**Business Analytics with R**

**BUAN 6356.006 Fall 2022**

**Group – 09**

**Final Project Report**

Telco Churn Prediction

## Moksh Mehool Mehta (mxm220009)

## Rithwik Reddy Koripelly (rxk210087)

## Spandana Aelapati (sxa220091)

## John Hart (jmh131130)

## Koti Reddy Gangasani (kxg220019)

# Abstract

Data analysis on customer churn is crucial to understanding and addressing the issue of churn in any business. This study uses data on customer behavior and characteristics to analyze customer churn in a specific industry.

Customer churn analysis and prediction in the telecom industry is a problem nowadays since the sector must examine customer behavior to identify those who will cancel their subscription. Because acquiring new customers costs more than keeping existing ones, machine learning techniques and algorithms are crucial for businesses in today's commercial environment.

In the telecom industry, customer churn can significantly impact a company's revenue and profitability. In this study, we analyze customer churn in the telecom industry using customer behavior and characteristics data. We use machine learning algorithms to identify factors associated with a higher likelihood of churn and develop strategies to reduce churn and improve customer retention.

In this project, we create and compare classification models using machine learning techniques for forecasting customer churn, such as lazy learning, random forest, and logistic regression.

Our goal is to find a data-driven solution that would enable us to lower churn rates, which will boost customer happiness and business income.

Table of Contents

1.0 Executive Summary

2.0 Project Motivation and Background

3.0 Data Description

4.0 Exploratory Data Analysis

5.0 BI Model

6.0 Model Evaluation

7.0 Conclusion

8.0 Future Enhancements

9.0 References

10.0 Project Presentation Link

# Executive Summary

In the telecommunications sector, customer churn can have a significant impact on a company's revenue and profitability. In this study, we analyze customer churn in the telecommunications sector using data on customer behavior and characteristics. We use machine learning algorithms to identify factors that are associated with a higher likelihood of churn and develop strategies to reduce churn and improve customer retention.

We begin by collecting and cleaning data on customer behavior and characteristics. This includes information on customer demographics, usage patterns, and interactions with the company. We then use machine learning algorithms including logistic regression, k nearest neighbors and random forests to identify patterns and trends in the data and develop predictive models to forecast churn. Finally, we analyzed the performance of all three models using the ROC curve.

People who are most likely to depart the firm soon are identified using the churn model. In addition to client retention, churn results can be used efficiently for various other goals. Based on our findings, we recommend several strategies to reduce churn and improve customer retention. These include offering competitive pricing, providing high-quality customer service, and implementing loyalty programs. We also suggest regular monitoring of customer satisfaction and implementing targeted interventions to prevent churn among at-risk customers.

Overall, our study provides valuable insights into customer churn in the telecommunications sector and offers practical strategies for reducing churn and improving customer retention. By implementing these strategies, companies in the telecommunications sector can improve their bottom line and maintain a competitive edge in the market.

# Project motivation and Background

Our team always wanted to implement the learnings of the topics learned in the BA with R subject on a problem set where the results/ findings would help or create impact among valued customers and people in general. Our main agenda is to develop a business intelligence model that can help any company profile in their business decision making to make their way into profits. One such idea that reflected to us was to solve the problem that are facing by telecommunication industry’s churn rate. It was a great opportunity for our group to leverage our knowledge of Business Analytics on to something which is a serious problem out there in the market and has a lot of scope to improve our findings so that the churn rate decreases as much as possible.

# Data Description

Among 19 variables in the data set, we ran separate reports on Senior Citizens, Monthly Charges, and Tenure. We viewed these variables as having the most significant impact on whether a customer chose to leave the service or not. Senior citizens have the lowest phone usage, customers always want to avoid high phone costs, and customers that have been using the service a longer time help show loyalty to the brand. We found these to be important variables to analyze separately as well. We believed that they had a strong impact on the model and wanted to see the data trend. Other variables that we examined extra were Payment Method, Contract, and Dependents.

During our regression model analysis, we wanted to look at key variables and see if their Individual churn rates were statistically significant. We performed this by looking at their individual p values of the customer churning of not churning based on the effect of the variable. This helped us verify and remove any unnecessary variables such as gender and phone services.

R Packages Used

* ggplot2
* Latice
* Caret
* Plyr
* pROC

ggplot was used for numerous plot graphs of various variables. The Latice, caret, and plyr packages were used for the regression and BI models that we created on the dataset. pROC package constructs the ROC curve which is critical in determining what model is appropriate for evaluations and decision making. Logistic Regression showed the highest performance amongst the models with the highest AUC (area under curve) line.

# Exploratory data analysis

Exploratory data analysis is a process of examining and summarizing a dataset to understand its content and context better. It involves several different techniques, such as visualizing the data, summarizing its main characteristics, and identifying any trends or patterns that may be present.

The goal of exploratory data analysis is to provide a better understanding of the dataset and to identify any potential issues or problems that may need to be addressed before further analysis can be done.

It is important to explore the data before analyzing it for the following reasons:

* First, exploring the data can help us gain a better understanding of what the data contains, and how it is structured. This can help us identify any potential issues or problems with the data, such as missing values or errors, which can affect the accuracy of your analysis.
* Additionally, exploring the data can also help us identify any trends or patterns in the data, which can provide valuable insights and help us develop more effective analysis techniques.
* Finally, exploring the data can also help us develop a better understanding of the context in which the data was collected, which can help us interpret our results more accurately.

Therefore, we have tried to explore the Telco churn data using R and plotted the following charts:

Firstly, we wanted to find out how many customers decided to leave the company in the previous month.

Here we can see that 1869 customers out of the 7032 decided to leave in the previous month.

Chart, bar chart

Description automatically generated

We also plotted a histogram and boxplot of the monthly charges. We found that the distribution was extremely right skewed with the highest number of customer class is paying less than 25$ a month.

We can also see from the boxplot that the median monthly charges are around the 70$ mark and also there are a lot of outliers on both the lower and the higher side of the boxplot.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# Feature Selection

Feature selection is the process of choosing a subset of relevant features to use in model construction. There are several reasons why feature selection is important in the process of building a machine learning model. First, using a subset of relevant features can improve the performance of the model. This is because the model will be able to focus on the most important characteristics of the data, rather than being distracted by irrelevant or redundant features. This can lead to better predictions and improved accuracy.

Secondly, feature selection can help to reduce the complexity of the model. This is important because overly complex models can be difficult to interpret and can be more prone to overfitting. By selecting only, the most relevant features, the model can be simpler and more interpretable, which can be beneficial in a variety of applications.

Also, it can help prevent overfitting, which occurs when a model performs well on the training data but poorly on new, unseen data. This can happen when the model is too complex and has learned to make predictions based on random noise or irrelevant features in the training data.

In this project we have implemented the Chi Square test as our feature selection methodology.

The chi-square test is a statistical test that is often used in feature selection to evaluate the relationship between two categorical variables. It is often used in machine learning to assess the significance of a feature in predicting the outcome of a model.

In a chi-square test, the null hypothesis is that there is no relationship between the two variables, and the alternative hypothesis is that there is a relationship.

The chi-square test provides a p-value, which can be used to evaluate the strength of the relationship between the two variables. If the p-value is small, it indicates that the relationship between the two variables is significant, and the feature should be included in the model.

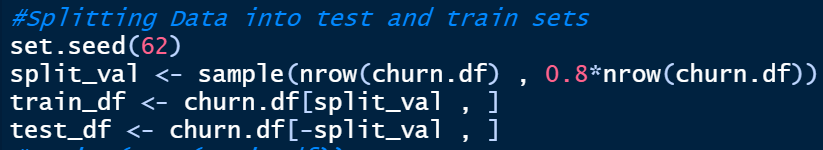
In the figure below, we can see that the p-value or the significance of all the columns except the Phone Service and gender are less than our predetermined threshold of 0.5. Hence, we decided to ignore those two columns to simplify our model.

Chart, bar chart

Description automatically generated

# BI Model

* Using the sample() method, the data was split into training data (80%) and testing data (20%).
* A machine learning model is trained using the train() function.
* *The repeated sampling technique is cv (repeated k-fold cross validation)*.



#### 5625 rows were allocated to the training set(train\_df) used to fit the machine learning model.

#### 1407 rows were allocated to the testing set(test\_df) used to evaluate the fit machine learning model.

* Utilizing the trainControl() function to manage the train() method's computation. The repeated sampling technique is cv (repeated k-fold cross validation). use three repeats and a 10-fold cross validation. Synthetic Minority Oversampling Technique is the sampling technique (SMOTE)
* The following Models were used in the study Logistic Regression, Random Forest, & KNN.

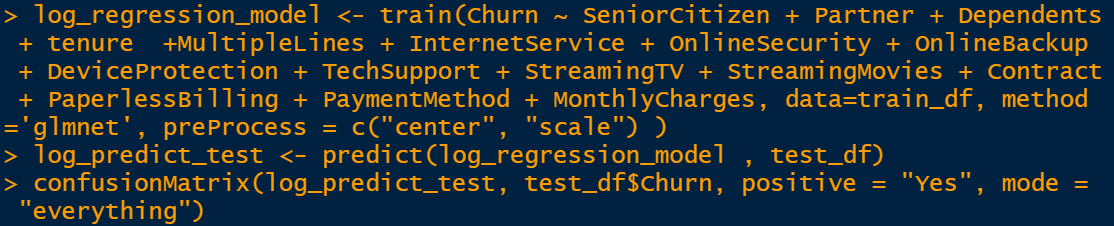
## Logistic Regression

Logistic regression is a statistical method for binary classification, which means it is used to predict the outcome of a binary dependent variable based on one or more independent variables. It is a type of regression analysis that is used to model the probability of a certain event occurring, such as whether a person will have a certain disease or not. In this case, since we want to predict whether a customer would be staying or leaving the telecommunications company, it is particularly useful.

In logistic regression, the dependent variable is always binary, meaning it can take only two values, such as 0 and 1 or "yes" and "no". The goal of logistic regression is to find the best-fitting model that describes the relationship between the dependent variable and the independent variables. This is done by estimating the probabilities that the dependent variable will take on each of its possible values, given the values of the independent variables.

In the case of logistic regression, the glmnet method can be used to fit a logistic regression model and predict the probability of a binary dependent variable based on one or more independent variables. It can also be used to determine the optimal threshold value for making predictions based on the probabilities output by the model.

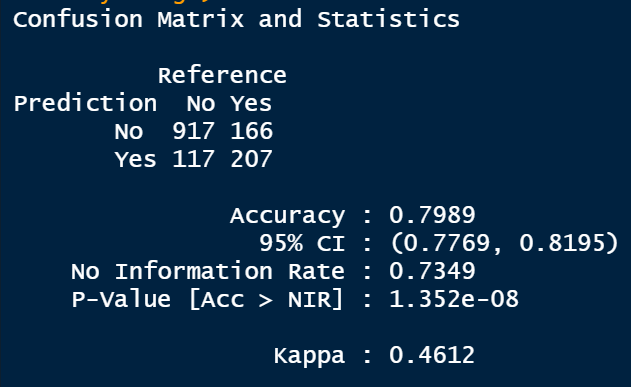
The accuracy of Logistic regression is highest among the three BI models that we implemented.



We also worked on understanding the underlying coefficients of the logistic regression model. The coefficients of a logistic regression model are used to represent the relationship between the independent variables and the dependent variable. The coefficients indicate the strength and direction of the relationship, and they can be used to make predictions about the dependent variable based on the values of the independent variables. A positive coefficient indicates that as the value of the independent variable increases, the probability of the dependent variable taking on a certain value also increases. On the other hand, a negative coefficient indicates that as the value of the independent variable increases, the probability of the dependent variable taking on a certain value decreases.



### Output

Text

Description automatically generated

Text

Description automatically generated

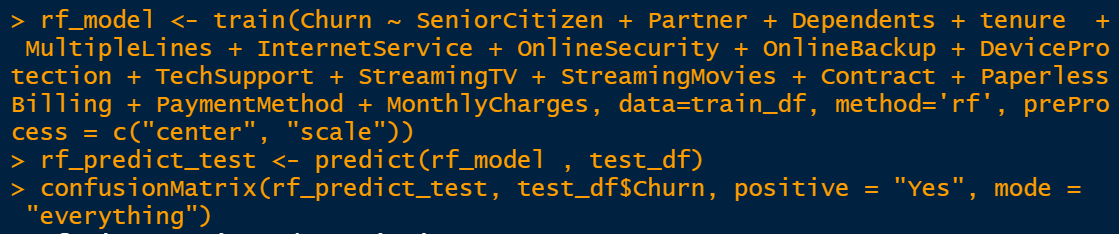
## Random Forest

Random forest is an ensemble learning method for classification and regression. It is a type of machine learning algorithm that builds multiple decision trees during training and combines their predictions to make a final prediction. Each decision tree in a random forest is trained on a different subset of the training data, using a different subset of the features, and using a different random seed. This means that each tree in the forest is a slightly different model, and they all make slightly different predictions. When making a prediction for a new data point, the random forest takes the average or majority vote of all the trees in the forest to make a final prediction.

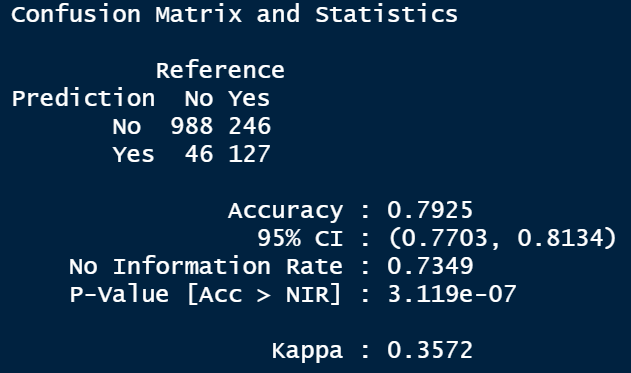
This approach of building multiple models and combining their predictions can often lead to more accurate and stable predictions than using a single model. Random forest is a powerful and widely used machine learning algorithm that can be applied to many different types of data.

Random forest is well-suited to predict customer churn because it can handle large and complex datasets, it can handle both numerical and categorical data, and it can make accurate predictions.

In the case of customer churn, the dataset may be large and may include many different types of information about the customers, such as their demographic information, their purchase history, and their interactions with the company. This information can be used to train a random forest model to predict which customers are likely to churn. Because random forest is an ensemble method, it can build multiple decision trees during training and combine their predictions to make a more accurate and stable final prediction. This can be particularly useful for predicting customer churn, where even small improvements in prediction accuracy can have a significant impact on the bottom line.



### Output

Text

Description automatically generated

## K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) is a machine learning algorithm that can be used for classification and regression tasks. In the context of predicting customer churn, KNN can be used to identify patterns and trends in customer behavior and characteristics that are associated with a higher likelihood of churn.

To use KNN for churn prediction, the algorithm first needs to be trained on a dataset of customer data that includes information on whether each customer has churned. The training data is used to identify the characteristics and behavior patterns that are most predictive of churn. Once the algorithm has been trained, it can then be applied to new data to predict whether a given customer is likely to churn.

It is predicated on the notion that the observations most "similar" to a given data point are those that are closest to it in the data set, allowing us to categorize unanticipated points based on the values of the existing points that are closest to them.

One advantage of using KNN for churn prediction is that it is a non-parametric method, which means that it makes no assumptions about the underlying distribution of the data. This makes it well-suited for data that may have complex or non-linear patterns. Additionally, KNN is a simple and easy-to-understand algorithm, which makes it a good choice for businesses that are new to using machine learning for churn prediction.

Text

Description automatically generated

### Output

Text

Description automatically generated

Text

Description automatically generated

# Model Evaluation

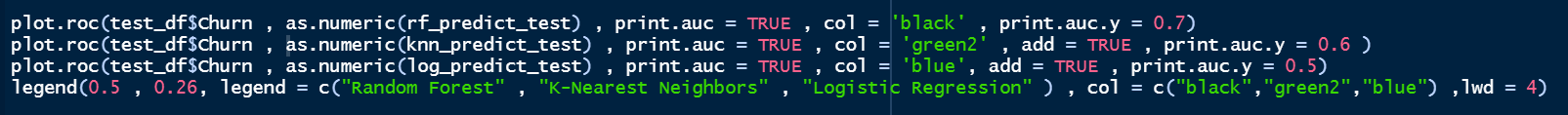
## Receiver Operating Characteristics (ROC Curve)

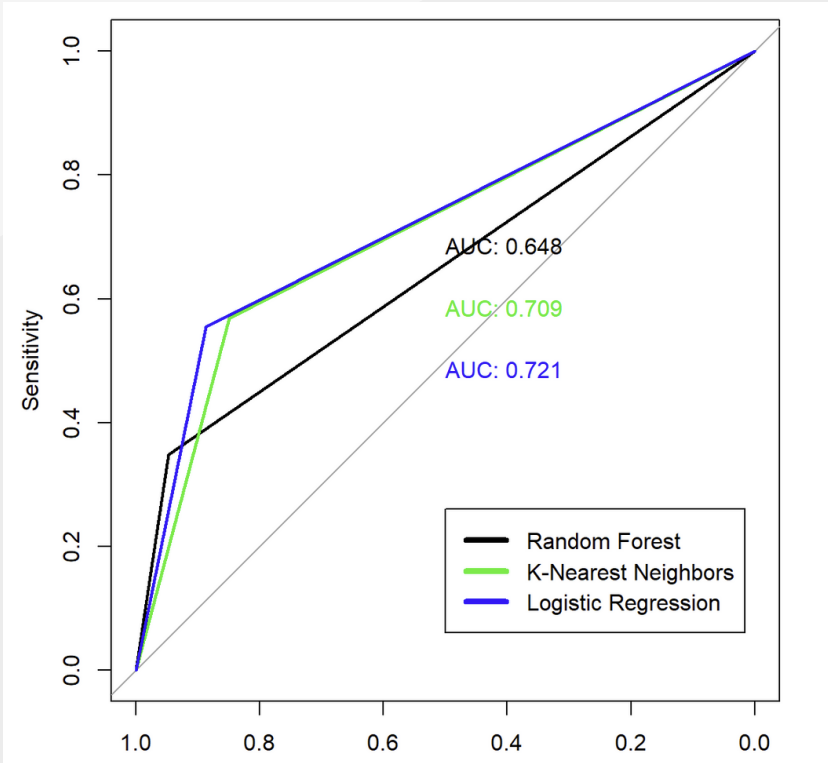
It is one of the most crucial evaluation criteria for assessing the effectiveness of any classification model. It is also spelled AUROC (Area Under the Receiver Operating Characteristics)

An indicator of performance for classification issues at different threshold levels is the AUC-ROC curve. AUC, or Area under the curve, signifies the level or measurement of separability, and ROC is a probability curve. It reveals how well the model can differ across classes. The higher the AUC, the more accurate the model is at classifying 0 classes as 0 and 1 class as 1. By analogy, the model is more effective at differentiating between patients with the condition and those who do not have it, the higher the AUC.

The true positive rate (TPR) is plotted against the false positive rate (FPR) on the ROC curve, with FPR on the x-axis and TPR on the y-axis.

* The Area under the curve measures how well the various groups of values can be distinguished. In contrast, the ROC curve represents the data's probability curve when it is plotted using the plot.roc() function.
* The greatest AUC value was for Logistic Regression (AUC = 0.721)





Hence, the best model for Telco Churn Prediction is Logistic Regression and the Area Under the Curve (AUC) is 72.1%.

# Conclusion

This kind of study in the telecom industry is to help businesses increase profits. Telecom businesses might have a clear perspective on the situation and offer enticing incentives to clients to keep using their services. The results show that by utilizing machine learning approaches, our suggested churn model performed better and produced better results.

We test the difference in means of churn and no churn customers by tenure, Monthly Charges and Total Charges. We also computed the correlation coefficient of tenure, Monthly Charges and Total Charges. Considering correlation values, Total Charges has a strong relationship with Monthly Charges and tenure. Hence, Total Charges is excluded when training model.

After analyzing customer churn in the telecommunications sector, we have found that several factors are associated with an increased likelihood of churn, such as being a younger customer, streaming movies, having multiple lines and the type of payment method. These are concluded using the coefficients of logistic regression. Our findings conclude that the best model of telco customer churn prediction is Logistic Regression due to it having the highest accuracy and favorable sensitivity and specificity.

Based on these findings, we can conclude that the telecommunications sector faces significant challenges in retaining customers, particularly younger customers, and those on lower-priced plans. To address these challenges, companies in this sector may need to focus on improving the customer experience, offering more competitive pricing and plans, and finding ways to engage and retain their most valuable customers.

# Future Enhancements

More columns can be removed to enhance interpretability. To further enhance performance, other machine learning models, such as deep learning and neural networks, can be used. Using regularization to avoid overfitting, more sophisticated optimization techniques, and nonlinear transformations of the input features are just a few ways to enhance logistic regression. Additionally, performance is frequently enhanced using ensembles of various logistic regression models.

# References

1. Abhishek and Ratnesh ,“Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques”, In the proceedings of 2017 International Conference on Advanced Computation and Telecommunication, Bhopal, India, 2017.

2. Abinash and Srinivasulu U ,“Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco”, In the proceedings of 2017 International Conference on Inventive Computing and Informatics , Coimbatore, India, pp. 721-725, 2017.

3. Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. "Machine learning amalgamation of Mathematics, Statistics and Electronics". International Research Journal on Advanced Science Hub, 2, 7, 2020, 100-108. doi: 10.47392/irjash.2020.72

4. Alae and El Hassane , “A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers”, In the proceedings of Intelligent Systems and Computer Vision , Fez, Morocco, 2017.

5. Salini Suresh; Suneetha V; Niharika Sinha; Sabyasachi Prusty; Sriranga H.A. "Machine Learning: An Intuitive Approach In Healthcare". International Research Journal on Advanced Science Hub, 2, 7, 2020, 67-74. doi: 10.47392/irjash.2020.67

# Project Presentation Link

<https://cometmail-my.sharepoint.com/:v:/g/personal/mxm220009_utdallas_edu/EUzCqOtB2ilAhVQYSwLtspIBnBelUWwyuwfh5aOhz4s4pA?e=qHaQ96>