Georgia State University

J. Mack Robinson College of Business



**Telecom Customer Churn Prediction**

Spring 2023

Graphical user interface

Description automatically generated

By:

Akhil Golla  
Krishna Reddy  
Parag Parab  
Sravya Basireddy  
Srirag Nair K M

Tejaswini Chintala

INDEX

|  |  |
| --- | --- |
| **Topic** | **Page Number** |
| Project Description | 3 |
| Data Exploration and Preprocessing | 4 |
| Graphs or plots to explore the data | 5 |
| Data Partition | 5 |
| Model  Result and Discussion  Summary | 9  10  10 |

**Project Description:**

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. **Telecom Customer Churn Prediction** helps one identify such customers so that telecom companies can initiate corrective actions to retain them.

**Business Problem:**

This simulated situation concerns a telecom company that wants to understand why its customers unsubscribe to their services and predict who is most likely to leave next.

We use the word “churn” to represent such voluntary unsubscriptions. This project focuses explicitly on the cause of customers churning the telecom services.

Going through data of past customers and assessing the key attributes can give great insight into churn trends. There would be certain common threads connecting all the customers who have churned.

Dataset is from Kaggle. The dataset consists of 7043 rows and 21 columns. It includes attributes such as customer id, gender, senior citizen, partner, dependents, tenure, phone service, multiple lines, internet service, online security, and many more. The features are of type Categorical, Boolean, and Numeric. Churn is the categorical and discrete target feature in the dataset.

The first part of this report: Descriptive and Exploratory Data Analysis

The second part of this report: Predictive Analysis

This paper illustrates an ensemble model approach to generate the chances of customers leaving. Ensemble modeling involves training multiple models and combining their predictions to derive the predictions submitted to Kaggle. The specific ensemble approach illustrated is called model stacking.

**Goal:** This project aims to develop a model(s) to predict which customer is going to churn the telecom company. The plan is as below:

* Assemble the data and explore it.
* Clean variables, build what is needed.
* Models: Logistic, Decision Tree, Random Forest, and Neural Network
* Choose the best model and predict the entry.

Diagram

Description automatically generated

**Data Exploration and Preprocessing:**

Identify the output variable and input variables.

*churn* is our target variable and the dependent variable for prediction, fields in the dataset include:

|  |  |
| --- | --- |
| * Monthly Charges | * Paperless Billing |
| * Gender | * Payment Method |
| * Partner | * Tenure |
| * Dependents | * Internet Service |
| * Contract | * Online Security |

**Data Architecture:**

**Diagram

Description automatically generated**

**Data Cleaning:**

* **Converting categorical variables to Yes/No & removing unnecessary fields**

Graphical user interface, text, application

Description automatically generated

**1. Summary of the dataset**

The dataset looks like

Text

Description automatically generated

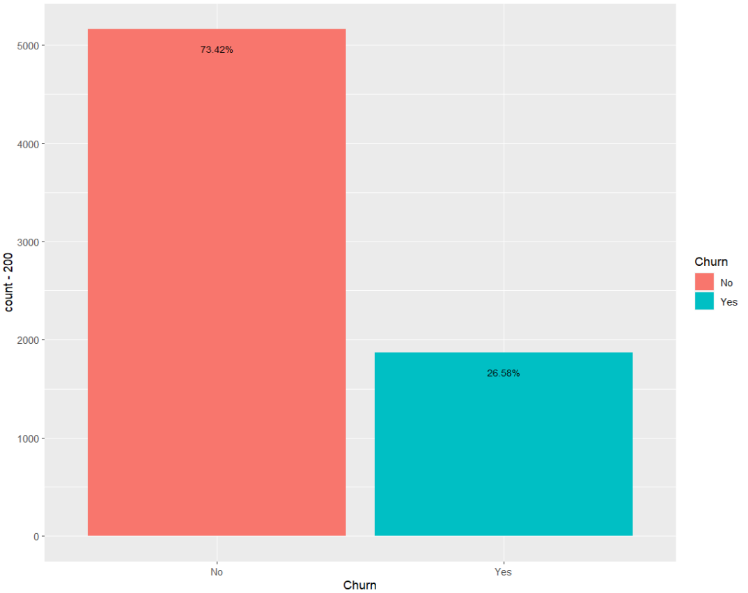
**2. Correlation**

Below is the correlation plot with the attributes TotalCharges, MonthlyCharges, and tenure.

Chart, bubble chart

Description automatically generated

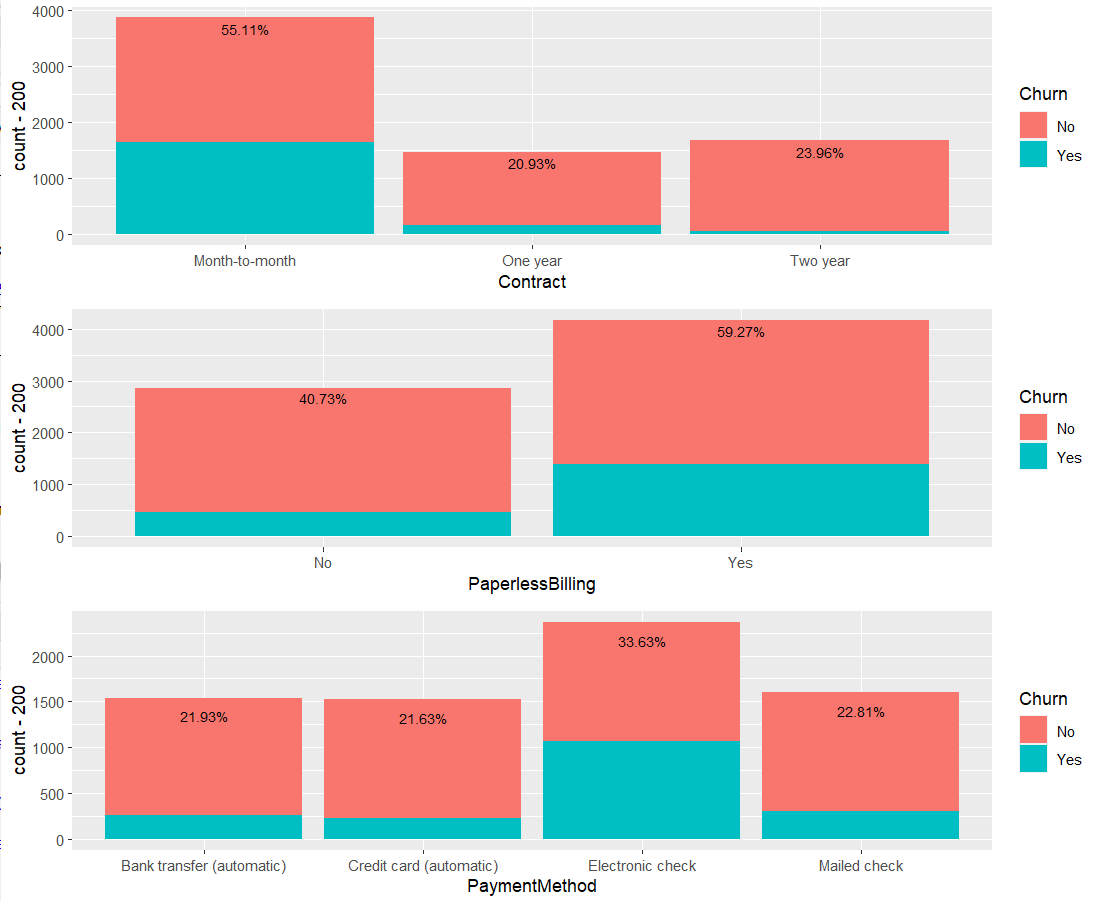
**3. Bar Plots (Churn vs. Count):**

****

From the data, we find that the percentage of customers churning is 26.58% whereas the percentage of customers not churning is 73.42%.

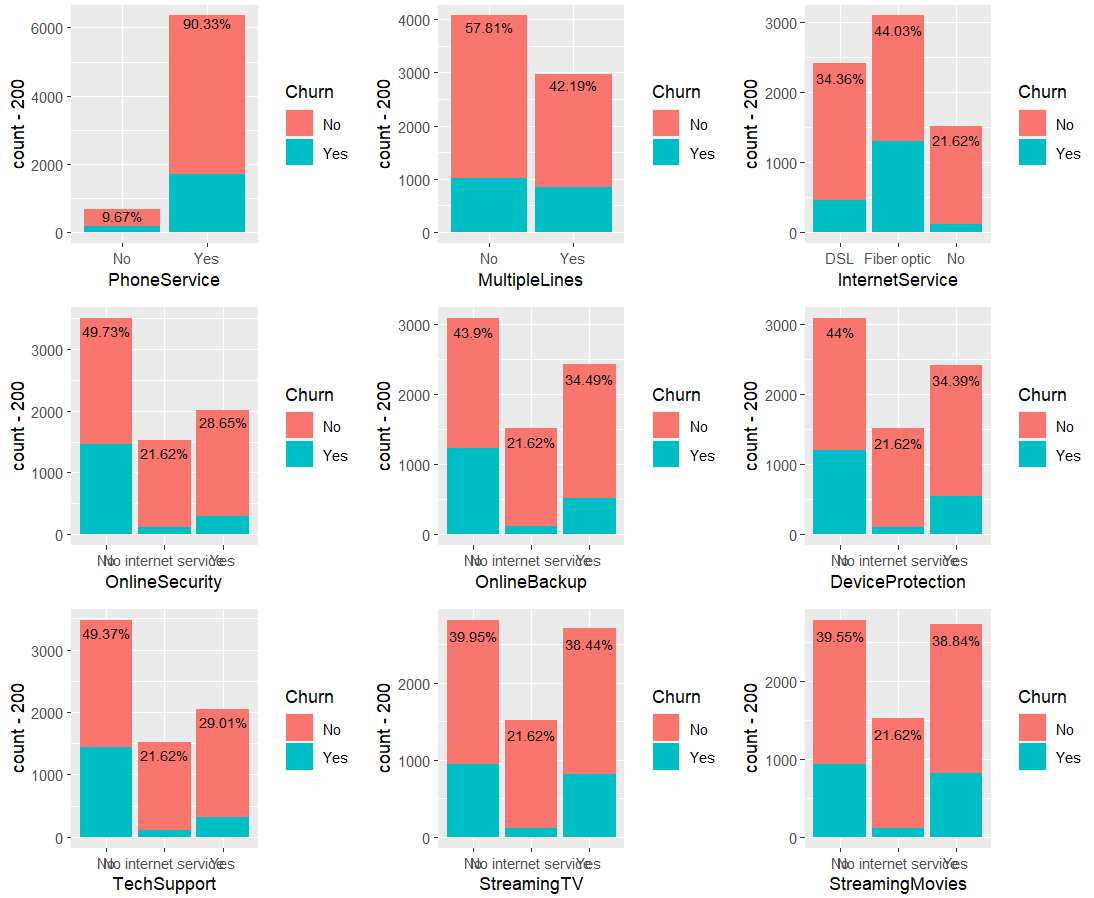
**4. Bar Plots (gender, SeniorCitizen, Partner, Dependents, Contract, PaperlessBilling, and PaymentMethod vs. Churn):**

Chart, bar chart, treemap chart

Description automatically generated 

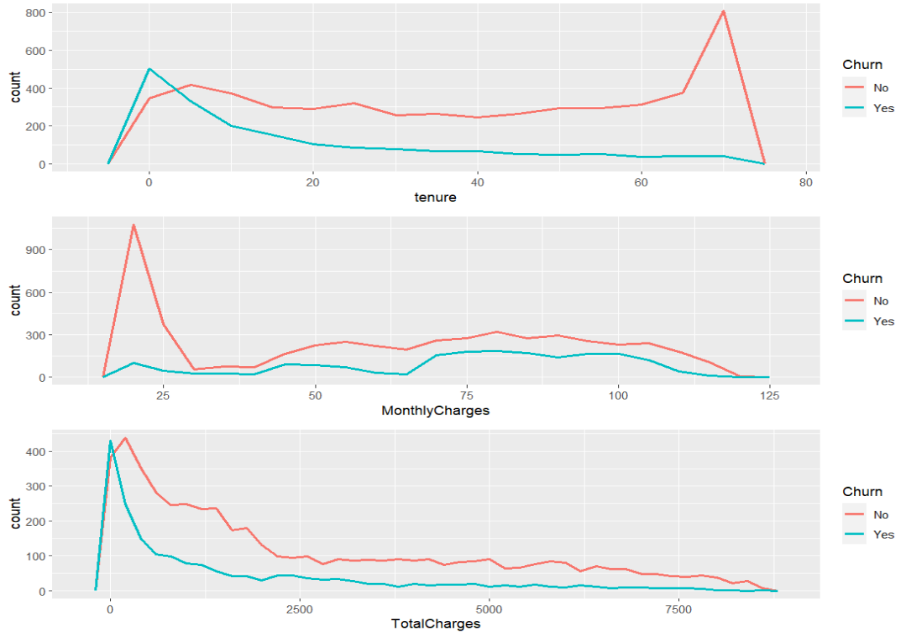
The above bar plots show the correlation of the attributes gender, SeniorCitizen, Partner, and Dependents**,** Contract, PaperlessBilling, and PaymentMethod with respect to the output variable Churn.

**5. Bar Plots (Various Telecom Plans vs. Churn):**



The above bar plot shows the correlation of various telecom mobile plans with respect to the Churn.

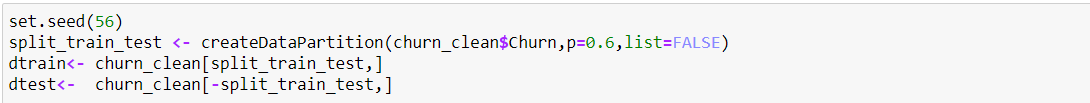
**7. Line Plots (tenure, MonthlyCharges, and TotalCharges vs. Churn):**

****

The above line plot shows the correlation of tenure, MonthlyCharges, and TotalCharges with respect to the Churn. We can observe that the Churn count is less among Customers with more tenure and more among customers with less tenure.

Similarly, we can observe the trends of customer churn with respect to MonthlyCharges and TotalCharges as well.

**Data Partition**We did data partition by using the Caret partitioning function.



**Models and their comparison:**

The models we will compare are the following:

* Logistic Regression
* Decision Tree
* Random Forest
* Neural Network
* Ensemble method

## Reasons for specific Model Selection

**Logistic Regression**

Logistic Regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary result, which can take two values: true/false, yes/no, and so on. As our output ‘churn’ is either yes or no, we used logistic Regression.

Text

Description automatically generated Text

Description automatically generated Text

Description automatically generated

The accuracy of the logistic regression came out to be 81.22% for the test data.

**Decision Tree**

Decision tree is a tree-like model used for classification purpose. As Decision trees are intuitive and easy to understand, we have used Decision Tree model to make it easy to interpret the results.

**Graphical user interface, application, Word

Description automatically generated**

The accuracy of decision tree turned out to be 79.23% for the test data.

**Text

Description automatically generated** **Text

Description automatically generated**

**Random Forest**

Random forests or random decision forests are an ensemble learning method for classification, Regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the categories or mean prediction of the individual trees.

Graphical user interface, text, application

Description automatically generated

The accuracy of Random Forest model turned out to be 79.45%.

Chart

Description automatically generated Text

Description automatically generated

**Neural Network**

A robust implementation must consider feature engineering, data cleaning, and cross-validation.

Graphical user interface, text, application

Description automatically generatedA picture containing chart

Description automatically generated

**Ensemble Method:**

Ensemble modeling involves training multiple models and combining their predictions to derive the projection. We firstly apply the ensemble to the model using averaging. We also tried using a weighted average.

## Model Comparison:

## We can see that Logistic Regression is doing better than the other models below.

Table

Description automatically generated

# **Result and Discussion**

Results from Logistic Regression came out to be the best. Decision tree and Random Forest are also performing well but comparatively, have a lower accuracy.

# **Summary**

By Analyzing the accuracies of Logistic Regression, Decision Tree, and Random Forest, we can observe that the accuracy of Logistic Regression is the best with is 81.22%. So, we could conclude that the Logistic Regression Model is the best prediction model to classify customer churn based on the parameters from the data set we have.

# **Test File:**



# **R Code:**



**PPT:**

****