**GBA 6210-DATA MINING FOR BUSINESS ANALYTICS**

**DR. SHUNO ZENG**

**CASE STUDY**

**CUSTOMER RETENTION-TELECO CUSTOMER CHURN DATASET**

**GROUP 8**

**MEMBERS NAME**

**SHREYA SETH**

**NIDHI PATHAK**

**ZIMENG LIU**

**JAYARAM SRINIVASAN**

**SUNEET THAKUR**

**Executive Summary**

**Summary:**

The project focuses on predicting customer churn for a telecommunications (Telco) company using a classification model built in R. Customer churn, or the rate at which customers leave a service, is a critical metric for Telco companies, as it directly impacts revenue and profitability. By identifying factors that contribute to customer churn, the company can take proactive measures to retain customers and improve service offerings. The project aims to build a predictive model that can accurately classify customers as churners or non-churners, allowing the company to target interventions more effectively and reduce churn rates.

**Dataset Description:**

The dataset used for this project contains 7043 customer records with 21 columns. Each row represents a customer, and each column contains specific customer attributes such as gender, senior citizen status, partner status, dependents status, tenure, contract type, billing method, payment method, monthly charges, total charges, and churn status.

**Approaches and Techniques:**

Data Exploration where we explored the dataset and its statistic summary to understand the distribution of variables and relationships between them. We then preprocessed the data, including removing outliers to ensure data quality. Next, we performed data and dimension reduction using unsupervised learning techniques, such as PCA (Principal Component Analysis), to identify important components that contribute most to the data's variance. This helped us reduce the dataset's dimensionality while retaining important information. Further, to evaluate our models, we used cross-validation, specifically 10-fold cross-validation, to partition the data into subsets for model evaluation. This approach ensured that our models were robust and generalizable.

**Algorithms Used:**

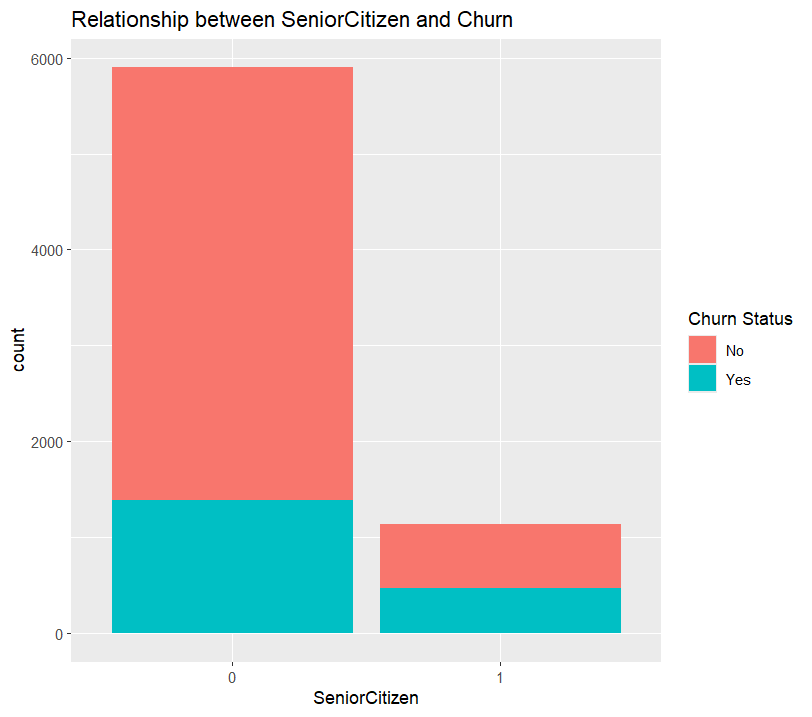
* Data Loading and Exploration: You loaded the dataset and explored its dimensions, first few rows, and specific subsets of data.
* Data Visualization: You created various plots to visualize the relationships between different variables and the churn status.
* Data Preprocessing: You handled missing values in the TotalCharges column and removed outliers using the IQR method.
* Dimensionality Reduction: You used Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and visualize the variance explained by each principal component.
* Data Partitioning: You partitioned the dataset into training and testing sets for model evaluation using cross-validation.
* Modeling: You used several algorithms for classification, including logistic regression and random forest. You trained these models on the training data and evaluated them on the testing data, calculating accuracy and generating ROC curves.
* Decision Trees: You built a decision tree model to predict churn based on tenure.
* Model Evaluation: You evaluated the models using confusion matrices and calculated metrics such as accuracy, sensitivity, specificity, precision, FDR, and FOR for both training and validation sets.
* Libraries: You used various R libraries such as caret, rpart, rpart.plot, e1071, gains, randomForest, ggplot2, gplots, reshape, GGally, and MASS for data manipulation, visualization, and modeling.

**Task 1:**

Data exploration

* Data loading and Overview: Started with loading the Telco-Customer-Churn dataset into RStudio using the read.csv function. This allowed us to access the data for analysis. Next, we used the ‘**dim**’ function to determine the dimensions of the dataset, providing us with the total number of rows and columns. Following this, we used the ‘**head**’ function to preview the first six rows of the dataset, giving us a glimpse of the data structure and contents. This initial overview helped us understand the size and format of the dataset before delving into further analysis.
* Subsets Creation: Subset some specific columns and rows to understand the data structure and content of that particular subset and how it is impacting our goal. We identified 11 missing values in the 'TotalCharges' column using the is.na function. To address these missing values, we used the mean function to calculate the average 'TotalCharges' and fill in the missing values with this average. This allowed us to ensure that the missing values did not impact our analysis.
* Packages installation and loading : Next, I installed several R packages, including 'forecast', 'ggplot2', 'gplots', 'reshape', 'GGally', and 'MASS', to visualize individual variables and compare them to each other and our target variable, 'Churn'. This helped me identify which variables have a significant impact on the dataset and are more likely to contribute to churn.
* Visualization of Key Variables and Interpretation:

1. BarPlot (SeniorCitizen and Churn)



Interpretation

Insights:

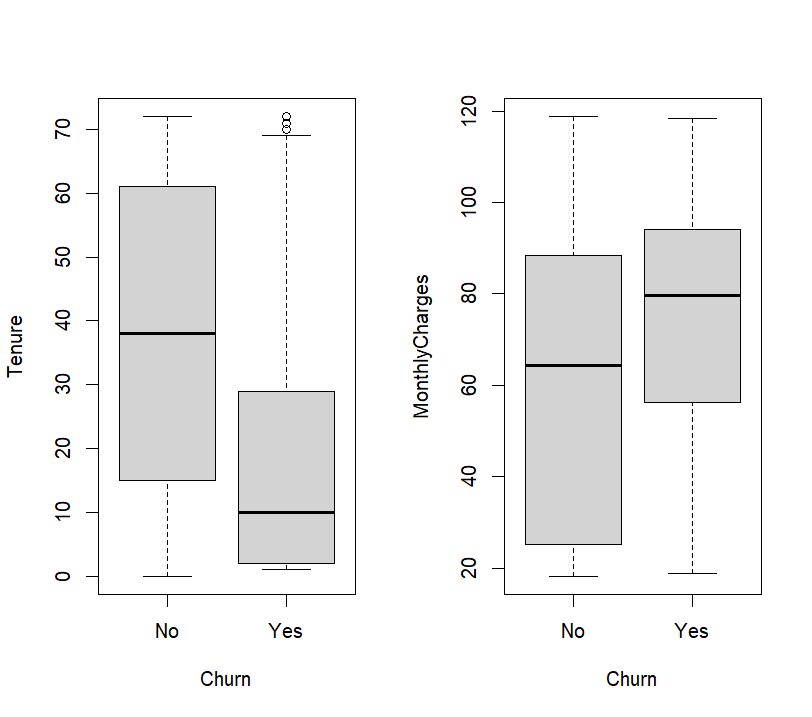
Senior Citizen:

The bar for SeniorCitizen=0 (non-senior citizens) is noticeably higher than the bar for SeniorCitizen=1 (senior citizens), indicating that the company has more non-senior citizen customers than senior citizens.

Churn:

The higher churn rate among non-senior citizens compared to senior citizens suggests that senior citizens are potentially more loyal customers to the company, with lower chances of churning. This observation could guide the focus of analysis towards younger customers, as they appear to be more likely to churn compared to senior citizens.

1. Side By Side Box plot (tenure Vs Churn and MonthlyCharges Vs Churn)



Interpretation

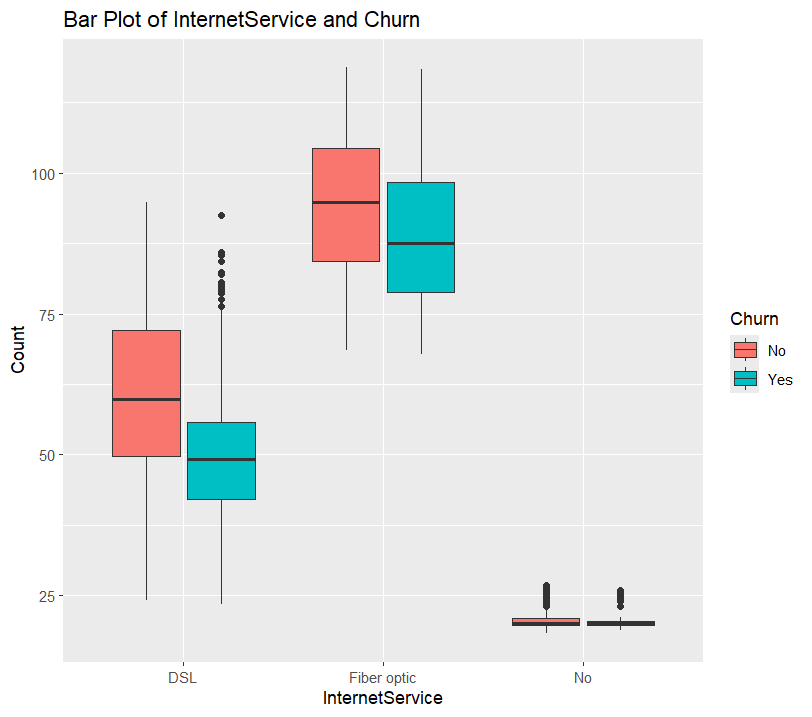
Insights (Tenure Vs Churn)

1. Tenure and Churn: Customers with lower tenure (less time with the company) are more likely to churn. This could indicate that newer customers are still evaluating the service and may be more willing to switch if they are not satisfied.

2. Monthly Charges and Churn: Customers with higher monthly charges are more likely to churn than those with lower charges. This could imply that customers who are paying more for the service may have higher expectations or may be more sensitive to price changes, leading them to churn if they are not fully satisfied with the service or if they find better deals elsewhere.

Combined Insights:

These observations can be potentially connected. Customers with lower tenure (newer) might be more price-sensitive and more likely to churn if they have a higher monthly charge. They might not have yet experienced the full value of the service or become invested enough to justify the higher cost compared to potentially cheaper alternatives.

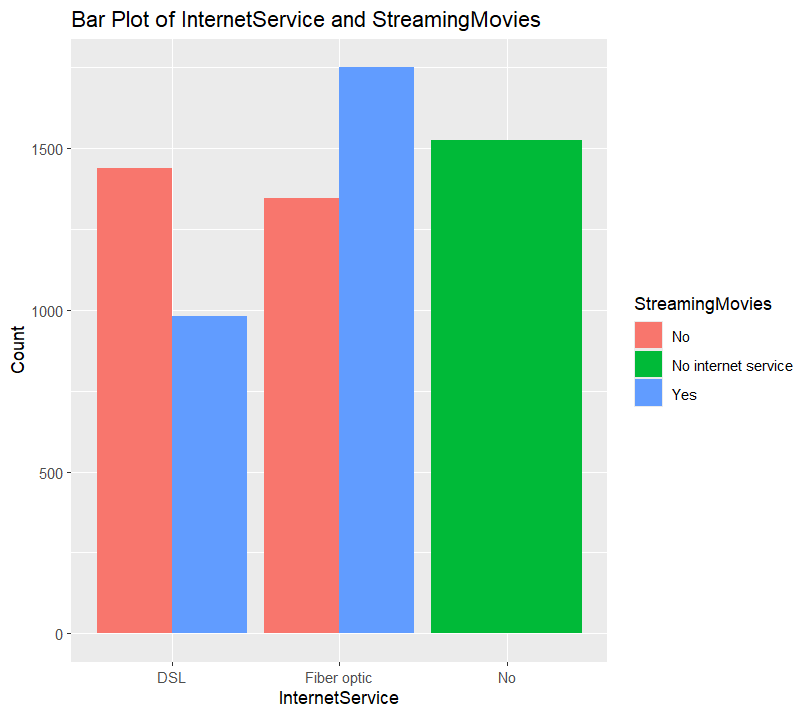
1. Boxplot (InternetService with MonthlyCharges, fill = Churn)

Interpretation:

Insights:

* Fiber Optics: The boxplot shows a potentially higher churn rate for fiber optic customers compared to DSL. This aligns with our assumption that higher costs might be a factor.
* DSL: The churn rate for DSL seems to be spread out, with a wider box (representing the interquartile range) and potentially some outliers on the higher end (potentially high-cost plans). This suggests churn might be more prevalent for mid-range or high-priced DSL plans.
* No Internet: The churn rate for "No" (likely customers who don't have internet service). It doesn’t show any major insights which says that there are fewer people or may be old people who buys no Internet Service as they less usage of the devices.

1. Barplot(InternetService Vs StreamingMovies/)



Interpretation:

Insights:

Further curiosity and to learn more about what other reasons could be for the fiber optics to have high churning rate we created a bar plot comparing two variable Streaming Movies and Internet Service. Though Streaming Service doesn’t definitely prove that it is the sole cause of Churn but still give a deep insight. Highly Chance that the Customers might be Price Sensitive. The fiber optics are probably the high selling service and using additional streaming line their high internet usage can lead for the higher Charges.

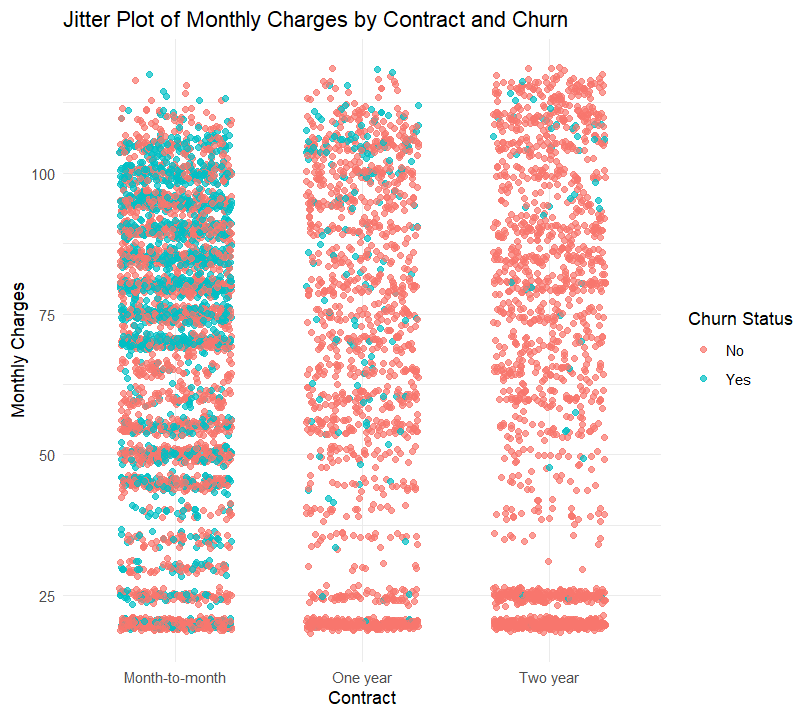
* Value Perception and Usage Data: Analyzing usage data for both internet and streaming services can reveal if customers are utilizing the full potential of their fiber optic plans and the bundled streaming services.
* Underutilized Services: If customers are not fully utilizing the features of the bundled services (e.g., not watching all the channels in a streaming TV package), they might perceive lower value and be more likely to churn.
* Personalized Recommendations: Based on usage data, the ISP could offer personalized recommendations for streaming content or suggest alternative bundled packages that better match customer needs. This can improve customer satisfaction and reduce churn.

**Suggestions:** Also, to retain the higher Charges and save Customers company can come with the target promotions

 Discounted streaming service subscriptions: Partnering with streaming services to offer discounted subscriptions bundled with fiber optic plans.

 Tiered Bundles: Creating tiered bundles with different combinations of internet speeds, streaming services, and other value-added features at competitive prices. This caters to different spending preferences within the high-spender segment.

1. Jitter Plot (Monthly Charges by Contract and Churn)



Interpretation

We opted for a Jitter Plot to compare numerical values with categorical data, as the fit of bar plots and box plots didn't seem appropriate. We explored various plot types and found the Jitter Plot to be the most suitable, providing a clear view of the large set of Monthly Charges and their relation to Contract types. The Jitter Plot handles the overlapping data values effectively, preventing them from obscuring each other, and presents each data point's relationship to the Contract type more clearly

Insights:

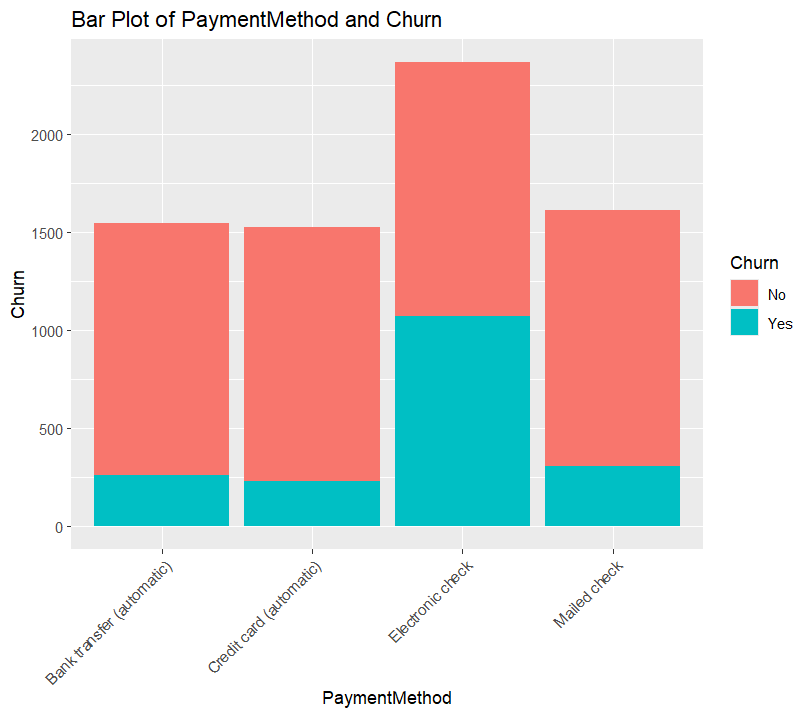
* Churn and Month-to-Month Contracts: We identified a higher churn rate for month-to-month contracts, particularly in the $60-$100 price range.
* Loyalty in Longer Contracts: Customers on yearly and two-year contracts seem to have lower churn rates, suggesting higher satisfaction and loyalty.
* Possible Reasons for Month-to-Month Churn: We identified two potential explanations for churn in month-to-month contracts with the Monthly Charges

1. Low Tenure and Service Exploration: New customers might be exploring the service and churn if it doesn't meet their expectations.
2. Price Sensitivity and Competitor Options: Customers on month-to-month contracts might be more price-sensitive and susceptible to better deals from competitors.

Suggestion:

**Contract Flexibility:** Offering short-term contracts (3-month, 6-month) could be an option for price-sensitive customers who want more stability than month-to-month but aren't ready for a year-long commitment.

1. PaymentMethod Vs Churn



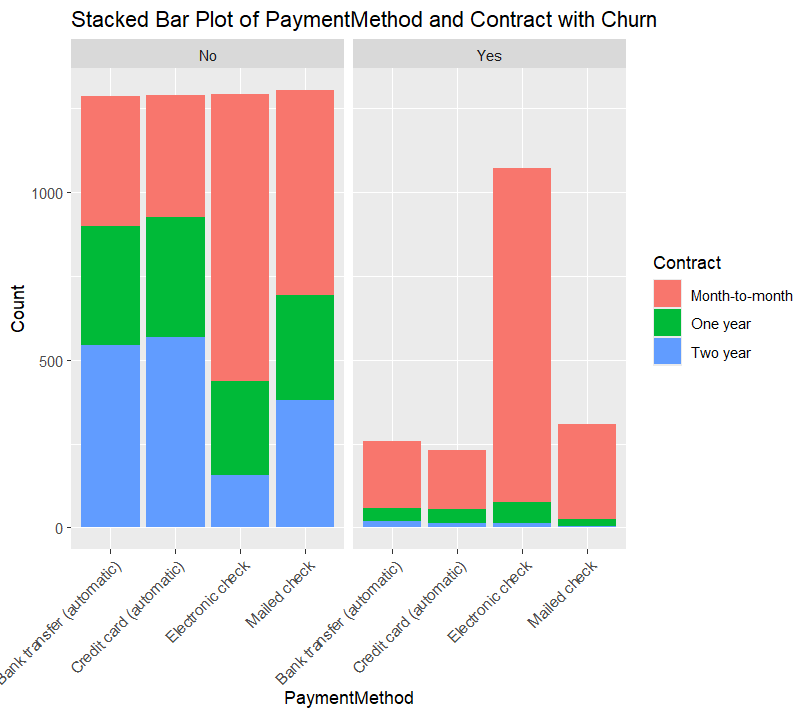
Interpretation

Insights:

* Electronic Checks and Churn: You're right - electronic checks have the highest usage but also the highest churn rate compared to other payment methods.
* Churn Rate for Electronic Checks: It's interesting to note that the churn rate for electronic checks is still relatively close to the non-churning rate. This suggests an opportunity to address churn before it becomes a major problem.

Before making any assumptions about why electronic Check has the highest churn rate we compared Payment Method with another variables to see where customers are highly paying with the electronic Check and we considered that Contract month-to-month had the higher number compare to any other variables and there churn rate are higher too.

1. Stacked Bar Plot of PaymentMethod and Contract with Churn



Interpretation

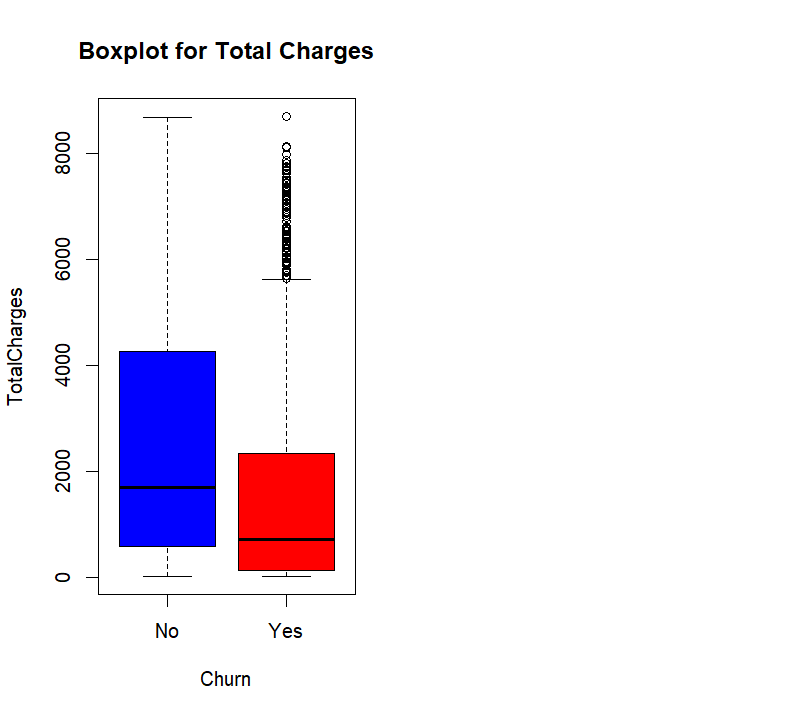
Insights

In our earlier analysis, we observed that customers with MonthlyCharges and Churn we observed in that month-to-month contracts had the highest churn rates. We speculated that this could be due to price sensitivity and the stress of monthly payments, or because new customers were exploring services and switching if they found better deals elsewhere. However, when we further investigated payment methods, we found that customers using electronic checks had the highest churn rates too.

To understand this better, we plotted a scatter plot and noticed that while month-to-month contracts had a high proportion of non-churning customers, they also showed the highest churn rates. This suggests that the prevalence of month-to-month contracts might also influenced by the use of electronic checks.

One possible reason for this could be perceived security risks associated with electronic checks. Customers may view electronic checks as less secure, leading them to switch to other payment methods or service providers. Additionally, the potential for failed payments due to insufficient funds in the linked account could result in service disruptions, late fees, and customer frustration, all of which could contribute to churn.

1. Boxplot(TotalCharges Vs Churn)

****

Insights:

* Churn and Total Charges: The boxplot shows a trend of higher churn rates for lower total charges. This seems counterintuitive as one might expect price sensitivity to lead to more churn for higher charges. However, there are a large number of outliers in the high total charges category.

Total charges are strongly related to monthly charges, so any factors affecting monthly charges also influence total charges. In our analysis, we observed that customers who churned tended to have moderate monthly charges. Factors such as contract type, tenure, and internet service choice played roles in churn rates related to monthly charges.

• For instance, high-demand internet services may come with higher costs, potentially leading some customers to churn. We also noted that customers with shorter tenure and high monthly charges were more likely to churn compared to those with very high tenure.

• This indicates that customers who are price-sensitive or those who are new to the service may struggle with managing their monthly payments. They may be dissatisfied with the service or prefer the flexibility of month-to-month contracts, but the stress of managing payments and potential late fees could prompt them to leave the company. Due to all these reasons the churn rate of Total Charges with low charges effecting more.

• To address this issue, the company could consider offering longer contract terms, such as quarterly contracts, to alleviate the burden of monthly payments. They could also provide education about electronic payment options to simplify the payment process. Additionally, offering promotional deals on high-demand services could attract and retain customers.

Moving on the churn side, the presence of outliers in total charges indicates that some customers with exceptionally high charges may be churning. This raises questions about the reasons behind their churn. Are these customers dissatisfied with the service despite the higher charges? Or are there other factors at play, such as billing errors or unexpected charges?

Further curiosity we decided to analyze more into these outliers.

**Task 2**

**Data Preprocessing**

Our primary objective was to identify and address outliers that could affect our analysis. Upon investigation, we found that the outliers in the TotalCharges column were actually missing values. Further analysis we looked up the rows where these missing values were located and it revealed that these missing values were associated with customers who had a tenure of 0 and were also paying MonthlyCharges and had taken services to use . This suggests that these customers were likely new active and had recently started using the company's services. Therefore, the outliers we identified were not errors but rather a natural variation in the dataset

> # Subset the dataframe to show rows where TotalCharges is NA

> na\_rows <- data.df[is.na(data.df$TotalCharges), ]

> # Print the subsetted dataframe

> print(na\_rows)

customerID gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines

489 4472-LVYGI Female 0 Yes Yes 0 No No phone service

754 3115-CZMZD Male 0 No Yes 0 Yes No

937 5709-LVOEQ Female 0 Yes Yes 0 Yes No

1083 4367-NUYAO Male 0 Yes Yes 0 Yes Yes

1341 1371-DWPAZ Female 0 Yes Yes 0 No No phone service

3332 7644-OMVMY Male 0 Yes Yes 0 Yes No

3827 3213-VVOLG Male 0 Yes Yes 0 Yes Yes

4381 2520-SGTTA Female 0 Yes Yes 0 Yes No

5219 2923-ARZLG Male 0 Yes Yes 0 Yes No

6671 4075-WKNIU Female 0 Yes Yes 0 Yes Yes

6755 2775-SEFEE Male 0 No Yes 0 Yes Yes

InternetService OnlineSecurity OnlineBackup DeviceProtection

489 DSL Yes No Yes

754 No No internet service No internet service No internet service

937 DSL Yes Yes Yes

1083 No No internet service No internet service No internet service

1341 DSL Yes Yes Yes

3332 No No internet service No internet service No internet service

3827 No No internet service No internet service No internet service

4381 No No internet service No internet service No internet service

5219 No No internet service No internet service No internet service

6671 DSL No Yes Yes

6755 DSL Yes Yes No

TechSupport StreamingTV StreamingMovies Contract PaperlessBilling

489 Yes Yes No Two year Yes

754 No internet service No internet service No internet service Two year No

937 No Yes Yes Two year No

1083 No internet service No internet service No internet service Two year No

1341 Yes Yes No Two year No

3332 No internet service No internet service No internet service Two year No

3827 No internet service No internet service No internet service Two year No

4381 No internet service No internet service No internet service Two year No

5219 No internet service No internet service No internet service One year Yes

6671 Yes Yes No Two year No

6755 Yes No No Two year Yes

PaymentMethod MonthlyCharges TotalCharges Churn

489 Bank transfer (automatic) 52.55 NA No

754 Mailed check 20.25 NA No

937 Mailed check 80.85 NA No

1083 Mailed check 25.75 NA No

1341 Credit card (automatic) 56.05 NA No

3332 Mailed check 19.85 NA No

3827 Mailed check 25.35 NA No

4381 Mailed check 20.00 NA No

5219 Mailed check 19.70 NA No

6671 Mailed check 73.35 NA No

6755 Bank transfer (automatic) 61.90 NA No

After visualizing the outliers, we used the quartile IQR method to check for any remaining outliers. The absence of outliers after this process indicates that our approach effectively managed the data points.

To address churn among senior citizens and long-term customers, we recommend that the company introduce low-cost service plans. These plans could help retain these valuable customers and improve overall customer satisfactionTo address churn among senior citizens and long-term customers, we recommend that the company introduce low-cost service plans. These plans could help retain these valuable customers and improve overall customer satisfaction.Top of Form

Task 3

Data and dimension reduction

To reduce the complexity, remove redundancy and Visualize High dimensional data we chose algorithm PCA (Principal Component Analysis) which reduce high dimensional data into lower dimensional space while preventing the most important patterns and improve model performance.

* The first step we begun with Converting “Churn" variable to numeric: We convert the "Churn" variable from categorical ("Yes" or "No") to numeric (1 or 0) using the ifelse function. This conversion is necessary for PCA, which requires numeric input.
* Remove rows with missing values: We remove any rows with missing values using the na.omit function. Missing values can interfere with PCA analysis.
* Next, we specify the numeric variables (TotalCharges, MonthlyCharges, tenure, Churn, and SeniorCitizen) that we want to include in the PCA analysis.
* Perform PCA: We use the prcomp function to perform PCA on the selected numeric variables. The scale = TRUE argument scales the variables to have unit variance before performing PCA. The highest variances were seen in the PC1 and PC2.

> pca\_result

Standard deviations (1, .., p=5):

[1] 1.4990354 1.1788917 0.9136846 0.6853477 0.2420460

Rotation (n x k) = (5 x 5):

PC1 PC2 PC3 PC4 PC5

TotalCharges 0.6516556 0.01036373 0.11086725 -0.07581986 0.74645657

MonthlyCharges 0.4459354 0.46833639 0.31097911 0.58179442 -0.38289676

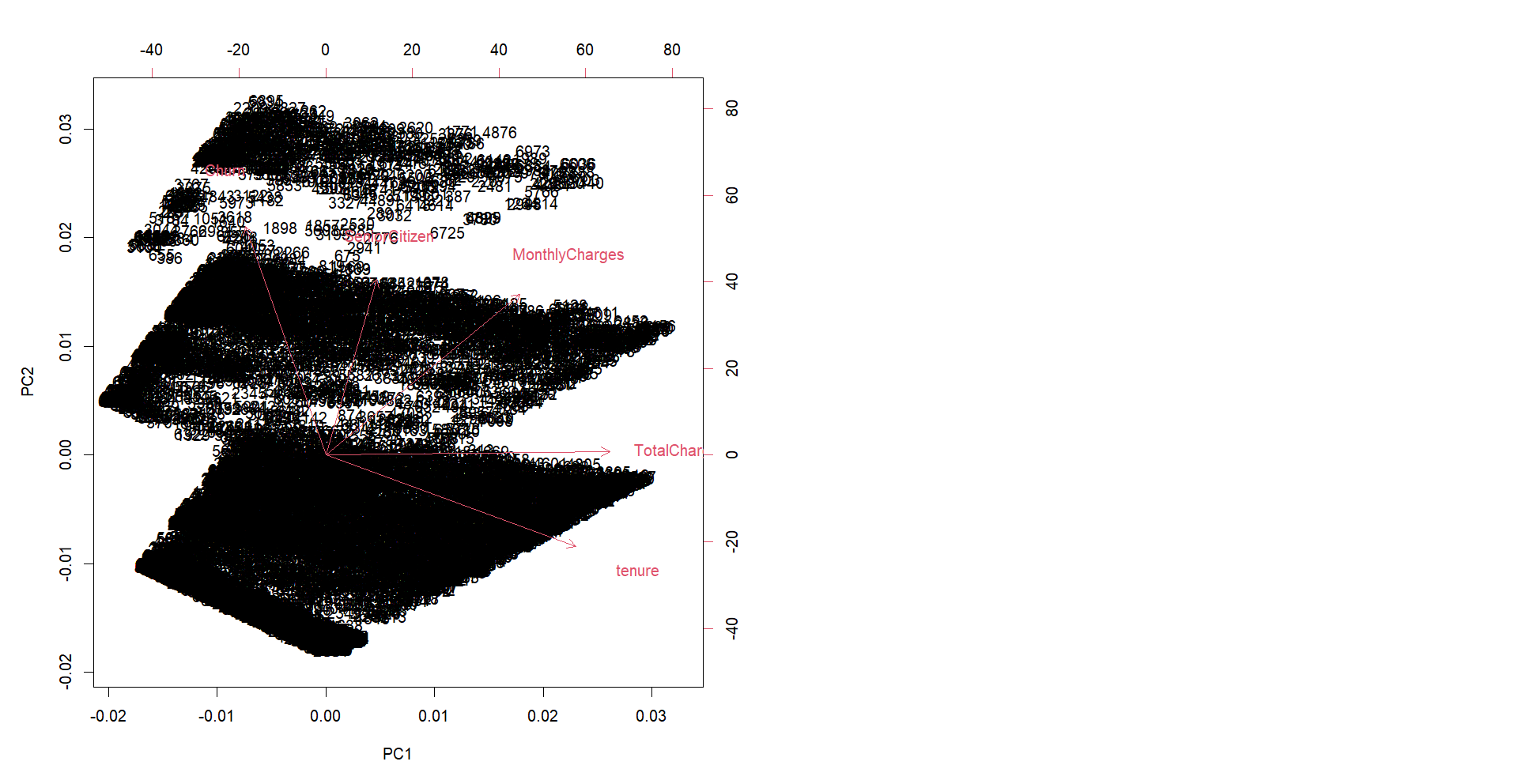
tenure 0.5735197 -0.26685806 -0.06348003 -0.54837328 -0.54324850

Churn -0.1839556 0.66728116 0.40823548 -0.59440787 0.03031962

SeniorCitizen 0.1171415 0.51388349 -0.84871345 -0.04162495 0.01242784

* Biplot: We create a biplot to visualize the principal components and their relationships

with the original variables. This plot helps us understand the structure of the data and the contribution of each variable to the principal components.



Interpretation

Insights

* **TotalCharges:** The position of the "TotalCharges" point suggests a strong positive correlation with PC1 and a weaker positive correlation with PC2.
* **MonthlyCharges:** The position of the "MonthlyCharges" point might indicate a moderate positive correlation with both PC1 and PC2.
* **tenure:** The "tenure" point might show a positive correlation with PC1 and a negative correlation with PC2.
* **Churn:** The "Churn" point likely has a strong positive correlation with PC2 and a weaker negative correlation with PC1.
* **SeniorCitizen:** The "SeniorCitizen" point might show a negative correlation with PC3 (not shown on the axes) and a possible positive correlation with PC2.

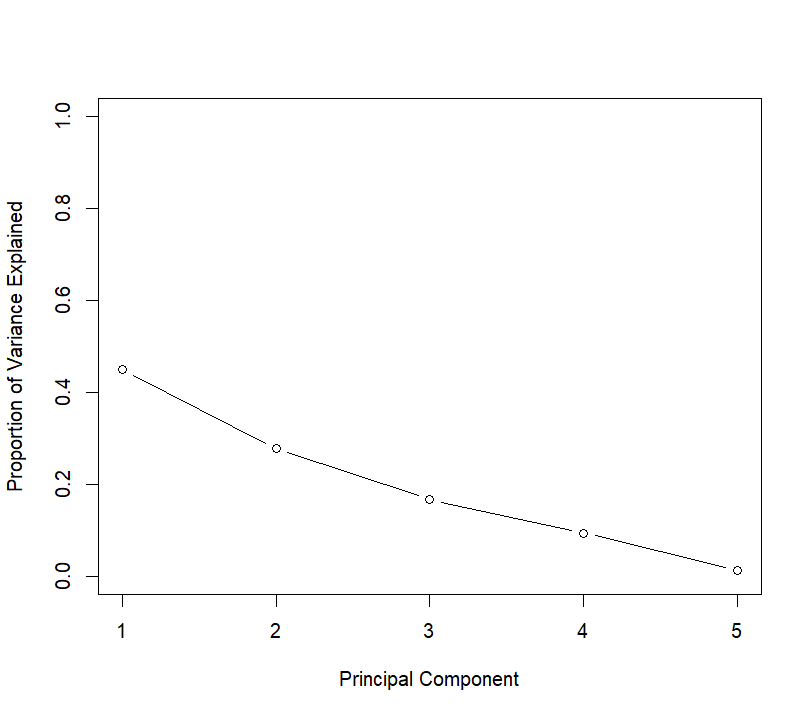
**Contract-Related Factors:** Given that PC2 captures variance related to churn and senior citizen status, PC1 might capture factors related to customer contracts that influence churn but are not directly tied to senior citizens. These could include:

* **Tenure:** Customers with longer tenure (represented by the "tenure" variable) might churn less often as they might have established a routine and found value in the service. This means that customers with the lower tenure has more chance of churning.
* **Total Charges:** Customers with higher total charges (represented by the "TotalCharges" variable) might also churn less often if their plan reflects a higher commitment level or specific needs they are satisfied with.
* **Monthly Charges (to a lesser extent):** Monthly charges might also play a role on PC1, but potentially to a lesser extent compared to PC2 (which captures churn). Customers with higher monthly charges might indicate a higher level of service usage, potentially leading to less churn.

**Interpretation in case of PC2**

* **PC2 Captures Churn and SeniorCitizen Variance:** The high variance of PC2 being explained by "Churn" and "SeniorCitizen" suggests that these two variables are strongly correlated and contribute significantly to the overall variability captured by PC2.
* **Potential Underlying Relationship:** There might be an underlying relationship between customer churn and senior citizen status that PC2 is highlighting. For example, senior citizens might be more likely to churn due to factors like changes in service needs, technological challenges, or price sensitivity.

Further to be more accurate for choosing right PCs for that we did square standard deviation of the principal component. Standard deviation is the square root of variance, so squaring it gives us the variance explained by each PC. And the output we got for each PCs displayed that the PC1 and PC2 had the highest variances of around 72% than PC3, PC4 and PC5.



Observation:

* PC1 appears to explain the largest proportion of the variance in the data (highest bar).
* The PVE values decrease for subsequent PCs, indicating they explain progressively smaller portions of the variance.
* Visualizing the exact elbow point is difficult but as we can see in PC1 and PC2 where we see the highest of the variance of 45% and 28% which I got through the var( **pce**) calculation we did earlier but after PC3and PC4 onwards seems to have a much lower impact and less significant portion.

Task 4

This task focuses on implementing k-fold cross validation, to evaluate model performance. Following steps are taken to achieve the same:

1. **k <- 10:** This line initializes the variable **k** to 10, indicating that we want to perform 10-fold cross-validation.
2. **n <- nrow(data.df)**: This line calculates the total number of observations in the dataset data.df and stores it in the variable n.
3. **subset\_size <- floor(n / k)**: This line calculates the size of each fold by dividing the total number of observations by the number of folds (**k**). The floor function ensures that we get an integer value.
4. **remaining <- n %% k**: This line calculates the number of remaining observations that cannot be evenly distributed among the folds.
5. **folds <- vector("list", k)**: This line creates an empty list folds to store the folds.
6. **indices <- sample(n)**: This line shuffles the indices of the dataset data.df to ensure randomness in the selection of observations for each fold.
7. Partitioning the dataset into k folds:

The for loop iterates over each fold (i) from 1 to **k**.

Inside the loop:

**start\_index and end\_index** are calculated to determine the range of indices for the current fold.

If there are remaining observations (remaining) and the current fold index is less than or equal to the remaining folds, one extra observation is added to the current fold.

**fold\_indices** stores the shuffled indices corresponding to the current fold.

The observations corresponding to **fold\_indices** are extracted from the dataset data.df and stored in the folds list.

1. Performing cross-validation:

Another **for** loop iterates over each fold (i) from 1 to **k**.

Inside the loop:

The training data is created by combining all folds except the current one **(folds[-i])** using **do.call(rbind, ...)**.

The validation data is set to the current fold (**folds[[i]]**).

The model is trained using the training data (**train\_data**) and evaluated using the validation data (**validation\_data**).

For demonstration purposes, the number of observations in each fold is printed.

This process ensures that each observation in the dataset is used for both training and validation exactly once, making the evaluation more robust and reliable.

+ print(paste("Number of observations in fold", i, "=", num\_observations))

+ }

[1] "Number of observations in fold 1 = 704"

[1] "Number of observations in fold 2 = 704"

[1] "Number of observations in fold 3 = 703"

[1] "Number of observations in fold 4 = 703"

[1] "Number of observations in fold 5 = 703"

[1] "Number of observations in fold 6 = 703"

[1] "Number of observations in fold 7 = 703"

[1] "Number of observations in fold 8 = 703"

[1] "Number of observations in fold 9 = 703"

[1] "Number of observations in fold 10 = 703"

Task 5

In Task 5 of the project, the goal was to create and evaluate two classification models for predicting customer churn, focusing specifically on the relationship between churn and customer tenure. The process began by utilizing R to handle the dataset, ensuring that the necessary libraries such as **dplyr**, **caTools**, **rpart**, and **rpart.plot** were loaded to facilitate data manipulation and analysis.

The initial step involved loading the data and selecting only the relevant 'Churn' and 'Tenure' columns. This subset of data was then divided into training and testing datasets using the **sample.split** function, with 80% of the data allocated for training to ensure a robust learning process. The decision tree model was constructed using the **rpart** function, focusing on 'Churn' as the dependent variable influenced by 'Tenure'. This model's simplicity and interpretability make it ideal for understanding the factors contributing to churn.

The decision tree, as visualized in the provided diagram, highlights critical tenure thresholds that significantly impact churn likelihood. For instance, customers with less than 17 months of tenure exhibit a higher propensity to churn, which decreases markedly as tenure increases beyond this point. This pattern is evident in the distribution of tenure by churn status shown in the histogram, where shorter tenure correlates with a higher churn rate.

Following the model construction, predictions were made on the testing dataset, and the model's accuracy was calculated to be approximately 75.23%. This performance metric indicates a relatively high level of predictiveness, considering the model's reliance on a single predictor variable.

The integration of these analyses—visualized through both the decision tree and tenure distribution histogram—provides a clear and actionable insight into how tenure influences customer churn. These visual tools are not only beneficial for the technical understanding of the model's functionality but also serve as excellent communication aids in presenting these findings to stakeholders, thereby bridging the gap between data science and business strategy implementation.

A graph of a number of months

Description automatically generated

A diagram of a tree

Description automatically generated

A screenshot of a computer code

Description automatically generated

Task\_6

**MODEL EVALUATION**

The provided R code uses a classification model to predict customer churn for a telco company, which aligns with the goal of proactive churn management. The model is made to predict and reduce customer turnover, which is a major concern for telecom firms because it affects revenue and profitability. It does this by utilizing the Naive Bayes algorithm.

The highlights:   
  
1. Model for Predicting Customer Churn: This model uses past data and customer behavior to identify which customers are likely to churn and to calculate the likelihood of turnover.

2. Methodology: The Naive Bayes technique is chosen because it works well with categorical data and provides a good strategy for the Telco Customer Churn dataset.   
  
3. Assessment of the Model: The model's performance is evaluated using custom metrics on training and validation sets, such as accuracy, sensitivity, specificity, precision, false discovery rate (FDR), false omission rate (FOR), area under the ROC curve (AUC), and lift.

4. Cross-Validation: Using a "for" loop, the "sample" function, and the "rbind" function, the project creates 10-fold cross-validation from scratch. Through data partitioning into several folds and iterative training and testing of the model on various subsets of the data, this technique ensures robust model evaluation.

"Model Evaluation and Performance Comparison"

1. Loading Libraries: This section loads the necessary libraries for the analysis, including caret, pROC, and e1071.

2. Reading Data: The dataset is read from the CSV file into R dataframe data.df.

3. Converting Target Variable: The target variable Churn is converted into a factor variable, which is necessary for modeling in R.

4. Partitioning Data: The dataset is split into training and validation sets using the createDataPartition function from the caret package. 70% of the data is allocated to the training set and the remaining 30% to the validation set.

5. Building Naive Bayes Model: A Naive Bayes classification model is built using the naiveBayes function from the e1071 package. The target variable Churn is predicted based on all other variables in the dataset.

6. Applying Model to Training and Validation Sets: The trained Naive Bayes model is applied to both the training and validation sets to generate predictions.

7. Defining Evaluation Metrics Function: A custom function calculate\_metrics is defined to compute various evaluation metrics such as accuracy, sensitivity, specificity, precision, false discovery rate (FDR), false omission rate (FOR), area under the ROC curve (AUC), and lift.

8. Calculating Metrics: The calculate\_metrics function is applied to both the training and validation predictions to compute evaluation metrics.

9. Displaying Metrics: Finally, the evaluation metrics for both training and validation sets are printed to the console.

Interpretation:

Accuracy: The proportion of correctly classified instances. In training, it's around 68% and in validation, it's around 73%.

Sensitivity: Also known as recall or true positive rate, it measures the proportion of actual positives that are correctly identified as such. In training, it's about 64%, and in validation, it's about 69%.

Specificity: It measures the proportion of actual negatives that are correctly identified as such. In training, it's around 79%, and in validation, it's around 82%.

Precision: The proportion of true positives among the instances that were classified as positive. In training, it's around 90%, and in validation, it's around 91%.

FDR: False Discovery Rate, the proportion of false positives among the instances that were classified as positive. In training, it's around 10%, and in validation, it's around 9%.

FOR: False Omission Rate, the proportion of false negatives among the instances that were classified as negative. In training, it's around 56%, and in validation, it's around 51%.

AUC: Area under the ROC curve, a measure of the classifier's ability to discriminate between positive and negative classes. In training, it's around 0.72, and in validation, it's around 0.76.

Lift: The ratio of the model's performance to a random classifier. In both training and validation, it's around 2.4 and 2.6, respectively.

OUTPUTS:

1] "Training Metrics:"

> print(training\_metrics)

$Accuracy

Accuracy

0.6801866

$Sensitivity

Sensitivity

0.6394257

$Specificity

Specificity

0.7929717

$Precision

Pos Pred Value

0.8952455

$FDR

Pos Pred Value

0.1047545

$FOR

Neg Pred Value

0.5571672

$AUC

Area under the curve: 0.7162

$Lift

Sensitivity

2.408715

[1] "Validation Metrics:"

> print(validation\_metrics)

$Accuracy

Accuracy

0.7267992

$Sensitivity

Sensitivity

0.6945876

$Specificity

Specificity

0.8160714

$Precision

Pos Pred Value

0.9127858

$FDR

Pos Pred Value

0.08721423

$FOR

Neg Pred Value

0.50913

$AUC

Area under the curve: 0.7553

$Lift

Sensitivity

2.619588