

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:

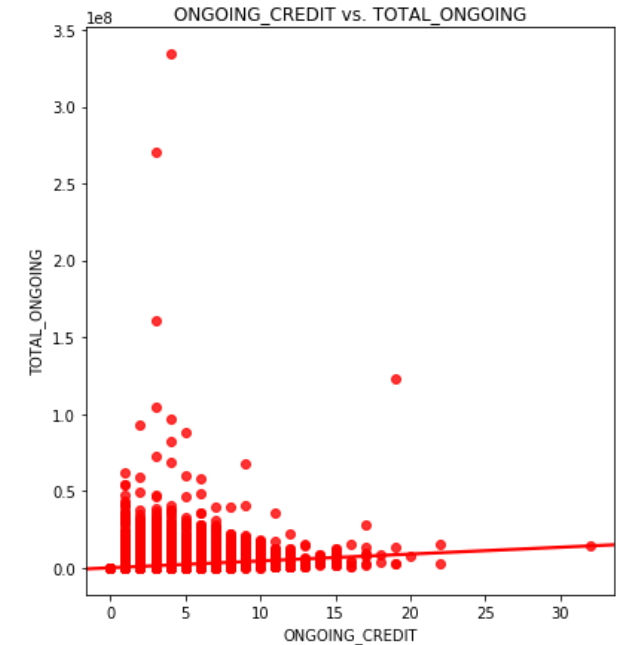
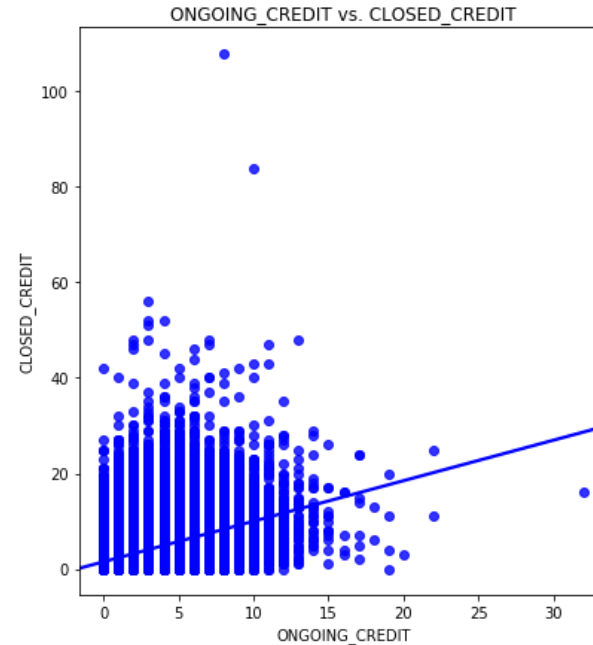
- Strong correlation between DAYS_TERMINATION and DAYS_DECISION which indicates multicollinearity.
 - DAYS_DECISION removed
- Strong correlation between AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE which indicates multicollinearity.
 - AMT_GOODS_PRICE removed
- There was a total of 12 correlations, only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

AMT_INCOME_TOTAL	1	0.16	0.19	0.16	0.074	0.028	0.021	0.016	0.015	0.018	0.03	0.06	0.024	0.086	0.033	0.027	0.15	0.015	0.13	0.061	0.077	0.037	0.016	0.007	0.043	0.067	0.009	0.008	0.02	0.001	0.013	0.027	0.037	0.008	0.014
AMT_CREDIT	0.16	1	0.77	0.99	0.1	0.009	0.028	0.063	0.016	0.074	0.037	0.13	0.11	-0.1	0.066	0.56	0.37	0.052	0.13	0.05	0.096	0.059	0.036	0.077	0.037	0.098	0.037	0.039	0.06	0.001	0.022	0.055	-0.1	0.066	0.21
AMT_ANNUITY	0.19	0.77	1	0.77	0.12	0.039	0.042	0.076	-0.02	0.064	0.027	0.13	0.081	-0.13	0.014	-0.063	0.48	0.038	0.124	0.001	0.099	0.033	-0.023	0.022	0.031	0.099	-0.037	0.037	0.034	0.011	0.021	0.009	-0.086	0.011	0.19
AMT_GOODS_PRICE	0.16	0.99	0.77	1	0.1	0.012	0.026	0.061	-0.02	0.076	0.04	0.14	0.12	-0.11	0.005	-0.53	0.37	-0.053	0.13	0.049	0.1	0.064	-0.028	0.066	0.039	0.095	-0.032	-0.036	0.063	-0.001	0.014	0.054	-0.1	0.099	0.18
REGION_POPULATION_RELATIVE	-0.074	0.1	0.12	0.1	1	0.054	0.011	-0.024	0.024	0.044	0.004	0.2	0.067	0.54	0.025	-0.016	-0.051	0.015	0.041	0.015	0.029	0.041	-0.025	0.003	0.018	0.051	0.01	-0.018	0.002	0.006	-0.026	-0.03	0.001	0.040	0.005
DAYS_REGISTRATION	-0.028	0.009	0.039	0.012	0.054	1	0.045	0.17	0.003	0.057	-0.096	-0.06	-0.11	0.08	-0.023	0.028	-0.033	0.017	0.092	0.027	0.032	0.005	0.011	0.085	0.031	-0.054	0.033	0.015	0.004	0.039	0.18	0.33	0.047	0.1	0.001
OWN_CAR_AGE	-0.021	0.028	0.042	0.026	0.011	0.045	1	0.088	0.005	0.024	0.016	0.01	-0.051	0.026	-0.017	0.007	-0.02	0.008	0.03	0.018	0.015	0.021	0.014	0.002	0.055	-0.014	-0.012	-0.014	0.012	0.002	0.069	0.085	-0.011	0.013	0.006
CNT_FAM_MEMBERS	-0.016	0.063	0.076	0.061	0.024	0.17	0.088	1	0.001	0.027	0.023	0.001	0.068	0.03	-0.014	0.007	0.057	0.015	0.074	0.034	0.021	0.019	-0.014	0.008	0.016	0.009	0.003	0.006	0.001	0.014	0.88	0.28	-0.068	0.023	0.002
DEF_30_CNT_SOCIAL_CIRCLE	-0.015	0.018	0.02	-0.02	0.002	0.003	0.005	0.001	1	0.001	0.033	0.032	-0.02	0.018	0.014	0.002	0.015	0.006	0.008	0.008	0.008	0.008	0.011	0.002	0.004	0.015	0.002	0.017	0.012	0.014	0.007	0.004	0.014	0.002	0.023
DAYS_LAST_PHONE_CHANGE	-0.019	0.074	0.066	0.076	0.044	0.057	0.024	0.027	0.001	1	0.064	0.2	0.087	0.026	0.12	0.036	0.013	0.06	0.066	0.09	0.048	0.13	0.12	0.03	0.095	-0.18	0.21	0.17	0.025	0.14	0.005	0.083	0.13	0.089	0.022
EXT_SOURCE_3	-0.03	0.037	0.027	0.04	0.004	0.096	-0.016	-0.027	0.033	0.064	1	0.094	0.11	-0.012	0.081	0.022	0.094	-0.11	-0.11	0.37	0.013	-0.05	0.021	0.004	0.029	0.005	-0.04	-0.035	0.041	-0.038	-0.04	-0.18	-0.053	-0.11	0.003
EXT_SOURCE_2	0.06	0.13	0.13	0.14	0.2	-0.06	0.001	0.001	0.032	-0.2	0.094	1	0.13	-0.29	0.018	-0.046	-0.035	-0.052	0.069	0.014	0.05	0.04	-0.018	0.004	0.024	0.051	-0.023	-0.025	0.028	0.004	-0.016	-0.092	-0.085	-0.051	0.007
EXT_SOURCE_1	-0.024	0.11	0.081	0.12	0.067	-0.11	-0.051	-0.068	-0.02	0.087	0.11	0.13	1	-0.078	0.023	-0.078	0.034	0.046	0.03	0.029	0.044	0.075	-0.019	0.009	0.034	0.095	-0.029	0.012	0.031	-0.028	-0.098	-0.36	-0.018	0.089	0.065
HOUSING	-0.086	-0.1	-0.13	-0.11	-0.54	0.08	0.026	0.03	0.018	0.026	-0.012	-0.29	0.078	1	-0.013	0.003	0.081	0.017	0.042	0.031	0.029	0.054	0.029	0.002	0.011	-0.035	0.005	0.012	0.006	-0.01	0.025	0.003	0.043	0.005	-0.014
ENQUIRIES	-0.033	0.066	0.014	0.005	0.025	-0.023	-0.017	-0.014	0.014	-0.12	-0.085	0.018	0.023	-0.013	1	0.004	-0.076	0.099	0.07	0.2	0.05	0.21	-0.22	-0.016	0.005	0.3	0.14	0.19	0.079	0.23	-0.03	-0.072	-0.017	0.065	0.021
PAYMENT_RATE	-0.027	-0.56	-0.063	-0.53	-0.016	0.028	0.007	0.007	0.002	0.036	0.002	0.046	0.078	0.003	0.004	1	0.03	0.032	-0.06	-0.093	0.028	0.044	0.031	-0.01	-0.01	-0.048	0.019	0.017	0.051	0.012	0.021	0.092	0.055	0.019	0.007
ANN_INCOME_PERC	-0.15	0.37	0.48	0.37	-0.051	0.033	-0.02	0.057	0.015	-0.013	0.094	-0.035	0.034	-0.081	-0.076	-0.03	1	0.005	-0.13	-0.14	-0.052	-0.063	0.025	0.004	0.038	-0.061	-0.072	-0.07	-0.019	0.017	0.002	0.081	-0.013	0.023	0.17
LAST_ONGOING_APP	-0.015	0.052	-0.038	0.053	0.015	0.017	0.008	0.015	0.006	0.06	0.11	0.052	0.046	0.017	0.099	0.032	0.005	1	-0.062	0.067	0.03	0.065	0.015	0.004	0.011	-0.041	0.027	0.018	0.005	0.016	0.004	0.044	0.047	0.065	-0.014
TOTAL_ONGOING	0.13	0.13	0.12	0.13	0.041	0.092	0.03	0.074	0.008	0.066	0.11	0.069	0.03	0.042	0.07	-0.06	0.13	0.062	1	0.39	0.22	0.02	-0.014	3e-05	0.027	0.038	-0.043	0.051	0.012	0.011	0.051	0.059	0.093	0.009	0.02
ONGOING_CREDIT	-0.061	0.054	0.000	0.049	0.019	0.027	0.018	0.034	0.008	-0.09	-0.37	0.014	0.029	0.031	0.2	-0.093	-0.14	-0.067	0.39	1	0.14	0.46	-0.049	0.066	0.036	0.052	-0.051	-0.045	0.001	-0.035	0.016	0.009	0.098	0.043	0.023
TOTAL_CLOSED	-0.077	0.096	0.099	0.1	0.029	0.032	0.015	0.021	0.009	0.048	0.13	0.05	0.044	-0.029	0.05	-0.028	-0.052	-0.03	0.22	0.14	1	0.28	-0.014	0.001	0.02	0.047	-0.014	0.021	0.024	0.016	0.01	-0.014	0.062	0.026	0.015
CLOSED_CREDIT	-0.037	0.059	0.033	0.064	0.040	0.005	0.021	0.019	0.009	0.13	0.05	0.04	0.075	0.054	0.21	-0.044	-0.063	0.065	0.2	0.46	0.28	1	-0.052	0.005	0.041	0.085	-0.035	-0.041	0.037	0.098	0.074	0.087	-0.13	-0.13	0.032
MONTHS_BALANCE	-0.016	0.036	-0.023	0.028	0.025	0.001	0.014	-0.015	0.011	0.12	0.021	0.018	0.019	0.029	0.22	0.031	0.025	0.015	-0.014	0.044	0.014	0.052	1	-0.1	0.051	0.16	-0.066	0.097	0.035	0.097	0.002	0.017	0.043	0.002	0.093
SK_DP	-0.007	0.077	0.002	0.066	0.003	0.008	0.002	0.008	0.002	0.03	0.004	0.004	0.009	0.002	0.016	-0.01	0.044	0.004	3e-05	0.006	0.001	0.005	-0.1	1	0.053	-0.018	-0.027	0.029	0.002	0.018	0.004	0.012	0.010	0.003	0.009
NET_PAID	-0.004	0.037	0.031	0.039	0.018	-0.031	0.005	-0.016	0.004	0.095	0.029	0.024	0.034	-0.010	0.005	-0.01	0.038	-0.011	0.027	0.036	0.02	0.041	-0.051	0.053	1	0.005	-0.066	0.060	0.005	-0.018	-0.022	-0.006	0.026	0.030	0.005
PREV_CREDIT	-0.067	0.098	0.099	0.095	0.051	-0.054	-0.014	0.009	0.015	-0.18	0.005	0.051	0.095	-0.035	0.3	-0.048	-0.061	0.041	0.038	0.052	0.047	0.085	-0.16	-0.018	0.005	1	0.063	0.22	0.1	-0.41	-0.055	-0.17	-0.061	0.045	0.018
DAYS_DECISION	-0.009	0.037	0.037	0.032	-0.01	0.033	-0.014	0.003	0.026	0.21	0.04	0.023	0.025	0.005	0.14	0.019	-0.072	0.027	0.043	0.051	0.014	-0.035	-0.066	0.027	0.066	0.63	1	0.93	0.066	0.075	0.041	0.042	0.058	0.042	0.075
DAYS_TERMINATION	-0.008	0.039	-0.037	0.036	0.013	0.015	-0.014	0.006	0.001	0.17	-0.035	0.025	-0.012	0.012	0.19	0.017	-0.07	0.018	-0.051	-0.045	0.021	-0.041	-0.097	-0.029	0.066	0.22	0.93	1	0.052	-0.16	0.008	0.096	0.048	0.035	-0.077
NET_PAYMENT	-0.02	0.06	0.034	0.063	0.002	0.043	0.012	0.001	0.012	-0.025	0.041	0.028	0.031	0.006	0.079	-0.051	-0.019	0.054	0.013	0.001	0.024	0.037	-0.038	0.002	0.006	0.1	0.066	0.052	1	-0.066	-0.014	0.032	0.008	0.003	0.003
PAYMENT_TIME	-0.001	0.001	0.011	0.003	0.006	0.039	0.002	0.014	-0.014	0.14	-0.038	0.004	0.028	-0.01	-0.23	0.012	0.017	0.016	0.011	-0.035	0.001	-0.098	0.097	0.018	-0.018	-0.41	-0.075	-0.16	-0.066	1	0.023	0.11	0.021	0.057	0.002
CNT_CHILDREN	-0.013	0.002	-0.023	0.014	0.024	0.18	0.069	0.88	0.000	0.009	0.04	0.016	0.098	0.025	0.03	0.021	0.028	0.044	0.051	0.016	0.01	0.007	0.002	0.002	0.023	-0.055	0.004	0.008	0.01	0.023	1	0.33	-0.042	0.043	0.016
DAYS_BIRTH	-0.027	-0.059	0.009	0.054	-0.03	0.33	0.085	0.28	0.000	0.083	-0.18	0.092	-0.36	0.009	0.071	0.092	-0.081	0.044	0.059	0.009	-0.014	-0.087	0.001	0.012	0.064	0.17	0.042	0.098	0.032	0.11	0.33	1	0.000	0.027	0.051
DAYS_EMPLOYED	-0.037	-0.1	-0.086	-0.1	0.001	0.047	-0.011	-0.068	0.014	0.13	-0.053	-0.085	-0.081	0.004	0.017	0.055	0.013	0.047	-0.093	-0.098	0.062	-0.13	0.043	-0.011	-0.028	-0.061	0.058	0.048	0.008	0.021	-0.042	0.004	-0.1	0.034	-0.021
DAYS_ID_PUBLISH	-0.008	0.006	-0.013	0.009	0.004	0.1	0.013	-0.021	0.002	0.089	-0.11	0.051	0.089	0.054	0.065	0.015	-0.023	0.065	-0.009	0.043	0.026	-0.13	0.003	0.003	0.031	-0.045	0.042	0.035	0.008	0.057	-0.028	0.27	0.034	-0.1	0.046
DOCUMENTS	-0.014	0.21	0.19	0.18	-0.009	0.001	-0.006	0.002	0.023	-0.022	0.038	0.007	0.006	0.001	0.021	0.007	0.17	-0.014	0.02	0.023	0.015														

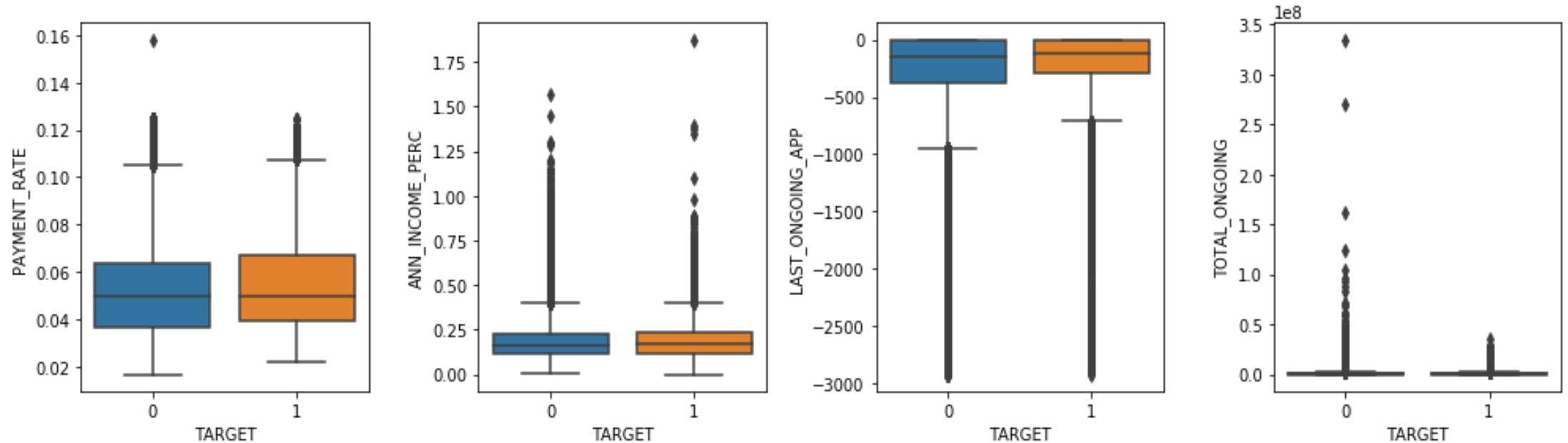
Correlation

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

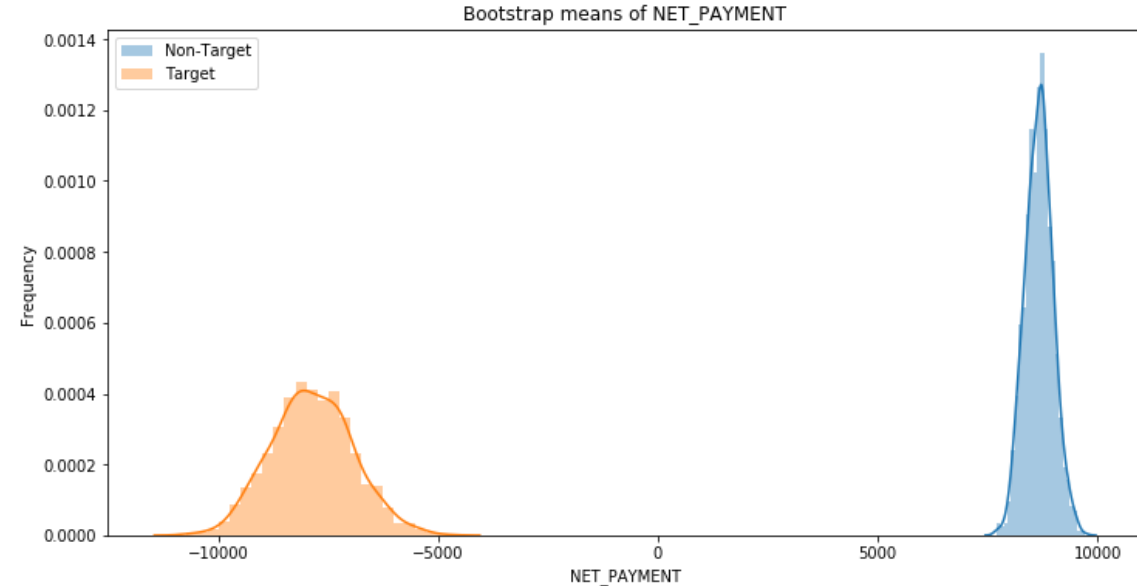


Key Points:

- 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.

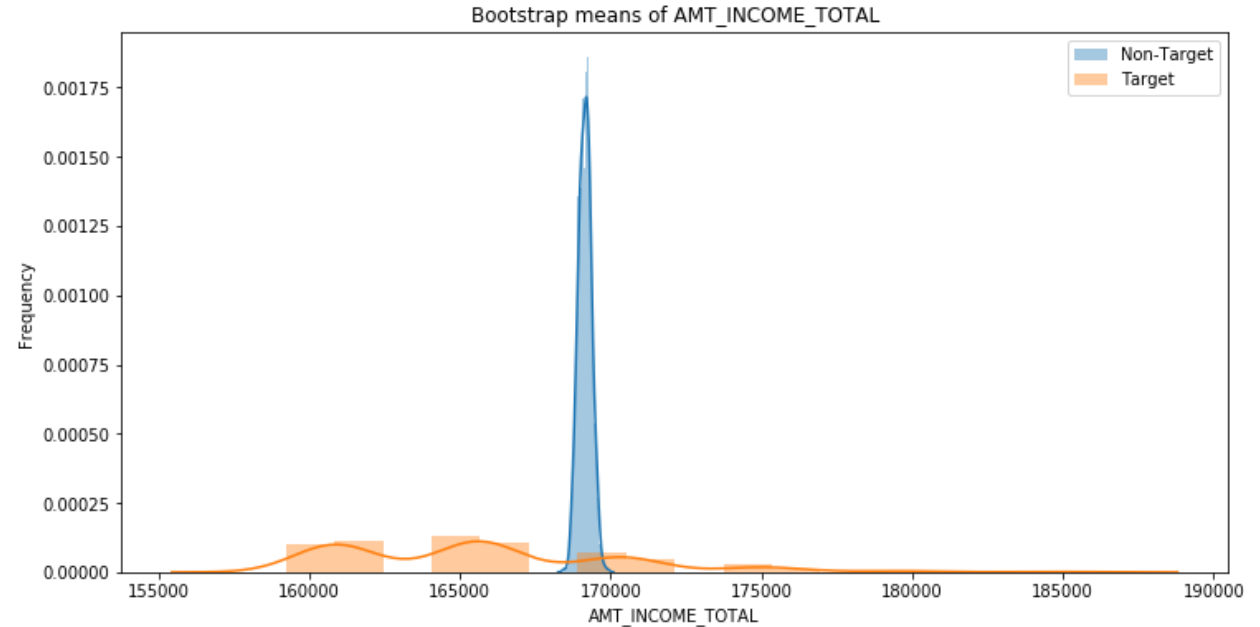
Inferential Statistics

- 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

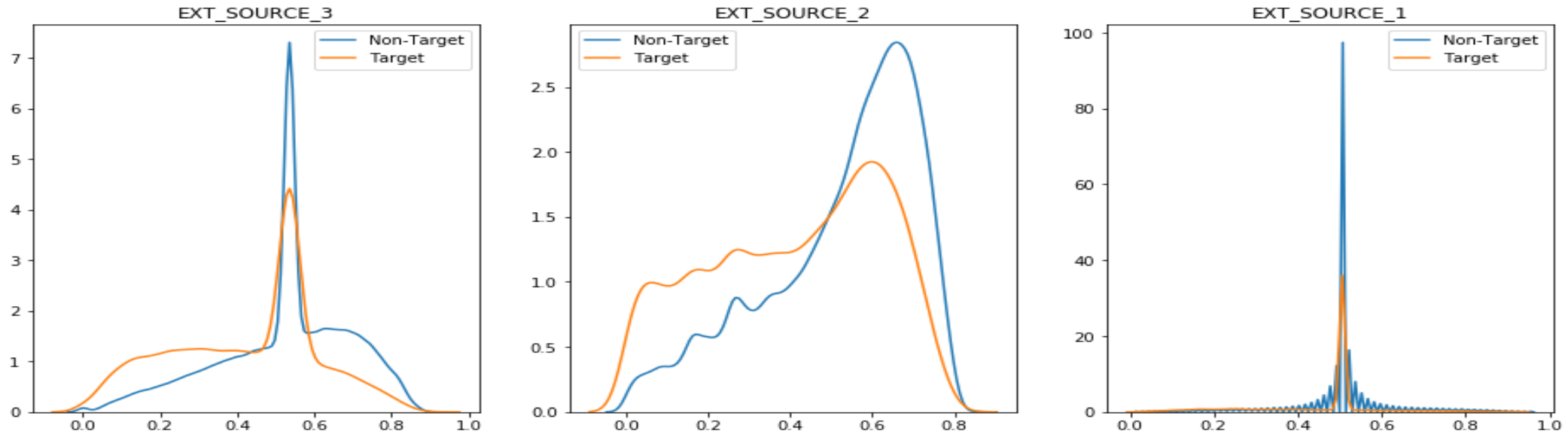


Total Income

- 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.

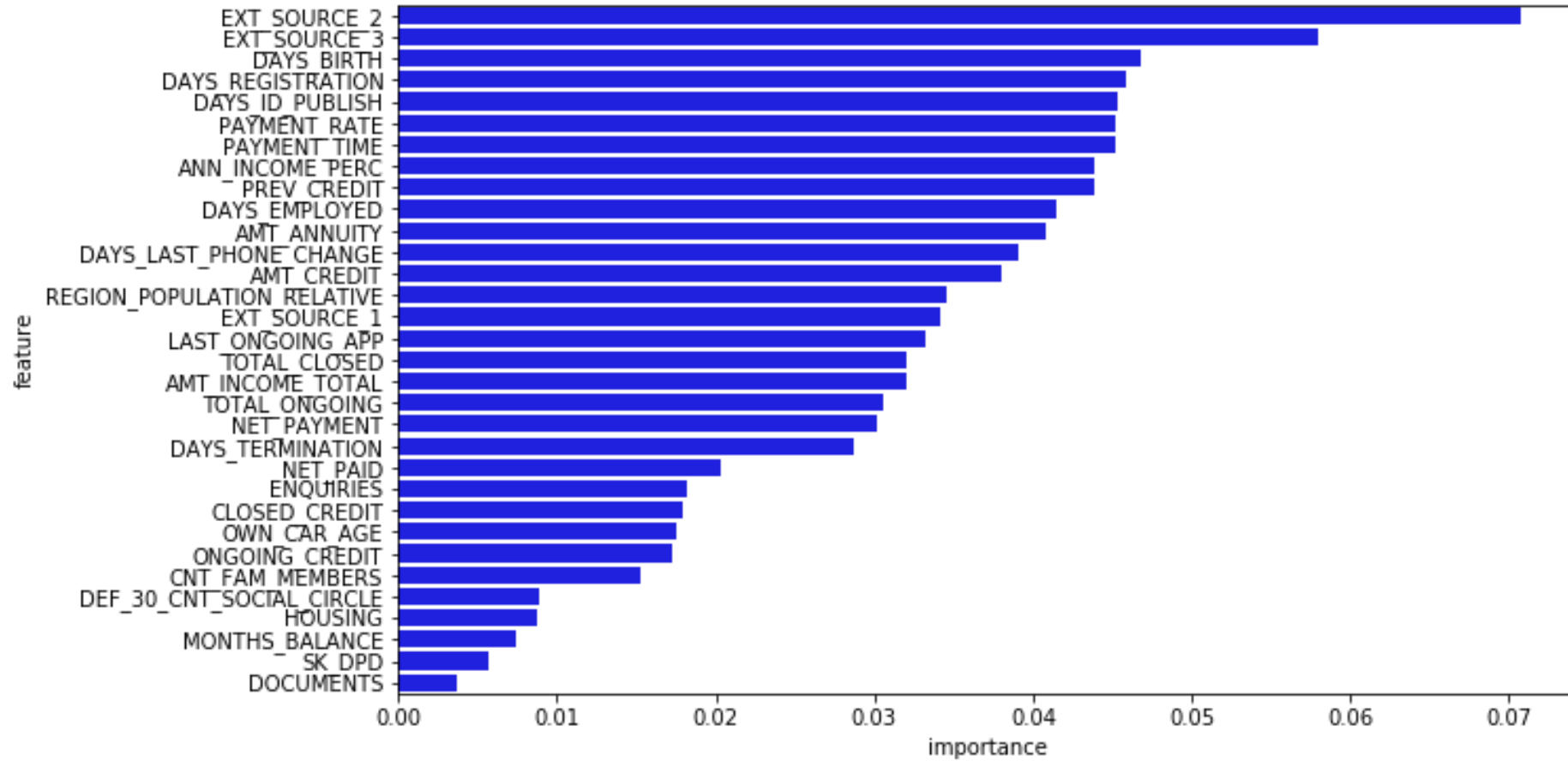


External Source Scoring

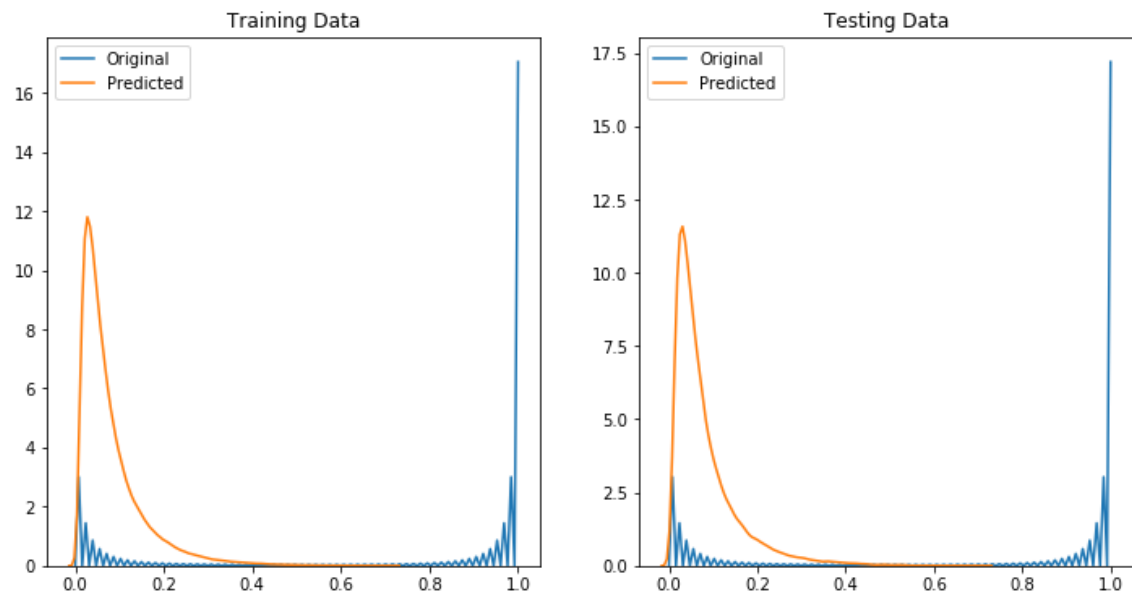


- External Source Scoring shows the highest magnitude of correlation.
 - In the training data, the target group has a higher population of lower scores than the non-target group.

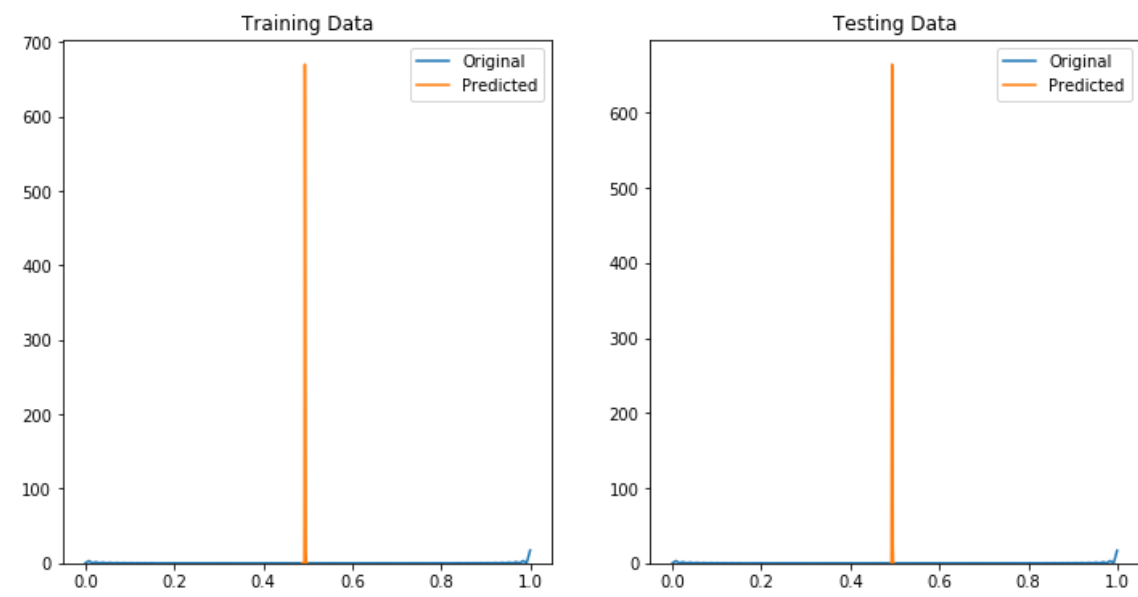
Numerical Feature Importance



Logistic Regression

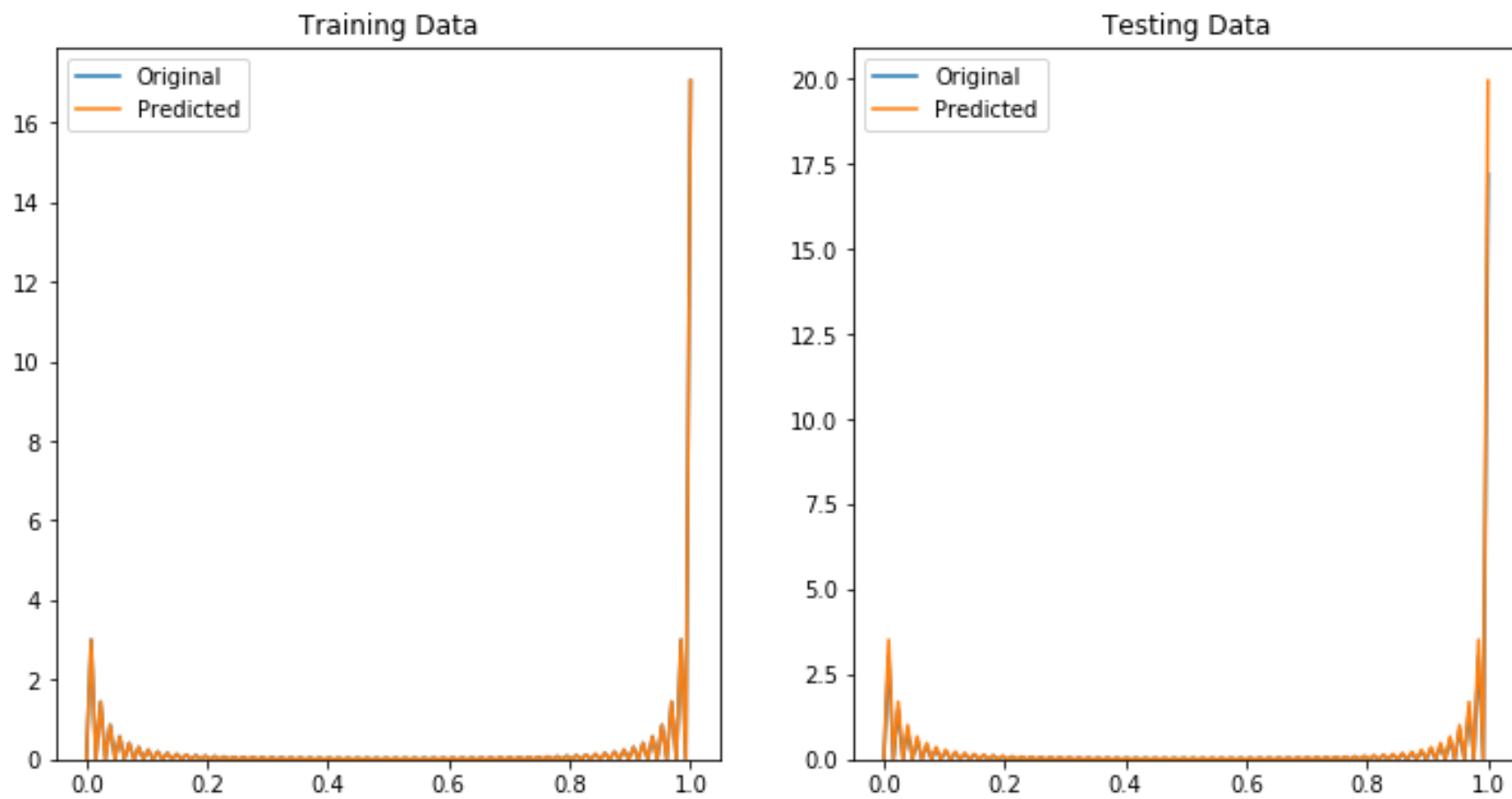


Logistic Regression

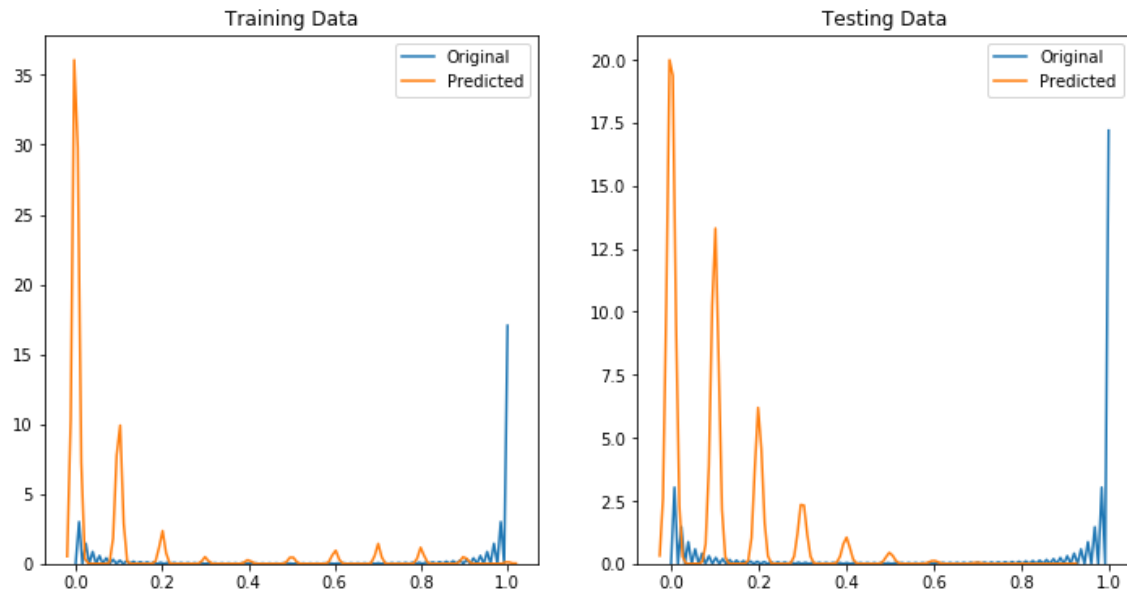


Logistic Regression with AdaBoost

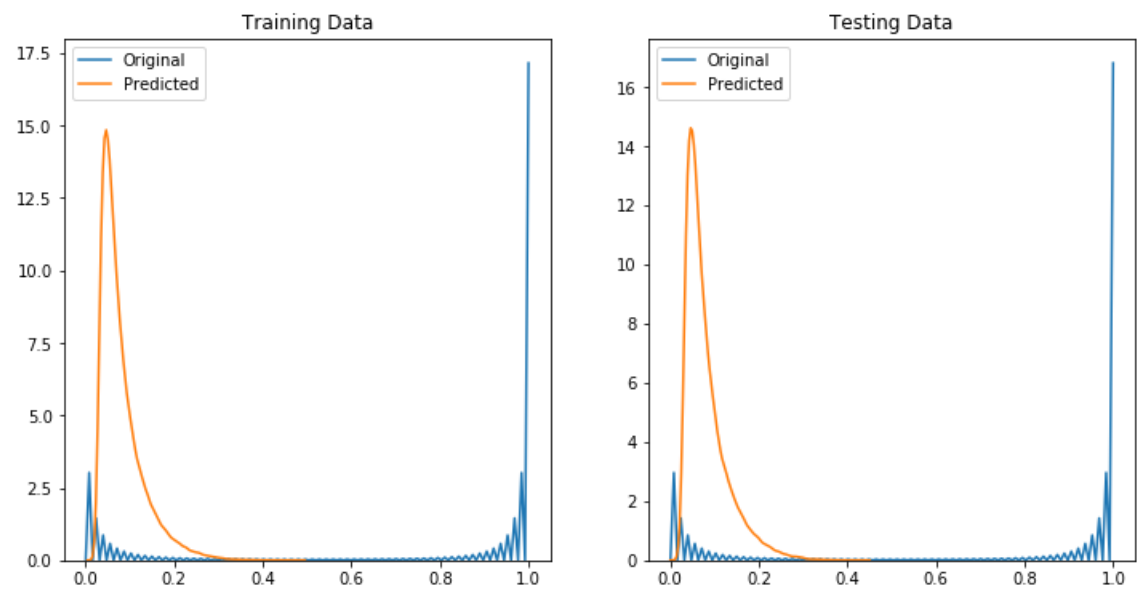
Decision Tree Classifier



Random Forest Classifier

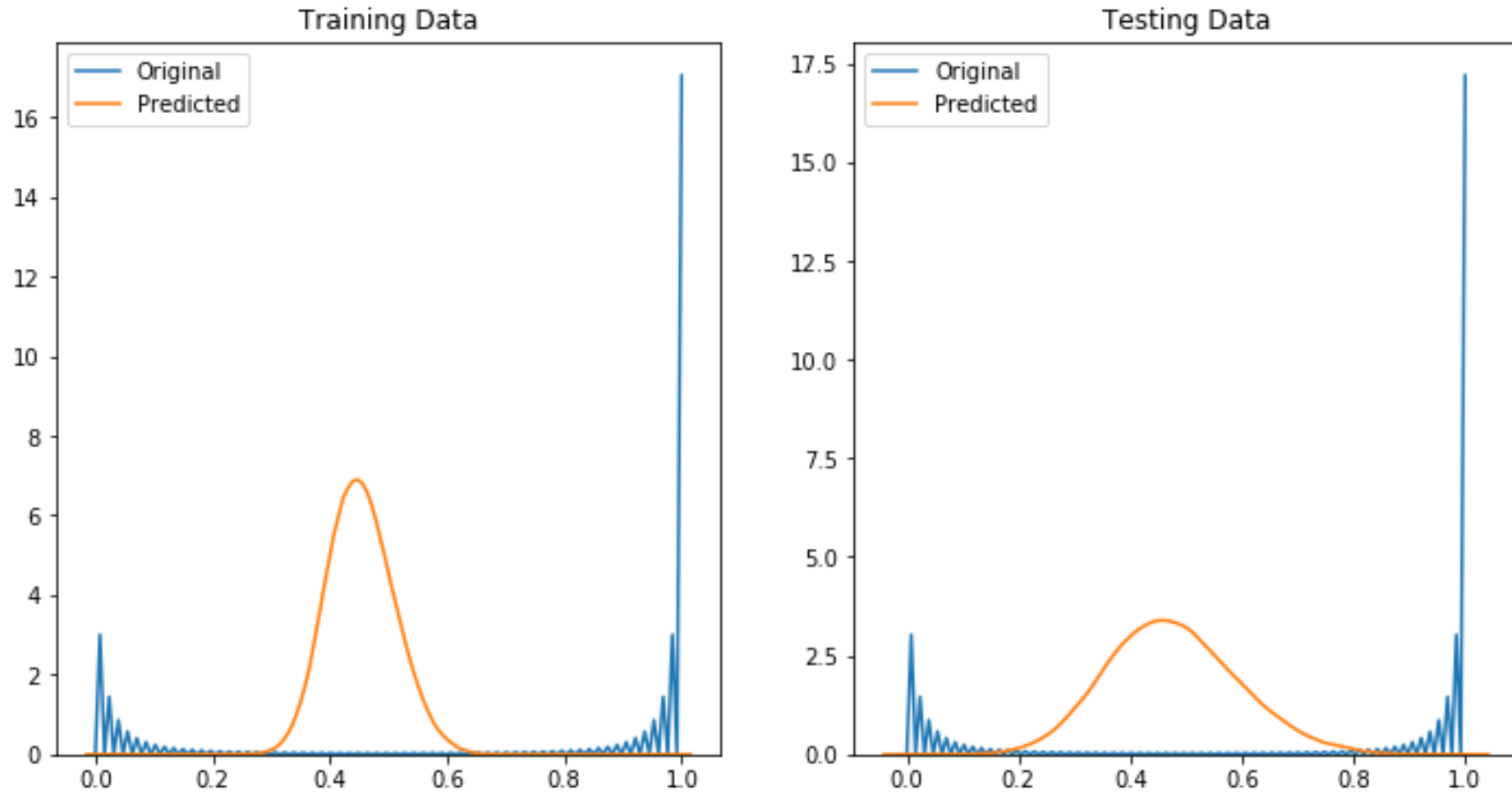


Pre-Hyperparameter tuning

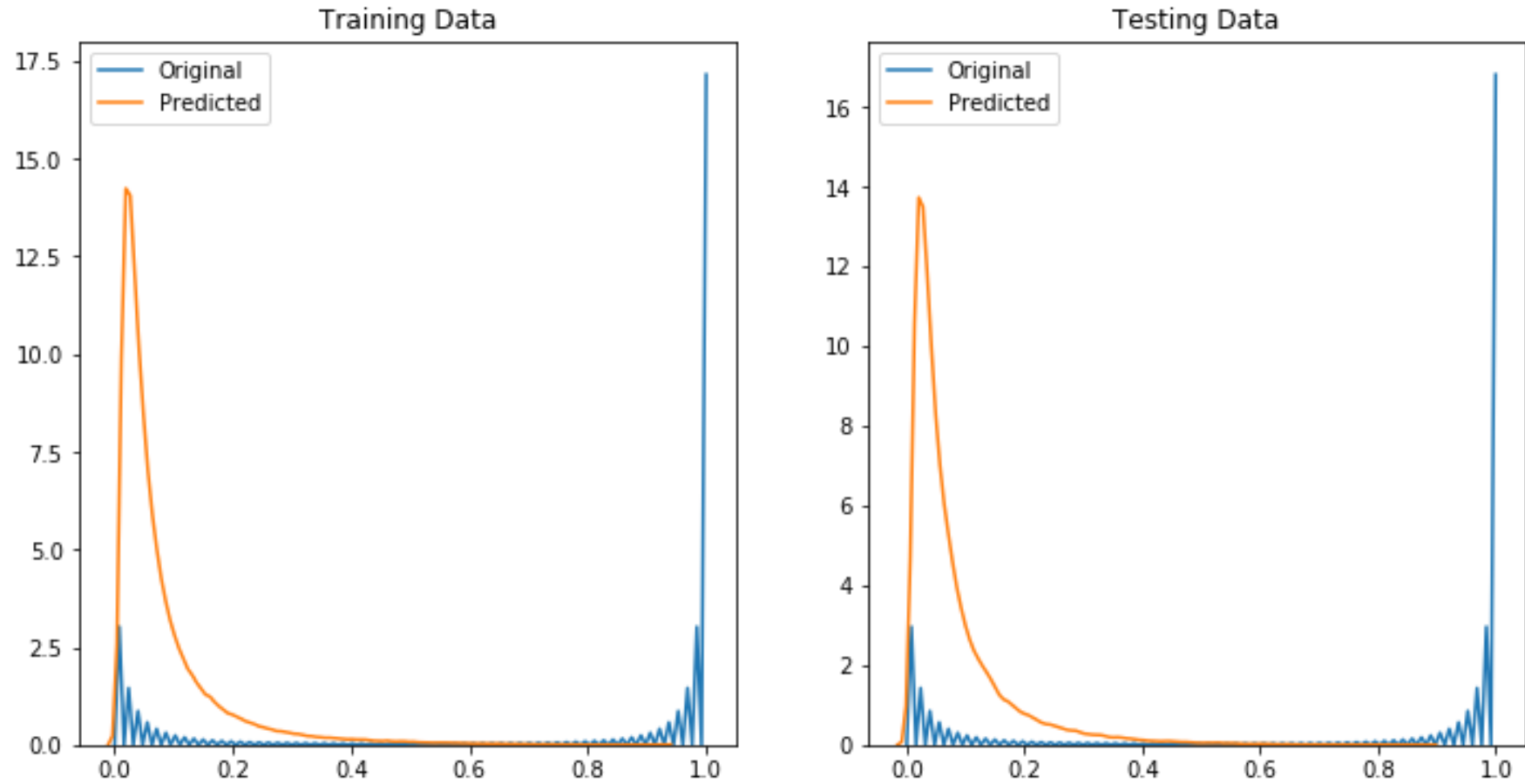


Post-Hyperparameter tuning

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.