Predicting Churn in Telecom:An Analytical Perspective

MATH 40024/50024: Computational Statistics

October 19, 2023

ACADEMIC INTEGRITY: Every student should complete the project by their own. A project report having high degree of similarity with work by any other student, or with any other document (e.g., found online) is considered plagiarism, and will not be accepted. The minimal consequence is that the student will receive the project score of 0, and the best possible overall course grade will be D. Additional consequences are described at http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism (http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism) and will be strictly enforced.

Instruction

Goal: The goal of the project is to go through the complete data analysis workflow to answer questions about your chosen topic using a real-life dataset. You will need to acquire the data, munge and explore the data, perform statistical analysis, and communicate the results.

Report: Use this Rmd file as a template. Edit the file by adding your project title in the YAML, and including necessary information in the four sections: (1) Introduction, (2) Computational Methods, (3) Data Analysis and Results, and (4) Conclusion.

Submission: Please submit your project report as a PDF file (8-10 pages, flexible) to Canvas by **11:59 p.m. on December 10**. The PDF file should be generated by “knitting” the Rmd file. You may choose to first generate an HTML file (by changing the output format in the YAML to `output: html_document`) and then convert it to PDF. **20 points will be deducted if the submitted files are in wrong format.**

Grade: The project will be graded based on your ability to (1) recognize and define research questions suitable for data-driven, computational approaches, (2) use computational methods to analyze data, (3) appropriately document the process (with R code) and clearly present the results, and (4) draw valid conclusions supported by the data analysis.

Example topics:

- Post-Hurricane Vital Statistics (https://www.biorxiv.org/content/10.1101/407874v2)
- Tidy Tuesday (https://github.com/rfordatascience/tidytuesday)

Datasets: I suggest to work on a dataset with at least thousands of observations and dozens of variables. You may consider (but are not restricted) to use the following data repositories: Data.gov (https://catalog.data.gov/dataset), Kaggle (https://www.kaggle.com/datasets), FiveThirtyEight (https://data.fivethirtyeight.com/), ProPublica (https://www.propublica.org/datastore/datasets), and UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)

Introduction [15 points]

Predicting Churn in Telecom

Introduction:

In the ever_changing telecommunications industry,where consumer preferences evolve rapidly, identifying and mitigating customer attrition is a crucial pursuit for sustainable business success.The phenomenon known as customer churn, which occurs when subscribers stop using services, affects revenue streams and is an indication of both market competitiveness and customer satisfaction.This project embarks on a comprehensive analysis of Telco customer churning,seeking to identify the essential factors influencing consumer decisions in the field of telecommunications services.

Telco Customer Churn Analysis: Research Questions

In this project, we explore a series of research questions aimed at uncovering insights into customer churn within the Telco telecommunications dataset. The analysis involves investigating various aspects. Where, each research question contribute to a deeper understanding of the factors influencing customer churn.

Research Questions:-

=>How do diverse combinations of demographic, service-related, and contractual variables interact to influence customer churn across various scenarios, and and what detailed trends and correlations emerge from the complex interaction of these factors?

=>How do changes influence the predictive models for Telco customer churn, and what insights can be gained regarding the most influential variables and their interactions in predicting customer churn behavior?

=>To what extent does the predictive performance of the model, trained on Telco customer churn data, vary when assessing model performance on distinct training and testing subsets, and how does the stability of performance impact the reliability of model predictions across different data partitions?

Identifying Telco Customer Churn: The Power of Data-Driven Computational Analysis

In the telecommunications industry, comprehending customer churn is crucial for maintaining a resilient customer base.In order to address important inquiries regarding churning patterns inside the Telco customer dataset, a computational, data-driven methodology is essential.This research explores the significance of applying computational methods for discovering insights that are not only comprehensive but also actionable. The Telco customer churn dataset likely contains a large volume of data with numerous variables. Utilizing a computational approach enables the efficient processing and analysis of this extensive information,which might be challenging or impossible to handle manually. A strong foundation for making meaningful inferences is provided by statistical tests and modeling approaches, which are frequently implemented computationally.These techniques aid in determining the importance of relationships and making informed decisions based on evidence. As datasets grow in size, a computational approach remains flexible.In order to handle the growing volume and complexity of data and enable consistent application of analysis across huge datasets, this is essential. A statistically rigorous and interactive environment is provided by computational tools, especially R.By exploring and visualizing the data iteratively, analysts can make dynamic modifications to their analyses.

Dataset Overview:-

The Telco customer churn dataset offers a comprehensive overview of consumers within a telecommunications company,capturing numerous aspects of their interactions, services utilized, and the critical outcome of figuring out whether they have churned or not.Based on this dataset, which serves as the foundation for our analysis, allowing us to gain insights into customer behavior and factors influencing churn.

Variables In the Dataset:-

The dataset includes identifiers and descriptors that collectively contribute to a comprehensive understanding of customer behavior.

CustomerID: A unique identifier assigned to each customer, facilitating individual tracking and analysis.

Gender: Indicates the gender of the customer, providing a demographic dimension for segmentation and understanding customer preferences.

SeniorCitizen: A binary variable specifying whether the customer is a senior citizen, which can influence service preferences and needs.

Partner: Indicates whether the customer has a partner, which could be relevant in understanding household dynamics and subscription patterns.

Dependents: A binary variable denoting whether the customer has dependents, offering insights into family-oriented service preferences.

Tenure: Represents the duration for which a customer has been subscribed to the telecommunications services, a crucial metric for understanding customer loyalty.

PhoneService: Indicates whether the customer has subscribed to phone services, a fundamental component of telecommunication offerings.

MultipleLines: Describes whether the customer has multiple phone lines, providing insights into the level of service subscribed.

InternetService: Specifies the type of internet service subscribed by the customer, such as DSL, fiber optic, or none.

OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport: These variables represent different aspects of additional services like online security, backup, device protection, and technical support, which contribute to the overall service package. StreamingTV,StreamingMovies: Indicate whether the customer has opted for streaming TV and movies, reflecting preferences for entertainment services.

Contract: Describes the type of contract the customer has, such as month-to-month, one year, or two years, providing insights into long-term commitment.

PaperlessBilling: A binary variable indicating whether the customer has opted for paperless billing, reflecting preferences for digital transactions.

PaymentMethod: Specifies the method of payment chosen by the customer, offering insights into payment preferences.

MonthlyCharges: Represents the amount charged to the customer on a monthly basis.

TotalCharges: Indicates the cumulative total charges incurred by the customer over their subscription period.

Churn: The target variable indicating whether the customer has churned (yes or no), serving as a critical metric for understanding customer retention.

Computational Methods [30 points]

Preparing Telco Customer Churn Data: Essential Data Wrangling Steps:-

Data loading and inspection,Data cleaning,Feature Engineering

Explanatory Data analysis:-

Visualization Techniques,Correlation Analysis,chi-square test

Modelling Techniques:-

Logisitic Regression,Random Forest

Resampling Techniques:-

k-fold cross-validation,Boot strapping

Evaluation Metrics for Data Analysis Quality:-

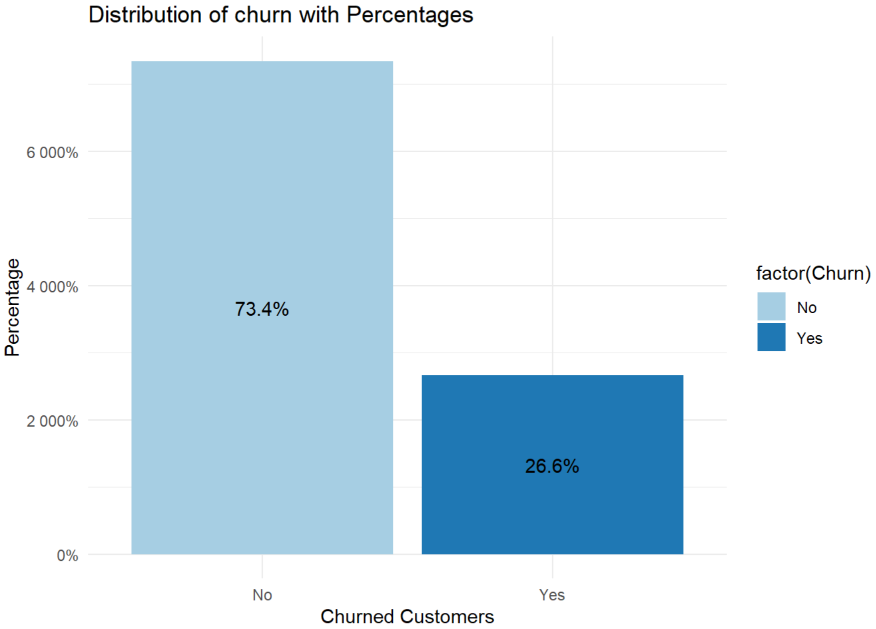
Accuracy,Precision, Recall,F1-Score,ROC-AUC(Receiver Operating Characteristic-Area Under The Curve),Confusion Matrix

Data Analysis and Results [40 points]

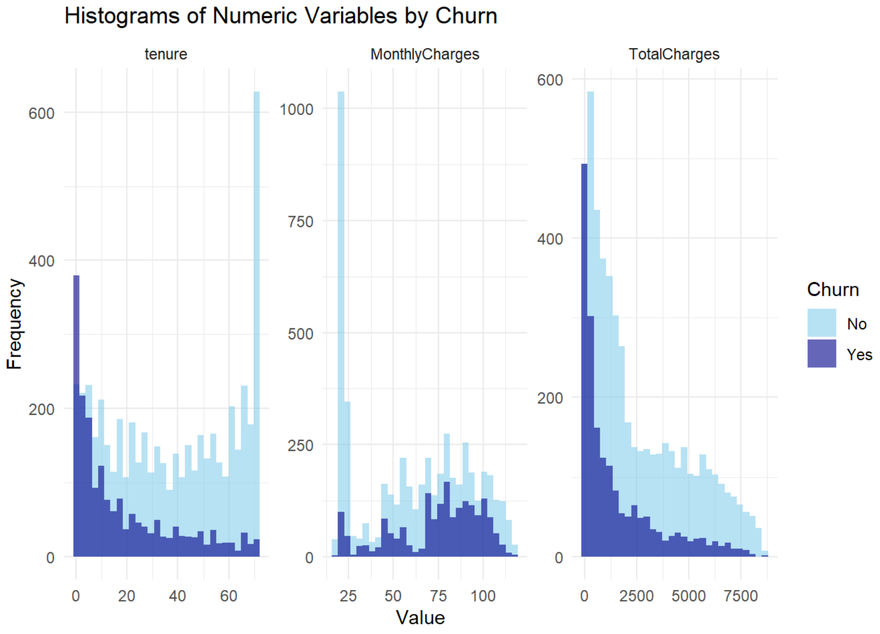
Explanatory Data Analysis

Libraries Used for the analysis:-

library(tidyverse),library(readxl),library(gridExtra),library(corrplot),library(pROC),library(caret),library(rsample),library(ggplot2),library(boot),library(reshape2),library(r



“The visual representation of the data reveals that 26.6% of customers have churned, while the majority, accounting for 73.4%, have not churned.”

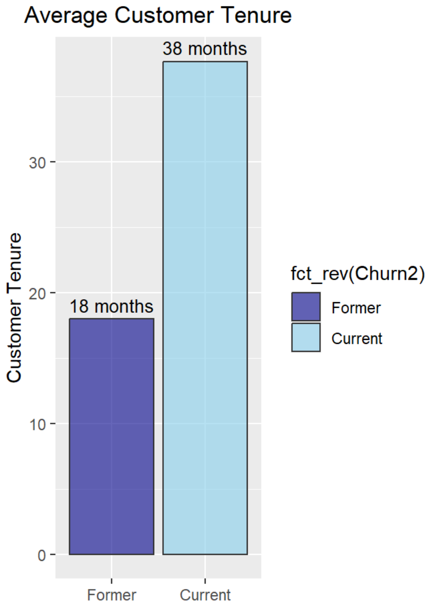


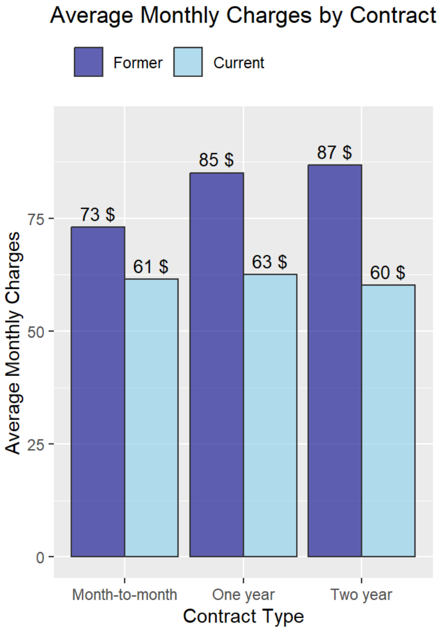
The histograms depicts the following: Less Tenure:The distribution of tenure for churned customers is concentrated towards lower values.

Higher Monthly Charges:Churned customers show a trend towards higher values in the distribution of monthly charges.

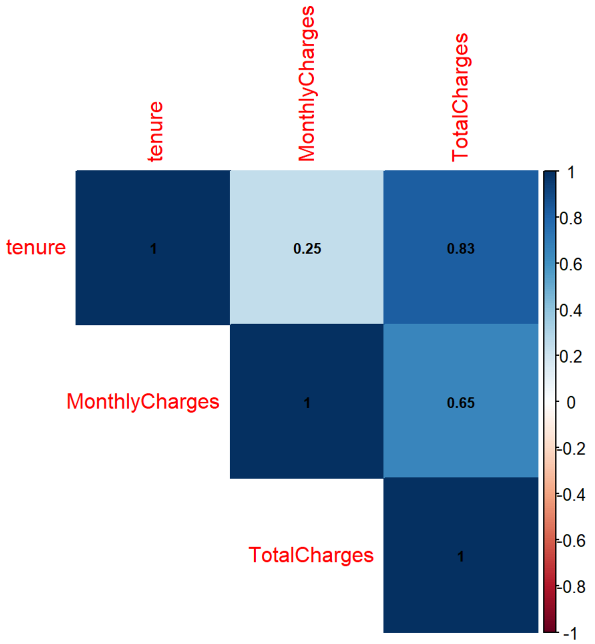
Less Total Charges:Churning customers may have lower total charges, indicating a potential pattern of early exits before accumulating higher total charges.

These observations may indicate that customers with shorter tenure, higher monthly charges, and lower total charges are more likely to churn.

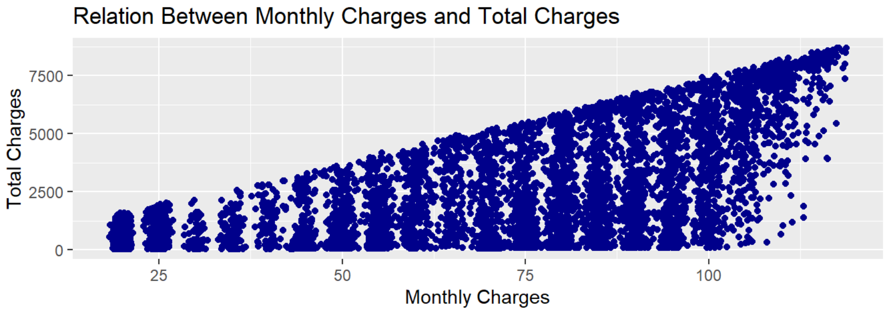


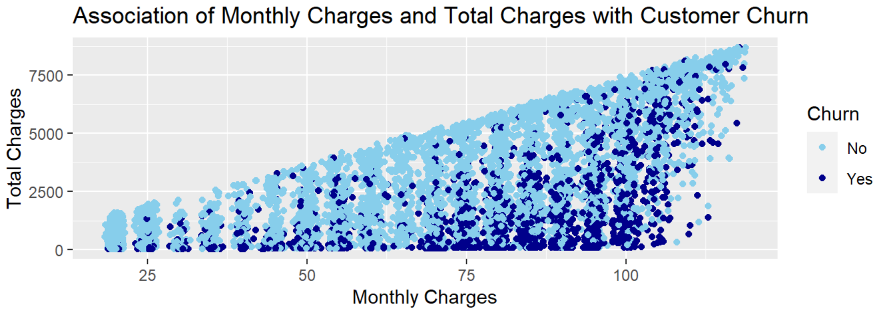


“The visualizations above depict the average tenure and monthly charges for both existing and former customers of Telco. Current customers, on average, have maintained their services for slightly over three years, while those who opted to leave had an average service duration of around 18 months.Furthermore, the data implies a trend where average monthly charges decrease for current customers across different contract durations, potentially indicating that recent customers, especially those on longer contracts, benefit from more favorable pricing.”

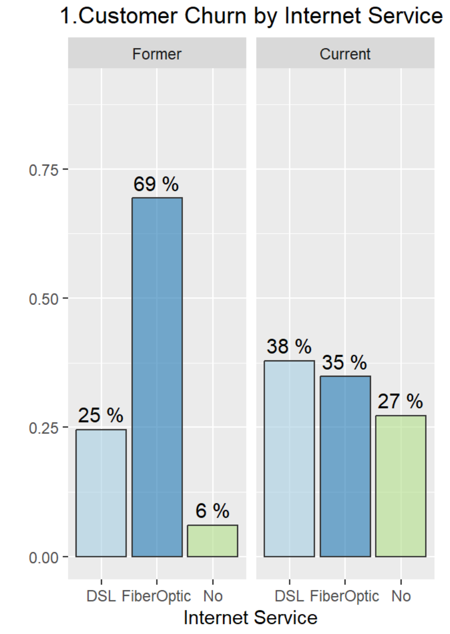


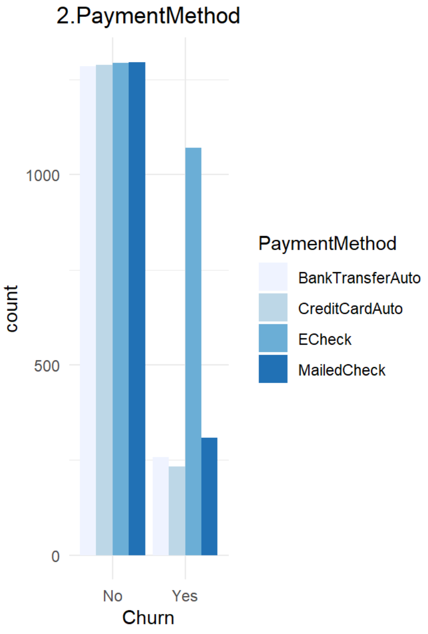
“The analysis reveals interesting connections between key variables in the telco customer churn dataset. Firstly, there is a weak positive correlation of about 0.25 between customer tenure and monthly charges, indicating a slight tendency for monthly charges to increase as the length of subscription (tenure) grows.Secondly, a strong positive correlation of approximately 0.83 exists between tenure and total charges. This implies that customers with longer subscriptions tend to accumulate higher total charges, which aligns with expectations.Lastly, there’s a moderate positive correlation around 0.65 between monthly charges and total charges. This suggests that customers with higher monthly charges also tend to have higher total charges.In simple terms, the correlation matrix offers valuable insights into how numeric variables interact with each other in the dataset, providing a clearer understanding of their relationships.” “Further analysis and modeling may be warranted to explore these relationships more deeply and understand the interplay between the categorical variables and customer churn.”





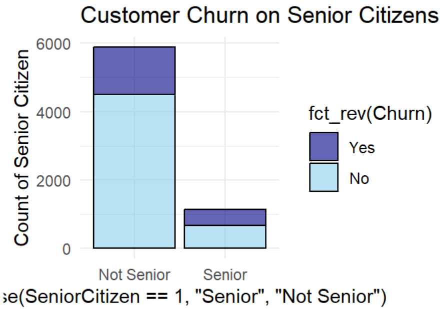
“The observed trend indicates a positive correlation between monthly charges and total charges. As the monthly charges increase, there is a corresponding rise in total charges. This pattern is associated with a higher likelihood of customer churn.”

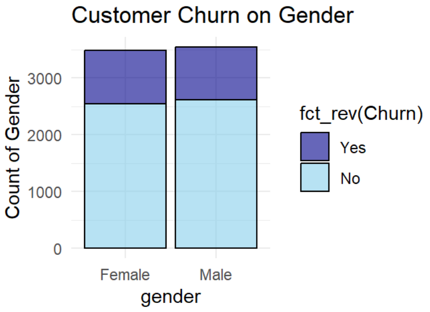


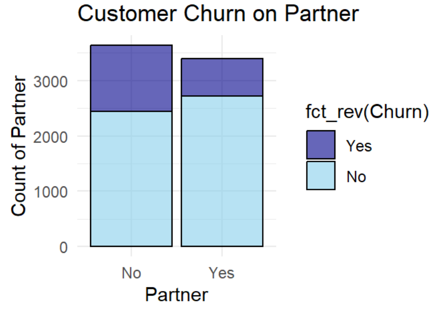


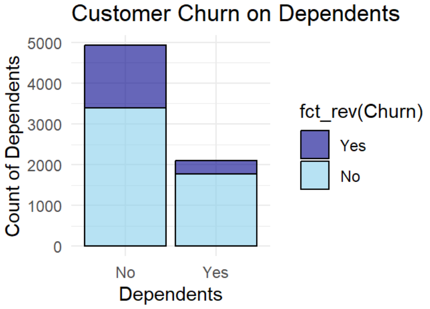
1.”First illustration depicts that among customers who left the internet service, a significant majority, 69%, had opted for Fiber Optic. In contrast, 25% had chosen DSL, and a smaller proportion, 6%, had no Internet service.Among current customers, 38% use DSL, making it the most popular choice, followed by 35% with Fiber Optic. Interestingly, 27% of present customers do not have an Internet service subscription.The higher proportion of Fiber Optic among departing customers could indicate potential dissatisfaction or issues related to this specific service offering.”

2.”From 2nd illustration the count of non-churning customers is relatively uniform across all four payment methods. In contrast, customers who have churned predominantly opted for the electronic check payment method.”

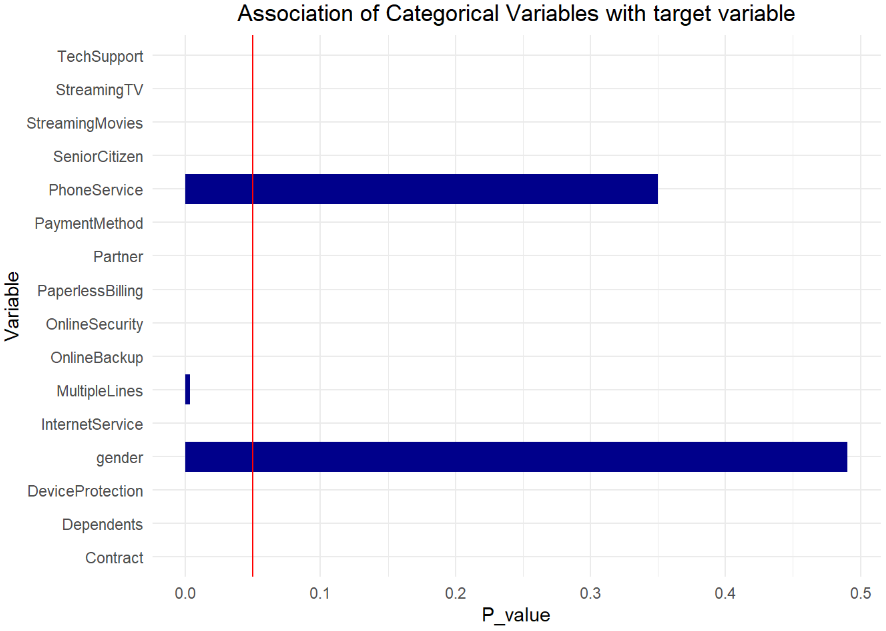








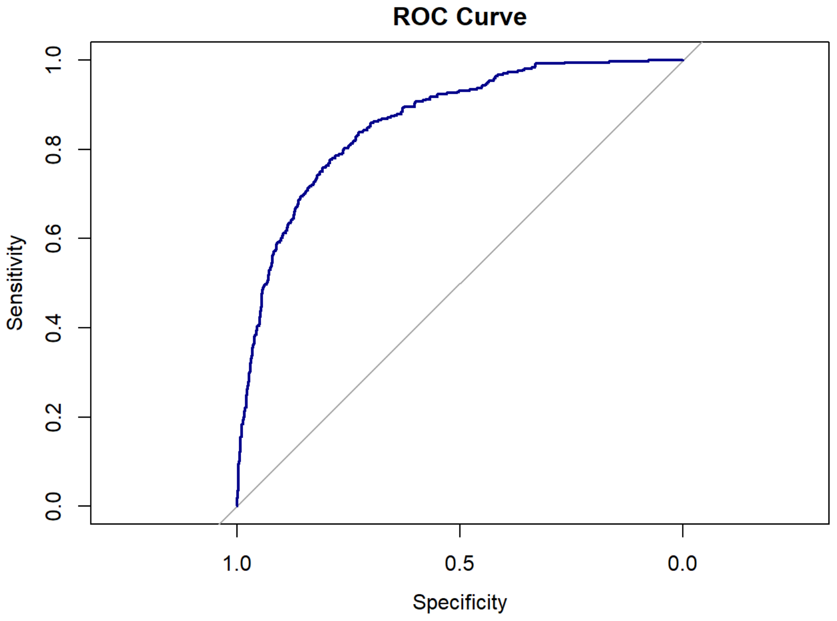
“The demographic factors of Senior Citizen, Partner status, and Dependents status may play significant roles in influencing customer churn. However, gender alone does not appear to be a prominent factor affecting churn behavior based on the observed patterns in the provided bar plots.”



“The analysis of the telco customer churn dataset brings forth meaningful insights. Almost all variables, except ‘Gender’ and ‘PhoneService,’ show a strong statistical connection with customer churn, evidenced by p-values below the widely accepted threshold of 0.05. This indicates the potential significance of these variables in predicting churn behavior.”

“Further analysis and modeling may be warranted to explore these relationships more deeply and understand the interplay between the categorical variables and customer churn.”

Modelling



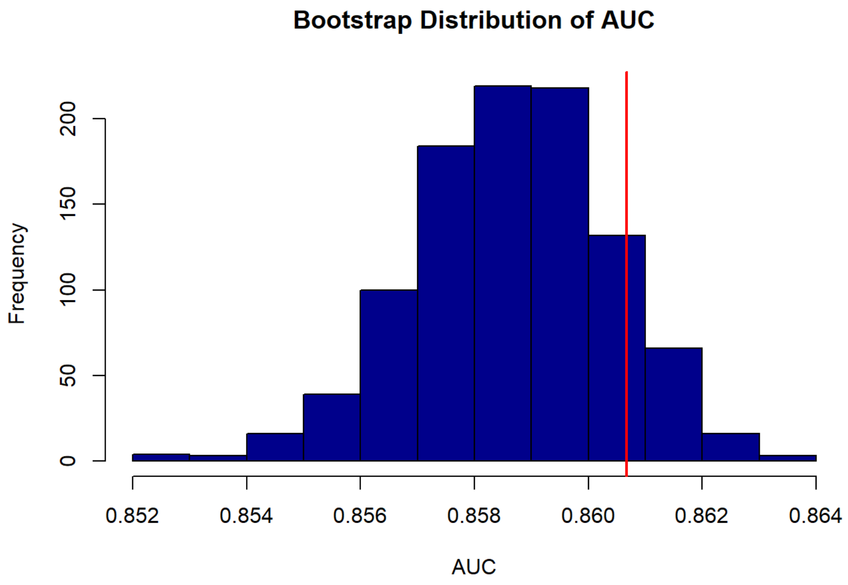
```
## The average evaluation metrics across all 10 folds after performing k-fold cross-validation:
```

```
## Average Precision: 0.8907767
```

```
## Average Recall: 0.8426472
```

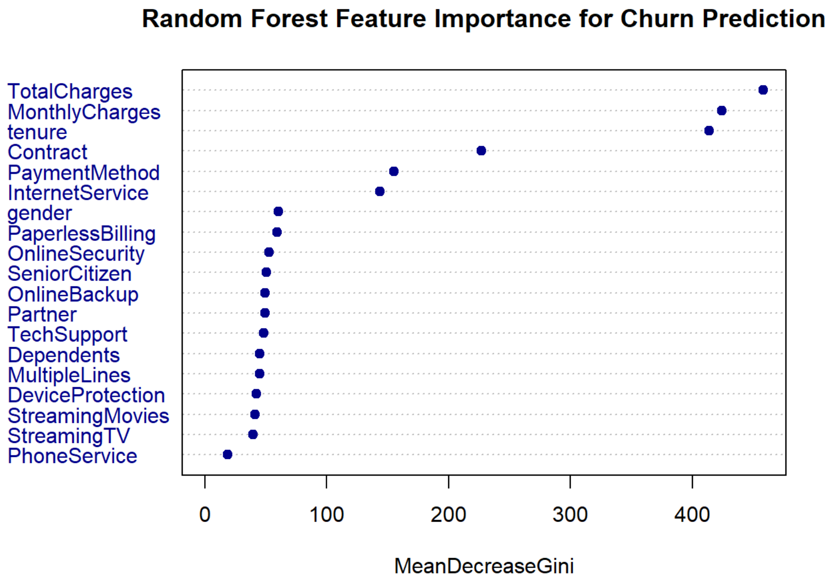
```
## Average F1-score: 0.8659643
```

```
## Average Accuracy: 0.7980468
## Average AUC: 0.8413718
```



The model exhibits consistent and robust predictive performance across all 10 folds in k-fold cross-validation. With an average precision of 0.8908, it excels in identifying positive cases, complemented by an average recall of 0.8426, indicating effective capture of actual positives. The balanced F1-score of 0.866 signifies a harmonious trade-off between precision and recall. The overall accuracy of 0.798 underscores the model's reliability in predicting both positive and negative cases, while the average AUC of 0.8414 confirms its discriminative ability. Additionally, bootstrapping validates the model's robustness, with the original AUC of approximately 0.86 falling within the confidence interval. This underscores the consistent and reliable predictive performance, emphasizing the significance of accounting for variability in customer behavior. The application of bootstrapping enhances model robustness, providing realistic confidence intervals and validating AUC stability. These combined findings underscore the model's effectiveness and practical utility in predicting customer churn for informed decision-making.

Feature Importance



The above illustration provides valuable insights into the significance of various features in anticipating customer churn. Among the variables, 'MonthlyCharges' and 'TotalCharges' demonstrate the highest importance with values of 431.08 and 440.59, respectively, indicating that these financial factors significantly contribute to the model's ability to differentiate between churn and non-churn instances. 'Tenure' follows closely with a substantial importance of 403.58, emphasizing the significance of customer tenure in predicting churn. Other notable factors include 'Contract' (255.38), 'InternetService' (119.04), and 'PaperlessBilling' (60.70), suggesting that the type of contract, internet service, and billing method also play crucial roles. These findings highlight the key drivers influencing customer churn in the telecommunications dataset, providing valuable guidance for retention strategies and business decision-making.

Conclusion [15 points]

Analysis of Research Questions:-

The analysis of the Telco customer churn dataset has provided valuable insights into the factors influencing customer attrition in the telecommunications industry. Through a comprehensive data analysis workflow, including data wrangling, exploratory data analysis, and predictive modeling, the study has illuminated key patterns and relationships within the dataset. The exploratory data analysis revealed several interesting observations: ->A higher proportion of churned customers had shorter tenure, higher monthly charges, and lower total charges. ->Internet service, contract type, and payment method were identified as significant factors associated with customer churn. ->Demographic factors such as being a senior citizen, having a partner, and having dependents were explored, but their impact on churn seemed less pronounced. The predictive modeling, particularly using logistic regression and random forest, demonstrated strong performance in identifying customer churn. The k-fold cross-validation and bootstrapping approaches ensured robustness and reliability in evaluating the model’s performance. The feature importance analysis highlighted the critical role of financial variables (MonthlyCharges, TotalCharges), tenure, and contract-related features in predicting churn.

Scope and Generalizability of the Analysis:-

Scope:

The examination of Telco customer churn involves a thorough exploration of demographic factors, financial metrics, service-related variables, and contractual details. This holistic approach provides a comprehensive understanding of customer attrition in the telecommunications industry. The practical significance lies in identifying key patterns and relationships, offering actionable insights for strategic decision-making in the Telco sector. The analysis not only uncovers the complexities of customer churn but also presents a practical framework for tailored retention strategies. Methodological rigor, including k-fold cross-validation and bootstrapping, enhances the reliability and applicability of predictive models across diverse scenarios, ensuring the validity of the findings.

Generalizability:

The analysis’s applicability reaches beyond the Telco customer churn dataset, offering insights transferable to the broader telecommunications sector. The methodologies used, from data wrangling to predictive modeling, are adaptable to similar contexts, allowing entities in comparable environments to optimize customer retention strategies. The findings, emphasizing financial variables, contract features, and service metrics, extend relevance to industries facing similar challenges. This versatility enables comparative studies and informs strategies across diverse sectors dealing with customer attrition dynamics.

Potential limitations and possibilities for improvement:

The analysis has limitations, relying on dataset representativeness and potential bias. Encoding strategies for categorical variables may impact model performance, and the absence of temporal information limits the ability to capture evolving trends. Additionally, logistic regression assumes linear relationships, potentially overlooking non-linear patterns. Improvements could involve addressing bias in encoding, incorporating temporal features, and exploring more complex models for more detailed relationships. To enhance model performance, explore feature engineering techniques, including creating new variables and incorporating time-related features. Experiment with ensemble methods like gradient boosting, optimize hyperparameters, and consider stacking models for improved predictions. Integrate external data sources for context, and implement continuous monitoring to adapt to evolving churn patterns. Additionally, employ validation metrics like area under the precision-recall curve for a thorough evaluation, especially in imbalanced datasets.