Home Credit Default Risk

Eric Huynh

Data Capstone 2

Problem & Overview:

- Countless numbers of loan/credit applications are refused each day by financial institutions
  - Due to monetary risk of credit defaults by clients
- Financial institutions are stringent on qualifications to request large loans.
  - Limits applicants to only clients that already have a strong credit record.
  - Forces clients with little to non-existent credit history to rely on unreliable lenders to meet their financial needs.

Solution and Objective

- Certain financial institutions have strived to expand the financial inclusion to those with insufficient or non-existing credit histories.
  - Utilize a variety of alternate data to gauge and measure the clients’ risk profiles without relying on the conventional methods.
- Objective:
  - Draw reliable variables from the alternate data that can reliably gauge a clients’ risk profile.
  - Develop a predictive model that can accurately classify clients as high-risk or low-risk for credit loans

Potential Clients

- Financial Institutions:
  - Expand their loans applicant pool by measuring new clients' risk profile with alternate data.
  - Higher yields both short-term and long-term as positive financial experience will bring in reoccurring clients.
- Loan Applicants:
  - Understand alternative attributes that can help them develop towards becoming a low-risk applicant despite insufficient credit histories.
  - Also applicable for anyone applying for a larger loan they may not currently qualify for under traditional requirements.

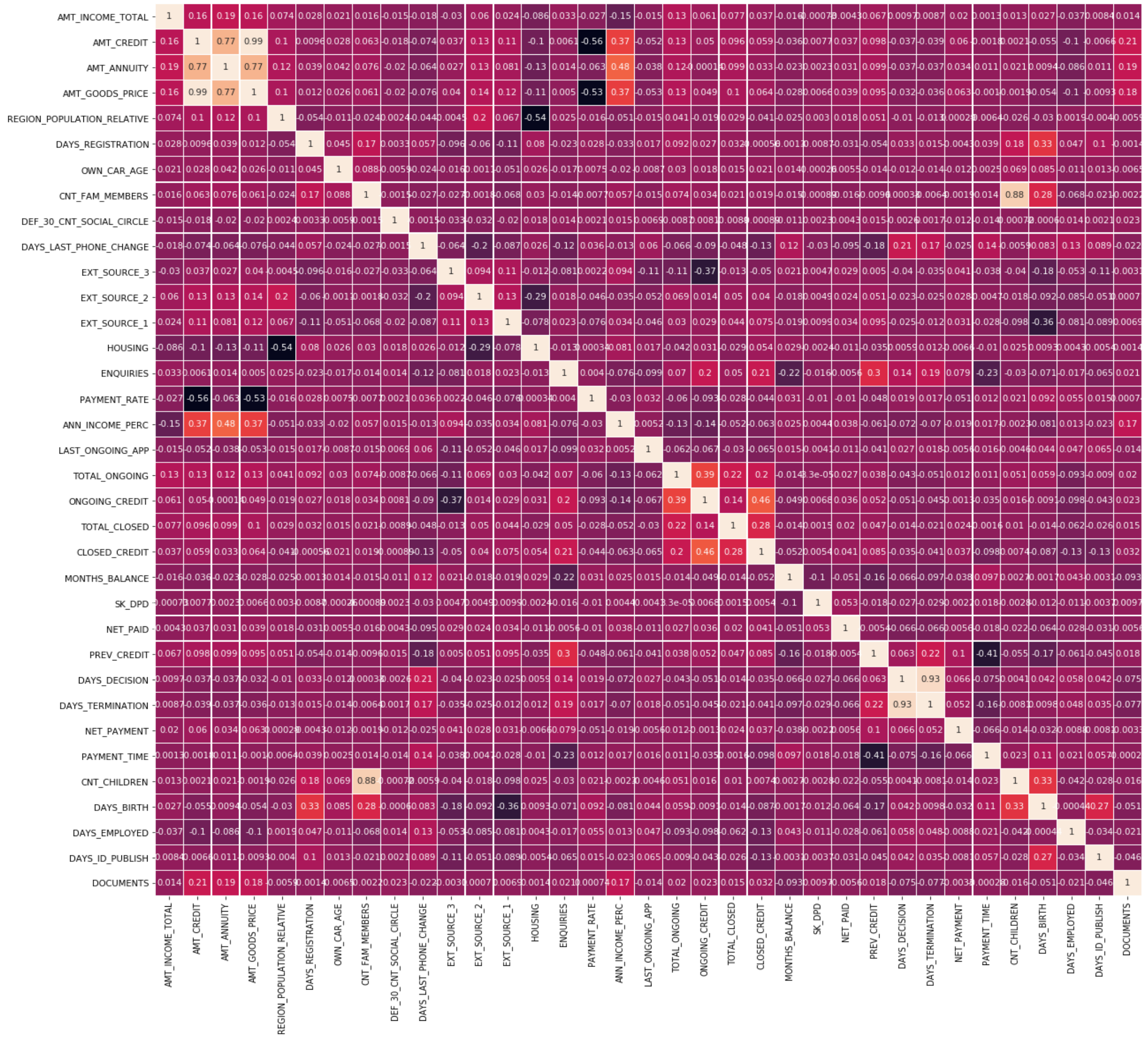
Data from Kaggle Competition

- Training Data:
  - Consists of 307,511 unique client IDs and 122 categories.
- 6 Supplemental Datasets (only 4 used):
  - “bureau.csv” – information regarding all previous credits
  - “credit_card_balance.csv” – previous credit information with Home Credit
  - “installments_payments.csv” – repayment history for approved credit by Home Credit
  - “previous_application.csv” – all previous applications made to Home Credit
- 300+ heterogenous mixture of both categorical and quantitative features

Numerical Correlation

Key Points:

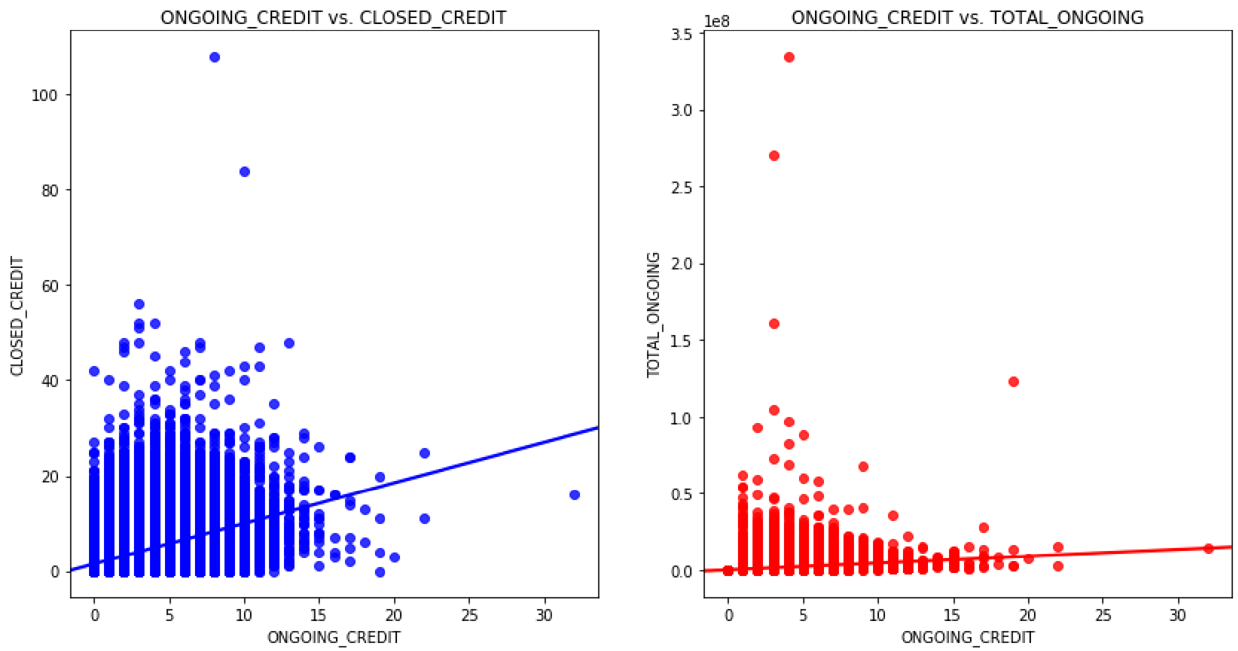
- Strong correlation between DAYS_TERMINATION and DAYS_DECISION which indicates multicollinearity.
    - DAYS_DECISION removed
- Strong correlation between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE which indicates multicollinearity.
    - AMT_GOODS_PRICE removed
- There was a total of 12 correlations, only 3 were acted on



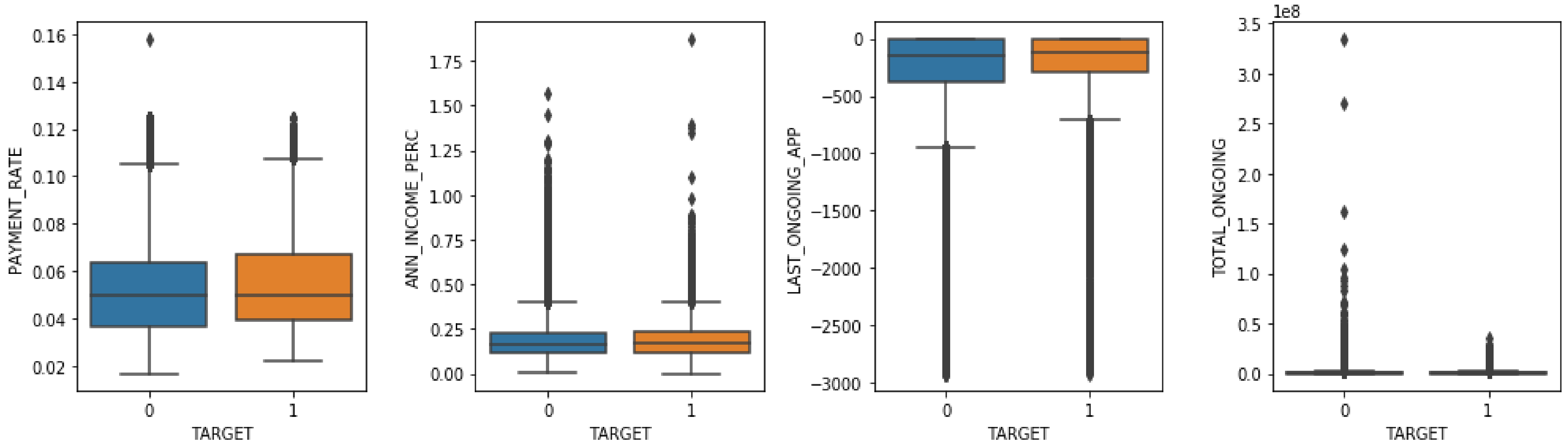
Correlation

Key Points:

- Despite the “positive correlations” seen in these two sets of variables, the distribution of the data appears fairly random.
- This is likely due to the extreme values despite the outliers already being scaled downward.



Numerical Variables as Predictors

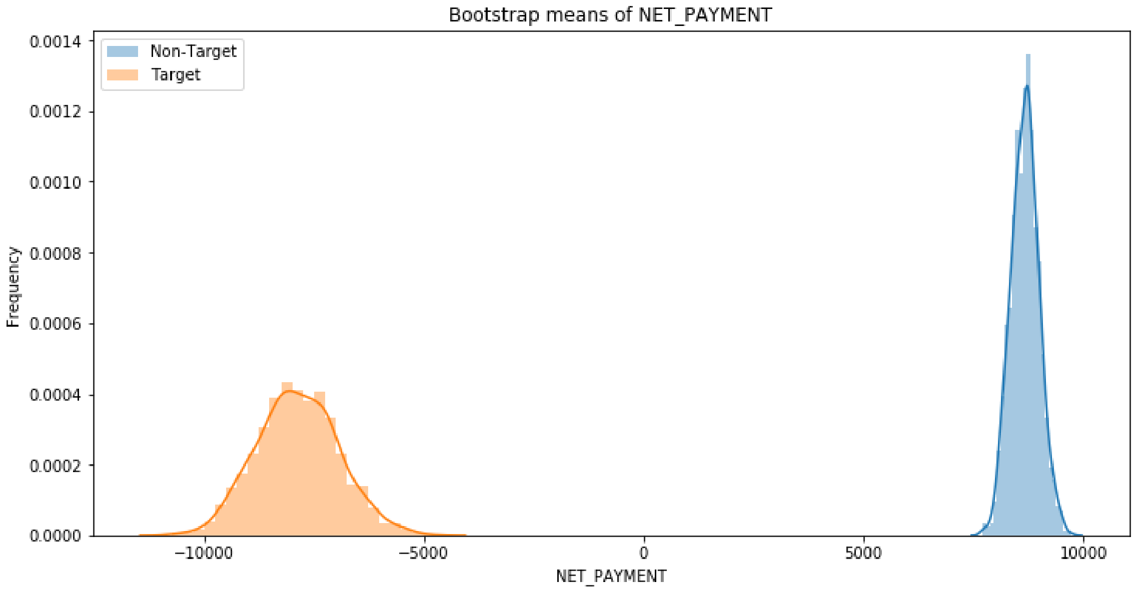


Key Points:

- Visual representations of the data were obscure and difficult to draw insights from.
- Instead, I chose to utilize inferential statistics to find statistically significant differences between target and non-target groups.

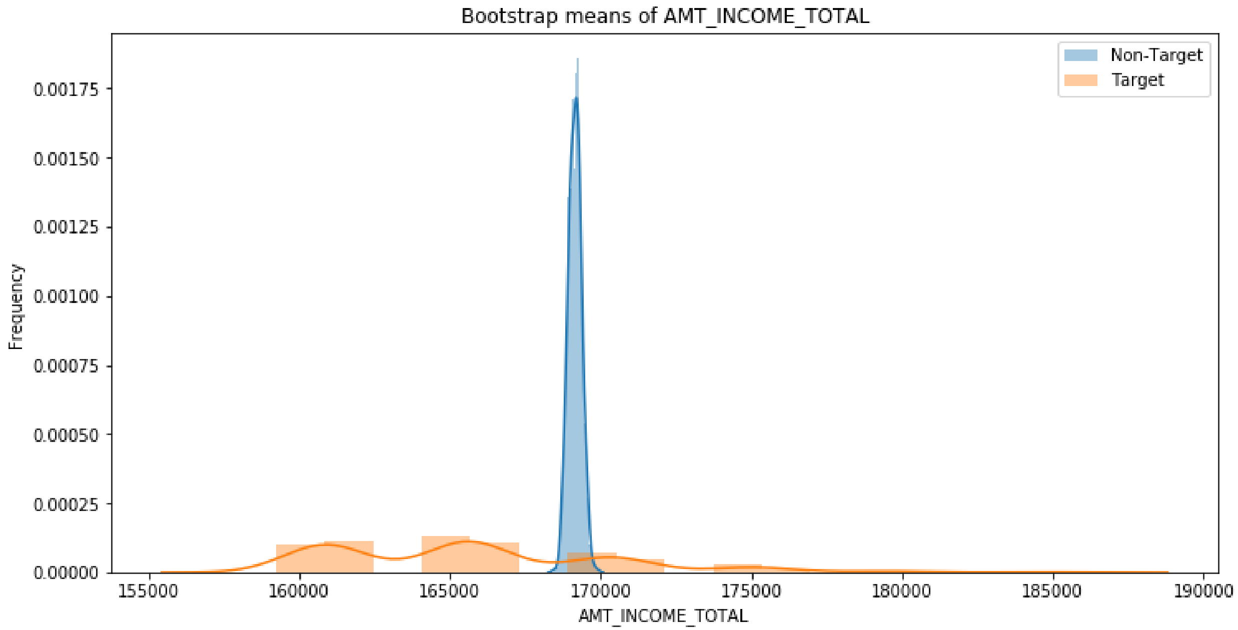
Inferential Statistics

- Two-sided t-test (p-value) and Frequentist bootstrap approach (95% confidence interval)
    - Applied to numerical and binary categorical data.
    - 5 variables failed to show statistically significant differences:
        - OWN_CAR_AGE, ENQUIRIES, SK_DPD, DAYS_TERMINATION, VALID_MOBILE

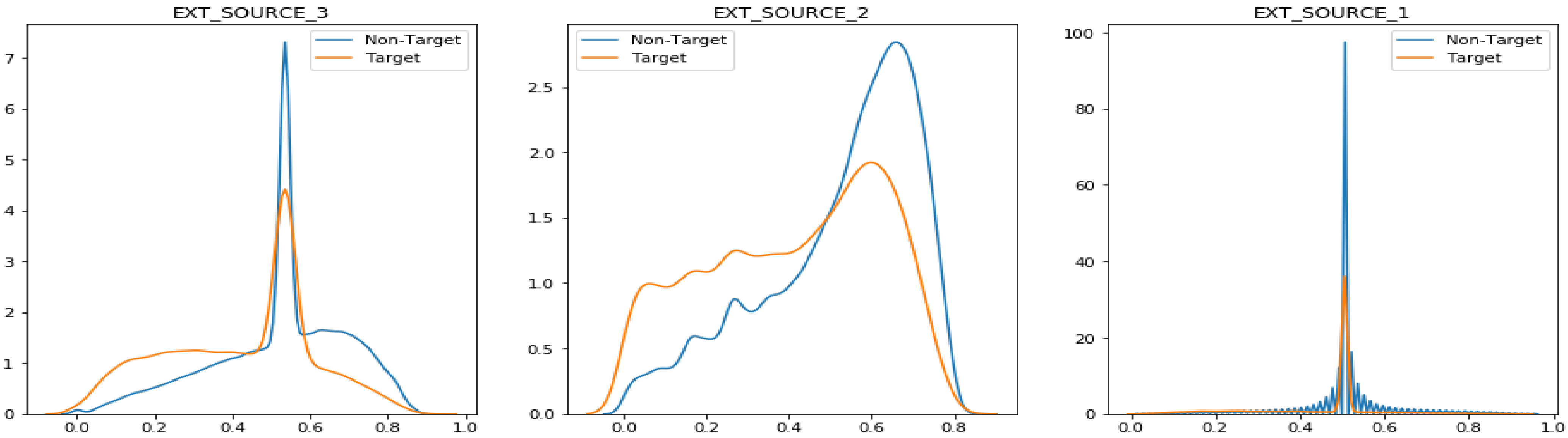


Total Income

- Non-target:
  - The means of the bootstrap samples are tightly fit around a small range
- Target:
  - The distribution of the bootstrap samples are right skewed and multi-peaked.

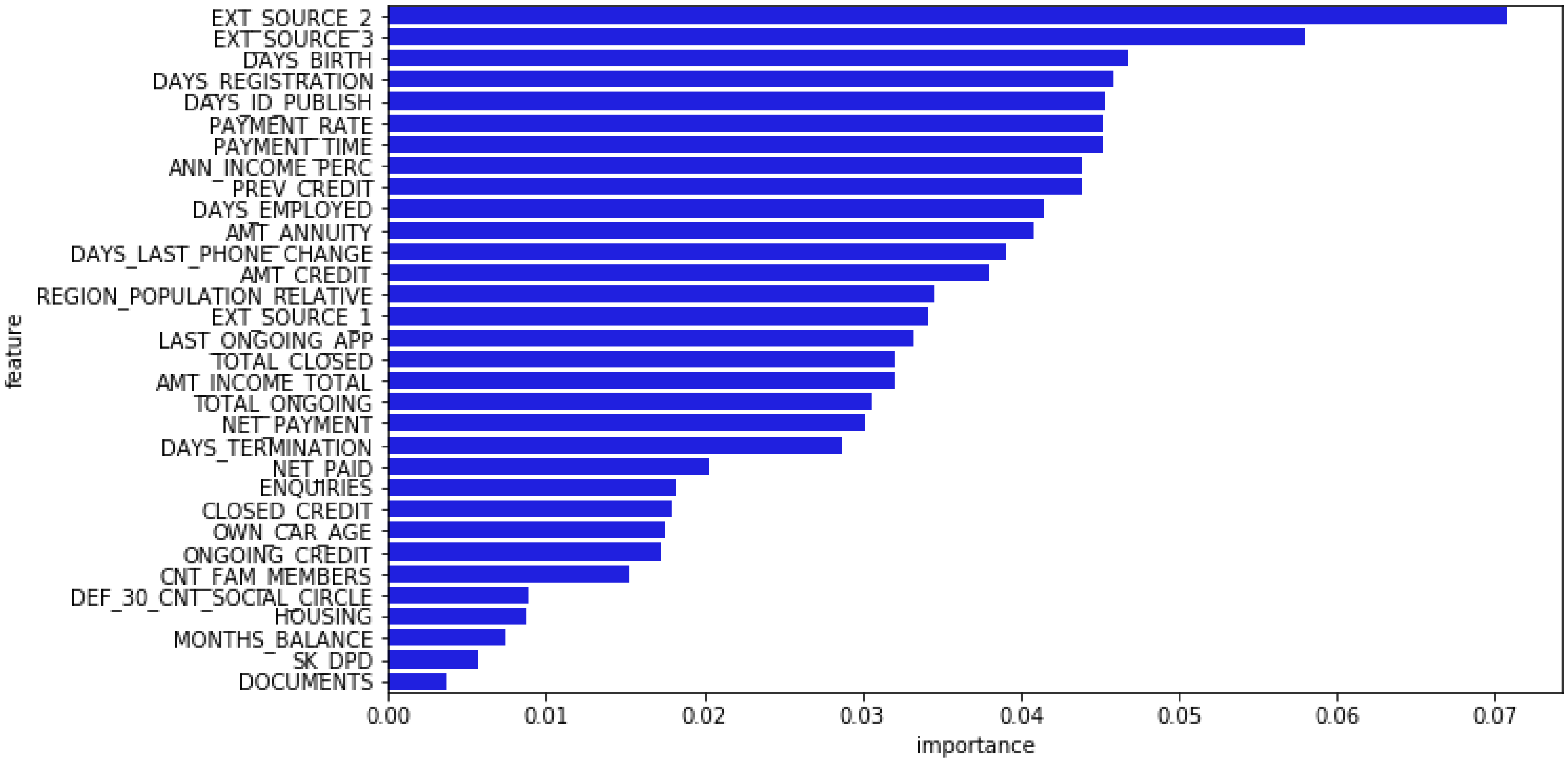


External Source Scoring

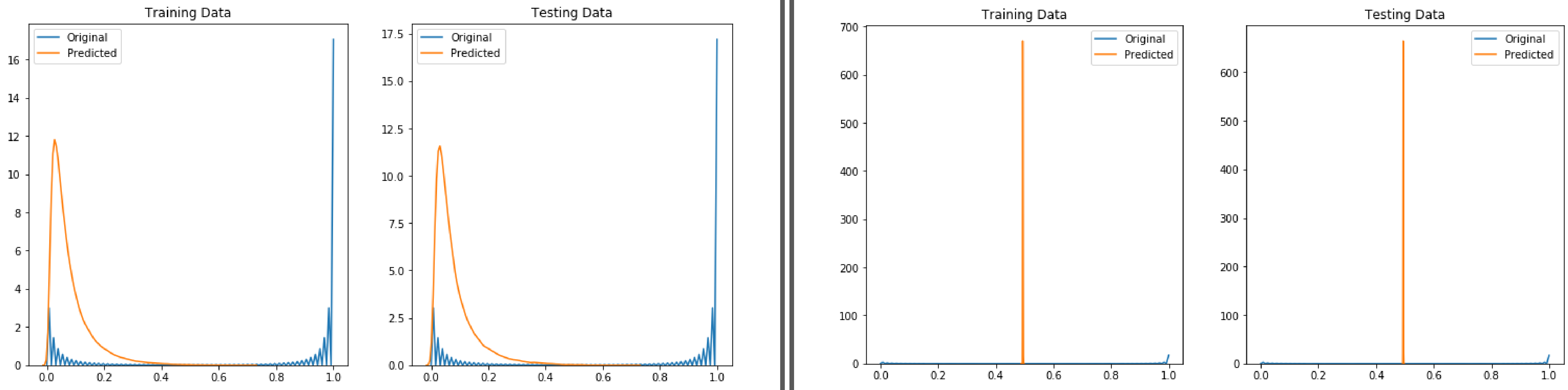


- External Source Scoring shows the highest magnitude of correlation.
  - In the training data, the target group has a higher population of lower scores than the non-target group.

Numerical Feature Importance



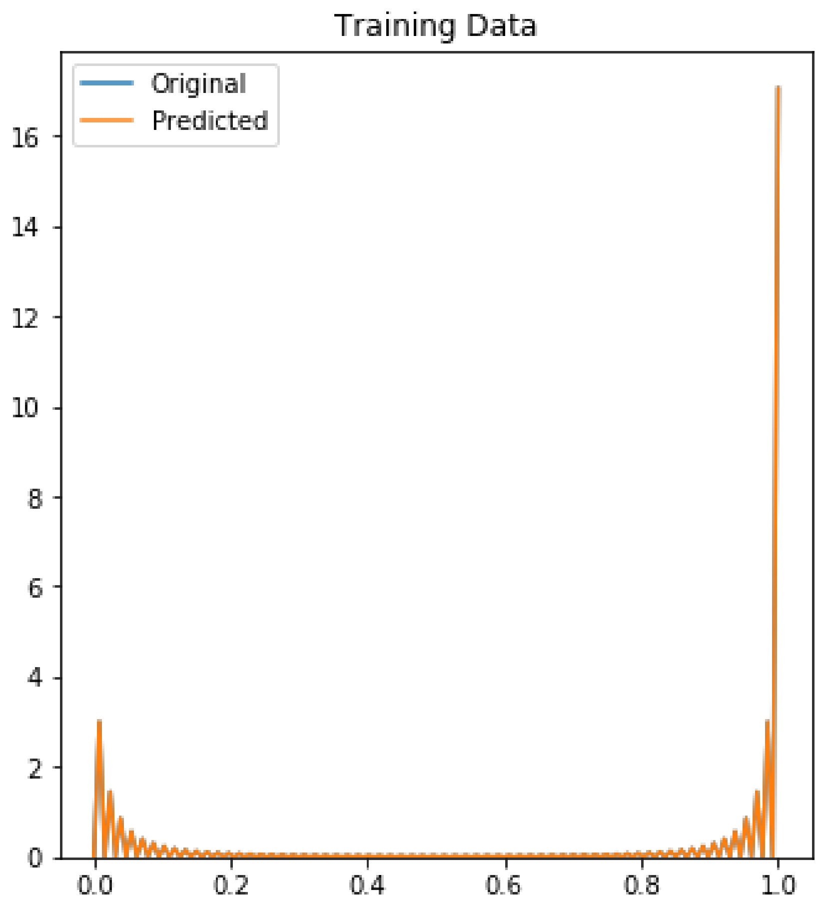
Logistic Regression

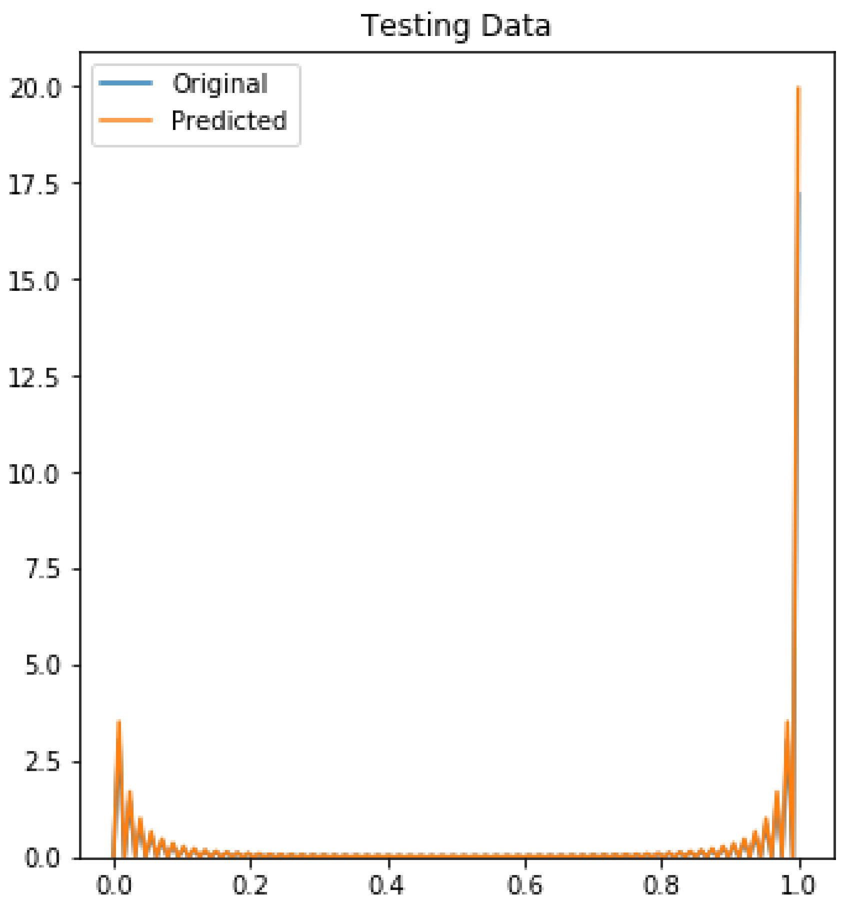


Logistic Regression

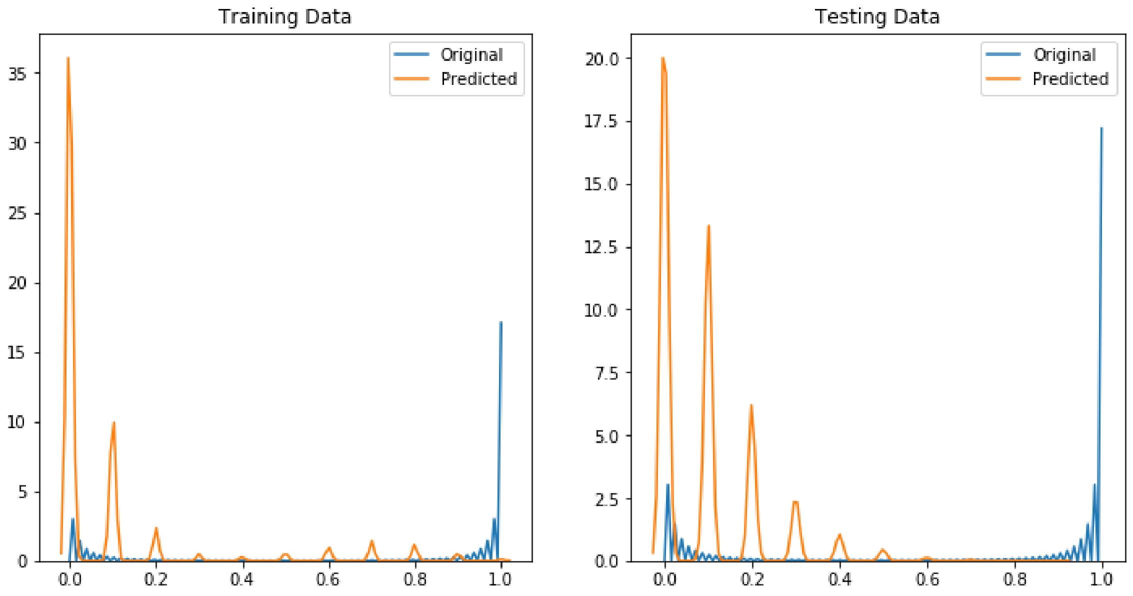
Logistic Regression with AdaBoost

Decision Tree Classifier

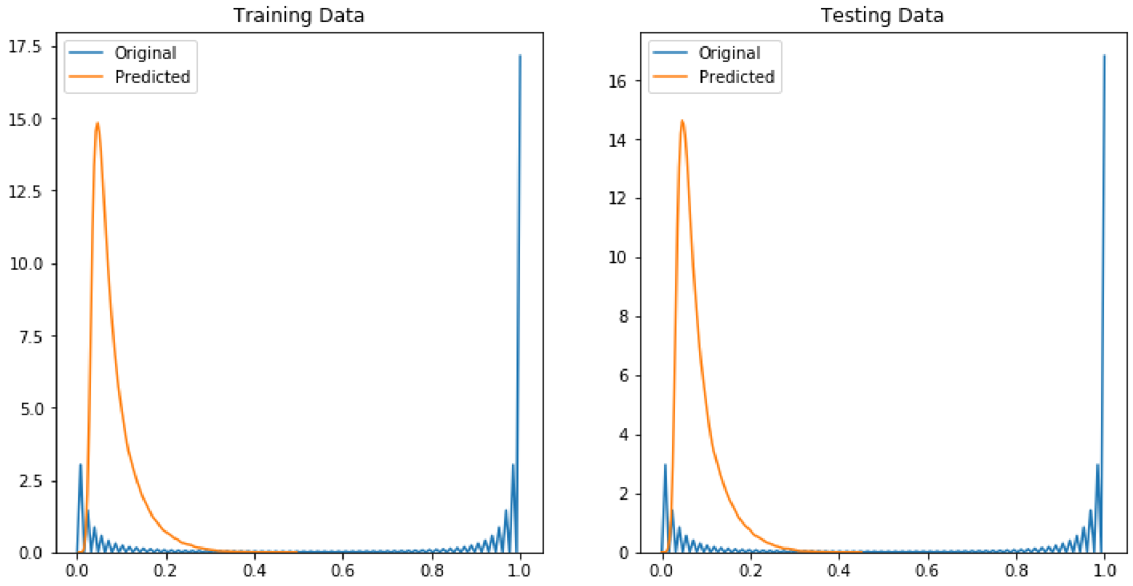




Random Forest Classifier

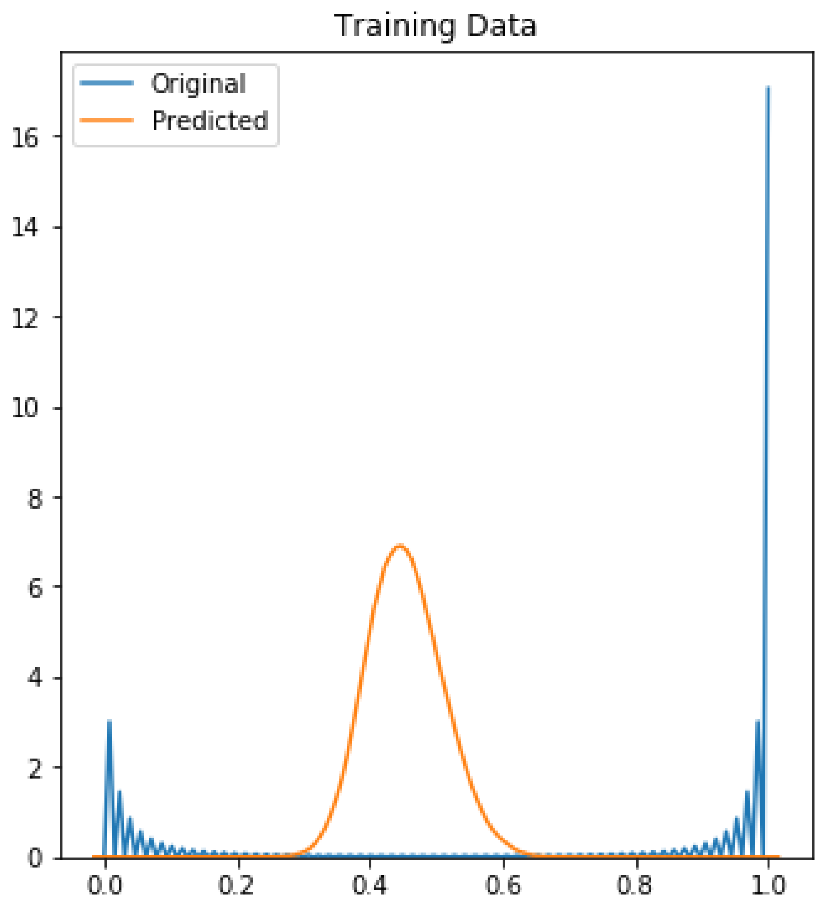


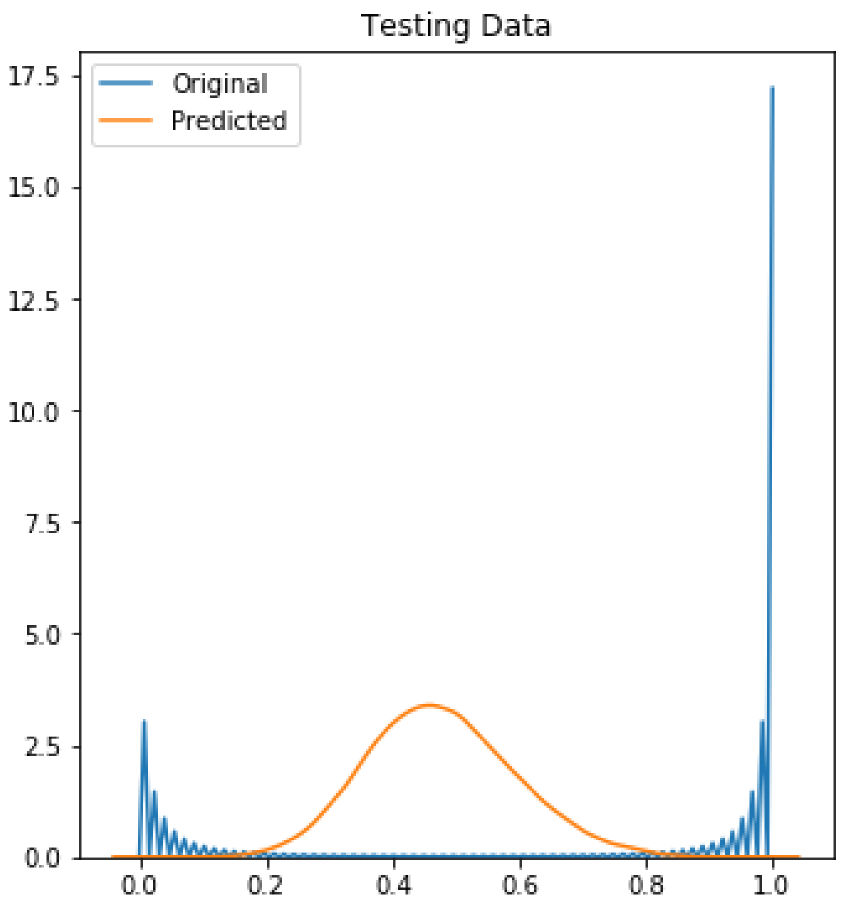
Pre-Hyperparameter tuning



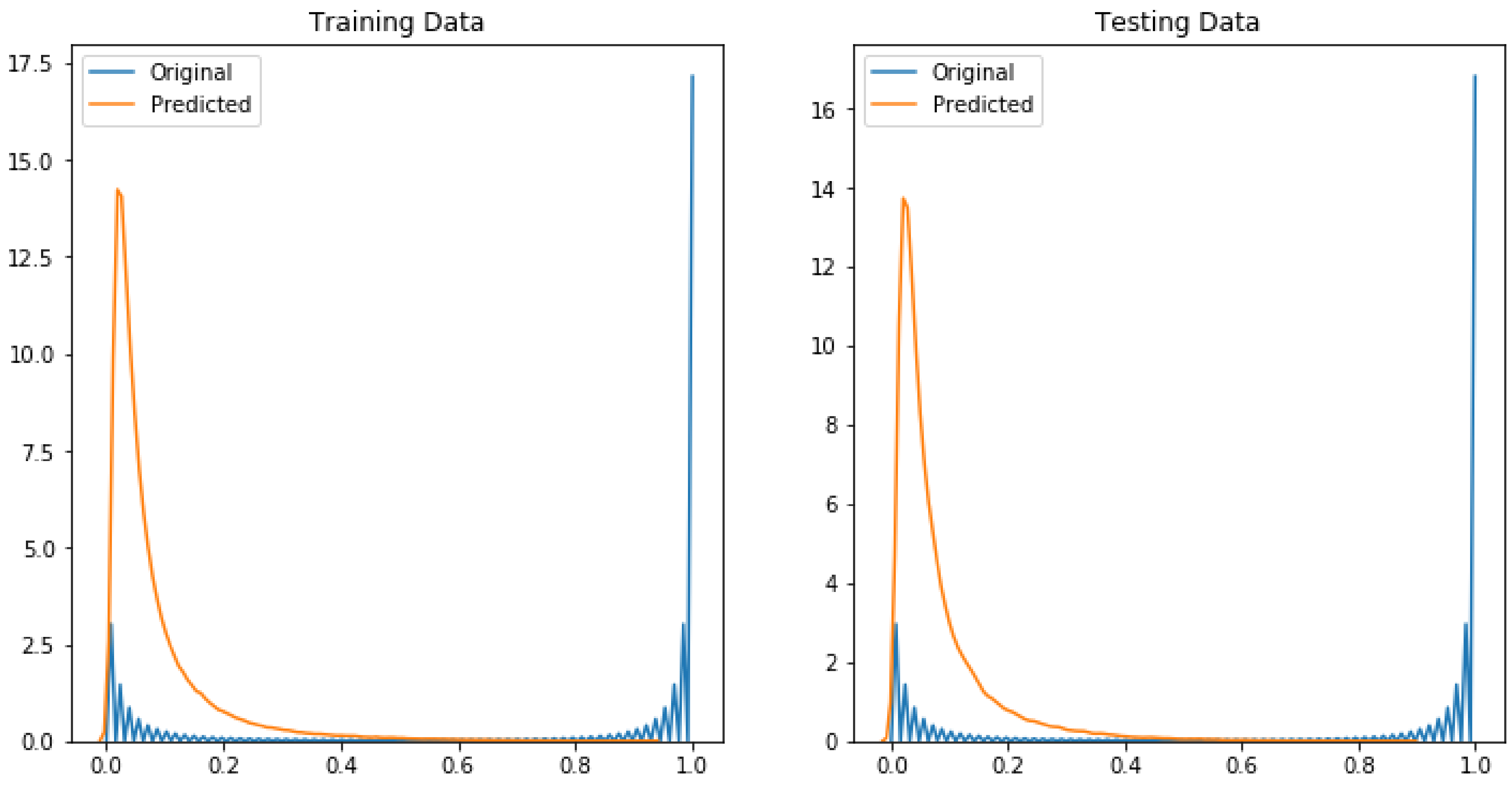
Post-Hyperparameter tuning

LinearSVC (Support Vector Classification)





XGBClassifier



Metric Scoring

- Scoring was performed via Area Under the ROC Curve.
- Of the five attempted algorithms, DecisionTreeClassifier performed the worst early on and was no longer tested on.
- LogisticRegression performed slightly better than the LinearSVC, with XGBClassifier outperforming the rest.
- Currently, the highest score is 0.761.