Assignment 8 - part 2

Business Case: In this case study we are trying to build a predictive model where we are trying to predict if the donor is willing to donate. This model will help the non-profit organization to spend the required resources to contact only those donors who are willing to donate. Helping them to reduce the reach out cost by only targeting potential donors. We are taking the dataset for Paralyzed Veterans of America (PVA) which has historical data about donors such as demographics , donor history , and other such factor which may help to predict the likelihood of donation.

Payoff Matrix:

Assumptions:

The average amount of donation: $15.6

Cost of reaching out:$0.68

If a person is willing to donate and we reach out to an individual then that person will for donate for sure, i.e., assuming 100% success rate for people willing to donate.

True Positive: The model predicts that the person will donate, so we reach out to them to get the donation.

TP: 15.6(donation amount)-0.68(reach out cost)=$14.92

False Positive: The model predicts that the person is willing to donate, however in reality they are not. We reach out to them, but they do not donate.

FP: -$0.68(reach out cost)

True Negative: The model predicts that the person will not donate, so we do not reach out to them.

TN: $0

False Negative: Th model predicts that the person will not donate, however in reality if we would have reached out they would have. This in reality costs nothing , but can also be looked as an opportunity cost of -$14.92.

FN:$0

	Predicted 0	Predicted 1
Actual 0	$0	-$0.68
Actual 1	$0	$14.92

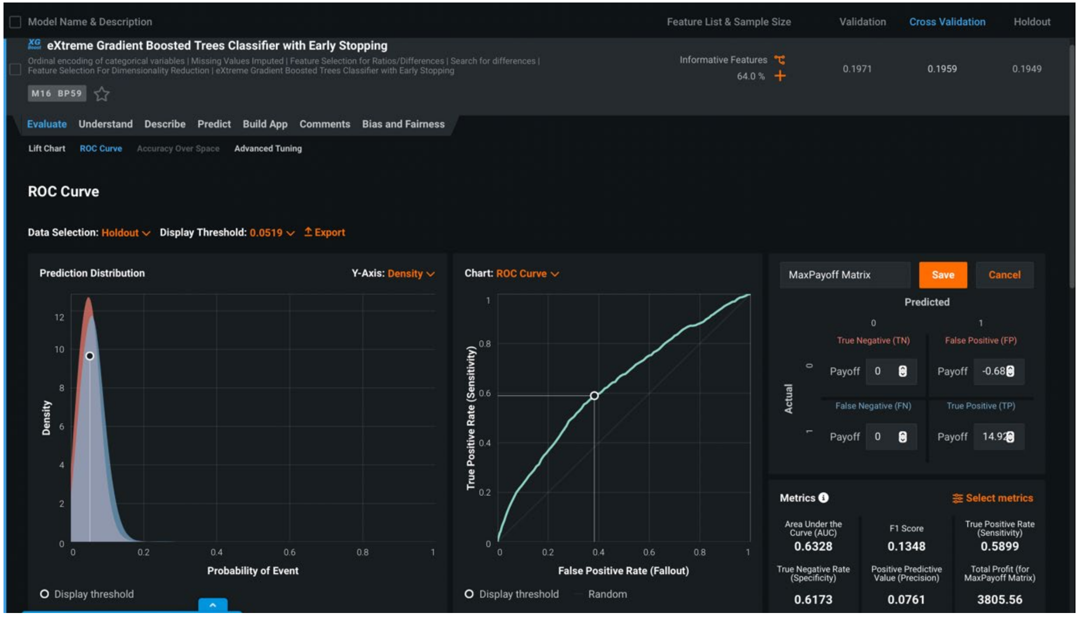
Target Leak:

We are building two models in our case :

- Classification model with TARGET_B, where we are classifying if the person is willing to donate or not. Here, the target leak is TARGET_D, here we removed it, as if the donation amount is greater than 0 then the target variable for this dataset will be 1 and 0 in other case.
- Regression model with TARGET _D to predict the amount the donor is likely to donate. Here the target leak is TARGET _B , so we have taken only those records from the dataset where the person has donated to the non-profit organization. That is we removed records with TARGET _B=0 and then removed 'TARGET _B' from the dataset.

Models for Target B

1. Gradient Boosted Trees Classifier with Early Stopping



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
eXtreme Gradient Boosted Trees Classifier with Early Stopping Ordinal encoding of categorical variables | Missing Values Imputed | Feature Selection for Ratios/Differences | Search for differences | Feature Selection For Dimensionality Reduction | eXtreme Gradient Boosted Trees Classifier with Early Stopping M16 BP59	Informative Features 64.0 %	0.1971	0.1959	0.1949

ROC Curve

Data Selection: Holdout Display Threshold: 0.0519 Export

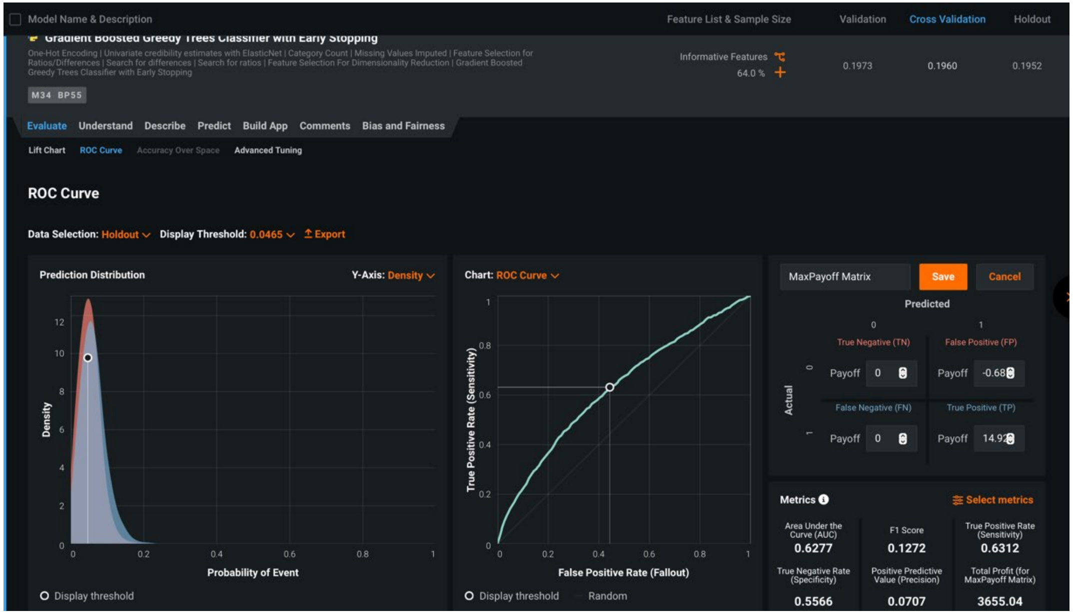
MaxPayoff Matrix

		Predicted 0	Predicted 1
Actual	0	True Negative (TN) Payoff 0	False Positive (FP) Payoff -0.68
	1	False Negative (FN) Payoff 0	True Positive (TP) Payoff 14.92

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6328	0.1348	0.5899
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for MaxPayoff Matrix)**
0.6173	0.0761	3805.56

2. Gradient Boosted Greedy Trees Classifier with Early Stopping



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Greedy Trees Classifier with Early Stopping One-Hot Encoding | Univariate credibility estimates with ElasticNet | Category Count | Missing Values Imputed | Feature Selection for Ratios/Differences | Search for differences | Search for ratios | Feature Selection For Dimensionality Reduction | Gradient Boosted Greedy Trees Classifier with Early Stopping M34 BP55	Informative Features 64.0 %	0.1973	0.1960	0.1952

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart ROC Curve Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout Display Threshold: 0.0465 Export

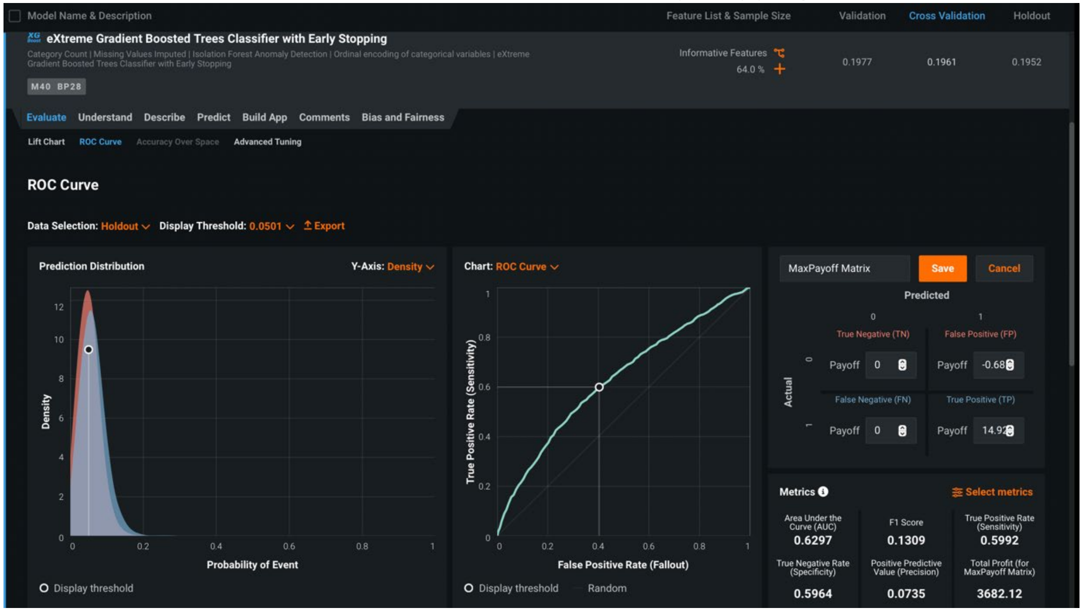
MaxPayoff Matrix Save Cancel

		Predicted 0	Predicted 1
Actual	0	True Negative (TN) Payoff 0	False Positive (FP) Payoff -0.68
	1	False Negative (FN) Payoff 0	True Positive (TP) Payoff 14.92

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6277	0.1272	0.6312
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for MaxPayoff Matrix)**
0.5566	0.0707	3655.04

3. eXtreme Gradient Boosted Trees Classifier with Early Stopping



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
eXtreme Gradient Boosted Trees Classifier with Early Stopping Category Count | Missing Values Imputed | Isolation Forest Anomaly Detection | Ordinal encoding of categorical variables | eXtreme Gradient Boosted Trees Classifier with Early Stopping M40 BP28	Informative Features 64.0 %	0.1977	0.1961	0.1952

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart ROC Curve Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout Display Threshold: 0.0501 Export

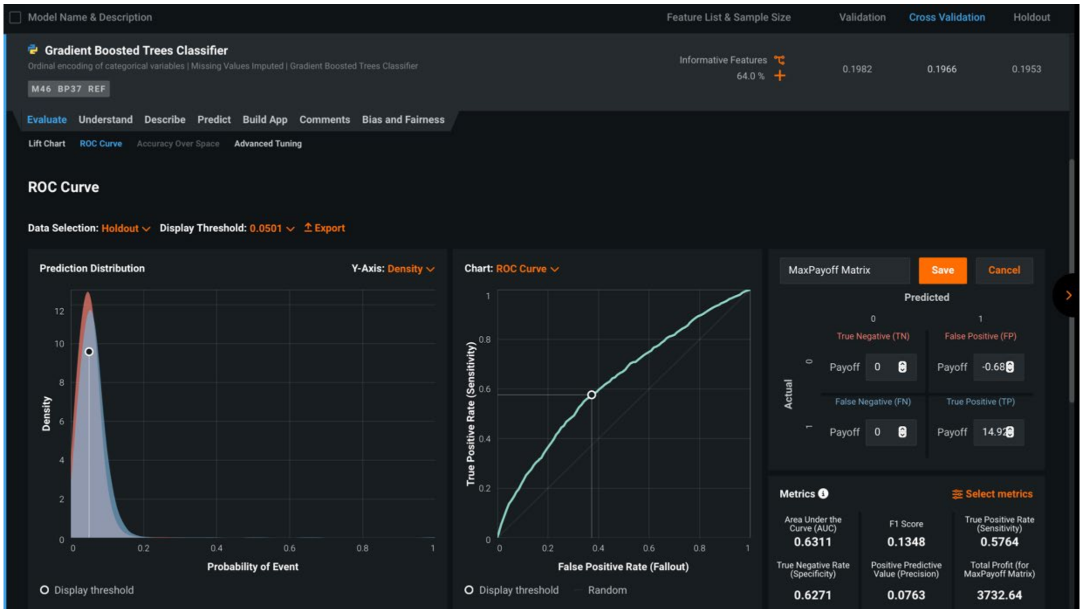
MaxPayoff Matrix Save Cancel

		Predicted 0	Predicted 1
Actual	0	True Negative (TN) Payoff 0	False Positive (FP) Payoff -0.68
	1	False Negative (FN) Payoff 0	True Positive (TP) Payoff 14.92

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6297	0.1309	0.5992
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for MaxPayoff Matrix)**
0.5964	0.0735	3682.12

4. Gradient Boosted Trees Classifier



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Gradient Boosted Trees Classifier M46 BP37 REF	Informative Features 64.0 %	0.1982	0.1966	0.1953

Data Selection: Holdout Display Threshold: 0.0501

MaxPayoff Matrix

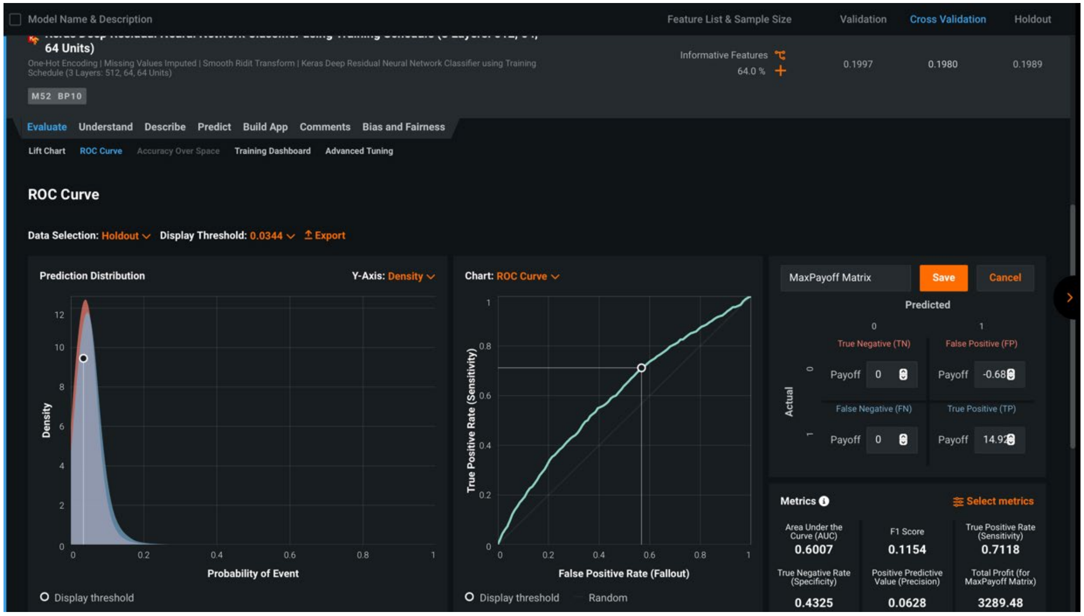
		Predicted 0	Predicted 1
Actual	0	True Negative (TN) Payoff 0	False Positive (FP) Payoff -0.68
Actual	1	False Negative (FN) Payoff 0	True Positive (TP) Payoff 14.92

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6311	0.1348	0.5764

True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for MaxPayoff Matrix)
0.6271	0.0763	3732.64

5. Keras Deep Residual Neural Network Classifier using Training Schedule (3 Layers: 512, 64, 64 Units)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
64 Units) One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Keras Deep Residual Neural Network Classifier using Training Schedule (3 Layers: 512, 64, 64 Units) M52 BP10	Informative Features 64.0 %	0.1997	0.1980	0.1989

Data Selection: Holdout Display Threshold: 0.0344

MaxPayoff Matrix

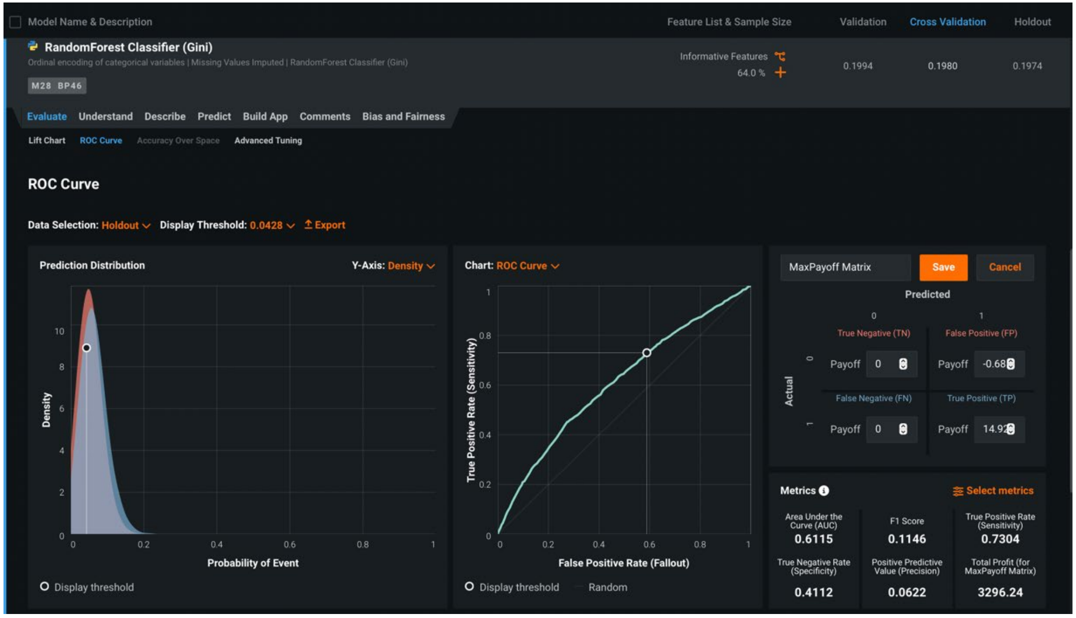
		Predicted 0	Predicted 1
Actual	0	True Negative (TN) Payoff 0	False Positive (FP) Payoff -0.68
Actual	1	False Negative (FN) Payoff 0	True Positive (TP) Payoff 14.92

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6007	0.1154	0.7118

True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for MaxPayoff Matrix)
0.4325	0.0628	3289.48

6. RandomForest Classifier (Gini)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M28 BP46	Informative Features 64.0 %	0.1994	0.1980	0.1974

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart ROC Curve Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout Display Threshold: 0.0428 Export

Chart: ROC Curve

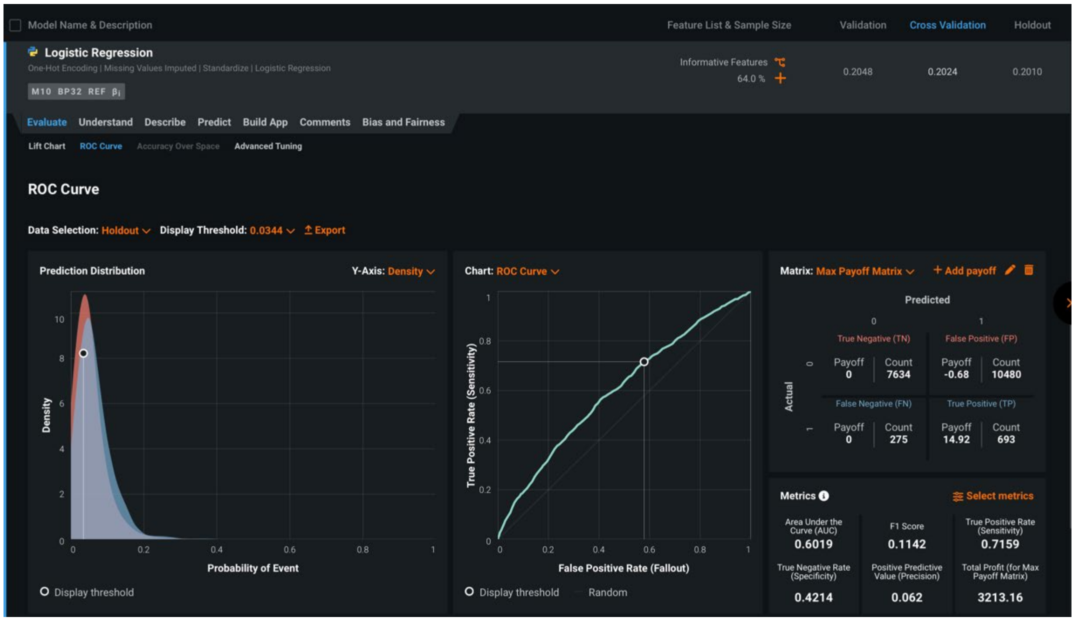
MaxPayoff Matrix Save Cancel

		Predicted 0	Predicted 1
		True Negative (TN)	False Positive (FP)
Actual	0	Payoff 0	Payoff -0.68
		False Negative (FN)	True Positive (TP)
Actual	1	Payoff 0	Payoff 14.92

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6115	0.1146	0.7304
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for MaxPayoff Matrix)**
0.4112	0.0622	3296.24

7. Logistic Regression



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M10 BP32 REF βi	Informative Features 64.0 %	0.2048	0.2024	0.2010

Evaluate Understand Describe Predict Build App Comments Bias and Fairness

Lift Chart ROC Curve Accuracy Over Space Advanced Tuning

ROC Curve

Data Selection: Holdout Display Threshold: 0.0344 Export

Chart: ROC Curve

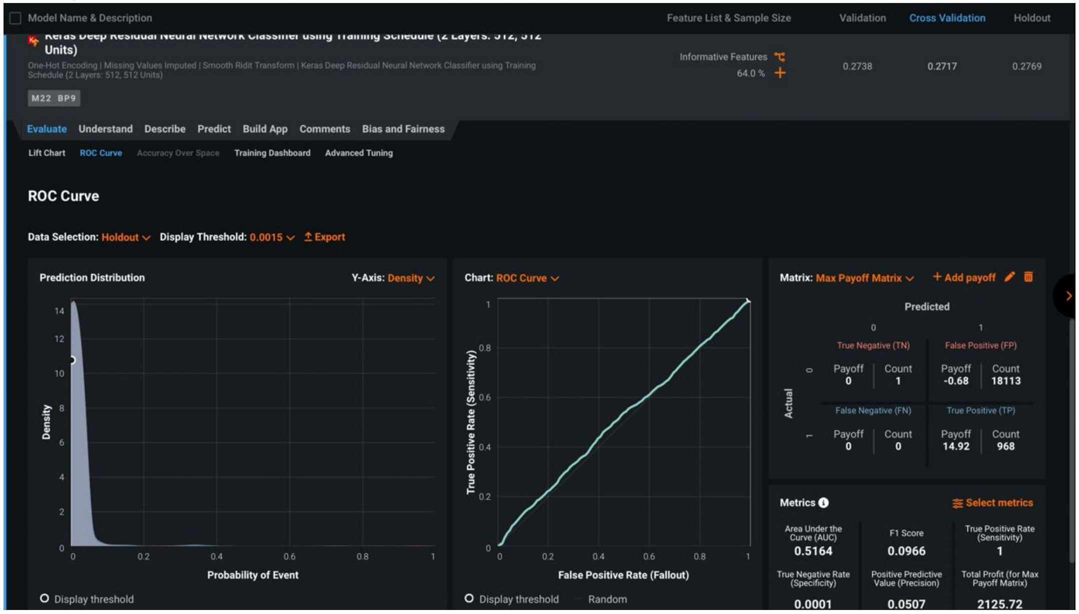
Matrix: Max Payoff Matrix + Add payoff

		Predicted 0		Predicted 1	
		True Negative (TN)		False Positive (FP)	
		Payoff	Count	Payoff	Count
Actual	0	0	7634	-0.68	10480
		False Negative (FN)		True Positive (TP)	
		Payoff	Count	Payoff	Count
Actual	1	0	275	14.92	693

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6019	0.1142	0.7159
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Max Payoff Matrix)**
0.4214	0.062	3213.16

8. Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units) One-Hot Encoding | Missing Values Imputed | Smooth Ridit Transform | Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units) M22 BP9	Informative Features 64.0 %	0.2738	0.2717	0.2769

ROC Curve

Data Selection: Holdout Display Threshold: 0.0015 Export

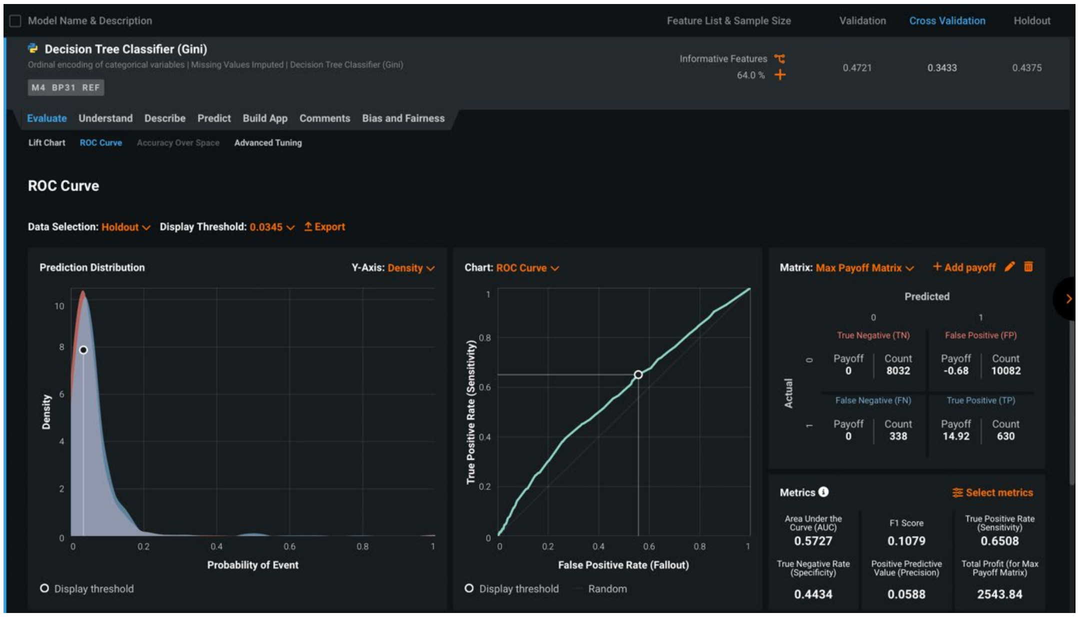
Predicted

Actual	0	1
0	True Negative (TN) Payoff 0 | Count 1	False Positive (FP) Payoff -0.68 | Count 18113
1	False Negative (FN) Payoff 0 | Count 0	True Positive (TP) Payoff 14.92 | Count 968

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.5164	0.0966	1
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Max Payoff Matrix)**
0.0001	0.0507	2125.72

9. Decision Tree Classifier (Gini)



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M4 BP31 REF	Informative Features 64.0 %	0.4721	0.3433	0.4375

ROC Curve

Data Selection: Holdout Display Threshold: 0.0345 Export

Predicted

Actual	0	1
0	True Negative (TN) Payoff 0 | Count 8032	False Positive (FP) Payoff -0.68 | Count 10082
1	False Negative (FN) Payoff 0 | Count 338	True Positive (TP) Payoff 14.92 | Count 630

Metrics Select metrics

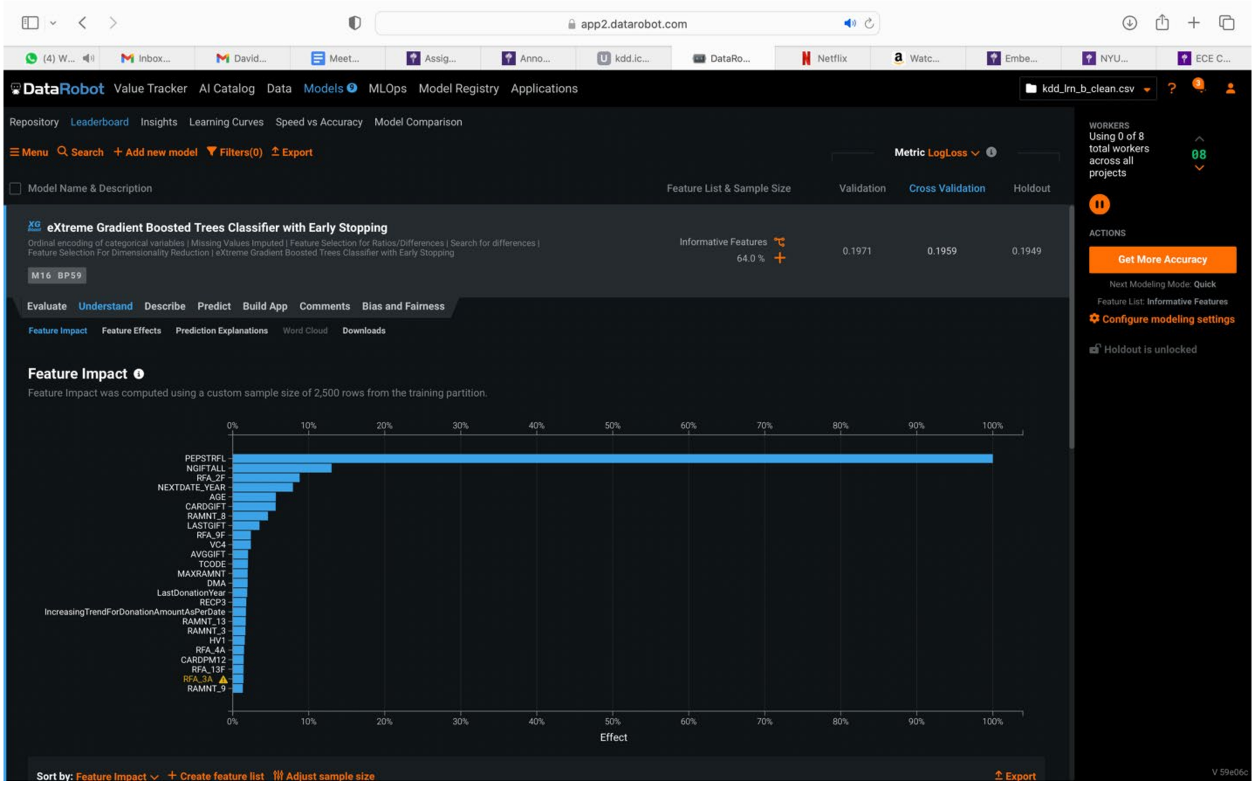
Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.5727	0.1079	0.6508
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Max Payoff Matrix)**
0.4434	0.0588	2543.84

Model Evaluation table for target B

Model	ROC AUC	F1	Recall	Specificity	Precision	Max Payoff
eXtreme Gradient Boosted Trees Classifier withEarly Stopping(Threshold: 0.0519)	0.6328	0.1348	0.5899	0.6173	0.0761	$3805.56
Gradient Boosted Greedy Trees Classifier with Early Stopping(Threshold: 0.0465)	0.6277	0.1272	0.6312	0.5566	0.0707	$3655.04
eXtreme Gradient Boosted Trees Classifier with Early Stopping(Threshold: 0.0501)	0.6297	0.1309	0.5992	0.5964	0.0735	$3682.12
Gradient Boosted Trees Classifier(Threshold: 0.0501)	0.6311	0.1348	0.5764	0.6271	0.0763	$3732.64
Keras Deep Residual Neural Network Classifier using Training Schedule (3 Layers: 512, 64, 64 Units)(Threshold: 0.0344)	0.6007	0.1154	0.7118	0.4325	0.0628	$3289.48
RandomForest Classifier (Gini)(Threshold: 0.0482)	0.6115	0.1146	0.7304	0.4112	0.0622	$3296.24
Logistic Regression(Threshold: 0.0344)	0.6019	0.1142	0.7159	0.4214	0.062	$3213.16
Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units)(Threshold: 0.0015)	0.5164	0.0966	1	0.0001	0.0507	$2125.72
Decision Tree Classifier (Gini)(Threshold: 0.0345)	0.5727	0.1079	0.6508	0.4434	0.0588	$2543.84

The best performing model is Gradient Boosted Trees Classifier with maximum pay off of $3805.56. Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to minimize the donations received . Maximum payoff metric assigns costs and benefits to different types of correct and incorrect predictions (true positives/true negatives and false positives/false negatives) and help evaluate the required profit/losses based on the given case. We need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

Feature impact for target B:



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
eXtreme Gradient Boosted Trees Classifier with Early Stopping Ordinal encoding of categorical variables | Missing Values Imputed | Feature Selection for Ratios/Differences | Search for differences | Feature Selection For Dimensionality Reduction | eXtreme Gradient Boosted Trees Classifier with Early Stopping M16 BP59	Informative Features 64.0 %	0.1971	0.1959	0.1949

Feature Impact

Feature Impact was computed using a custom sample size of 2,500 rows from the training partition.

WORKERS
Using 0 of 8 total workers across all projects

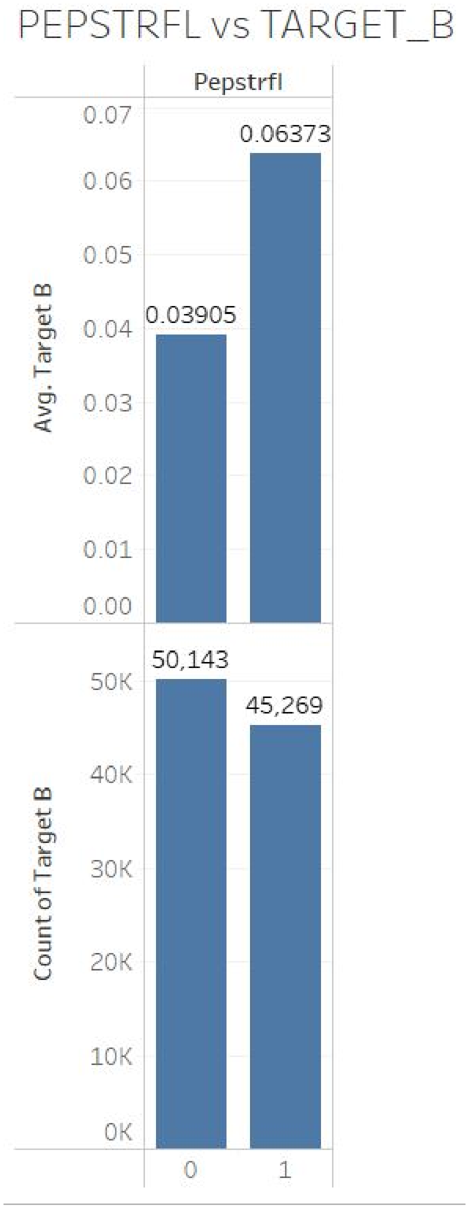
Next Modeling Mode: Quick

Feature List: Informative Features

Holdout is unlocked

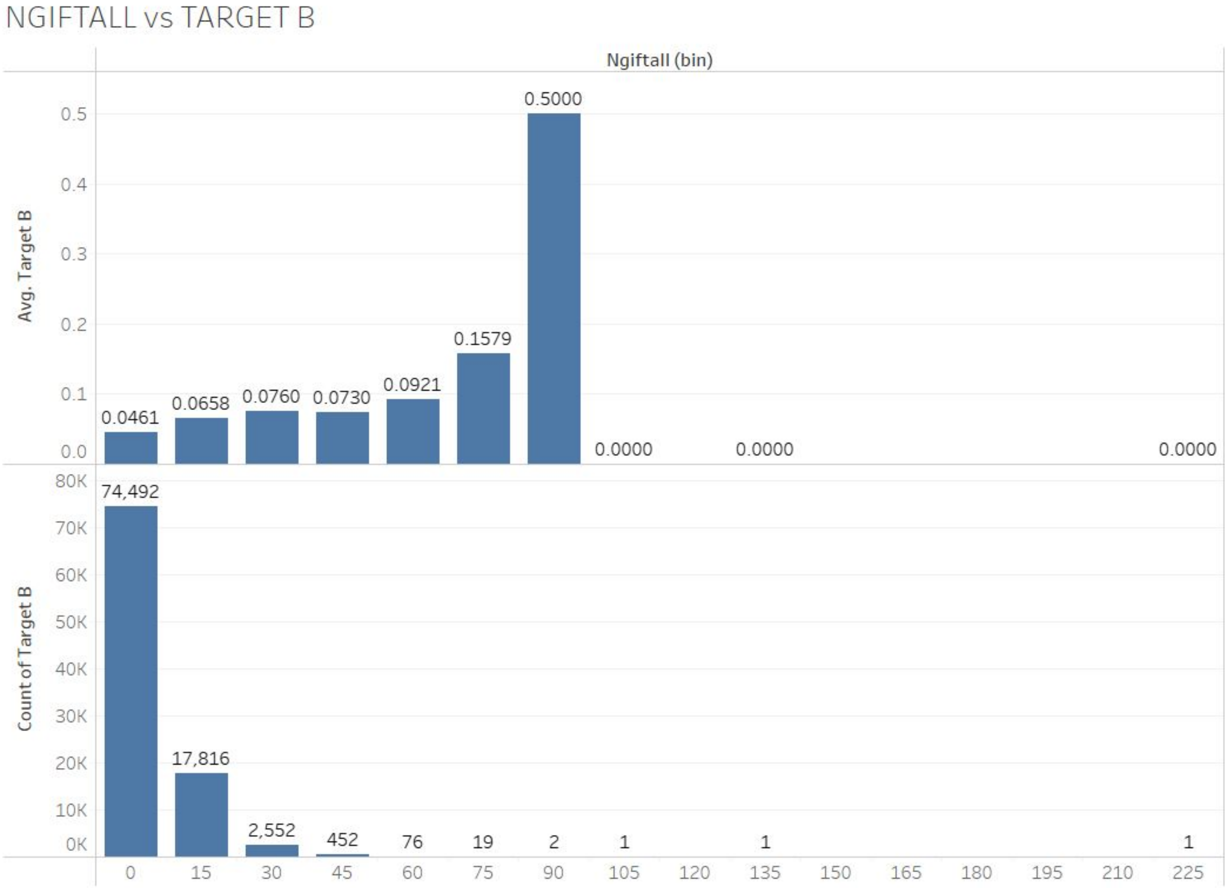
Top predictor for target B:

1. PEPSTRFL



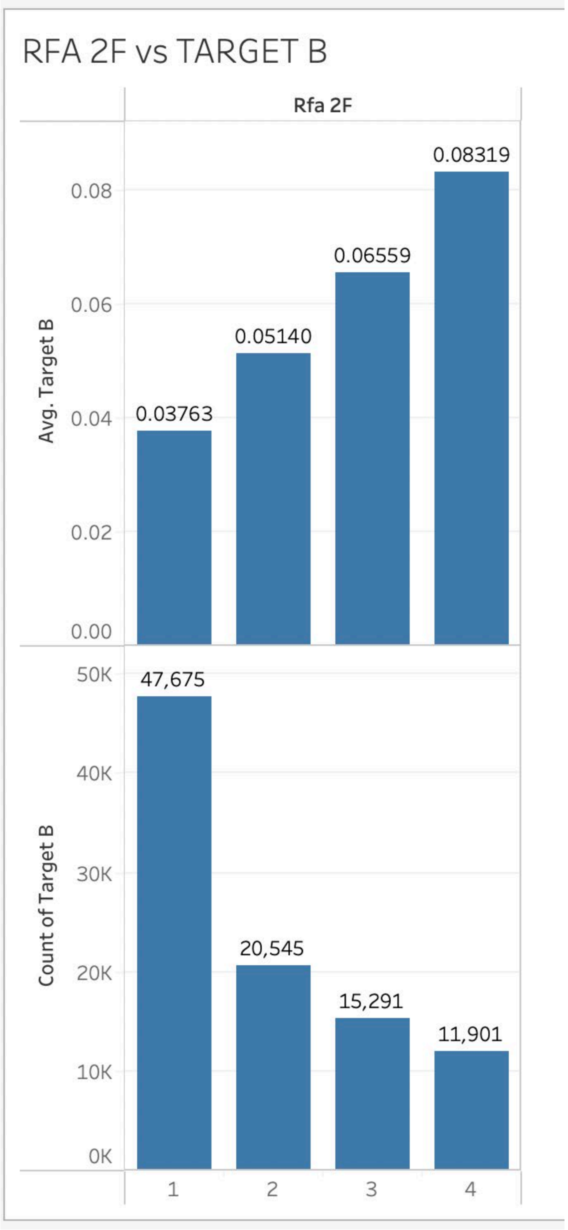
The count for both 0 and 1 is nearly same when compared to the size of the dataset but we can conclude from the above graph that the value for PEPSTRFL at 1(count 45,269) is 1.6 times the average donation rate when compared at PEPSTRFL at 0(count 50,143).

2. NGIFTALL



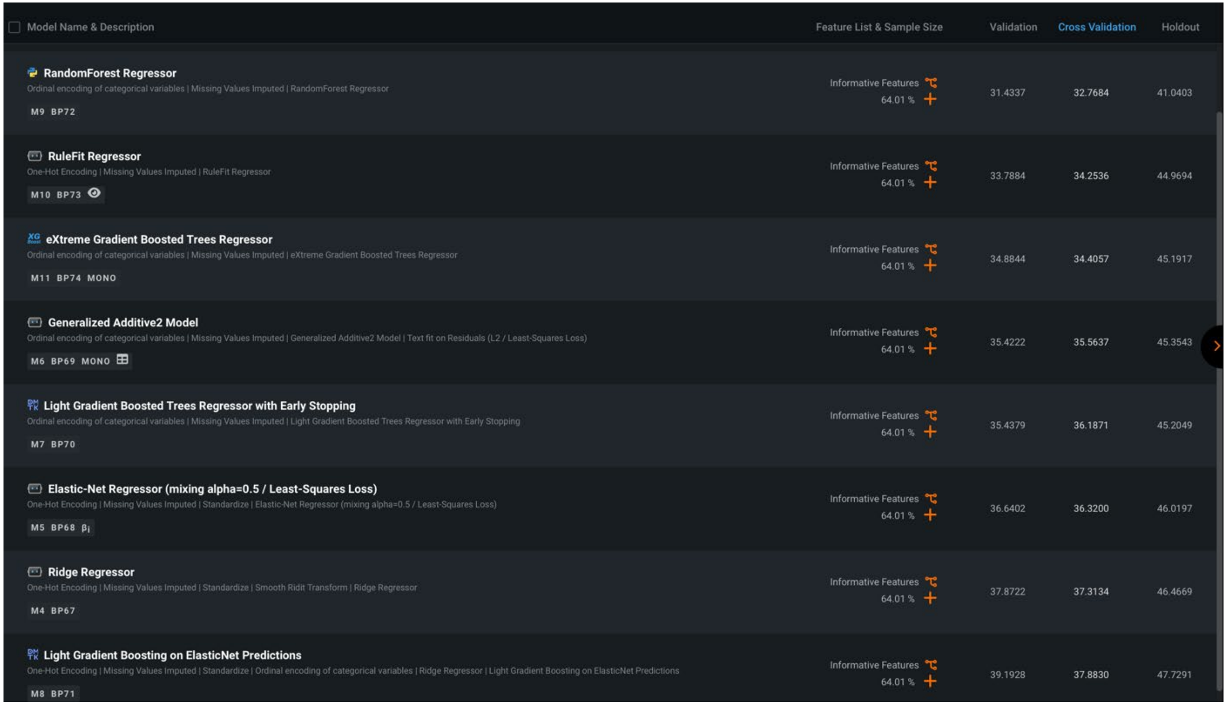
Count of lifetime gift by donor till date is captured in NGIFTALL feature. As can be seen in the above graph higher the count of total gift higher is the donation rate (with few exception). 78% approx. is the range for gift between bin 0-15. Even though the count is low but we can clearly see that from bin of [75-90] to [90-105] we can see a jump of 216% in the average donation rate.

3. RFA 2F



RFA 2F shows status as of 97NK promotion date. F shows the frequency code for RFA_2. As can be seen as the count decreases the frequency is increasing. With count of 47,675 at 3.7% it increases to 83.19% at 11,901 count. Which is a roughly increase of 80%.

Model for target D



Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
RandomForest Regressor Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Regressor M9 BP72	Informative Features 64.01 %	31.4337	32.7684	41.0403
RuleFit Regressor One-Hot Encoding | Missing Values Imputed | RuleFit Regressor M10 BP73	Informative Features 64.01 %	33.7884	34.2536	44.9694
eXtreme Gradient Boosted Trees Regressor Ordinal encoding of categorical variables | Missing Values Imputed | eXtreme Gradient Boosted Trees Regressor M11 BP74 MONO	Informative Features 64.01 %	34.8844	34.4057	45.1917
Generalized Additive2 Model Ordinal encoding of categorical variables | Missing Values Imputed | Generalized Additive2 Model | Text fit on Residuals (L2 / Least-Squares Loss) M6 BP69 MONO	Informative Features 64.01 %	35.4222	35.5637	45.3543
Light Gradient Boosted Trees Regressor with Early Stopping Ordinal encoding of categorical variables | Missing Values Imputed | Light Gradient Boosted Trees Regressor with Early Stopping M7 BP70	Informative Features 64.01 %	35.4379	36.1871	45.2049
Elastic-Net Regressor (mixing alpha=0.5 / Least-Squares Loss) One-Hot Encoding | Missing Values Imputed | Standardize | Elastic-Net Regressor (mixing alpha=0.5 / Least-Squares Loss) M5 BP68 βi	Informative Features 64.01 %	36.6402	36.3200	46.0197
Ridge Regressor One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Ridge Regressor M4 BP67	Informative Features 64.01 %	37.8722	37.3134	46.4669
Light Gradient Boosting on ElasticNet Predictions One-Hot Encoding | Missing Values Imputed | Standardize | Ordinal encoding of categorical variables | Ridge Regressor | Light Gradient Boosting on ElasticNet Predictions M8 BP71	Informative Features 64.01 %	39.1928	37.8830	47.7291

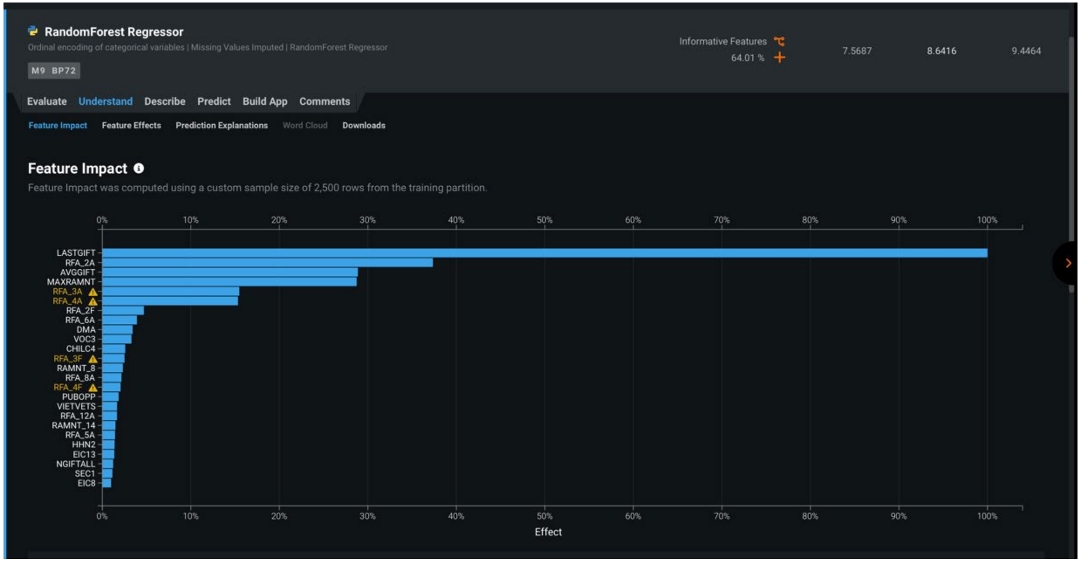
Model Evaluation table for target D

Model	R2	RMSE	MAE	MAPE
RandomForest Regressor	0.3924	9.4464	4.4351	41.0403
Light Gradient Boosted Trees Regressor with Early Stopping	0.3541	9.7394	4.6931	45.2049
Generalized Additive2 Model	0.3449	9.8085	4.7916	45.3543
eXtreme Gradient Boosted Trees Regressor	0.3369	9.8680	4.6103	45.1917

RuleFit Regressor	0.3040	10.1101	4.7499	44.9694
Ridge Regressor	0.3623	9.6775	4.8498	46.4669
Light Gradient Boosting on ElasticNet Predictions	0.3367	9.8699	4.9479	47.7291
Elastic-Net Regressor (mixing alpha=0.5 / Least-Squares Loss)	0.3406	9.8405	4.7682	46.0197

Here the best performing model is Random Forest Regressor. The best metric to evaluate the same is R2. As R2 metric helps to understand how well the data fits the model . As R2 is a standardize metric with values ranging between 0 and 1 it makes it easier to understand. In our case the best evaluated model Random Forest Regressor has the R2 value of 0.3924.

Feature impact for target D:



RandomForest Regressor

Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Regressor

M9 BP72

Informative Features 64.01 %

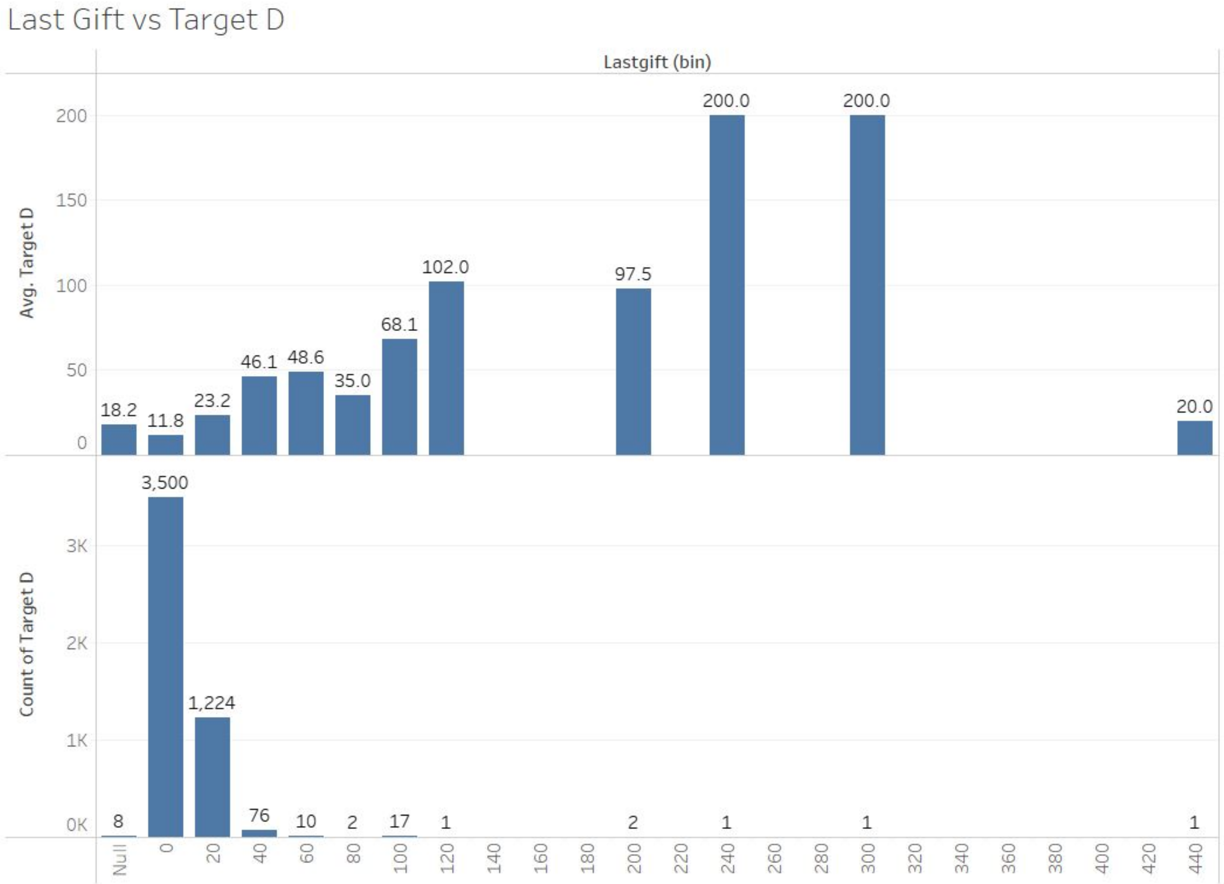
7.5687 8.6416 9.4464

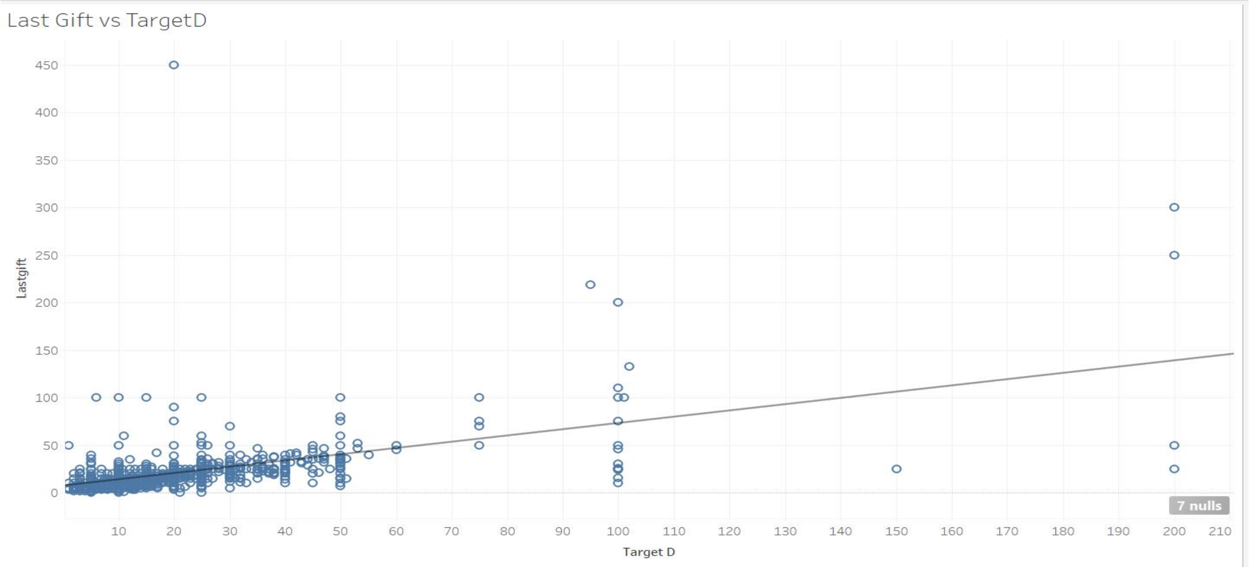
Feature Impact

Feature Impact was computed using a custom sample size of 2,500 rows from the training partition.

Top predictor for target D:

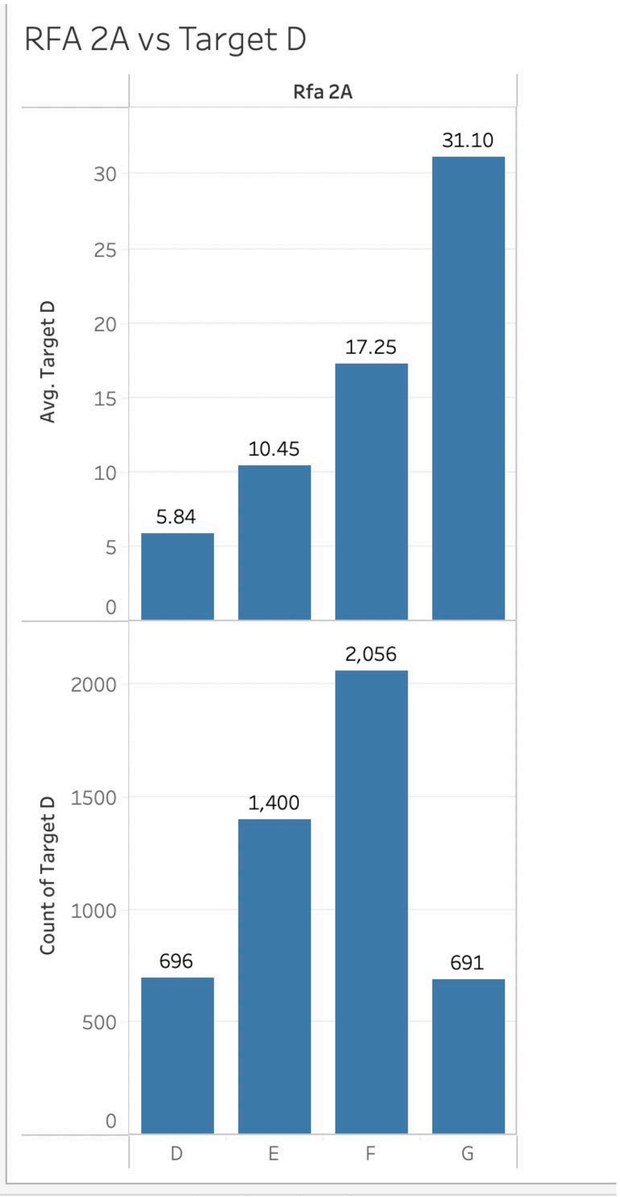
1. Last Gift


Last Gift vs Target D


Last Gift vs TargetD

This feature indicates the dollar amount of the last gift received. The last donation amount is between $0-$40 for 97.54% of total donors. There is an increase in the average donation amount when compared to the previous donations. Average amount of last gift and average donation amount is close.

2. RFA 2A



RFA 2 indicates Donor's RFA status as of 97NK promotion date. Where RFA 2A indicates donation amount code for RFA_2. Where the count is highest at F for 2056 with percentage of 17.25% which is the second highest. Whereas the highest percentage is observed at G for a count of 691 for 31.10%. Which is 13.25% higher.

Where D-G indicate the below:

D=$5.00 - $9.99

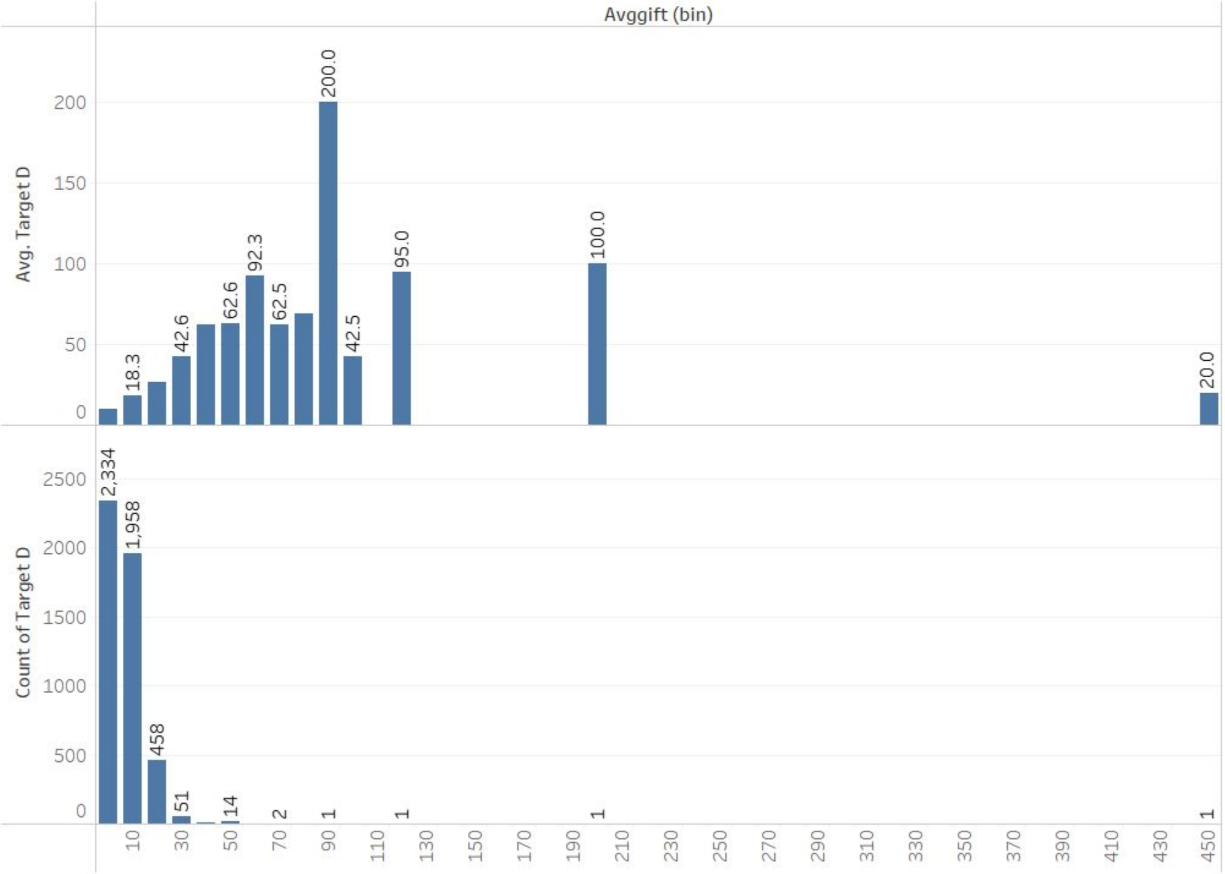
E=$10.00 - $14.99

F=$15.00 - $24.99

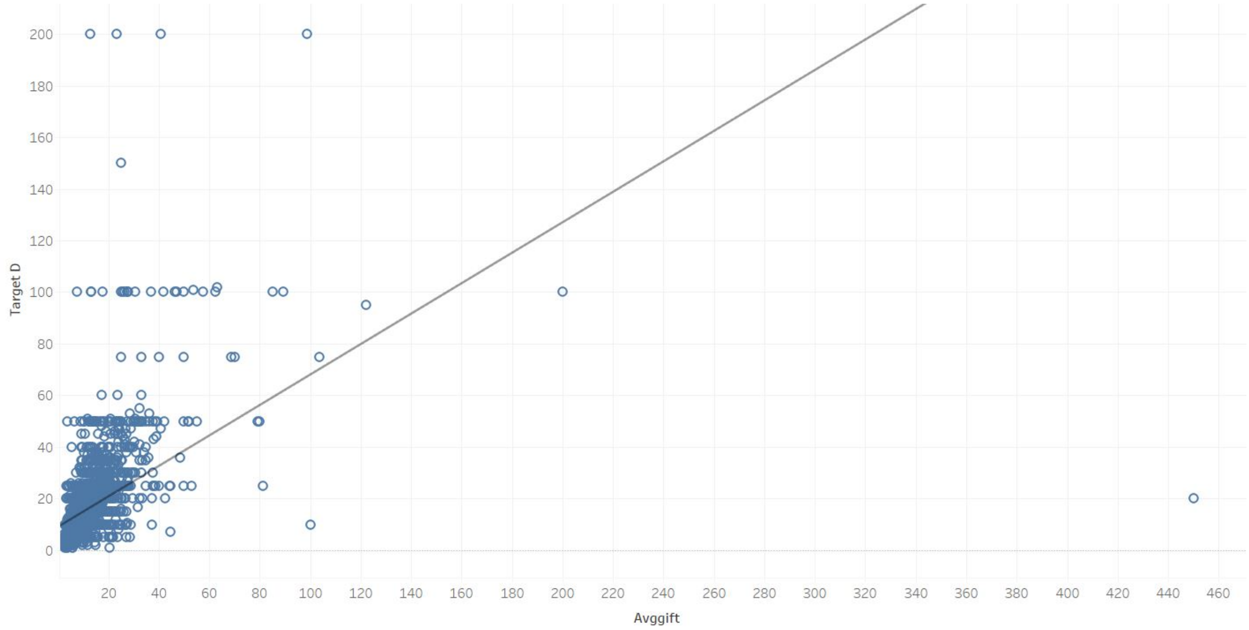
G=$25.00 and above

3. Avg gift

Avg Gift vs TargetD



Avg Gift vs Target D



This feature indicates Average dollar amount of gifts to date . Where bin [10-30] has the highest count and indicates the total of 18.3% . Highest percent is observed at bin 90 for 200% but this is an exception as count is 1. Highest percentage is observed at bin count of 50 for 62.6% at a count of 14 followed by bin 30 for 42.16% at a count of 51.

Top 20 predicted donors for Threshold of 0.0519

row_id	Prediction_B	PredictedLabel	Prediction_D	Donation forcast
57111	0.23670886	1	56.491831	13.3721169
44742	0.24338761	1	44.2244533	10.763684
18554	0.10872028	1	91.9603438	9.99795433
52049	0.15477118	1	58.3347897	9.02854424
54798	0.29127762	1	28.9267083	8.42570275
12030	0.07533555	1	97.7892548	7.36700776
17428	0.08979941	1	79.7825119	7.16442262
28950	0.0701474	1	100.703152	7.06406428
33604	0.07751193	1	90.853161	7.04220376
20474	0.10184731	1	68.2445526	6.9505241
56738	0.07471199	1	90.4831346	6.76017505
19025	0.25974491	1	25.4261256	6.60430671
4488	0.07731543	1	84.3728413	6.52332251
2742	0.08885829	1	71.9215429	6.39082532
5691	0.08473988	1	74.7711235	6.33609594
38764	0.23081344	1	27.3250512	6.30698907
45069	0.06114031	1	102.7057	6.27945834
34264	0.06343739	1	95.2653021	6.04338187
17480	0.06756824	1	89.2258907	6.0288364