Business Case

Paralyzed Veterans of America (PVA) is a nonprofit organization which aims to help injured veterans. They send out mailings asking for donations to their previous donors. It would be helpful to them to have a priority mailing list so that they could reach out to the right donors which are most likely to donate the most. Hence, the problem is to find the right people to target from their existing donor lists.

Handling Target Leaks

The learning dataset cup98LRN.txt contains 95,412 records and 481 fields, which was loaded into Tableau Prep Builder.

Here, data preprocessing (green) - is the process where we change the variable type and remove features for DataRobot to read the data properly.

Target B is about whether the donor donated or not and Target D is about how much the donor donated. For Target B - we removed Target D as it was a target leak for B.

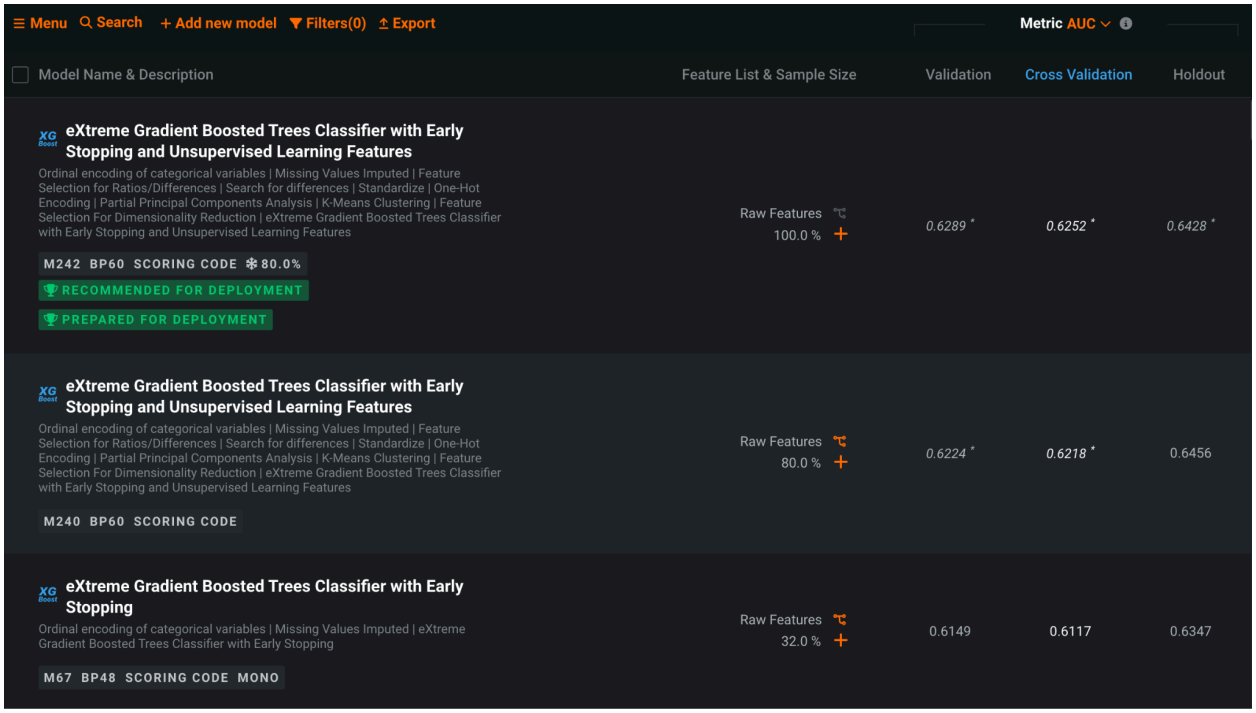
For Target D - we kept only the donors that donated for the campaign by filtering them from Target B. It was done because the percent of donors is 5% amongst all the donors that were mailed for the campaign.

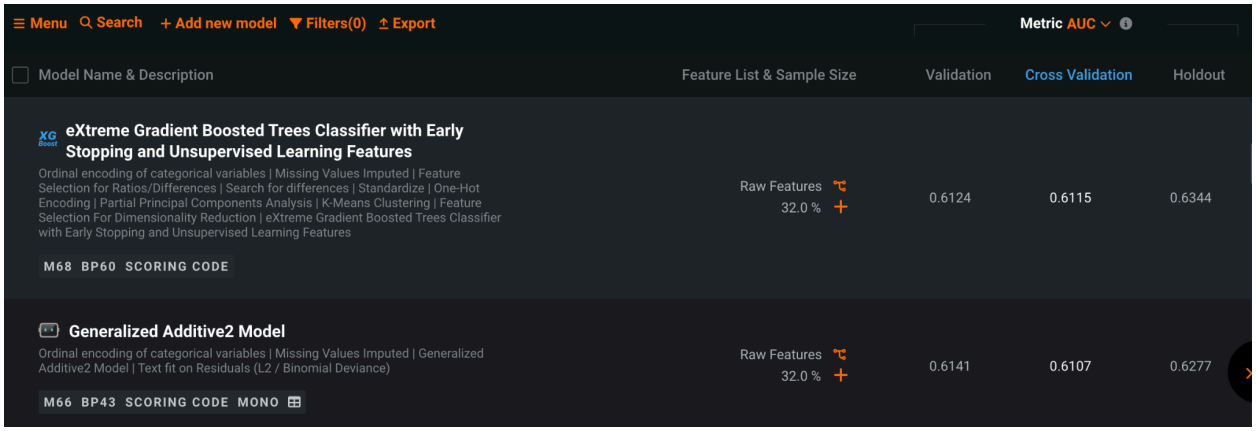
Models evaluated for Target_B

For Target_B, the metric that we used for model evaluation was AUC. DataRobot was on Autopilot and the top 5 models were:

- eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242)
- eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M240)
- eXtreme Gradient Boosted Trees Classifier with Early Stopping (M67)

- eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M68)
- Generalized Additive2 Model (M66)



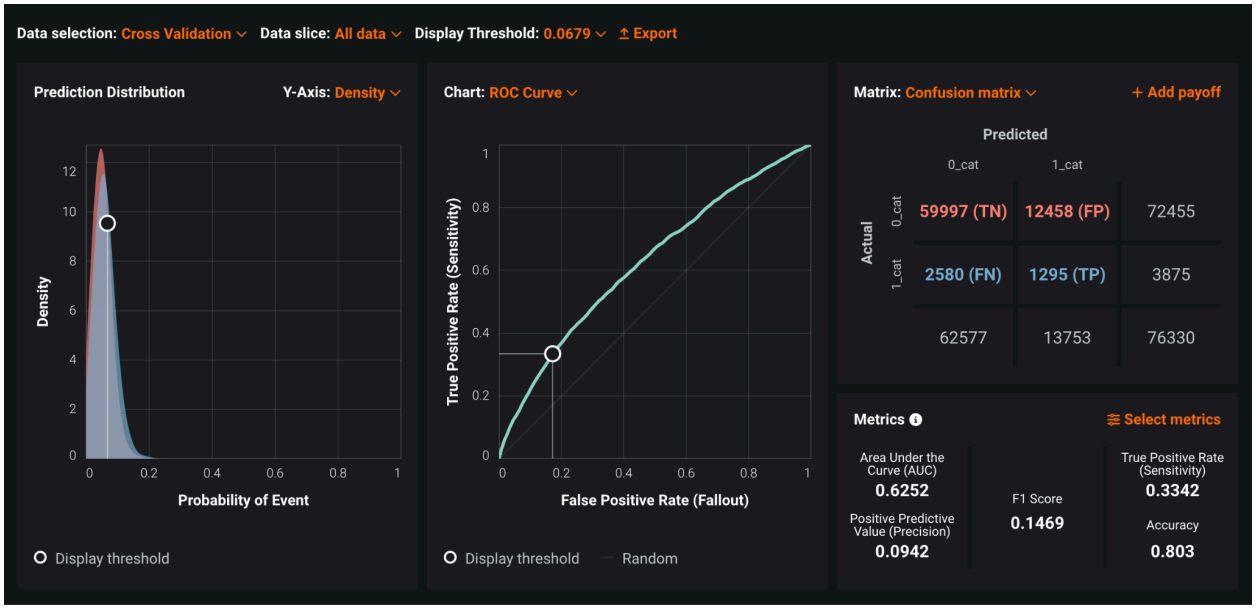


Cross Validation Metrics:

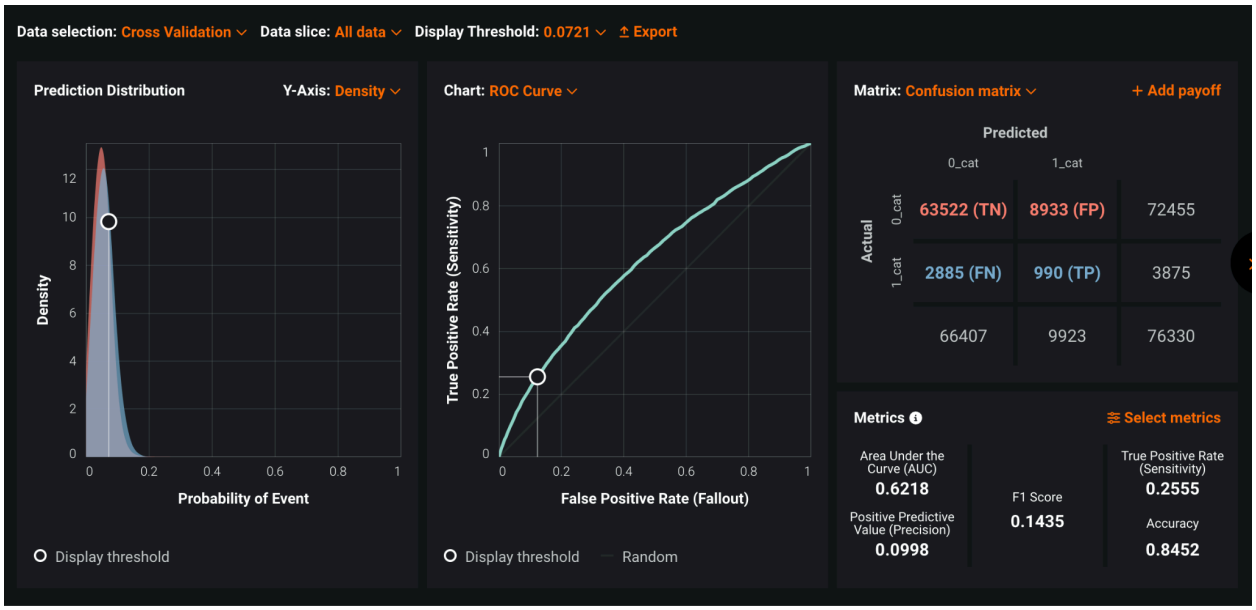
Models	ROC AUC	LogLoss	Accuracy	F1	Precision	Recall	Threshold
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242)	0.6252	*0.1956*	0.803	0.1469	0.0942	0.3342	0.0679
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M240)	0.6218	*0.196*	0.8452	0.1435	0.0998	0.2555	0.0721
eXtreme Gradient Boosted Trees Classifier with Early Stopping (M67)	0.6117	0.1969	0.7828	0.1354	0.0848	0.335	0.0643
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M68)	0.6115	0.1969	0.8613	0.1373	0.1003	0.2175	0.0744
Generalized Additive2 Model (M66)	0.6107	0.197	0.8443	0.1379	0.0959	0.2452	0.0704

Here, exTreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242) performs the best considering ROC AUC and LogLoss. The F1 score looks low for all the models, 0.1469 for the first model.

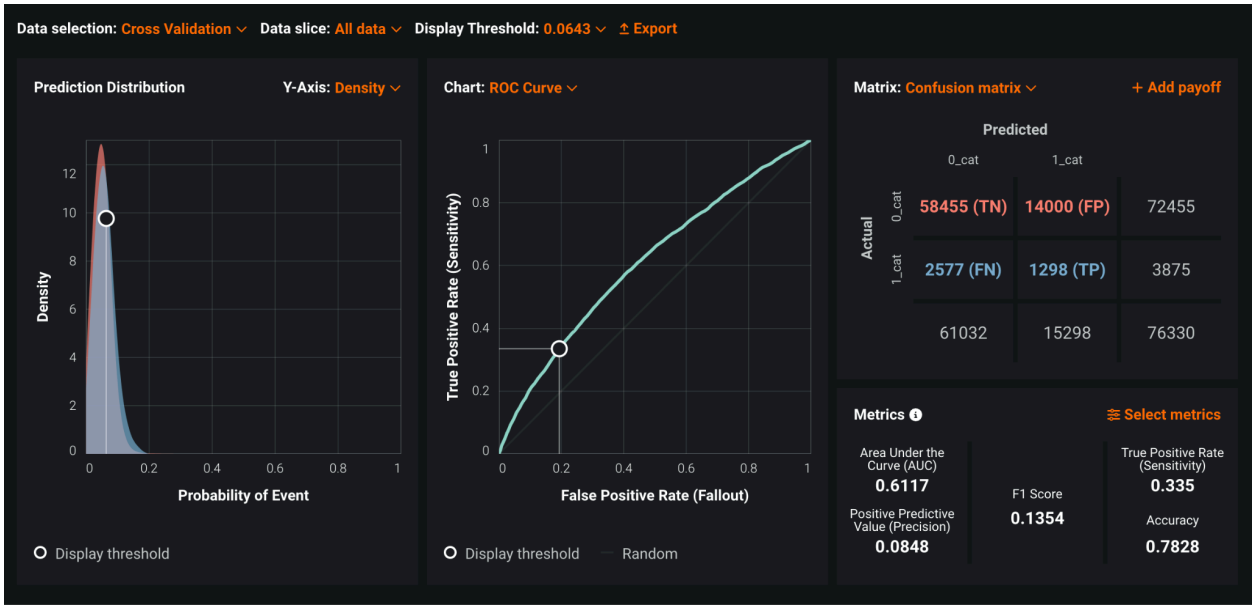
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242)



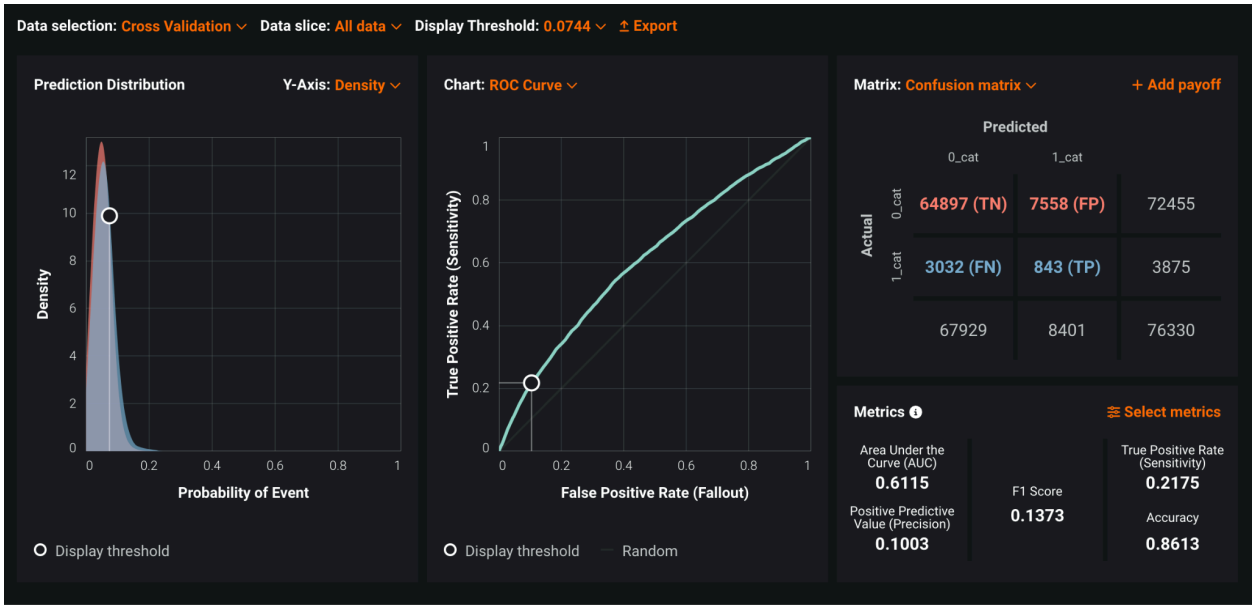
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M240)



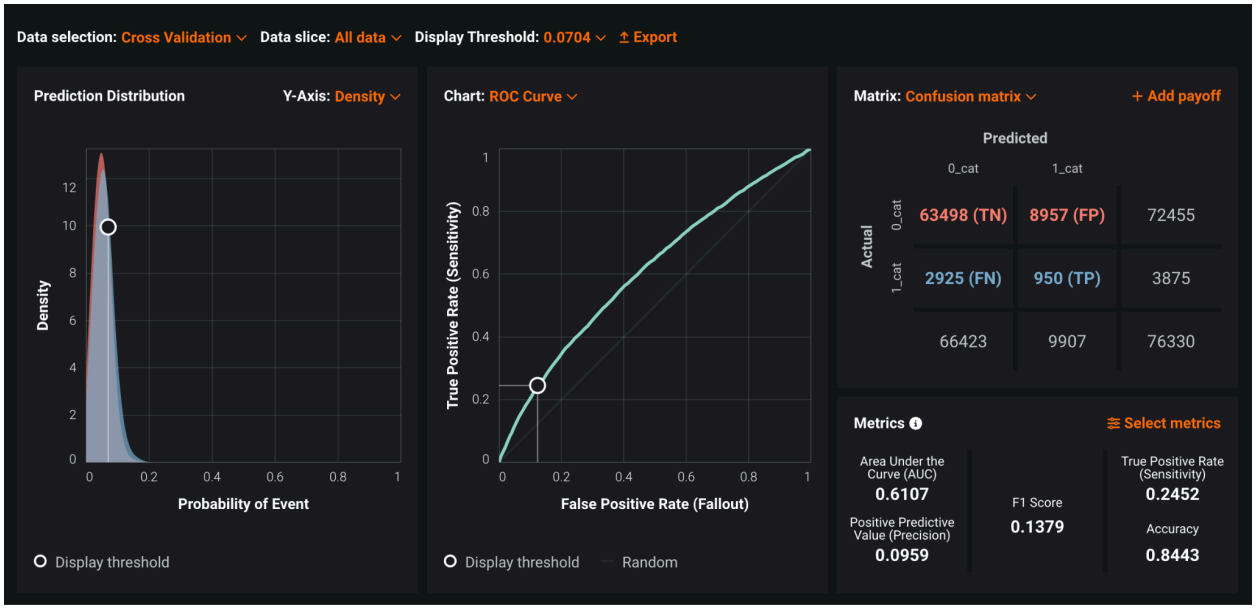
eXtreme Gradient Boosted Trees Classifier with Early Stopping (M67)



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M68)



Generalized Additive2 Model (M66)



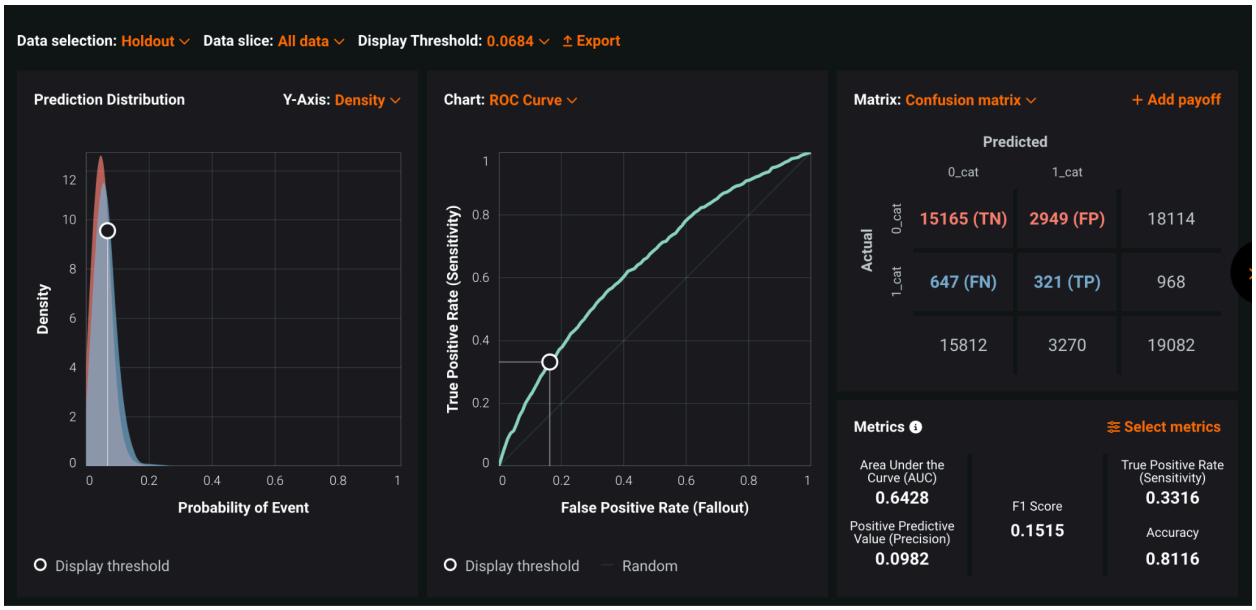
Holdout Metrics:

Models	ROC AUC	LogLoss	Accuracy	F1	Precision	Recall	Threshold
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242)	0.6428	0.1944	0.8116	0.1515	0.0982	0.3316	0.0684
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M240)	0.6456	0.1943	0.8299	0.1573	0.1051	0.313	0.0694
eXtreme Gradient Boosted Trees Classifier with Early Stopping (M67)	0.6347	0.1952	0.8615	0.1439	0.1048	0.2293	0.0754

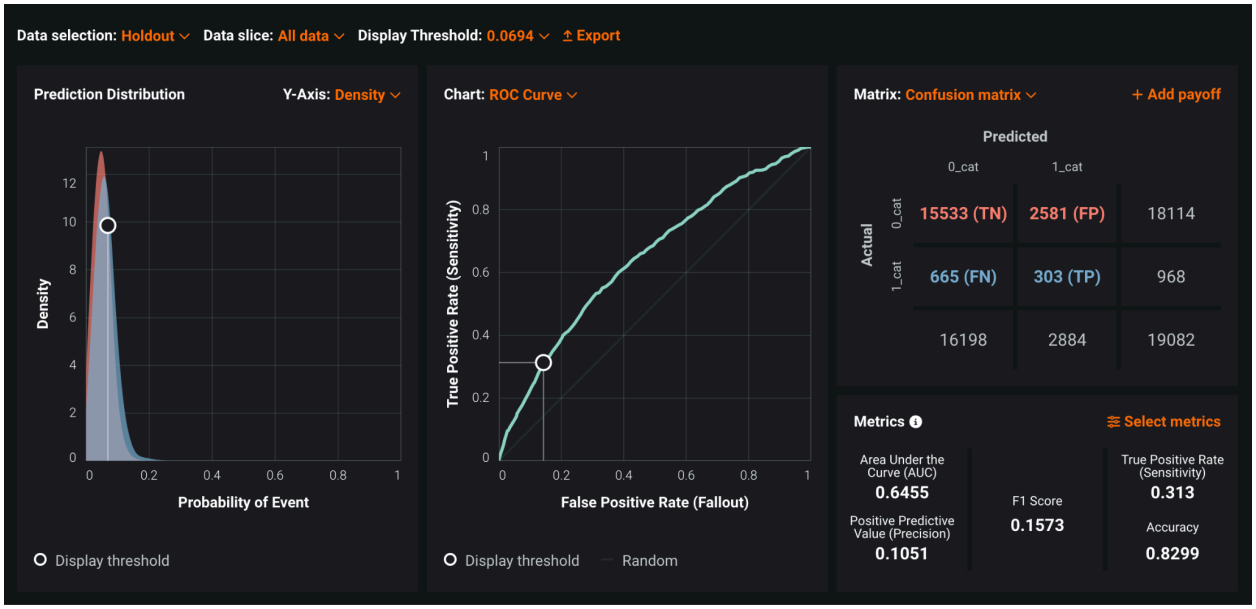
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M68)	0.6344	0.1955	0.8102	0.1458	0.0945	0.3192	0.0686
Generalized Additive2 Model (M66)	0.6277	0.1965	0.8015	0.1446	0.0925	0.3306	0.0629

Here, exTreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242) performs the best considering ROC AUC and LogLoss. The model has F1 as 0.1515 which isn't a good indicator and accuracy of 0.8116 which is fine. It looks like the model may have predictive value.

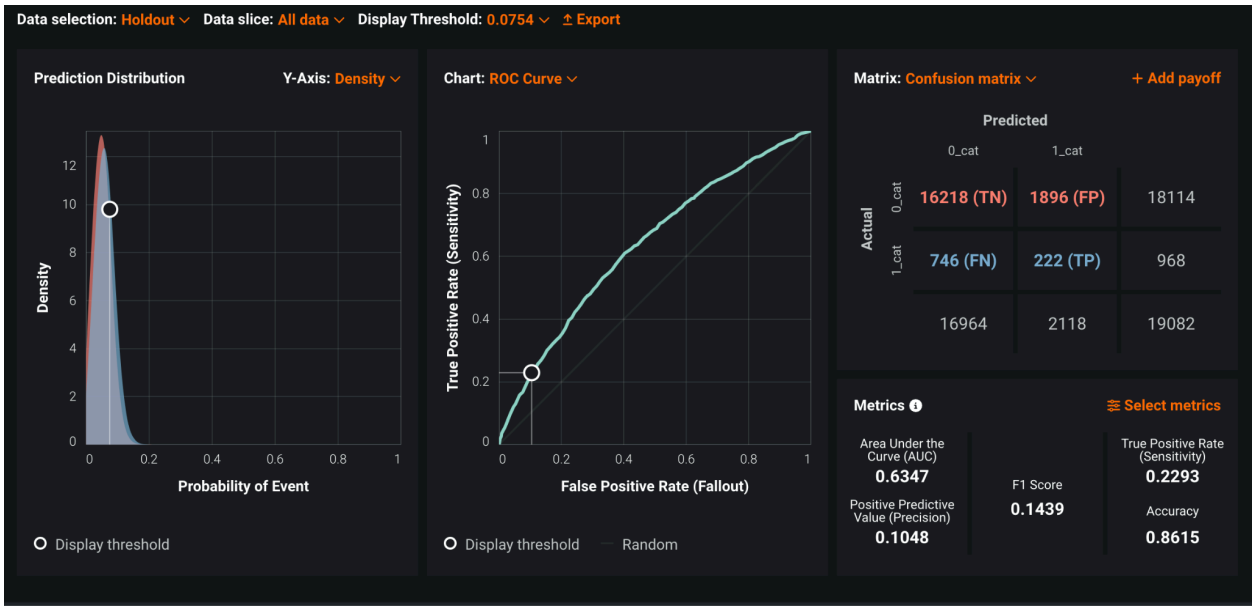
eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242)



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M240)

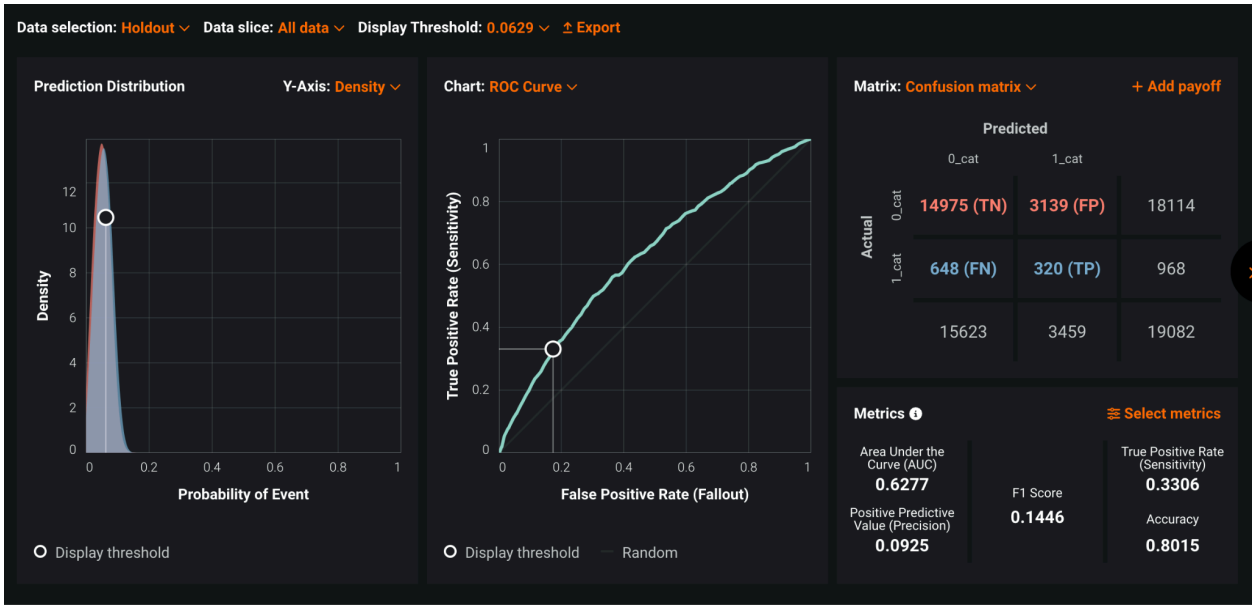


eXtreme Gradient Boosted Trees Classifier with Early Stopping (M67)



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M68)

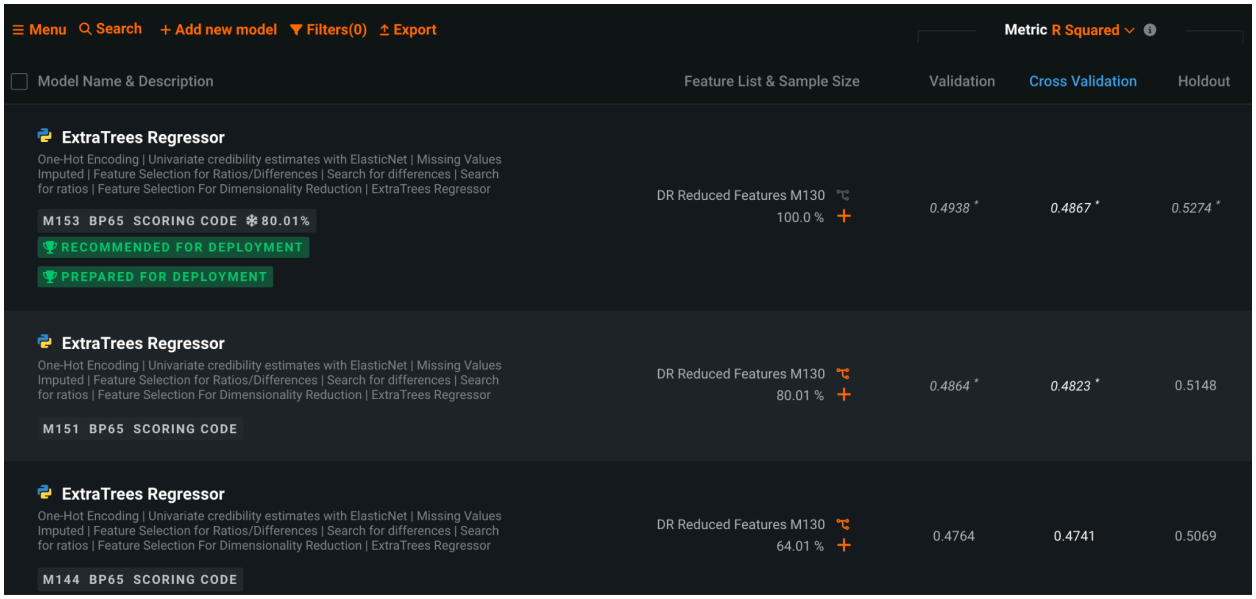
Generalized Additive2 Model (M66)

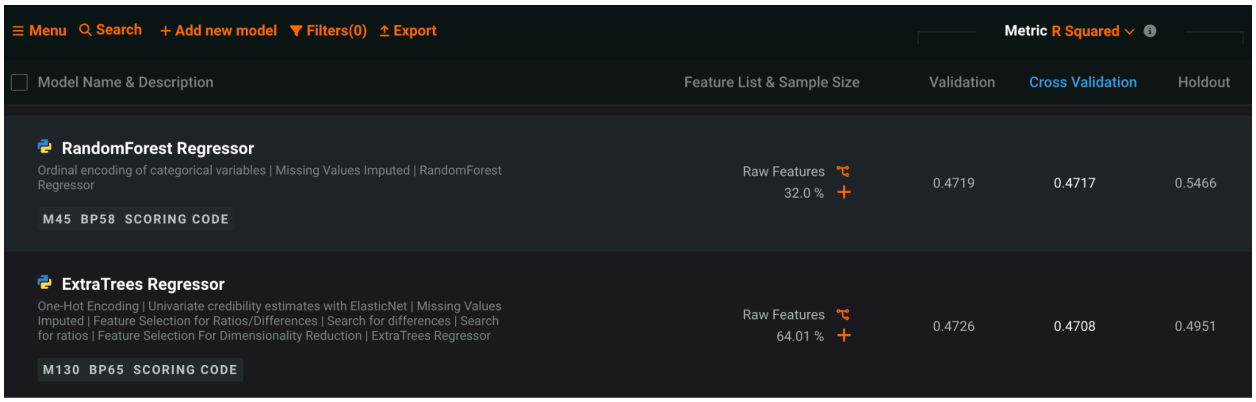


Models evaluated for Target_D

For Target_D, the metric that we used for model evaluation was R-squared. DataRobot was on Autopilot and the top 5 models were:

- ExtraTrees Regressor (M153 BP65)
- ExtraTrees Regressor (M151 BP65)
- ExtraTrees Regressor (M144 BP65)
- RandomForest Regressor (M45 BP58)
- ExtraTrees Regressor (M130 BP65)





Cross Validation Metrics:

Models	R-squared	RMSE	MAE	MAPE
ExtraTrees Regressor (M153 BP65)	0.4867	8.9999	4.2938	33.6994%
ExtraTrees Regressor (M151 BP65)	0.4823	9.0390	4.2938	33.8419%
ExtraTrees Regressor (M144 BP65)	0.4741	9.1137	4.3527	35.3844%
RandomForest Regressor (M45 BP58)	0.4717	9.1365	4.3239	34.2042%
ExtraTrees Regressor (M130 BP65)	0.4708	9.1412	4.2939	34.7830%

For cross validation, the best performing model is ExtraTrees Regressor (M153 BP65) considering R-squared and RMSE. The MAE and MAPE values are considerable and it looks like the model has predictive value.

Holdout Metrics:

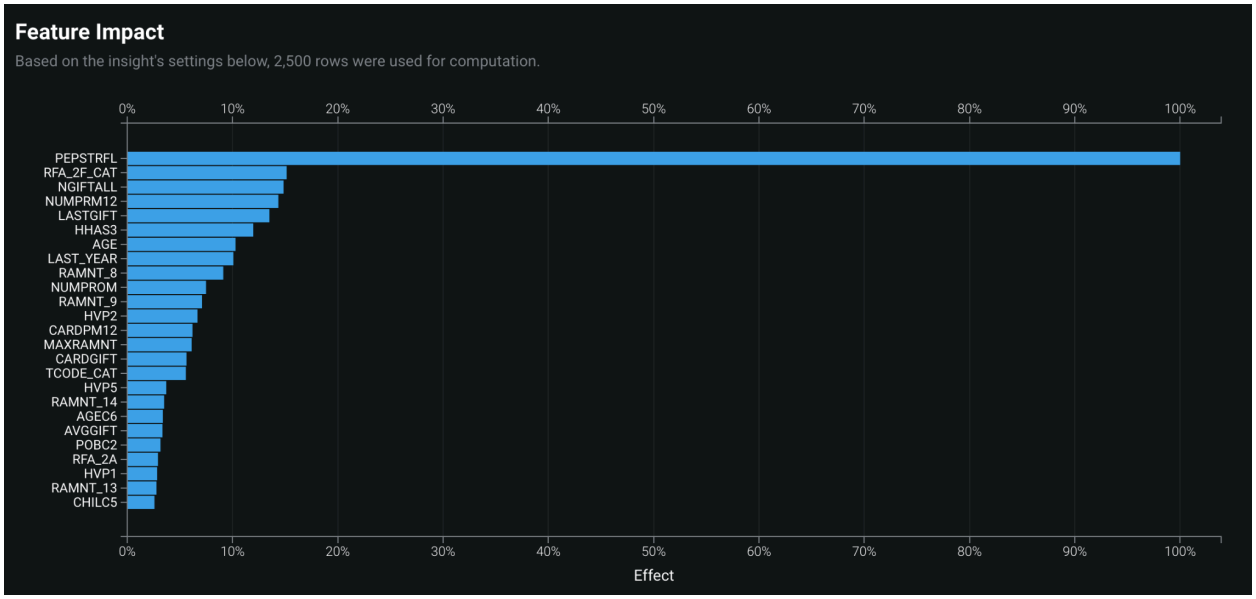
Models	R-squared	RMSE	MAE	MAPE
ExtraTrees Regressor (M153 BP65)	0.5274	8.1009	4.1893	39.1638%
ExtraTrees Regressor (M151 BP65)	0.5148	8.2087	4.2225	39.6769%

ExtraTrees Regressor (M144 BP65)	0.5069	8.2751	4.2628	40.2017%
RandomForest Regressor (M45 BP58)	0.5466	7.9350	4.0425	40.2017%
ExtraTrees Regressor (M130 BP65)	0.4951	8.3732	4.2995	41.2619%

For Holdout as well, the best performing model is ExtraTrees Regressor (M153 BP65) considering R-squared and RMSE. The MAE and MAPE values are considerable and it looks like the model has predictive value.

Most Impactful Features for Target_B

Using eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features (M242), here are the most impactful features for Target B.

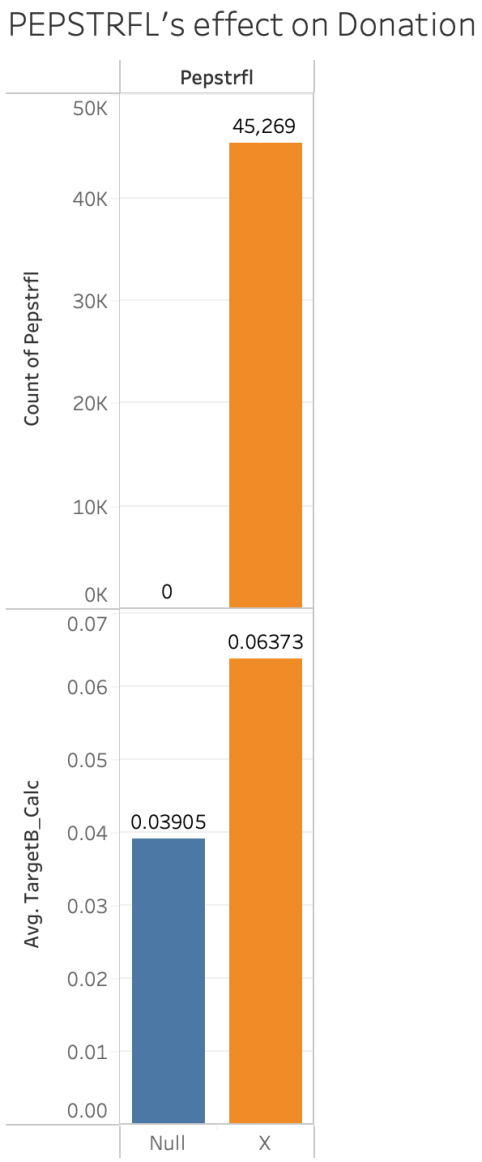


PEPSTRFL has a 100% feature impact on Target B. It may be because of a perfect correlation.

The top 3 features apart from PEPSTRL are:

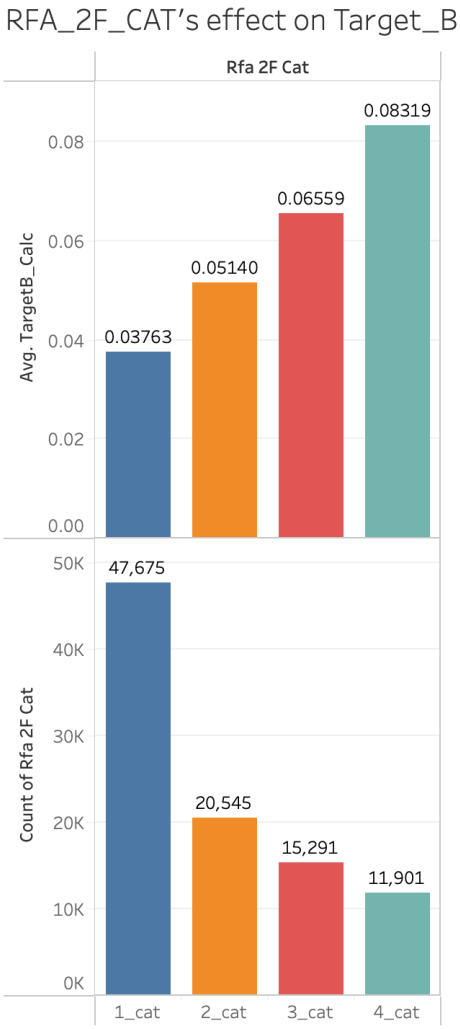
1. RFA_2F_CAT
2. NGIFTALL
3. NUMPRM12

Considering PEPSTRFL:



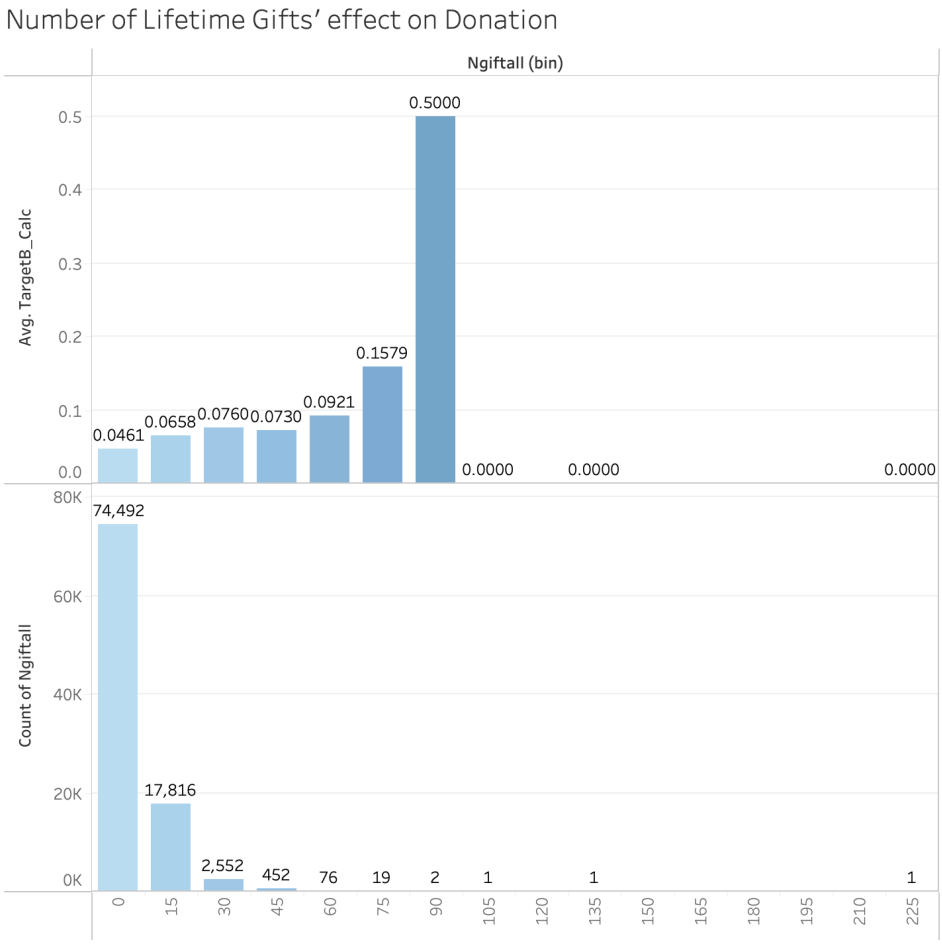
Summary: People who have PEP star RFA status are 60% likely to make a donation.

RFA_2F_CAT



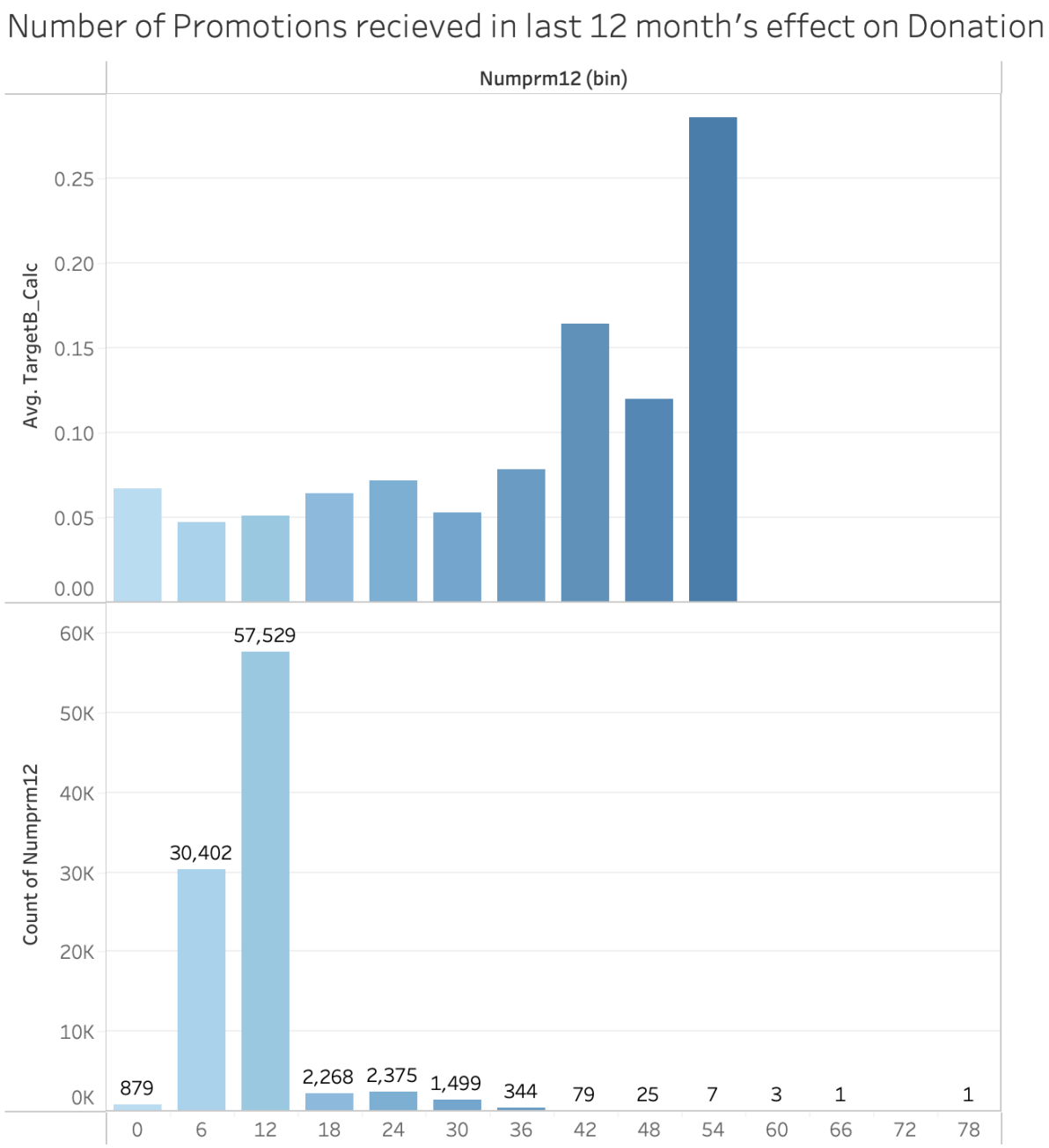
Summary: The percentage of people who would donate increased with the increase in the number of gifts in the recency period.

NGIFTALL



Summary: The number of lifetime gifts given till 45 to 50 is around 0.7% of people donating and people who have made more than 90 gifts are around 50% likely to donate.

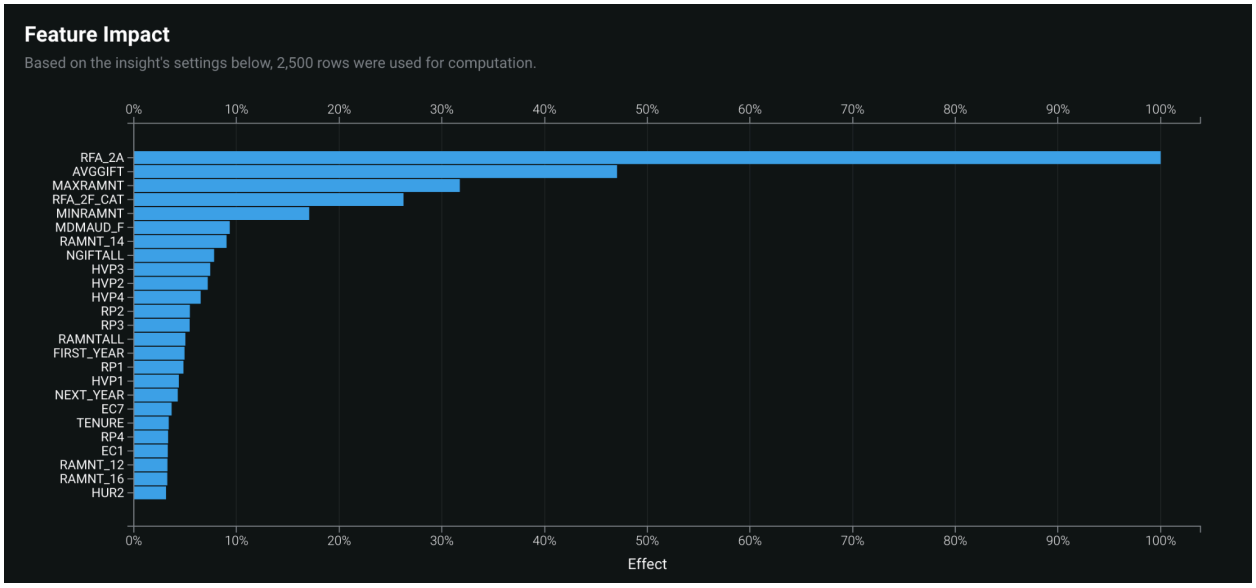
NUMPRM12



Summary: If people received more than 48 up to 55 promotions they were twice likely to make a donation than the rest.

Most Impactful Features for Target_D

Using , ExtraTrees Regressor (M153 BP65) here are the most impactful features for Target B.

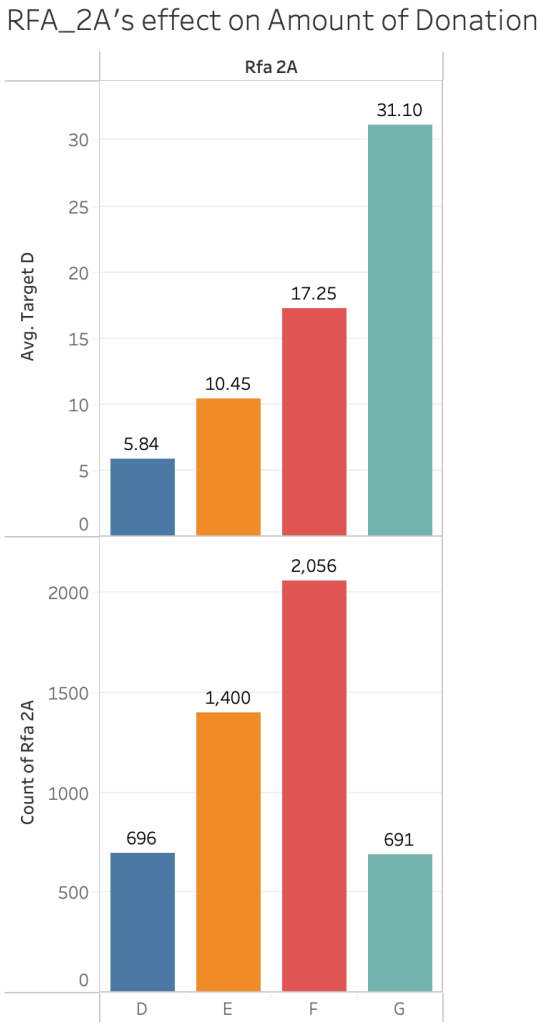


Here, RFA_2A has a 100% feature impact on Target B. It may be because of a perfect correlation.

The top 3 features apart from RFA_2A are:

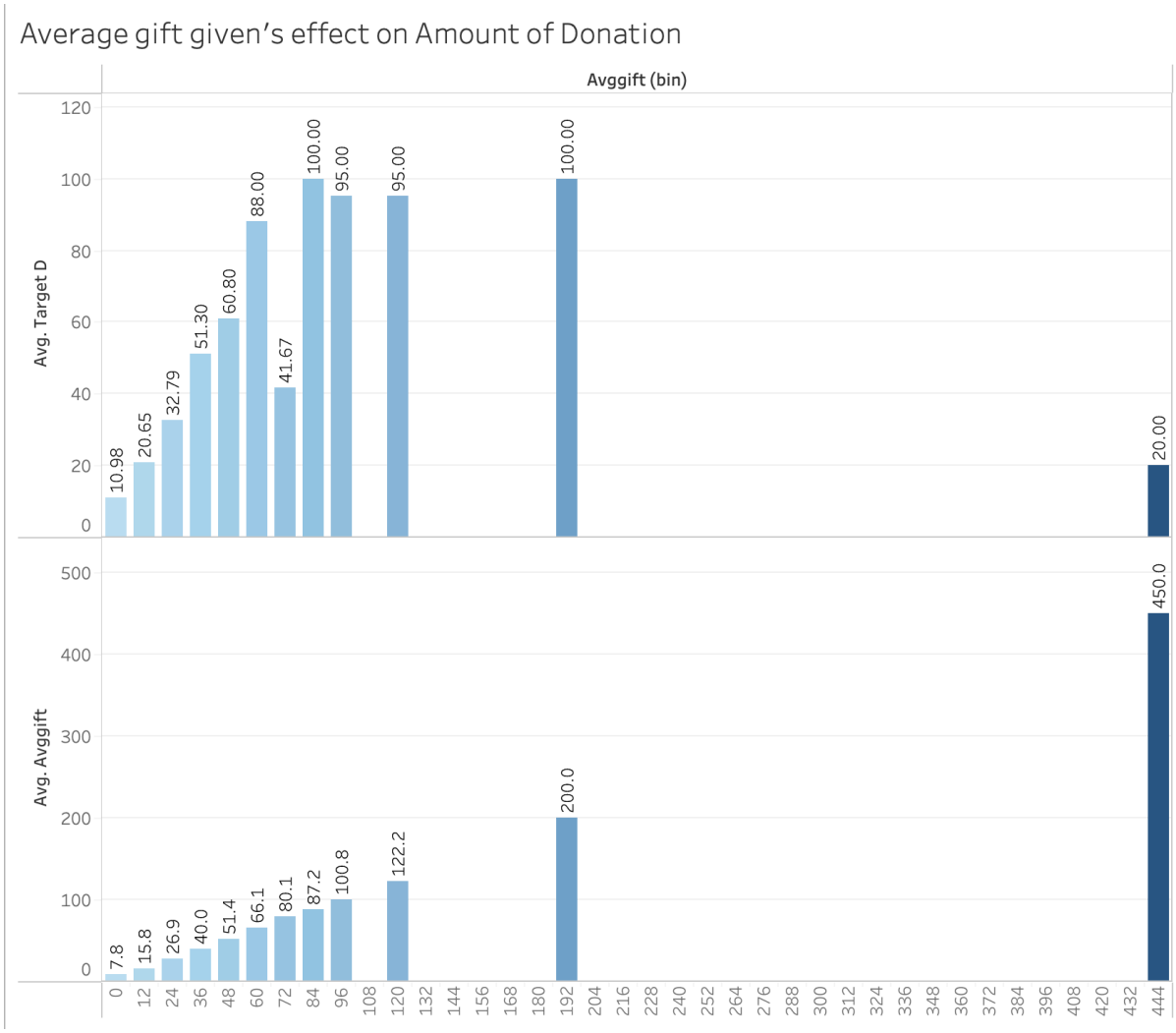
1. AVGGIFT
2. RFA_2F_CAT
3. MINRAMNT

Considering RFA_2A here,



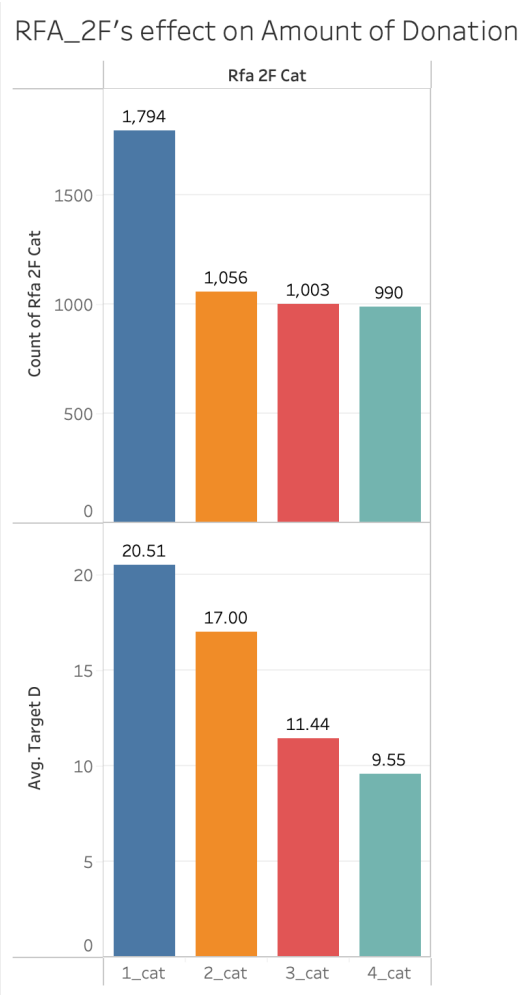
Summary: The average donation is the highest amongst Group G which have donated 25 and above.

AVGGIFT



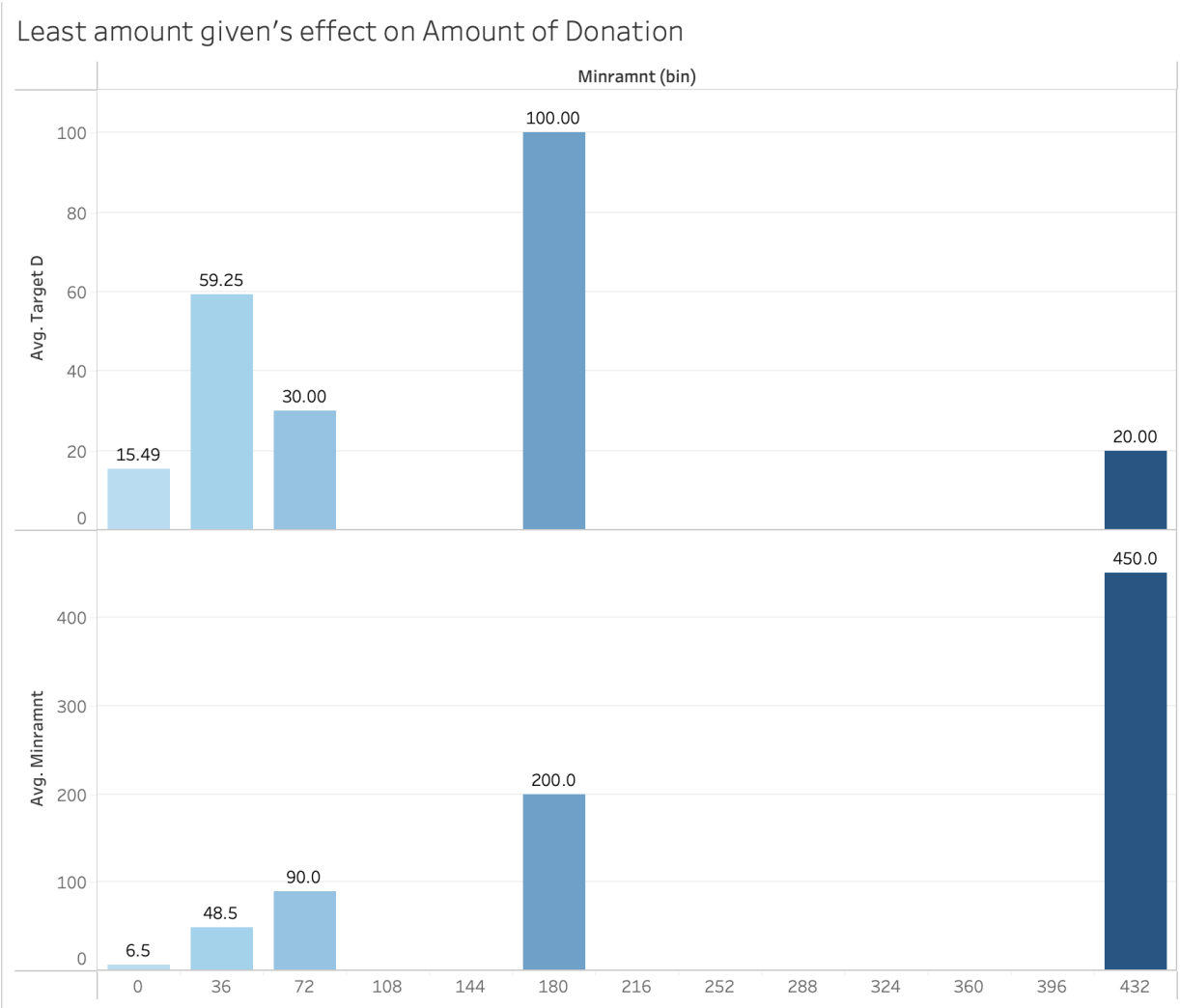
Summary: Average donation increased with increase in average gift till $120 dollars.

RFA_2F_CAT



Summary: The people who gave less frequent donations were likely to donate more than the ones who donate regularly.

MINRAMNT



Summary: When the minimum donation amount was till $180, then the average donation amount shot up around $100.

List of Top 20 Predicted Donors

Priority list	Row_ID	Probability of Donation (Target B)	Predicted Amount of the Gift (Target D)	Donation Forecast
1	53495	0.099	120.248	11.918
2	94491	0.393	28.469	11.182
3	39507	0.420	26.308	11.055
4	93435	0.117	94.055	11.048
5	71306	0.221	49.005	10.829
6	71464	0.124	86.789	10.726
7	20009	0.115	91.654	10.586
8	2371	0.106	92.737	9.789
9	87524	0.177	53.926	9.540
10	26872	0.087	105.157	9.103
11	11263	0.097	93.112	9.020
12	75588	0.106	84.119	8.936
13	30854	0.094	93.136	8.744
14	33593	0.437	20.000	8.732
15	67201	0.333	26.192	8.715
16	43338	0.087	97.117	8.436
17	75989	0.424	19.850	8.412
18	72629	0.434	19.258	8.358
19	95893	0.147	55.843	8.201
20	68663	0.415	19.374	8.045