Case Study – Predict Potential Donors

Business Case: -

We're in the process of developing a predictive model aimed at determining the inclination of donors to contribute. The objective is to assist non-profit organizations in optimizing their resource allocation by targeting individuals likely to donate. By leveraging historical data from Paralyzed Veterans of America (PVA), encompassing donor demographics, contribution history, and related factors, we aim to accurately identify potential donors. This approach enables the organization to minimize outreach expenses while maximizing the effectiveness of their campaigns.

Payoff Matrix: -

As mentioned in the provided documents,

- The average amount of donation is $15.61.
- Package cost (including the mail cost) is $0.68 per piece mailed.

Assumption –
100% people who are willing to donate and are reached out through the campaign make the donation.

True Positive - Model predicts that if reached out the person would donate, and the person donates.
cost = amount of donation made - mailing cost = 15.61 - 0.68 = $14.93

False Positive - Model predicts that the person would donate and hence the mail is sent to them, but they don't donate.
cost = mailing cost = $-0.68

True Negative - Model predicts that the person would not donate and hence we do not reach out to them.
Cost = $0

False Negative - Model predicts that the person would not donate and hence doesn't send the campaign to the person, but the person donates.
cost = $0
This could also be considered as a missed opportunity cost ($-14.92)

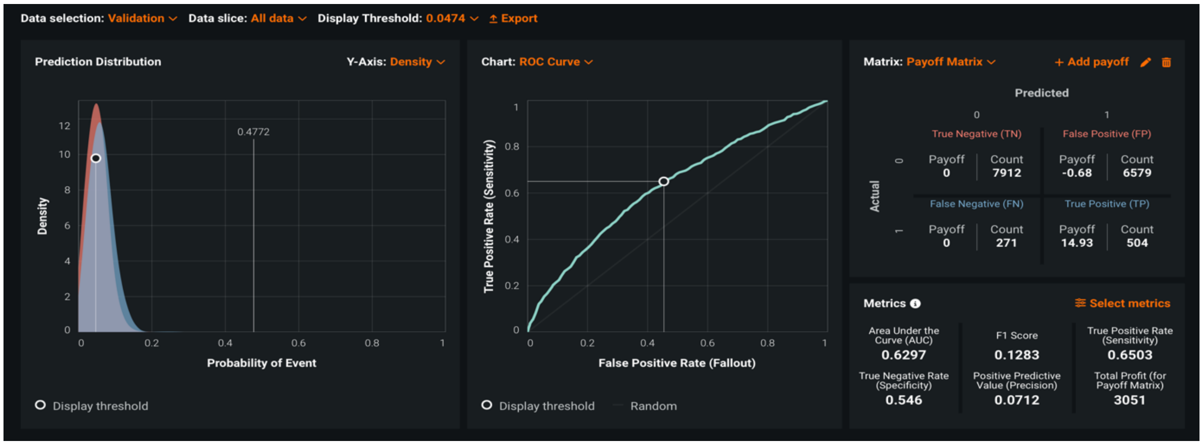
	Predicted 0	**Predicted 1**
Actual 0	$0	$-0.68
Actual 1	$0	$14.93

1. **A clear explanation of how you handled target leaks.**
   To handle target leaks we have followed the below steps-
   - For the classification model build on TARGET_B we have removed TARGET_D from the feature set as if we know the amount of donation made (TARGET_D) we know the value for TARGET_B which signifies if the donation is made or not.
   - For the regression model build on TARGET_D we have first removed all rows which have value of TARGET_B = 0 which is for the scenario where no donation is made and then we have removed TARGET_B. This means the model would be trained only for values where a donation was made and would hence help to predict the amount of donation.

2. **A summary table of models evaluated criterion for best model selection and the choice of the best model for each – the probability of a gift (TARGET_B) and the predicted donation amount (TARGET_D)**

   Models for TARGET_B –

   Light Gradient Boosted Trees Classifier with Early Stopping



Data selection: Validation   Data slice: All data   Display Threshold: 0.0474   Export

Prediction Distribution   Y-Axis: Density

Chart: ROC Curve

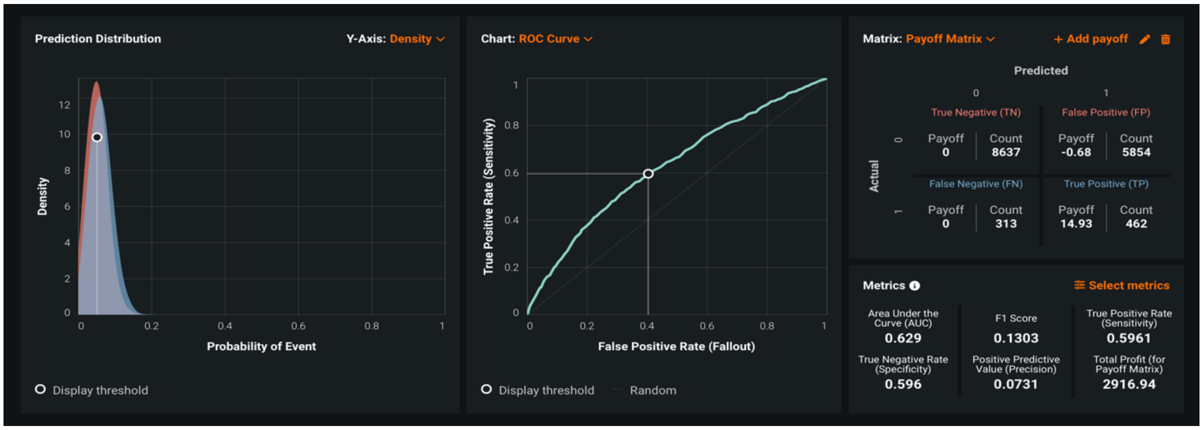
Matrix: Payoff Matrix   + Add payoff

		Predicted 0		Predicted 1	
		True Negative (TN)		False Positive (FP)	
Actual 0		Payoff 0	Count 7912	Payoff -0.68	Count 6579
		False Negative (FN)		True Positive (TP)	
Actual 1		Payoff 0	Count 271	Payoff 14.93	Count 504

Metrics   Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6297	0.1283	0.6503
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Payoff Matrix)**
0.546	0.0712	3051

   eXtreme Gradient Boosted Trees Classifier with Early Stopping



Prediction Distribution   Y-Axis: Density

Chart: ROC Curve

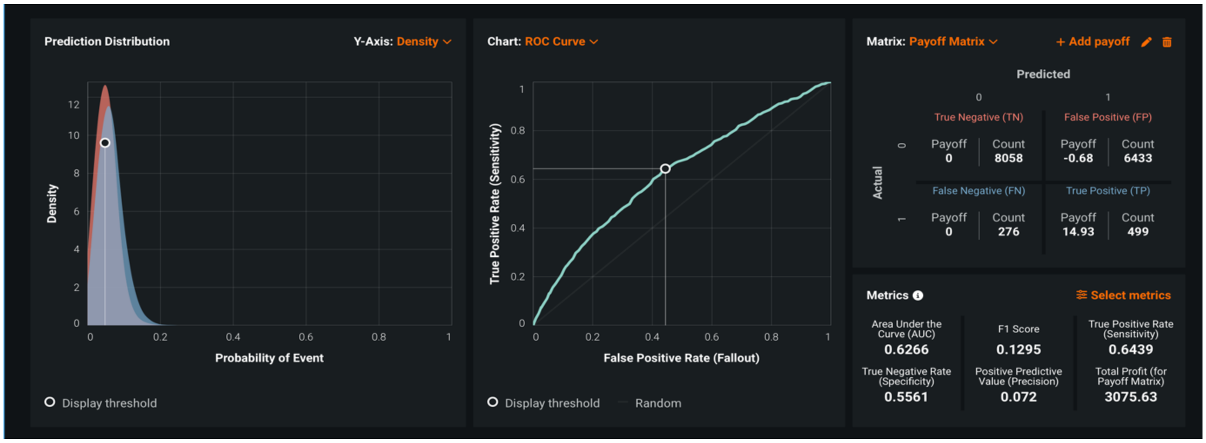
Matrix: Payoff Matrix   + Add payoff

		Predicted 0		Predicted 1	
		True Negative (TN)		False Positive (FP)	
Actual 0		Payoff 0	Count 8637	Payoff -0.68	Count 5854
		False Negative (FN)		True Positive (TP)	
Actual 1		Payoff 0	Count 313	Payoff 14.93	Count 462

Metrics   Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.629	0.1303	0.5961
True Negative Rate (Specificity)	**Positive Predictive Value (Precision)**	**Total Profit (for Payoff Matrix)**
0.596	0.0731	2916.94

Light Gradient Boosting on ElasticNet Predictions



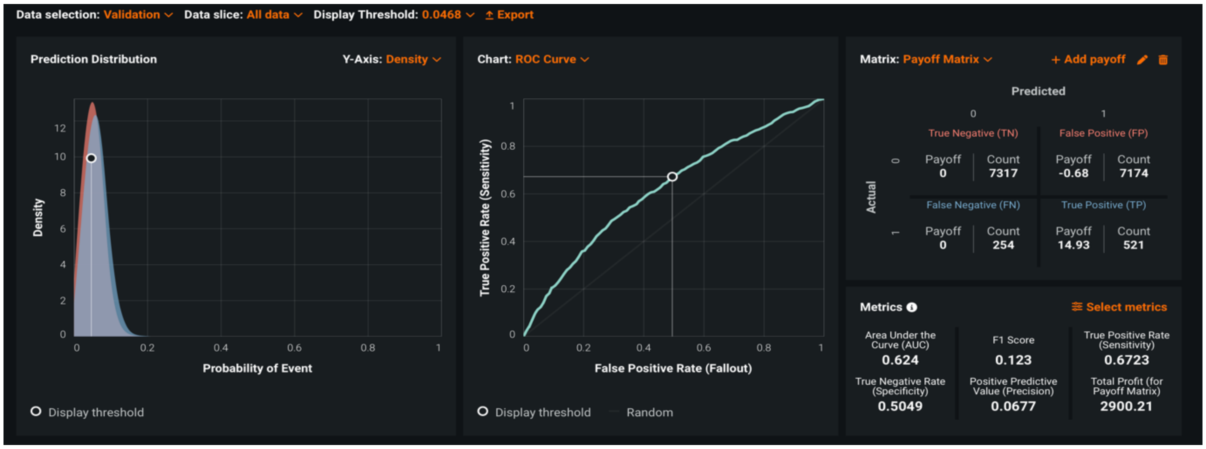
Predicted

Actual	0	1
	True Negative (TN)	False Positive (FP)
0	Payoff 0 · Count 8058	Payoff -0.68 · Count 6433
	False Negative (FN)	True Positive (TP)
1	Payoff 0 · Count 276	Payoff 14.93 · Count 499

Metrics ⓘ Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6266	0.1295	0.6439
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.5561	0.072	3075.63

Elastic-Net Classifier (mixing alpha=0.5/Binomial Deviance)


Data selection: Validation Data slice: All data Display Threshold: 0.0468 Export

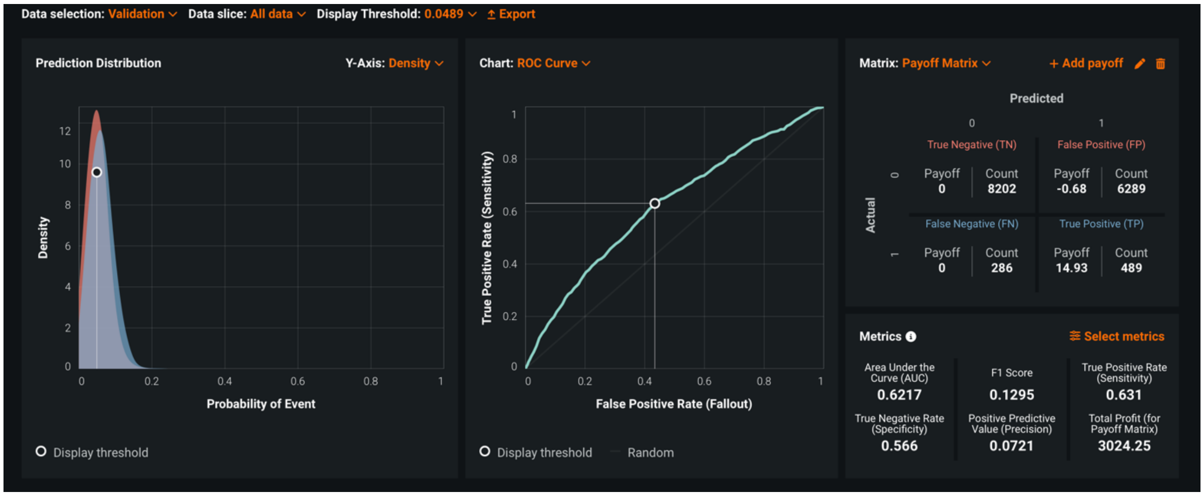
Predicted

Actual	0	1
	True Negative (TN)	False Positive (FP)
0	Payoff 0 · Count 7317	Payoff -0.68 · Count 7174
	False Negative (FN)	True Positive (TP)
1	Payoff 0 · Count 254	Payoff 14.93 · Count 521

Metrics ⓘ Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.624	0.123	0.6723
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.5049	0.0677	2900.21

Elastic-Net Classifier (L2/Binomial Deviance)


Data selection: Validation Data slice: All data Display Threshold: 0.0489 Export

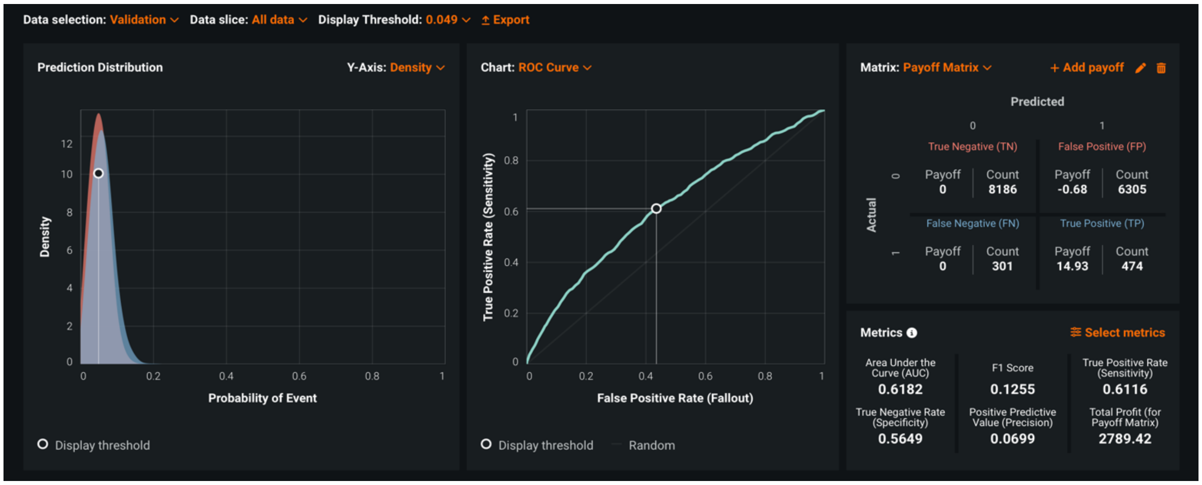
Predicted

Actual	0	1
	True Negative (TN)	False Positive (FP)
0	Payoff 0 · Count 8202	Payoff -0.68 · Count 6289
	False Negative (FN)	True Positive (TP)
1	Payoff 0 · Count 286	Payoff 14.93 · Count 489

Metrics ⓘ Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6217	0.1295	0.631
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.566	0.0721	3024.25

Generalized Additive2 model



Data selection: Validation Data slice: All data Display Threshold: 0.049 Export

Prediction Distribution Y-Axis: Density

Chart: ROC Curve

True Positive Rate (Sensitivity)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
False Positive Rate (Fallout)
Display threshold
Random

Matrix: Payoff Matrix + Add payoff

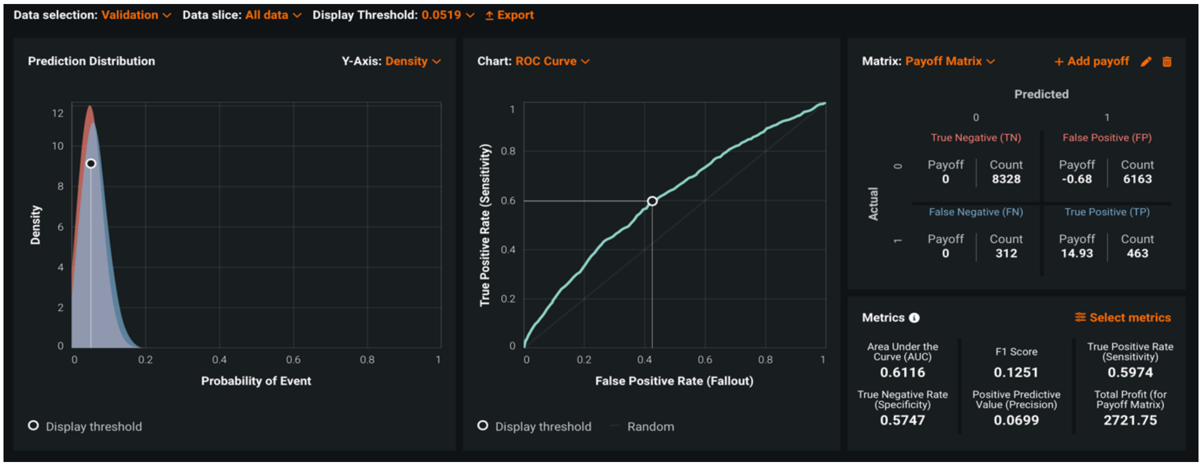
Predicted

Actual	0		1	
	True Negative (TN)		False Positive (FP)	
0	Payoff 0	Count 8186	Payoff -0.68	Count 6305
	False Negative (FN)		True Positive (TP)	
1	Payoff 0	Count 301	Payoff 14.93	Count 474

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6182	0.1255	0.6116
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.5649	0.0699	2789.42

RandomForest Classifier (Gini)



Data selection: Validation Data slice: All data Display Threshold: 0.0519 Export

Prediction Distribution Y-Axis: Density

Chart: ROC Curve

True Positive Rate (Sensitivity)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
False Positive Rate (Fallout)
Display threshold
Random

Matrix: Payoff Matrix + Add payoff

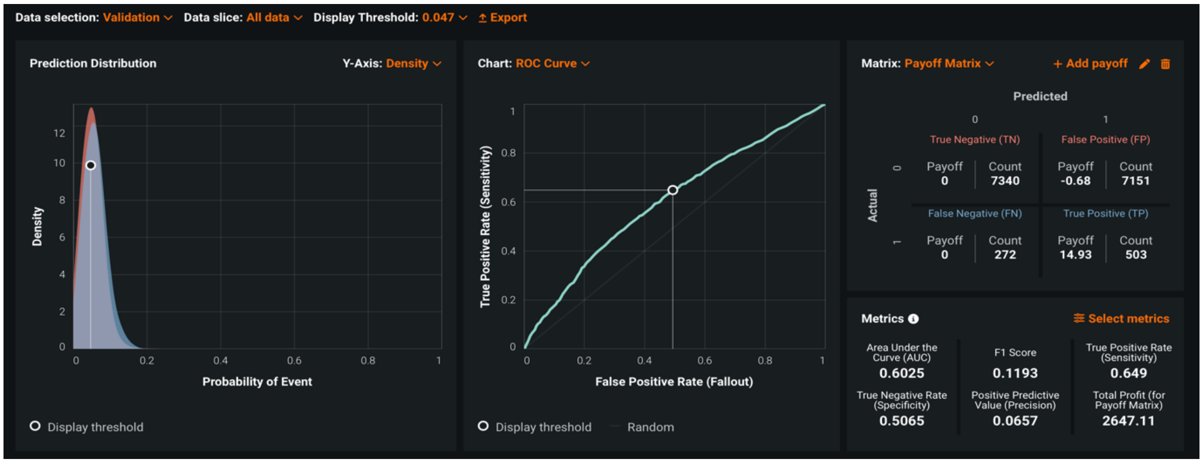
Predicted

Actual	0		1	
	True Negative (TN)		False Positive (FP)	
0	Payoff 0	Count 8328	Payoff -0.68	Count 6163
	False Negative (FN)		True Positive (TP)	
1	Payoff 0	Count 312	Payoff 14.93	Count 463

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6116	0.1251	0.5974
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.5747	0.0699	2721.75

RuleFit Classifier



Data selection: Validation Data slice: All data Display Threshold: 0.047 Export

Prediction Distribution Y-Axis: Density

Chart: ROC Curve

True Positive Rate (Sensitivity)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
False Positive Rate (Fallout)
Display threshold
Random

Matrix: Payoff Matrix + Add payoff

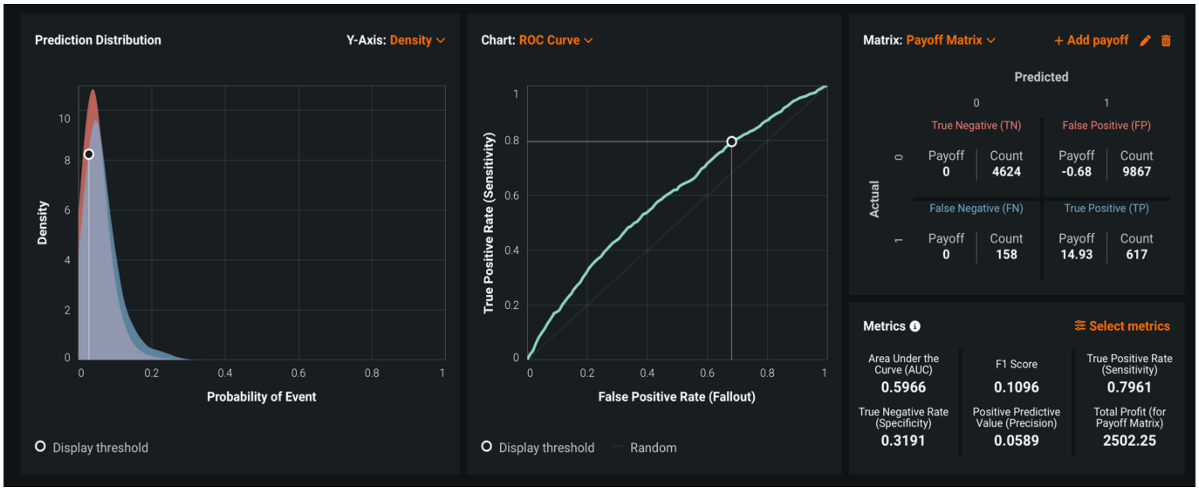
Predicted

Actual	0		1	
	True Negative (TN)		False Positive (FP)	
0	Payoff 0	Count 7340	Payoff -0.68	Count 7151
	False Negative (FN)		True Positive (TP)	
1	Payoff 0	Count 272	Payoff 14.93	Count 503

Metrics Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6025	0.1193	0.649
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.5065	0.0657	2647.11

Keras Slim Residual Neural Network Classifier using Training Schedule (1 Layer: 64 units)



Matrix: Payoff Matrix + Add payoff

		Predicted 0		Predicted 1	
		True Negative (TN)		False Positive (FP)	
		Payoff	Count	Payoff	Count
Actual	0	0	4624	-0.68	9867
		False Negative (FN)		True Positive (TP)	
		Payoff	Count	Payoff	Count
Actual	1	0	158	14.93	617

Metrics ☰ Select metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.5966	0.1096	0.7961
True Negative Rate (Specificity)	Positive Predictive Value (Precision)	Total Profit (for Payoff Matrix)
0.3191	0.0589	2502.25

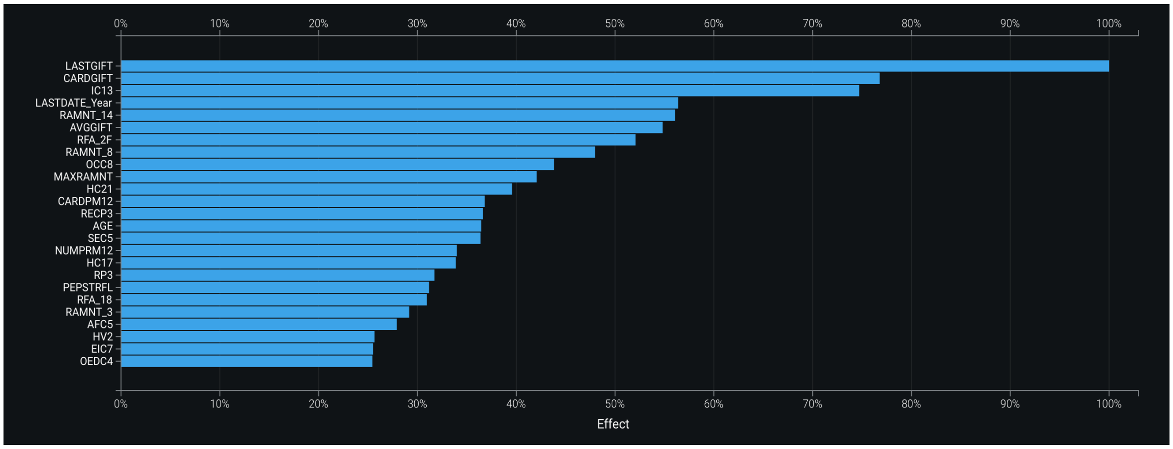
Summary table of models for TARGET_B –

Model	ROC AUC	F1	Recall	Specificity	Precision	Max Payoff
Light Gradient Boosted Trees Classifier with Early Stopping	0.6297	0.1283	0.6503	0.546	0.0712	$3051
eXtreme Gradient Boosted Trees Classifier with Early Stopping	0.629	0.1303	0.5961	0.596	0.0731	$2916.94
Light Gradient Boosting on ElasticNet Predictions	0.6266	0.1295	0.6439	0.5561	0.072	$3075.63
Elastic-Net Classifier (mixing alpha=0.5/Binomial Deviance)	0.624	0.123	0.6723	0.5049	0.0677	$2900.21
Elastic-Net Classifier (L2/Binomial Deviance)	0.6217	0.1295	0.631	0.566	0.0721	$3024.25
Generalized Additive2 model	0.6128	0.1255	0.6116	0.5649	0.0699	$2789.42

RandomForest Classifier (Gini)	0.6116	0.1251	0.5974	0.5747	0.0699	$2721.75
RuleFit Classifier	0.6025	0.1193	0.649	0.5065	0.0657	$2647.11
Keras Slim Residual Neural Network Classifier using Training Schedule(1 Layer: 64 units)	0.5966	0.1096	0.7961	0.3191	0.0589	$2502.25

The best performing model for TRAGET_B is the Light Gradient Boosting on ElasticNet Predictions model as it gives the highest payoff of $3075.63. The metric used here for consideration is the maximum payoff as it evaluates and assigns values to all types of correct and incorrect predictions (true positive, true negative, false positive and false negative) and hence provide a better estimation of the profit or loss made. Hence, our best model is Light Gradient Boosting on ElasticNet Predictions and the metric used for evaluation is the maximum payoff metric.

Feature impact for model for TARGET_B –

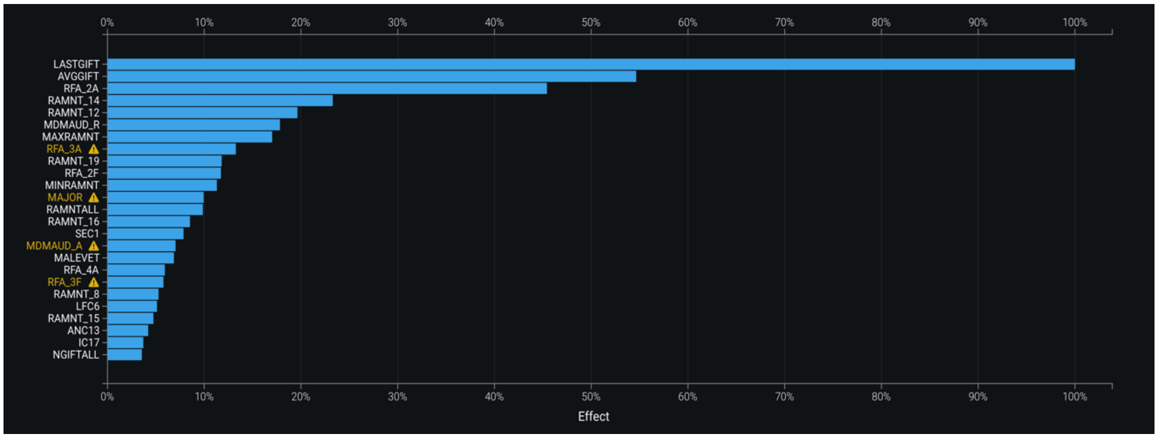


Summary table of models for TARGET_D –

Model	RMSE	MAE	MAPE	R2
Light Gradient Boosted Trees Regressor with early stopping	8.8689	4.3896	36.1837	0.5146
eXtreme Gradient Boosted Trees Regressor	8.9965	4.2511	34.6524	0.4963
RandomForest Regressor	9.0183	4.2078	33.2241	0.4987
Light Gradient Boosting on ElasticNet Predictions	9.1440	4.5430	37.5429	0.4709
Generalized Additive2 model	9.2165	4.5743	38.0471	0.4718
Ridge Regressor	9.2379	4.5762	37.1183	0.4634
Elastic-Net Regressor (mixing alpha=0.5/Least-Squares Loss)	9.4715	4.5337	37.1136	0.4110
RuleFit Regressor	9.8949	4.5920	35.4756	0.3835

The best model for TARGET_D is the Light Gradient Boosted Trees Regressor with early stopping as it has the highest R2 score of 0.5146. The metric used here for evaluation is R2 as it helps us measure the proportion of variance in the target variable as explained by the model. Also, it is a standardized metric ranging from 0 to 1 which makes it easier to understand. Hence our best model is Light Gradient Boosted Trees Regressor with early stopping and our metric used for evaluation is R2.

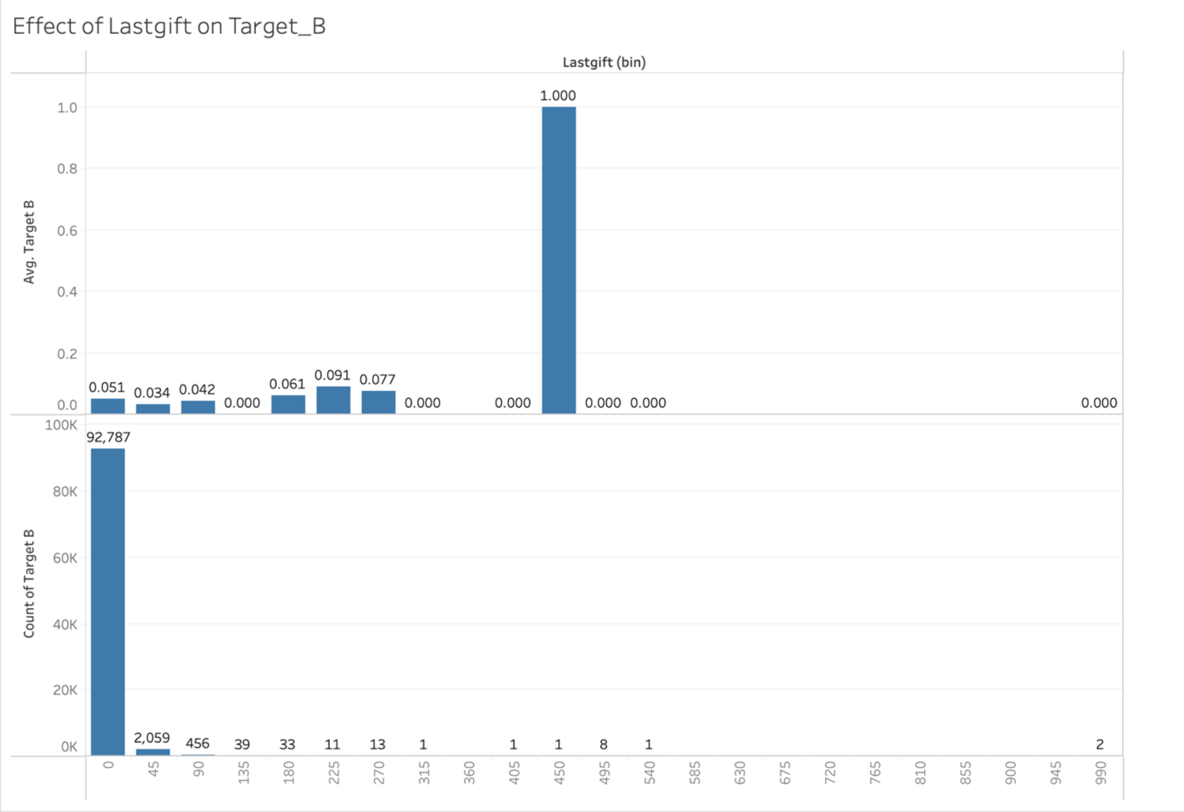
Feature impact for Model for TARGET_D –



3. **A visualization and 1-2 sentence summary of the most important predictors for target B and target D**

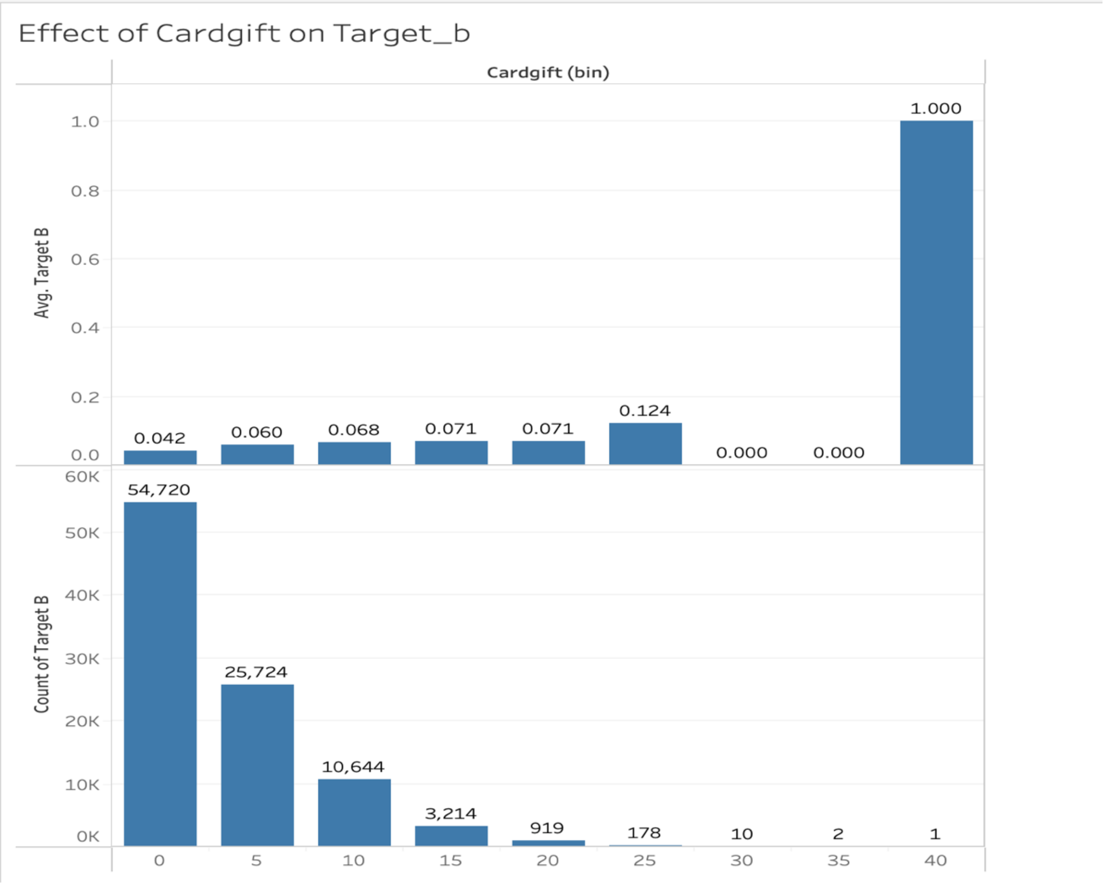
Most important predictors for TARGET_B-

1. Lastgift



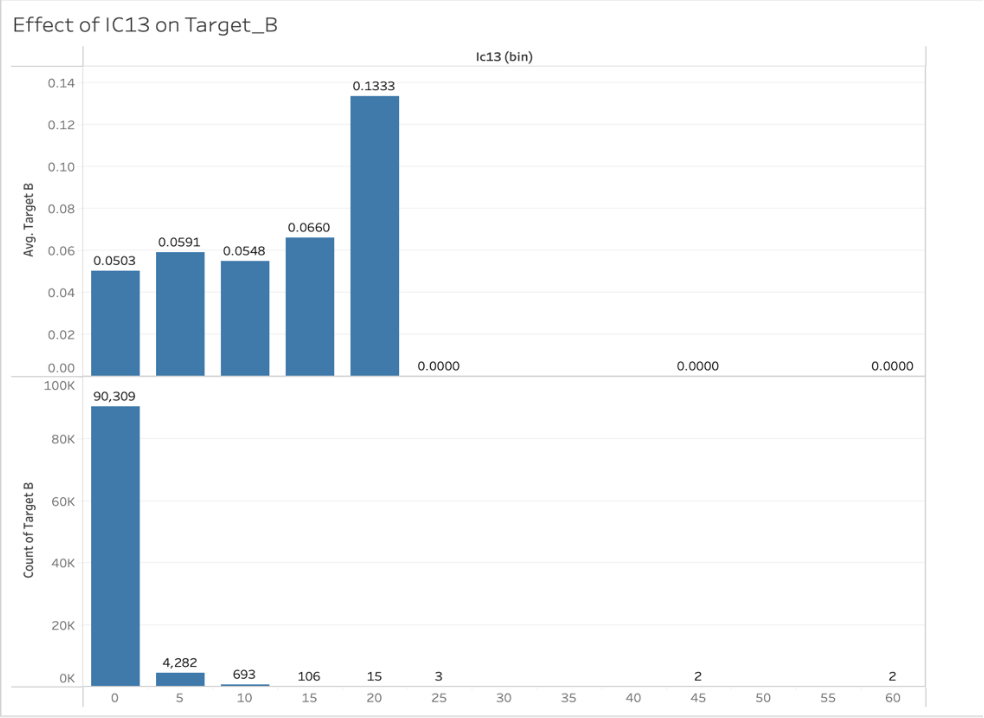
Even though the count in the 45 bin is higher (2059), the percentage value for that is 3.4 only. The bin for 225 has only 11 records but still records a percentage of 9.1%. The highest percentage is 100 for 450 bin but that is an exceptional case as we have only 1 value for that.

2. Cardgift



Cardgift represents number of lifetime gifts to card promotions to date. The count for cardgift in the bin of 5 is highest (25724) but its percentage is 6% which is almost 2 times less than percentage for the bin of 25 which has only 178 records which is almost 145 times lesser than the count in the bin of 5. Even though the count decreases as the bin for cardgift increases, the percentage increases.

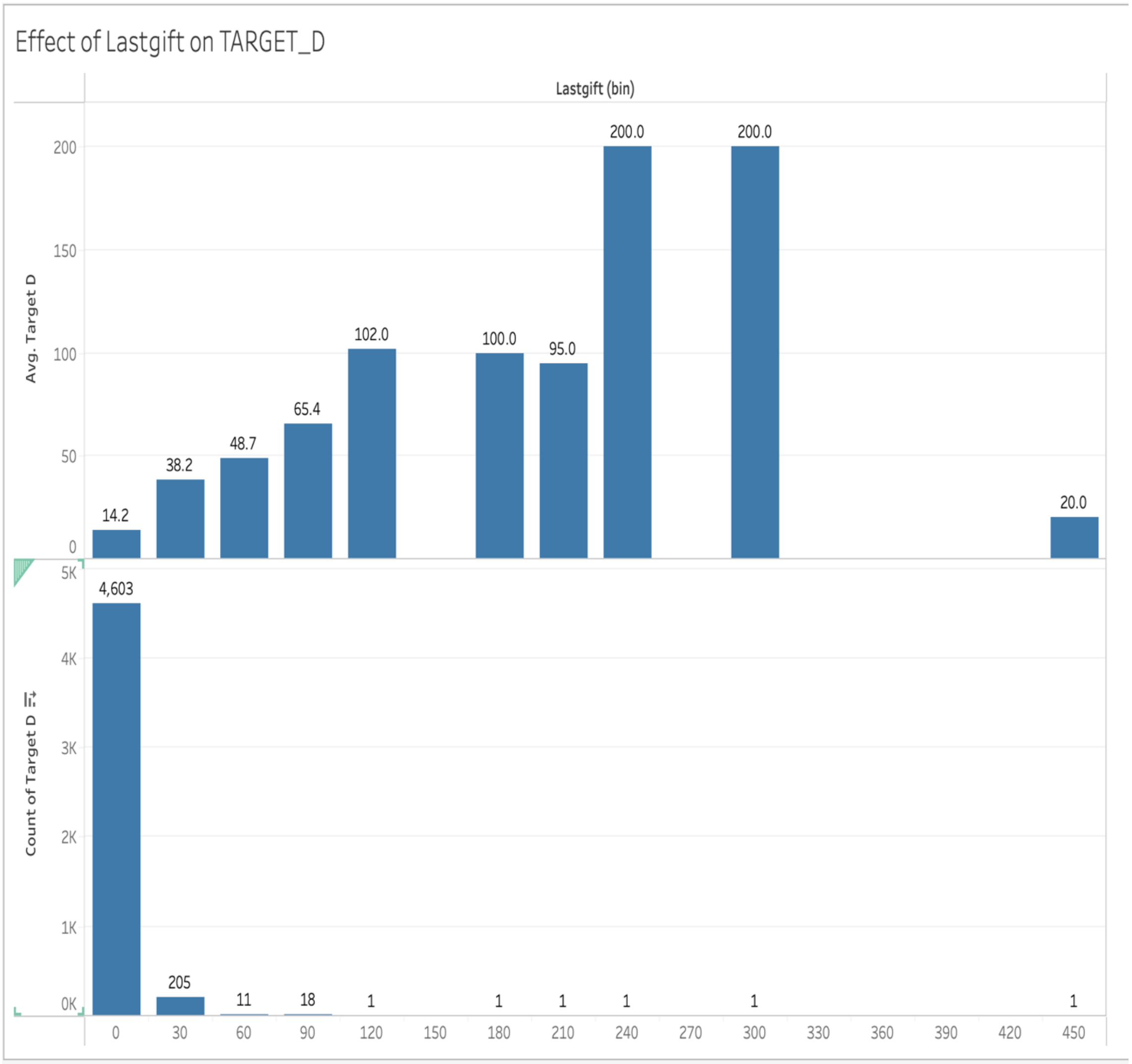
3. IC13



IC13 represents percent households with Income $125,000 - $149,999. Even though the bin for 5 has a count of 4282 which is almost 285 times more than the bin for 20 (count=15), the percentage for bin 20(13.33%) is almost 2.25 times larger than that for bin 5(5.91%)

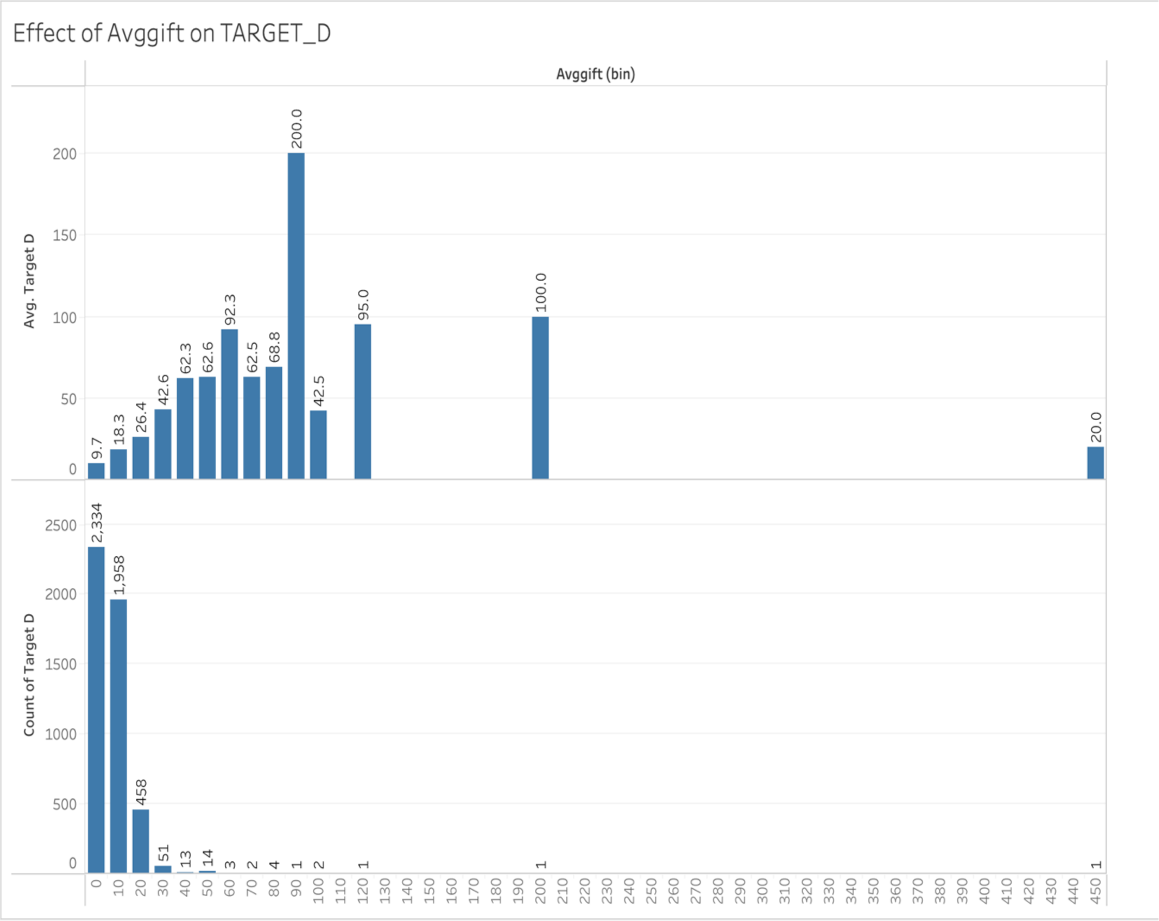
Most important predictors for TARGET_D –

1. Lastgift feature



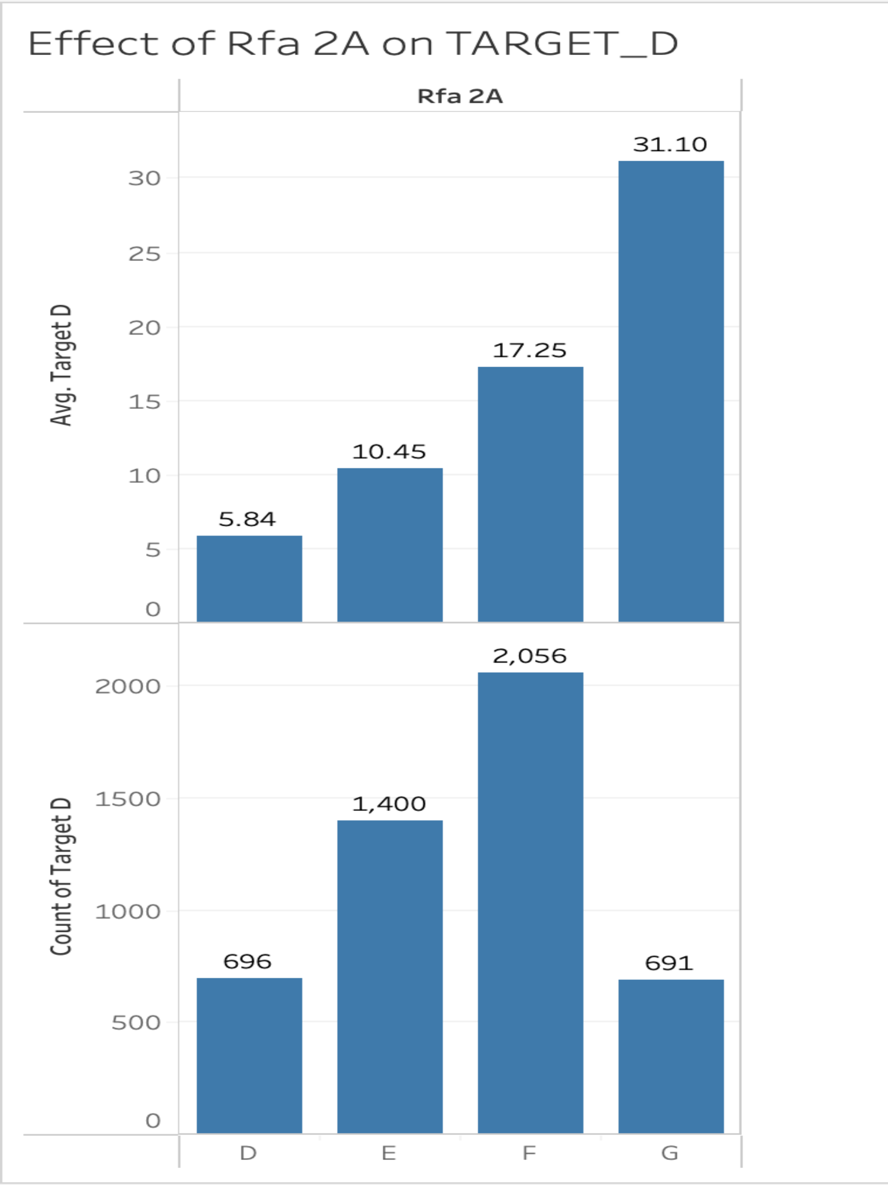
Lastgift represents dollar amount of the most recent gift. The bin for 30 has the highest count of records of 205 people with a percentage of 38.2 which is almost 1.7 times lower than the percentage (65.4) of the bin of 90 which has 18 records. The bins of 240 and 300 have only one record but a percentage of 200 which is an exceptional case.

2. Avggift feature



Avggift represents the average dollar amount of gifts to date. The bins of avggift that have higher value of count have a lower value of percentage, bins 10-30 have much higher count of records (1958,458 and 51) but they have low percentage (18.3, 26.4, 42.6). The bin of 90 is an exceptional case as its count is 1 but has a percentage of 200.

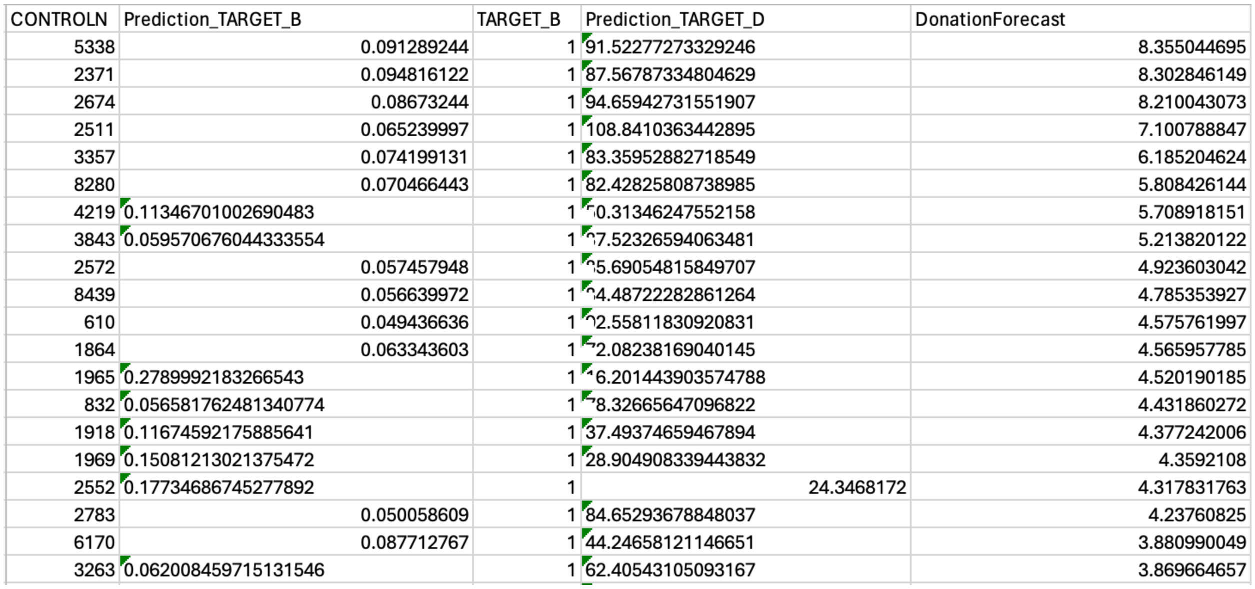
3. Rfa 2A feature



RFA 2 indicates donor’s RFA status as of 97NK promotion date where RFA 2A stands for donation made. The percentage is highest for G values (31.10) even though its count is least (691). The highest count is for F values (2056) but its percentage is only 17.25% which is 1.8 times lesser as compared to G values.

And the above values are denoted as –
D=$5.00 - $9.99
E=$10.00 - $14.99
F=$15.00 - $24.99
G=$25.00 and above

4. **List top 20 predicted donors from the scoring data set in a table (CONTROLN, DonationForecast, TARGET_B, TARGET_D) – order the list by DonationForecast in descending order.**



CONTROLN	Prediction_TARGET_B	TARGET_B	Prediction_TARGET_D	DonationForecast
5338	0.091289244	1	91.52277273329246	8.355044695
2371	0.094816122	1	87.56787334804629	8.302846149
2674	0.08673244	1	94.65942731551907	8.210043073
2511	0.065239997	1	108.8410363442895	7.100788847
3357	0.074199131	1	83.35952882718549	6.185204624
8280	0.070466443	1	82.42825808738985	5.808426144
4219	0.11346701002690483	1	50.31346247552158	5.708918151
3843	0.059570676044333554	1	87.52326594063481	5.213820122
2572	0.057457948	1	85.69054815849707	4.923603042
8439	0.056639972	1	84.48722282861264	4.785353927
610	0.049436636	1	92.55811830920831	4.575761997
1864	0.063343603	1	72.08238169040145	4.565957785
1965	0.2789992183266543	1	16.201443903574788	4.520190185
832	0.056581762481340774	1	78.32665647096822	4.431860272
1918	0.11674592175885641	1	37.49374659467894	4.377242006
1969	0.15081213021375472	1	28.904908339443832	4.3592108
2552	0.17734686745277892	1	24.3468172	4.317831763
2783	0.050058609	1	84.65293678848037	4.23760825
6170	0.087712767	1	44.24658121146651	3.880990049
3263	0.062008459715131546	1	62.40543105093167	3.869664657