**TELECOM CUSTOMER CHURN PREDICTION**

Contents

[**Introduction and Problem Identification** 3](#_Toc183869881)

[**Business Problem** 3](#_Toc183869882)

[**Significance of the Problem in the Telecom Industry** 3](#_Toc183869883)

[**Objective of the Analysis** 3](#_Toc183869884)

[**Stakeholders** 3](#_Toc183869885)

[**Exploratory Data Analysis** 4](#_Toc183869886)

[**Data Distribution & Relationships between Features** 4](#_Toc183869887)

[**Summary Statistics** 9](#_Toc183869888)

[**Dataset Balance For Target Variable** 9](#_Toc183869889)

[**Data Preprocessing** 10](#_Toc183869890)

[**Handling Missing Values** 10](#_Toc183869891)

[**Removing Irrelevant Columns** 10](#_Toc183869892)

[**Handling Inconsistencies** 10](#_Toc183869893)

[**Outliers Removal** 10](#_Toc183869894)

[**Dividing Data into Training & Testing Sets** 11](#_Toc183869895)

[**Project Methodology & Algorithms** 11](#_Toc183869896)

[**Appropriate Model Selection & Justification** 11](#_Toc183869897)

[**Hyperparameter Tuning of Random Forest** 11](#_Toc183869898)

[**Model Evaluation** 11](#_Toc183869899)

[**Features Importance** 14](#_Toc183869900)

[**Analysis and Findings** 15](#_Toc183869901)

[**Discussion and Insights** 16](#_Toc183869902)

[**Recommendations** 16](#_Toc183869903)

[**Limitations of the Model** 16](#_Toc183869904)

# **Introduction and Problem Identification**

## **Business Problem**

Companies in several industries, especially telecom, struggle with client churn. Customer churn in telecom is the rate at which subscribers cancel or move providers. This leads to lost revenue and high acquisition and retention costs as organizations must constantly recruit new clients to replace those that depart. Understanding and forecasting churn is crucial for companies to retain and expand customers. Customer retention is now more crucial than acquisition in the competitive telecom sector. Churn is a continuous issue due to consumer choice and cheap switching costs in many sectors. Companies that fail to foresee and reduce churn risk losing a considerable amount of their market share, especially if they engage heavily in client acquisition. Long-term cost savings, client loyalty, and profitability can result from reducing churn. Telecom firms can boost profitability and long-term growth by targeting at-risk clients, personalizing offers, and improving customer happiness.

## **Significance of the Problem in the Telecom Industry**

Due to high client acquisition expenses, telecom churn is enormous. Getting a new customer is more expensive than keeping one. Thus, limiting churn helps organizations preserve income and decreases the need for aggressive marketing to recruit new consumers. Telecom firms must identify churn reasons to manage customer satisfaction, service quality, and competitive pressures, as high churn rates generally indicate these concerns. Telecom businesses can respond before customers depart by predicting churn. Early identification of at-risk clients allows businesses to give targeted promotions, discounts, or upgraded services to keep them. Predictive models can also discover customer behavior patterns and trends to improve service quality, customer experience, and operational efficiency. In the telecom industry, churn prediction is crucial since minimizing churn can boost profitability and market position.

## **Objective of the Analysis**

This investigation focuses on building and testing machine learning models to forecast telecom customer attrition. The research uses data to identify consumers likely to churn, allowing the organization to take preventative measures. This requires fitting Logistic Regression, Decision Tree, and Random Forest machine learning models. These models will be trained on past customer data to forecast churn based on customer traits and behaviors. This investigation will compare the accuracy, interpretability, and performance of two popular categorization models. Logistic Regression is a basic but effective binary classification model for churn prediction. Random Forest, an ensemble approach, handles complex datasets with great accuracy. These models are compared to find the best telecom customer churn prediction model.

## **Stakeholders**

This analysis will benefit numerous firm departments that manage client retention and commercial strategy. The Marketing Department can use predictive analytics to create customized campaigns and specials to retain high-risk customers. Understanding which consumers are prone to churn allows marketing teams to target their efforts to individual wants and concerns, improving retention methods. The study can enable the Customer Service Department proactively contact at-risk customers and suggest solutions to boost happiness and loyalty. Understand turnover factors to focus customer care on areas most likely to cause dissatisfaction and resolve issues before they cause churn. Finally, the Strategy Department will use churn prediction models to inform company strategies. Understanding churn patterns and customer attrition reasons helps the organization improve its strategy, service, and customer loyalty and profitability.

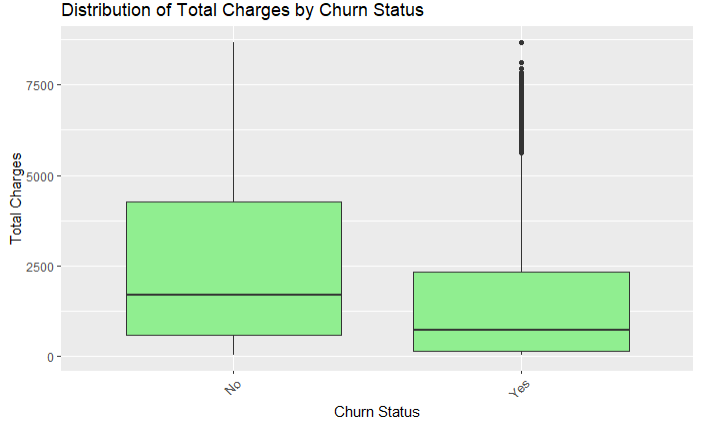
# **Exploratory Data Analysis**

## **Data Distribution & Relationships between Features**

The exploratory data analysis (EDA) was conducted to uncover patterns and trends in the data that may help us understand customer behavior and the factors that contribute to churn. Below is a detailed examination of several key findings from the analysis.

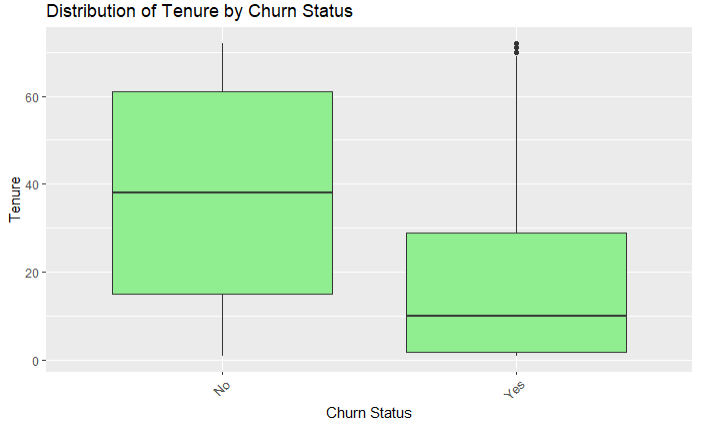
**Total Charges: Churners vs Non-Churners**

The data shows that churning customers pay less overall. This shows that long-term clients incur more charges. Churners may have shorter corporate ties, resulting in reduced charge accumulation. This suggests that customers who spend less may be more inclined to cancel their account owing to unhappiness or disengagement with the company's products.



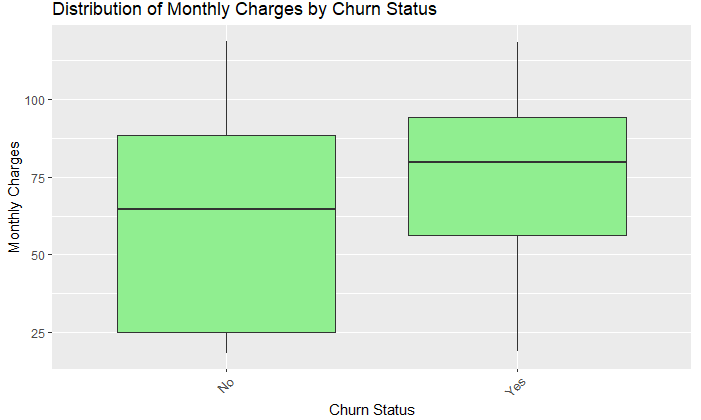
**Tenure: Churners vs Non-Churners**

Another interesting conclusion is that churners have a much shorter average tenure. Churners leave the company faster than keepers. This supports the idea that long-term consumers are less likely to leave, presumably because they know the company and are satisfied with the service. Turnover customers likely had a bad experience or troubles earlier in their tenure, therefore they left.



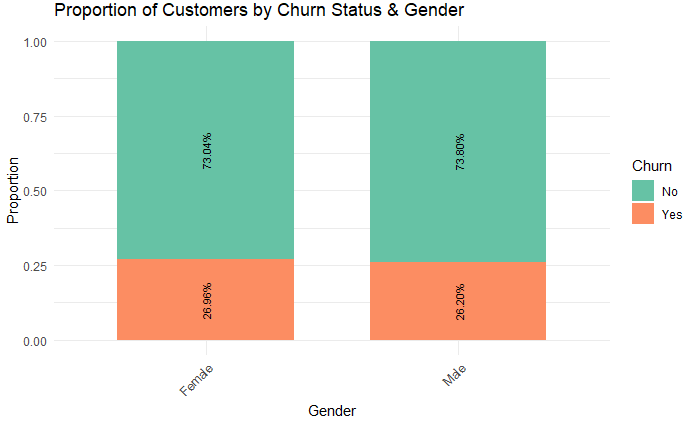
**Monthly Charges: Churners vs Non-Churners**

Churning consumers pay much more every month. This may sound paradoxical, but it shows that customers who pay more per month may be dissatisfied with the service they receive for the cost, which could increase turnover rates. It may also suggest that higher-paying consumers, such as those with premium plans, expect better service and depart faster if they don't.



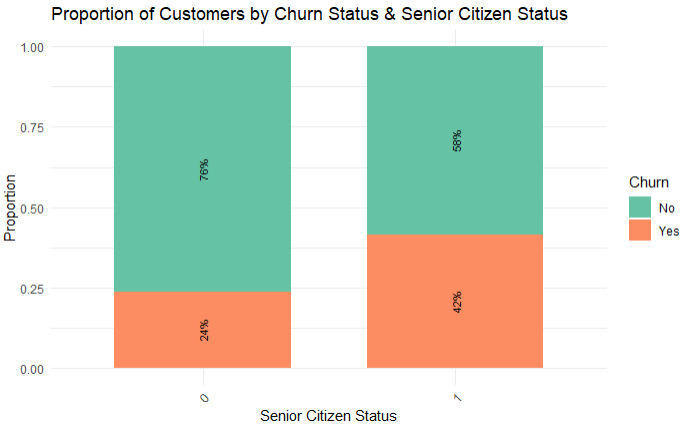
**Gender Distribution Among Churners**

Among the churned customer base, the gender distribution shows that 26.20% of churners are male, while 26.96% are female. This indicates a relatively balanced gender distribution among churners, with a slight difference in favor of females. The churn rate does not appear to vary significantly by gender, suggesting that other factors, such as service quality, contract type, or payment method, may play a more significant role in determining whether a customer churns.



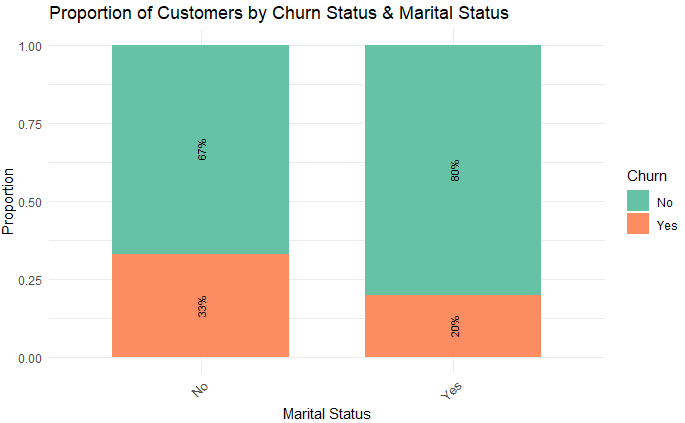
**Churn Among Senior Citizens**

Significantly, 42% of seniors churn, compared to 24% who do not. This is noteworthy and shows that seniors leave the service more often. Dissatisfaction with the service, shifting needs, or other personal reasons may cause this. It may also indicate that senior citizens are less loyal to the company and more price-sensitive, making them more prone to switch providers.



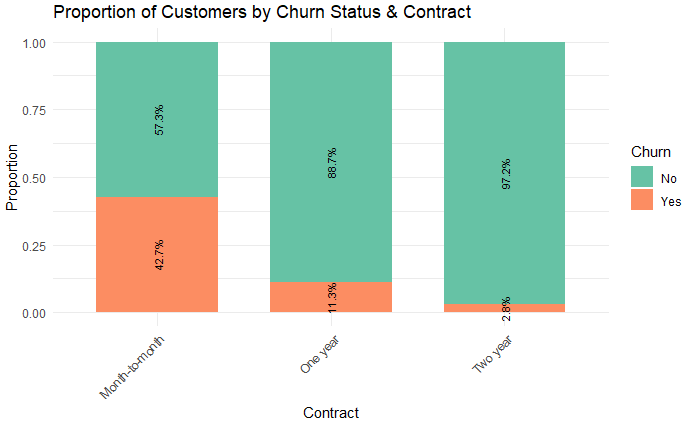
**Churn Among Married vs Non-Married Customers**

The attrition rate for married clients is 20% and non-married customers 33%. This suggests that married consumers are less likely to churn due to family plans, a propensity for stability, or more bundled services that strengthen the company-customer relationship. Non-married consumers, who may have more individual plans, churn more often, maybe due to less commitment or service dissatisfaction.



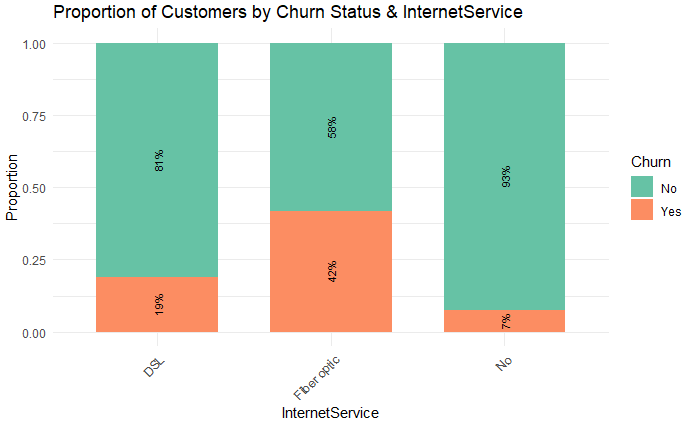
**Churn Rate by Contract Type**

Contract type greatly affects customer churn. The lowest turnover rate is 2.8% for two-year contracts. Longer-term contracts may keep clients with the company. A month-to-month contract has the greatest churn rate at 42.7%, while a one-year contract has 11.3%. Short-term contracts or no contract (month-to-month) are related with increased churn, indicating that customers without long-term commitments are more likely to transfer providers.



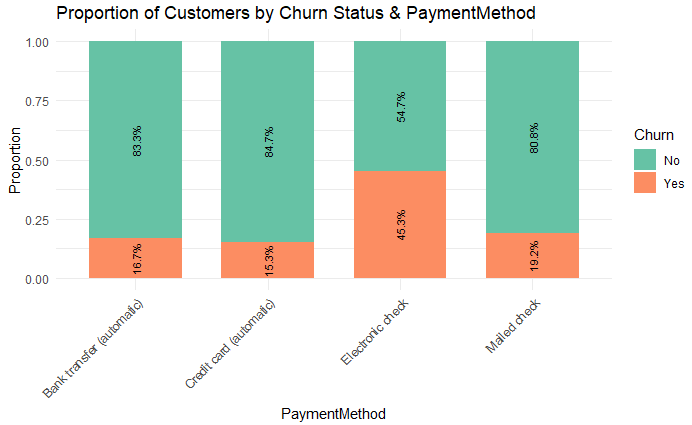
**Churn by Internet Service Type**

Internet service type affects churn. Only 19% of DSL consumers churn, while 42% of fiber optic subscribers do. This suggests that while fiber optic delivers higher speeds and maybe better service, cost, installation, and service quality may cause a higher churn rate. DSL, however slower, may be more stable or cheap for many customers, reducing turnover.



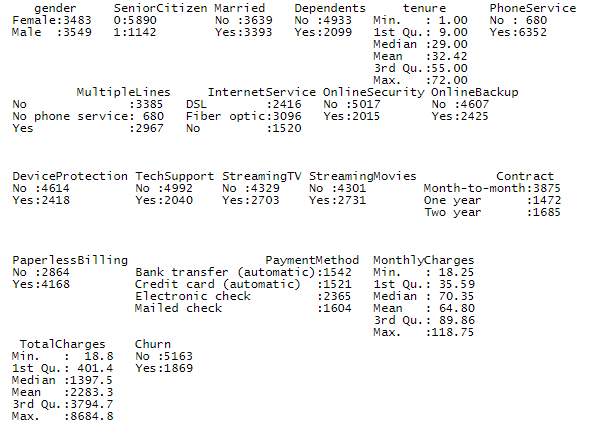
**Churn by Payment Method**

Payment method also affects churn. At 45.3%, electronic cheque consumers have the greatest churn rate, suggesting they are less engaged with the company. Mail-in check clients have a 19.2% turnover rate, credit card automatic payment consumers 15.3%, and bank transfer customers 16.7%. These data imply that customers who use automatic and convenient payment methods like credit cards and bank transfers are less likely to churn due to a higher commitment and convenience.



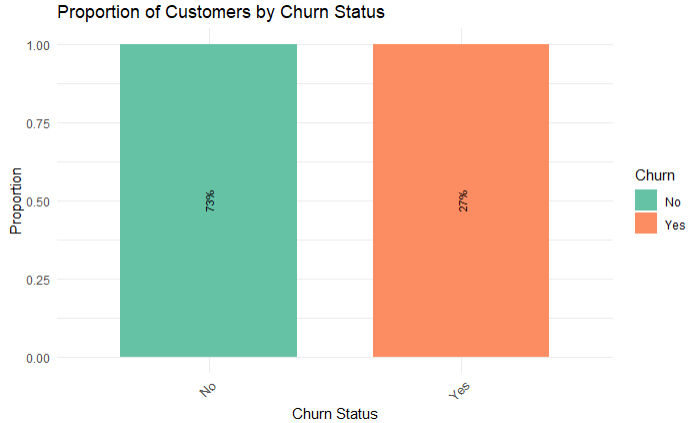
## **Summary Statistics**

The dataset consists of demographic and service-related information about customers. It shows that the number of male and female customers is nearly equal, with 3,483 females and 3,549 males. Most customers are not senior citizens (5,890 out of 7,032), with the median tenure being 29 months and a mean of 32.42 months, indicating that most customers have been with the service for several years. A majority of customers (63.3%) have phone service, while only 9.7% have no phone service. Internet services are predominantly provided via fiber optic (44.0%) and DSL (34.4%), with a notable portion of customers opting for paperless billing (59.2%). In terms of contract types, most customers are on a month-to-month plan (55.1%), with smaller proportions on one-year (20.9%) and two-year (24.0%) contracts. The dataset also includes payment method details, with a significant number of customers using electronic checks (33.6%) or bank transfers (21.9%). Monthly charges have a wide range, with the median at $70.35 and the mean at $64.80, indicating varied pricing based on the services chosen. The churn rate is about 26.5%, with 1,869 customers opting to cancel their service, while the majority (73.5%) remain subscribed. Total charges range from $18.80 to $8,684.80, with a median of $1,397.50.



## **Dataset Balance For Target Variable**

Overall, 1,869 customers have churned, while 5,163 have not. This suggests that 27% of customers have churned, a high rate. Strategies to retain consumers and reduce attrition need understanding the causes of this turnover. The high proportion of non-churners provides a robust client base for retention efforts.



# **Data Preprocessing**

## **Handling Missing Values**

To find NA values, use colSums(is.na(data)) to check for missing values. Na.omit(data) removes missing rows. This assures that the dataset only contains full cases without missing values.

## **Removing Irrelevant Columns**

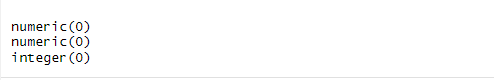
The customerID column is removed from the dataset. This column is likely a customer-specific identifier and not needed for modeling. The code finds all character columns in the dataset and outputs their unique values. To comprehend the categories and avoid discrepancies and surprises.

## **Handling Inconsistencies**

Character variables are transformed to replace "No internet service" with "No". This data cleansing procedure standardizes modeling values. Using as.factor(x), all character columns become factor variables. Categorical data must be quantitatively represented for machine learning models. Factor variables are common in R for categorical data. The binary SeniorCitizen column is specifically turned into a factor variable. This makes it a categorical variable in modeling rather than a continuous numeric characteristic.

## **Outliers Removal**

The outliers are checked in the dataset using IQR method and the results thus obtained are attached below. The IQR method labels all those observations as outliers if the value is either above the upper bound or below the lower bound. The IQR results shows that there are no outliers in either the monthly charges column or total charges column. Additionally, there are no outliers in the tenure column as well. Hence, the data is cleaned.



## **Dividing Data into Training & Testing Sets**

The data is divided into training and test sets. 70% of the data (trainIndex) is utilized for training and 30% for testing. This split lets you test the model on unknown data, assuring generalizability. These pretreatment steps clean the dataset, handle categorical variables, and prepare it for machine learning modeling.

# **Project Methodology & Algorithms**

## **Appropriate Model Selection & Justification**

**Logistic Regression**

Logistic Regression was used in this investigation because of its simplicity, interpretability, and success in binary classification problems like churn prediction. It predicts customer attrition based on their qualities as a linear model. It lets us determine how each predictor variable affects customer attrition. Logistic Regression, a simple model, can set the stage for more complicated models and reveal customer attrition trends. It is also computationally efficient and less likely to overfit than complex models.

**Random Forest**

Random Forest, a more advanced machine learning model, predicted churn. This ensemble approach constructs numerous decision trees and aggregates their predictions to improve accuracy and reduce overfitting. Data on customer churn is difficult, yet Random Forest can model non-linear relationships and handle many input factors. It helps select the most essential churn prediction features when relationships between features are unclear. Random Forest is more versatile and robust than Logistic Regression for high-dimensional datasets and complex variable interactions. However, it is harder to interpret than Logistic Regression. However, its capacity to understand complicated data patterns and manage noisy data makes it suited for this purpose.

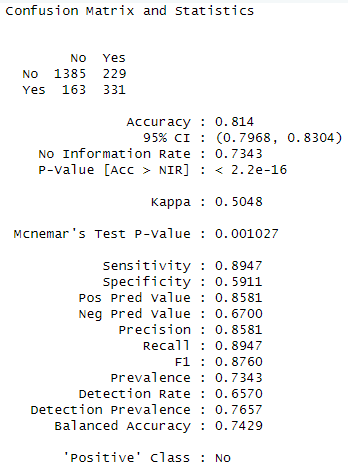
## **Hyperparameter Tuning of Random Forest**

We optimized the Random Forest model via hyperparameter adjustment. The mtry parameter, which regulates the amount of variables randomly sampled at each decision tree split, was modified. Choosing the right mtry value improves model performance and prevents overfitting. The R caret package simplifies model training and hyperparameter adjustment. We cross-validated a grid of mtry values from 2 to 16. This method trains the Random Forest model with the optimal number of variables at each split, resulting in a more balanced model. We examined each mtry value to find the one that minimizes model error. The caret package automatically cross-validates to get the optimum hyperparameter configuration. Logistic Regression and Random Forest models were trained on 70% of the data. The Logistic Regression model was trained directly, while the Random Forest model was hyperparameter tuned and the best mtry value was chosen using cross-validation.

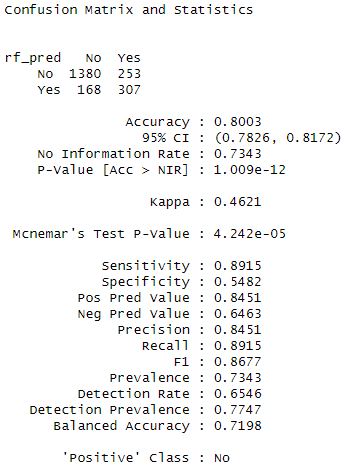
# **Model Evaluation**

In this section, we evaluate the performance of the three models used for predicting customer churn: Logistic Regression, Random Forest, and Tuned Random Forest. The evaluation metrics include accuracy, precision, recall, F1 score, and the confusion matrix, which provide insight into how well each model distinguishes between churners and non-churners. Below is a detailed analysis of the results for each model.

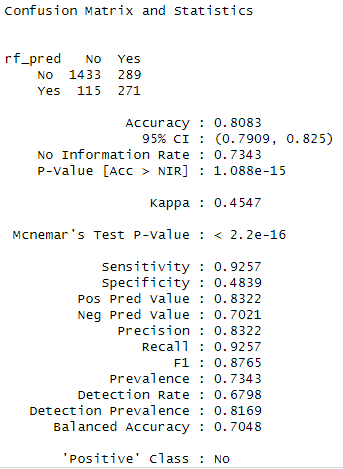
The Logistic Regression model correctly predicts churn or non-churn for 81.4% of the customers. Precision measures the proportion of true positive predictions (non-churners) out of all predicted positive instances. This means that 85.81% of the instances predicted as "No churn" are actually correct. Recall measures the proportion of true positives (non-churners) that were correctly identified. 89.47% of actual non-churners were correctly predicted by the model. The F1 score is the harmonic mean of precision and recall, providing a balance between the two. A value of 0.8760 indicates a strong model performance in terms of balancing precision and recall. Specificity is the proportion of true negatives (churners) correctly identified. The model identifies 59.11% of the churners correctly.



The Random Forest model achieves an accuracy of 80.03%, slightly lower than the Logistic Regression model, indicating that it correctly predicts 80.03% of the time. Precision measures how many of the instances predicted as "No churn" are correct. The precision of 84.51% is slightly lower than that of Logistic Regression, suggesting more false positives. The Random Forest model correctly identifies 89.15% of non-churners, which is similar to the recall of the Logistic Regression model. The F1 score of 0.8677, while strong, is slightly lower than the Logistic Regression model, indicating that Random Forest sacrifices a small amount of balance between precision and recall.

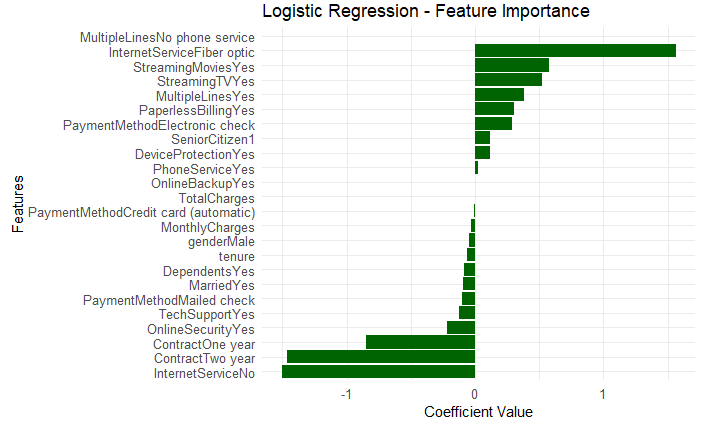


The Tuned Random Forest model shows a slight improvement in accuracy (80.83%) compared to the non-tuned Random Forest model. Precision is slightly lower than the non-tuned Random Forest model, indicating a higher number of false positives. The recall of 92.57% is notably higher than both the Logistic Regression and non-tuned Random Forest models. This suggests that the tuned model is very good at identifying non-churners. The F1 score of 0.8765 is similar to that of the Logistic Regression model, indicating a good balance between precision and recall.

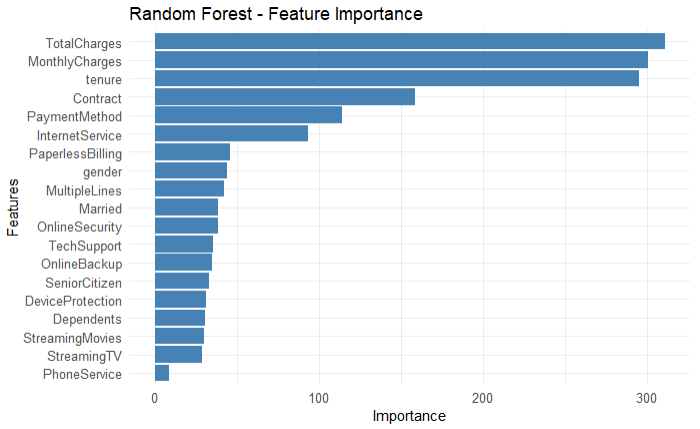


## **Features Importance**

The logistic regression model features importance shows that the multiple lines, internet service fiber optic and steaming movies status of yes are the most important features in predicting the churn.



The features importance plot for the random forest model is attached below. The most important features are total charges, monthly charges, tenure, and contract and payment method. These are the top 5 most important features in predicting customer churn.



The Logistic Regression model has the best precision and recall balance, but the Tuned Random Forest model has better recall and identifies more non-churners. However, the Tuned Random Forest model lacks specificity, making it less successful in churner detection. Random Forest performs well but has lesser specificity than Logistic Regression. Logistic Regression is a good churn prediction model since it balances precision and recall. Tuned Random Forest can be useful when the main goal is to limit the amount of non-churners misclassified as churners, but further modification may improve specificity.

# **Analysis and Findings**

**Summary of Analysis and Findings**

Logistic Regression had the highest precision and recall, 81.4%. This model balanced churner and non-churner identification best. Random Forest performed similarly with 80.03% accuracy, although it had lower churner prediction specificity than Logistic Regression. Despite losing specificity, the Tuned Random Forest model improved recall (92.57%) and non-churner detection. Overall accuracy was marginally higher than the non-tuned Random Forest model. Through feature analysis, several key factors were found to significantly impact customer churn:

* Senior Citizen Status: Senior citizens were more likely to churn, with higher recall rates indicating that age might be a crucial predictor for customer retention.
* Internet Service and Customer Support: Customers who did not have internet service or were dissatisfied with customer support were more likely to churn. These variables likely contribute to the dissatisfaction leading to churn.
* Contract Type and Payment Method: Customers with month-to-month contracts and those using paperless billing were found to have a higher churn rate. This indicates that customers with more flexible or non-committed service plans might be more likely to leave.
* Usage and Account Features: High usage of services such as additional phone lines or premium services also correlates with churn. Customers who did not engage with or use certain account features (like streaming services) were more likely to churn.

The Logistic Regression model provides a better balance between precision and recall, which is important in real-world applications where both false positives (incorrectly predicting no churn) and false negatives (failing to predict churn) must be minimized. Tuned Random Forest excelled at detecting non-churners (high recall), but its lower specificity indicated that it was more prone to false positives, incorrectly predicting churn when the customer was actually retained. The analysis identified several key predictors that influence churn likelihood, including:

* Contract Type: Month-to-month contracts were associated with higher churn rates, as customers had no long-term commitment to the service.
* Service Usage: Customers using multiple services or those who had more advanced service options (like streaming) were less likely to churn.
* Billing and Payment Methods: Customers who had digital billing options were more likely to remain, while those with paper billing showed higher churn rates.
* Customer Support Satisfaction: Dissatisfaction with customer support was strongly associated with a higher likelihood of churn, emphasizing the importance of maintaining good customer relationships.

# **Discussion and Insights**

## **Recommendations**

Based on the analysis and key predictors of churn, here are several strategies the company can implement to address and reduce customer churn:

* Since month-to-month contracts were strongly associated with higher churn rates, the company could incentivize customers to switch to one or two-year contracts by offering discounts or additional benefits (e.g., price freezes or bonus services). This would increase customer retention by locking in customers for a longer period.
* Introduce loyalty programs that reward customers for staying longer. For example, customers who renew for a year or more could receive benefits such as discounted rates, free upgrades, or priority customer service.
* Improving customer service satisfaction is key, as dissatisfaction in this area was a significant driver of churn. The company should focus on reducing wait times, improving response times, and offering more personalized support. Providing multiple communication channels (e.g., phone, live chat, email) and implementing proactive customer support solutions (such as AI-driven chatbots) can also improve the customer experience.
* Customers who are not fully utilizing their current service package may be at risk of churn. The company could analyze usage patterns and offer customized packages that match their needs more closely (e.g., more data, additional services, or better support for seniors or specific demographics).
* Using the insights from the analysis, the company can target high-risk customers with retention campaigns. For example, senior citizens or customers with month-to-month contracts could receive special offers or reminders about the benefits of longer-term contracts or service upgrades.
* Tailor communication to customers based on their usage patterns. For example, a customer who has been using a basic plan but has started showing signs of service dissatisfaction might be offered an upgraded plan with additional perks.

## **Limitations of the Model**

While the models developed in this analysis provide valuable insights into customer churn, there are several limitations that should be acknowledged:

* Missing or inaccurate data: Even though missing values were handled by removing incomplete records, this can still introduce bias if the missing data is not random. For example, if high-value customers were more likely to have missing data, the model might be less accurate for those segments.
* Limited features: The dataset used in this analysis may not include all relevant factors affecting churn. For instance, external factors such as market conditions, competitor offerings, or broader economic trends were not considered, which could affect the results.
* Imbalance in class distribution: The dataset had a higher number of non-churners compared to churners. While the model achieved good performance, the imbalance in the data might have led to biased predictions, especially in terms of sensitivity (false negatives). The models might be more accurate at predicting the majority class (non-churners) but less effective at detecting churners.