Telecom Customer Churn Analysis and Prediction

Introduction

What is Customer Churn?

Customer churn is defined as when customers or subscribers discontinue doing business with a firm or service.

Customers in the telecom industry can choose from a variety of service providers and actively switch from one to the next. The telecommunications business has an annual churn rate of 15-25 percent in this highly competitive market.

Individualized customer retention is tough because most firms have a large number of customers and can't afford to devote much time to each of them. The costs would be too great, outweighing the additional revenue. However, if a corporation could forecast which customers are likely to leave ahead of time, it could focus customer retention efforts only on these "high risk" clients. The ultimate goal is to expand its coverage area and retrieve more customers' loyalty. The core to succeed in this market lies in the customer itself.

Customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers.

Context

Predict the whether a customer churn or not by analyzing the behavior of different customers. Here we also analyze what are the reasons that make a customer to churn.

Motivation

Different reasons trigger customers to terminate their contracts, for example better price offers, more interesting packages, bad service experiences or change of customers' personal situations. From an organizational perspective, it is always cheaper to retain existing customer than to spend money to acquire new customer. We want to use Machine Learning models to predict whether a customer will retain or not.

Data Source

The dataset was collected from Kaggle named as Telco Customer Churn dataset. It was an IBM issued dataset.

Link: https://www.kaggle.com/blastchar/telco-customer-churn

Content

Each row in the dataset represents a customer, while each column contains customer's attributes described on the column Metadata.

The raw data contains 7043 rows (customers) and 21 columns (features).

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.

Column names:

- CustomerID: Customer ID unique for each customer.
- gender: Whether the customer is a male or a female.
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0).

- Partner: Whether the customer has a partner or not (Yes, No).
- Dependent: Whether the customer has dependents or not (Yes, No).
- PhoneService: Whether the customer has a phone service or not (Yes, No).
- MultipeLines: Whether the customer has multiple lines or not (Yes, No, No phone service).
- InternetService: Customer's internet service provider (DSL, Fiber optic, No).
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service).
- OnlineBackup: Whether the customer has an online backup or not (Yes, No, No internet service).
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service).
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service).
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service).
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service).
- Contract: The contract term of the customer (Month-to-month, One year, Two years).
- PaperlessBilling: The contract term of the customer (Month-to-month, One year, Two years).
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).

Next, there are 3 numerical features:

- Tenure: Number of months the customer has stayed with the company.
- MonthlyCharges: The amount charged to the customer monthly.
- TotalCharges: The total amount charged to the customer.

Finally, there's a prediction feature:

- Churn: Whether the customer churned or not (Yes or No).

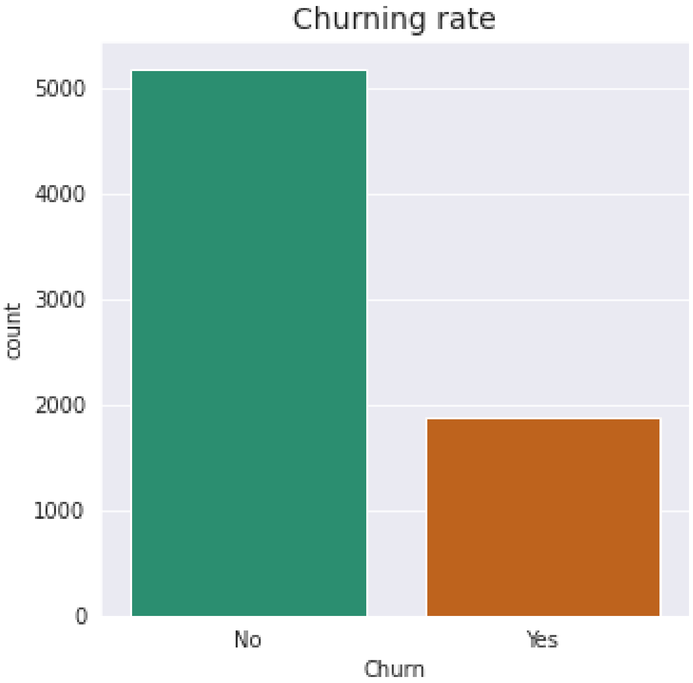
Methodology

The methodologies explained in this project are data collection which has already been discussed. Now, we move to data preprocessing and data cleaning. Then comes the analysis part along with the visualization and afterwards the prediction was done based on the model created.

- Data Collection
- Data Preprocessing and cleaning
- Data Analysis and Visualization
- Split dataset into train and test
- Model selection and evaluation
- Prediction

Data Cleaning, Data Analysis and Visualization

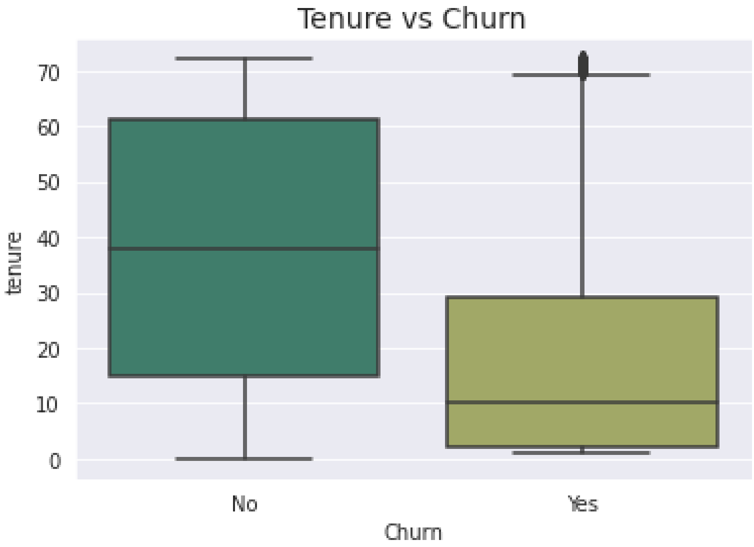
Before moving into cleaning, I have done some preprocessing of the data to observe the shape of the dataset and the summary of each column. The Total Charges and tenure column was found to int and object in their data type. It has been updated to float. The dataset was found to have 11 missing values in the Total Charges column and was filled using the mean of the Total Charges.



By visualizing the target variable Churn, the dataset is found to be **highly imbalanced**.

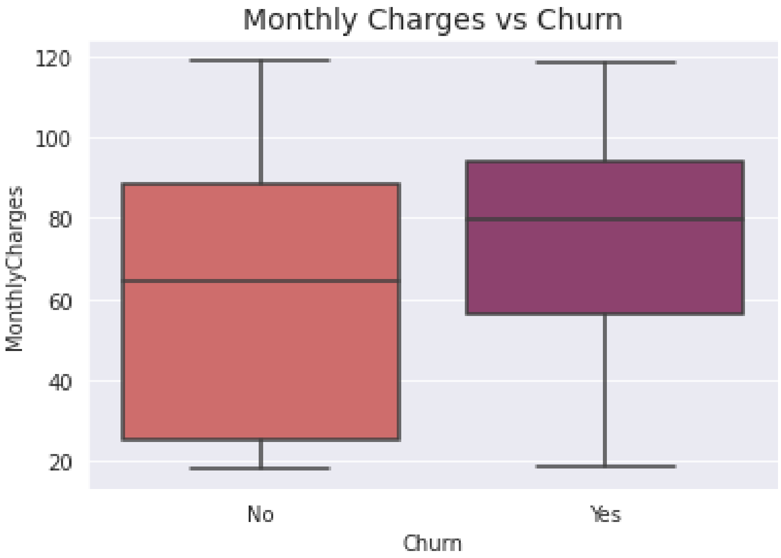
Tenure, Monthly Charges and Total Charges

By analyzing these three continuous variables in our dataset, outliers are detected among Tenure and Total Charges. A box plot was created to visualize all three of them.

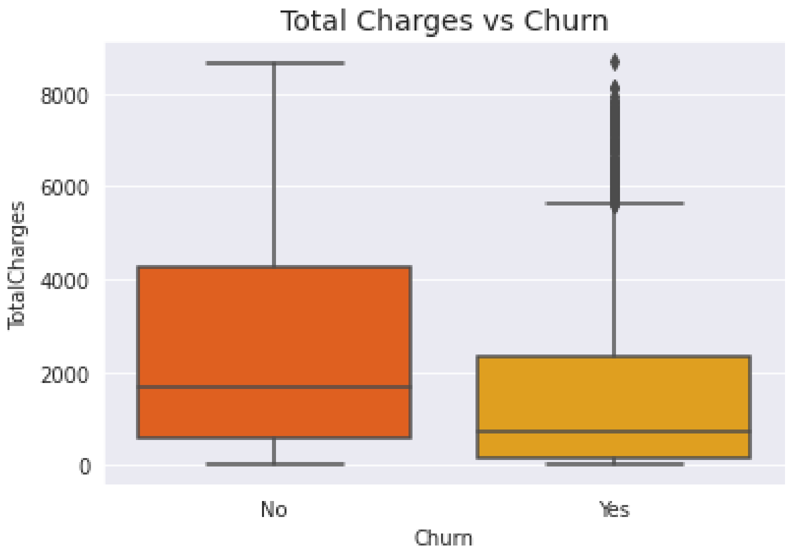


From this graph, we can see that the median tenure for the customers churned is around 10 months.

The median value of the Monthly charges of the churned customers is above 75, so



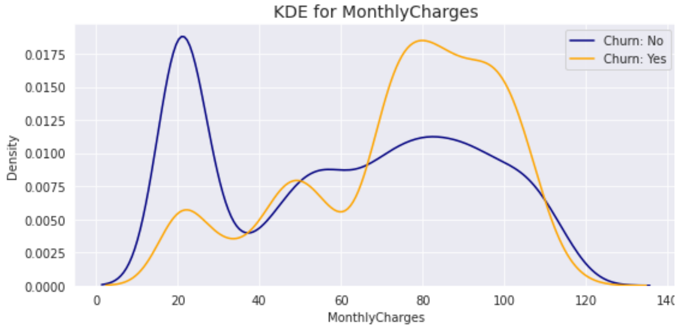
we can conclude that the **customers churned have high Monthly charges**.

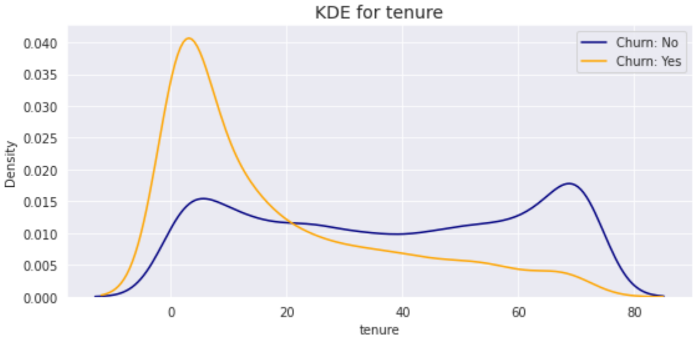


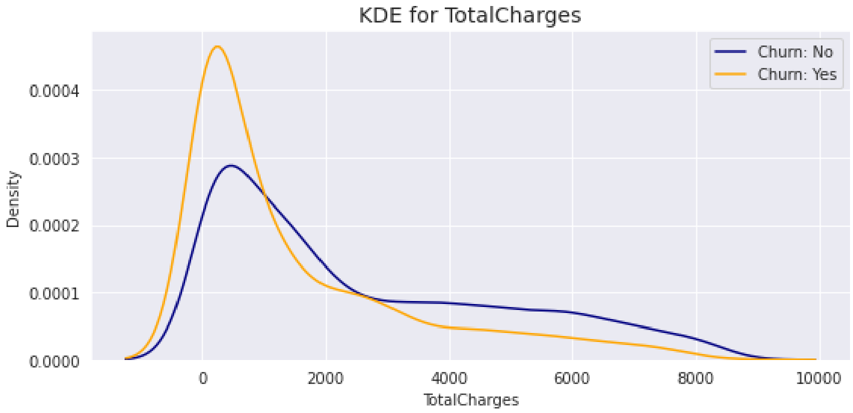
The median value of Total Charges for the customers churned is very low.

Probability distributions of Tenure, Monthly Charges and Total Charges

A Kernel Density Estimator (KDE) was plotted in order to visualize the distributions of the continuous variables.



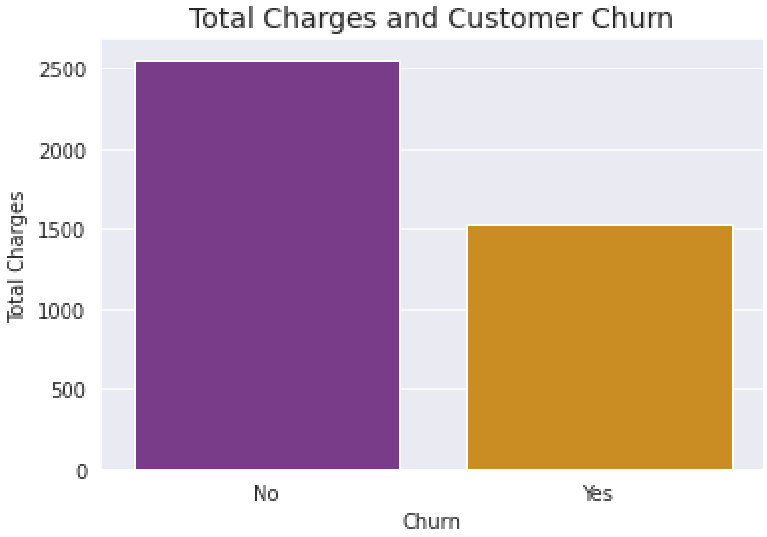




From these three plots, we can conclude that

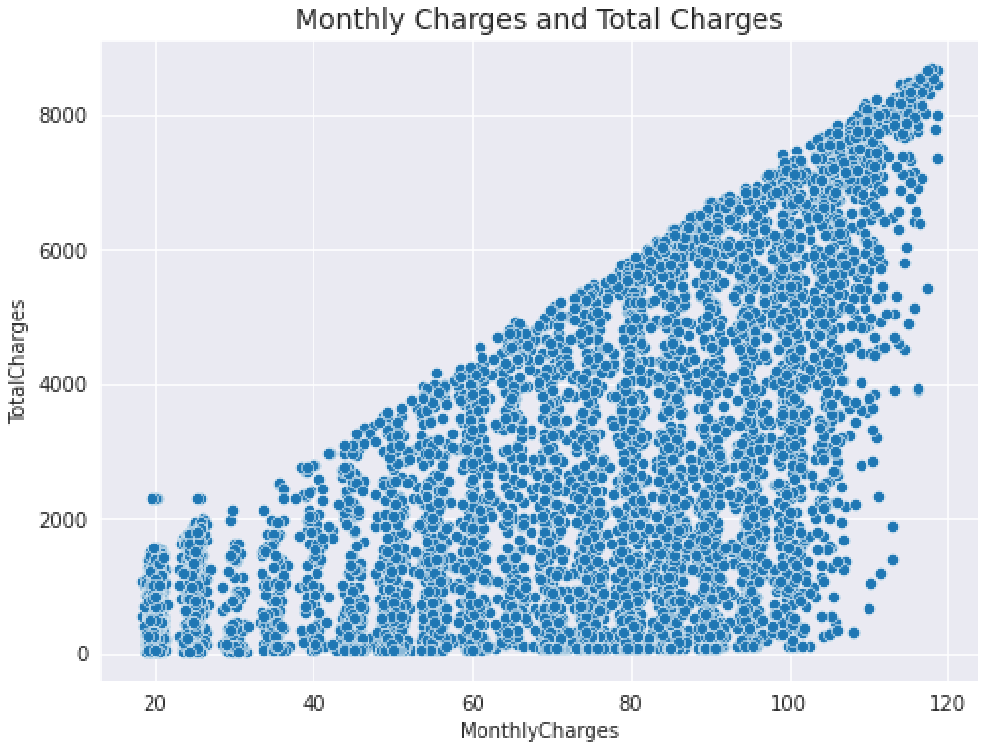
- Recent clients are more likely to churn
- Clients with high monthly charges churns faster
- Tenure and Monthly Charges can be probably identified as the important factor affecting churning.

Total Charges and Churning of customers



The customers churned were having Total Charges in an average of $1531.7 and also the margin of Total charges between the customers churned and remained are also very less.

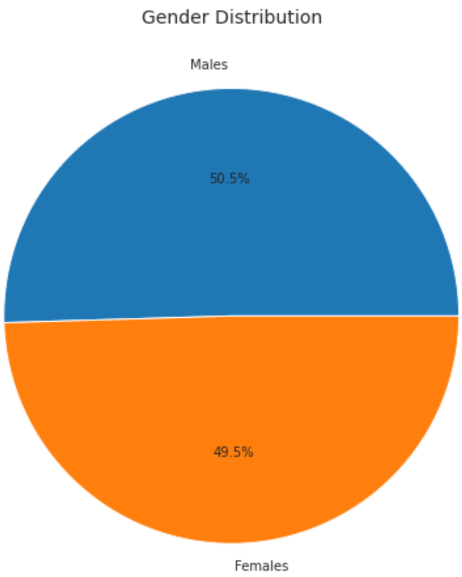
But while analyzing the **relationship between Monthly Charges and Total Charges,** it was seen that they are highly correlated.



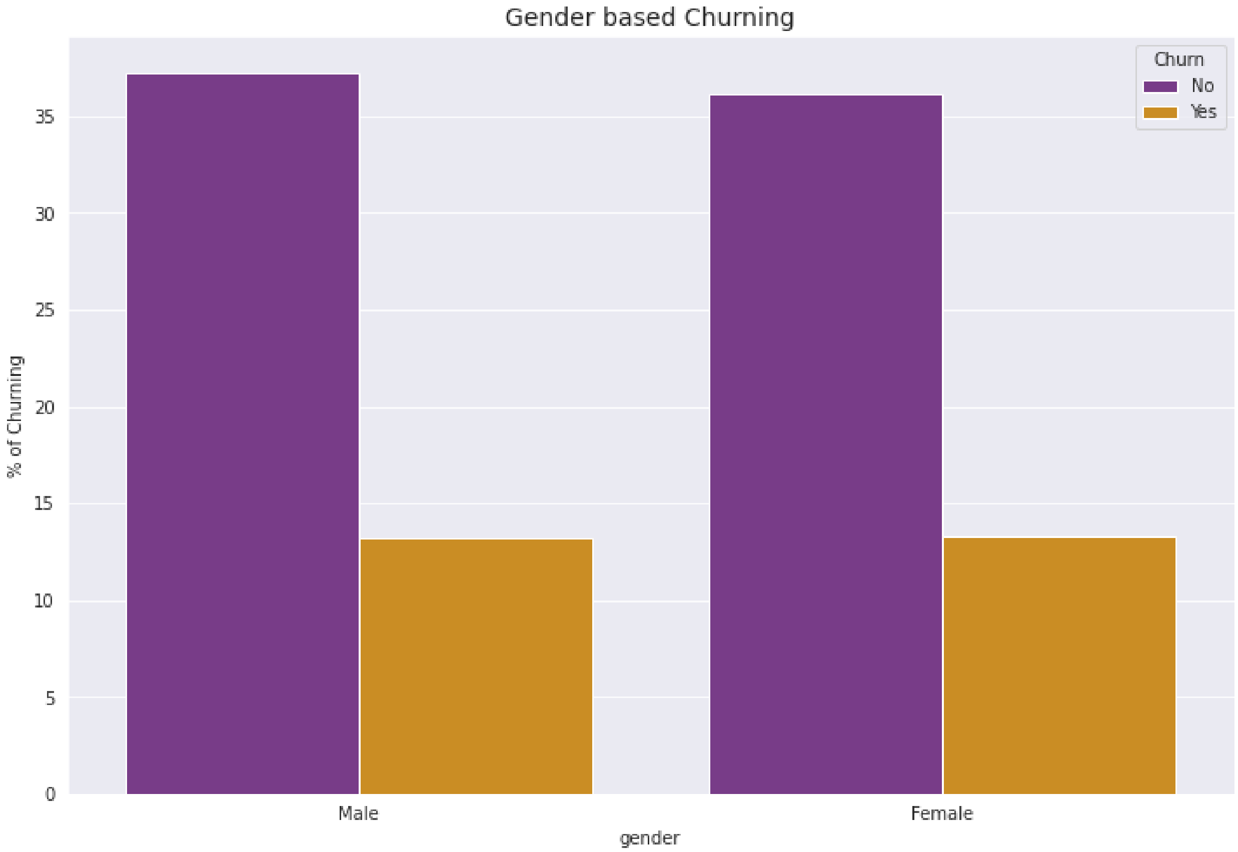
As the Monthly charge increases, the total charge also increases and the customers who churned also had high Monthly charges. **So, Total Charges may also be a reason of churning.**

Gender composition of customers

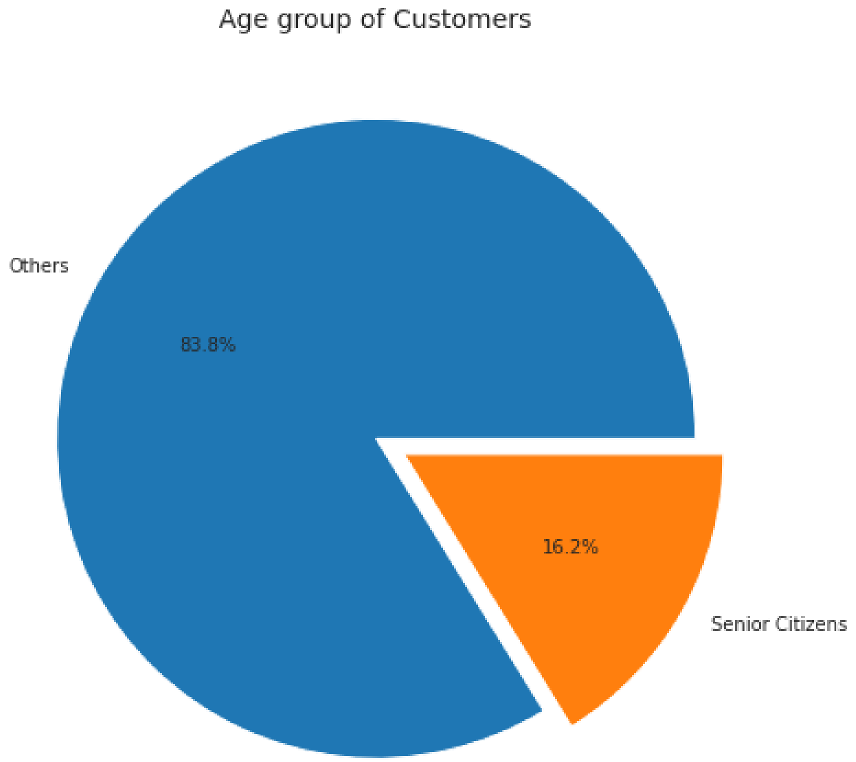
About **half of the customers in the dataset are females while the other half comprises of males**. A pie plot is used to visualize the percentage of male and female customers present in the dataset.



Amongst these customers, the rate of churning of male and female customers is almost the same. So, gender cannot be identified as a feature that affects churning.



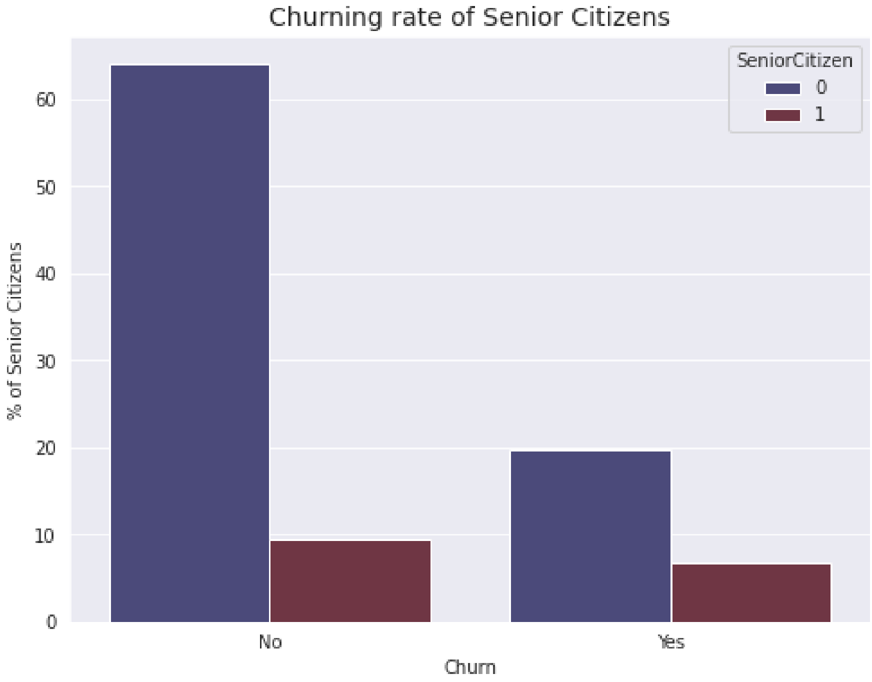
Age group of customers



Out of the whole dataset, only 16.2% are Senior citizens. 83.8% are younger people. Thus the **majority of the customers are younger people.**

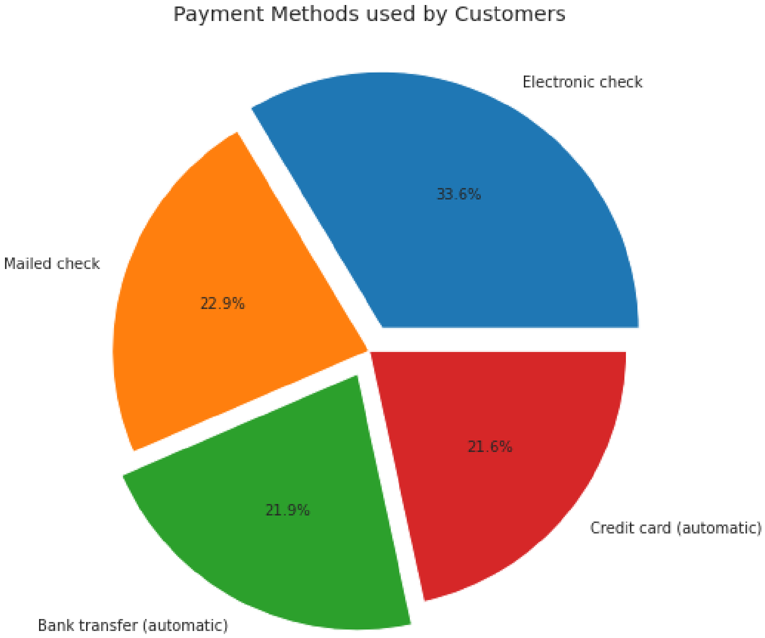
Churning rate of Senior citizens

It has been observed that the percentage of senior customers who left the company is very less compared to the other age groups. Only 6.7% of the senior citizens changed the company where as percentage of young customers left the company is 19.7% which is compared to be slightly higher. So the **young customers have more tendencies to change to other company.**

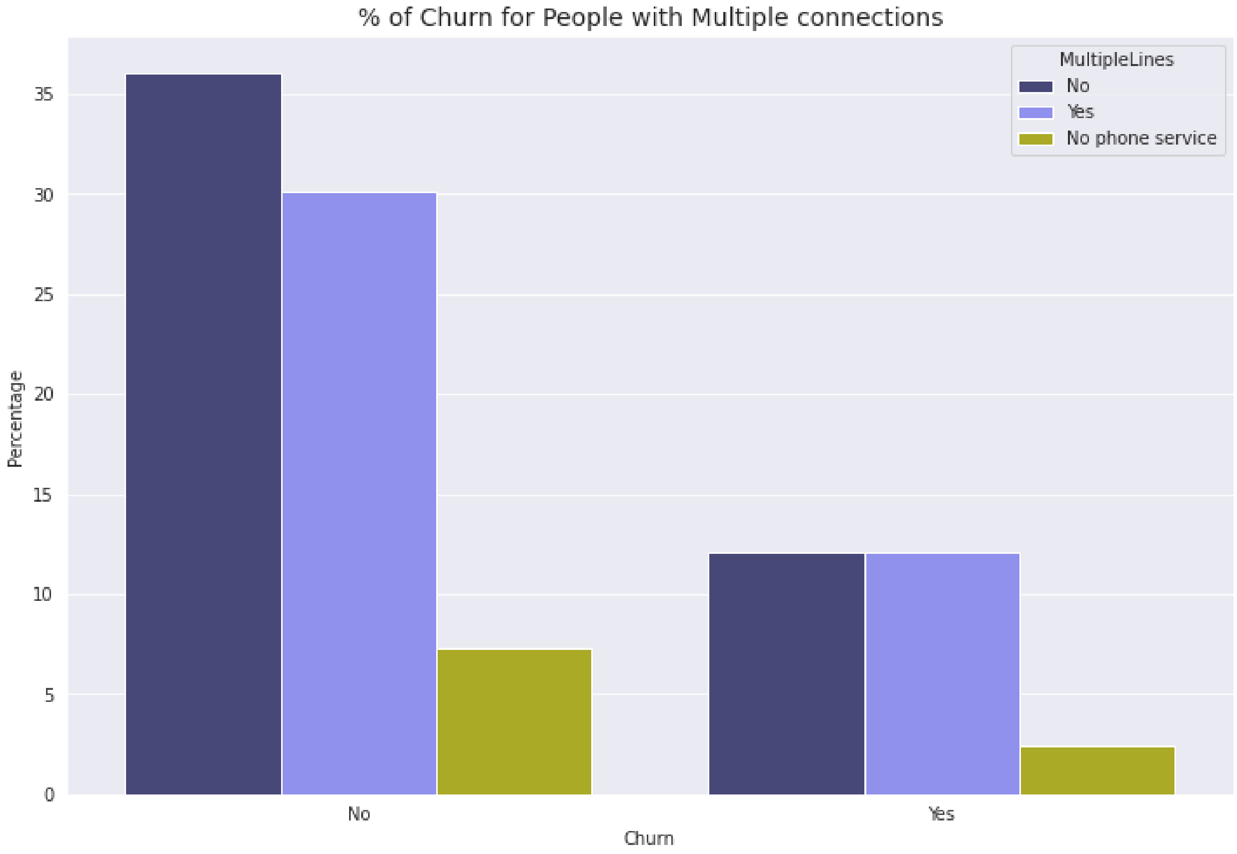


Payment methods

Electronic check is the most frequent payment method used by the customers. About 33.6% of the customers use an electronic check for their payment followed by Mailed check, Bank transfer and Credit card.



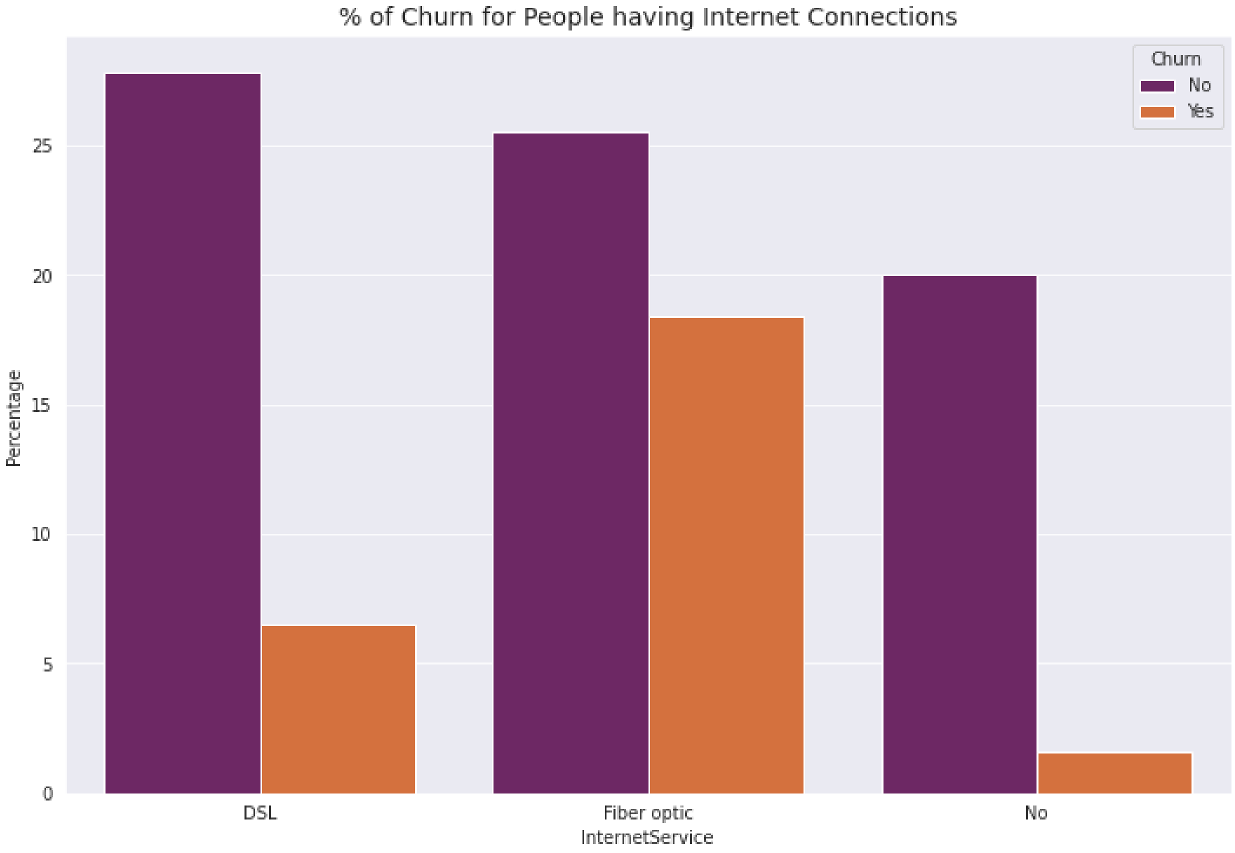
Churning rate of customers with Multiple lines



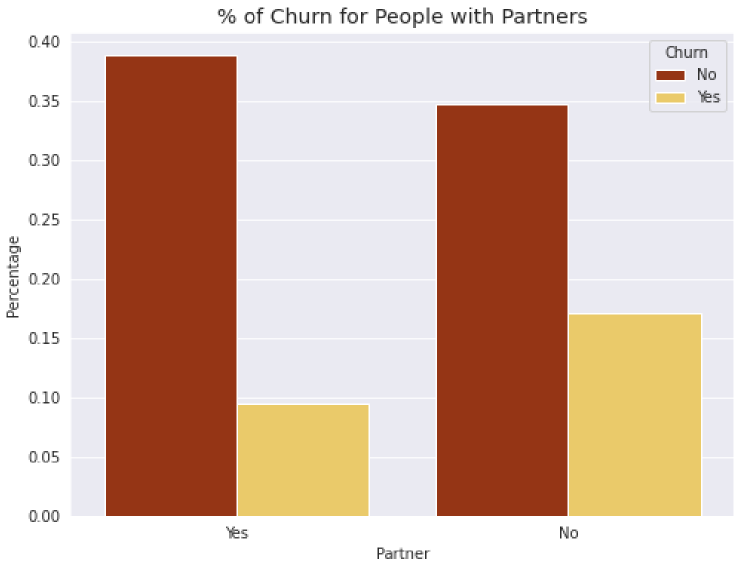
From this analysis, we found that some of the customers don’t have phone service. The customers with and without multiple lines have almost the same churn rate. So, this **cannot be identified** as a factor of churning.

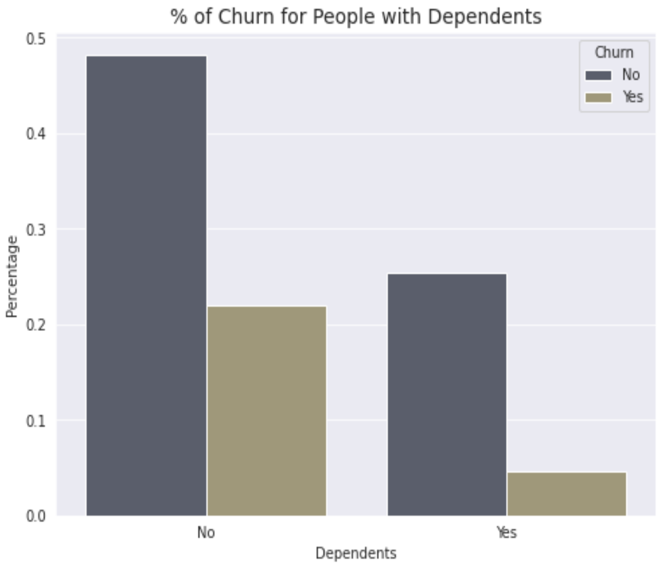
Internet service

This can be considered as an **underlying factor of churning**. The clients who has not subscribed to internet service has very low churning rate while the clients who have subscribed to different internet methods have a very high rate of churning. Also, the customers using **optic fiber are more probable to churn** than those using DSL connection.


% of Churn for People having Internet Connections

Partners and Dependents

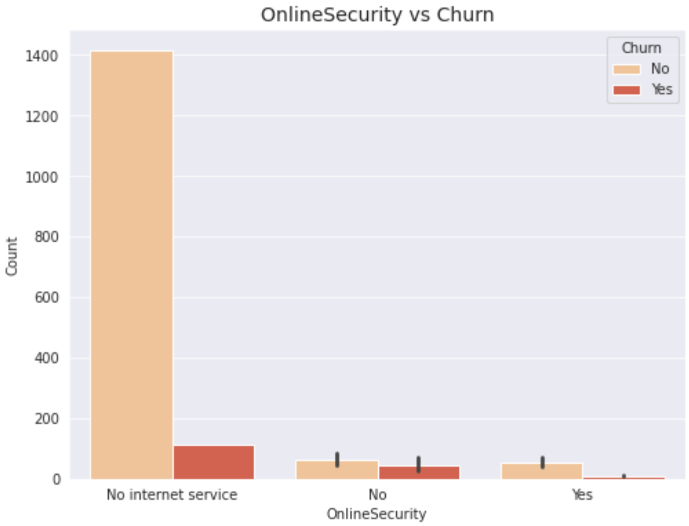

% of Churn for People with Partners

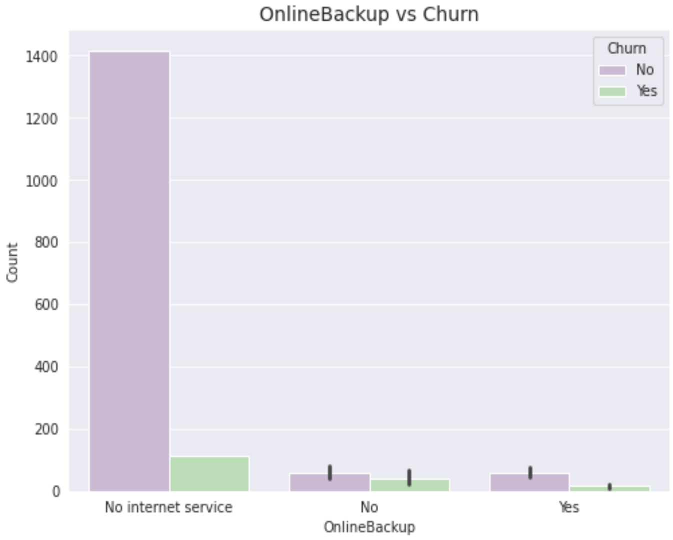

% of Churn for People with Dependents

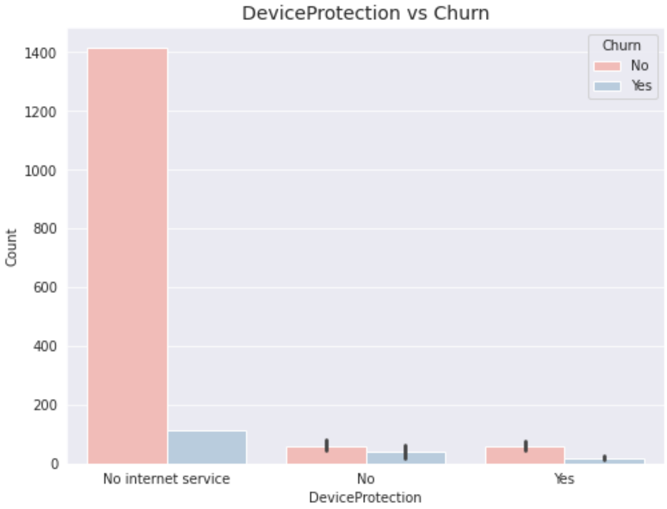
From the above two bar graphs, it is evident that the **clients with partners** are more likely to churn than that of the clients with dependents.

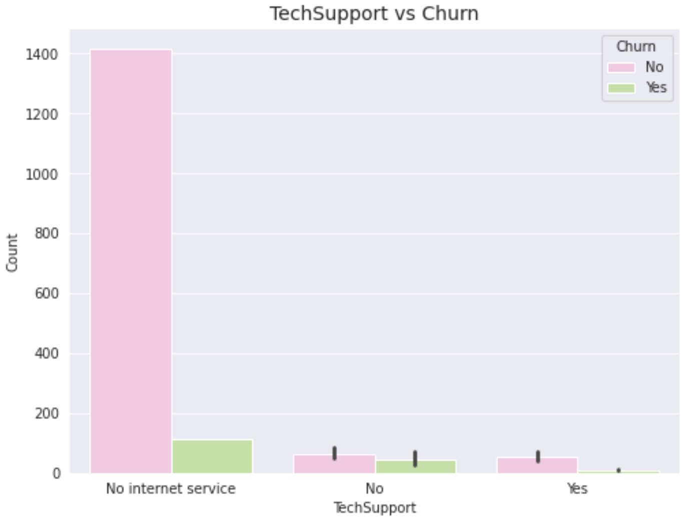
Additional Charges

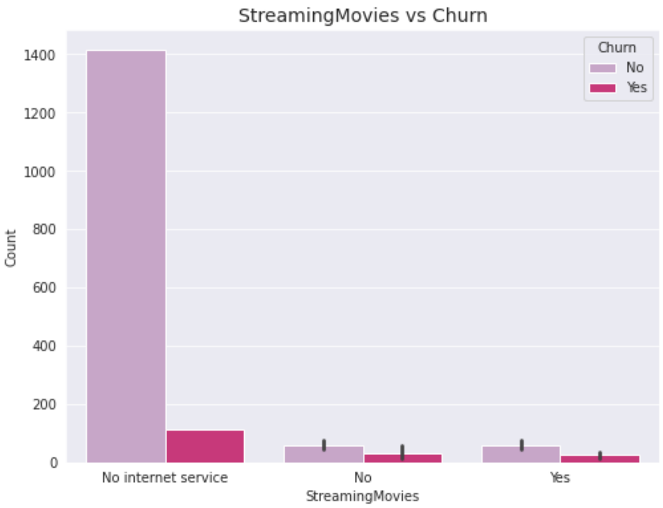
There are several additional charges subscribed by the customers mainly online security, online backup, device protection, tech support, and streaming TV and movies.

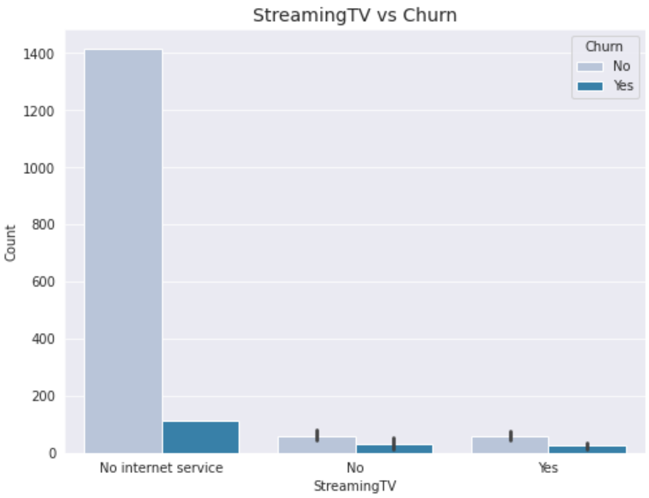












As visible from the above plots, additional charges don’t play a crucial role in the customer churn analysis.

Exploratory Data Analysis Concluding Remarks:

Let’s try to summarize some of the key findings from this EDA:

- The dataset is imbalanced with the majority of customers being active.
- Recent clients are more likely to churn and the contract period is around 10 months.
- Clients with high Monthly Charges churns faster.
- Monthly Charges and Total Charges are highly correlated.
- Half of the customers are females while the other half is males.
- Majority of the customers are younger people and have more tendencies to churn.
- The customers using optic fiber are more probable to churn than those using DSL connection.

- Clients with partners are more likely to churn.

Encode Categorical data

Any categorical variable that has more than two unique values have been dealt with one-hot encoding using get_dummies method in pandas here.

Split dataset into train and test

Now we need to separate the dataset into x and y values. y would be the ‘Churn’ column whilst x would be the remaining list of independent variables in the dataset. Let’s decouple the master dataset into training and test set with an 80%-20% ratio and separate ‘customerID’ from training and test data frames. It’s quite important to normalize the variables before conducting any machine learning (classification) algorithms so that all the training and test variables are scaled within a range of 0 to 1.

Dealing with class imbalance using Synthetic Minority Over-sampling Technique

This technique is followed to avoid over fitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class as an example and then new synthetic similar instances are created. These synthetic instances are then added to the original dataset. The new dataset is used as a sample to train the classification models.

Thus class imbalance has been dealt using SMOTE and the resampled dataset is used for training and testing.

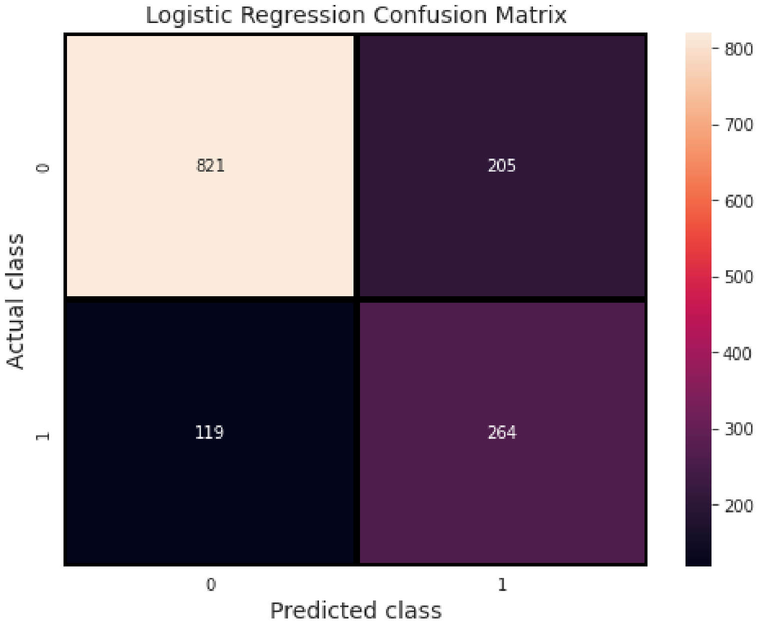
Model selection and evaluation

The models used in this scenario are

- Logistic Regression
- Decision Tree Classifier
- K- Nearest Neighbors

Logistic Regression

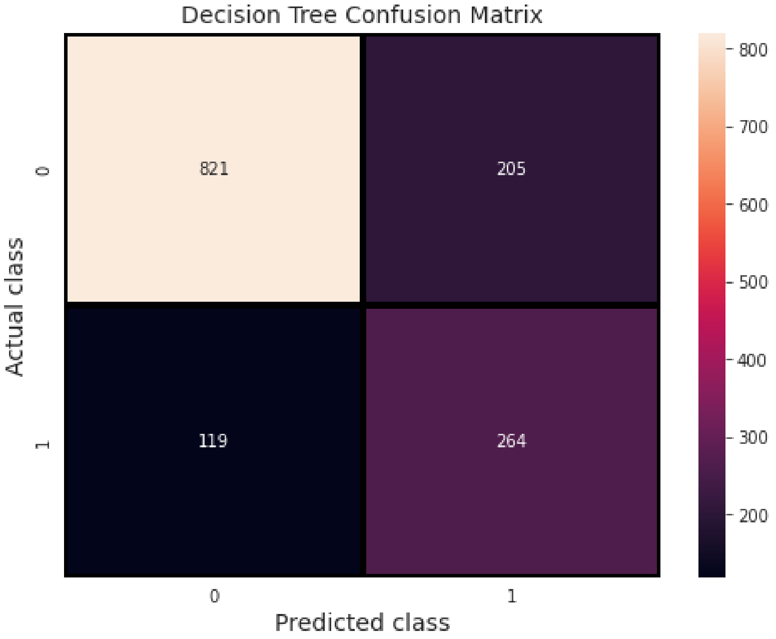
As we have a binary classification problem we applied logistic regression which is mainly used for binary classification problems. After applying our model we saw that the model was giving an accuracy of 77 % both for training and testing. But we cannot go with the accuracy score as we can get a high accuracy score even if our model predicts all the churners as non-churners. Instead of checking the accuracy score we will print the classification report to check the F1 score and plot the confusion matrix.



The model was able to predict 264 churners out of 383 churners and from the classification report the precision was found to be 0.56, recall was 0.69 and the F1 score was 0.62.

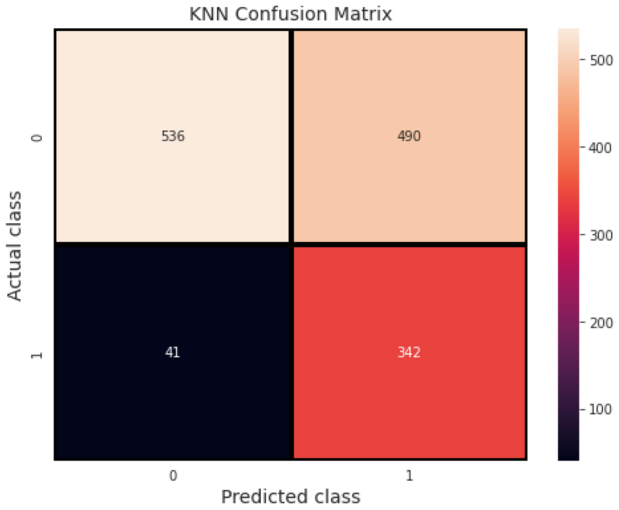
Decision Tree Classifier

Decision Tree algorithm was applied to the resampled dataset and we got an accuracy of 77% which is the same as that in Logistic Regression. Also the precision, recall and the F1 score was found to be the same. The model was able to predict 264 churners out of 383.



K- Nearest Neighbors

KNN was applied and the accuracy score got was too low. The accuracy was only 62% and the F1 score 0.56 but the recall was found to be 0.89. Also the Type 2 error was very high and the method was time consuming.



Conclusion

After performing the analysis and applying different Machine Learning Algorithms, we can say that out of the three models, decision tree and logistic regression were more accurate in predicting the values and we were able to get an accuracy of 77%.