



Capstone Project: Telco Churn

Jethro Low



Agenda

1. Problem Statement
2. Data Dictionary
3. Data Cleaning
4. EDA and Feature Engineering
5. Modelling
6. Shapley Values
7. Recommendations

Problem Statement

Problem Statement: From the perspective of a data analytics consultancy, to offer client (Telco) advise on how to reduce the churn rate.

Stakeholders: Client (Telco)

Method: Usage of machine learning methods to create a model to predict the churn.

Metric: Accuracy

Dataset: Kaggle Dataset.

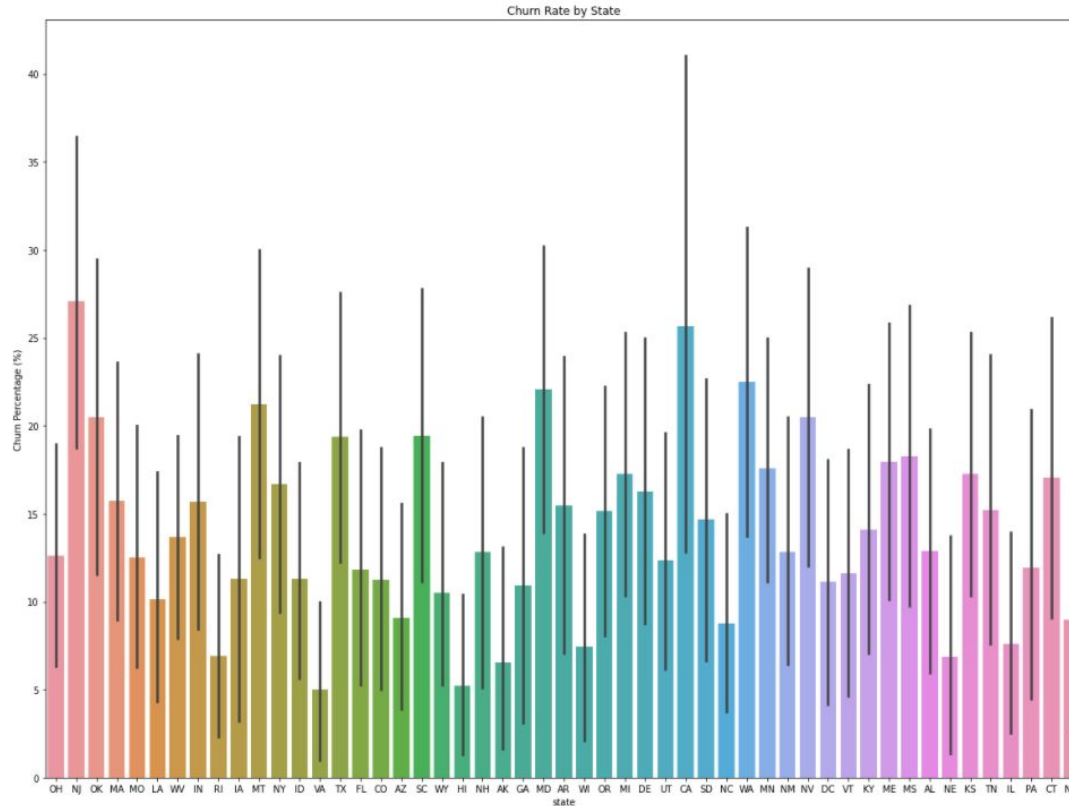
(Link: <https://www.kaggle.com/c/customer-churn-prediction-2020/overview>)

End-State: Advise that can be implemented to help the client reduce the churn rate.

Data Dictionary

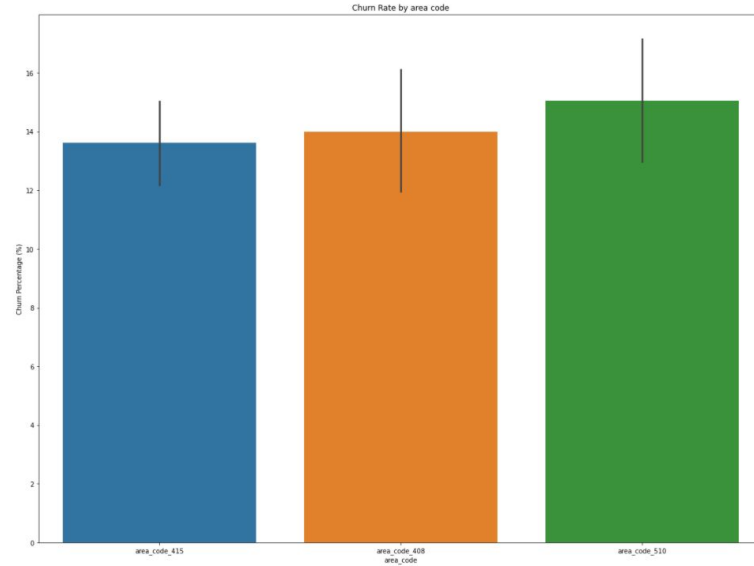
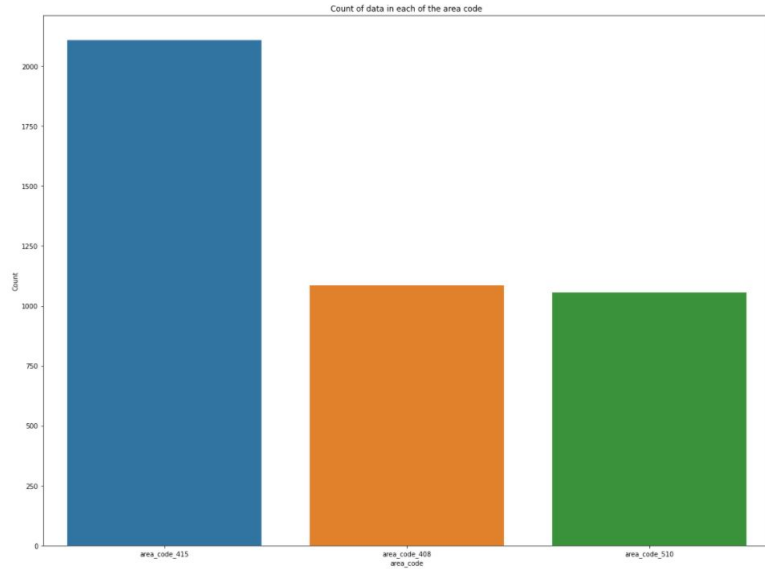
1.	"state", <i>string</i> .	2-letter code of the US state of customer residence
2.	"account_length", <i>numerical</i> .	Number of months the customer has been with the current telco provider
3.	"area_code", <i>string</i> ="area_code_AAA"	where AAA = 3 digit area code.
4.	"international_plan", (<i>yes/no</i>).	The customer has international plan.
5.	"voice_mail_plan", (<i>yes/no</i>).	The customer has voice mail plan.
6.	"number_vmail_messages", <i>numerical</i> .	Number of voice-mail messages.
7.	"total_day_minutes", <i>numerical</i> .	Total minutes of day calls.
8.	"total_day_calls", <i>numerical</i> .	Total minutes of day calls.
9.	"total_day_charge", <i>numerical</i> .	Total charge of day calls.
10.	"total_eve_minutes", <i>numerical</i> .	Total minutes of evening calls.
11.	"total_eve_calls", <i>numerical</i> .	Total number of evening calls.
12.	"total_eve_charge", <i>numerical</i> .	Total charge of evening calls.
13.	"total_night_minutes", <i>numerical</i> .	Total minutes of night calls.
14.	"total_night_calls", <i>numerical</i> .	Total number of night calls.
15.	"total_night_charge", <i>numerical</i> .	Total charge of night calls.
16.	"total_intl_minutes", <i>numerical</i> .	Total minutes of international calls.
17.	"total_intl_calls", <i>numerical</i> .	Total number of international calls.
18.	"total_intl_charge", <i>numerical</i> .	Total charge of international calls
19.	"number_customer_service_calls", <i>numerical</i> .	Number of calls to customer service
20.	"churn", (<i>yes/no</i>).	Customer churn - target variable.

EDA - State



- States that have a higher churn rate include WA, CA and NJ
- States that have a lower churn Rate include VA, HI and AK

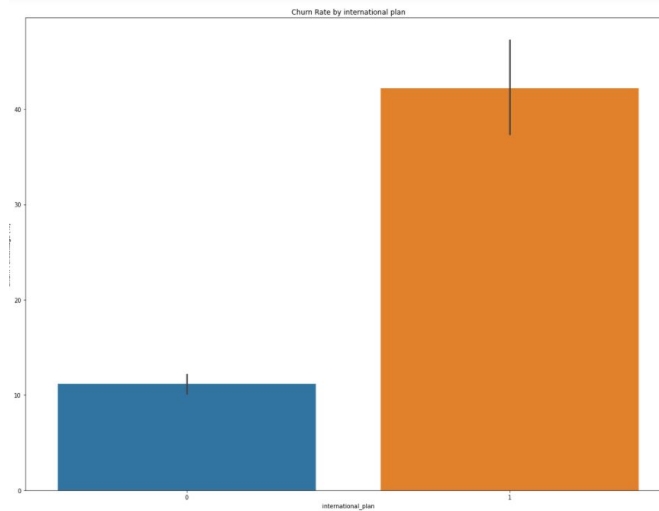
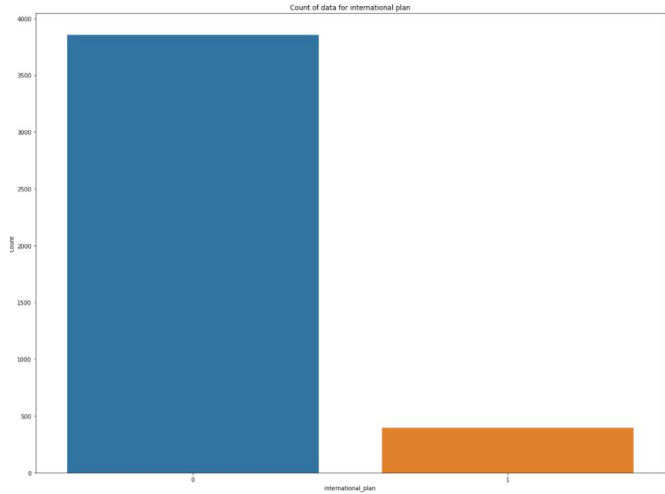
EDA - Area Code



Findings:

- There is not much difference in the churn percentage with respect to the area code.

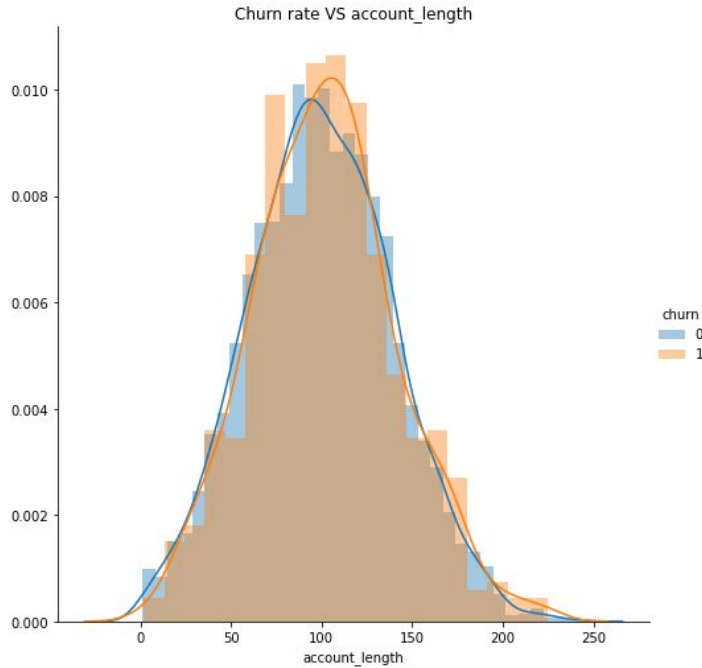
EDA - International Plan



Findings:

- Most people do not have an international plan.
- It seems that customers who have international plan have a much higher probability of churn.
- This makes sense. Please with international plan tend to travel a lot and hence would be more prone to churn

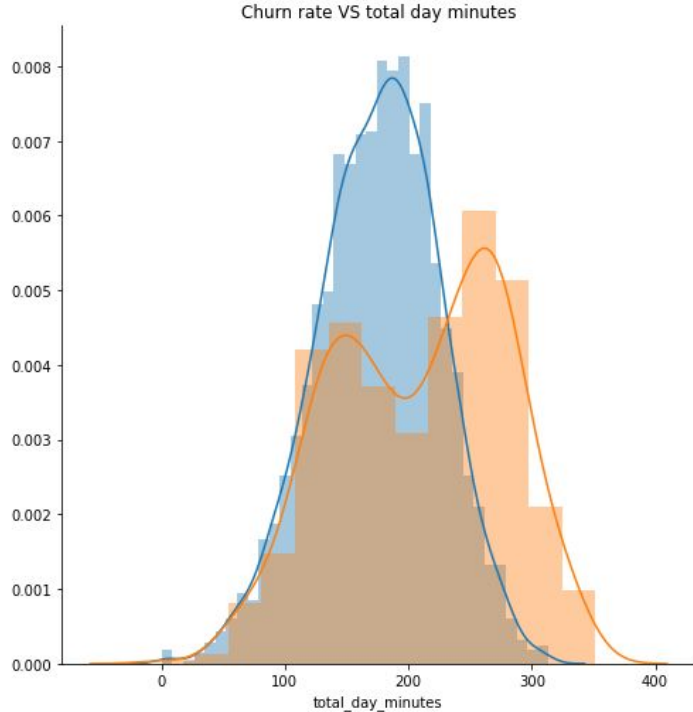
EDA - Account Length



Findings:

- Customers that churn tend to have a slightly higher account_length as compared to customers that do not churn
- This makes sense as people do need some time to get dissatisfied enough to leave.
- It is likely that the account length will be a poor predictor of churn.

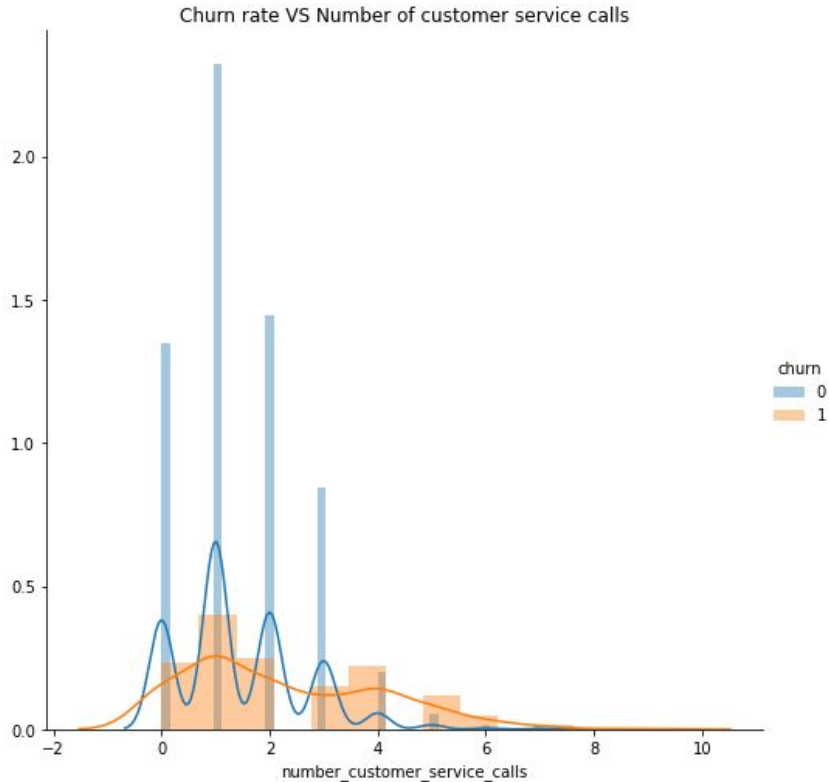
EDA - Total Day Minutes



Findings:

- This is an interesting finding. There is a clear distinction in the number of total_day_minutes between churn and no churn customers.
- Customers that churn tend to have a total day minutes of either 110-130 minutes or 270 to 290 minutes.
- New product offering required?

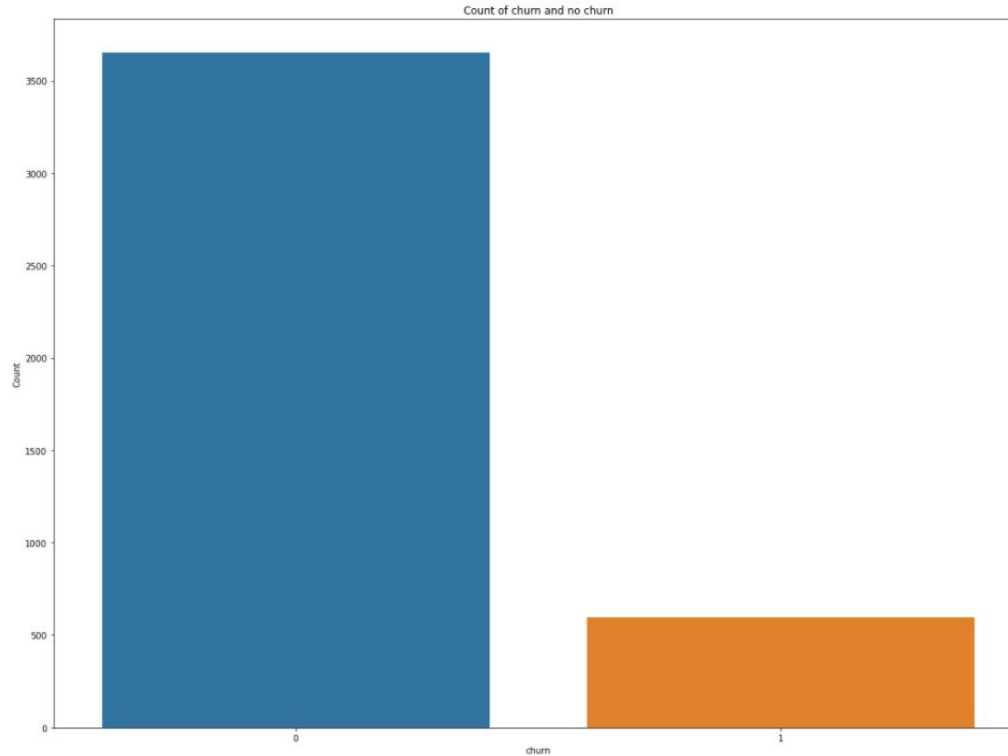
EDA - Customer Service Calls



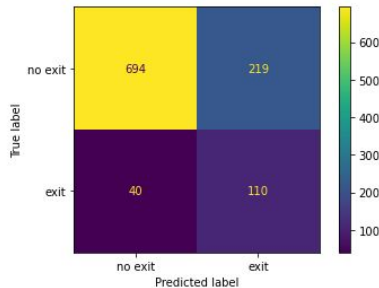
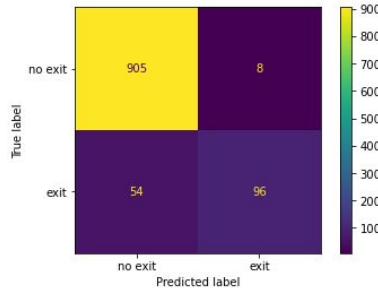
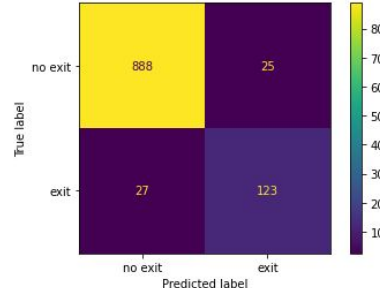
Findings:

- As expected, a higher proportion of churn customers had a higher number of customer service calls.
- This suggests that some of the churn could be due to dissatisfied service or experience with the Telco.

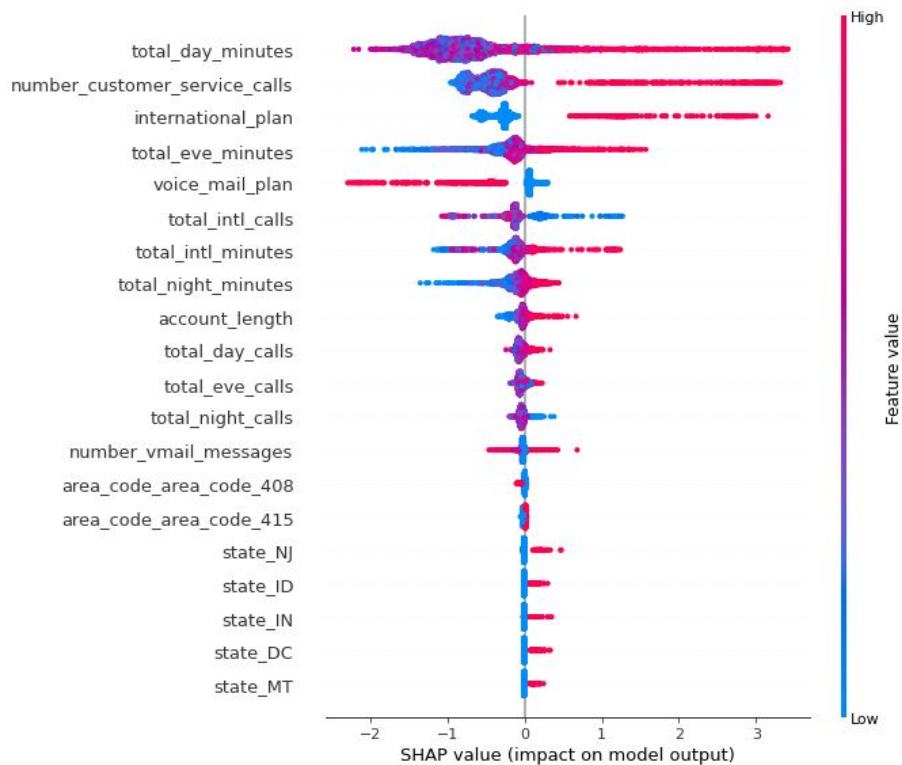
EDA - Imbalance Data



Modelling

Description	Logistic Regression	Random Forest	XGBoost																																				
Data Imbalance	Class Weights	Class Weights	scale_pos_weight																																				
Hyperparameter Tuning	No (Baseline Model)	GridSearch	Manual																																				
Train Score	0.80	0.99	0.96																																				
Test Score	0.76	0.94	0.95																																				
Confusion Matrix	 <p>Confusion Matrix for Logistic Regression. The y-axis is 'True label' with categories 'no exit' and 'exit'. The x-axis is 'Predicted label' with categories 'no exit' and 'exit'. The matrix shows 694 true negatives, 219 false positives, 40 false negatives, and 110 true positives. A color bar on the right indicates counts from 100 to 600.</p> <table><tr><th></th><th>Predicted label</th><th>no exit</th><th>exit</th></tr><tr><th>True label</th><th>no exit</th><td>694</td><td>219</td></tr><tr><th>exit</th><td>40</td><td>110</td><td></td></tr></table>		Predicted label	no exit	exit	True label	no exit	694	219	exit	40	110		 <p>Confusion Matrix for Random Forest. The y-axis is 'True label' with categories 'no exit' and 'exit'. The x-axis is 'Predicted label' with categories 'no exit' and 'exit'. The matrix shows 905 true negatives, 8 false positives, 54 false negatives, and 96 true positives. A color bar on the right indicates counts from 100 to 900.</p> <table><tr><th></th><th>Predicted label</th><th>no exit</th><th>exit</th></tr><tr><th>True label</th><th>no exit</th><td>905</td><td>8</td></tr><tr><th>exit</th><td>54</td><td>96</td><td></td></tr></table>		Predicted label	no exit	exit	True label	no exit	905	8	exit	54	96		 <p>Confusion Matrix for XGBoost. The y-axis is 'True label' with categories 'no exit' and 'exit'. The x-axis is 'Predicted label' with categories 'no exit' and 'exit'. The matrix shows 888 true negatives, 25 false positives, 27 false negatives, and 123 true positives. A color bar on the right indicates counts from 100 to 800.</p> <table><tr><th></th><th>Predicted label</th><th>no exit</th><th>exit</th></tr><tr><th>True label</th><th>no exit</th><td>888</td><td>25</td></tr><tr><th>exit</th><td>27</td><td>123</td><td></td></tr></table>		Predicted label	no exit	exit	True label	no exit	888	25	exit	27	123	
	Predicted label	no exit	exit																																				
True label	no exit	694	219																																				
exit	40	110																																					
	Predicted label	no exit	exit																																				
True label	no exit	905	8																																				
exit	54	96																																					
	Predicted label	no exit	exit																																				
True label	no exit	888	25																																				
exit	27	123																																					

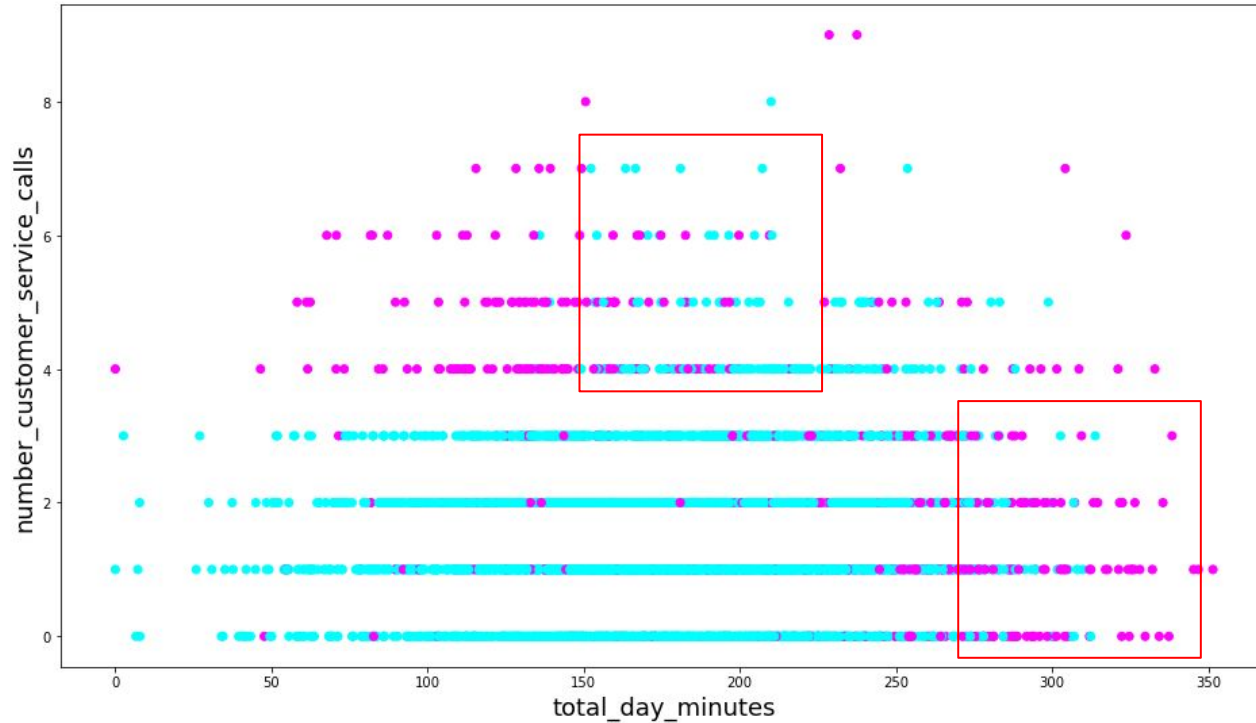
Shapley Values



Recommendation 1

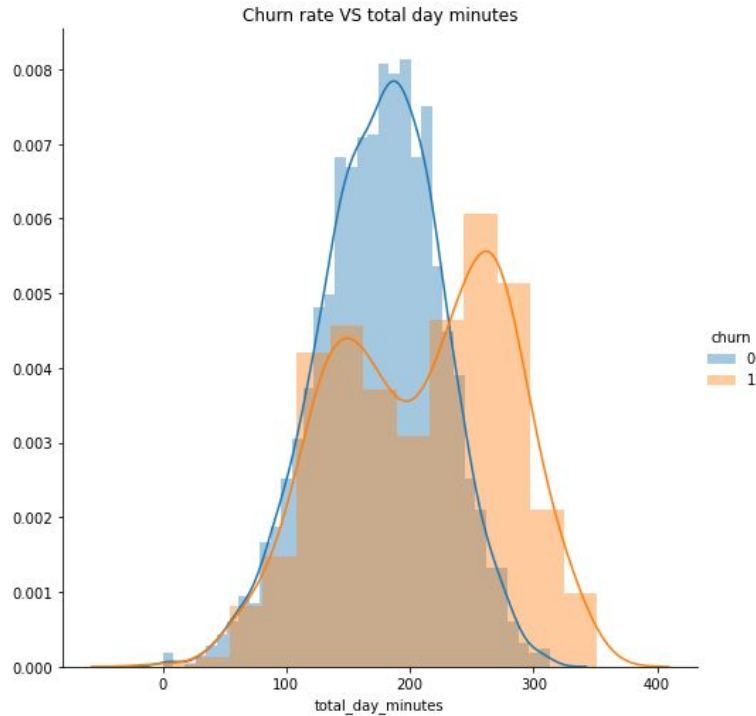
Pink - Churn

Blue - No churr



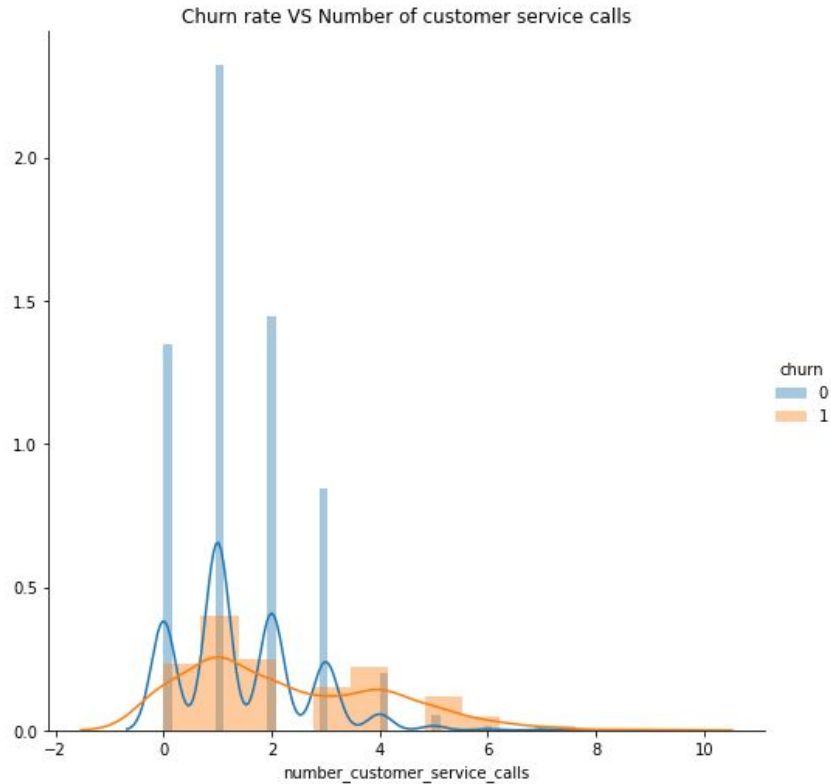
- Address those that may switch Telco but have yet to do so

Recommendation 2



- New product offering could address the lower and higher end of total day minutes

Recommendation 3



- Improve service or experience with Telco