IE 500: STATISTICAL MACHINE LEARNING

CUSTOMER CHURN PREDICTION AND ANALYSIS IN TELECOMMUNICATION

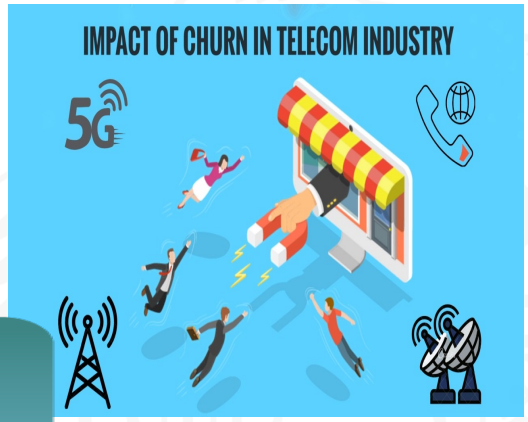
By:
Awnish Shankar 50542202

University at Buffalo The State University of New York

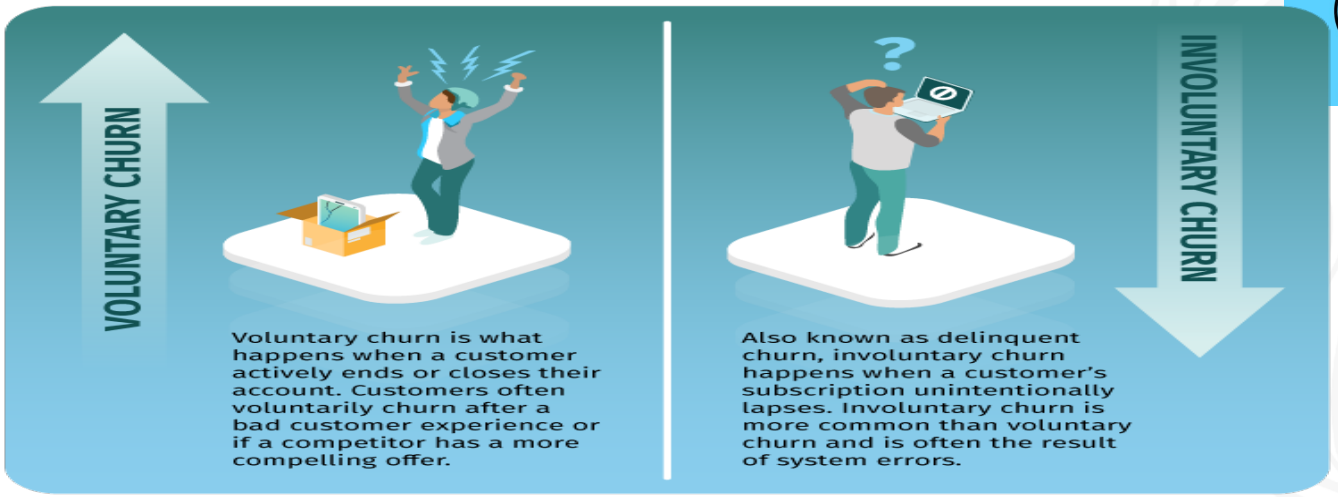
Introduction

Customer churn is the rate at which customers stop using your company's product or service during a certain time frame.

In some countries, the cost of customer acquisitions could be as high as 20 times the cost of customer retention.



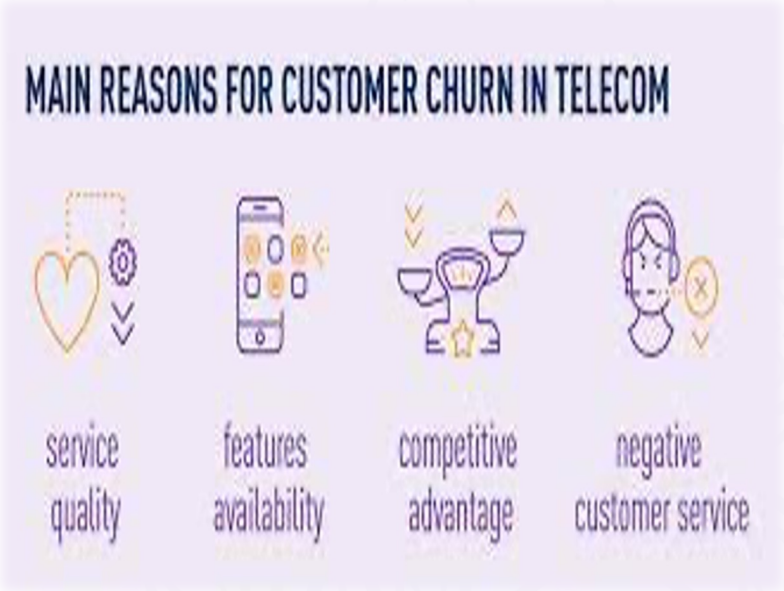
By:- National University of Sciences and Technology, Pakistan.



Size of the Problem

Figures From the Year 2022

- 400 M subscribers in the U.S. telecommunications industry for different Services.
- Churn rate was 1.9% across the top four carriers (AT&T, Verizon, T-Mobile, Sprint).
- The cost of new customer acquisition per carrier was $10.39 B.
- The yearly loss due to customer churn per carrier was just $780 M which is approx. 1/13th the cost of the customer acquisition.



Objective of the study

- The study aims to address the works in the domain of customer churn in telecommunication and propose and apply different ideologies to further elevate the predictive performance.
- To identify the relative importance of different variables, and how telecom companies can optimize resources, focusing efforts on these variables that have a significant impact on reducing customer churn.
- To propose applications of other significant models for further enhance predictive accuracy and to draw actionable insights into customer relationship management.

Data Introduction

- The dataset is taken from a project on Kaggle which claims to have used the IBM Sample dataset.
- The dataset has 7043 observations and 21 variables. Out of 21, 3 are numerical, and 18 are categorical variables.

Variables:

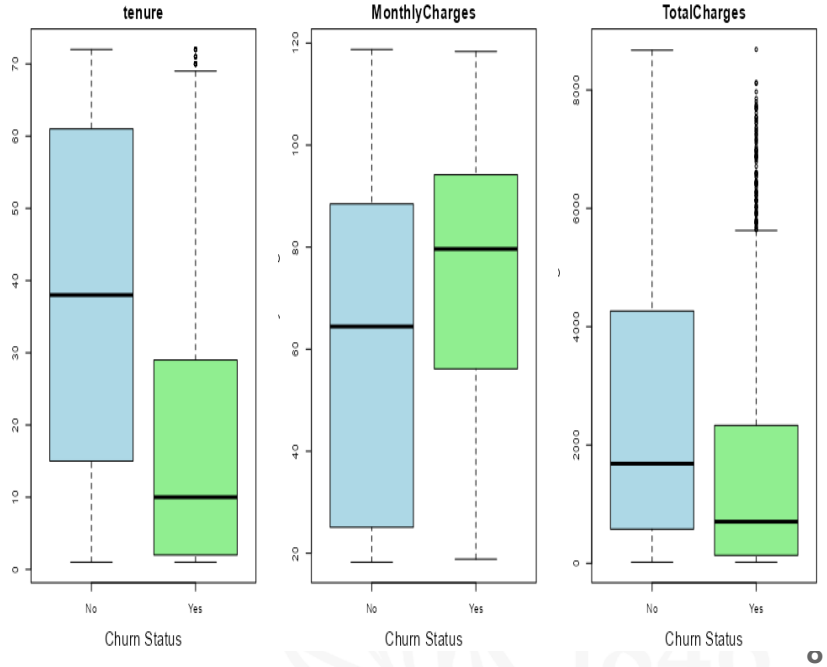
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges,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

Variable Name	Type	Explanation	Level(and its abbreviated form)
Customer ID	Categorical	Unique No.	
Gender (Categorical)	Categorical	Whether the customer is a male or a female	Male Female
Senior Citizen (Categorical)	Categorical	Whether the customer is a senior citizen or not	Yes No
Partner (Categorical)	Categorical	Whether the customer has a partner or not	Yes No
Dependents (Categorical)	Categorical	Whether the customer has dependents or not (Yes, No)	Yes No
Phone Service (Categorical)	Categorical	Whether the customer has a phone service or not (Yes, No)	Yes No
Multiple Lines (Categorical)	Categorical	Whether the customer has multiple lines or not	Yes, No, No phone service(NPS)
Internet Service (Categorical)	Categorical	Customer's internet service provider	DSL, Fiber optic(Fo), No
Online Security (Categorical)	Categorical	Whether the customer has online security or not	Yes, No, No internet service(NIS)
Online Backup (Categorical)	Categorical	Whether the customer has online backup or not	Yes, No, No internet service(NIS)
Device Protection (Categorical)	Categorical	Whether the customer has device protection or not	Yes, No, No internet service(NIS)
Tech Support (Categorical)	Categorical	Whether the customer has tech support or not	Yes, No, No internet service(NIS)
Steaming TV (Categorical)	Categorical	Whether the customer has streaming TV or not	Yes, No, No internet service(NIS)

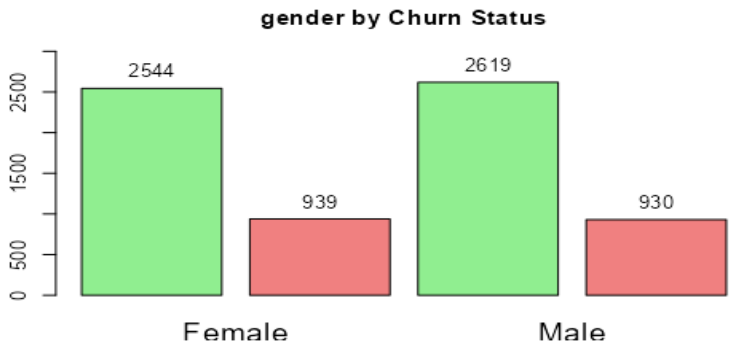
Steaming TV (Categorical)	Categorical	Whether the customer has streaming TV or not	Yes, No, No internet service(NIS)
Streaming Movies (Categorical)	Categorical	Whether the customer has streaming movies or not	Yes, No, No internet service(NIS)
Contract (Categorical)	Categorical	The contract term of the customer	Month-to-month(Mm), One year(Oy), Two year(Ty)
Online Paperless Billing (Categorical)	Categorical	Whether the customer has paperless billing or not	Yes No
Payment Method (Categorical)	Categorical	The customer's payment method	Electronic check(Ec), Mailed check(Mck), Bank transfer(Bk) (automatic), Credit card (automatic)(Cc)
Monthly charges (Numerical)	Numerical	The amount charged to the customer monthly	18.25……118.75
Total Charges (Numerical)	Numerical	The total amount charged to the customer	18.8…….. 8684.8
Tenure (Numerical)	Numerical	Number of months the customer has stayed with the company	1……72
Churn(Target Variable) (Categorical)	Categorical	Whether the customer churned or not	Yes No

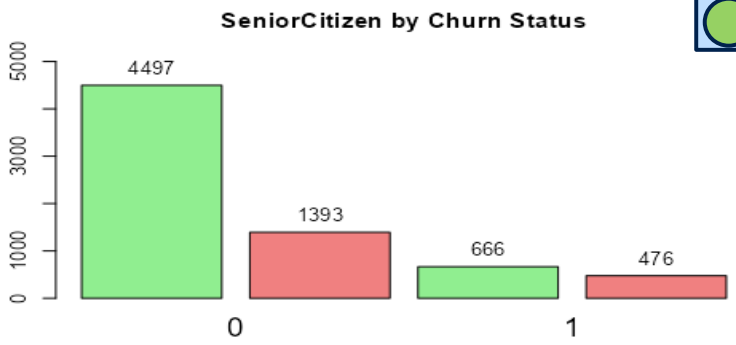
Data Exploratory Analysis

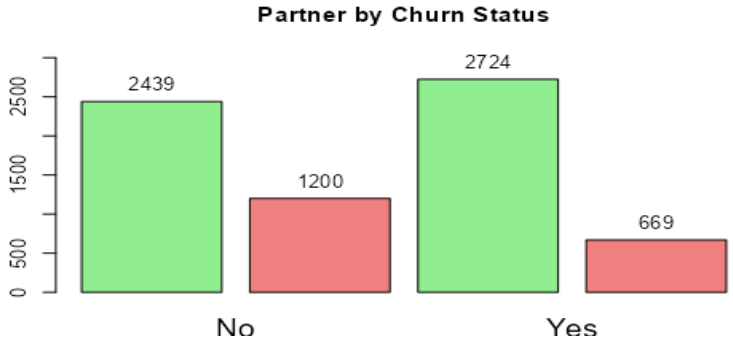
	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.25	0.83
MonthlyCharges	0.25	1.00	0.65
TotalCharges	0.83	0.65	1.00

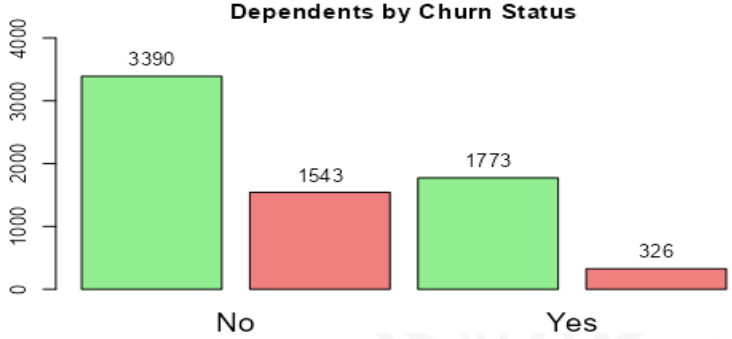


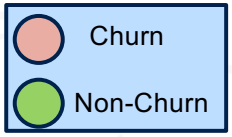
Data Exploratory Analysis



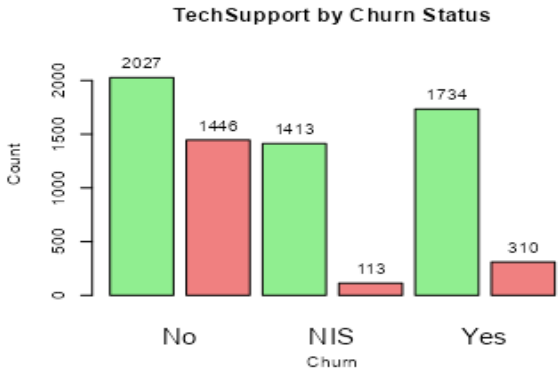


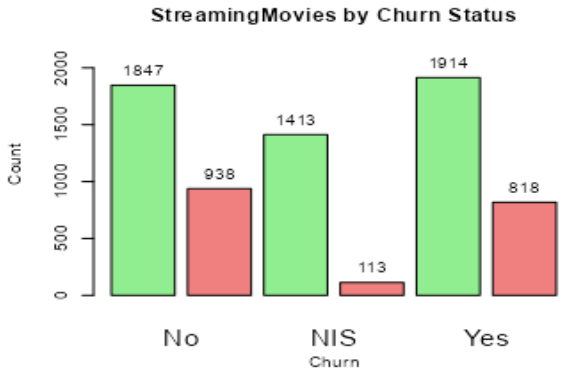


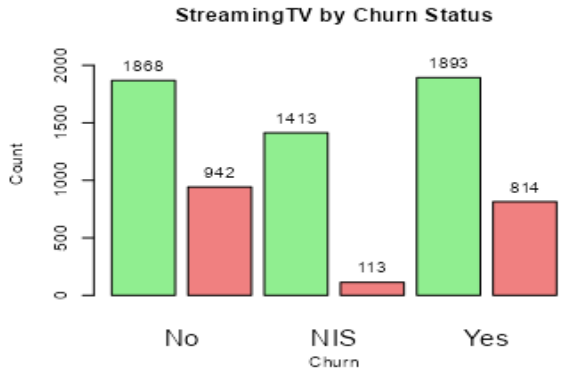


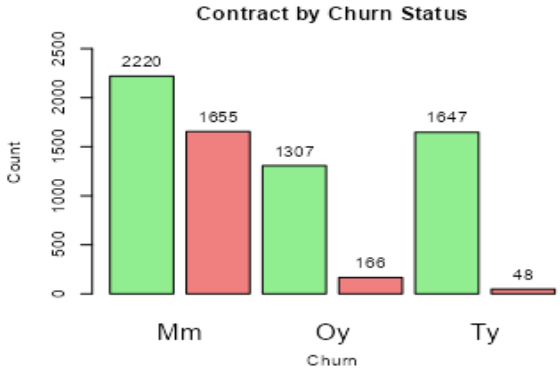


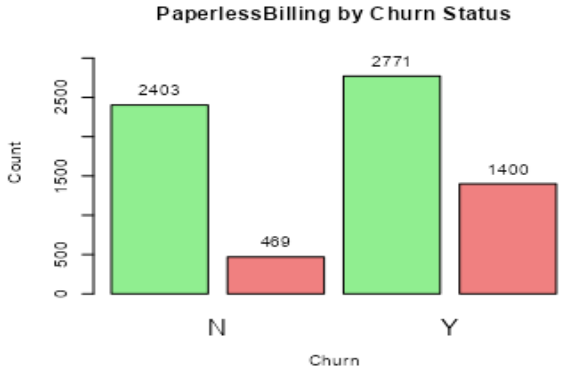
Data Exploration Analysis

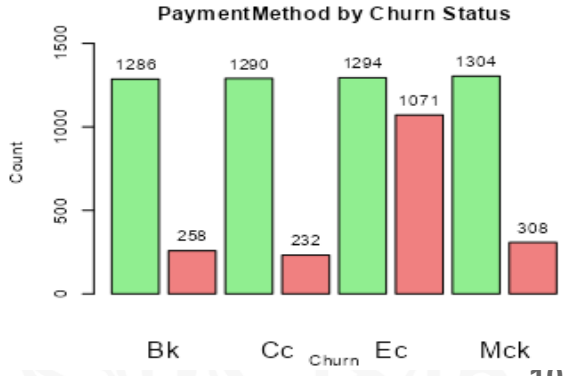




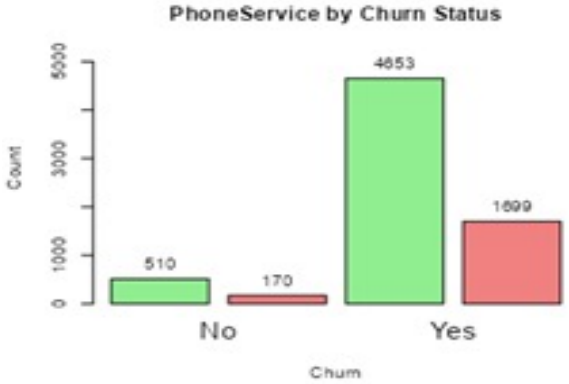


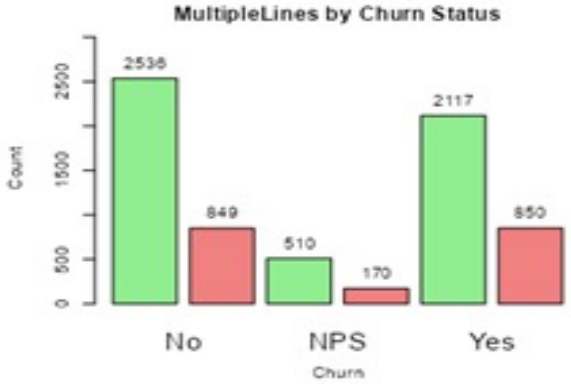


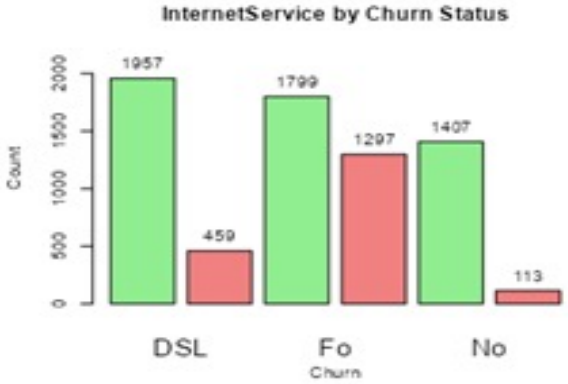


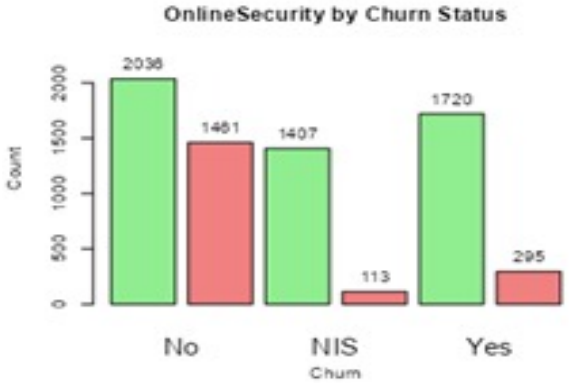


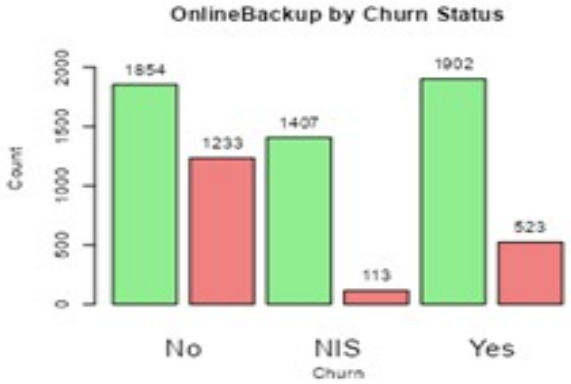
Data Exploration Analysis

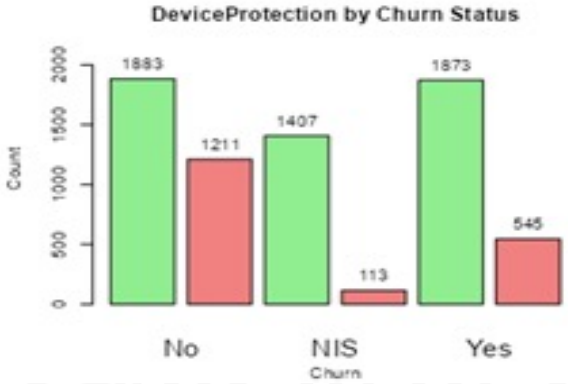




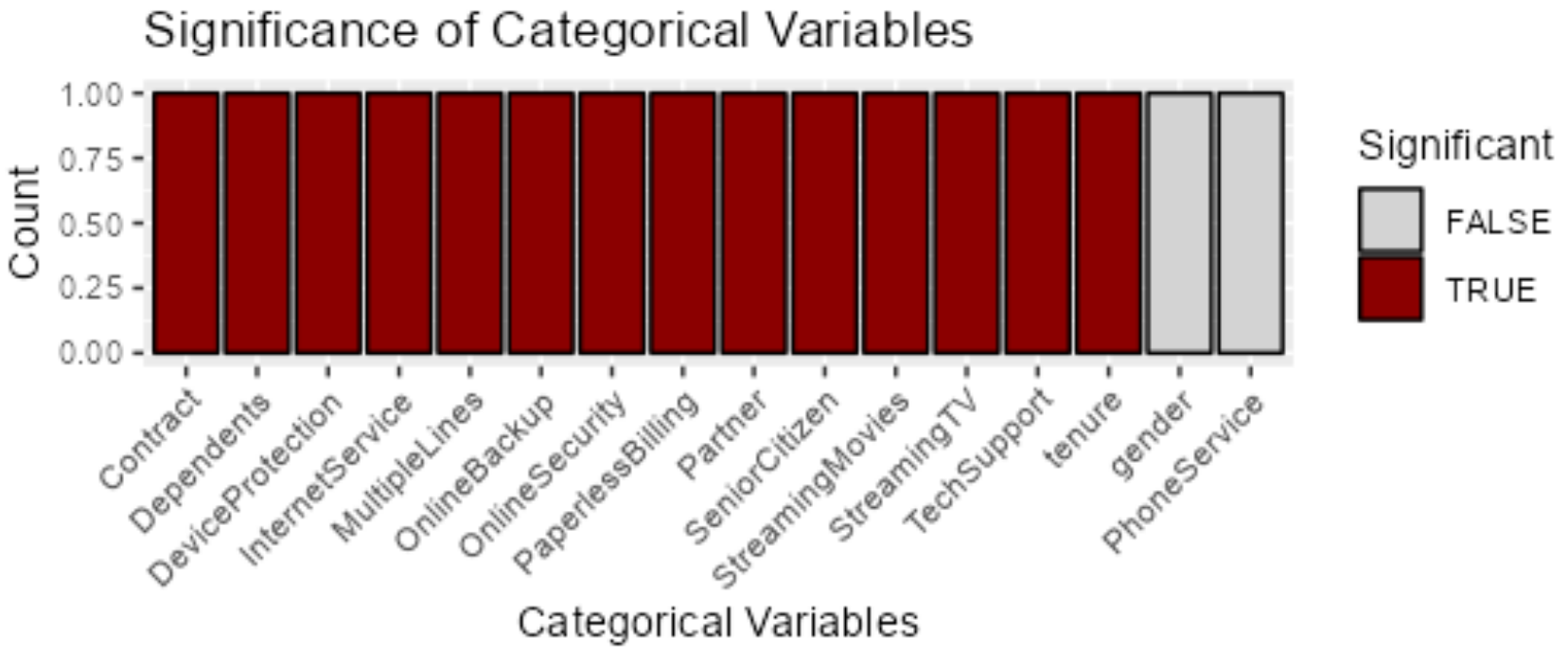




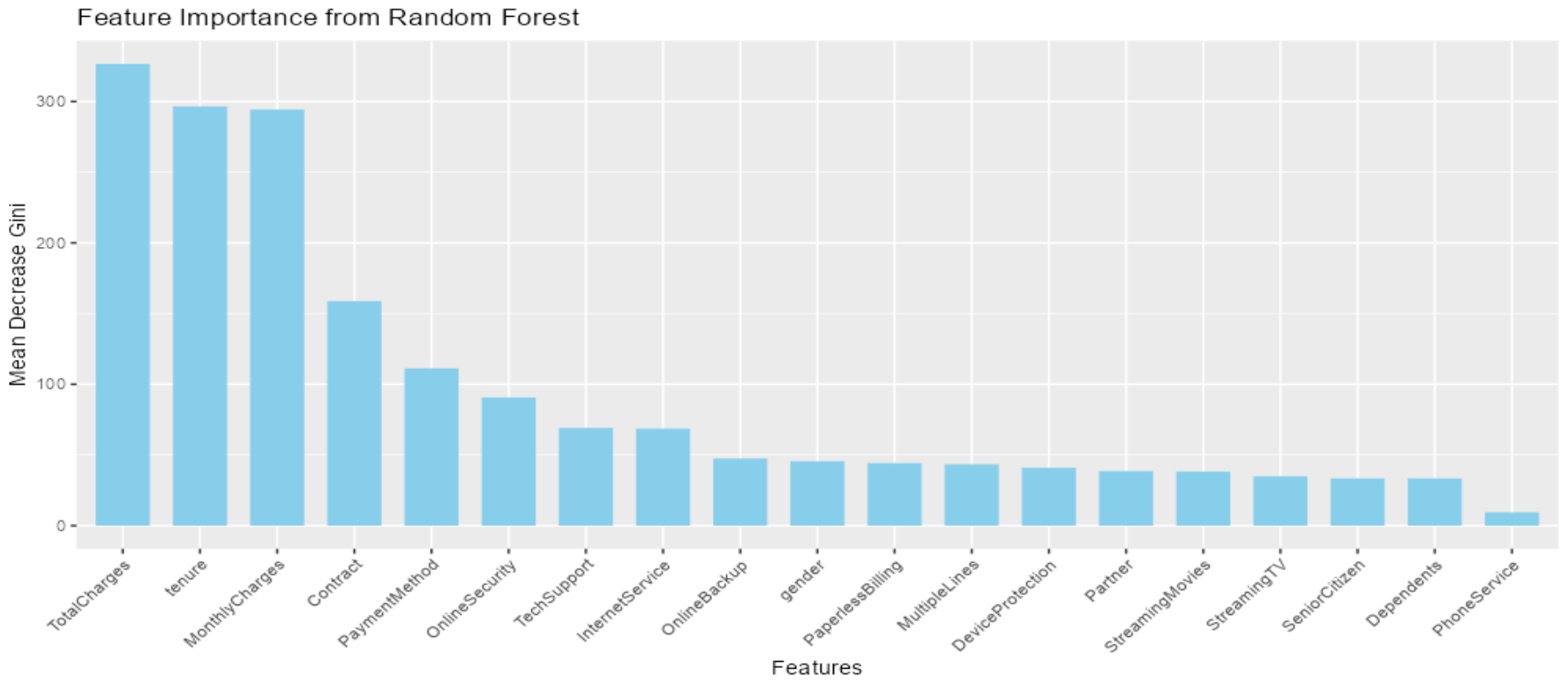




Identifying Significant variables



Data Exploration

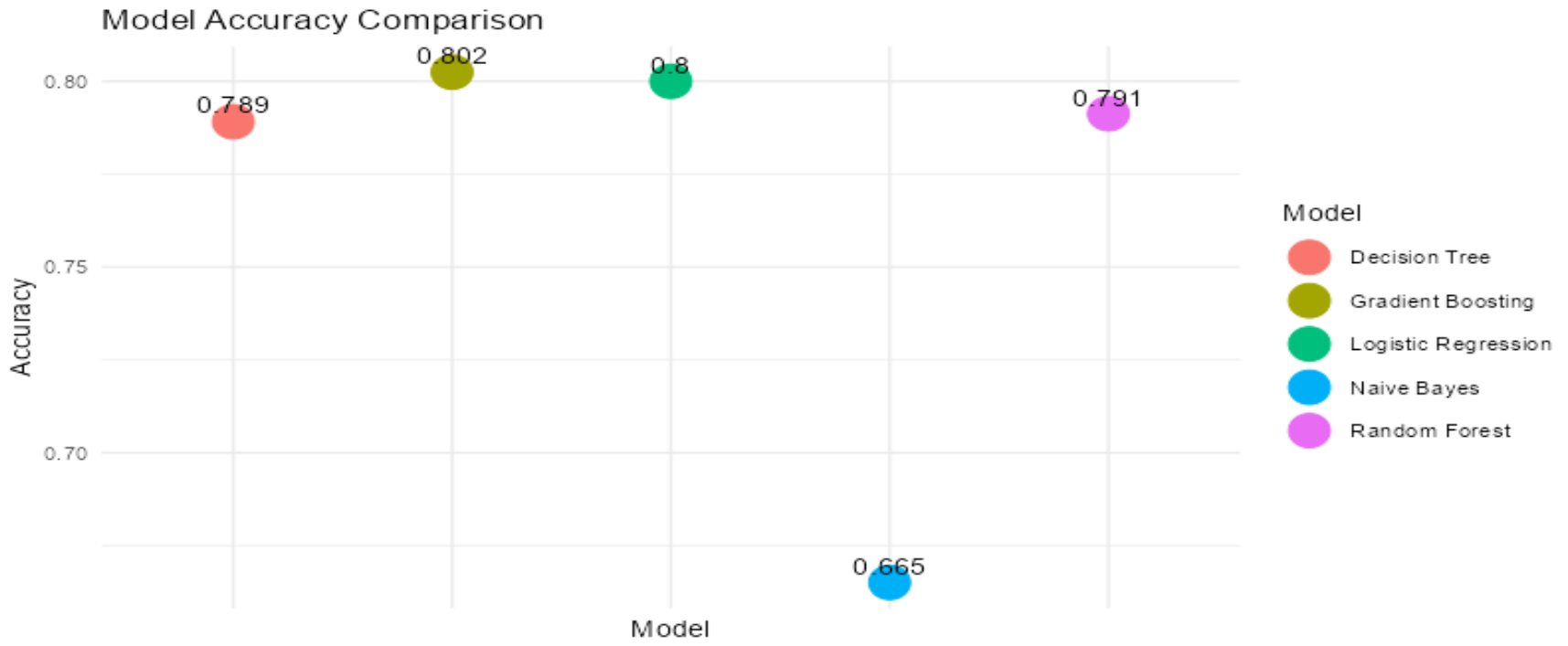


Models Performance Evaluation

- Dataset Split: 75% Train data and 25% Test data.
- All the models are trained and evaluated on a train-set using 5-fold cross-validation.
- For all the models, the variables gender and tenure are dropped as per the chi-square test.
- Logistic regression.
  - To address the collinearity, the highly correlated categorical variables such as streaming TV, online backup up, and device protection.
  - For numerical variables, Tenure was found to be substantially correlated with Total charges. The model was tested with and without tenure and found to have a 0.05 % increase in accuracy.

Model	Accuracy(%)
Gradient Boosting	80.02
Logistic regression	80
Random forest	79.1
Decision Tree	78.9
Naïve Bayes	66.5

Models Accuracy



Final Model

- **Logistic Regression**
- **Accuracy: 79.74 %**
- **Reason:**
  - Even though the gradient boost has higher accuracy on the train set, the increase is very marginal as compared to Logistic Regression.
  - Given the emphasis on customer retention in the telecommunications industry, logistic regression's straightforward interpretation and the ability to identify significant predictors make it a practical choice for informing strategic decisions in customer relationship management.

```
          Reference
Prediction   No  Yes
       No  1151  217
       Yes  139  250

               Accuracy : 0.7974
                 95% CI : (0.7778, 0.816)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : 4.248e-10

                  Kappa : 0.4516

 Mcnemar's Test P-Value : 4.484e-05

            Sensitivity : 0.8922
            Specificity : 0.5353
         Pos Pred Value : 0.8414
         Neg Pred Value : 0.6427
             Prevalence : 0.7342
         Detection Rate : 0.6551
   Detection Prevalence : 0.7786
      Balanced Accuracy : 0.7138

       'Positive' Class : No
```

Conclusion

- Gender and Phone Services are found to be insignificant variables among all the predictors.
- Given the type of problem, Logistic regression was found to be the most relevant among all 5 models with an accuracy of 79.74 % on the test set.
- **Future Work:** Want to apply time series analysis, Customer segmentation techniques, and deep learning methods to check if we can further increase the accuracy.

Thank You