Capstone Three Project – Final Report **Brintha S**
November 2021

Background and Problem Identification

Customer churn is the tendency of customers to leave the service provider or switch to a different provider due to several reasons. Compared to other fields, organizations in the telecommunication industry experience higher churn rates of around 20 – 40% every year (Ahn et al., 2020). Investigating on customer churn rate is important as it affects the KPIs and profitability of the company adversely. In general, acquiring new customers costs much more than retaining existing customers because of expenses relating to marketing promotions. Also, it is believed that existing customers bring new customers through positive word of mouth (Devriendt, et al., 2021). Therefore, this project will focus on predicting the churn rates of customers in a telecom organization. Subsequently, this will help businesses to segment customers who will most likely churn, and thereby to formulate targeted strategies to retain them for a long period. In order to predict if a customer will churn or not, this project uses features such as type of services offered (phone/internet/movie/TV streaming), support services such as tech support, back-up support, tenure period with the company, type of contract(monthly/yearly), monthly charges etc.

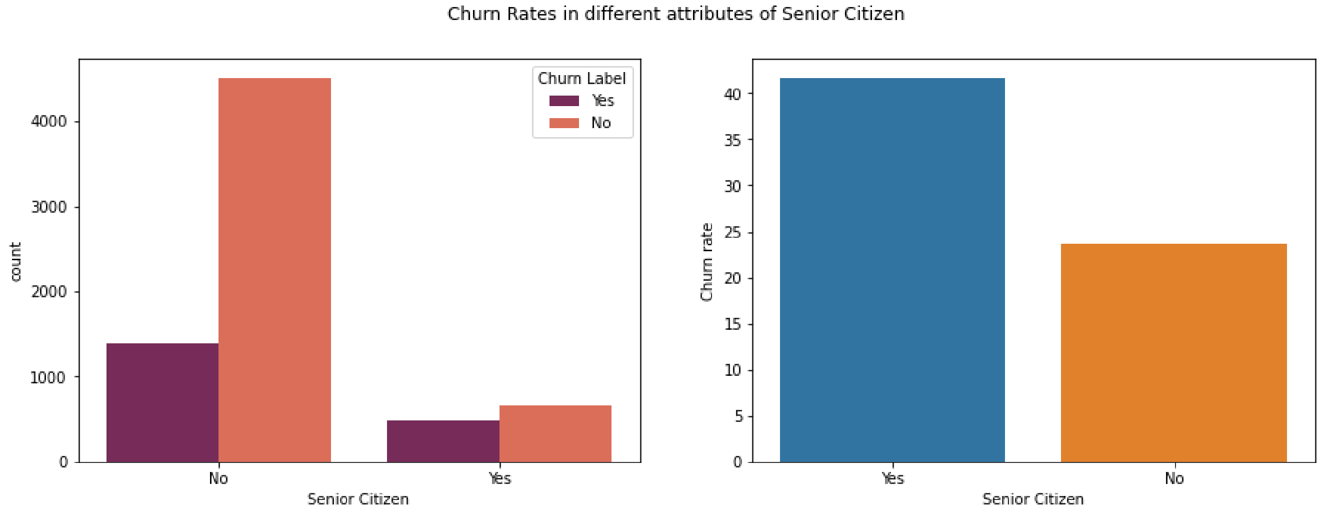
Data and Method

The dataset for the project is from a fictional Telecom company that offers home phone and internet services to customers in California in Q3. Data consists of information of around 7043 customers' service contracts – type of services offered, payment details etc. Data was extracted from IBM Cognos Analytics Data Collection. Source of the data is available from following link. https://community.ibm.com/accelerators/catalog/content/Customer-churn

Exploratory Analysis

The dataset initially consisted of around 30 columns, after choosing features that are deemed as useful in predicting churn rate, there were 17 columns. In general, there was no missing values. However, the total charges for some customers were missing, after determining that those consumers are on their first month, total charges were imputed with their monthly charges. Target variable namely Churn label is a binary variable with Yes (if they have churned) and No (if they haven't churned) options. Most of the other features were categorical except for variables; tenure months, total and monthly charges, Customer Lifetime Value (CLTV). Following plots summarize each feature in terms of how they are related to churn rate among customers. Features like Gender, having Streming_TV/ Streaming movies in their contract did not show much difference in churn, therefore their relationships have been omitted here. Below, on the left is a count plot which compares the count of 'Yes' and 'No' of Churn labels in different attributes of a particular categorical variable. The plot on the right is a bar plot that shows the churn rate as a % (for e.g., out of total senior citizens, how many had churned).

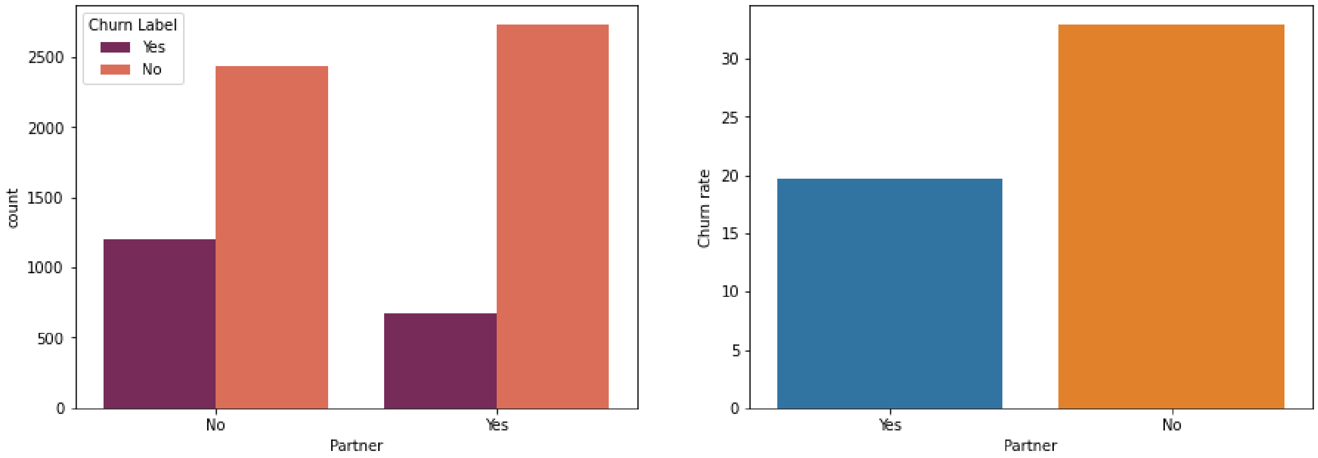
Senior Citizen:



Here the plots reveal that Senior Citizens, churn at least 1.5 times more than that of regular customers.

Partner:

This feature measures if the customer has any partner in holding this contract with the company.

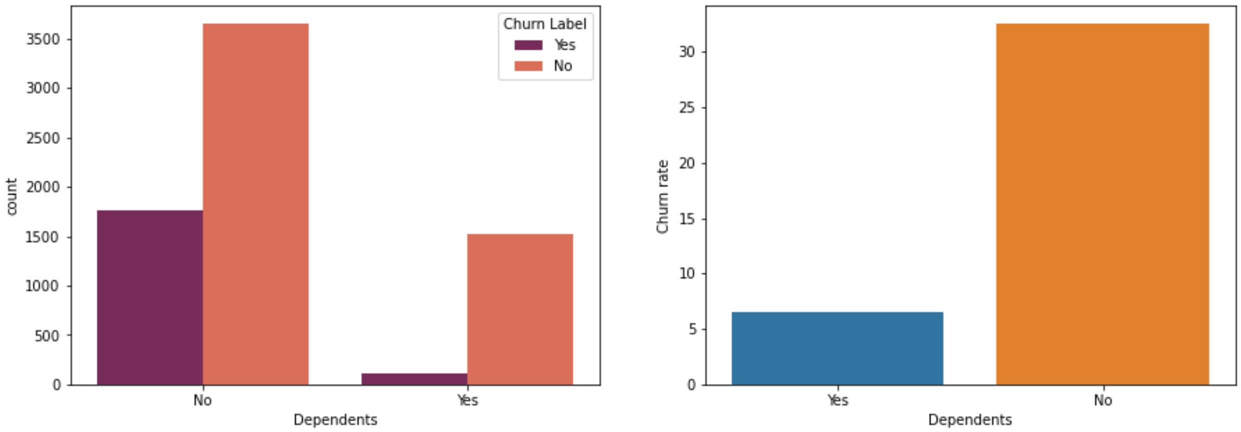


Churn Rates in different attributes of Partner

Plots show that when there's no partner, the churn rates are higher.

Dependents:

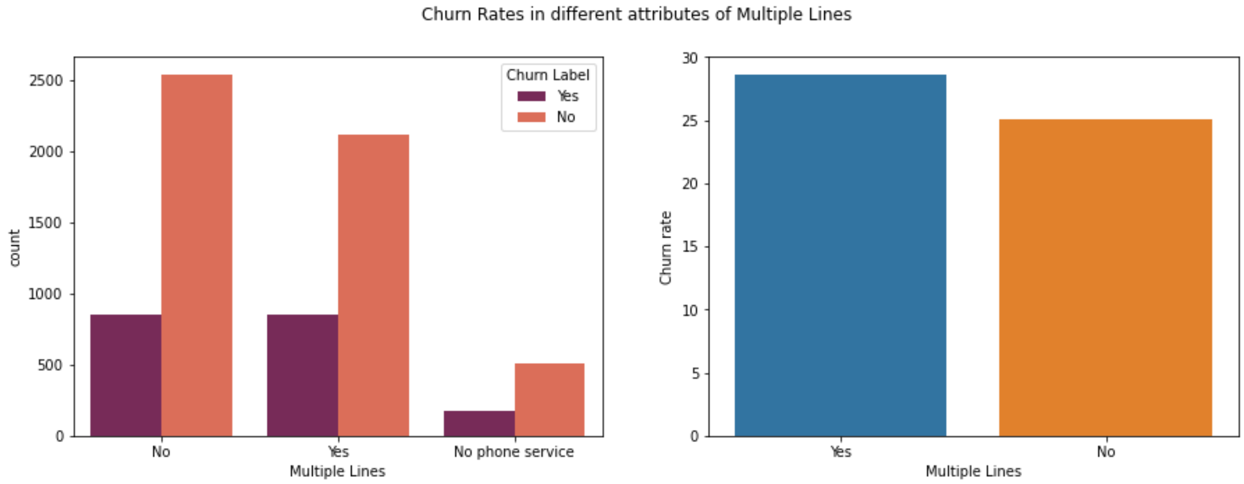
Those who do not have dependents are more likely (more than 4 times) to churn than those with dependents.



Churn Rates in different attributes of Dependents

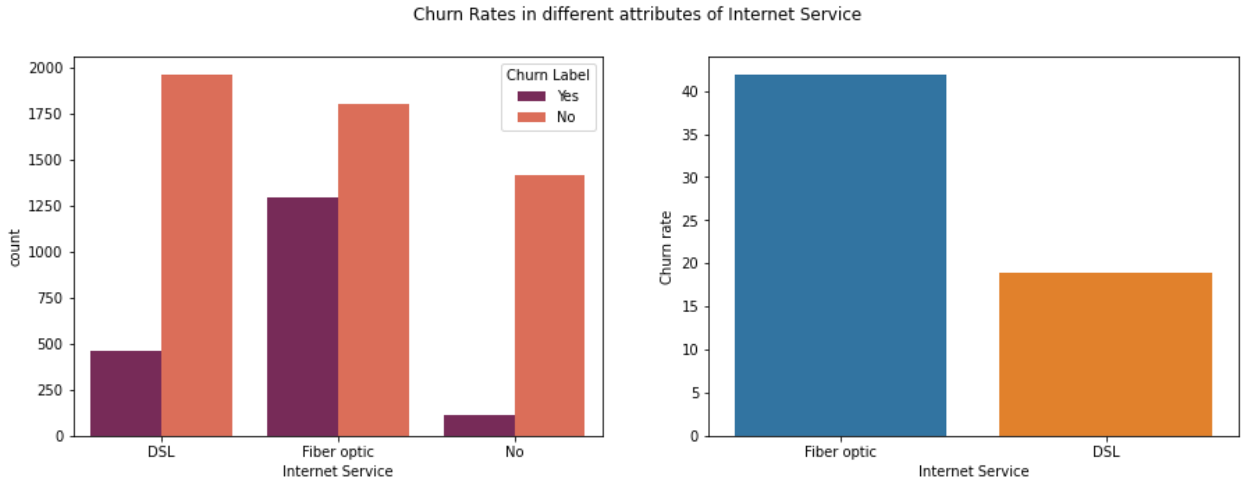
Multiple Lines:

This variable measures, if the customer has more than one phone line in contract. Again, there's only a slight difference in churn rates among those with multiple lines and single phone line.



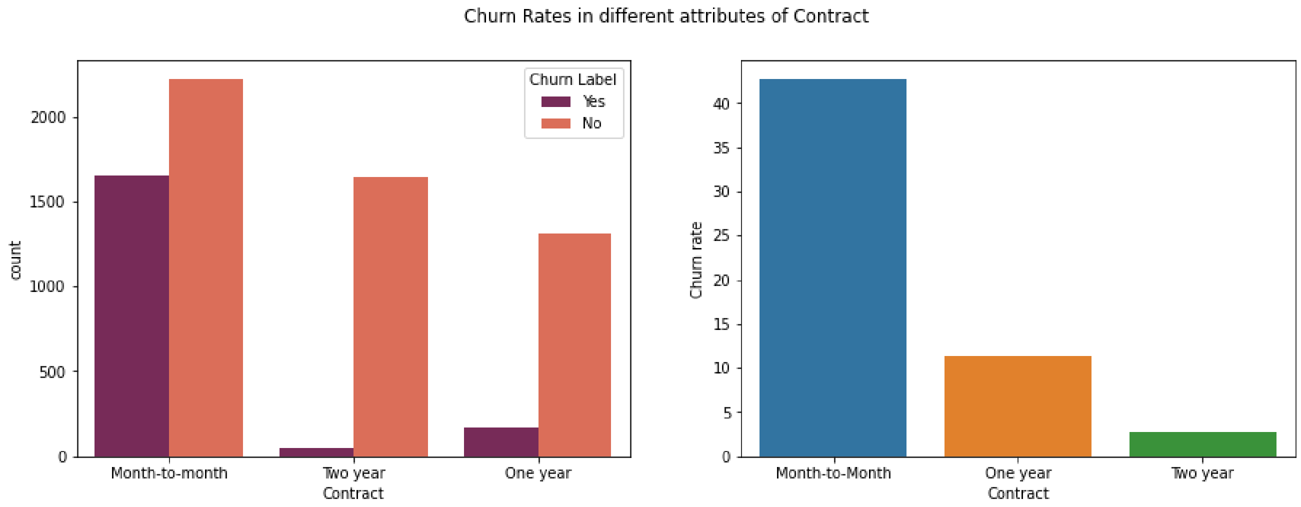
Internet Service:

This variable measures if the customers have internet service, and if they have, if it's DSL (Digital Subscriber Line) or Fiber optic in nature. It turns out that those who have fiber optic internet churn at least 2 times more than that of those with DSL.



Contract:

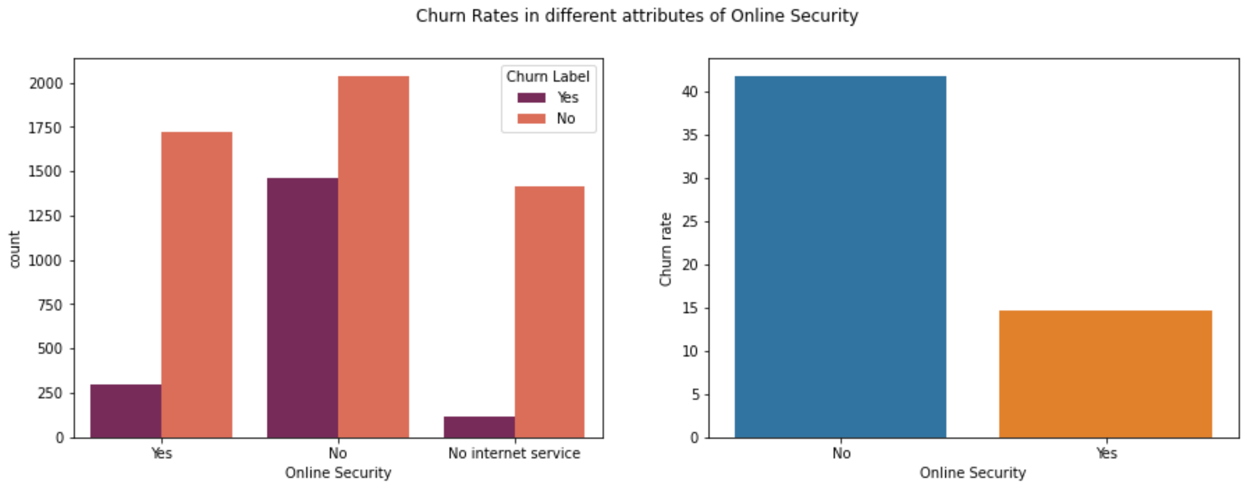
This variable measures, if the customer has a month-to-month or one-year or a two-year contract with the company. As expected, the month-to-month customer churn at least 4 times more than that of one year contract customers and 10 times more than the 2-year contract customers. One year contract ones churn at least 4 times more than that of two-year contract ones.



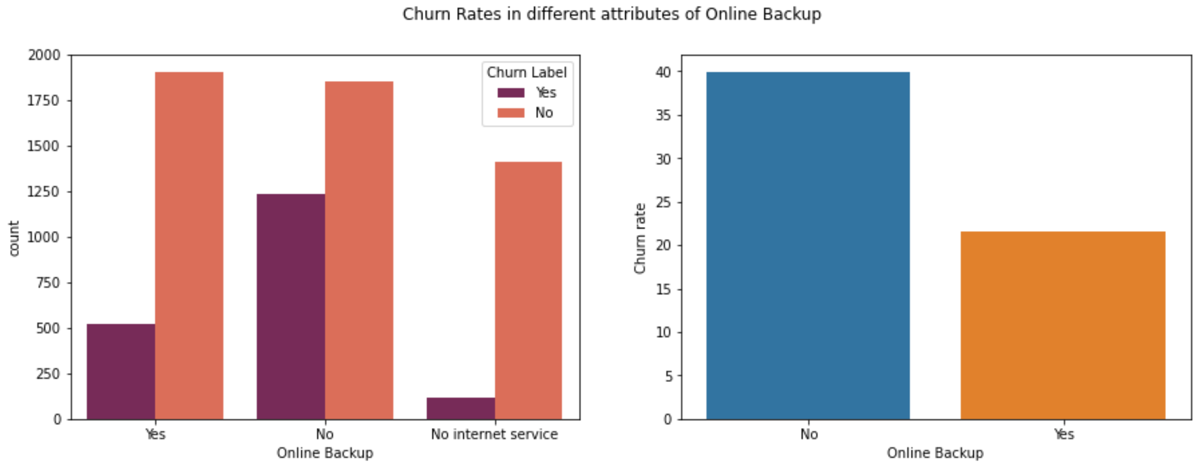
Other Support Services

These services measure if the customers have online security, online backup, tech support and device protection services in their contract.

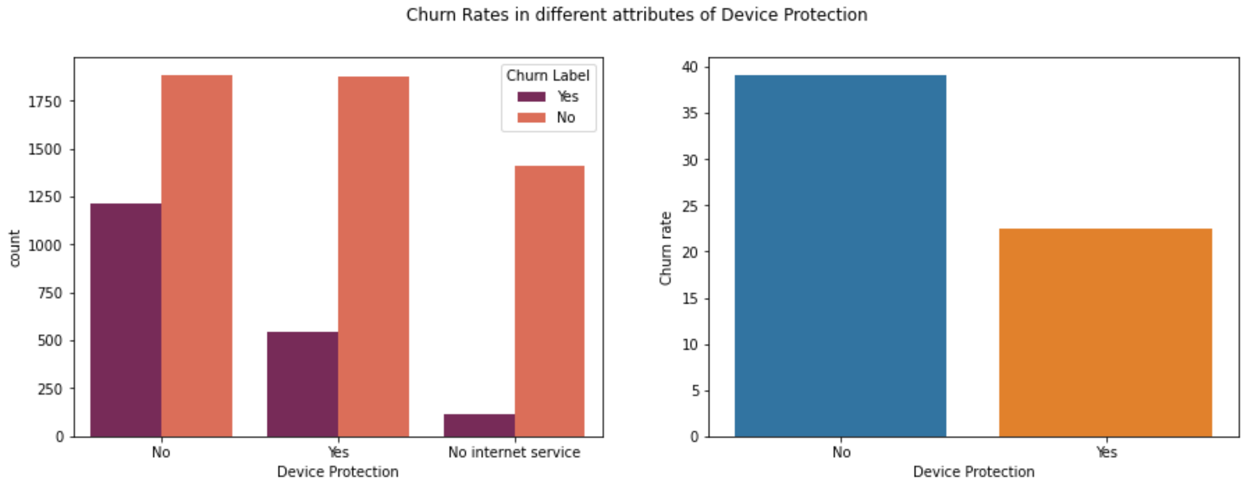
Online Security:



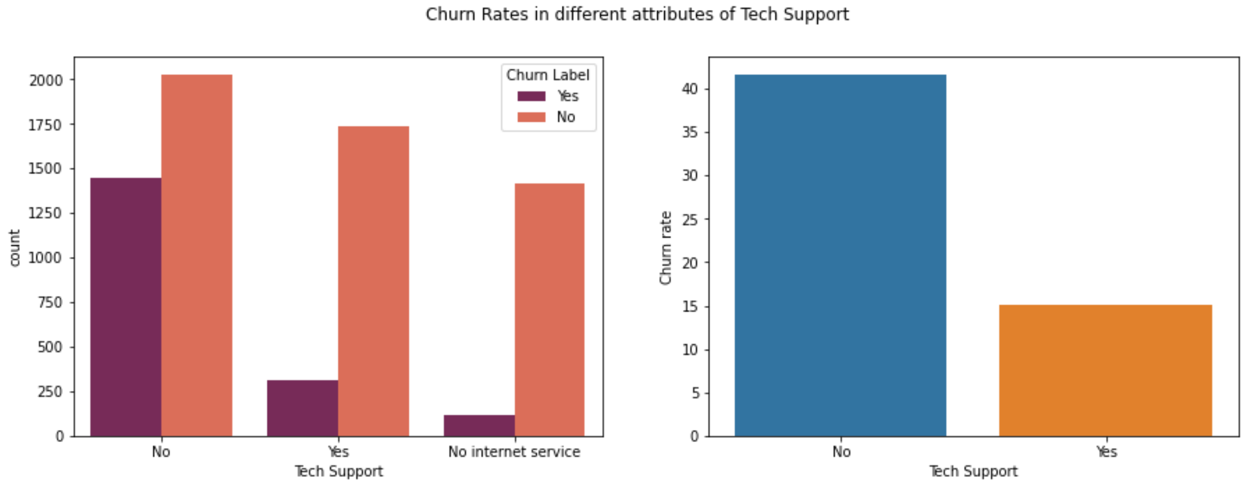
Online Backup:



Device Protection



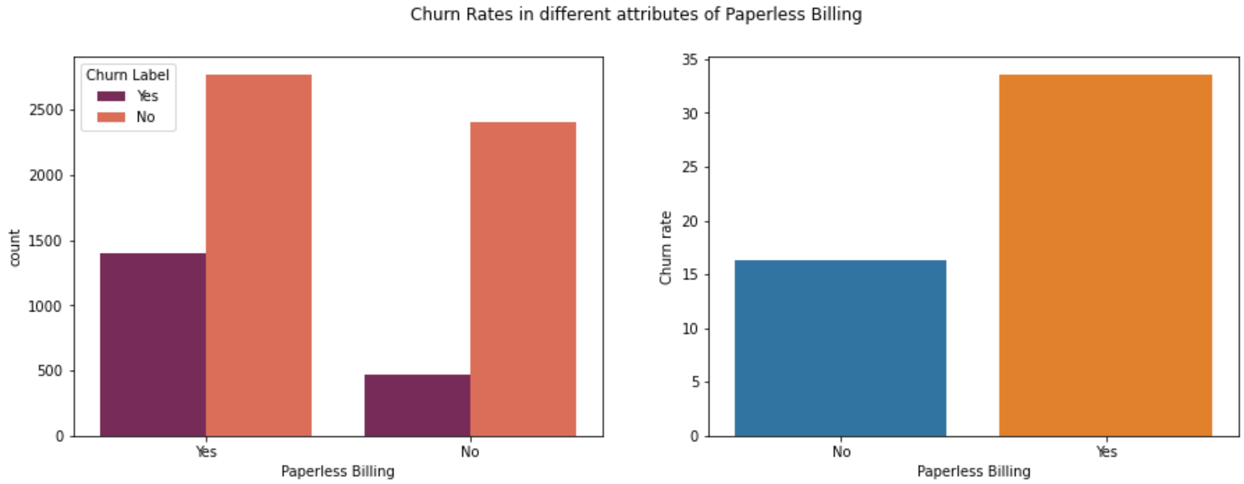
Tech Support:



From the above plots, it’s evident that those customers without support services churn more than that of the customers who have those support services.

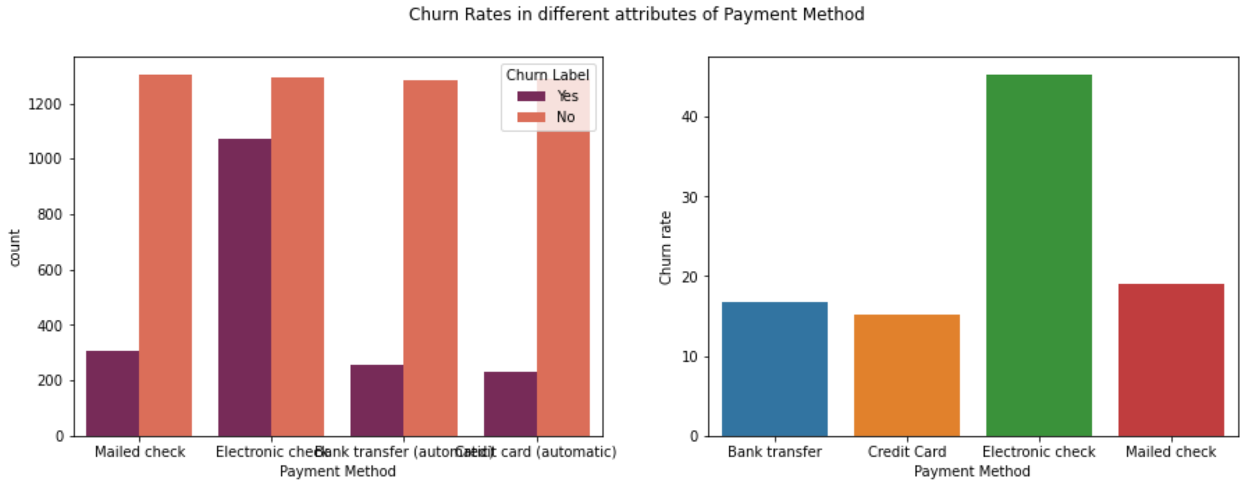
Paperless Billing:

Below plots show that customers with paperless billing churn more than that of those without paperless billing.



Payment Method:

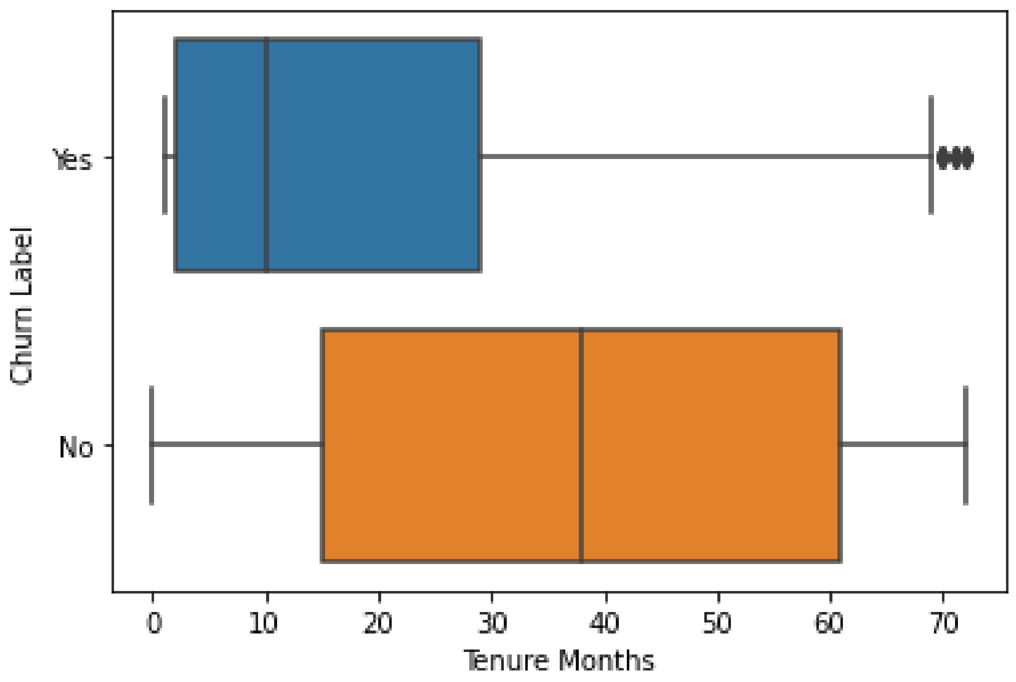
Payment method measures if the customer pays via bank transfer, credit card, electronic check or through mailed check. Plots show that customers who pay through electronic check churn at least 2-3 times more than that of customers with other modes of payments.



Numeric Features

Tenure Months:

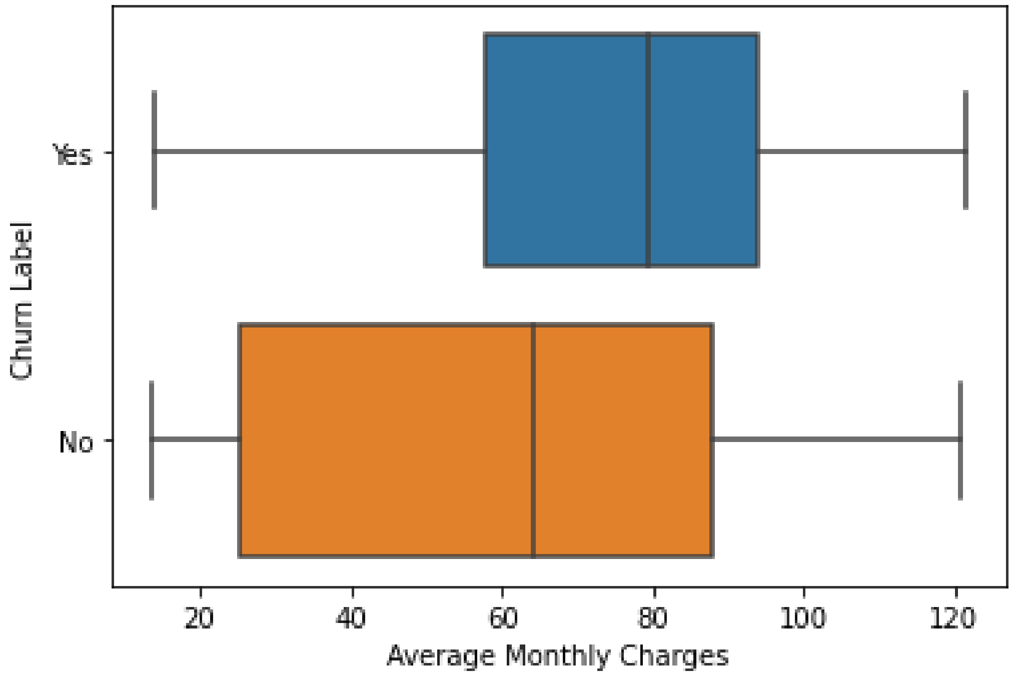
This variable measures how long the customer has been in contract with the company. Following boxplot depicts how tenure months are distributed among churners and non-churners. The churners have a mean of around 18 months (with a median of 10 months), whereas for non-churners, mean is 38 months (with a median equal to mean).



The tenure months distribution shows a normal curve for non-churners, but it is right skewed for churners, meaning most of those who have churned have lesser number of tenure months.

Average Monthly Charges

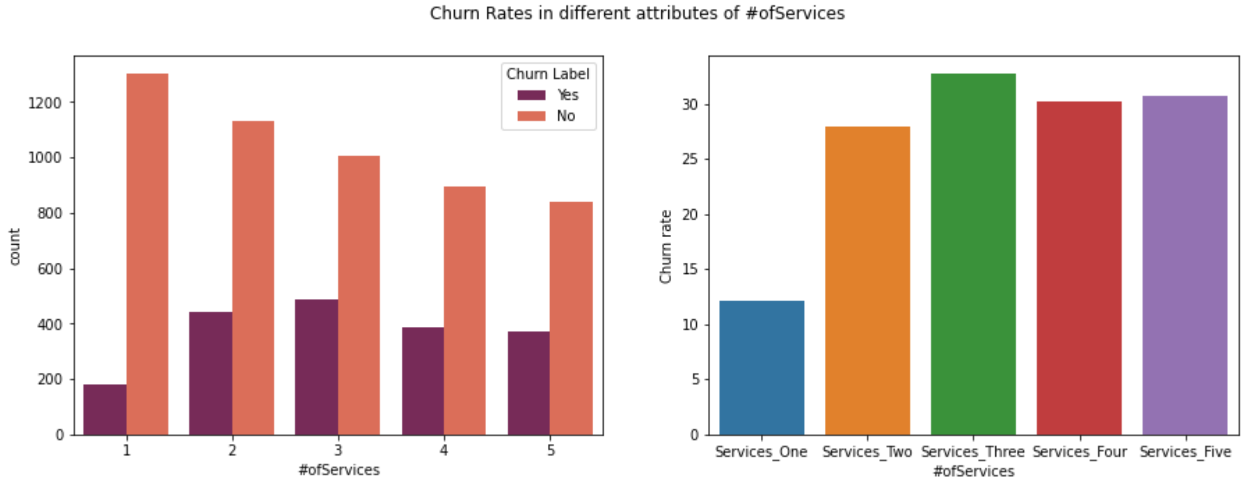
Average monthly charges were calculated by dividing total charges divided by total number of tenure months.



Above boxplot shows that the average monthly charges for churners is higher and the distribution is somewhat narrower than that of non-churners. Also, the median charges for churners is higher ($80) than that for non-churners ($64).

Total Number of Services

This feature was calculated by taking into whether the customer is having phone (single line/multiple line), internet, streaming TV, and streaming movies. It ranged from 1 to 5.



Plots above show that churn rates are lower when the customer has only one service and it increases when it becomes at least 2. On average, churn rates reach the highest when the number of services are 3.

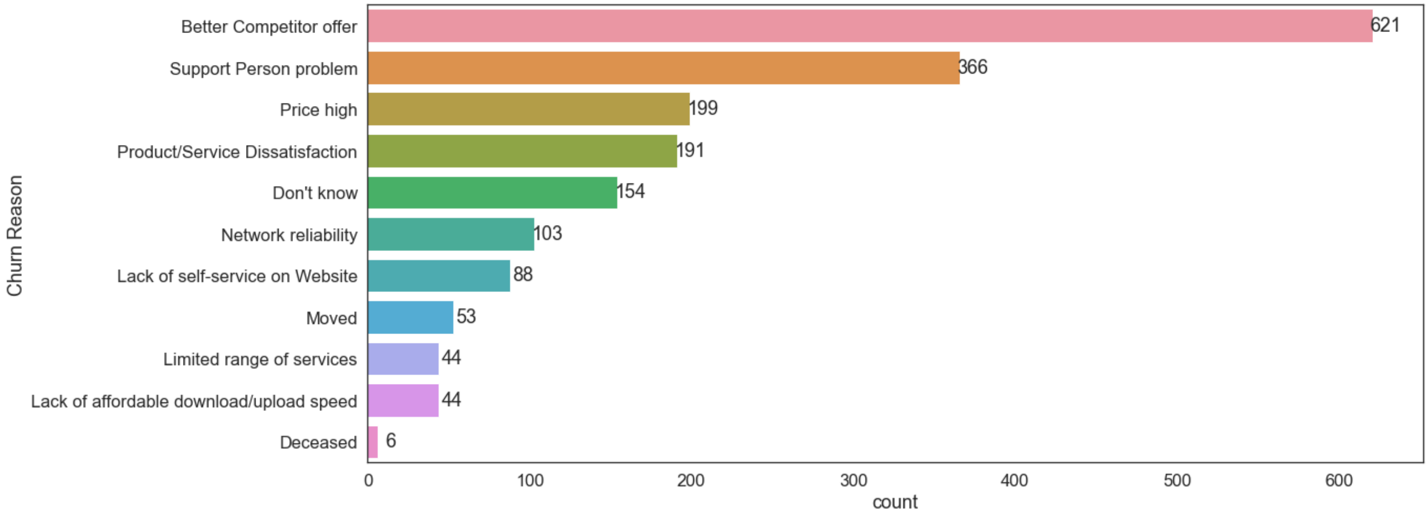
Churn rates in Different Cities/Neighborhoods

Dataset consisted of column showing the cities and neighborhoods of each customer. Based on that churn rates were calculated for each city/neighborhood. This was done by dividing the total number of churners in each city by total customers in the city. Using Tableau, the geographical distribution of churn rates in different cities/neighborhoods was made. This can be viewed in this link;

https://public.tableau.com/app/profile/b1453/viz/TelecomChurnratesinCitiesofCalifornia/Sheet1

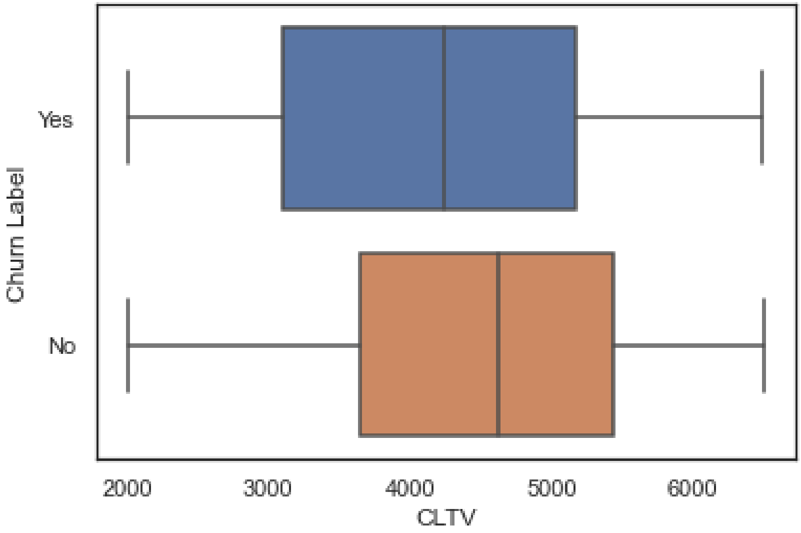
Exploring Reasons for Churning

Dataset also consisted of a column showing reasons for churning that were given by customers. Although this data cannot be used in modeling churning as these reasons have been given after they have churned, businesses can use it to learn what makes their customers leave and take steps to avoid that in the future.



The above countplot shows the reasons for churning with annotations of the number of churners with that reason. For instance, the top 3 reasons for churning were ‘Better competitor offer’, ‘Problems with the support person’ (lack of expertise and attitude of support personnel) and ‘High price’ (extra charges for long distance, extra data charges etc.).

Customer Lifetime Value (CLTV)



CLTV is the total value that a customer will bring to the company after deducting all the costs of servicing the customer. Although CLTV is a good measure to look at how much a company would lose when a customer leaves a company, this measure cannot be included in predicting churn probability as it’s a measure not so relevant from a customer’s perspective. The boxplot shows that on average CLTV is somewhat low for churners (low median and mean) than non-churners.

Modeling

Before modeling, following preprocessing steps were taken; creating dummy variables for categorical variables (last category in each feature was dropped), converting rest of the variables to numeric type, splitting train and test data (test-size was 0.2) and standardization of numeric features. Different models namely dummy regressor, ensemble methods of logistic regression and neural networks were attempted. Accuracy Score and AUC (Area Under the ROC Curve) were used as performance metrics. Accuracy score is the proportion of correct predictions divided by total number of predictions. ROC is plotted by showing true positive rates (y axis) against false positive rates (x axis). Hence, the further the curve is away from the x axis, the better will be the prediction. In essence, Area under the ROC evaluates how well a logistic regression model classifies two different outcomes (e.g., 0s and 1s) at all viable cutoffs. Usually, it ranges from 0.5 to 1, and the larger it is, the better the model in classification.

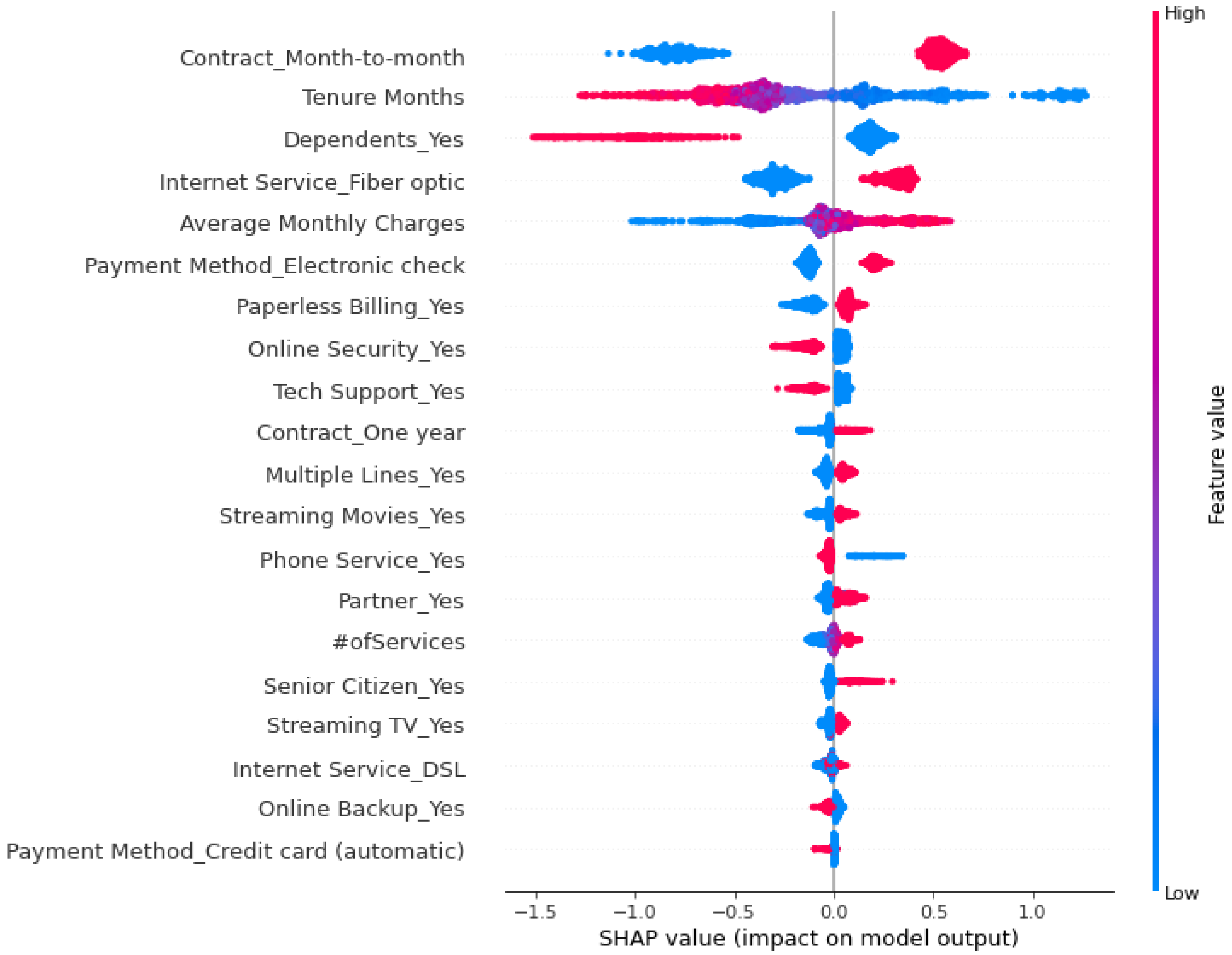
Initially the dummy regressor model was implemented. The most frequent value was used as the predicted label, which is 0s for this train data. With this model, the accuracy score was 0.743 for test data. Then I tried logistic regression, which improved the test data accuracy to 0.807 (AUC = 0.866). Later, a logistic regression with GridSearch CV was attempted, this didn't improve performance metrics in any way. Hence, I decided to use ensemble method called XGBoost. XGBoost/Extreme gradient boosting uses the gradient boosting decision tree algorithms; however, it is better in performance and speed compared to other gradient boosting techniques (Brownlee, 2021). With XGBoost algorithm, the accuracy is 0.808, and AUC is 0.862. This was followed by another XGBoost model with hyperparameters specified ('colsample_bytree' = 0.5,
'learning_rate = 0.05, 'max_depth' = 4, 'min_child_weight' = 2, 'n_estimators' = 100,
'subsample' = 0.8). This yielded Accuracy of 0.816 and AUC of 0.870. Subsequent to XGBoost models, I also tried a neural network model in Keras with an intention of improving the performance metrics. I used 2 hidden layers with 'relu' activation and 'sigmoid' activation for output model. With neural network, even though it improved the train data accuracy, it didn't

perform in par with XGBoost with hyperparameter tuning. Neural networks overfitted the train data that it performed poorly on the test data. Following table summarizes the results of different models used in classification.

Algorithm	train (accuracy)	train (AUC)	test (accuracy)	test (AUC)
Dummy Regressor (most frequent values)	0.743			
Logistic Regression	0.811	0.856	0.807	0.866
Logistic Regression with GridSearchCV	0.811	0.856	0.808	0.867
Neural Network (3 hidden layers)	0.820		0.805	0.857
XGBoost	0.823	0.885	0.808	0.862
XGBoost with hyperparameter tuning	0.826	0.883	0.816	0.870

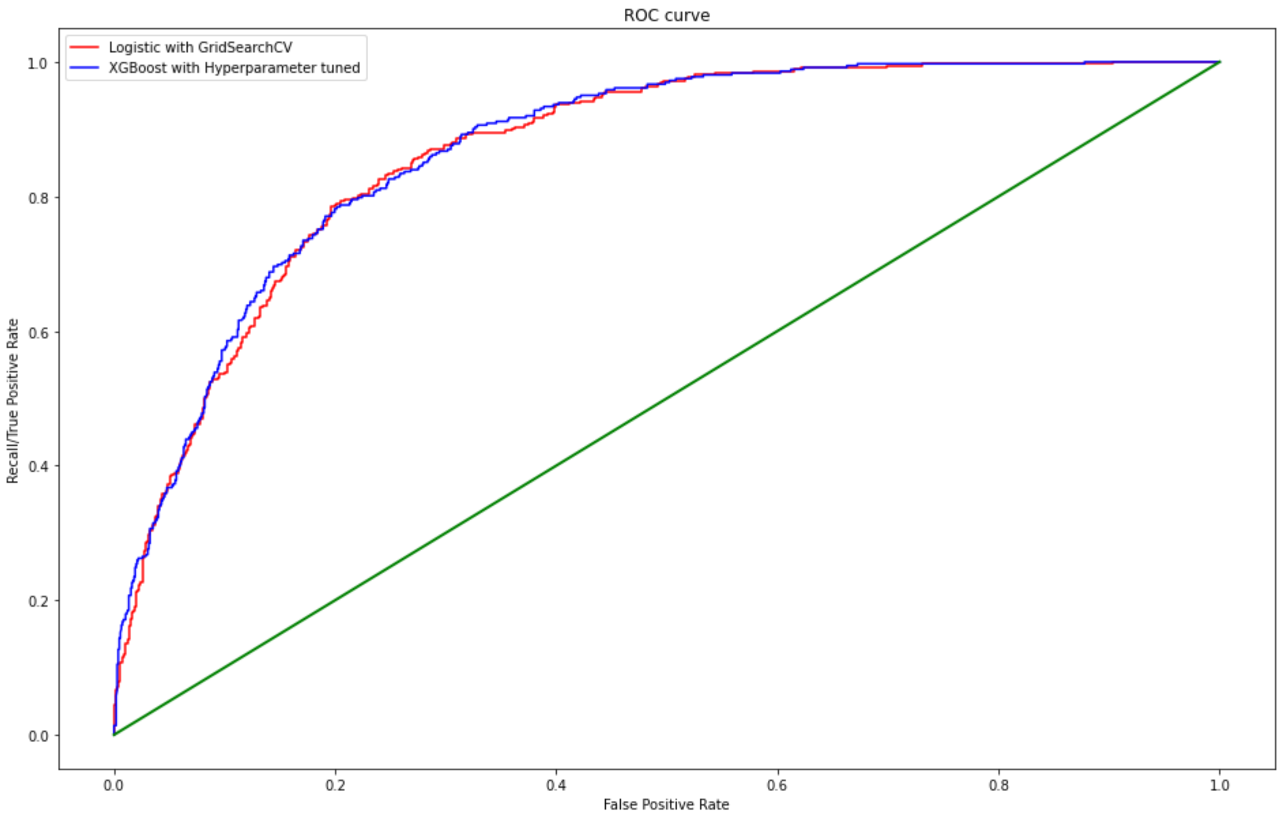
Finally, XGBoost with tuned hyperparameters was chosen as the final model because of higher test-data accuracy and AUC.

In order to better understand how feature importances contribute to its prediction, Shapley Additive Explanations (SHAP) were used. Summary_plot method of SHAP offers details about feature importance and the impact of Shapley values on the prediction. Red means high feature values and blue means low feature values. Depending on which side the red and blue values take they will either increase or decrease the predicted values.



Features like contract_Month-to-Month, Tenure Months, Dependents_Yes and having fiber optic internet are some of the most important key features that determine classification of samples here.

ROC below shows how the curve differs between logistic regression and XGBoost. Examining the graph reveals that there’s only a slight difference in AUC among them.



In classification models, other metrics such as precision, and recall are also used in evaluating a model. Precision measures the percentage of correctly predicted positives out of the total positive predictions. Recall is the percentage of correctly predicted positives out of the total actual positives. Accordingly, the finalized model produced the following precision and recall.

Outcome	Precision	Recall
0 (Not Churn)	0.86	0.90
1 (Churn)	0.67	0.57

	Predicted	
Actual	0	1
0	945	102
1	157	205

Out of total actual churners of 362, the model identifies 205 of them correctly, and hence the recall is 0.57. However, out of 1047 actual non-churners, model identifies 945 non-churners correctly and therefore the recall is 0.90. The model performs better in terms of finding correct non-churners compared to churners. In a Telecom industry, compared to identifying a non-churner as a churner, not being able to find a churner will affect the business more adversely.

That is, recall matters a lot. Therefore, to improve detecting higher number of true positives against total actual positives, I decided to reduce the decision boundary (usually it is 0.5) to the range between 0.3 - 0.4. After trying different values of decision boundaries to optimize a balance between the recall of non-churners and churners, it was decided that when decision boundary is 0.33 (as opposed to 0.5), the model identifies around 284 churners out of 362 total actual churners with a recall of 0.78 (earlier it was 0.57). While there's an improvement in the recall, the precision has gone down as we over predict the number of positives which leads to identifying a non-churner as a churner. Also, one should be careful of this, as it might lead to unnecessary promotional costs targeted at customers who may not churn anyway. Hence, it's important that we optimize the level of decision boundary so that the recall doesn't cost much of precision. Following tables show recall, precision (on the left) and confusion matrix (on the right) for decision boundary at 0.33.

Outcome	Precision	Recall
0 (Not Churn)	0.91	0.80
1 (Churn)	0.57	0.78

	Predicted	
Actual	0	1
0	833	214
1	78	284

Limitations and Future Directions

Having analyzed the reasons for churning, we learned that customers leave the company mainly due to competitor offers. Hence, incorporating, other competitive factors such as the number of companies in the area that offer the same services, their pricing strategies etc. might improve churn prediction. In addition, customers also mentioned that lack of service support and service personnel’s attitude as their reasons for churning. These differences could be included in prediction through features like customer’s satisfaction score after a service failure recovery or support service experience of a customer.

References

Ahn, J., Hwang, J., Kim, D., Choi, H. & Kang, S. (2020), S."A Survey on Churn Analysis in Various Business Domains," 2020, *IEEE Access*, 8, 220816-220839.

Brownlee, J. (2021). “A Gentle Introduction to XGBoost for Applied Machine Learning”. Access: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Devriendt, F., Berrevoets, J., & Verbeke, W. “Why you should stop predicting customer churn and start using uplift models”. 2021, *Information Sciences*, 548, 497-515.

IBM Cognos Analytics Data Collection (updated on July 14, 2021). Access: https://community.ibm.com/accelerators/catalog/content/Customer-churn.