

Assignment 5 - Customer churn prediction

Question 1:

Business Case:

Who is affected : The retention team at Telecom provider

Problem: Customer switching to another Telcom provider for better offers or services known as 'churning'.

Solution: Providing offers/incentives to customers which makes them to stay with their current telecom provider.

We'll focus our evaluation based on a data set containing 4251 values. Various models are used to make the analysis on given features. Models are further then compared against each other to find the better model to maximize the profit.

Payoff matrix:

- Monthly telecom plan for consumer: \$30
- Retention team offers 'buy for 6 months free for 6 months' offer, that is, customer pays for 6 months, and the next 6 months are free. This can be considered as 50% off yearly plan.
- Assuming each customer stays for 12 months with a telecom partner.
- Assuming that retention team pays \$50 hours to its employee, and they make 5 calls every hour. Which means \$10 cost.
- Assuming success rate for each retention is 100%.

True Positive:

Customer who may switch to new telecom plan are accurately identified and are offered discounted plans to opt. In our case they are offered 50% off on yearly plan (Purchase for 6 months , get 6 months free offer).

$$TP: 30 * 12 * 0.50 \text{ (discount)} - \$10 \text{ (call cost)} = \$170$$

False Positive: Incorrect target group of customers were identified who were not planning to switch to another telecom but still were offered discount.

$$FP: 30 * 12 * 0.50 \text{ (discount)} - \$10 = \$170$$

True Negative: Customer group who will not switch to another telecom partner were identified and were not offered any discount.

$$TN: 30 * 12 = \$360$$

False Negative: No calls were made, and no revenue was generated

FN: \$0

	Predicted 0	Predicted 1
Actual 0	360	170
Actual 1	0	170

Ethical and legal considerations:

The customers provide data while signing up for telecom connection, which includes personal information, the telecom provider needs to carefully consider evaluating which personal information can be used to make customized plans for users considering customer's data privacy.

Customer should be informed about the ways the information submitted by them can be used by the telecom providers.

Question 2:








Feature list:

<input type="checkbox"/> Feature Name	Data Quality	Index	Importance ↑	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
<input type="checkbox"/> total_day_minutes		7	<div><div></div></div>	Numeric	1,682	0	180	54.24	180	0	347
<input type="checkbox"/> total_day_charge		9	<div><div></div></div>	Numeric	1,682	0	30.57	9.22	30.67	0	58.96
<input type="checkbox"/> number_customer_service_calls		19	<div><div></div></div>	Numeric	10	0	1.57	1.31	1	0	9
<input type="checkbox"/> international_plan		4	<div><div></div></div>	Categorical	2	0					
<input type="checkbox"/> voice_mail_plan		5	<div><div></div></div>	Categorical	2	0					
<input type="checkbox"/> number_vmail_messages		6	<div><div></div></div>	Numeric	46	0	7.73	13.53	0	0	52
<input type="checkbox"/> total_intl_calls		17	<div><div></div></div>	Numeric	21	0	4.42	2.46	4	0	20
<input type="checkbox"/> total_intl_charge		18	<div><div></div></div>	Numeric	161	0	2.78	0.74	2.81	0	5.40
<input type="checkbox"/> total_intl_minutes		16	<div><div></div></div>	Numeric	161	0	10.28	2.75	10.40	0	20
<input type="checkbox"/> total_eve_minutes		10	<div><div></div></div>	Numeric	1,613	0	200	50.16	201	0	359
<input type="checkbox"/> total_eve_charge		12	<div><div></div></div>	Numeric	1,439	0	17.01	4.26	17.06	0	30.54
<input type="checkbox"/> total_night_minutes		13	<div><div></div></div>	Numeric	1,597	0	201	50.12	201	23.20	382
<input type="checkbox"/> total_night_charge		15	<div><div></div></div>	Numeric	941	0	9.03	2.26	9.05	1.04	17.19
<input type="checkbox"/> state		1	<div><div></div></div>	Categorical	51	0					
<input type="checkbox"/> area_code		3	<div><div></div></div>	Categorical	3	0					
<input type="checkbox"/> total_day_calls	1	8	<div><div></div></div>	Numeric	118	0	99.68	19.95	100	0	165
<input type="checkbox"/> total_eve_calls	1	11	<div><div></div></div>	Numeric	120	0	100	20.02	100	0	170
<input type="checkbox"/> account_length		2	<div><div></div></div>	Numeric	212	0	101	39.70	100	1	243
<input type="checkbox"/> total_night_calls	1	14	<div><div></div></div>	Numeric	124	0	99.77	19.99	100	33	170

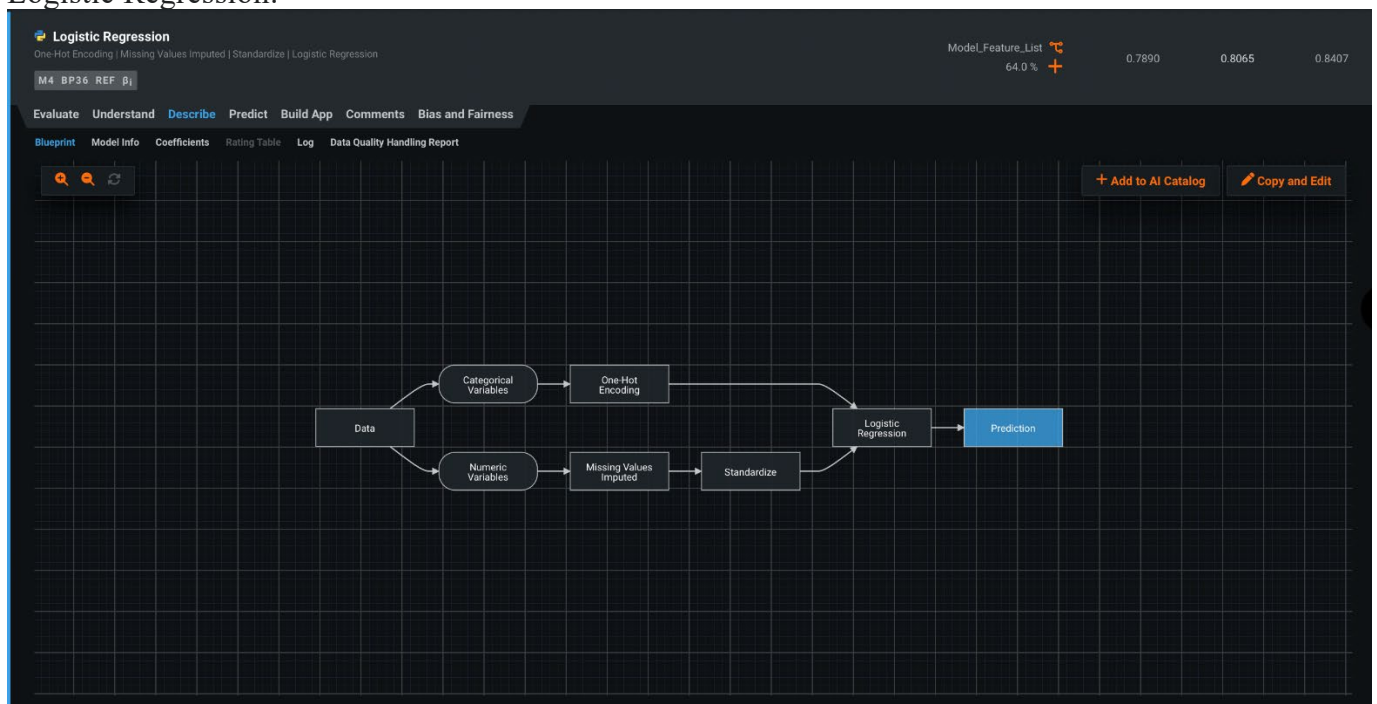
The following models were run on the given data:

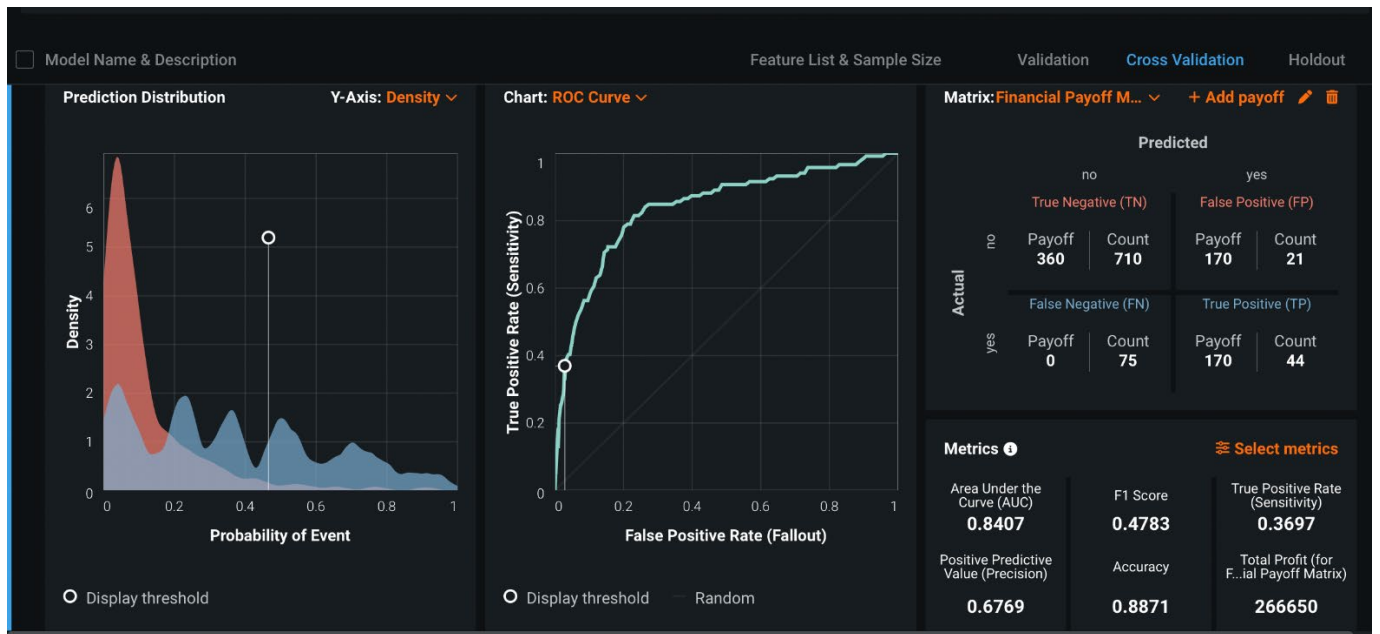
<input type="checkbox"/> Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
<div><div><div>XGBoost</div><div>eXtreme Gradient Boosted Trees Classifier with Unsupervised Learning Features</div><div>Ordinal encoding of categorical variables Missing Values Imputed Search for differences Standardize One-Hot Encoding Partial Principal Components Analysis K-Means Clustering eXtreme Gradient Boosted Trees Classifier with Unsupervised Learning Features</div><div>M46BP66</div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9111	0.9205	0.9083
<div><div><div>XGBoost</div><div>eXtreme Gradient Boosted Trees Classifier</div><div>Category Count Missing Values Imputed Isolation Forest Anomaly Detection Ordinal encoding of categorical variables eXtreme Gradient Boosted Trees Classifier</div><div>M22BP32</div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9165	0.9193	0.9025
<div><div><div></div><div>Gradient Boosted Trees Classifier</div><div>Ordinal encoding of categorical variables Missing Values Imputed Gradient Boosted Trees Classifier</div><div>M16BP41REF</div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9176	0.9192	0.9122
<div><div><div>XGBoost</div><div>eXtreme Gradient Boosted Trees Classifier</div></div></div>				

<input type="checkbox"/> Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
<div><div><div><div><div>XGBoost</div><div>eXtreme Gradient Boosted Trees Classifier</div><div>Ordinal encoding of categorical variables Missing Values Imputed Search for differences eXtreme Gradient Boosted Trees Classifier</div><div>M28BP65</div><div></div></div></div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9097	0.9187	0.9080
<div><div><div><div><div></div><div>RandomForest Classifier (Gini)</div><div>Ordinal encoding of categorical variables Missing Values Imputed RandomForest Classifier (Gini)</div><div>M64BP45REF</div></div></div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9076	0.9153	0.9057
<div><div><div><div><div></div><div>RandomForest Classifier (Entropy)</div><div>Ordinal encoding of categorical variables Category Count Missing Values Imputed RandomForest Classifier (Entropy)</div><div>M52BP63</div></div></div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9126	0.9152	0.9066
<div><div><div><div><div>XGBoost</div><div>eXtreme Gradient Boosted Trees Classifier (learning rate =0.01)</div><div>One-Hot Encoding Missing Values Imputed Search for differences Search for ratios eXtreme Gradient Boosted Trees Classifier (learning rate =0.01)</div></div></div></div></div>	Model_Feature_List 64.0 % <div><div></div><div>+</div></div>	0.9074	0.9151	0.8961

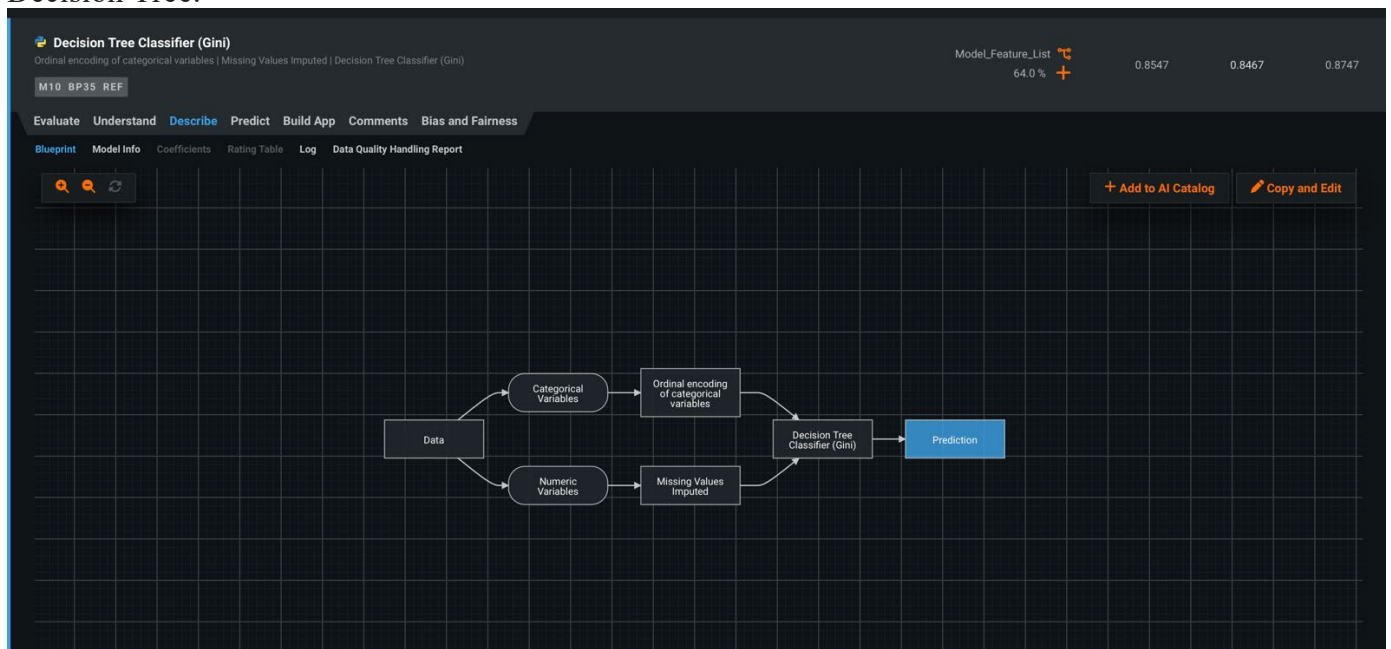
<input type="checkbox"/>	Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
	RandomForest Classifier (Entropy) (Shallow)				
	Ordinal encoding of categorical variables Category Count Missing Values Imputed RandomForest Classifier (Entropy) (Shallow)	Model_Feature_List 64.0 % 	0.9123	0.9139	0.9050
	M58 BP23				
	eXtreme Gradient Boosted Trees Classifier (learning rate =0.01)				
	One-Hot Encoding Univariate credibility estimates with ElasticNet Category Count Missing Values Imputed Search for differences Search for ratios eXtreme Gradient Boosted Trees Classifier (learning rate =0.01)	Model_Feature_List 64.0 % 	0.9003	0.9083	0.8904
	M34 BP29				
	Decision Tree Classifier (Gini)				
	Ordinal encoding of categorical variables Missing Values Imputed Decision Tree Classifier (Gini)	Model_Feature_List 64.0 % 	0.8547	0.8467	0.8747
	M10 BP35 REF				
	Logistic Regression				
	One-Hot Encoding Missing Values Imputed Standardize Logistic Regression	Model_Feature_List 64.0 % 	0.7890	0.8065	0.8407

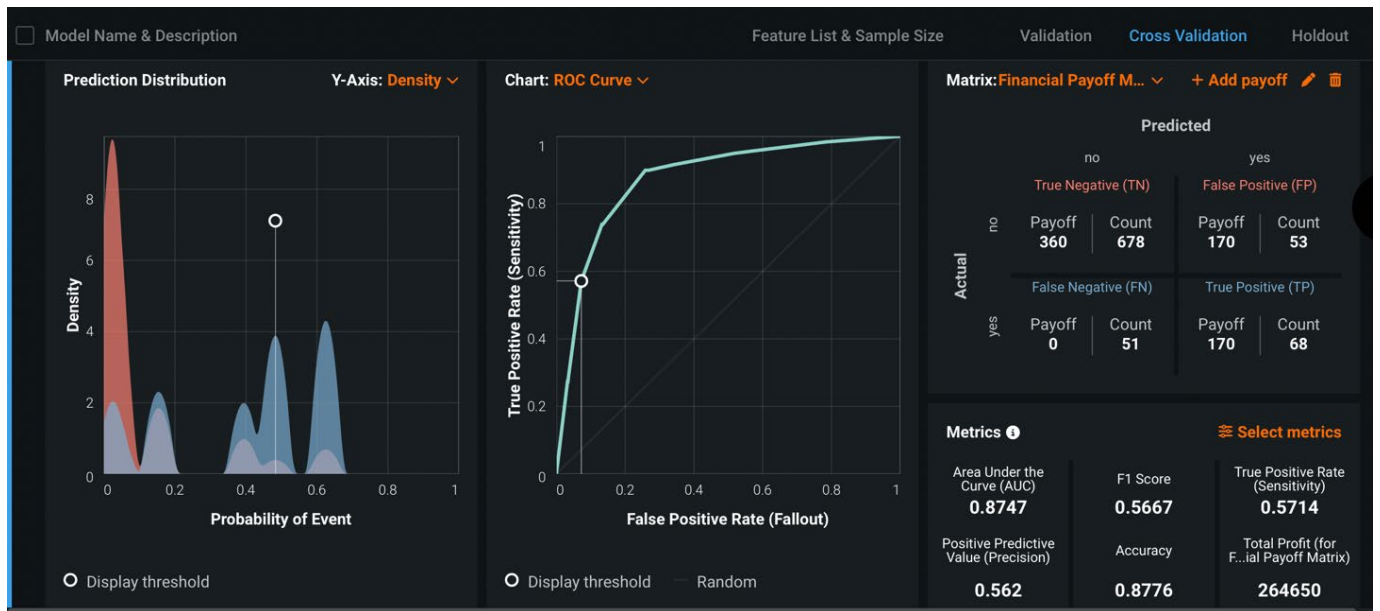
Logistic Regression:



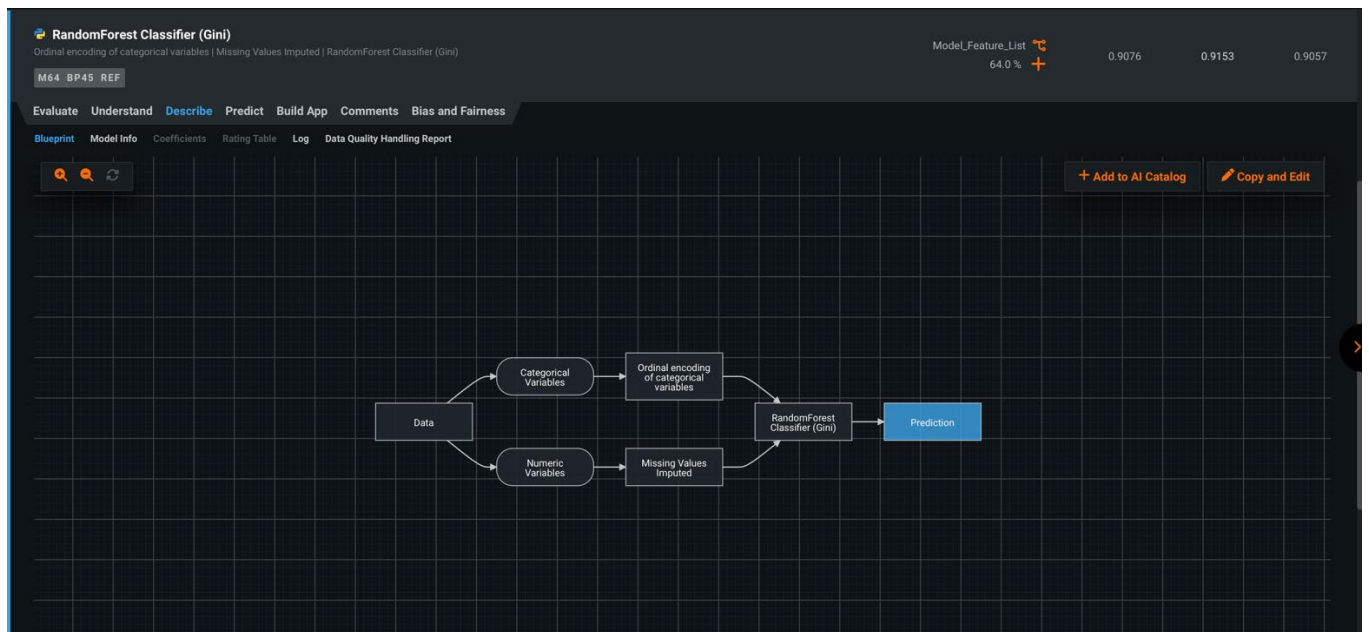


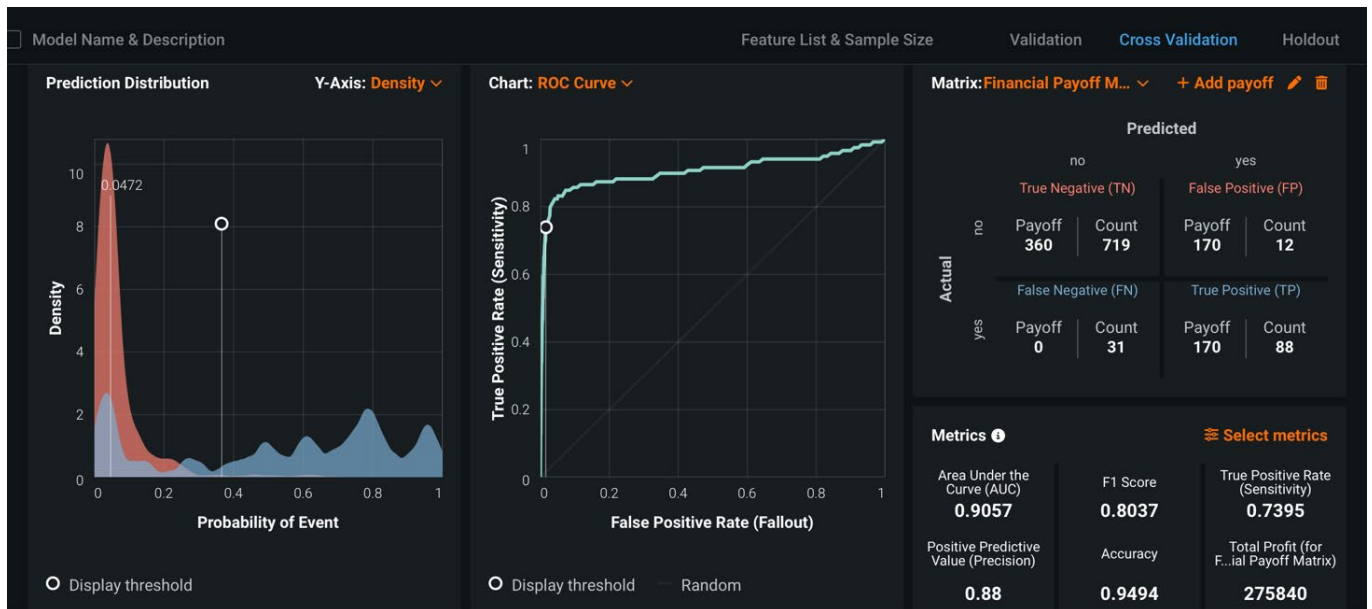
Decision Tree:



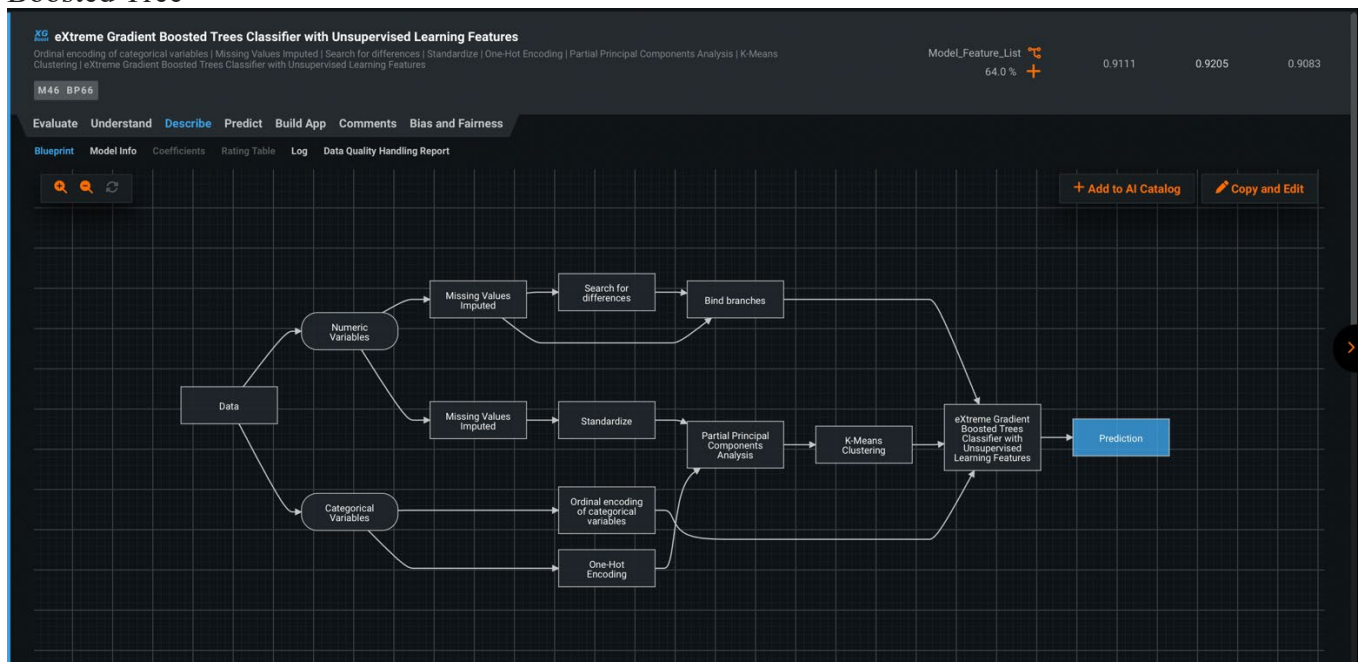


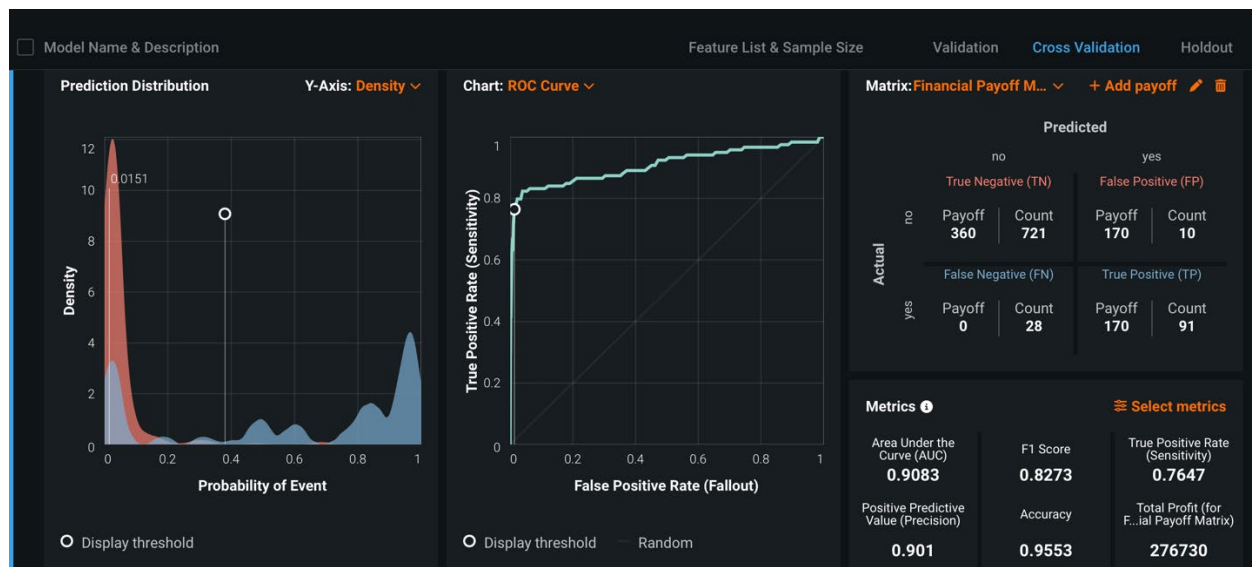
Random Forest:





Boosted Tree




Matrix: Financial Payoff M...
+ Add payoff

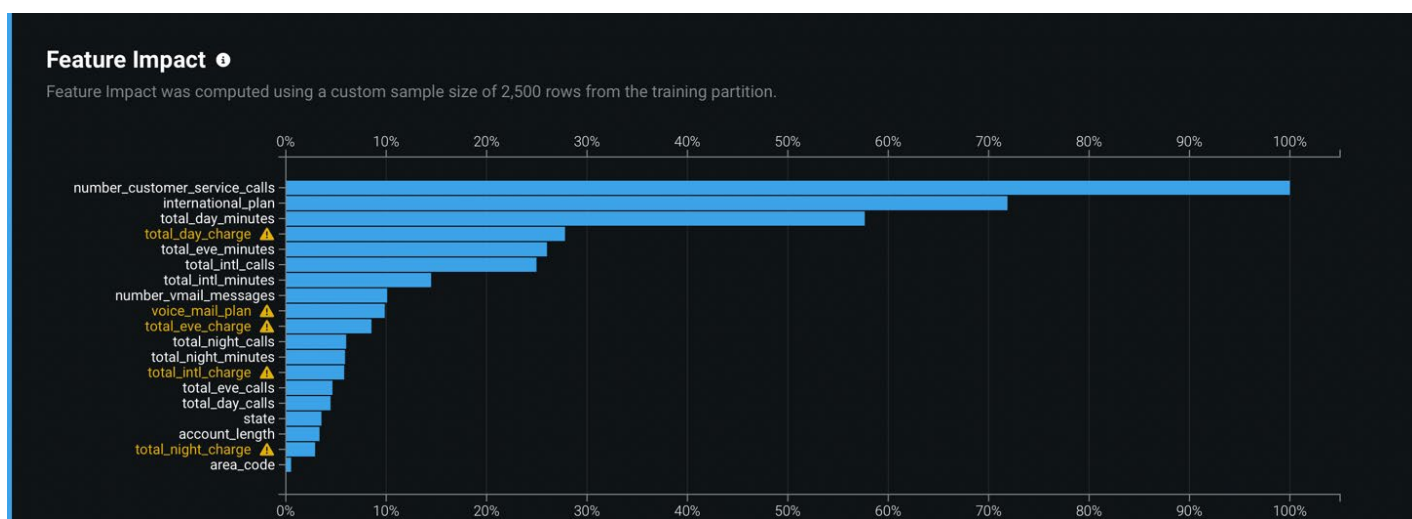
Metrics
Select metrics

Model	Recall	Precision	F1	Accuracy	Error	ROC AUC	Maximum Payoff
Logistic Regression (Threshold: 0.4657)	0.3697	0.6769	0.4783	0.8871	0.1129	0.8407	\$266,650
Decision Tree (Threshold: 0.4828)	0.5714	0.562	0.5667	0.8776	0.1224	0.8747	\$264,650
Boosted Tree (eXtreme Gradient Boosted Trees Classifier with Unsupervised Learning Features) (Threshold: 0.3803)	0.7647	0.901	0.8273	0.9553	0.0447	0.9083	\$276,730
Random Forest (Random Forest Classifier (Gini)) (Threshold: 0.3655)	0.7395	0.88	0.8037	0.9494	0.0506	0.9057	\$275,840

The best performing model is Boosted Tree Model. It provides the highest payoff metric of \$276,730. Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to evaluate the profit based on the number of customers who might consider changing the telecom provider and giving them incentives/offers to stay with them. 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. As our business case requires us to evaluate the potential customer who may opt for different telecom provider and be given a counter offer to stay, and hence we need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

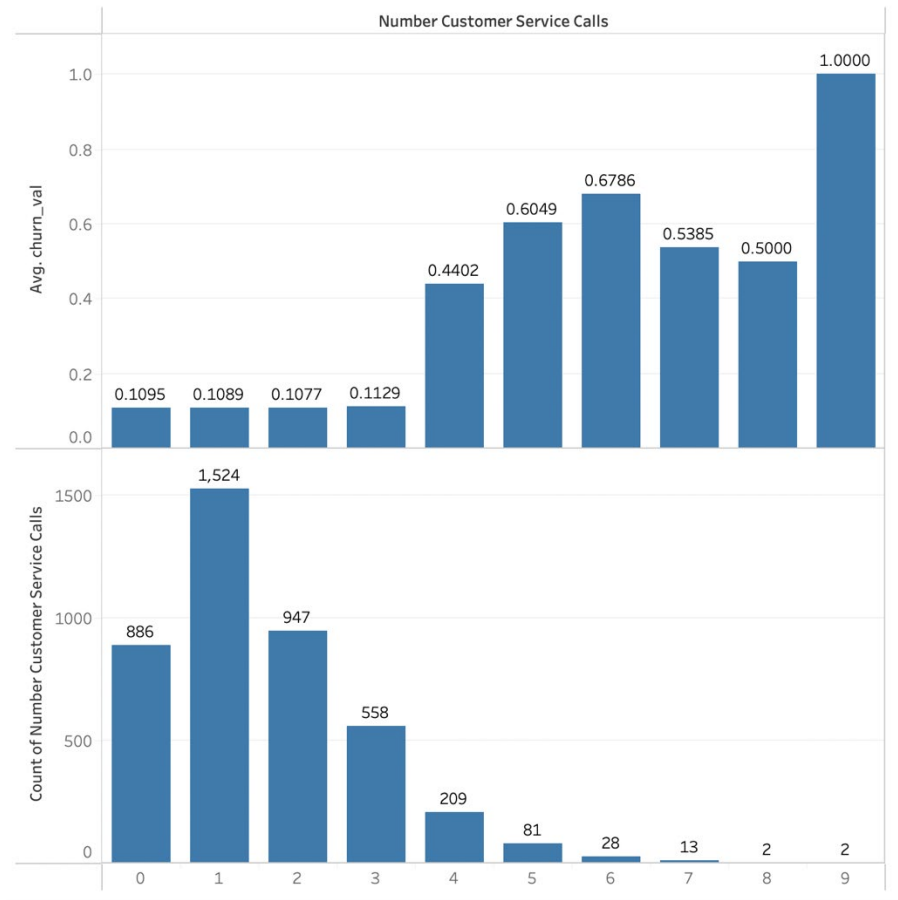
Question 3:

Few features have been marked as redundant features, however, removing the same has no effects on the model performance



Feature 1: Number of customer service calls:

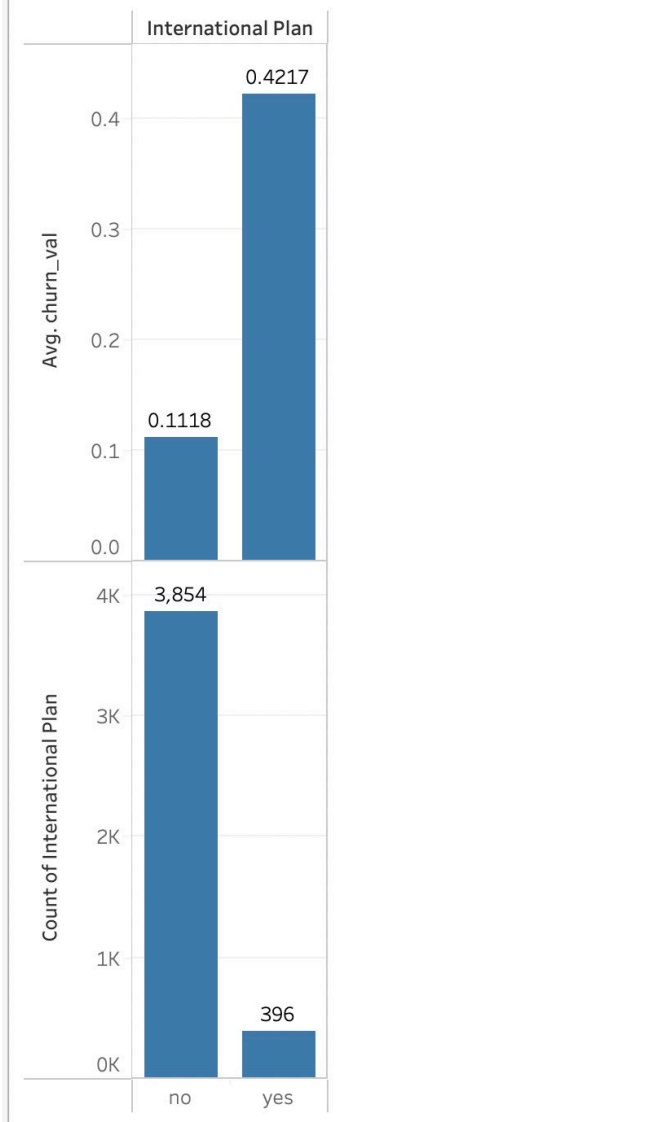
Impact of Customer service calls on Churn



It is observed that churn rate is low 11.29% for call counts upto 3. Churn rate increases by more than 4x once the call count is between 4-6. Churn rate decreases to 50% when then the call count is 7 or 8 and increases to 100% when count is 9 but the observed count is too low to make a generalized observation.

Feature 2: International Plan

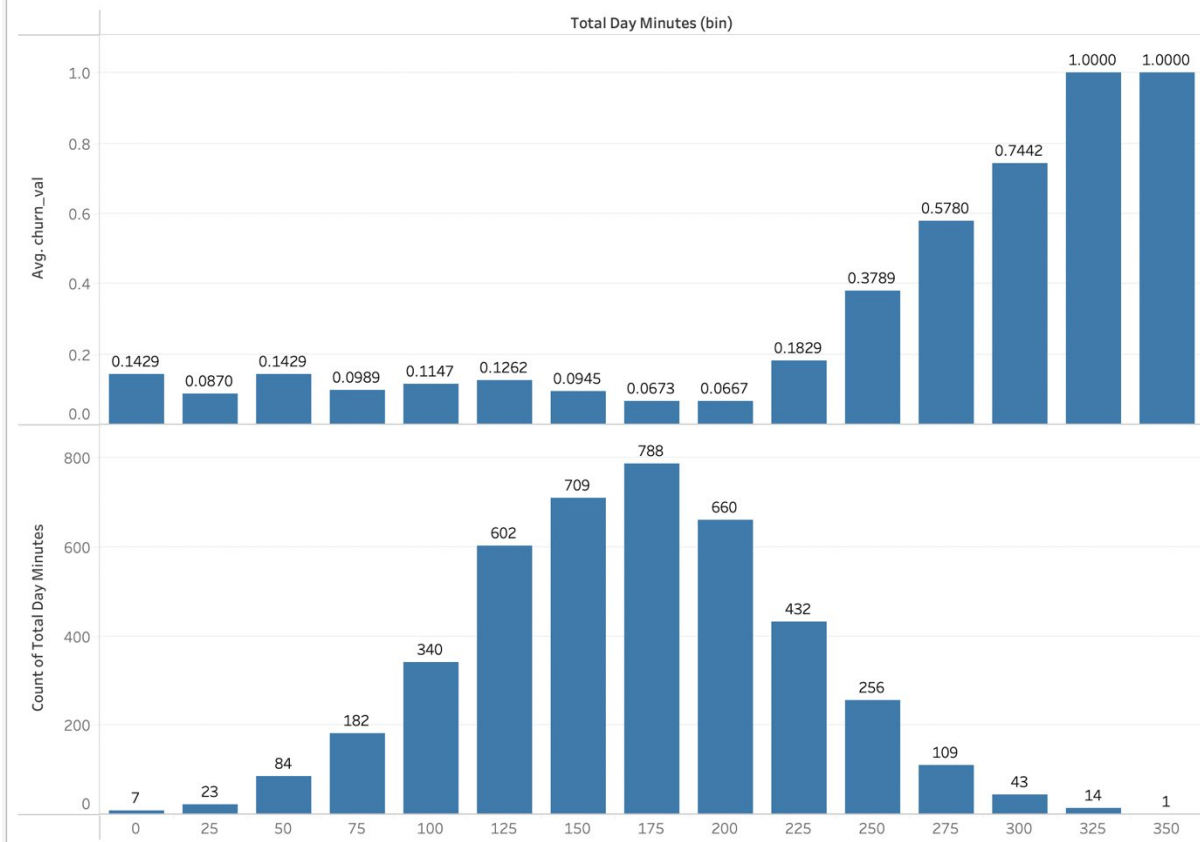
Impact of International Plan on Churn



It is observed that the even though number of customers who have not subscribed to international plan is high (count is 3854 in our case) churn rate is low for customers who do not have international plan that is 11.18%. However, the churn rate is 42.17% for people with international plans which is roughly 4 times.

Feature 3: Total day minutes

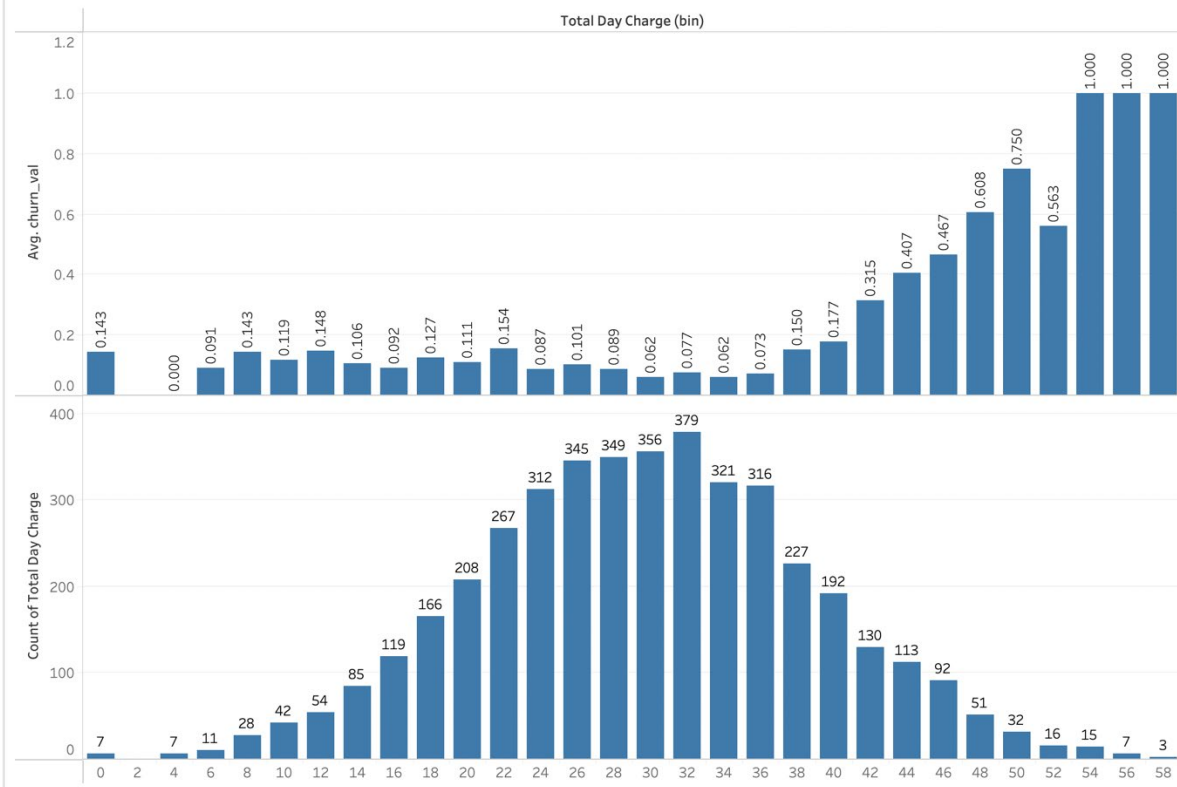
Impact of total day minutes on Churn



It is observed that as soon as total minutes of the day call increase there is an increase in the churn rate. Being around 14.29% for 0-25 minutes it increases to close to 100% . The churn rate at 350 minutes is 100% but the count is too low (1 in our case) to make any assumptions based on this.

Feature 4: Total Day charge

Impact of total day charge on Churn



It is observed as the total day charge increases the churn rate also increases. Starting at 14.3% for \$0-\$2 to 75% at \$50-\$52. Which is an increase of approximately 5x percentage.

Question 4:

Observed Effects	Recommendations
<p><u>Feature: Number of customer service call</u> It is observed that churn rate is low 11.29% for call counts upto 3. Churn rate increases by more than 4x once the call count is between 4-6. Churn rate decreases to 50% when then the call count is 7 or 8 and increases to 100% when count is 9 but the observed count is too low to make a generalized observation.</p>	<p>Customers issues should be solved on priority. They should not call multiple times to follow up, as observed the churn rate increases if one has to call multiple times.</p>
<p><u>Feature: International Plan</u> It is observed that the even though number of customers who have not subscribed to international plan is high (count is 3854 in our case) churn rate is low for customers who do not have international plan that is 11.18%. However, the churn rate is 42.17% for people with international plans which is roughly 4 times.</p>	<p>Telecom provider should come up with better international plans and provide offers on the same.</p>
<p><u>Feature: Total Day Minutes</u> It is observed that as soon as total minutes of the day call increase there is an increase in the churn rate. Being around 14.29% for 0-25 minutes it increases to close to 100% . The churn rate at 350 minutes is 100% but the count is too low (1 in our case) to make any assumptions based on this.</p>	<p>Customers need better plans for longer calls. Offers should be planned around the minutes offered to a customer.</p>
<p><u>Feature: Total Day charge</u> It is observed as the total day charge increases the churn rate also increases. Starting at 14.3% for \$0-\$2 to 75% at \$50-\$52. Which is an increase of approximately 5x percentage.</p>	<p>Customers should be provided offers/ incentives for using the service more this will increase the traffic of users using the services.</p>