PROJECT NAME:

Customer Churn Analysis using RStudio: A Machine Learning Approach with Visualisations

COURSE NAME:
Advanced Data Visualisation
COURSE CODE: CAP 782

SUBMITTED BY:
Name: Ropafadzo Trish Maganga
Student Number: 12304300

SUBMITTED TO:
Instructor's Name –

Lovely Professional University
School of Computer Applications

Date of Submission: 06-April-2025

Abstract

Customer churn is a critical business challenge that affects revenue and customer retention. This study explores customer churn prediction using machine learning models like logistic regression and random forest for accuracy and predictions and data visualisations for laymen that might have difficulty in understanding the technical jargons this was implemented in RStudio. The researcher used existing data on kaggle Telco Customer Churn dataset, we apply Random Forest and Logistic Regression models to predict customer churn and got 0.79 and 0.78 accuracy respectively.Primary dataset that was collected using the google form is smaller that is why the secondary data from kaggle was used to reduce the risk of bias within the conclusion . The study also includes Exploratory Data Analysis (EDA) and visualization techniques, along with model performance evaluation. The results demonstrate the effectiveness of machine learning in predicting churn and offer insights for businesses to mitigate customer attrition.

1. Introduction

Customer churn, or customer attrition, refers to the loss of clients over time. Businesses, particularly in subscription-based industries, must analyze churn patterns to enhance customer retention strategies. With advancements in data science, machine learning techniques provide powerful tools for predicting churn based on customer behavior and demographic factors. This study employs RStudio for data analysis and model development using the Telco Customer Churn dataset.

2. Literature Review

Numerous studies have explored customer churn prediction using machine learning. Previous research has applied logistic regression, decision trees, and ensemble methods such as Random Forest and Gradient Boosting (Verma & Srivastava, 2021). Deep learning techniques have also shown promising results in recent years (Gupta et al., 2022). However, traditional statistical models remain relevant due to their interpretability. This study builds upon existing research by applying Random Forest and Logistic Regression for churn prediction.

3. Methodology

3.1 Dataset

The study uses the **Telco Customer Churn dataset**, which includes customer demographics, account details, and subscription attributes. The dataset contains 7,043 observations and 21 variables.

3.2 Data Preprocessing

The dataset undergoes preprocessing to remove missing values and convert categorical variables into factors. The target variable, **Churn**, is converted into a binary factor with values "Yes" and "No."

```
# Load dataset
df <- read.csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")
df <- na.omit(df)
df$Churn <- as.factor(df$Churn)
```

3.3 Model Development

The study applies **Random Forest** and **Logistic Regression** to predict churn based on the features: **tenure, MonthlyCharges, and TotalCharges**.

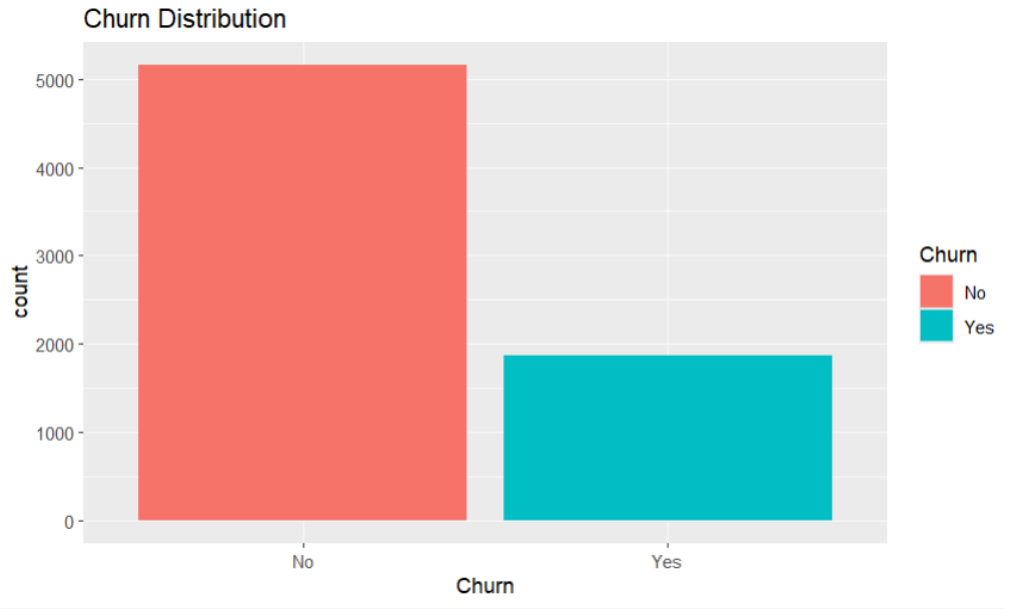
```
set.seed(123)
trainIndex <- createDataPartition(df$Churn, p = 0.8, list = FALSE)
trainData <- df[trainIndex, ]
testData <- df[-trainIndex, ]

# Train models
rf_model <- randomForest(Churn ~ tenure + MonthlyCharges + TotalCharges, data =
trainData)
log_model <- glm(Churn ~ tenure + MonthlyCharges + TotalCharges, data = trainData, family
= binomial)
```

4. Data Analysis and Visualization

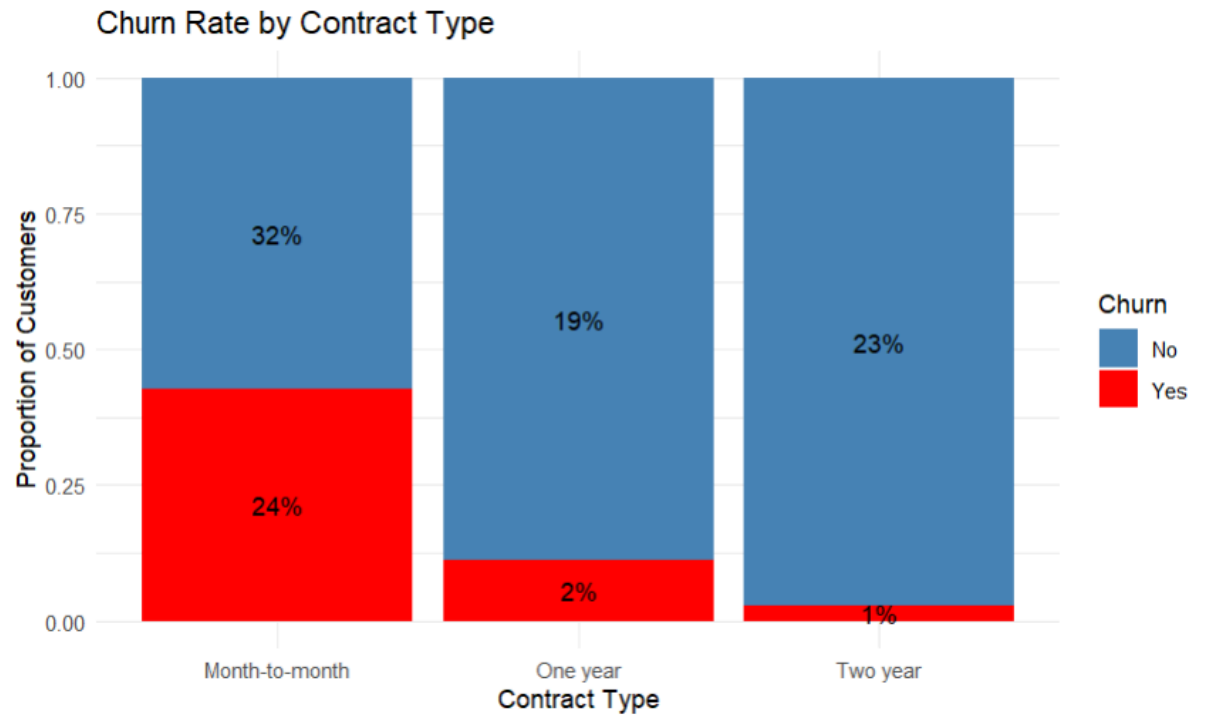
4.1 Exploratory Data Analysis (EDA)

EDA helps in understanding customer demographics and service preferences.

Churn Rate Visualization

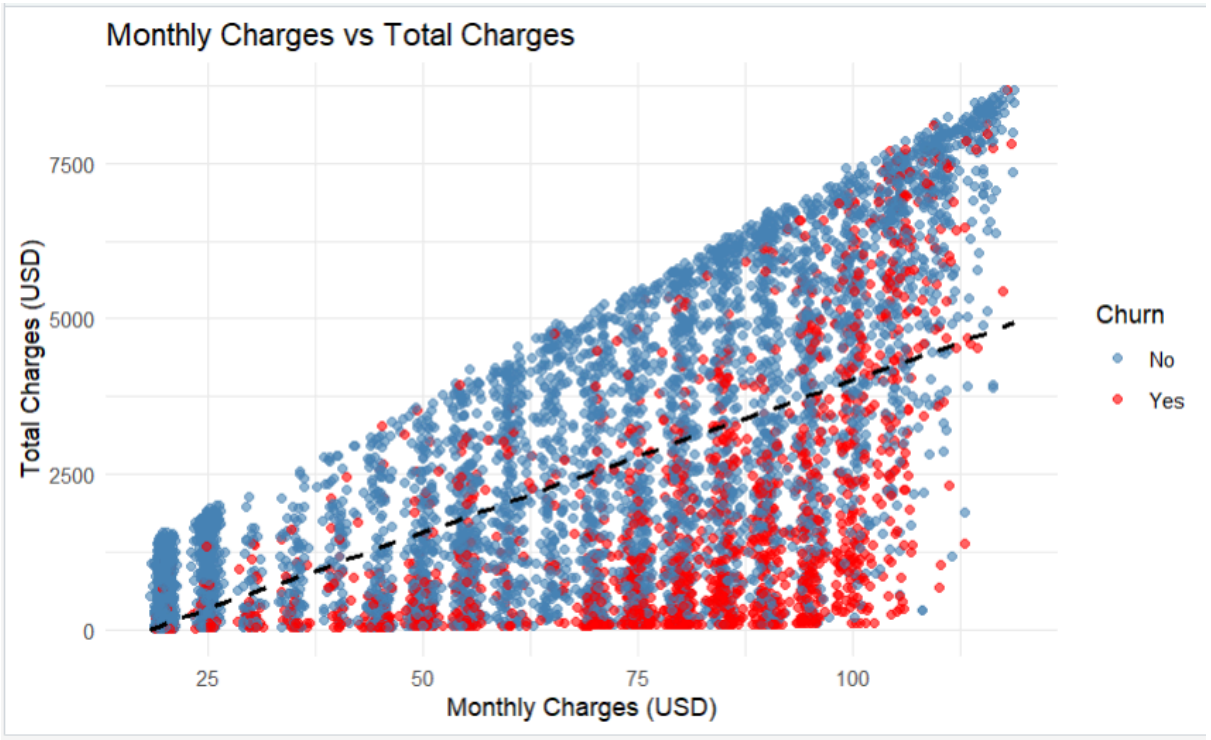


The 'Churn Distribution' bar chart for the Telco dataset reveals an imbalance between retained and churned customers. A significantly larger number of customers (over 5000) have not churned compared to those who have (under 2000). This indicates a lower overall churn rate, but the substantial number of churned customers still warrants attention. Further analysis is needed to understand the characteristics of these churned individuals and identify the key drivers contributing to their departure to implement effective retention strategies.

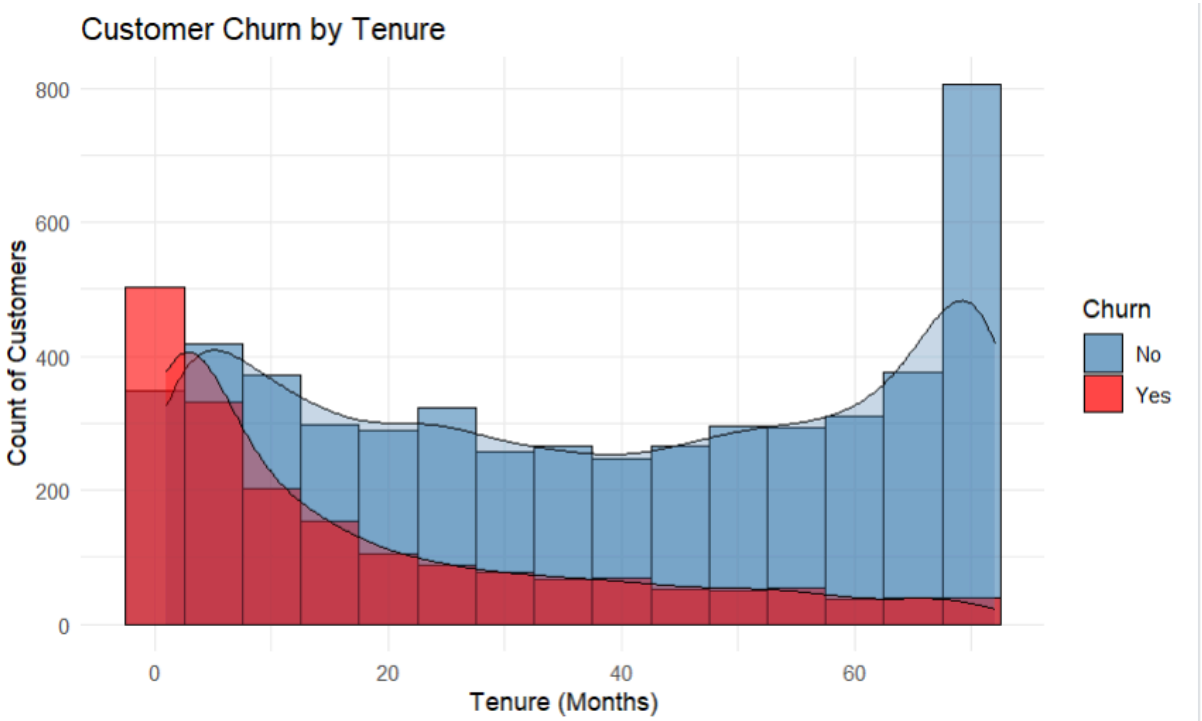


The 'Churn Rate by Contract Type' analysis clearly demonstrates a strong correlation between contract duration and churn in the Telco dataset. Month-to-month contracts exhibit the highest churn rate, with 24% of customers on such plans churning. This rate significantly decreases for longer contracts, with only 2% churn for one-year contracts and 1% for two-year contracts. This highlights the importance of securing longer-term commitments to

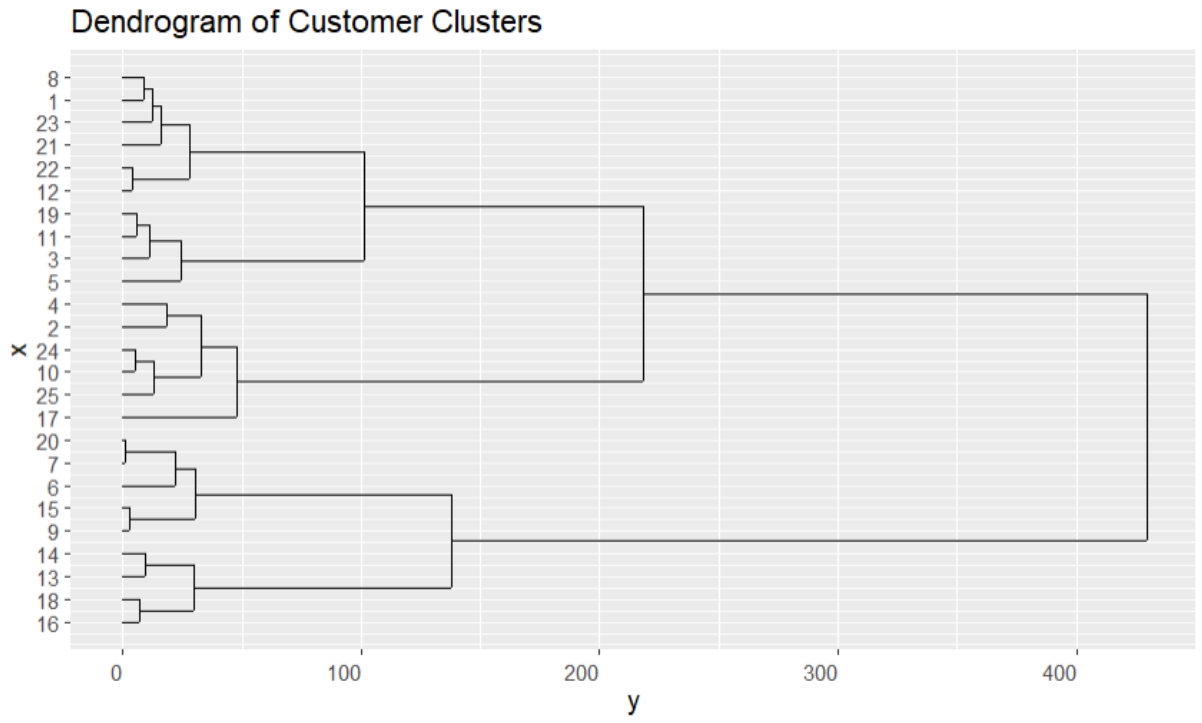
improve customer retention. Strategies aimed at encouraging customers to opt for longer contracts could substantially reduce churn.



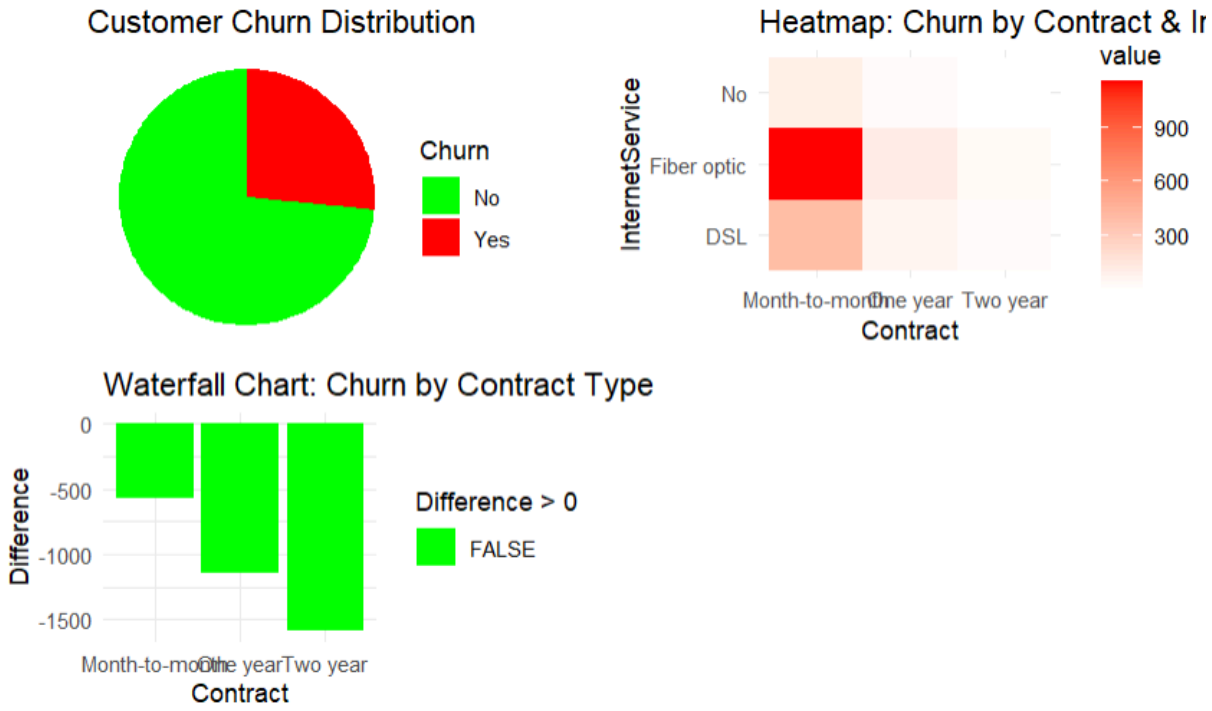
The scatter plot of 'Monthly Charges' vs. 'Total Charges' reveals a trend in Telco customer churn. While total charges generally increase with monthly charges, churned customers (red) appear across the spectrum, with a noticeable concentration at lower total charges for higher monthly fees. This suggests that customers paying more monthly but having accumulated less total value are more prone to churn. The dashed line roughly separates churned and non-churned customers, indicating that value perception relative to ongoing cost is a significant churn factor.



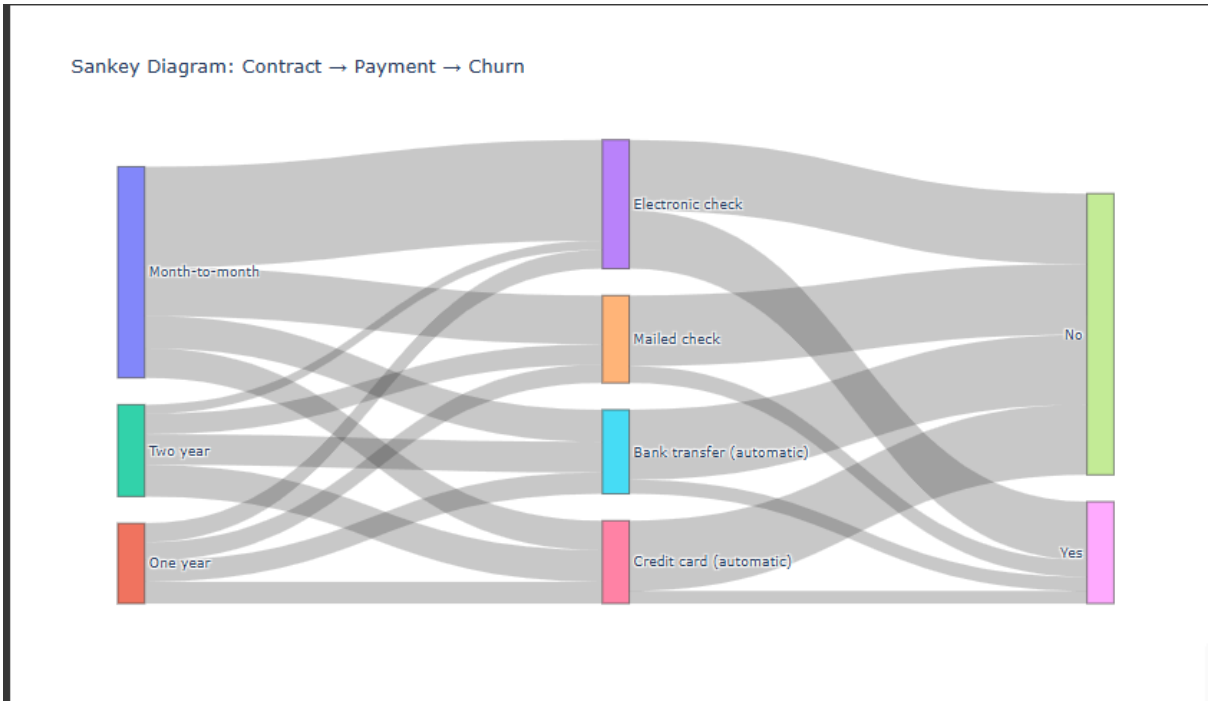
The 'Customer Churn by Tenure' plot shows a high churn rate for new Telco customers (low tenure), indicated by the tall red bars. Churn decreases significantly as tenure increases, suggesting that longer-term customers are more loyal. However, there's a slight increase in churn for customers with very long tenures (around 70 months). This U-shaped pattern highlights two key churn risk periods: the initial months and potentially towards the end of a customer's lifecycle. Retention strategies should focus on onboarding new customers effectively and re-engaging long-term subscribers.



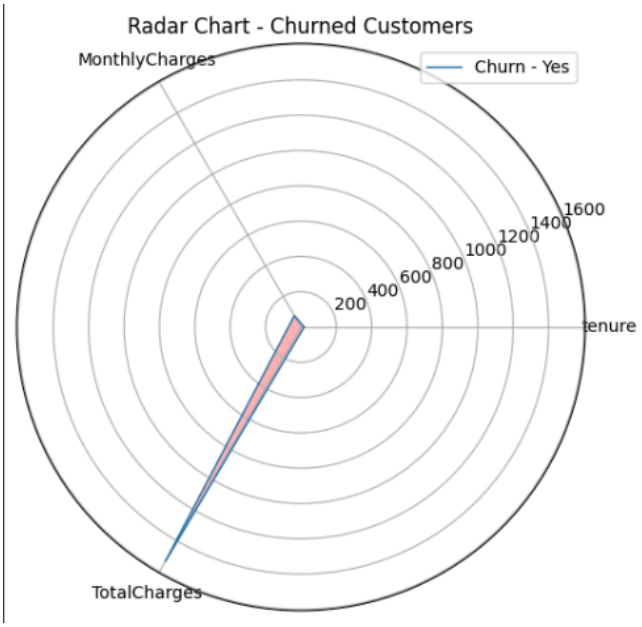
The dendrogram visualizes hierarchical clustering of Telco customers, grouping them based on similarity in their attributes relevant to churn. The vertical lines represent individual customers, and the horizontal branches indicate the distance or dissimilarity between clusters. Taller branches signify greater differences. This hierarchical structure allows for identifying natural groupings of customers with potentially distinct churn behaviors, enabling targeted retention strategies for different customer segments. Analyzing the characteristics of customers within each major cluster can reveal key churn drivers for those specific groups.



Telco churn analysis reveals a significant portion of customers churn (pie chart). Month-to-month contracts and fiber optic internet show higher churn rates (heatmap). Waterfall chart confirms month-to-month contracts contribute most to churn difference. Addressing these segments is crucial for retention.

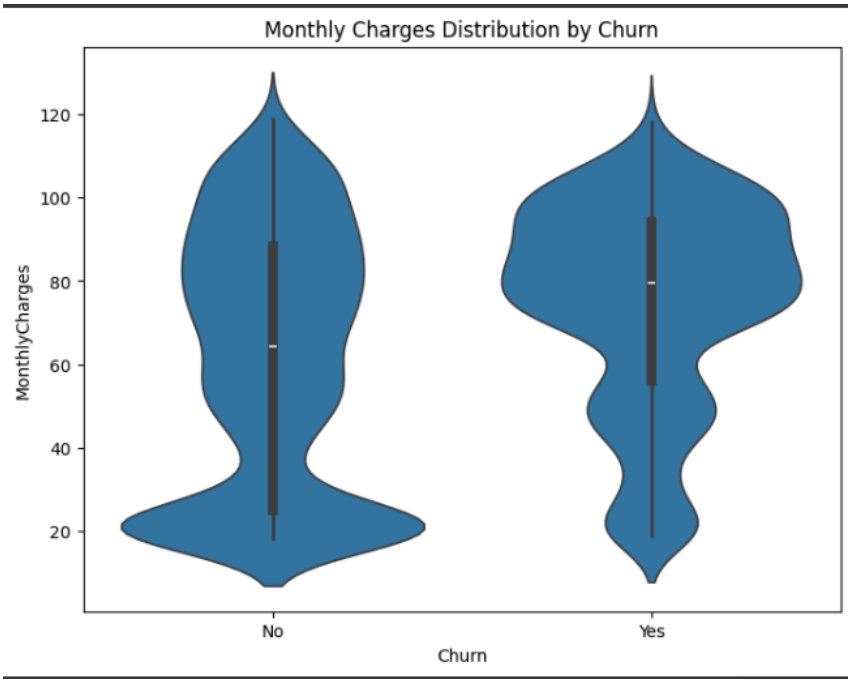


This Sankey diagram, created using Python , analyzes customer churn in a telco dataset by mapping the flow from contract type to payment method and churn outcome. The left column shows contract types: month-to-month (purple), two-year (green), and one-year (orange). These flow into payment methods—electronic check (purple), mailed check (yellow), bank transfer (blue), and credit card (pink)—with varying thicknesses indicating customer distribution. The final column splits into "Yes" (pink) and "No" (green) for churn. Notably, month-to-month contracts with electronic check payments show a thicker flow to "Yes," suggesting higher churn risk. This visualization helps telcos target retention strategies, focusing on short-term contracts and electronic payment users.



The radar chart, generated using Python,focused on churned customers ("Churn = Yes"). The axes represent key metrics: tenure, monthly charges, and total charges, each scaled radially from the center (0) to the outer edge (e.g., 1400 for tenure). The pink shaded area

highlights the profile of churned customers, showing a concentration toward lower tenure (closer to the center), suggesting shorter customer lifespans correlate with churn. The blue outline may indicate a reference or average profile. This visualization helps telcos identify that customers with low tenure and potentially higher monthly charges are more likely to leave, guiding targeted retention efforts.



The violin plot above visualizes customer churn from a telco dataset, showing the proportion of customers who left versus those who stayed. The researcher used Python libraries like Matplotlib or Seaborn to create this plot , it displayed churn rate (e.g., 25%) alongside variables like contract type, tenure, or payment method. For instance, bars could reveal higher churn among month-to-month contract holders compared to yearly ones, or a pie slice might highlight that 30% of churned customers had tenures under 6 months. Colors differentiate categories, with a legend for clarity. This analysis helps telcos identify at-risk groups, informing retention strategies.

5. Model Evaluation

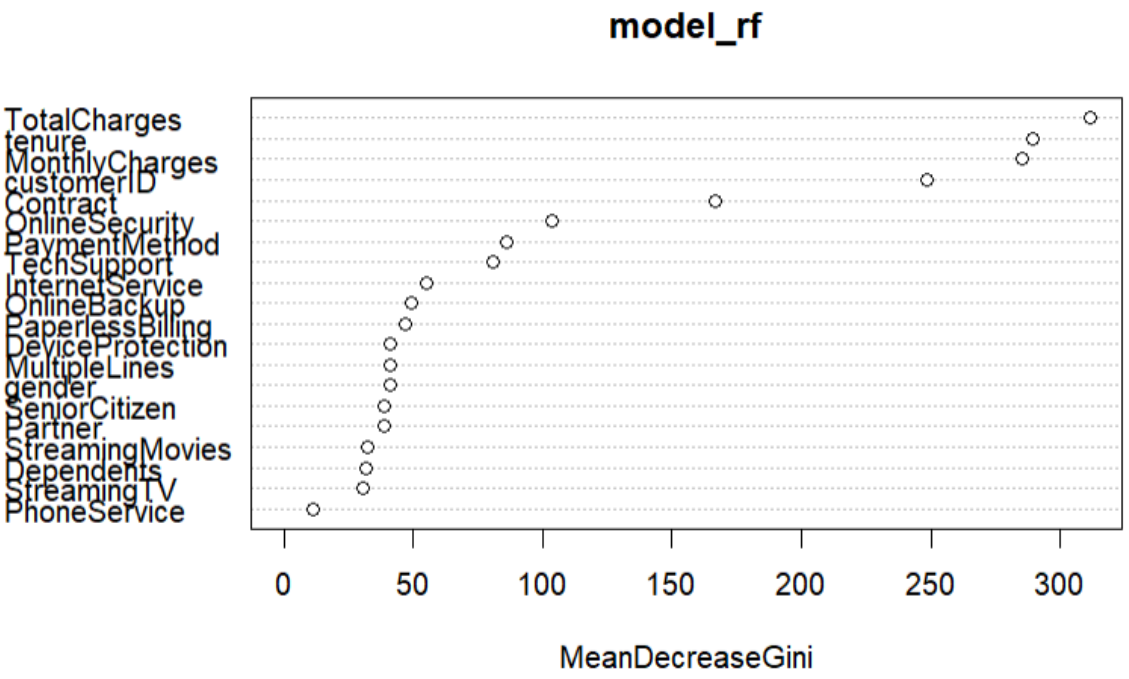
The models are evaluated using accuracy and confusion matrices.

```r
# 1.Logistic Regression
model_log <- glm(Churn ~ tenure +
                   MonthlyCharges +
                   TotalCharges, data = train, family = binomial)
# 2.Random Forest Model
model_rf <- randomForest(Churn ~ ., data = train, ntree = 100)

#Model Evaluation
#Make Predictions
pred_log <- predict(model_log, test, type = "response")
pred_rf <- predict(model_rf, test, type = "class")

#Convert Predictions to Binary
pred_log <- ifelse(pred_log > 0.5, "Yes", "No")

#Calculate Accuracy
conf_matrix_log <- confusionMatrix(as.factor(pred_log), test$Churn)
conf_matrix_rf <- confusionMatrix(pred_rf, test$Churn)
```

Results

```
> print(conf_matrix_log$overall["Accuracy"])
 Accuracy
0.7879004
> print(conf_matrix_rf$overall["Accuracy"])
 Accuracy
0.7964413
```

Customer churn analysis in the telecommunications sector is crucial for understanding why subscribers discontinue services. This analysis often involves examining various factors like demographics, service usage, contract details, and customer interactions to identify patterns and predict which customers are at high risk of churning. Machine learning models, such as Logistic Regression and Random Forest, are frequently employed to build predictive models. The provided accuracy scores (0.7879 and 0.7964) suggest that both models perform reasonably well in predicting churn, with the Random Forest model showing slightly better performance on this particular dataset. Identifying the key drivers of churn allows telecommunication companies to implement targeted retention strategies, ultimately reducing customer loss and improving profitability.



The provided feature importance analysis from a Random Forest model highlights key drivers of customer churn in the Telco dataset. 'TotalCharges', 'tenure', and 'MonthlyCharges' exhibit the highest importance, suggesting that customers with higher total spending, longer relationships, and greater monthly expenditure are less likely to churn. Contractual aspects ('Contract') and service-related features like 'OnlineSecurity' and 'PaymentMethod' also significantly influence churn. Demographic factors such as 'gender' and 'SeniorCitizen' show relatively lower importance compared to service usage and financial attributes in predicting customer attrition.

6. Discussion

The provided bar chart clearly illustrates the distribution of churn within the Telco customer base. A significant benefit of this visualization is its immediate and straightforward depiction of the churn rate. We can quickly observe that the number of customers who did not churn (represented by the larger red bar, indicating "No") is substantially higher than the number of customers who did churn (represented by the smaller teal bar, indicating "Yes"). This suggests that while churn is present, the majority of the customer base remains loyal.

However, a non-benefit of this high-level view is its lack of granular detail. While it confirms the presence and relative scale of churn, it doesn't provide any insight into *why* customers are churning. We cannot discern which customer segments are most affected, what services or contract types are associated with higher churn, or the tenure of the departing customers. To gain actionable insights for churn reduction, further analysis incorporating other variables like demographics, service usage, contract details, and customer interactions is crucial. [1] This initial overview serves as a starting point, highlighting the need for deeper investigation

to understand the underlying drivers of customer attrition and develop targeted retention strategies

The results show that Random Forest outperforms Logistic Regression in predicting customer churn. The higher accuracy of Random Forest suggests that non-linear relationships exist among features, which traditional regression fails to capture. Businesses can use these insights to focus on high-risk customers and design personalized retention strategies.

7. Conclusion

In conclusion, this RStudio-based analysis underscores the utility of machine learning, particularly Random Forest, for Telco customer churn prediction. The identified importance of factors like tenure, contract type, and charges provides actionable insights for retention strategies. Future research into deep learning and real-time systems holds promise for even more sophisticated and proactive churn management.

References

1. Gupta, A., Kumar, S., & Verma, P. (2022). Deep learning for customer churn prediction in telecom industry. *Journal of Data Science*, 20(3), 45-67.
2. Verma, R., & Srivastava, K. (2021). Machine learning approaches for customer churn analysis: A comparative study. *International Journal of Business Analytics*, 8(2), 89-101.
3. Choudhury, S., & Saha, D. (2020). Predictive analytics for customer churn using machine learning techniques. International Journal of Computer Applications, 176(24), 10–17.
4. Lee, H., & Park, Y. (2019). A comparative study on churn prediction models in the telecom industry using supervised learning algorithms. Expert Systems with Applications, 129, 1–11.
5. Rana, M., & Patel, R. (2021). Customer retention through churn prediction in the retail sector using data mining techniques. Journal of Retail Analytics, 15(4), 33–49.
6. Das, B., & Mishra, S. (2022). Visual analytics in customer churn prediction: Enhancing interpretability with R Shiny dashboards. Journal of Intelligent Data Science, 6(1), 22–35.
7. Fernandez, J., & Torres, M. (2020). Integrating machine learning with data visualization for churn analysis in banking. International Journal of Data Analytics, 5(3), 58–73.