Q1. Business Case.

In the telecommunications industry, the companies face revenue loss when customers discontinue their services. The goal is to analyze models to understand why customers leave and take proactive measures to retain them.

We will use a regression model, decision trees, random forest and boosted trees to analyze the data. The dataset has a total of 20 features and 4250 data points.

Financial implications:

Assumed Revenue (Through service plans)	
Customer pays	$135
Time period	1 year
Money earned	$135

The revenue will be generated by the service provided by the telecommunication companies. They earn ~$135 from each customer annually. Thus, money earned is $135.

Assumed Cost	
Discount provided	$40.5
Cost of call	$20
Time period (years)	1 year
Total Cost incurred	$60.5

30% on $135(currently paid by the customer)

To keep the customers that are predicted to churn, we will provide them a discount of 30% on their current plan. Hence, 30% of $135 is $40.5.

Also, when we make the call we give the person in charge $60/hr which is $20 per call (if the person makes 3 calls in each hour).

Hence, it occurs at a cost of $60.5 in total.

Payoff Matrix

TP	The model says customer will churn, we give discount - 30% = $40.5 from $135 earnt per customer, the total will be ($135 - $40.5) - $20 (cost of each call) = $74.5
TN	The model predicts the customer doesn't churn, the company doesn't take any action hence, no cost involved but it is still getting revenue from the customer = $135.
FP	The customer doesn't churn but the model falsely predicts they would so, the company gives discount of 30% = $40.5 from $135 earnt per customer, the total will be ($135 - $40.5) - $20 (cost of each call) = $74.5
FN	The model predicts the customer doesn't churn, but they do, and the company takes no action.

	Predicted		
Actual		0	1
	0	TN (135)	FP (74.5)
	1	FN (0)	TP (74.5)

Q2. Data Preprocessing and exploratory data analysis.

Our target variable is churn as we want to know whether the customer will continue to use the service plan or not.

Also, we haven't excluded any features or changed the var type of the features as they are all identified correctly.

Hence, we have taken all 20 features provided in the data set for the 4250 data points.

We will compare logistic regression, decision trees, random forest and boosted trees.

There are some features that have many unique values like total_eve_minutes, total_eve_charge, total_night_minutes which are 1613, 1439 and 1597 respectively. Even though these seem like large numbers, it isn't high cardinality compared to the dataset size we have.

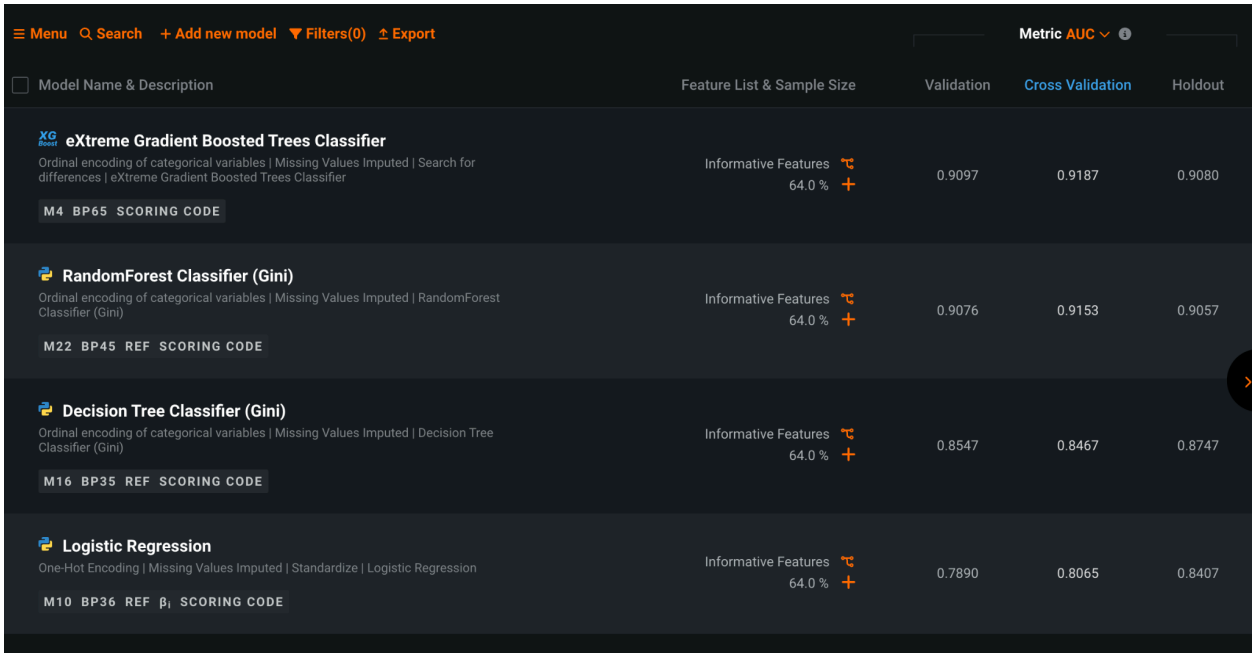
Dataset

Menu Search Feature List: All Features View Raw Data + Create feature list 1-20 of 20

Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
churn		20	Target	Categorical	2	0					
total_day_minutes		7		Numeric	1,682	0	180	54.24	180	0	347
total_day_charge		9		Numeric	1,682	0	30.57	9.22	30.67	0	58.96
number_customer_service_calls		19		Numeric	10	0	1.57	1.31	1	0	9
international_plan		4		Categorical	2	0					
voice_mail_plan		5		Categorical	2	0					
number_vmail_messages		6		Numeric	46	0	7.73	13.53	0	0	52
total_intl_calls		17		Numeric	21	0	4.42	2.46	4	0	20
total_intl_charge		18		Numeric	161	0	2.78	0.74	2.81	0	5.40
total_intl_minutes		16		Numeric	161	0	10.28	2.75	10.40	0	20

Q3. Models and metrics.

The models we have considered here are logistic regression, decision trees, random forest and boosted trees.



Menu Search + Add new model Filters(0) Export Metric AUC

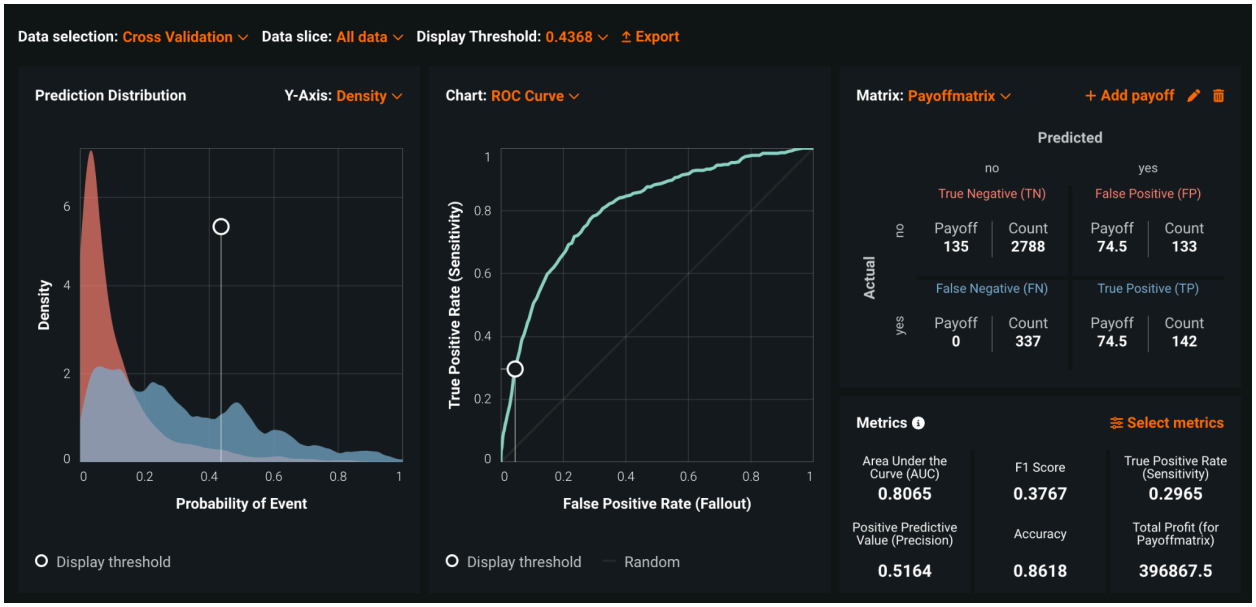
Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
eXtreme Gradient Boosted Trees Classifier Ordinal encoding of categorical variables | Missing Values Imputed | Search for differences | eXtreme Gradient Boosted Trees Classifier M4 BP65 SCORING CODE	Informative Features 64.0 %	0.9097	0.9187	0.9080
RandomForest Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | RandomForest Classifier (Gini) M22 BP45 REF SCORING CODE	Informative Features 64.0 %	0.9076	0.9153	0.9057
Decision Tree Classifier (Gini) Ordinal encoding of categorical variables | Missing Values Imputed | Decision Tree Classifier (Gini) M16 BP35 REF SCORING CODE	Informative Features 64.0 %	0.8547	0.8467	0.8747
Logistic Regression One-Hot Encoding | Missing Values Imputed | Standardize | Logistic Regression M10 BP36 REF β_i SCORING CODE	Informative Features 64.0 %	0.7890	0.8065	0.8407

The values for AUC look good for both cross validation and holdout. It can be seen that considering AUC, Boosted Trees Classifier performs the best here, and Random Forest has a very close value of AUC to it too.

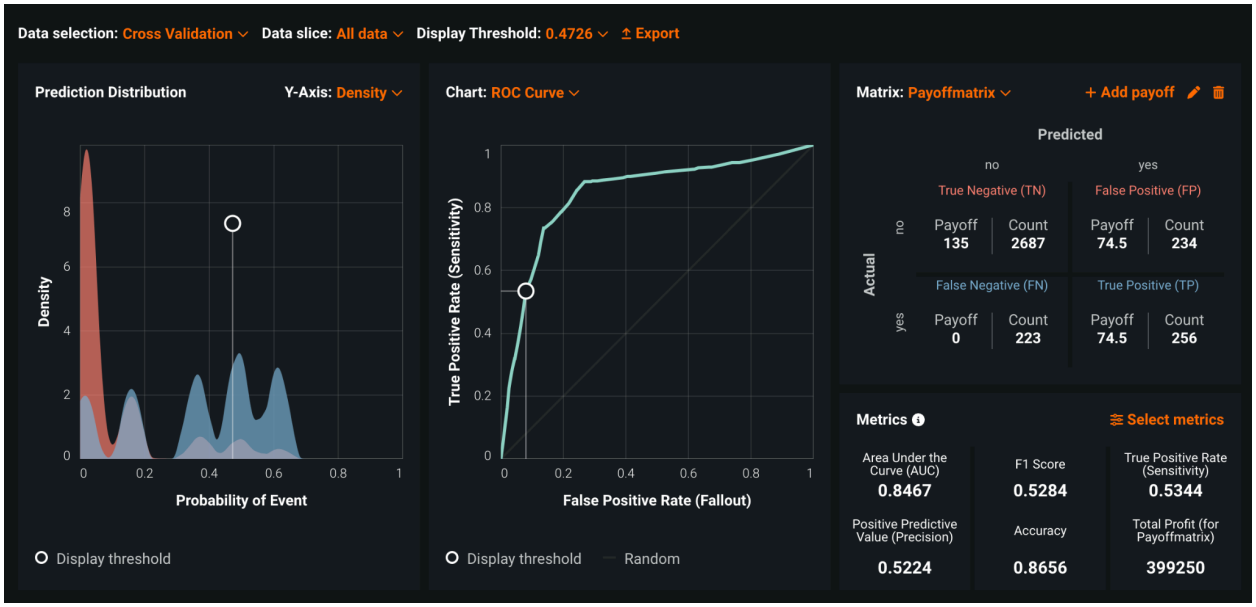
Cross Validation comparison of all models:

All the values are captured by keeping the maximum payoff.

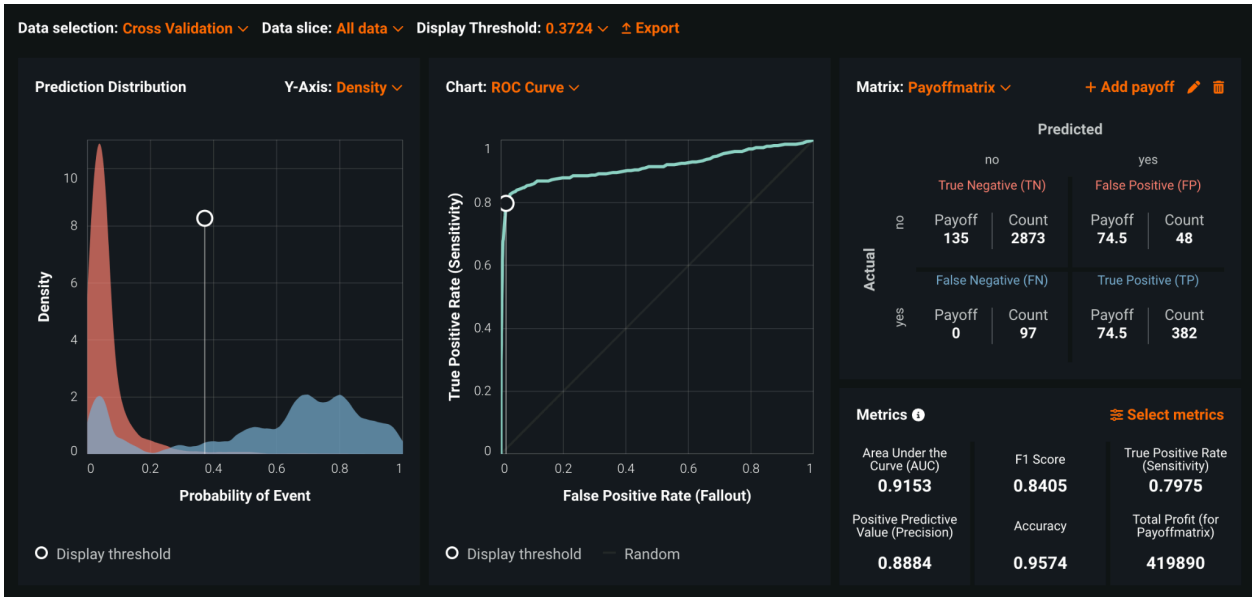
Logistic Regression



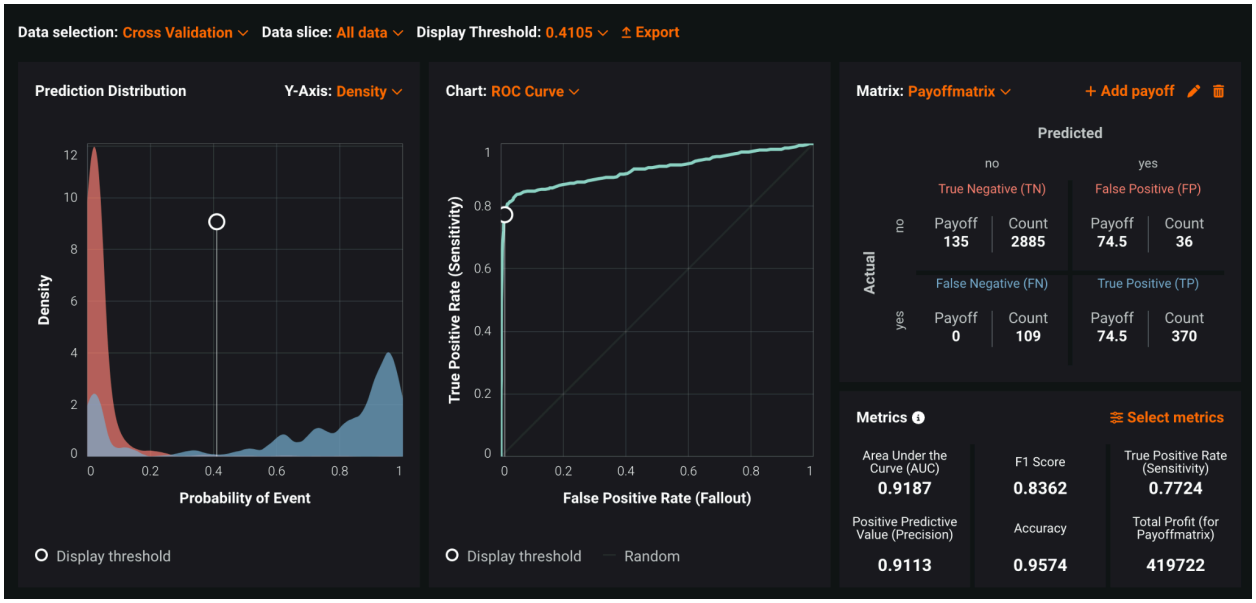
Decision Tree Classifier



Random Forest Classifier



Boosted Trees Classifier



	Cross Validation Metrics at Maximum Payoff			
	Logistic Regression	*Decision Tree*	*Random Forest*	*Boosted Trees*
Recall	0.2965	0.5344	0.7975	0.7724
Precision	0.5164	0.5224	0.8884	0.9113

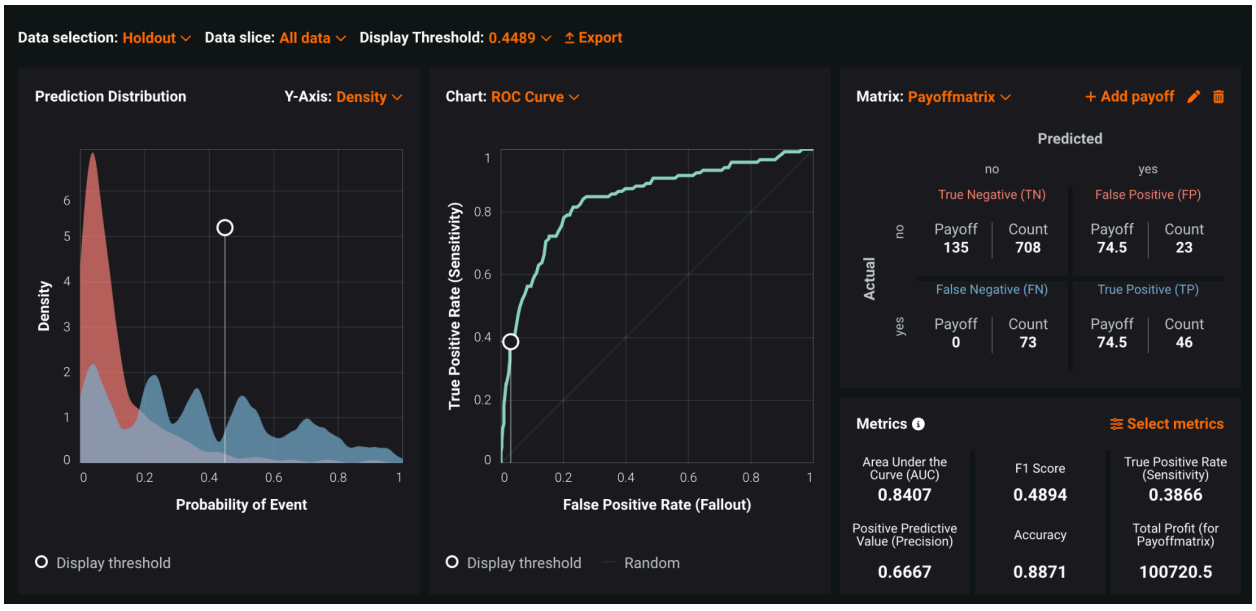
F1	0.3767	0.5284	0.8405	0.8362
Accuracy	0.8618	0.8656	0.9574	0.9574
ROC AUC	0.8065	0.8467	0.9153	0.9187
Maximum Payoff	$3,96,867.5	$3,99,250	$4,19,890	$4,19,722
Threshold	0.4368	0.4726	0.3724	0.4105

For cross validation, the payoff of Random Forest is the most. But, considering all the other metrics like precision and ROC AUC, Boosted Trees performs better. The payoff metric's difference for Random Forest and Boosted Trees isn't much. Overall, the boosted trees classifier performs better than other models.

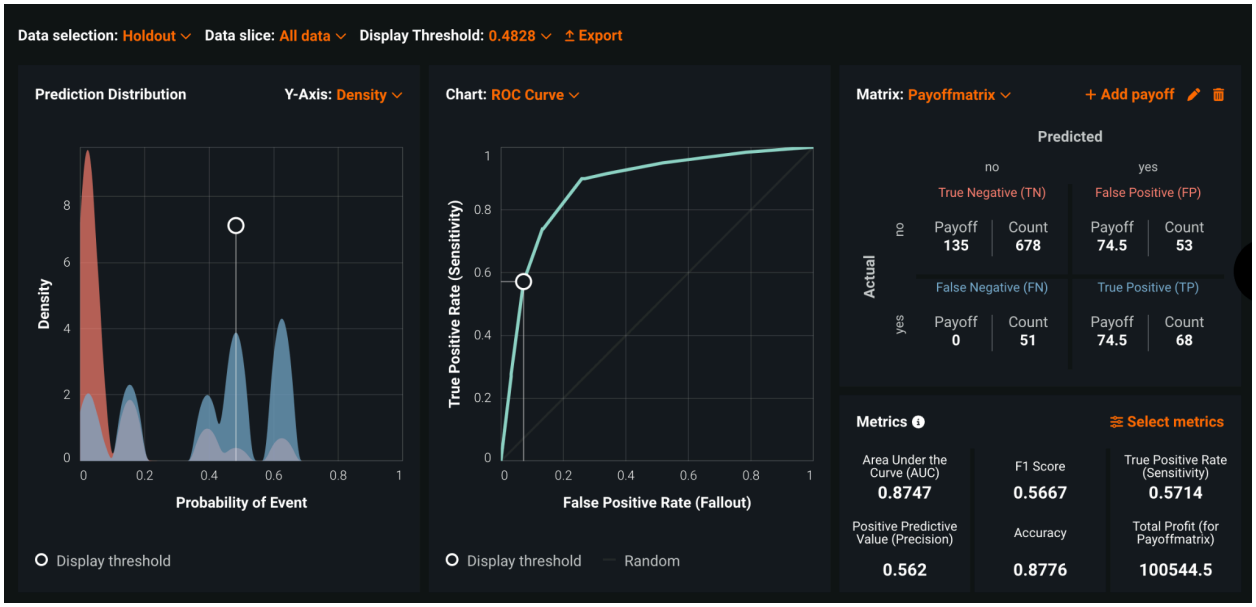
Holdout comparison of all models:

All the values are captured by keeping the maximum payoff.

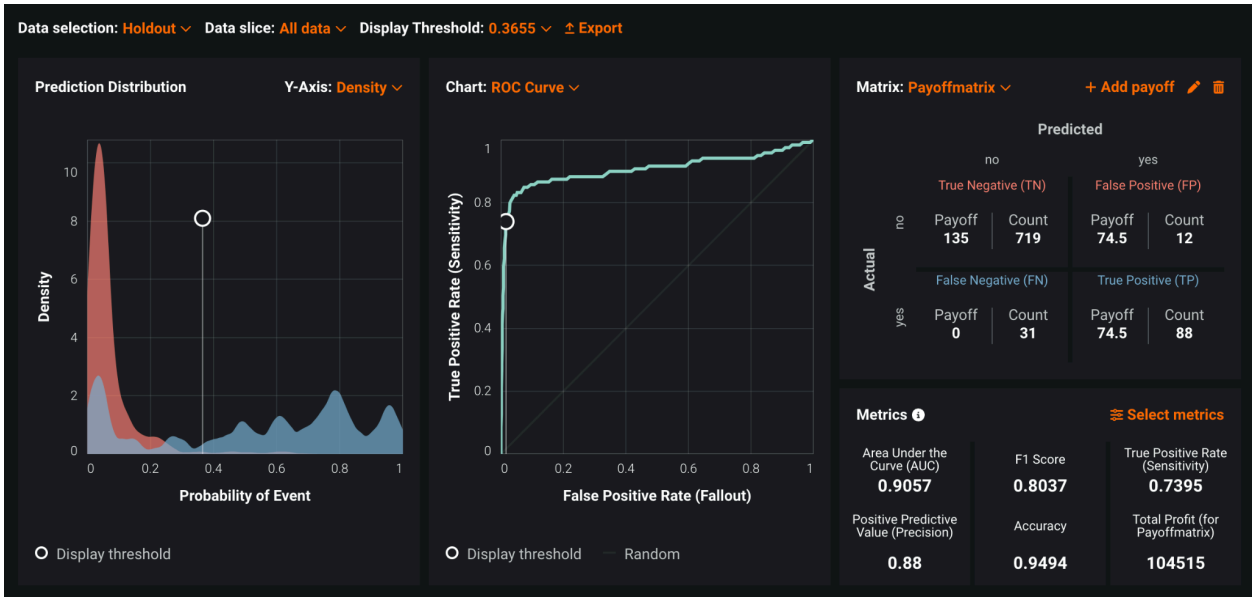
Logistic Regression



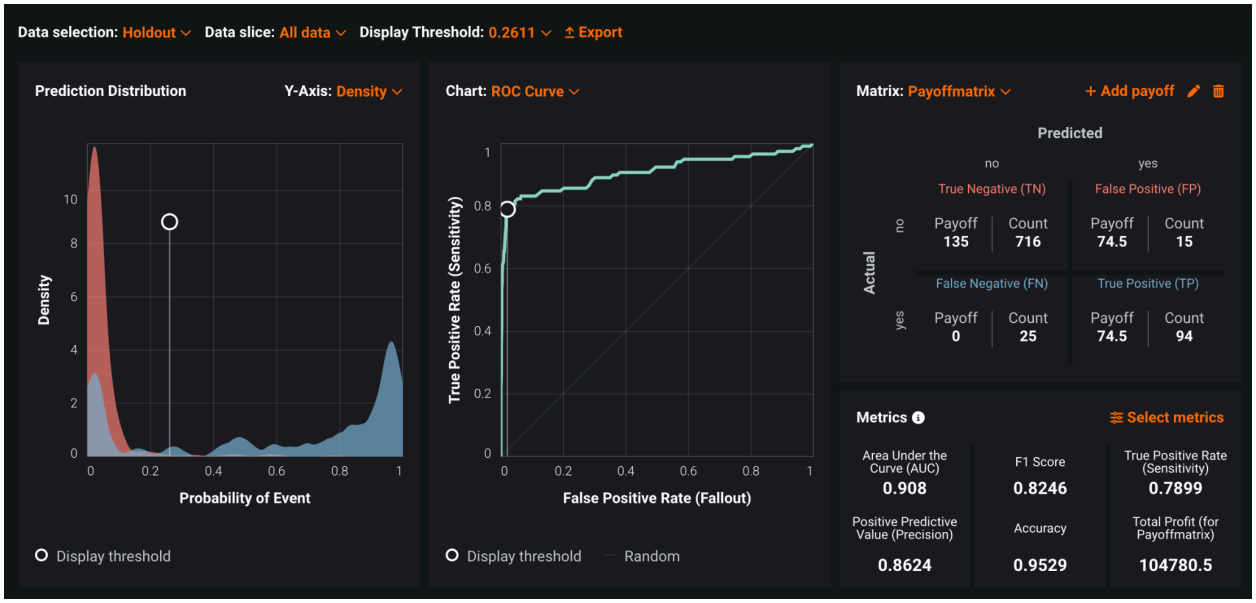
Decision Tree Classifier



Random Forest Classifier



Boosted Trees Classifier



	Holdout Metrics at Maximum Payoff			
	Logistic Regression	*Decision Tree*	*Random Forest*	*Boosted Trees*
Recall	0.3866	0.5714	0.7395	0.7899
Precision	0.6667	0.562	0.88	0.8624
F1	0.4894	0.5667	0.8037	0.8246
Accuracy	0.8871	0.8776	0.9494	0.9529
ROC AUC	0.8407	0.8747	0.9057	0.908
Maximum Payoff	$1,00,720.5	$1,00,544.5	$1,04,515	$1,04,780.5
Threshold	0.4489	0.4828	0.3655	0.2611

For holdout, Boosted trees have the most payoff amount.

All the metrics of Boosted Trees perform better than others except for precision where Random forest has a better value by a slight value. But, overall Boosted trees perform better than others.

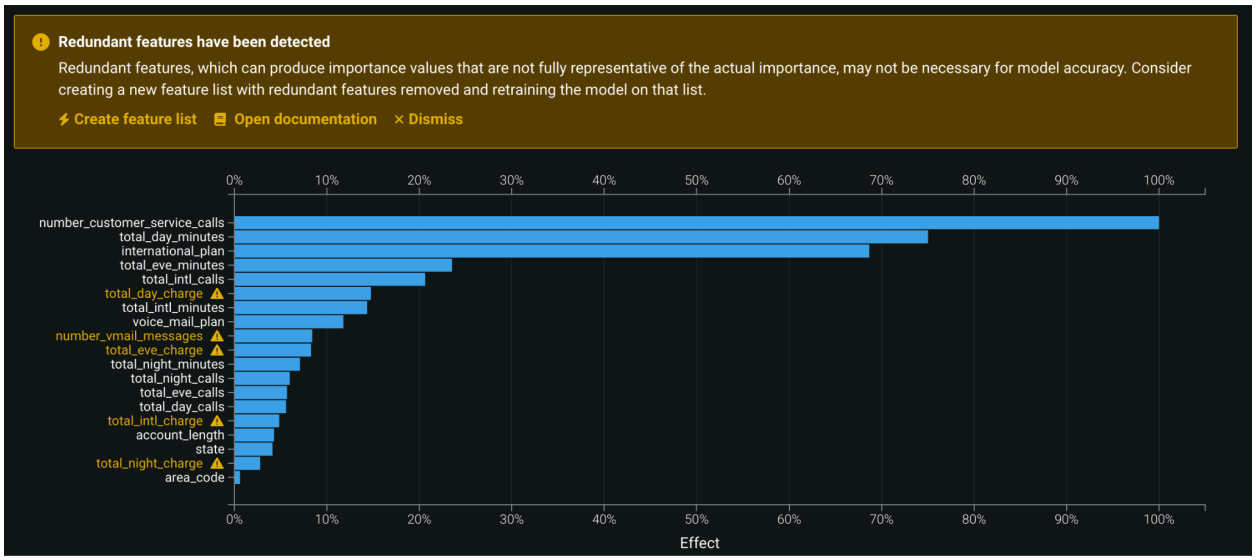
Q4. Top 4 predictors of customer churn.

The best performing model was Boosted Trees and according to the same model, the features that impact churn the most are selected.

The top 4 here are number_customer_service_calls, total_day_minutes, international_plans, and total_eve_minutes.

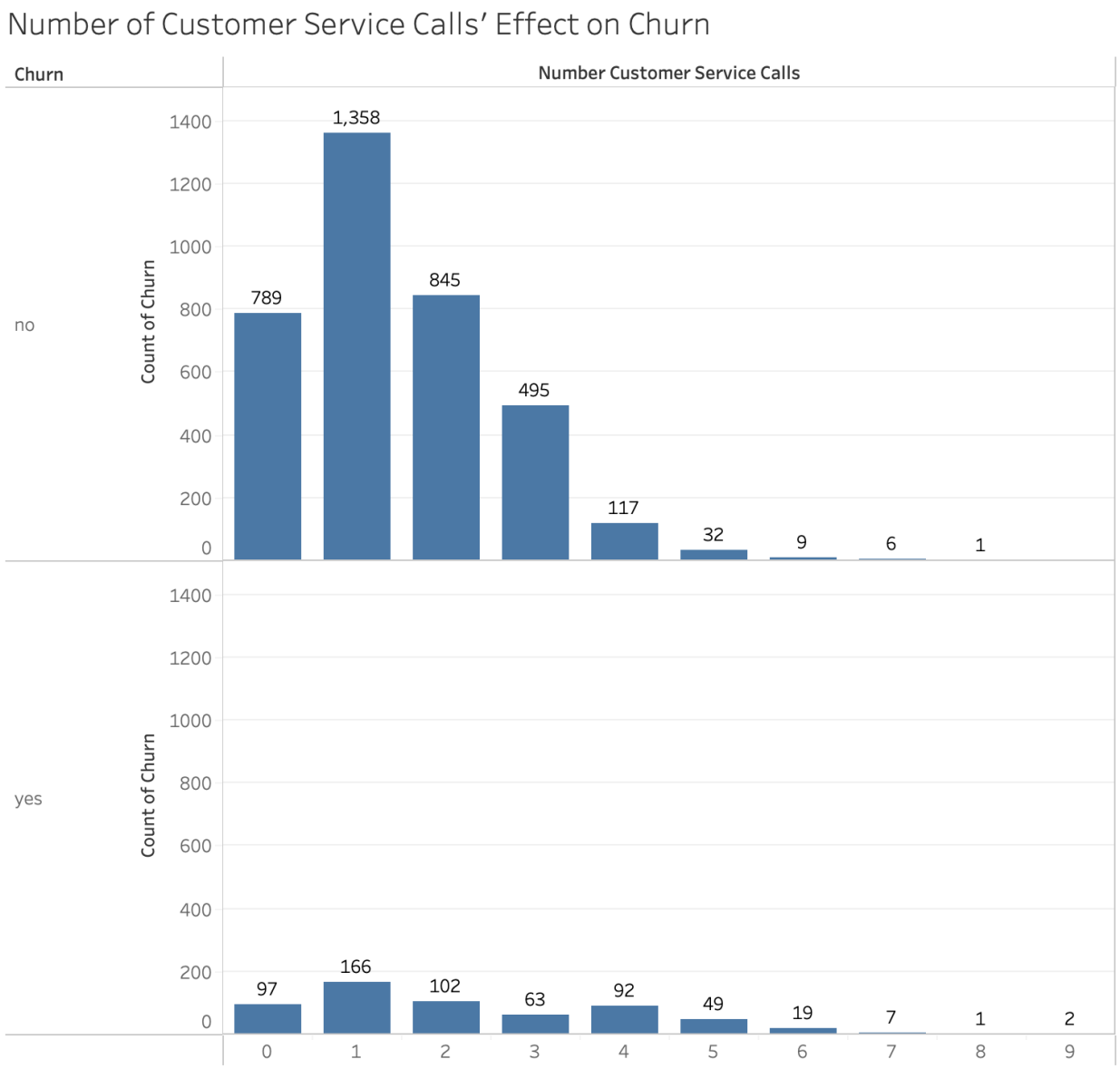
Here, we can see that there are some features which are redundant and are in high correlation with each other.

The first four features don't have any feature that is being redundant hence we consider the same four mentioned above.



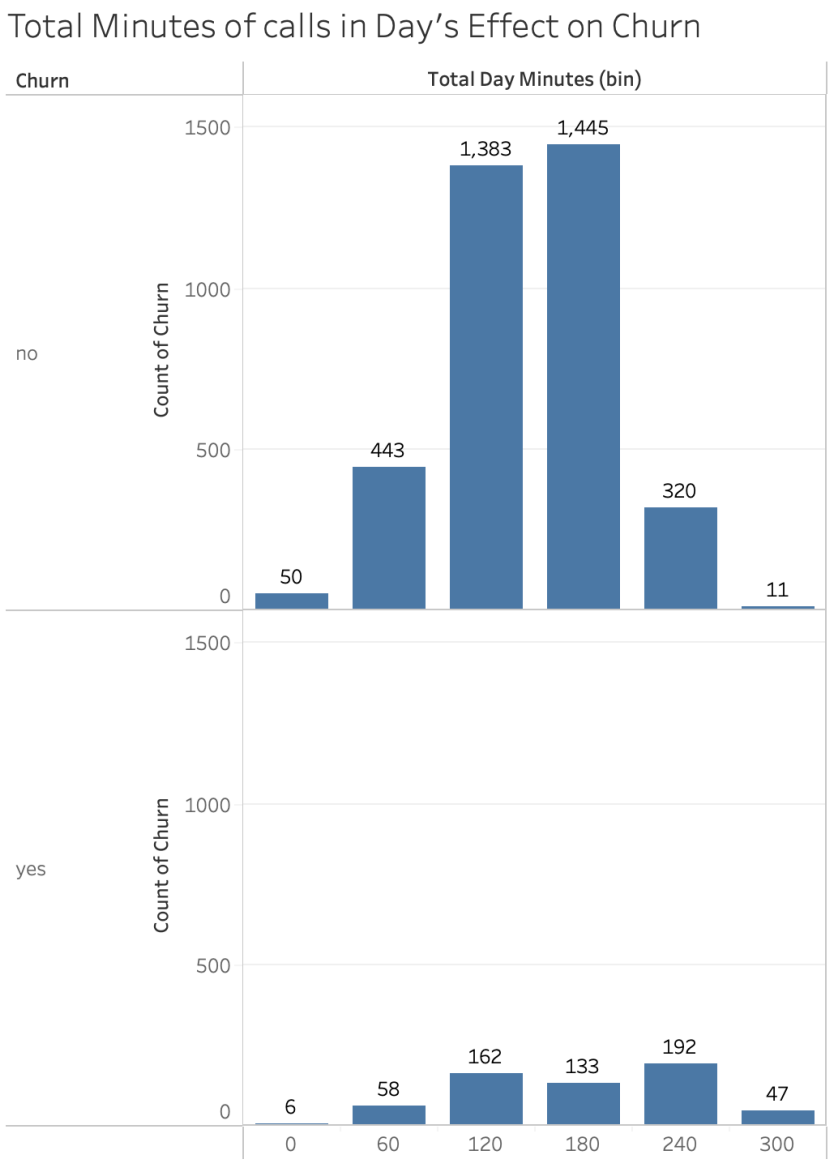
Feature impact list

1st Feature - number_customer_service_calls



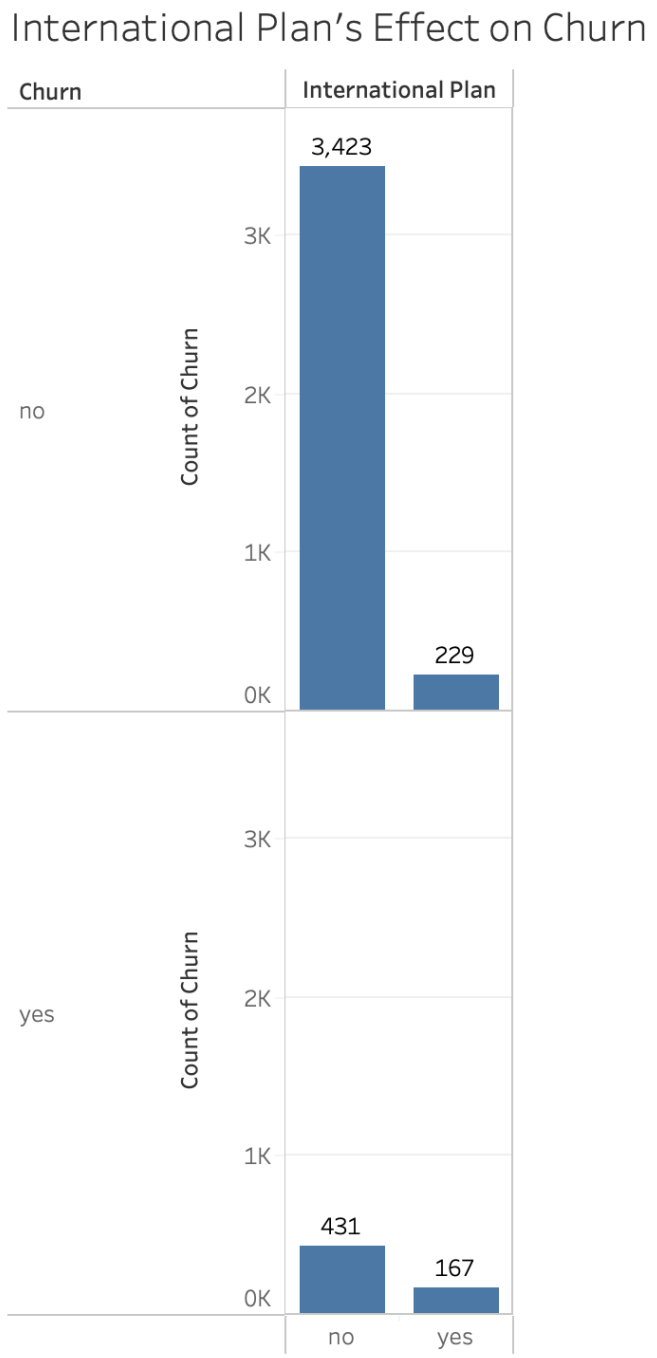
Almost 12% of customers churn with or without customer calls. After a lot of customer service calls, the churn rate increases than the rate of customers staying.

2nd Feature - total_day_minutes



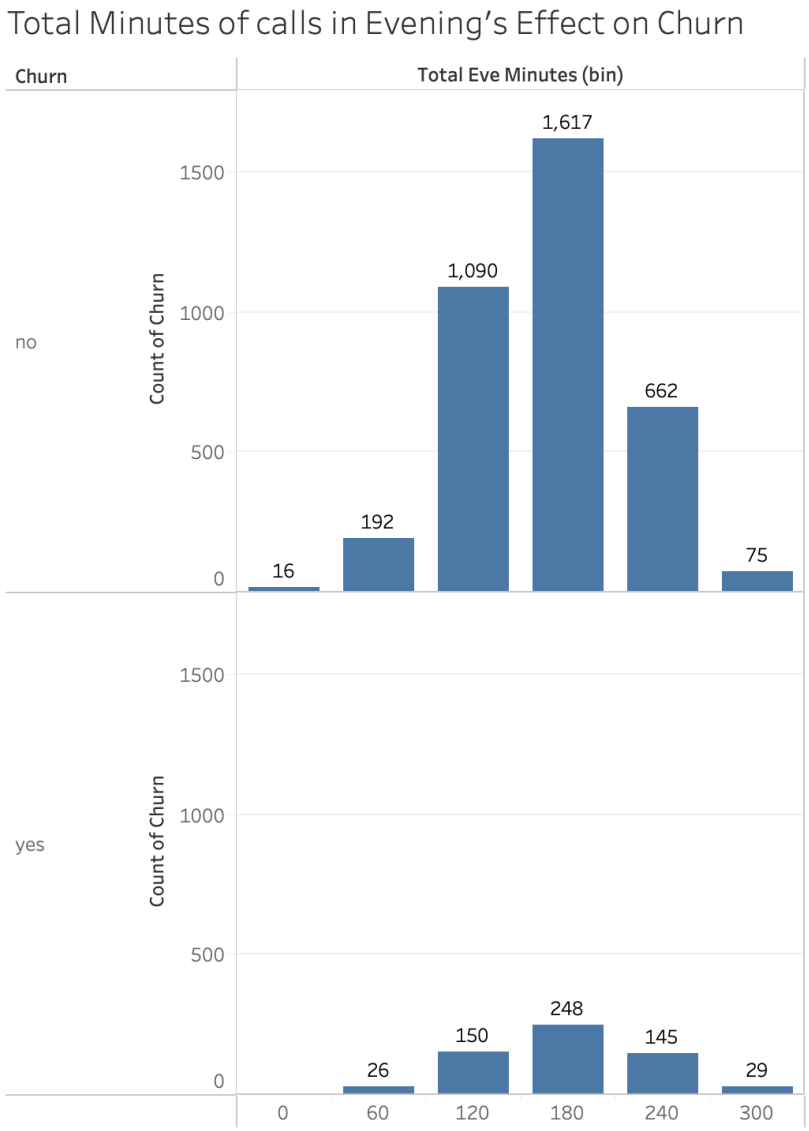
Upto 200 minutes (more than 3 hours) of day's talktime, the churn rate seems to be less than 13% but when the day talk time increases, the churn rate increases too in comparison to total people talking that long.

3rd Feature - international_plans



Only 12% of customers who don't have an international plan churn whereas 73% of customers that have international plans are likely to churn.

4th Feature - total_eve_minutes



People that talk for 180 minutes or less at the evening time are ~15% likely to churn, whereas the ones that talk more than that are likely to churn by ~20% or more.

Q5. Actionable insights on top 4 predictors of customer churn.

1st Feature - number_customer_service_calls

If the issues of the customers are solved in less number of calls then they are less likely to churn, hence improving the customer service by training them better would be an option to decrease churn rate.

2nd Feature - total_day_minutes
People who talk for more than ~3 hours during daytime should be presented with deals and offers as they are more likely to churn.

3rd Feature - international_plans
More than ~70% of customers that have international plans churn, but this may be the case due to them shifting permanently or using a local plan in some other regions. Here, there isn't much actionable insight.

4th Feature - total_eve_minutes
Here too, people that talk for more than 3 hours, like the ones in day time should be provided with offers and deals as they are more likely to churn.