Objective: Our aim is to develop a churn prediction model based on several characteristics which assist telecom operators to predict customers who are most likely subject to churn.

For a company acquiring a new customer is more expensive than retaining the existing ones and losing customers mean losing revenue. Predicting customer churn is critical for telecom companies to retain customers.

Dataset (https://www.kaggle.com/datasets/blastchar/telcochurn):

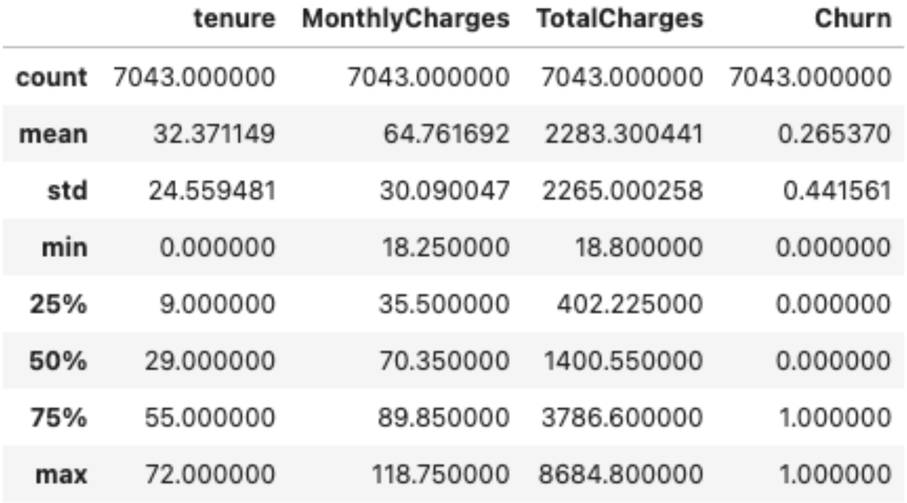
The Telco data set includes information about.

- 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, senior citizen and if they have partners and dependents.

Data Preprocessing:

- Telco dataset contains 11 blank/missing values for column 'Total Charges'. Since values are missing in small proportion, we will perform Mean imputation.
- We have converted our outcome variable 'Churn' into 0 and 1 values where 0 indicates 'No Churn' and 1 indicated 'Churn'.
- We will exclude first column 'Customer ID' from our analysis as that does not provide any value.

Summary:

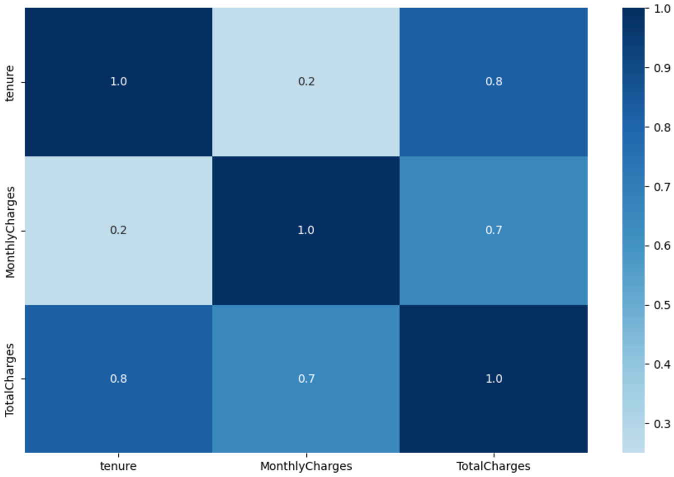


	tenure	MonthlyCharges	TotalCharges	Churn
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2283.300441	0.265370
std	24.559481	30.090047	2265.000258	0.441561
min	0.000000	18.250000	18.800000	0.000000
25%	9.000000	35.500000	402.225000	0.000000
50%	29.000000	70.350000	1400.550000	0.000000
75%	55.000000	89.850000	3786.600000	1.000000
max	72.000000	118.750000	8684.800000	1.000000

- Telco Dataset consists of 21 columns and 7043 observations
- Average tenure of a customer is ~32 months.

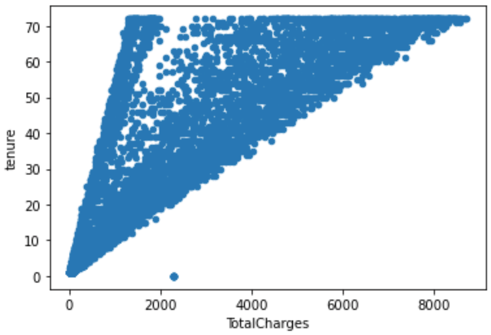
- Average monthly charge paid by a customer is around $65.
- Average total charge paid by a customer is around $2283.

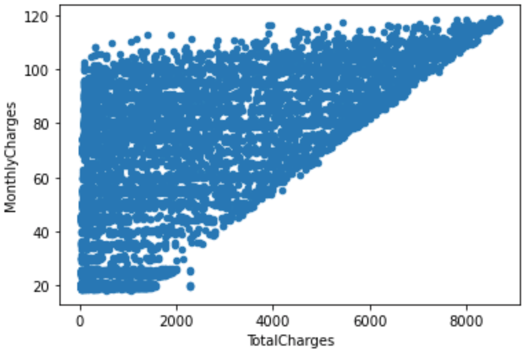
Heatmap to show correlation between numerical measures:



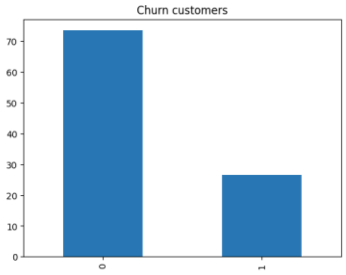
We can see from above heatmap that Total charges paid by a customer is positively highly correlated with Tenure and Monthly charge. This means that increase in Tenure or Monthly charge of a customer may also increase Total charge paid by a customer.

This can also be seen by plotting a scatter plot between the two.





Percentage of Churn:

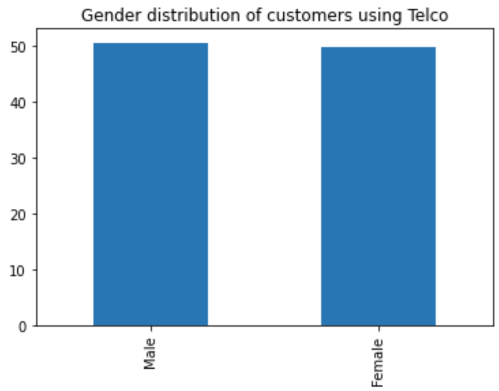


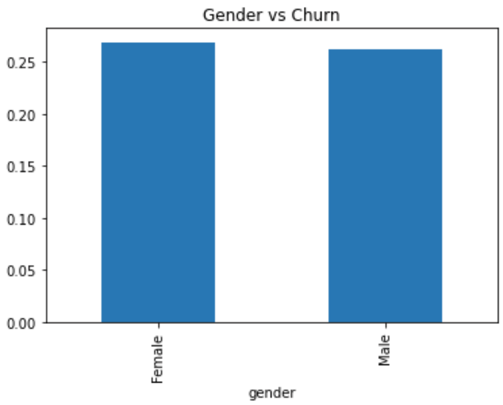
In our dataset, around 73 % of the customers are not churned whereas around 27% customer churned.

Exploring Demographic Data:

We will explore the demographic data first like gender, senior citizen, partner and dependent status of the customers.

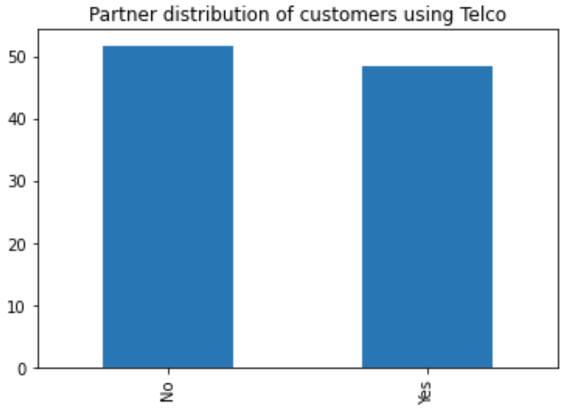
Gender:

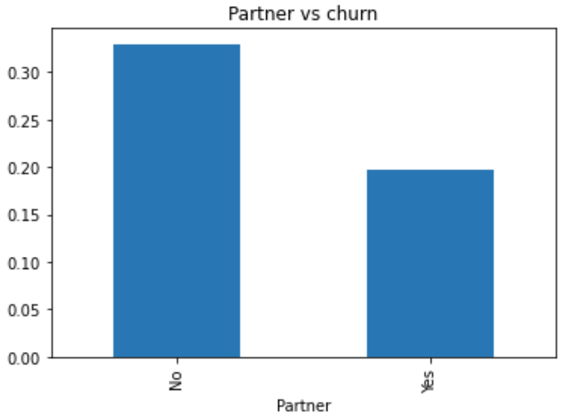




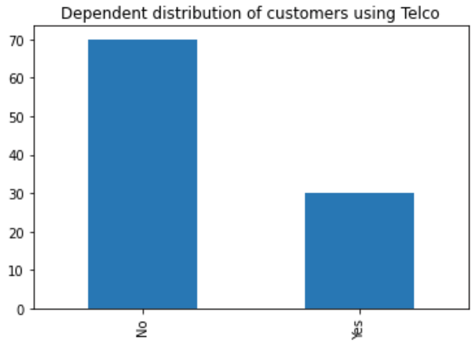
Above graphs shows that there is almost equal percentage of male and female who are Telco's customer and almost equal level of churn for both male and females.

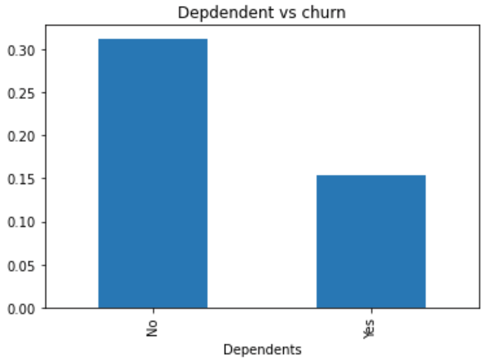
Partner and Dependent:





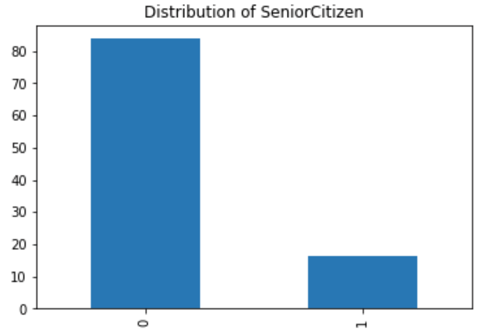
From above graph we can see that there is almost equal number of customers with/without a partner. Percentage of churn is more for customers who do not have a partner.

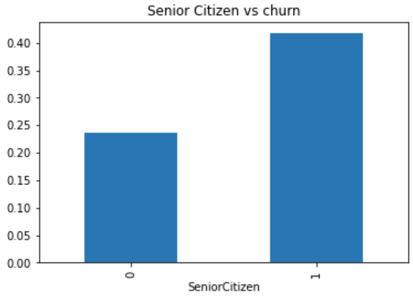




As per above graph, we can see that majority of customers do not have a dependent (almost 70%) where as only ~30% of the customers have a dependent. Percentage of Churn is more for customer who do not have a dependent as compared to customers who have a dependent.

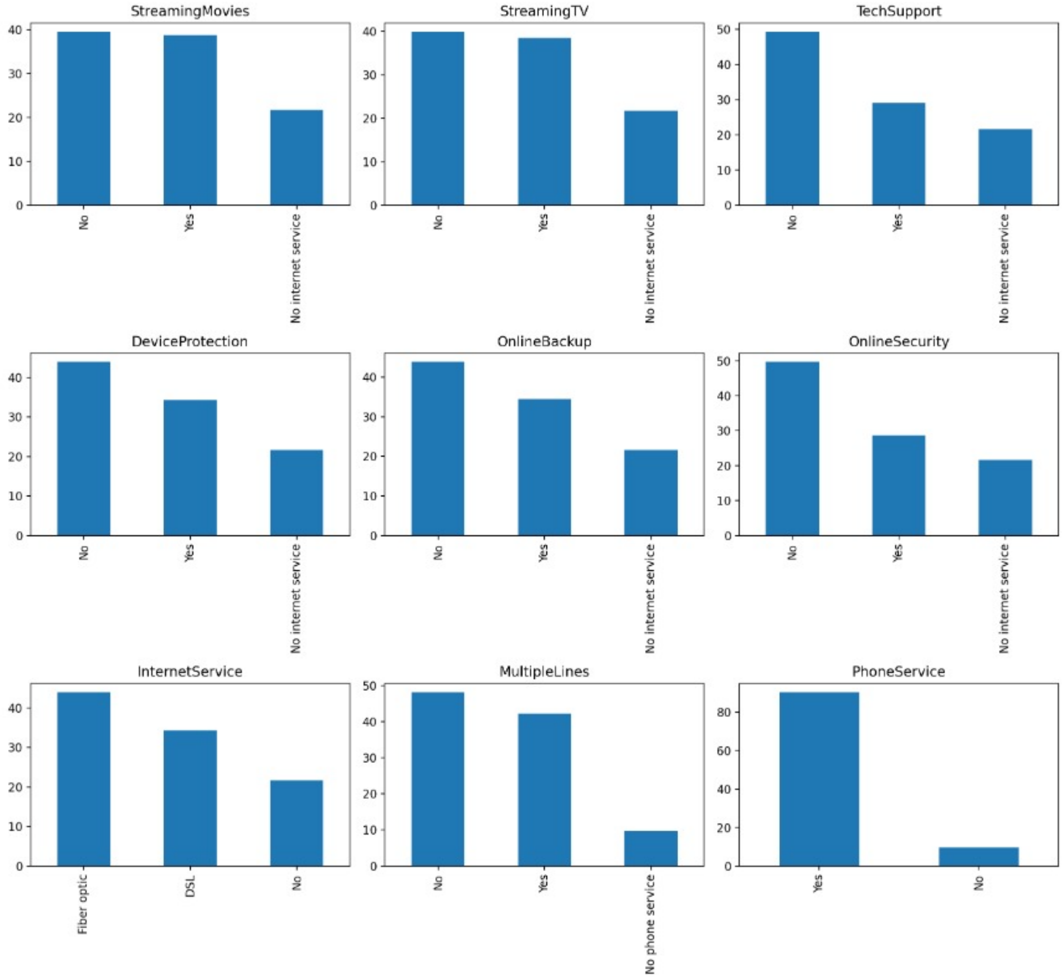
Senior Citizen:





Above graph shows that percentage of senior citizen is very low in dataset. Only ~16% of the customers are senior citizens. Percentage of Churn is more for senior citizen customers.

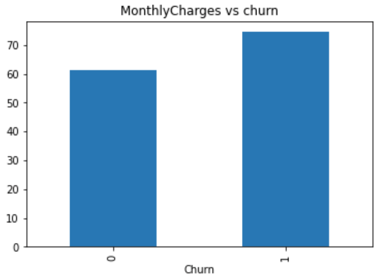
Exploring Service information in the Dataset:

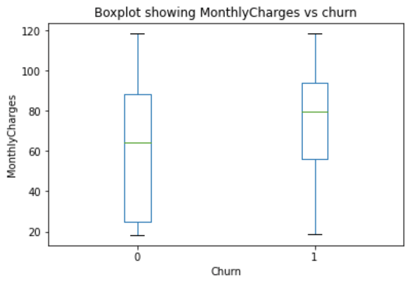


From above graphs we infer that:

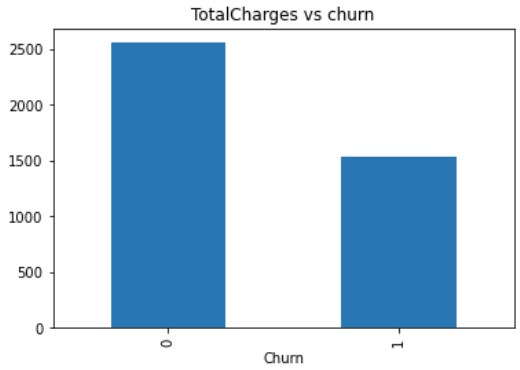
- Most of the Telco customers use Phone service. Out of all the customers who have opted for a phone service, there is almost an equal split between the customers who have opted for a mutiple line vs people who haven’t.
- Customers are majorly using Fiber Optic Internet Service followed by DSL. Almost ~78% customers haven’t opted for Internet Service. After opting for internet service, most of the customers have not signed up for service that comes with it like Online Security/Backup, Device protection and Tech support.
  Out of all the customers who have signed up for Internet service, there is almost an equal percentage of customers who have/have not signed up for Streaming TV/movies.

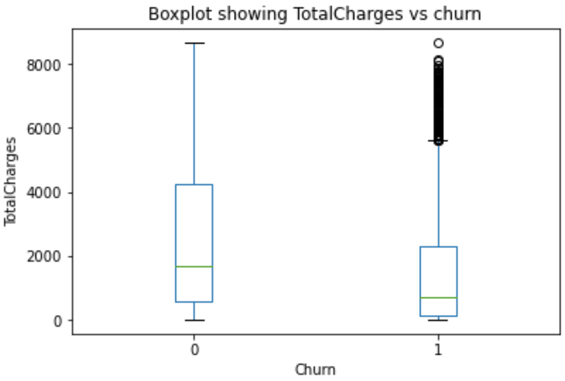
<u>Exploring Customer Account information:</u>



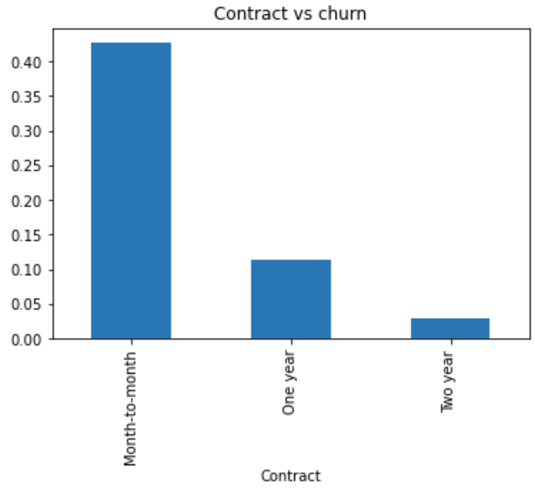


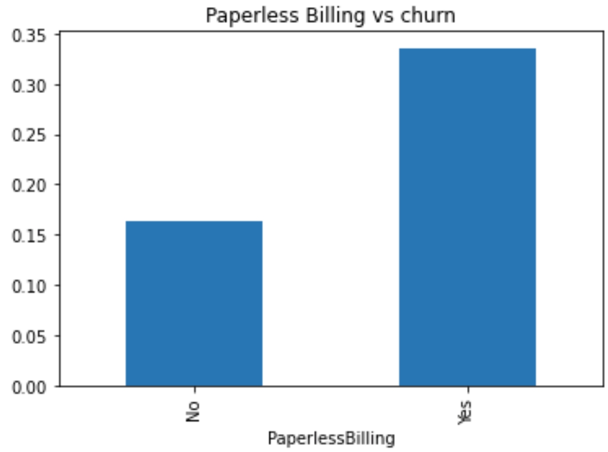
Customers with higher average monthly charges are getting churned.

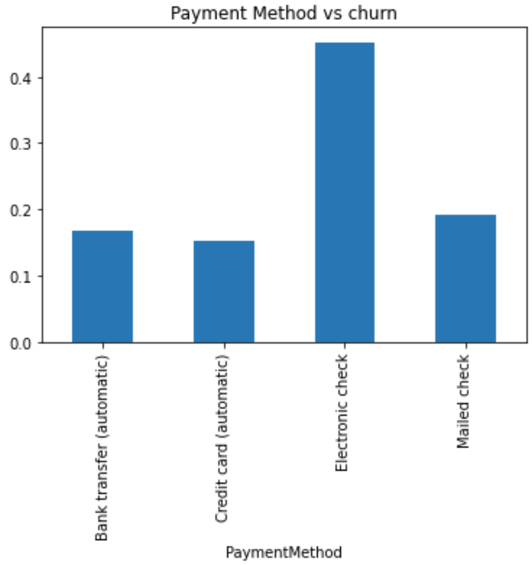




Customers with low average total charges are getting churned. From the boxplot, we can see that Total charges for churned customer have some outliers. This means that there are some customers who are paying high total charges but getting churned.







- Month-to-month contract plan has the highest customer churn rate and two-year plan has the highest retention.
- Customers who are using paperless billing are churning more than those not using paperless billing.
- Customers using Electronic check as payment method have higher churn rate.

METHODS:

Since Churn is a categorical variable, we will use algorithms like KNN, Classification trees, Random forest, Boosted trees and Logistic regression for predicting customer churn.

KNN: KNN is a supervised machine learning algorithm mostly used for classification. It classifies the new data points based on the similarity measure of the earlier stored data points.

We have categorized the output variable and created dummies wherever required.

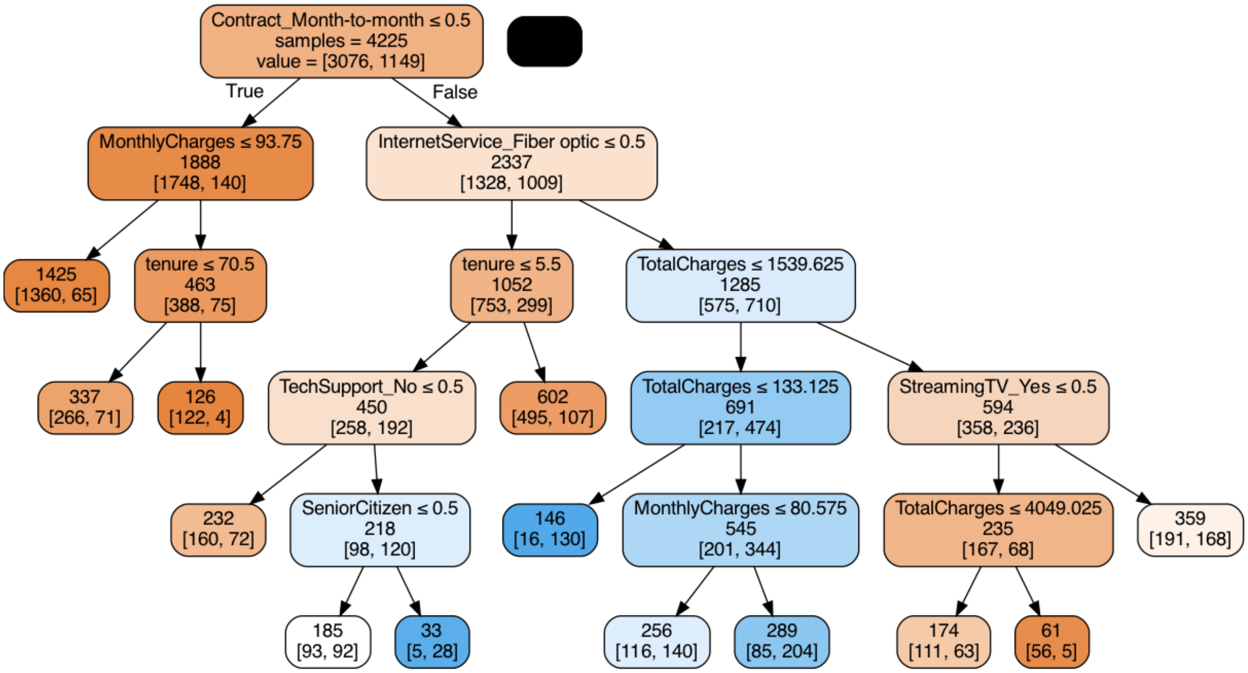
We normalized the numerical data and split the data into training and validation set.

We initially checked validation accuracy on K = 3 which is 75.15%

To get the optimal K, we have trained the classifier for different values of K (1 to 20). We are getting the best accuracy at K = 16 which can lead to uncertainty if neighbors are 50% of one and 50% of the second class.

The next best accuracy at an odd K is **79% at K = 13**.

Classification trees: A classification tree helps us to determine a set of if-then logical (split) conditions that permit accurate prediction or classification of cases.



The above classification tree is a pruned tree, obtained by performing a grid search and fine tuning the parameters.

Parameters used are:

- Maximum Depth : 7
- Minimum impurity decrease : 0.0011
- Minimum sample split : 10

Accuracy of this tree is : 79.52%

Random Forest: Random forest is a method for improving predictive power by combining multiple classifiers or prediction algorithms.

Results from a Random Forest cannot be displayed in a tree-like diagram, thereby losing the interpretability that a single tree provides.

Accuracy using Random Forest: 79.21%

Boosted Trees: In Boosted Trees, a sequence of trees is fitted, so that each tree concentrates on misclassified records from the previous tree. Results from a Boosted tree also cannot be displayed in a tree-like diagram, thereby losing the interpretability that a single tree provides.

Accuracy using Boosted Tree: **80.02%**

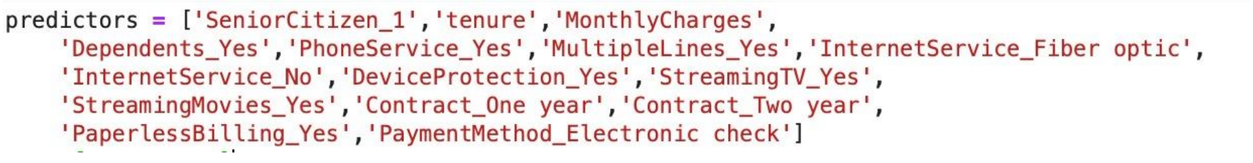
Logistic Regression: It is a supervised learning algorithm used to predict a dependent categorical target variable.

While building our logit we have removed the variable ‘Total charges’ to avoid multicollinearity.

Model1: We have created our first model using all 29 variables (including dummy variables). The accuracy on validation data of Model 1 is 80.34% .

Model2: Based on logit prediction of Model 1 and GLM regression results we have considered 15 important variables to build our model 2. The accuracy on validation data for Model 2 is **80.55%**

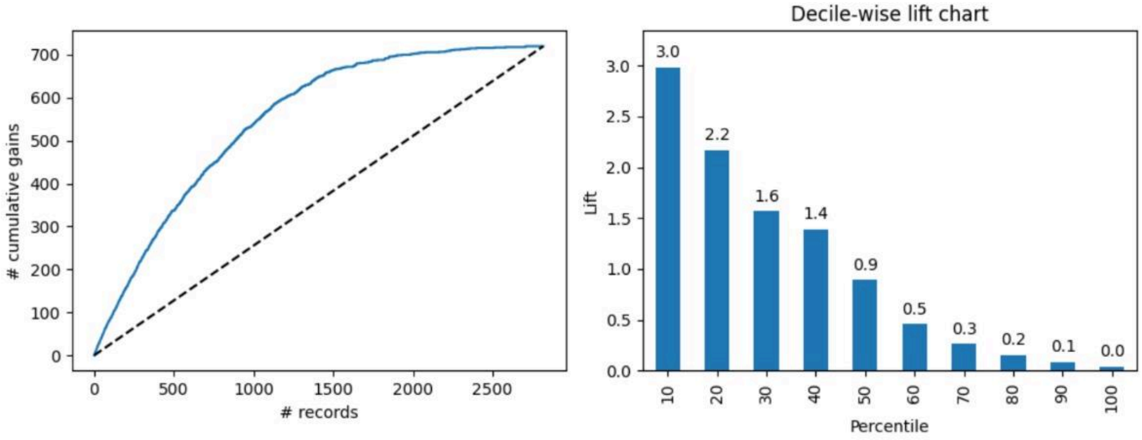
Variables used in building Model 2:



```
predictors = ['SeniorCitizen_1','tenure','MonthlyCharges',
    'Dependents_Yes','PhoneService_Yes','MultipleLines_Yes','InternetService_Fiber optic',
    'InternetService_No','DeviceProtection_Yes','StreamingTV_Yes',
    'StreamingMovies_Yes','Contract_One year','Contract_Two year',
    'PaperlessBilling_Yes','PaymentMethod_Electronic check']
```

Hence, we will use Model 2.

Cumulative Gains and Decile wise Lift Chart for Model 2



We can see that above cumulative gains chart is pulled towards upper left corner which shows that the model has a good predictive performance. Similarly, the decile wise lift chart also follows a staircase pattern (higher values in initial deciles) indicating a good model.

Comparison:

METHODS	ACCURACY
KNN	79%
CLASSIFICATION	79.52%
RANDOM FOREST	79.21%
BOOSTED TREES	80.02%
LOGISTIC REGRESSION	80.55%

We are getting maximum accuracy on Logistic Regression.

Recommendation:

We will recommend Telco company to use logistic model. Based on 15 important variables used in model 2, Telco can predict if a new customer will churn or not.

Based on our initial exploration and important variables of logit model(Model2), Telco can implement below-mentioned recommendations:

- Promote long-term contracts more in order to convert month to month customers to a longer contract period like one year or two years.
- Special combo pack for senior citizens as per their needs.
- Providing exciting offers to customers if they opt for a multiple line.
- Offer complimentary streaming service along with internet to target those customers who haven’t tried this service at all.