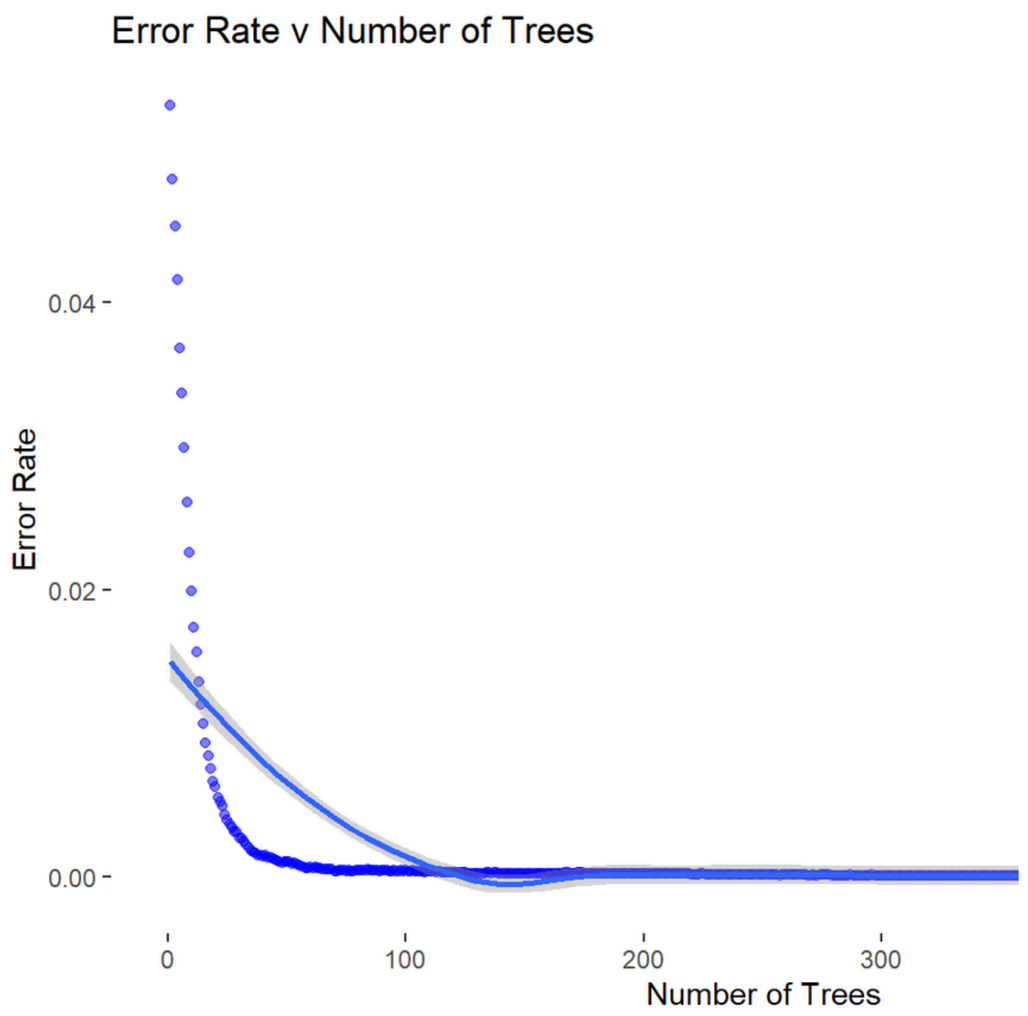
NB Model 3

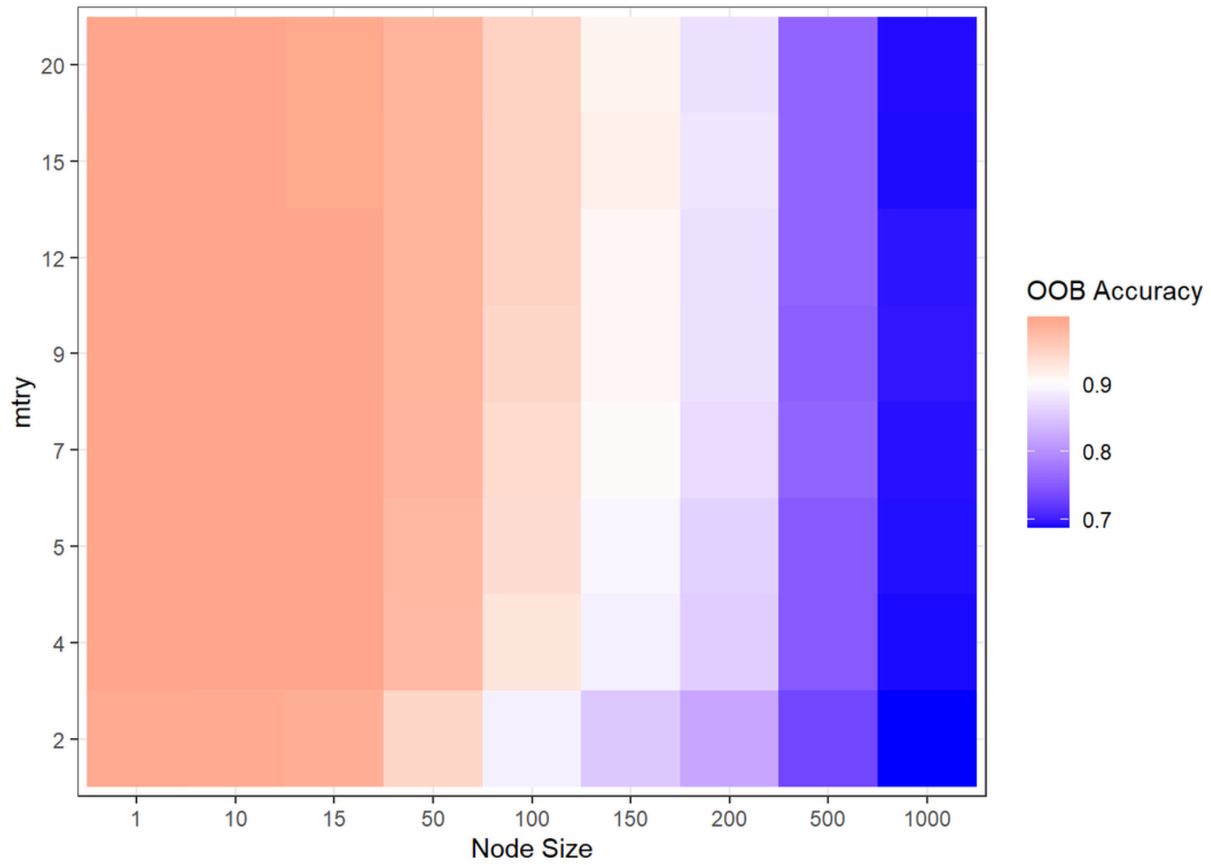
03 This final NB model will apply the parameters of model 2 to an upsampled training dataset.

GROUP 8

Random Forest & Decision Tree

To determine what the optimal number of trees to set the 'ntree' parameter of the Random Forest model is, we plotted the out-of-bag error rate against the number of trees. Then we tuned random forest hyperparameters and used cross-validation to estimate the test error rate.





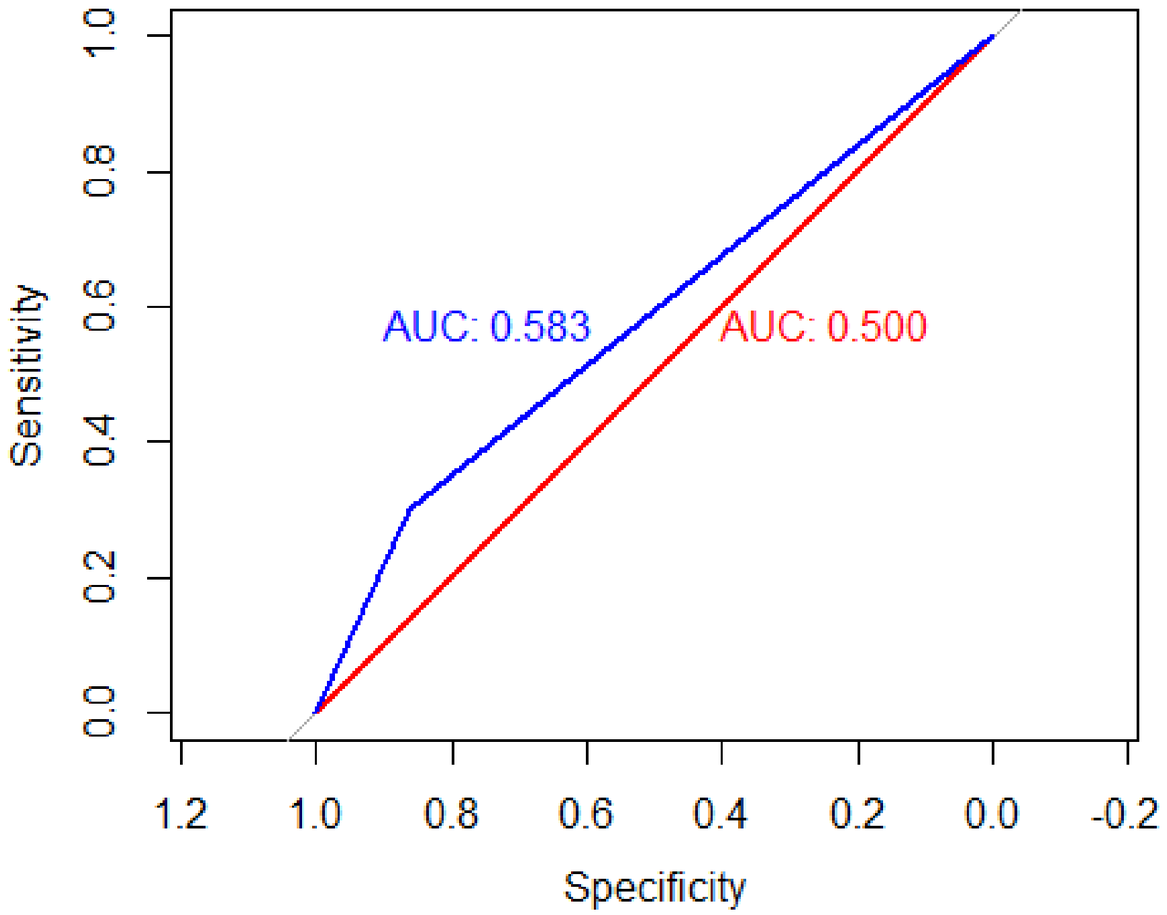
Heat Map

The heatmap depicts different combinations of two hyperparameters that we are tuning for our model: mtry and node size. When the node size value is low and the mtry value is high, for example when node size = 1 and mtry = 12, is when we see our largest OOB accuracy.

RESULTS

Our tuned Random Forest models performed the best of all the models that we tried; logistic regression, Naive Bayes and Decision Tree or default Random Forest model.

Findings

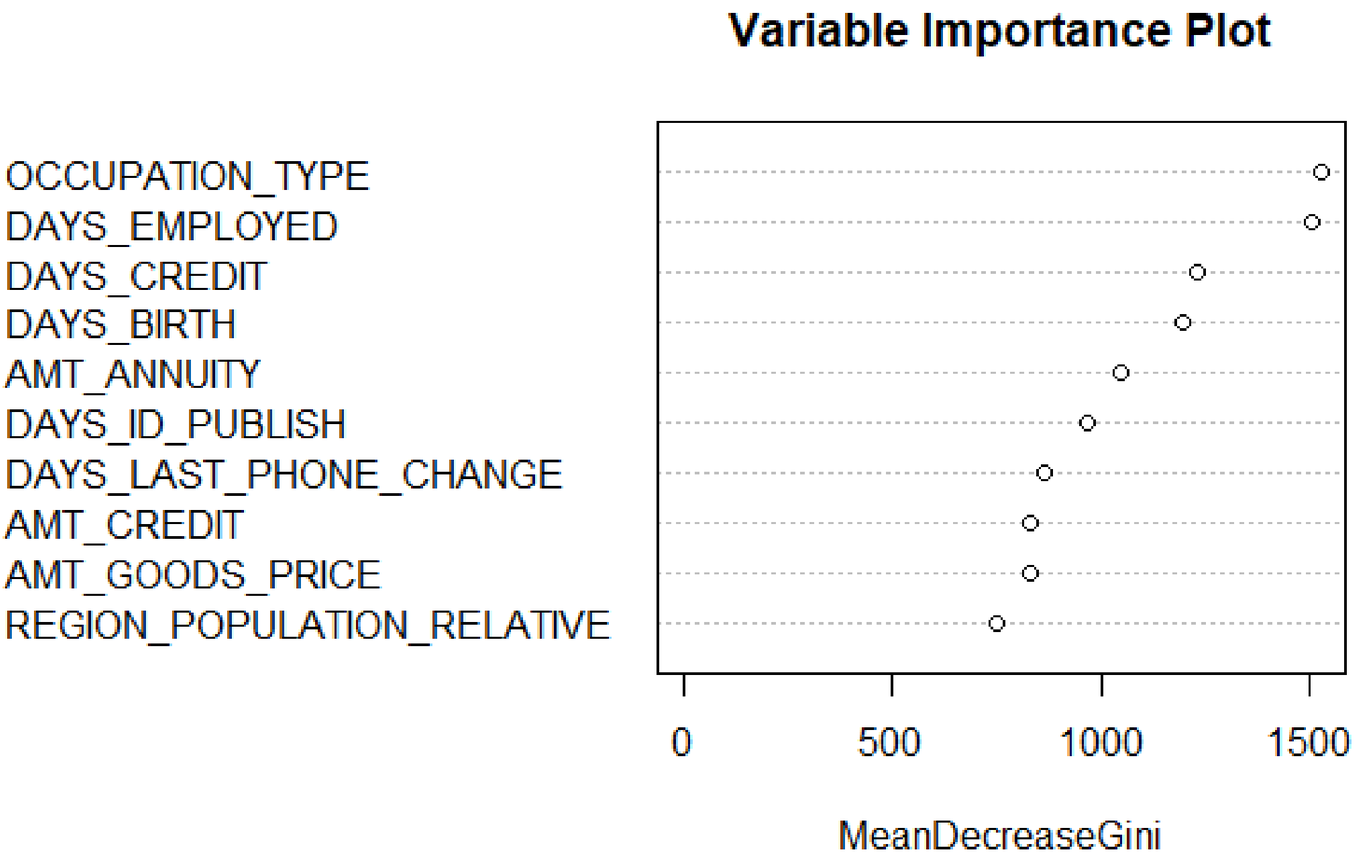


Final Peformance

Graph shows the differences in Accuracy between Model 1 - Always Predicting Majority Class, and Model 2 - Tuned Random Forest

Most Important Features

The graph shows the most important features identified in our model analysis.

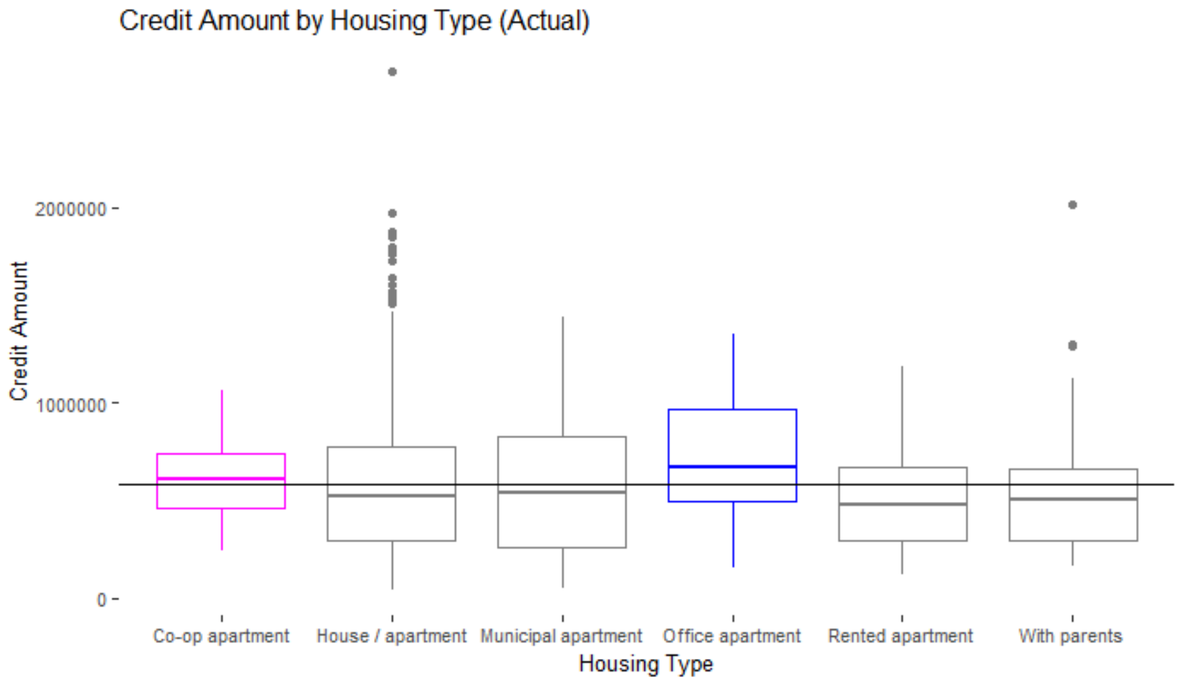


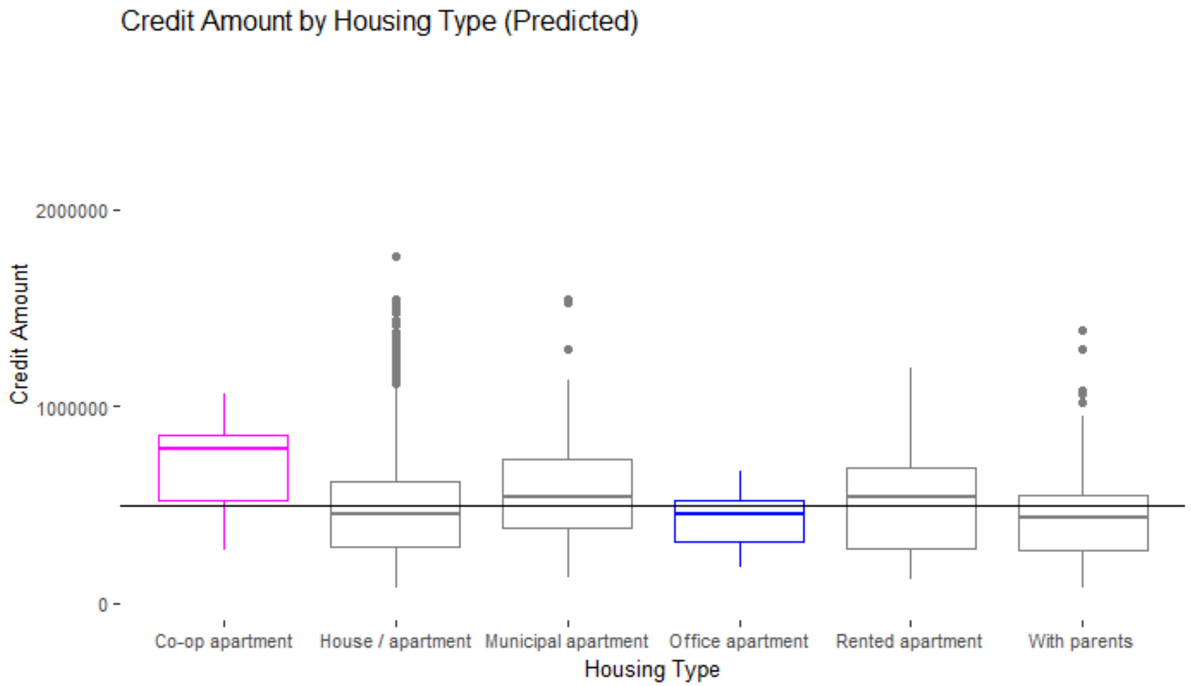
GROUP 8

Recommendations

01 Dedicating resources to contextualized data that is more reflective of underserved populations.

02 Include Currency ID as a factor.





GROUP 8

Business Application

$505,160,164

Estimated greater ROI of random forest model over majority classifier performance* in normalized currency value.

***Assumes simple interest rate of 18.912% (based on median prior median rate) and principal is fully recovered**

IS 6812 - GROUP 8

THANK YOU