ChurnEra: Transforming Telco Customer Relationships through Predictive Modeling

MATH 40028/50028: Statistical Learning

March 13, 2024

The telecommunications industry is constantly evolving, with rapid changes in consumer preferences. One critical aspect for sustainable business success is identifying and addressing customer attrition, also known as customer churn, where subscribers discontinue using services. Churn not only impacts revenue but also reflects market competitiveness and customer satisfaction levels.

This project aims to conduct a thorough analysis of Telco customer churn, focusing on identifying key factors that influence consumer decisions in telecommunications services. By using predictive analytics, models will be developed to predict customer churn and provide actionable insights for retention strategies. This will help Telco companies proactively manage churn rates, enhance customer satisfaction, and sustain long-term business growth.

Dataset Overview:-

The Telco customer churn dataset is a comprehensive collection of data points that provide valuable insights into customer behavior within a telecommunications company. It consists various aspects of customer interactions, services availed, and most importantly, whether customers have churned or not. The dataset serves as the foundation for our analysis and prediction endeavors, enabling us to delve deep into understanding the factors influencing churn and providing actionable strategies for customer retention.

The dataset comprises several key variables that contribute to a comprehensive understanding of customer behavior:

CustomerID, Gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn

Sampling Considerations and Target Population Analysis:-

The target population is the Telco customers, specifically those who may potentially churn.Sampling strategies like Stratified random sampling is used to partition the dataset into training and testing sets.Which ensures a balanced representation of classes in training and testing sets, mitigating class imbalances.Potential biases could arise from sampling methods mainly if the dataset is imbalanced i.e., there are significantly more non-churners than churners or vice versa,it could affect the model's performance and interpretation of the results.Feature engineering and model selection process plays a crucial role in addressing bias. However, rigorous statistical techniques can mitigate biases, ensuring robust predictive models.

Prediction Problems:-

=>Customer Churn Prediction:Predicting whether a customer will churn (cancel their subscription or leave the service) based on various customer attributes and behaviors and assessing the performance of various machine learning models in predicting customer churn.

=>Feature Engineering and Selection Strategies:Enhancing model performance by employing techniques such as selecting important features and creating new features that are more predictive of the target variable (churn).

=>Model Tuning and Optimization:Fine-tuning hyperparameters of models to achieve better performance and generalization on unseen data.

=>Resampling: Employing resampling techniques to assess the stability and reliability of statistical learning models by repeatedly drawing samples from the dataset.

=>Comparative Analysis of statistical learning models:Evaluating and comparing the effectiveness of different statistical learning algorithms for churn prediction.

Data Splitting and Utilization Strategy:-

The following steps provides an outline about splitting the data and other plans to use the data effectively:-

1.Stratified Random Sampling

2.Exploratory Data Analysis

3.Feature Engineering

4.Model Selection and Evaluation

5.Hyperparameter Tuning

6.Feature Selection

7.Model Validation

Understanding the Data: Exploratory Data Analysis (EDA):-

Exploratory data analysis for the training set undergoes following steps:-

->Handling missing values to ensure data completeness for further analysis.

->Detecting outliers by employing techniques such as box plots to identify outliers,considering their impact on the analysis and decision on treatment.

->Utilizing visualization techniques like histograms,pie chart and bar plots to understand the distribution and class balance of the target variable.

-> Applying chi-square test to assess the association between categorical variables and the target variable, understanding their significance in predicting churn.

->Using correlation analysis to investigate the association between numeric variables, evaluating their relevance in predicting churn.

->Calculating summary statistics such as mean, median, and quartiles for numeric variables and frequencies for categorical variables, aiding in understanding data distribution.

->Exploring opportunities for feature engineering, creating new variables or transformations to enhance the predictive power of the models.

->Visualizing the associations of interaction terms with the target variable, exploring potential synergies or impacts on churn prediction.

Statistical Learning Approaches:-

Logistic Regression, Random Forest, Support Vector Machines(SVM), Linear Discriminant Analysis (LDA), Decision Trees, Naive Bayes.

Feature Engineering Strategies:-

Feature generation(creating new variables or combining existing ones), Principle Component Analysis(Dimensionality Reduction Technique), Feature Selection(random forest feature importance,forward/backward selection).

Advanced Model Tuning Strategies:-

Regularization , Cost Optimization , Hyperparameter Tuning , Pruning.

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.

Model Assumptions and Applicability to Prediction:-

Statistical learning methods like Logistic Regression, Random Forest, SVM, LDA, Decision Tree, and Naive Bayes are applicable to customer churn prediction due to their ability to handle binary classification tasks. Feature engineering techniques enhance model performance by capturing relevant churn indicators. Model tuning optimizes algorithms for better predictive accuracy and generalization. Resampling assesses model stability. Comparative analysis aids in selecting the most effective algorithm for churn prediction based on metrics like accuracy, precision, recall, F1-score, and AUC. These methods collectively offer a comprehensive approach to address the complex nature of customer churn prediction.

Packages Utilized:-

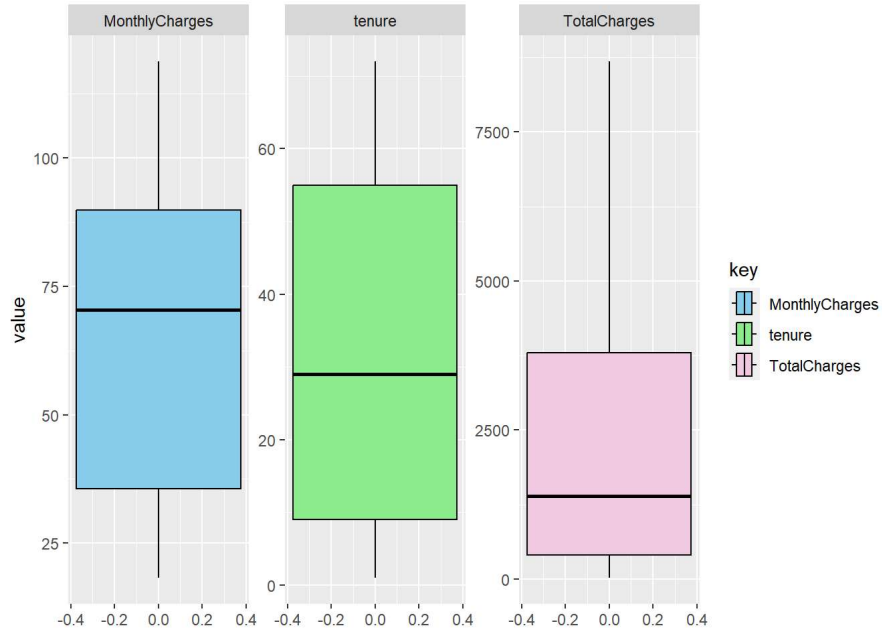
library(tidyverse),library(readxl),library(dplyr),library(gridExtra),library(corrplot),library(pROC),library(caret),library(rsample),library(reshape2),library(ggplot2),library(t

Exploring Telco Customer Churn Dataset:-

gender <fct>	SeniorCitizen <fct>	Partner <fct>	Dependents <fct>	tenure <dbl>	PhoneService <fct>	MultipleLines <fct>	InternetService <fct>
Female	0	Yes	No	1	No	No	DSL
Male	0	No	No	34	Yes	No	DSL
Male	0	No	No	2	Yes	No	DSL
Male	0	No	No	45	No	No	DSL
Female	0	No	No	2	Yes	No	FiberOptic
Female	0	No	No	8	Yes	Yes	FiberOptic

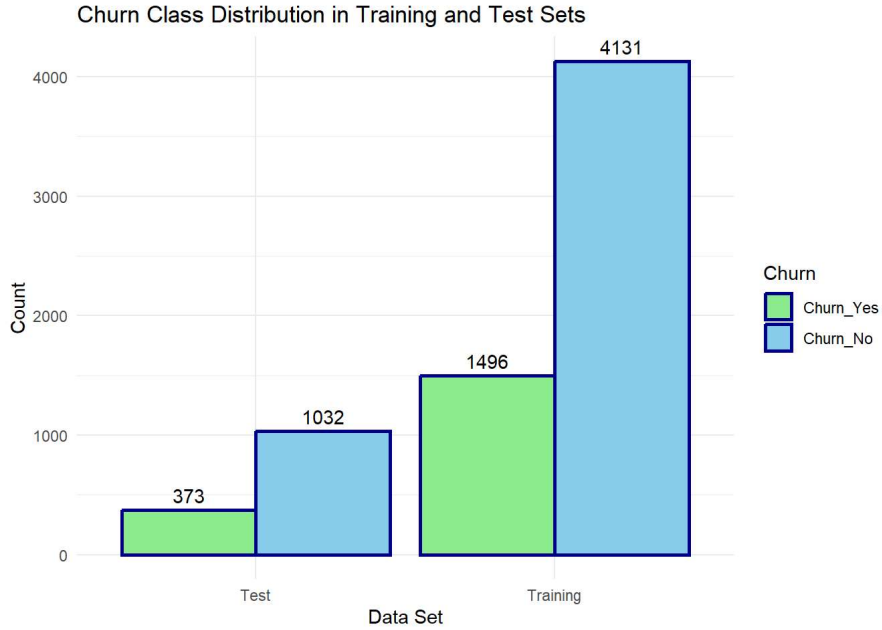
6 rows | 1-8 of 20 columns

Outlier Detection and Boxplot Analysis for Numerical Variables:-



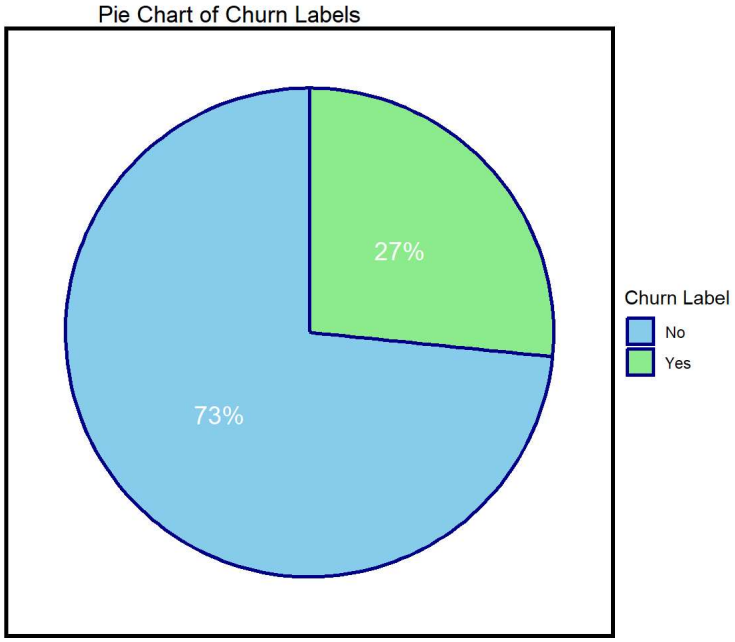
Data Partitioning/splitting Using Stratified Random Sampling:-

Stratified sampling for data splitting ensures a balanced representation of classes in training and testing sets, mitigating class imbalances. It partitions data into 80% training and 20% testing, maintaining proportional class distributions to improve model performance.



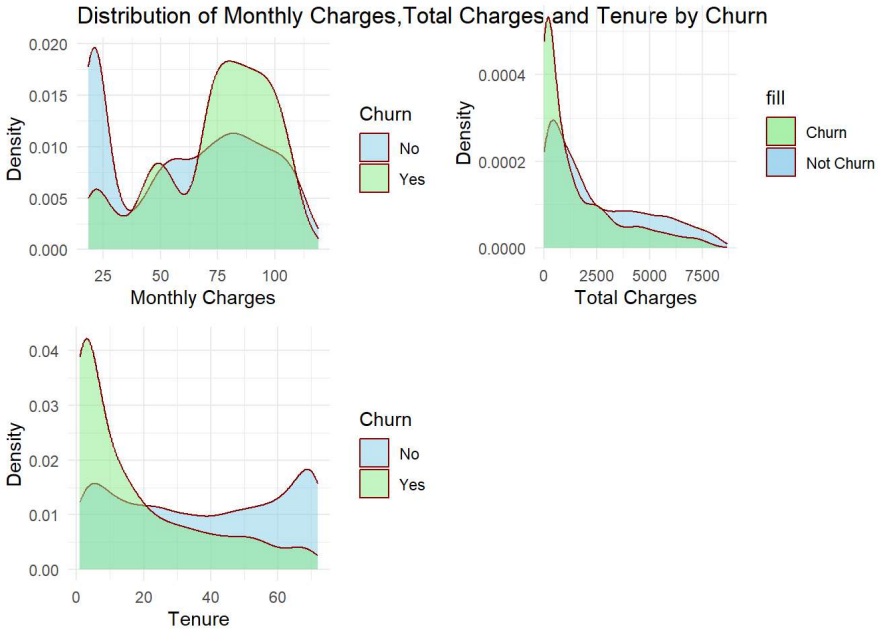
The bar plot illustrates the distribution of churn (yes/no) in the training and test sets after stratified sampling. It shows a proportional representation of churners and non-churners in both sets, ensuring a balanced dataset for training predictive models.

Exploratory data analysis using the training set:-



The pie chart visually depicts that 27% of customers have experienced churn, whereas the larger portion, comprising 73%, has not churned.

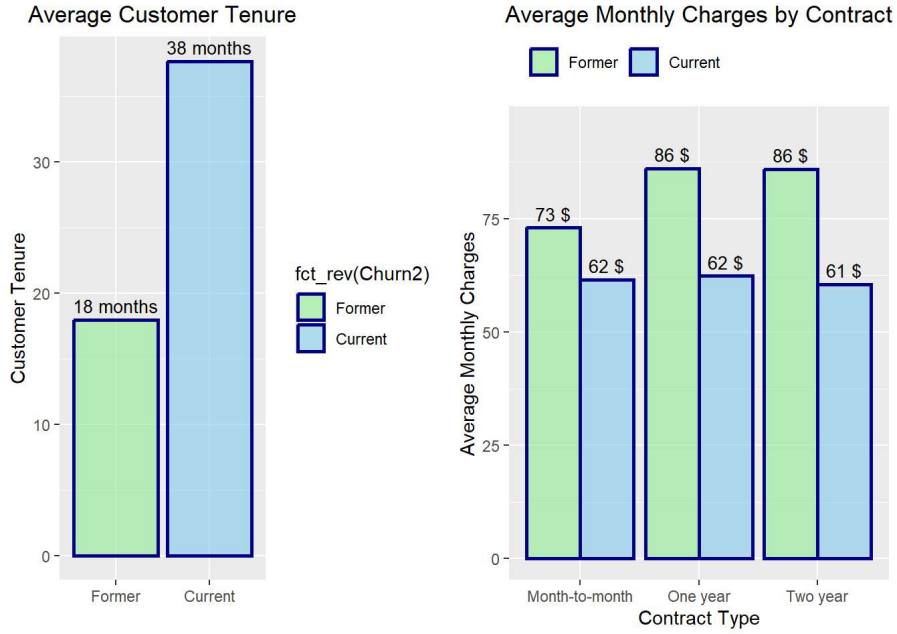
Exploring Associations Between Numeric Variables and Churn:-



The density plot reveals the following insights:

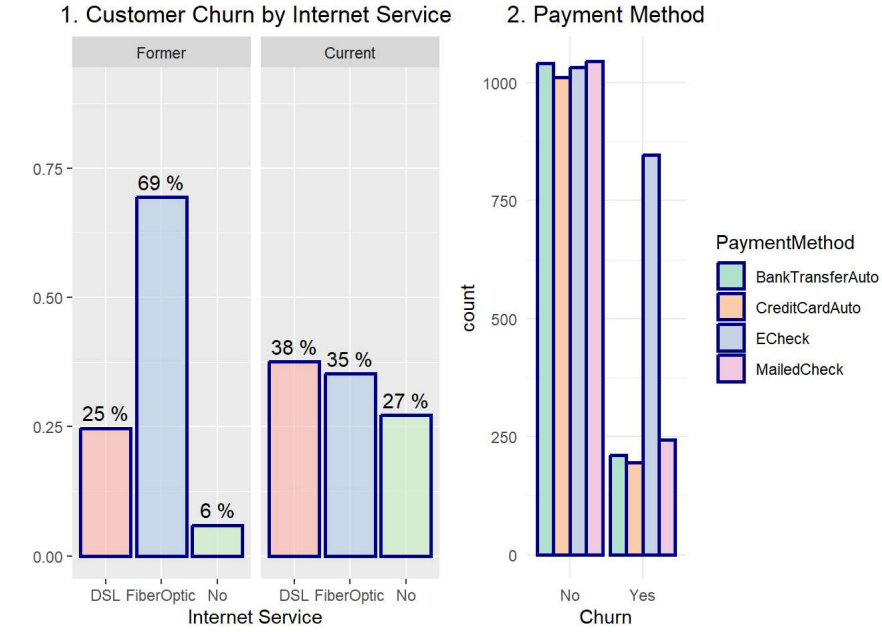
- Shorter Tenure: Churned customers tend to have shorter tenure periods, with a concentration towards lower values.
- Higher Monthly Charges: The distribution of monthly charges is skewed towards higher values among churned customers.
- Lower Total Charges: Churned customers exhibit lower total charges, suggesting a tendency for early exits before accumulating significant total charges.

These findings suggest that customers with shorter tenure, higher monthly charges, and lower total charges are more prone to churning.



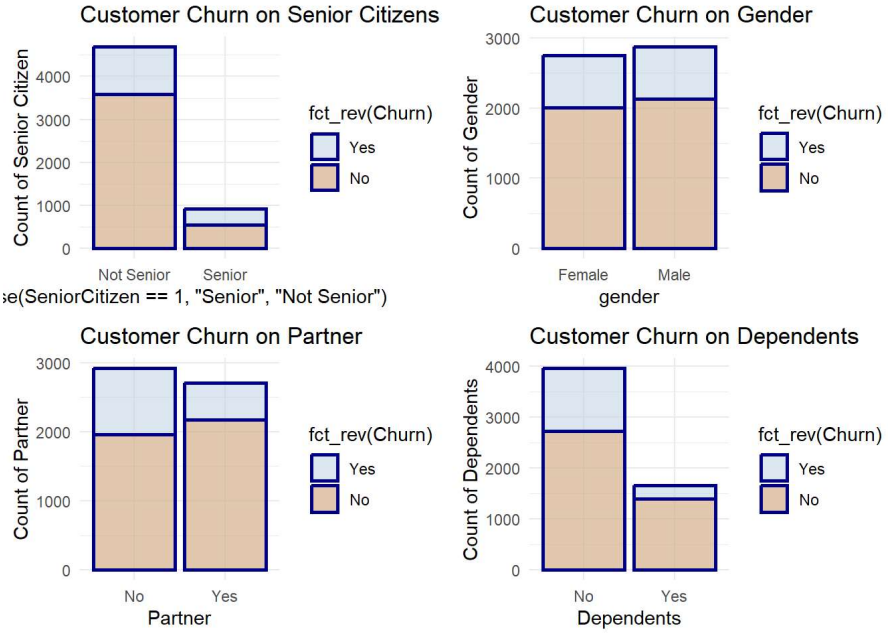
The visualizations illustrate the average tenure and monthly charges for existing and former Telco customers. On average, current customers have a service tenure of slightly over three years, while former customers stayed for around 18 months. Moreover, there's a trend suggesting lower average monthly charges for current customers across various contract durations, possibly indicating better pricing for recent customers, particularly those on longer contracts.

Exploring Associations Between Categorical Variables and Churn:-



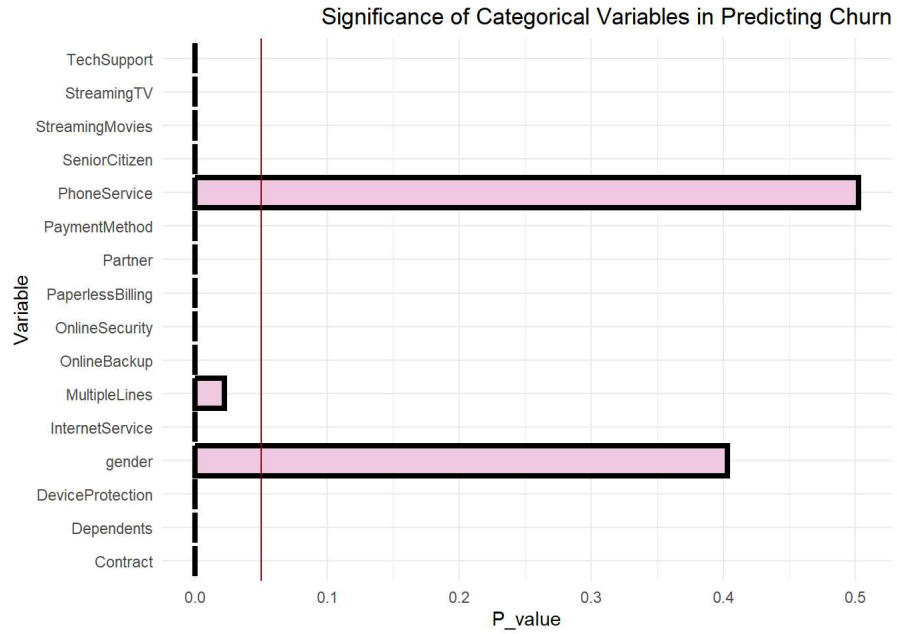
1.The initial visualization reveals that a considerable 69% majority of customers who discontinued their internet service had selected Fiber Optic. In contrast, 25% had opted for DSL, and a smaller fraction of 6% did not have any Internet service. Among existing customers, DSL is the most popular choice at 38%, followed by Fiber Optic at 35%. Notably, 27% of current customers lack an Internet service subscription. The elevated presence of Fiber Optic among departing customers might suggest dissatisfaction or concerns related to this particular service option.

2.The second visualization indicates that the number of customers who haven't churned remains fairly consistent across all four payment methods. On the other hand, churned customers predominantly chose the electronic check payment method.



Demographic elements like being a Senior Citizen, having a Partner, and having Dependents are likely influential in customer churn dynamics. Conversely, gender does not seem to be a decisive factor in churning behaviors, as indicated by the trends seen in the bar charts.

Assessing the Significance of Categorical Variables in Predicting Churn Using Chi-Square Test:-



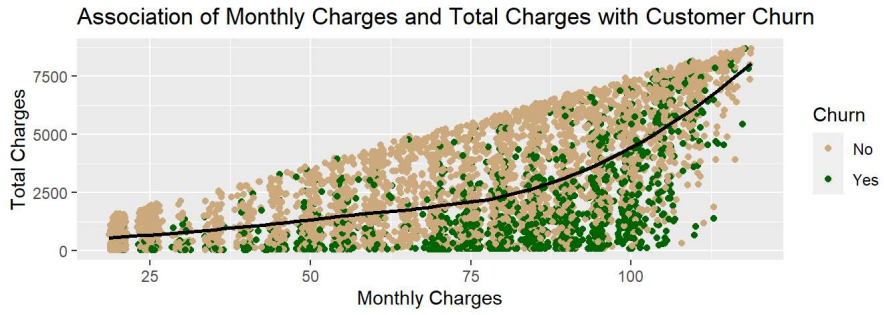
The examination of the telco customer churn dataset reveals valuable findings. Most variables, excluding ‘Gender’ and ‘PhoneService,’ demonstrate a significant statistical correlation with customer churn, as indicated by p-values below the commonly accepted threshold of 0.05. This suggests the potential importance of these variables in predicting churn patterns.

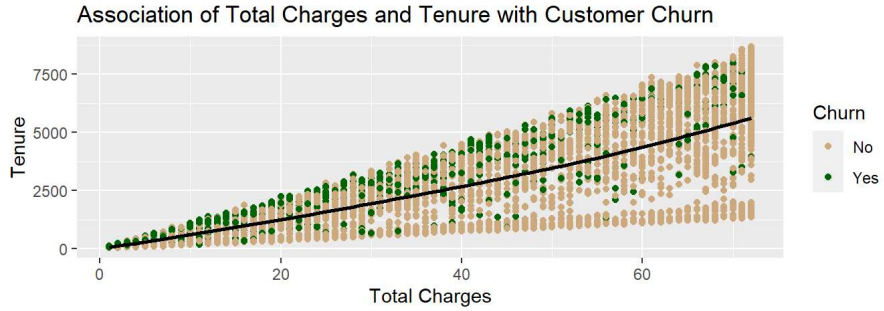
Additional investigation and modeling could be necessary to delve deeper into these correlations and grasp the interaction between the categorical variables and customer churn more comprehensively.

Correlation Analysis of Numeric Variables:-

The analysis reveals interesting relationships within the telco customer churn dataset. Customer tenure shows a weak positive correlation (0.25) with monthly charges, indicating a tendency for charges to rise with longer subscriptions. A strong positive correlation (0.83) exists between tenure and total charges, indicating higher costs for longer subscriptions. Additionally, a moderate positive correlation (0.65) links monthly charges to total charges, suggesting higher monthly charges result in higher overall costs. This correlation matrix illuminates numeric variable interactions, warranting further modeling to delve into categorical variable impacts on churn.

Smoothed Scatterplots Showing Associations with Customer Churn-Interaction terms:-



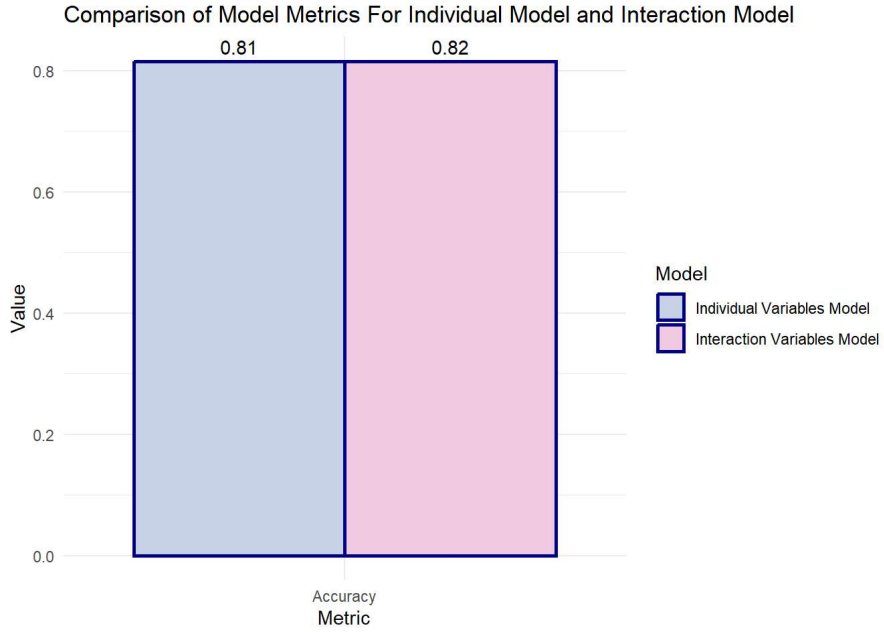


The data suggests a clear positive correlation between monthly charges and total charges, where an increase in monthly charges is linked to a rise in total charges. This trend is indicative of a higher probability of customer churn.

Predictive analysis:-

Comparing And Assessing The Effectiveness Of Model Performance: Traditional(Individual) Predictors vs. Interaction Features based on Correlation Analysis-Feature Generation:-

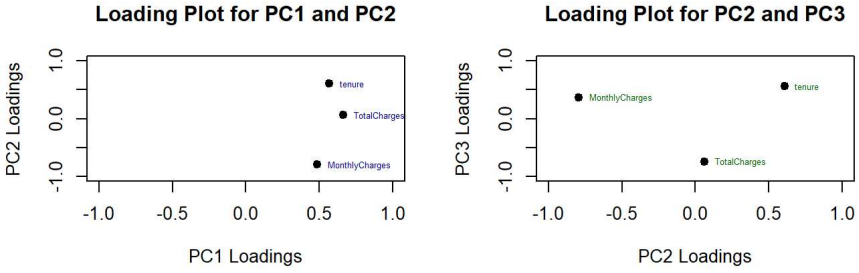
Model	Accuracy	Precision
Logistic Regression Model(individual) variables	0.81	0.68
Logistic Regression Model (Interaction) variables	0.82	0.69

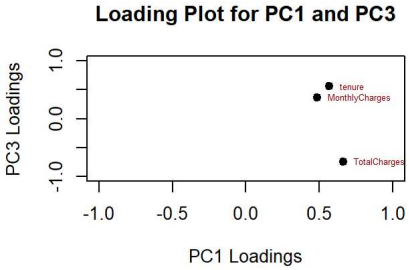


The bar plot comparing the individual model and the interaction model reveals interesting insights. While both models show high accuracy rates of around 0.81 for the individual model and around 0.82 for the interaction model, the slight improvement in accuracy with the interaction model suggests that considering the interaction effects among variables contributes to better predictive performance.

Performing Principal Component Analysis(PCA)-Dimensionality reduction technique:-

```
## Importance of components:
##                           PC1    PC2     PC3
## Standard deviation     1.4771 0.8713 0.24279
## Proportion of Variance 0.7273 0.2531 0.01965
## Cumulative Proportion  0.7273 0.9804 1.00000
```





PC1, with a standard deviation of 1.477, represents a combination of information from tenure, MonthlyCharges, and TotalCharges.A high positive loading for TotalCharges in PC1 suggests that customers with higher total charges tend to exhibit characteristics captured by PC1, which could include longer tenure and higher monthly charges. PC2, with a standard deviation of 0.871, shows a contrast between tenure and the charges a positive loading for tenure and negative loadings for the charges indicate a pattern where longer tenure is associated with lower monthly and total charges. PC3, with a standard deviation of 0.243, reveals an orthogonal relationship between tenure and TotalCharges.Positive loading for tenure and negative loading for TotalCharges suggest an opposite influence; as tenure increases, TotalCharges tend to decrease.

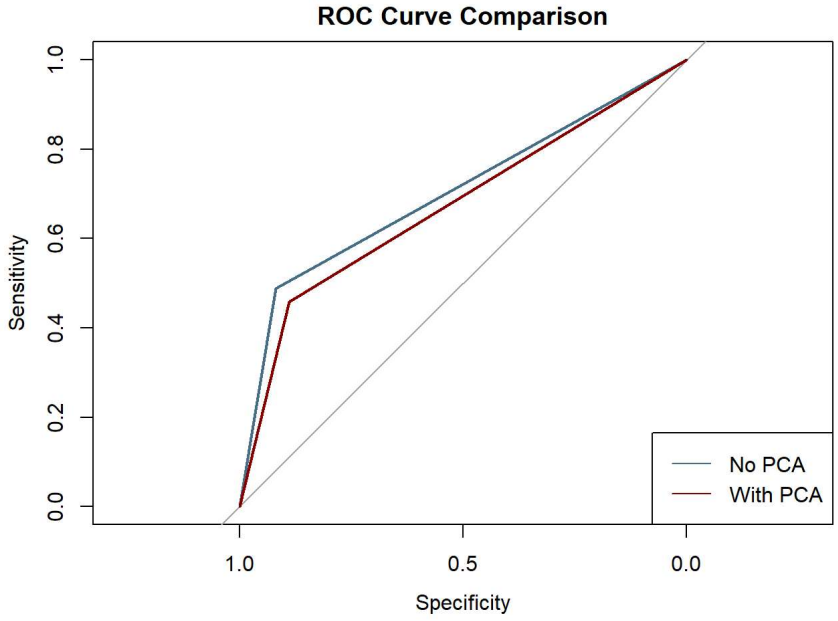
Comparison of Random Forest Models with and without PCA:-

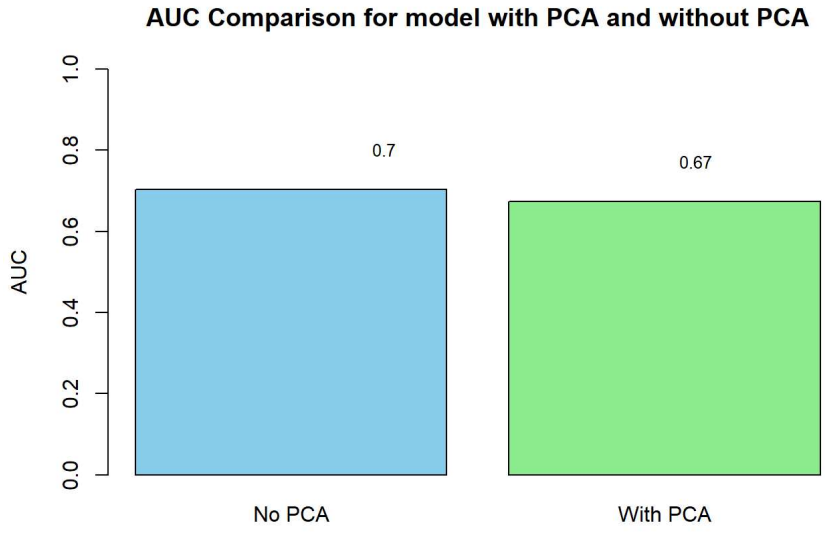
Model	Accuracy	Precision
Random Forest Model without PCA	0.80	0.66
Random Forest Model with PCA	0.78	0.61

The comparison of Random Forest models with and without Principal Component Analysis (PCA) reveals insights into their predictive performance. The model without PCA shows higher accuracy and precision, indicating better overall predictive capability. However, the model with PCA demonstrates slightly lower accuracy and precision, suggesting potential trade-offs between dimensionality reduction and predictive accuracy. To further enhance model performance, hyperparameter tuning is planned, leveraging cross-validation and a grid search approach to optimize the Random Forest algorithm's parameters. This iterative process aims to improve model accuracy and precision, addressing potential limitations observed in the initial model evaluations.

Optimizing Random Forest Models with Hyperparameter Tuning and Cross-Validation: A Comparative Analysis of PCA vs. Non-PCA Approaches for Churn Prediction:-

Model	Accuracy	Precision	Recall	F1_Score	AUC
Tuned Random Forest Model without PCA(5-foldCV)	0.80	0.83	0.92	0.87	0.70
Tuned Random Forest Model with PCA(5-foldCV)	0.78	0.82	0.89	0.85	0.67



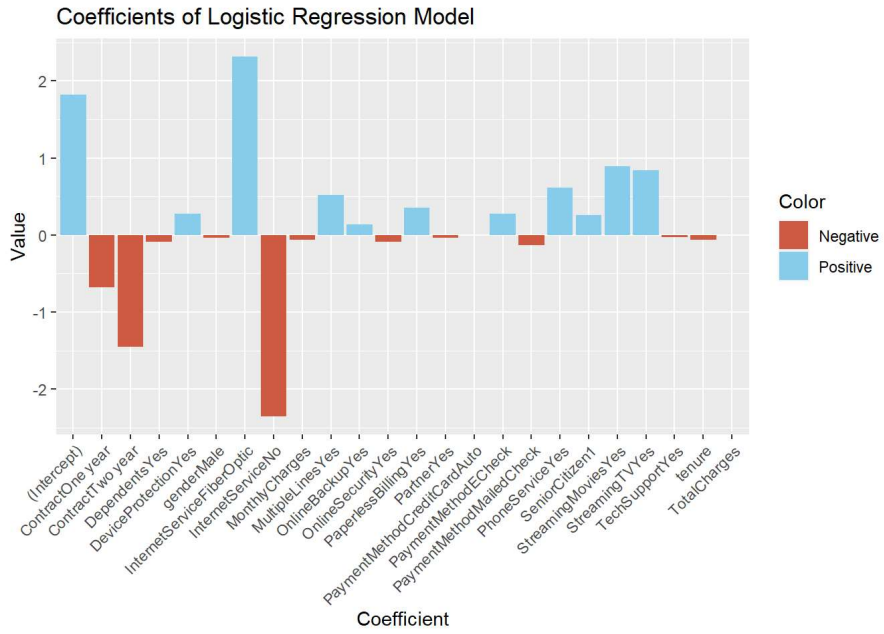


The tuned Random Forest models, both with and without PCA, exhibit respectable performance. Without PCA, the model achieved an accuracy of approximate 80%, demonstrating robustness in predicting churn. The precision and recall values for the positive class (churn) indicate a high proportion of correctly identified churn cases among predicted churn instances, with an F1 score of 0.873, reflecting a balance between precision and recall. The Area Under the Curve (AUC) value of around 70% signifies a good ability to discriminate between churn and non-churn instances. However, with PCA, while the accuracy slightly decreased to around 77%, the precision, recall, and F1 score remained high, indicating consistent performance. The AUC value of Around 67% suggests slightly reduced discriminative ability compared to the model without PCA but still demonstrates reasonable predictive power. Both models are good choices, but the PCA model stands out for reducing complexity while still predicting well.

Logistic Regression Model with Forward Selection: Predicting Churn with Subset Selection-Feature Selection:-

Model	Accuracy	Precision	Recall	F1_Score	AUC
Logistic Regression Model with Forward Selection	0.81	0.68	0.58	0.62	0.74

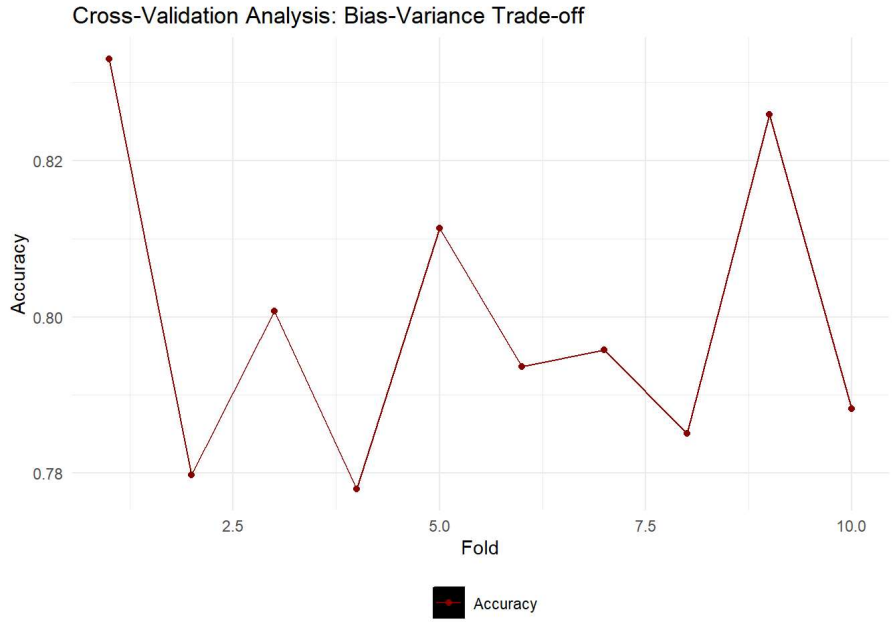
The logistic regression model, constructed using forward selection, highlights key predictors influencing customer churn. Significant variables such as SeniorCitizen, Partner, tenure, MultipleLines, InternetService, Contract, and PaymentMethod play crucial roles in predicting churn likelihood. The model achieved an accuracy of around 81%, with notable precision (68%) and recall (58%). These metrics indicate the model's ability to correctly identify churn cases while minimizing false positives. The AUC score of 0.74 further confirms the model's reasonable discriminatory performance. Overall, the model provides valuable insights into factors driving customer churn in the telecommunications industry.



The plot illustrates the impact of various predictor variables on customer churn prediction in the telecommunications sector using logistic regression. Red bars indicate factors reducing churn likelihood, such as longer contract durations ("Contract_Two year" and "Contract_One year"). In contrast, blue bars signify variables increasing churn risk, like "InternetService_Fiber optic," suggesting fiber optic users are more likely to churn. Bar lengths reflect the relative influence strengths of predictors on customer retention or attrition

Cross-Validation Performance Metrics for Logistic Regression with Forward Selection:-

Model	Accuracy	Precision	Recall	F1_Score
Logistic Regression Model with Forward Selection(10-foldCV)	0.8	0.64	0.56	0.69



The bias-variance trade-off plot illustrates how the accuracy of our logistic regression model changes with varying levels of complexity, represented by the number of folds in cross-validation. As the complexity increases, initially the model shows high variance (overfitting), leading to reduced accuracy. However, at an optimal complexity level, around the middle folds in this case, we achieve the highest accuracy, indicating a good

balance between bias and variance. Beyond the point, increasing complexity leads to decreased accuracy due to higher bias (underfitting). This plot guides us in selecting an appropriate model complexity that maximizes accuracy while avoiding overfitting or underfitting issues.

Linear Discriminant Analysis (LDA) Model Evaluation:-

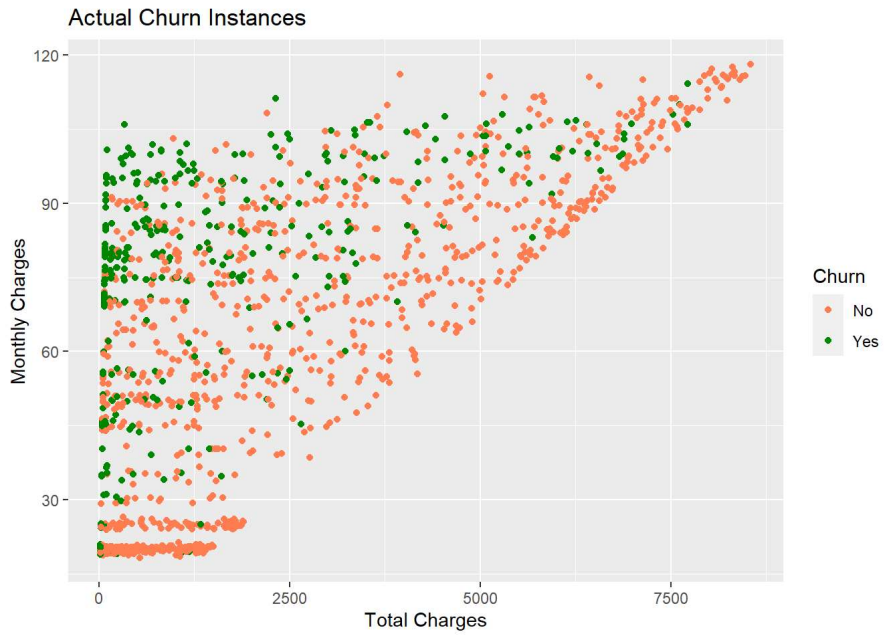
```
## Accuracy: 0.8049822
```

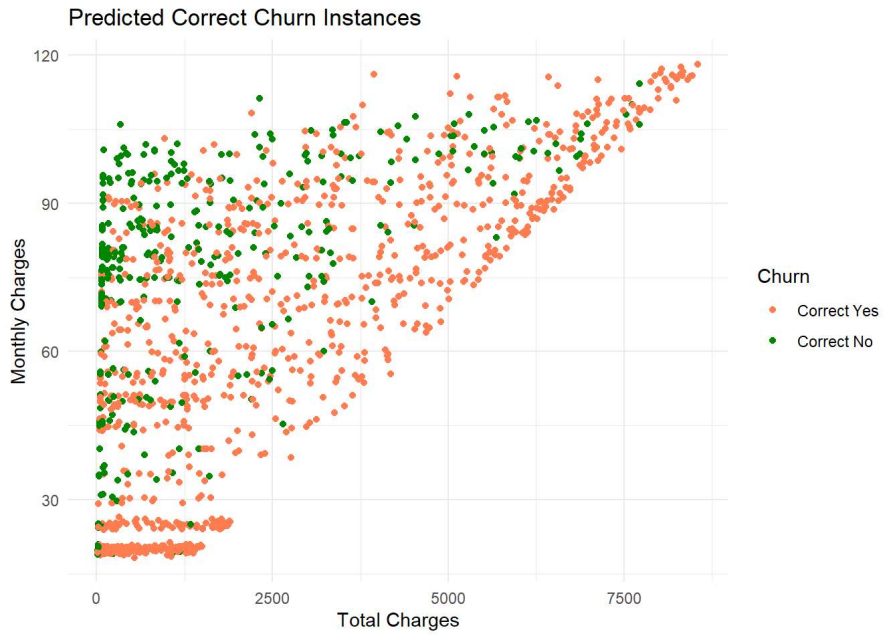
```
## Precision: 0.6468843
```

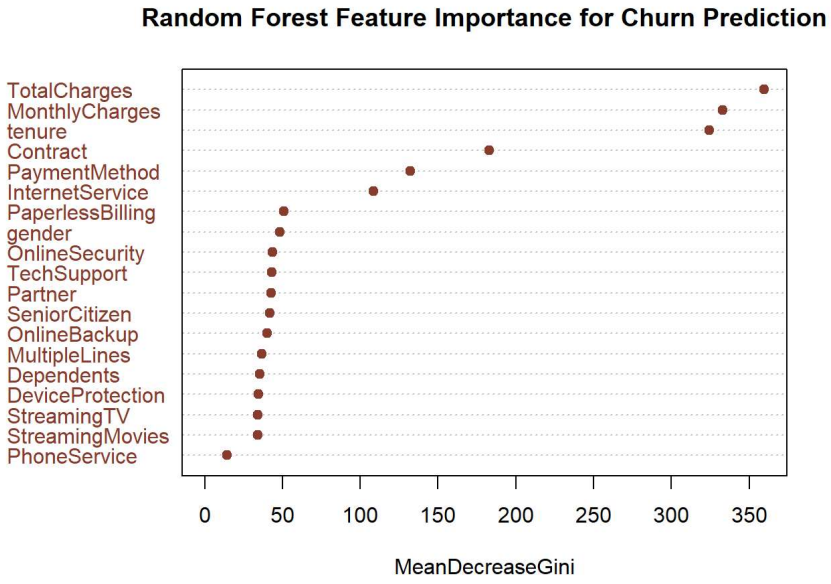
```
## Recall: 0.5844504
```

```
## F1 Score: 0.6140845
```

```
##          Reference
## Prediction  No Yes
##        No  913 119
##        Yes 155 218
```







Based on Random Forest feature importance, key predictors like Contract, TotalCharges, PaymentMethod, MonthlyCharges, InternetService, and Tenure significantly impact churn prediction. By selecting these top important features, an accurate Linear Discriminant Analysis (LDA) model can be built for effective churn prediction in telecom data.

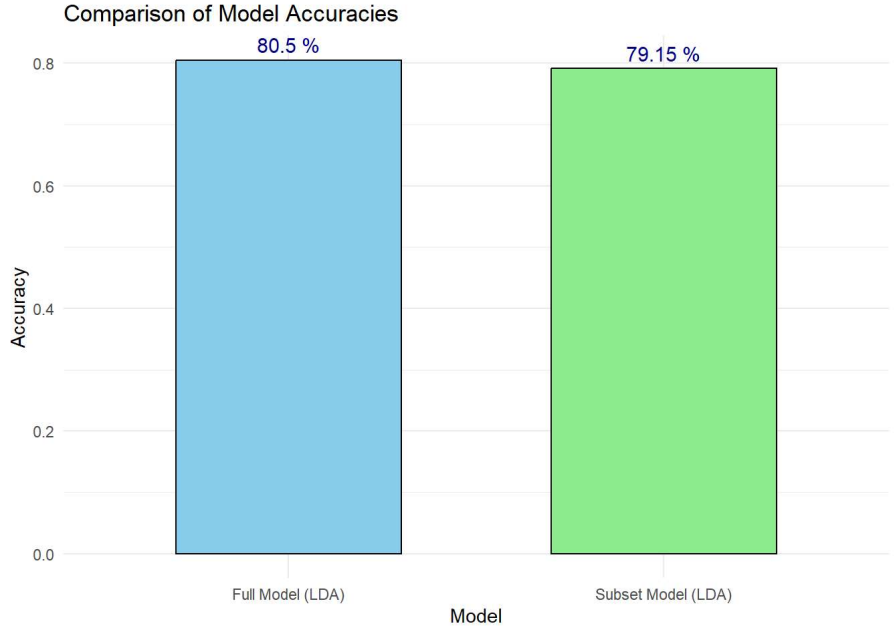
Linear Discriminant Analysis (LDA) Model with Important Features-Subset Selection:-

```
## Accuracy with Selected Features (LDA): 0.7914591
```

```
## Precision with Selected Features (LDA): 0.630719
```

```
## Recall with Selected Features (LDA): 0.5174263
```

```
## F1 Score with Selected Features (LDA): 0.5684831
```



The full Linear Discriminant Analysis (LDA) model achieved an accuracy of around 81%, indicating strong predictive performance across all features. The subset LDA model, focusing on important variables obtained from random forest feature importance, achieved a slightly lower accuracy of around 79%. This suggests that while the subset model maintains high accuracy, there is a minor trade-off in predictive power compared to the full model.

SVM Model for Churn Prediction: Evaluating Performance and Discriminatory Power:-

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  961 198
##        Yes  71 175
##
##                Accuracy : 0.8085
##                  95% CI : (0.787, 0.8288)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 5.054e-11
##
##                   Kappa : 0.4492
##
##  Mcnemar's Test P-Value : 1.562e-14
##
##             Sensitivity : 0.9312
##             Specificity : 0.4692
##          Pos Pred Value : 0.8292
##          Neg Pred Value : 0.7114
##              Prevalence : 0.7345
##          Detection Rate : 0.6840
##    Detection Prevalence : 0.8249
##       Balanced Accuracy : 0.7002
##
##        'Positive' Class : No
##
```

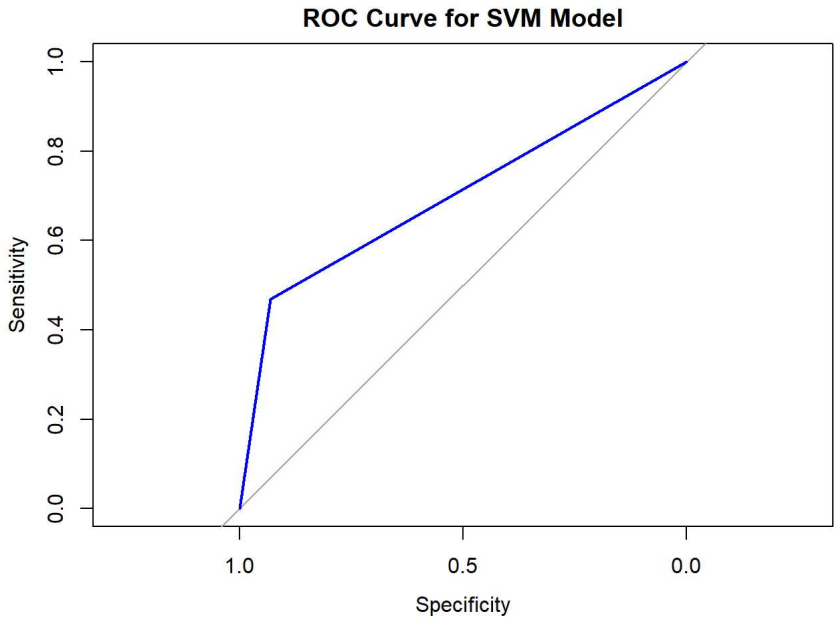
```
## Accuracy: 0.8085409
```

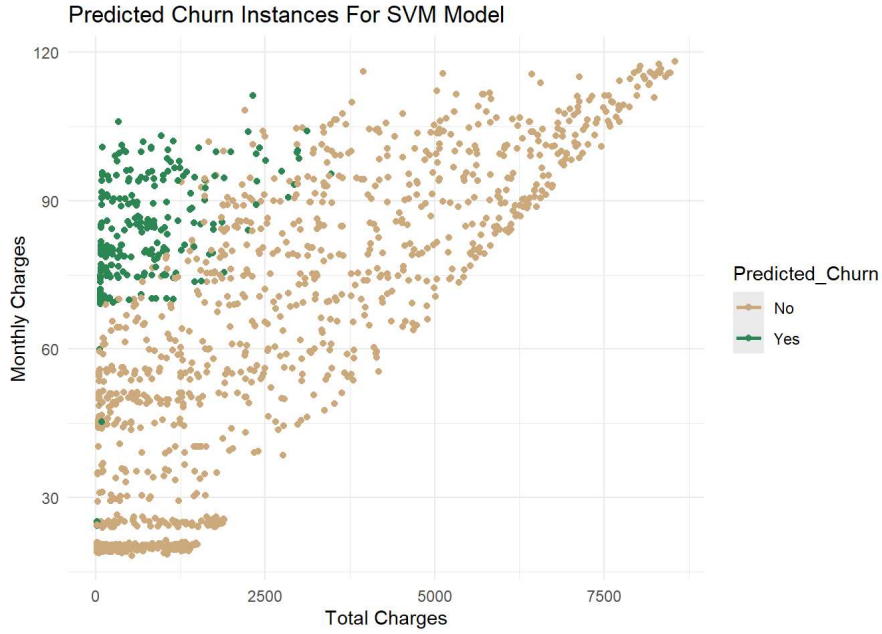
```
## Precision: 0.8291631
```

```
## Recall: 0.9312016
```

```
## F1-score: 0.877225
```

```
## AUC: 0.7001852
```





The SVM model demonstrates decent performance, with an accuracy around 81%, showcasing its ability to predict churn effectively. Precision stands at approximately 83%, indicating that when it predicts churn, it is correct about eight times out of ten. The recall, roughly 93%, suggests that the model captures a vast majority of actual churn cases. The F1-score, at about 87%, reflects a good balance between precision and recall. The AUC, around 70%, signifies a moderate ability of the model to distinguish between churn and non-churn instances.

Enhanced Churn Prediction: Regularized SVM Model with Cost Optimization and Cross-Validation:-

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  961 198
##        Yes  71 175
##
##                Accuracy : 0.8085
##                  95% CI : (0.787, 0.8288)
##     No Information Rate : 0.7345
##     P-Value [Acc > NIR] : 5.054e-11
##
##                   Kappa : 0.4492
##
##  Mcnemar's Test P-Value : 1.562e-14
##
##             Sensitivity : 0.9312
##             Specificity : 0.4692
##          Pos Pred Value : 0.8292
##          Neg Pred Value : 0.7114
##              Prevalence : 0.7345
##          Detection Rate : 0.6840
##    Detection Prevalence : 0.8249
##       Balanced Accuracy : 0.7002
##
##        'Positive' Class : No
##
```

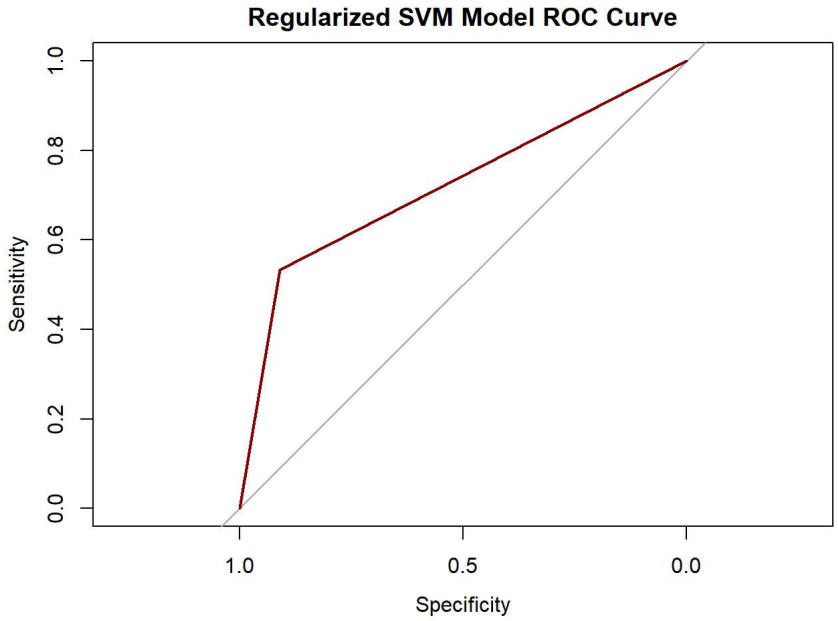
```
## Accuracy after regularization: 0.8099644
```

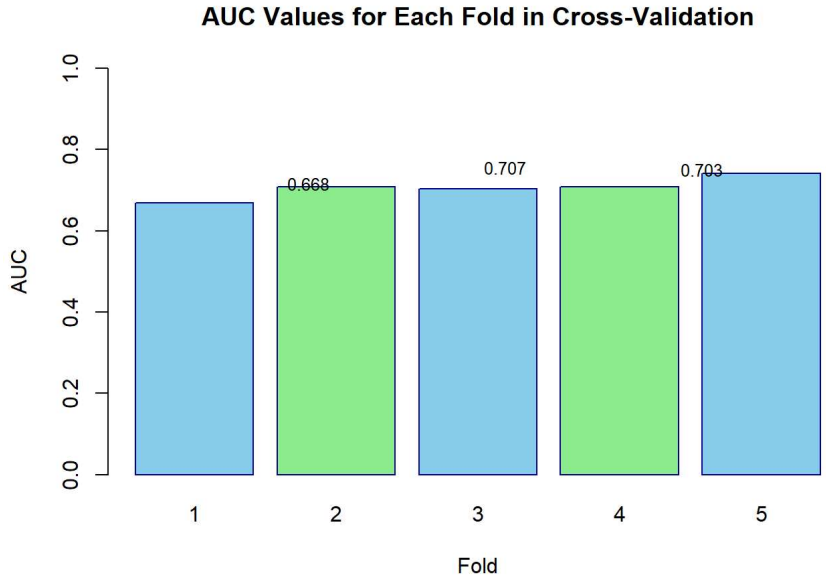
```
## Precision after regularization: 0.8291631
```

```
## Recall after regularization: 0.9312016
```

```
## F1-score after regularization: 0.877225
```

```
## AUC after regularization: 0.7216979
```






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

```
## Average Accuracy: 0.8011449
```

```
## Average Precision: 0.9097078
```

```
## Average Recall: 0.834477
```

```
## Average F1-score: 0.8704439
```

```
## Average AUC: 0.7055462
```

The SVM model with regularization,cost optimization and cross-validation significantly improved performance,indicating better overall predictive capability compared to the non-regularized SVM model. With cross-validation,it achieves approximately 80% accuracy, highlighting its ability to classify correctly. Moreover, with high precision around 91%, it's adept at identifying true positives. The model maintains a balance with recall at about 83%, implying it captures a good portion of actual positives. The F1-score, an overall measure, sits around 87%, indicating a strong balance between precision and recall. The AUC score of roughly 71% suggests a decent ability to distinguish between classes. Overall, these metrics signify a well-performing SVM model with optimized parameters and reliable generalization ability.

Decision Tree Model for Telecom Churn Prediction:-

```
## [1] "Accuracy: 0.792170818505338"
```

```
## [1] "Precision: 0.827433628318584"
```

```
## [1] "Recall: 0.906007751937985"
```

```
## [1] "F1 Score: 0.864939870490287"
```

```
## [1] "AUC: 0.737353177795656"
```

```
## [1] "Confusion Matrix:"
```

```
##          Reference
## Prediction  No Yes
##        No  935 195
##        Yes  97 178
```

Enhanced Decision Tree Model with Pruning and Regularization:-

```
## [1] "Accuracy: 0.795017793594306"
```

```
## [1] "Precision: 0.809484193011647"
```

```
## [1] "Recall: 0.942829457364341"
```

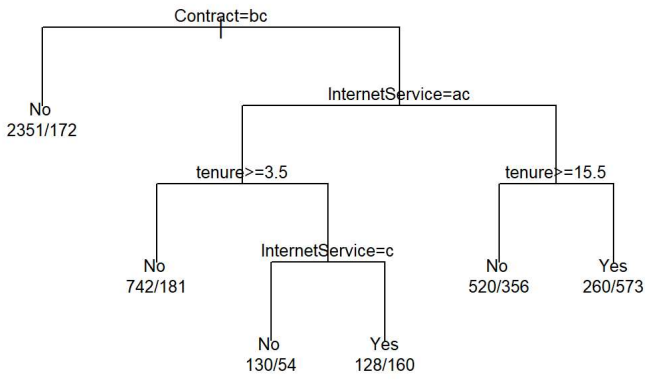
```
## [1] "F1 Score: 0.871083258728738"
```

```
## [1] "AUC: 0.759421899461489"
```

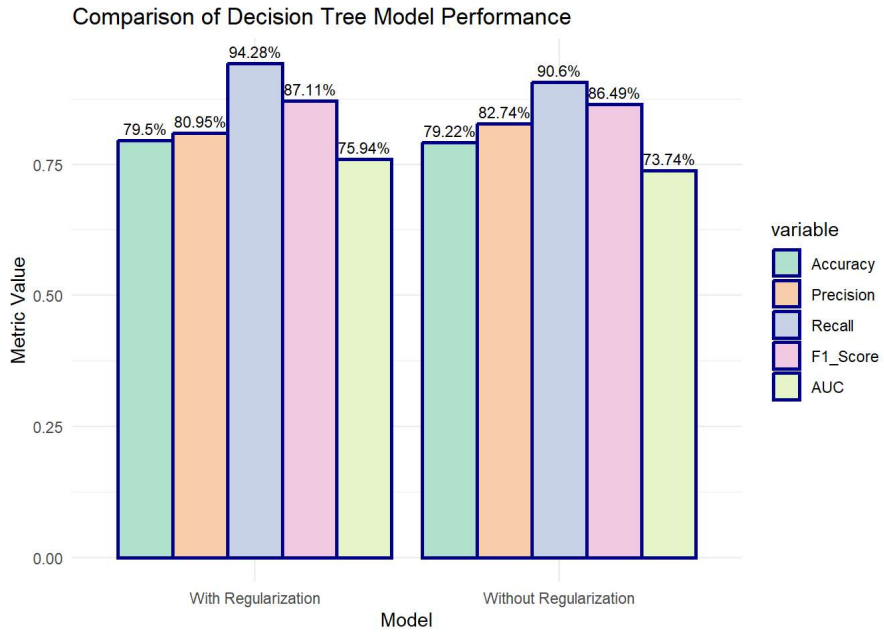
```
## [1] "Confusion Matrix:"
```

```
##          Reference
## Prediction  No Yes
##        No  973 229
##        Yes  59 144
```

Decision Tree Visualization:-



The decision tree plot illustrates the key factors influencing customer churn for a telecommunications company, focusing on a specific contract type (“b=one year,c=two year”). It splits the data based on internet service category (“a=DSL,c=No internet” or “c=No internet”) and customer tenure duration. The leaf nodes represent the predicted outcome - retaining the customer (“Yes”) or losing them (“No”), along with the corresponding data counts. Notably, customers with longer tenure periods and certain internet service packages appear more likely to be retained. This visual model can guide targeted retention strategies by identifying the feature combinations most correlated with churn risk.



The graph compares the performance of decision tree models with and without regularization techniques on the telco customer churn dataset. It evaluates various metrics such as accuracy, precision, recall, F1-score, and AUC. The model with regularization shows better performance across all metrics, with an accuracy of approx 80% , precision of around 81%, recall of around 94%, F1-score of around 87%, and AUC of around 76%. This indicates that regularization techniques like pruning can help prevent overfitting and improve the generalization ability of the decision tree model on this churn prediction task. The performance gains highlight the importance of appropriate regularization methods for building robust and accurate decision tree models.

Naive Bayes Model for Churn Prediction:-

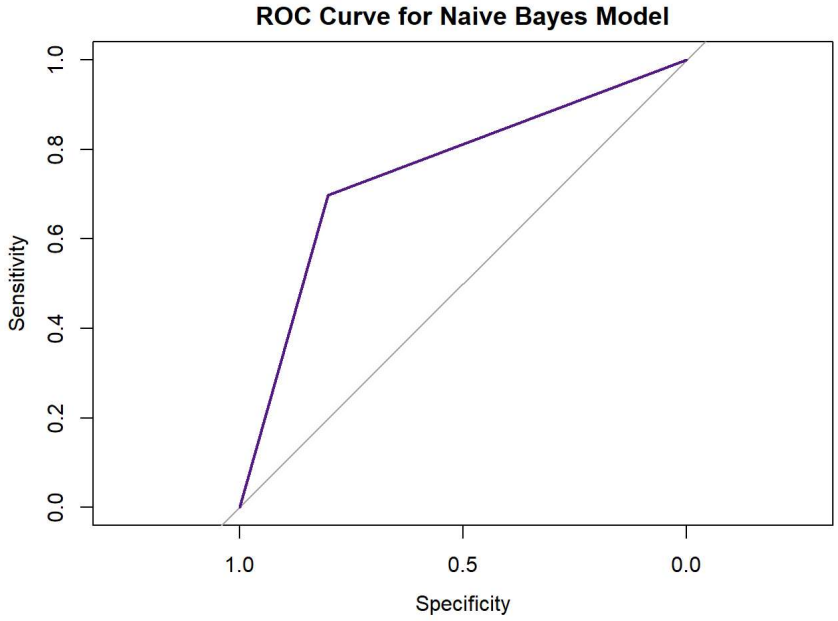
```
## Accuracy: 0.7743772
```

```
## Precision: 0.879915
```

```
## Recall: 0.8023256
```

```
## F1-score: 0.839331
```

```
## AUC: 0.7496883
```



The performance metrics for the Naive Bayes model indicate strong overall effectiveness. Its accuracy is high, suggesting that it correctly predicts outcomes in the dataset the majority of the time. The precision value indicates a low rate of false positives, implying that when the model predicts a positive result, it is usually correct. The recall value, although slightly lower than precision, indicates that the model captures a substantial portion of

actual positive cases. The F1-score, which balances precision and recall, shows a good overall balance between these metrics. Additionally, the AUC value signifies decent discrimination ability, indicating that the model can distinguish between classes effectively.

Bootstrapped Evaluation of Naive Bayes Model:-

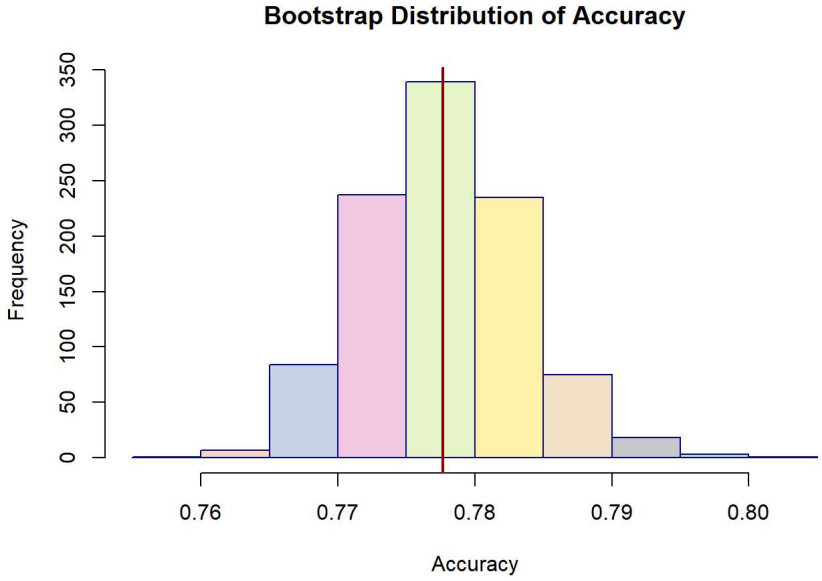
```
## Mean Accuracy: 0.7776574
```

```
## Mean Precision: 0.8838239
```

```
## Mean Recall: 0.8025973
```

```
## Mean F1-score: 0.8412389
```

```
## Mean AUC: 0.7556522
```



The Naive Bayes model demonstrated robust performance across 1000 bootstrap iterations. On average, it achieved high precision, indicating a low false positive rate. The recall score highlights its ability to capture a significant proportion of relevant instances. The F1-score reflects a harmonious balance between precision and recall, crucial for classification tasks. Additionally, the model's mean AUC score indicates its competence in distinguishing between classes. These consistent metrics suggest that the Naive Bayes algorithm is reliable, showcasing its potential for accurate predictive modeling without sacrificing generalization.

Analyzing Model Performance and Interpretation of Results:-

Model	Accuracy	Precision	Recall	F1_Score	AUC
Logistic Regression Model(individual) variables	0.81	0.68	-	-	-
Logistic Regression Model (Interaction) variables	0.82	0.69	-	-	-
Random Forest Model without PCA	0.80	0.66	-	-	-
Random Forest Model with PCA	0.78	0.61	-	-	-
Tuned Random Forest Model without PCA(5-foldCV)	0.80	0.83	0.92	0.87	0.7
Tuned Random Forest Model with PCA(5-foldCV)	0.78	0.82	0.89	0.85	0.67
Logistic Regression Model with Forward Selection	0.81	0.68	0.58	0.62	0.74
Logistic Regression Model with Forward Selection(10-foldCV)	0.80	0.64	0.56	0.69	-
Linear Discriminant Analysis (LDA) Model	0.80	0.65	0.58	0.61	-
Linear Discriminant Analysis (LDA) Model with Important Features	0.79	0.63	0.52	0.57	-
Support Vector Machines Model	0.81	0.83	0.93	0.88	0.7
Regularized SVM Model with Cost Optimization	0.81	0.84	0.91	0.88	0.72
Regularized SVM Model with Cost Optimization(5-fold CV)	0.80	0.91	0.83	0.87	0.71
Decision Tree Model	0.79	0.83	0.91	0.86	0.74
Decision Tree Model with Pruning and Regularization	0.80	0.81	0.94	0.87	0.76

Model	Accuracy	Precision	Recall	F1_Score	AUC
Naive Bayes Model	0.77	0.88	0.8	0.84	0.75
Naive bayes model with bootstrapping(1000)	0.78	0.88	0.8	0.84	0.75

The analysis of the Telco customer churn dataset yielded insightful results across various statistical learning models. Logistic Regression models, both with individual and interaction variables, showcased good accuracy at around 81-82%. Random Forest models performed well, especially after tuning without PCA, achieving an accuracy of 80% with high precision and recall, indicating a balanced prediction. Support Vector Machines (SVM) and Regularized SVMs demonstrated strong performance, emphasizing their suitability for this classification task with accuracy around 81% and good precision-recall trade-offs. Decision Tree models also performed decently, especially with pruning and regularization, achieving 80% accuracy and high precision-recall metrics. Naive Bayes models, while simpler, provided competitive results but lacked the robustness seen in other models. Overall,methods like Random Forest and SVM, along with tuned and regularized versions, showed the most promise for predicting customer churn effectively in this dataset.

Scope of Predictive Analysis:-

The predictive analysis conducted on the Telco customer churn dataset showcases a comprehensive exploration of various machine learning models' performance in predicting customer churn. The analysis encompasses logistic regression, random forest, SVM, linear discriminant analysis, decision tree, and naive Bayes models, along with feature engineering techniques like interaction variables and forward selection. The scope of this predictive analysis is broad, aiming to understand and predict customer churn based on a diverse set of predictors, including both numeric and categorical variables. The models exhibit varying levels of accuracy, precision, recall, and F1 scores, highlighting the importance of model selection and feature engineering in improving predictive performance. Additionally, techniques such as PCA and regularization are explored, further enhancing the predictive capabilities of certain models. Overall, the scope extends to optimizing model performance, understanding feature importance, and leveraging different algorithms to address the churn prediction challenge effectively.

Evaluating Model Generability:-

The predictive analysis shows promising generability based on the performance metrics across various models. Logistic Regression and Support Vector Machines (SVM) demonstrate consistent accuracy and high precision, indicating their reliability in predicting churn. Random Forest, despite lower precision, maintains competitive accuracy. Notably, Regularized SVM with Cost Optimization showcases improved recall, crucial for identifying churn cases. Decision Trees with pruning exhibit balanced precision and recall, enhancing their robustness. However, Naive Bayes, while achieving good precision, needs further improvement in recall for broader applicability. The tuning and feature engineering efforts, especially in Random Forest models and Logistic Regression with Forward Selection, contribute to model enhancement. Overall, the diverse models' performances and optimization efforts suggest a strong foundation for predictive generability, signaling potential applicability in real-world scenarios beyond the training data.

Limitations and Improvement Strategies:-

The results showcase promising accuracies across various models, especially with Regularized SVM, Logistic Regression (with Forward Selection), and Decision Trees. However, these metrics lack insights into model robustness and potential overfitting. To enhance analyses, consider incorporating feature importance techniques like SHAP values or Recursive Feature Elimination. Additionally, using ensemble methods such as Gradient Boosting Machines or Stacking can potentially boost predictive performance and provide better generalization on unseen data. Conducting further hyperparameter tuning, especially with complex models like SVM and Random Forest, could also yield improvements in model stability and predictive power.