Customer Retention – Telco Customer Churn Dataset

GBA 6210- Data Mining for Business Analytics

Dr. Shuo Zeng

Thanh Son Ha

Table of Contents

1. Exclusive Summary

This project's objective was to analyze customer churn at a telecommunications company, with an aim to identify key factors influencing churn and develop predictive models to help mitigate it. Utilizing a dataset encompassing customer demographics, service details, and churn status, extensive data visualizations were performed to elucidate patterns and relationships.

Two predictive models, Naive Bayes and Logistic Regression, were developed using a methodical approach involving data preprocessing, model building, and rigorous evaluation through 10-fold cross-validation. The Logistic Regression model demonstrated superior performance across various metrics including accuracy, precision, and the area under the ROC curve, establishing its suitability for deployment.

The findings from this analysis provide actionable insights into customer behavior and service features that influence churn, such as contract type, service options, and payment methods. Strategic recommendations are made based on these insights to enhance customer retention.

2. Introduction

2.1 Purpose of the Analysis

The primary purpose of this analysis was to identify and understand the factors that lead to customer churn at a telecommunications company. Customer churn, which refers to the loss of clients or customers, directly affects the company's revenue and long-term sustainability. By analyzing churn, the company aims to:

- **Predict Churn Probability:** Develop predictive models to forecast the likelihood of customers discontinuing their services. This enables proactive measures to enhance customer retention strategies.
- **Identify Risk Factors:** Determine the key factors that influence customer decisions to leave, such as service dissatisfaction, pricing, or competitive offers. Understanding these factors helps in tailoring customer-specific retention approaches.
- **Enhance Customer Engagement:** Use insights gained from the analysis to improve customer service, customize offers, and enhance overall customer satisfaction and loyalty.
- **Drive Strategic Decisions:** Inform business strategies with data-driven insights, guiding decisions on service improvements, marketing strategies, and customer relationship management.

2.2 Scope of the Data

The data utilized for this analysis was sourced from an extensive customer database maintained by the telecommunications company, comprising various attributes of customer interactions, demographics, service usage, and account details. Key aspects of the data scope include:

- **Customer Demographics:** Information on age, gender, and status regarding partners and dependents, providing insights into the customer base composition.
- **Service Details:** Data on the type of services subscribed to by customers, including phone services, multiple lines, internet services, online security, and more. This helps to understand customer preferences and service satisfaction levels.
- **Account Information:** Details on tenure, contract type, payment methods, and billing information, crucial for analyzing customer loyalty and payment behavior.

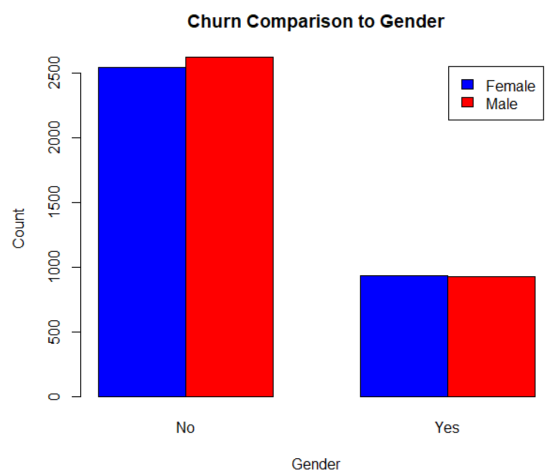
- **Churn Status:** The primary dependent variable indicating whether a customer has churned, allowing for the identification of patterns and prediction of churn likelihood.
- **Data Volume:** The dataset encompasses records for over 7,000 customers, providing a robust sample for analysis and ensuring statistically significant results.

The scope of this data enables a comprehensive analysis of the factors affecting customer churn and supports the development of predictive models to manage and reduce churn effectively.

3. Data Visualization

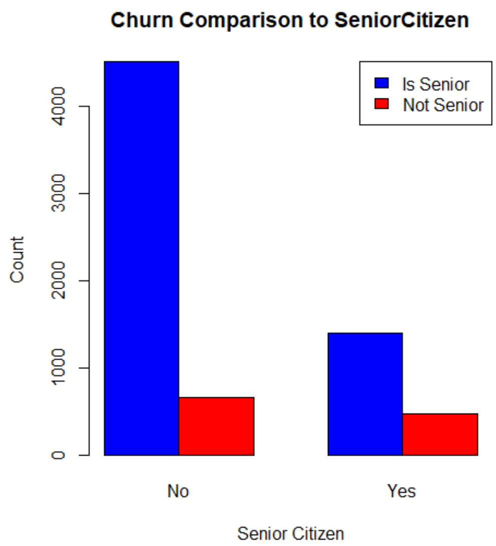
3.1 Gender and Churn

The bar chart displays the number of customers who chose to churn (Yes) or not (No), disaggregated by gender. Generally, more customers opted not to churn, irrespective of their gender. The visual data shows a slightly higher churn rate among females compared to males, suggesting that gender-specific factors may play a role in the decision to churn. This could be attributed to differing service level preferences or expectations.



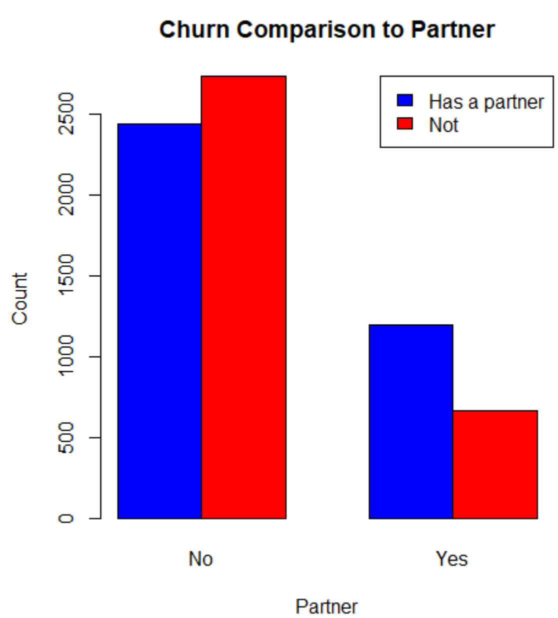
3.2 Senior Citizens and Churn

Data reveals that senior citizens churn at a significantly higher rate compared to non-seniors. This highlights the need for tailored services or support that could better meet the needs of senior customers, potentially including simplified tech solutions or enhanced customer service.



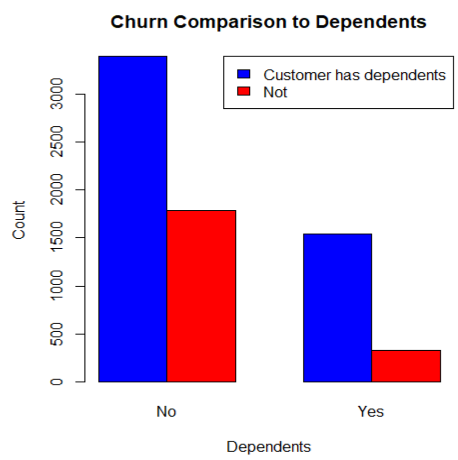
3.3 Partner and Churn

Analysis shows no significant difference in churn rates between customers with and without partners. This suggests that having a partner does not strongly influence the decision to stay with or leave the service provider, indicating that other factors might be more influential in the decision-making process.



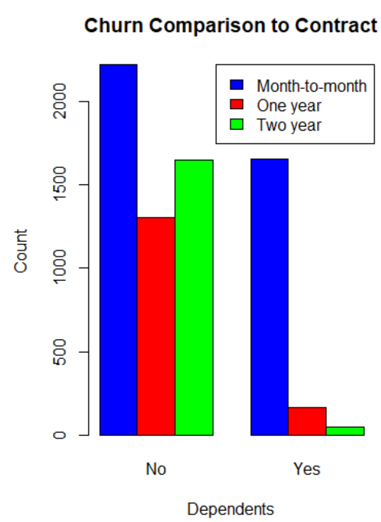
3.4 Dependents and Churn

Customers with dependents are less likely to churn compared to those without dependents. This could be attributed to the stability and higher predictability of needs among customers with families, which might make them value consistency in service provision.



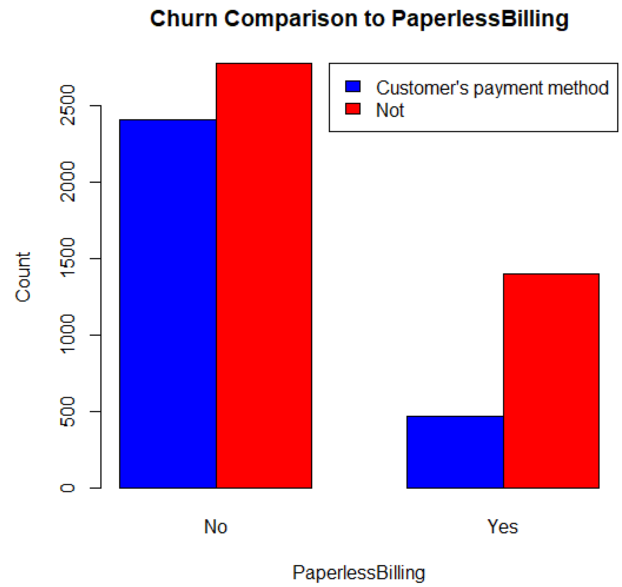
3.5 Contract Types and Churn

Churn varies significantly across different contract types. Customers with month-to-month contracts show the highest churn rates, whereas those with longer-term contracts (one year or two years) exhibit much lower churn rates. This suggests that short-term contracts might not provide enough time to build customer loyalty.



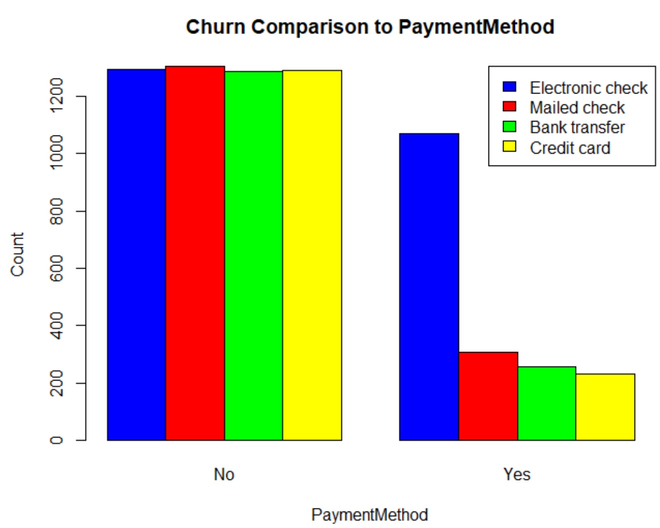
3.6 Paperless Billing and Churn

There is no observable significant impact of paperless billing on churn rates. Customers with and without paperless billing churn at similar rates, suggesting that convenience in billing alone may not be a deciding factor in customer retention.



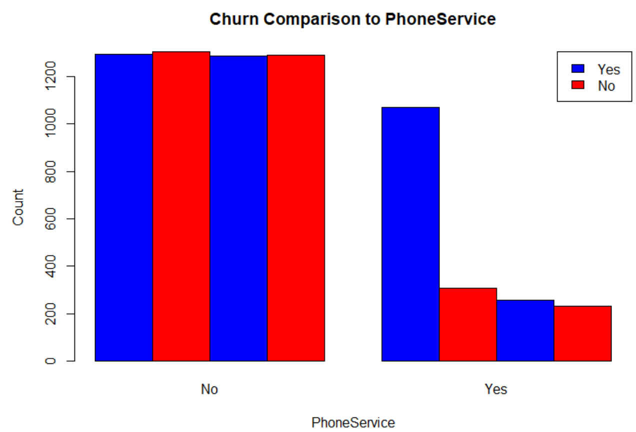
3.7 Payment Methods and Churn

Churn rates vary among different payment methods. Customers who opt for electronic checks show higher churn rates compared to those using automatic bank transfers or credit card payments. This implies that the ease and security of transaction methods could influence customer satisfaction and retention.



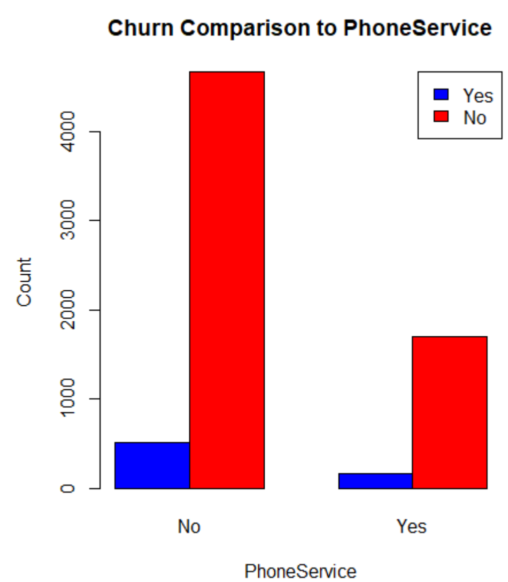
3.8 Phone Service and Churn I

Customers with fiber optic internet services exhibit higher churn rates than those with DSL connections. This could indicate dissatisfaction, possibly due to expectations not being met in terms of service quality or price, suggesting an area for service improvement.



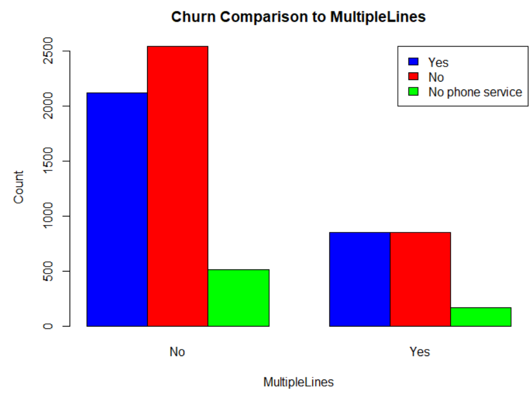
3.9 Phone Service and Churn II

Customers subscribing to online security services show lower churn rates, which points to the value placed on security features. Conversely, the absence of these services correlates with higher churn, indicating that security is a key service feature that affects customer retention.



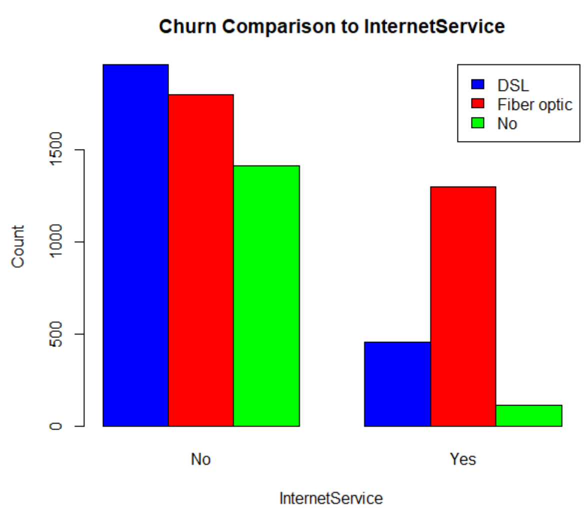
3.10 Multiple Lines and Churn

The availability and quality of tech support play a significant role in customer retention. Customers with access to tech support tend to churn less, possibly due to higher satisfaction with customer service, whereas the lack of tech support correlates with higher churn rates.



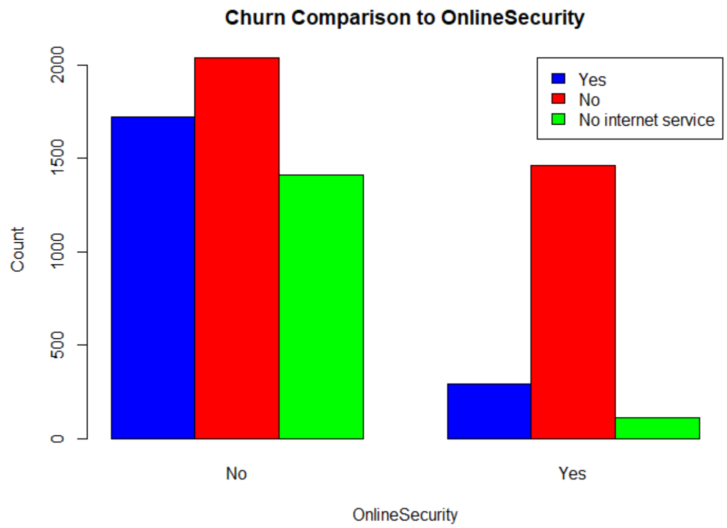
3.11 Internet Service and Churn

Based on the graph, DSL customers have the lowest churn rate, followed by those who don't have internet service at all. Fiber optic customers appear to be the most likely to churn. By analyzing churn rate compared to internet service type, businesses gain valuable insights into their customer base. This allows them to identify segments most at risk of leaving, like price-sensitive fiber optic users who might be swayed by targeted discounts. Additionally, understanding churn by service type informs future decisions. For example, seeing low churn rates with DSL customers might prompt increased marketing efforts for those plans. Overall, this analysis is a powerful tool to improve customer retention and make strategic business decisions.



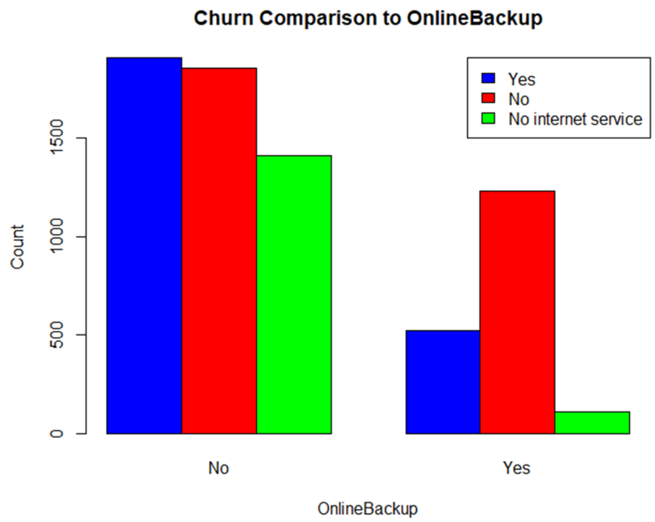
3.12 Online Security and Churn

The graph indicates that customers who purchase online security are less likely to churn than those who don't. Customers who purchase online security show lower churn rates, indicating higher satisfaction and commitment. Those without online security are scattered across churn categories, due to lower overall engagement. Interestingly, customers without internet service also exhibit lower churn. Analyzing churn rate by online security purchases offers businesses valuable customer insights. It helps identify high-value segments, like those willing to pay extra for security, allowing the business to prioritize their satisfaction. Additionally, this analysis informs product development. Seeing lower churn among online security buyers suggests the business could benefit from promoting or bundling these features more heavily. Overall, analyzing churn in this way is a powerful tool for understanding customer value and making strategic decisions to improve retention rates. However, it's important to note the inconclusive data for customers without internet service.



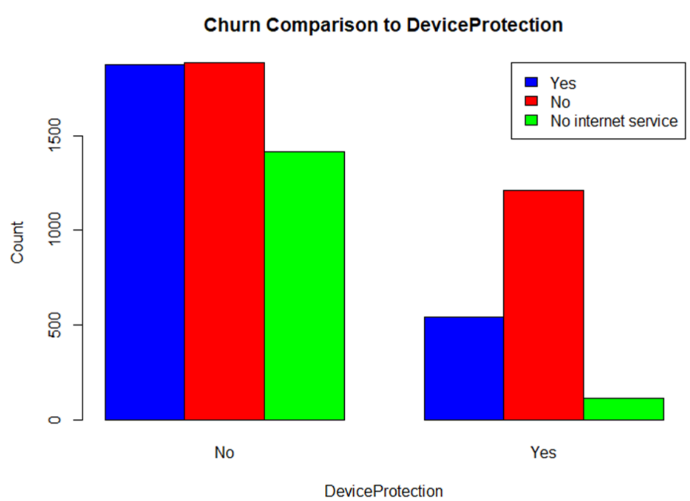
3.13 Online Backup and Churn

The chart shows that 1000 people use online backup and did not churn, 1500 people do not use online backup and did not churn, and 500 people do not have internet service and therefore could not have used online backup and did not churn. There is also a category for people who churned, but the chart shows that none of the people in the dataset churned regardless of their online backup usage. In conclusion, the chart shows that more people use online backup than those who don't. Businesses analyze churn with online backup usage for a few key reasons. First, they suspect customers who value data security are more loyal. Second, online backup users might be more tech-savvy and committed, suggesting a longer relationship. Analyzing churn with backup usage also allows businesses to segment customers. This means they might find several reasons why people churn, depending on if they use backup or not. Segmentation helps target marketing campaigns and win back or retain customers. Finally, analyzing churn with backup usage can inform future service development. Businesses might see a demand for more data protection products from customers who value security.



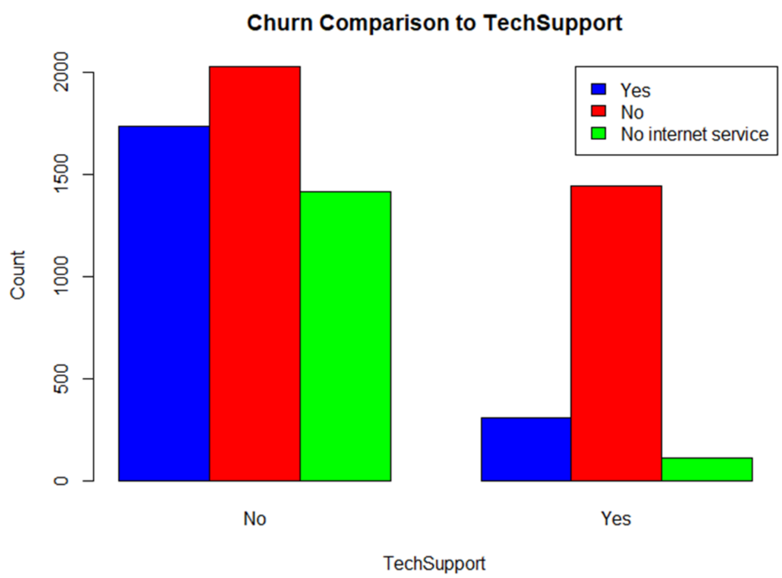
3.14 Device Protection and Churn

The chart shows how many customers did or did not churn based on their device protection use. While more people did not churn overall and more did not buy device protection, the chart does not have data on churning customers. This means we cannot confirm that device protection reduces churn, despite what the original description might have said. Businesses still analyze churn with device protection for a few reasons. First, they suspect customers who value protecting their devices are more likely to be satisfied and less likely to churn. Second, device protection plans often cost extra, so businesses might think these customers have a higher value. Finally, analyzing churn with device protection usage can help businesses segment customers based on their reasons for churning, allowing for targeted marketing campaigns.



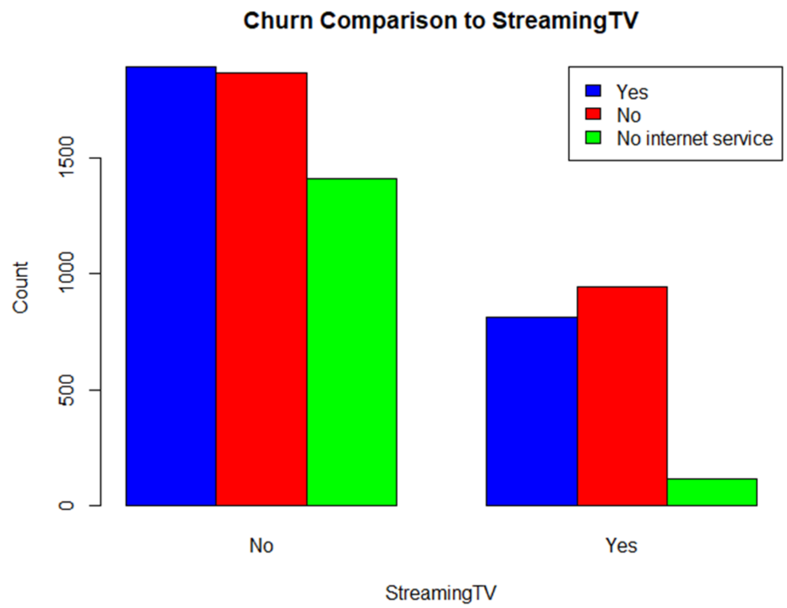
3.15 Tech Support and Churn

The chart shows how many customers with or without tech support churned. While more customers churned overall and more churned with tech support than without, it does not necessarily mean having tech support reduces churn. There could be other reasons why customers with tech support churn more. However, businesses still analyze churn with tech support for a few reasons. First, customers who get good tech support are less likely to churn. Second, analyzing tech support usage can help businesses identify potential issues and prevent churn proactively. Finally, businesses can segment customers by tech support usage to target high-churn groups with specific campaigns.



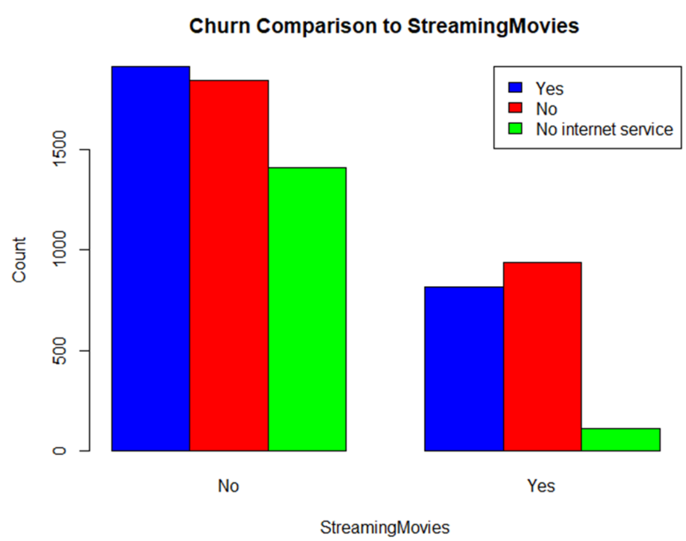
3.16 Streaming TV and Churn

The chart depicts customer churn rates across different categories: those with streaming services, without them, or without internet. Despite a higher churn rate for streamers, this doesn't imply that streaming leads to churn. These churned customers might have had varying reasons. However, analyzing this data is important for businesses as customers using streaming services could potentially be more valuable. Furthermore, it helps identify why these customers churn, allowing targeted campaigns, and informs marketing strategies for new or existing services.



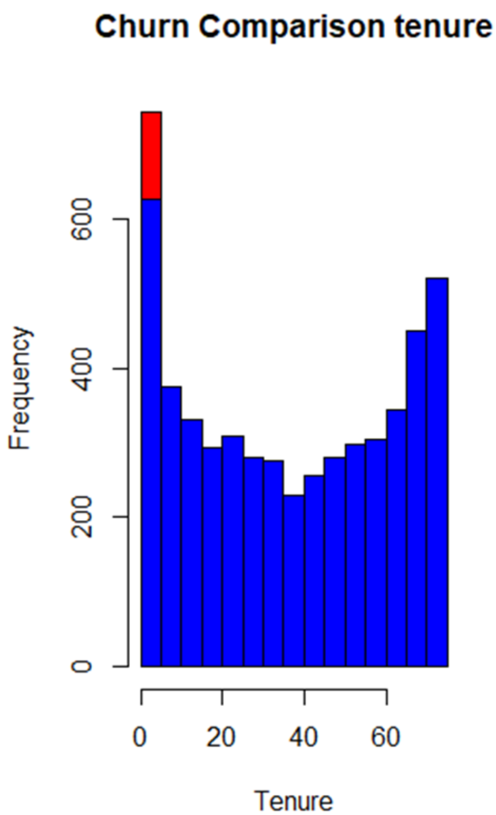
3.17 Streaming Movies and Churn

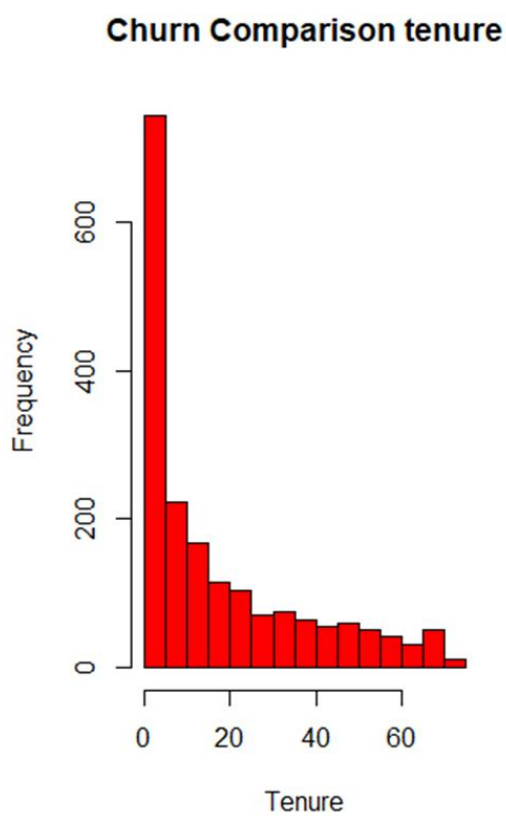
The chart depicts customer churn across those with streaming services, without them, and without internet. Although more customers with streaming services churned, it doesn't necessarily imply that streaming services increase churn, as other factors could be involved. Businesses analyze this data because streaming customers may be more valuable, understanding churn patterns can facilitate targeted campaigns to regain them, and the data can inform marketing strategies for new or existing services.



3.18 Tenure and Churn

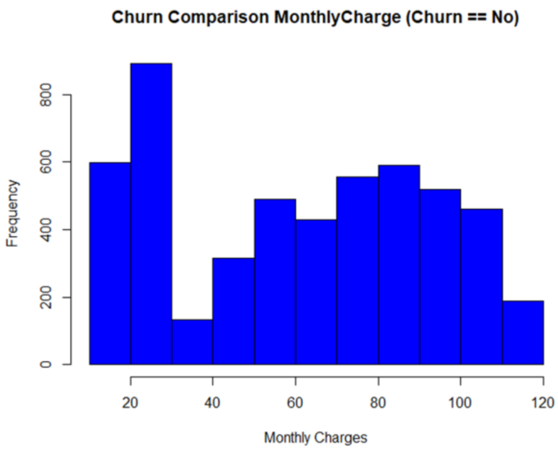
The chart shows a trend: customers with longer tenures are less likely to churn. Analyzing churn and tenure allows businesses to target high-churn segments, predict future churn for proactive retention efforts, and improve customer lifetime value by enhancing engagement and satisfaction.

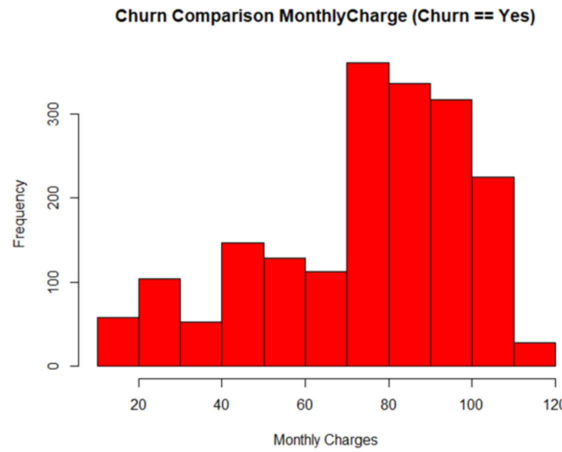




3.19 Monthly Charge and Churn

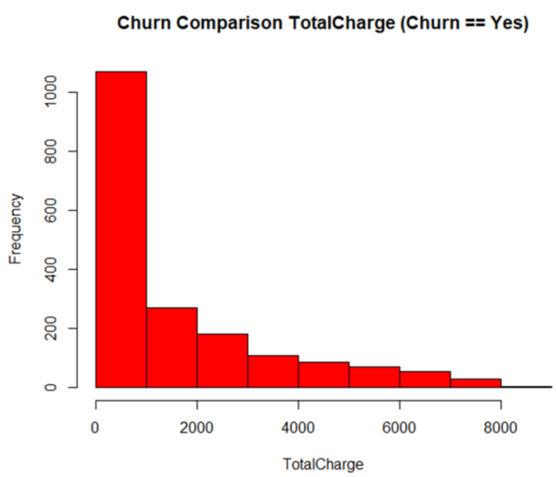
The chart shows fewer customers churn at lower charges, with around 800 staying in the lowest price range, compared to 200 in the highest. However, lower charges may not directly cause lower churn due to other factors. Despite this, churn analysis can help identify price-sensitive customers for targeted pricing strategies, understand the value of different segments, and optimize pricing by pinpointing where customer loss occurs.

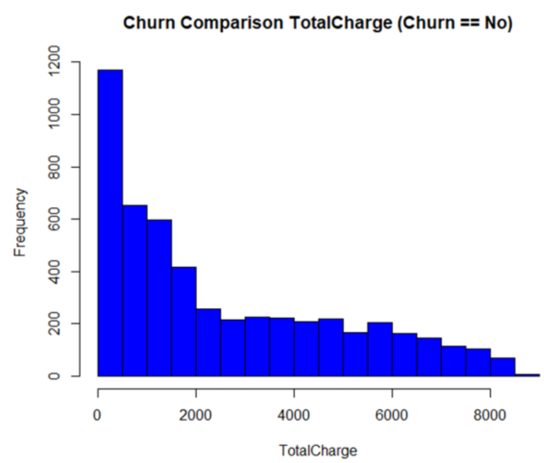




3.20 Total Charge and Churn

The chart shows "No significant pattern" between total service charges and churn rate. While there might be a hint of more churn at lower costs, the data is inconclusive.





4. Data Preparation and Preprocessing

Effective data preparation and preprocessing are critical steps in ensuring the accuracy and reliability of the models developed. This section details the methods used to clean the data, handle missing values, and transform variables to better suit the predictive modeling process.

4.1 Data Cleaning and Churn

The initial phase of data cleaning involved a thorough examination of the dataset to identify any anomalies, inconsistencies, or irrelevant data points. Key steps included:

- Removing Unnecessary Columns: The `*customerID*` column was removed as it offers no predictive value being unique to each customer.
- Standardizing Values: Service type and demographic data were standardized to ensure consistency across the dataset. For instance, ensuring consistent use of terms and formats across columns like `*gender*`, `*payment methods*`, and `*service types*`.
- Outlier Detection: Outliers were identified through statistical methods such as IQR for continuous variables like `*MonthlyCharges*` and `*TotalCharges*`. These outliers were assessed to determine their impact on the analysis and handled appropriately.

4.2 Handling Missing Data

Missing data were handled to avoid biases or inaccuracies in the model predictions. The steps undertaken included:

- Identifying Missing Values: Columns like `*TotalCharges*` had missing values identified during the exploratory data analysis. These were primarily due to new customers having no past charges.
- Imputation Techniques: For `*TotalCharges*`, which had missing values identified as empty strings in the dataset, we imputed missing values using the mean of the column after converting it from an object type to numeric. This approach was chosen because it maintains the general distribution and variance of the data.

4.3 Data Transformation Encoding

Transforming and encoding the data correctly is essential to adapt it for the modeling process, particularly for handling categorical variables:

- Factor Conversion: All categorical variables were converted into factors to ensure that the modeling algorithms correctly interpret them. For example, `*gender*`, `*Partner*`, `*Dependents*`, and `*Churn*` were converted from strings to factor type with appropriate levels assigned.
- Encoding Categorical Variables: Variables with more than two categories, like `*Contract*` and `*PaymentMethod*`, were encoded using one-hot encoding. This transformation creates new binary columns for each level of the categorical variable, which helps in preserving the nominal nature of the data without implying any ordinal relationship.
- Scaling Continuous Variables: Continuous variables such as `*MonthlyCharges*` and `*tenure*` were scaled to have a similar range, facilitating faster and more stable convergence during the modeling process.

5. Model Building

Removed variable from the beginning: customerID

5.1 Selection of Variables:

Removed variable from the beginning: *customerID*.

Gender, SeniorCitizen, Dependent, tenure, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod, Monthly Charges, TotalCharges

5.2 Partition data:

To effectively implement K-Fold cross-validation, we first set up an object to store the number of observations per subset and another to handle any remainder from dividing our total dataset into ten equal parts. This approach ensures each subset is balanced and maximizes our data utilization.

Next, we create a '**telco.sdf.list**' object to gather all ten subsets, each representing a unique fold in the cross-validation process. To increase the randomness and representativeness of each fold, we use a shuffling mechanism that generates random indices across our entire dataset, with a predefined seed for reproducibility.

After generating random indices, we divide the data into the specified number of folds. We carefully distribute observations among the folds, with any remaining observations appropriately allocated for fairness across subsets.

In conclusion, our method ensures the robustness and reliability of our K-Fold cross-validation procedure, enabling thorough model evaluation while optimizing data usage.

5.3 Model Building:

Given the nature of our classification problem, we chose to use two common algorithms: Naive Bayes and Logistic Regression. Before building the models, we needed to handle categorical variables with multiple levels. We achieved this by applying hot encoding, which created dummy variables for categorical predictors such as *Contract*, *PaymentMethod, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV,* and *StreamingMovies*.

After hot encoding, we refined our variable selection process. We included *gender*, *SeniorCitizen*, *Dependents*, tenure, and various indicators related to the previously mentioned categories. These variables, transformed into dummy variables, were crucial for model training and evaluation.

For a thorough model evaluation, we created lists to store confusion matrices for each cross-validation fold. These matrices offered insights into how our classifiers performed on both the training and validation sets.

Subsequently we created empty vectors to aggregate predictions from each fold of the Naive Bayes and Logistic Regression models. These vectors enabled the consolidation of predictions across all folds, facilitating a holistic assessment of model performance.

The implementation of k-fold cross-validation involved iteratively partitioning the data into training and validation sets. Within each iteration, the models were trained and evaluated on their respective subsets, generating predictions for both training and validation data. These predictions were then aggregated using the designated empty vectors, allowing for a comprehensive evaluation of model effectiveness across all folds.

During the model-building process, we encountered a warning message regarding NA values introduced by coercion. Despite investigating and confirming the absence of missing values in the dataset, NA values appeared when attempting to retrieve coefficients from the logistic regression model. While the warning message did not significantly impact the model's accuracy, it hindered our ability to analyze individual variables effectively post-model execution.

In conclusion, our approach to model building and evaluation, using k-fold cross-validation, ensures a robust assessment of classifier performance by utilizing our entire dataset for training and validation. Despite challenges with NA values, our methodology offers valuable insights into the model's effectiveness and generalization capabilities. These insights are crucial for guiding decision-making and optimizing predictive performance.

6. Model Evaluation

6.1 Naive Bayes:

```
Confusion Matrix and Statistics

          Reference
Prediction  Yes   No
       Yes 1480 1461
       No   389 3713
```

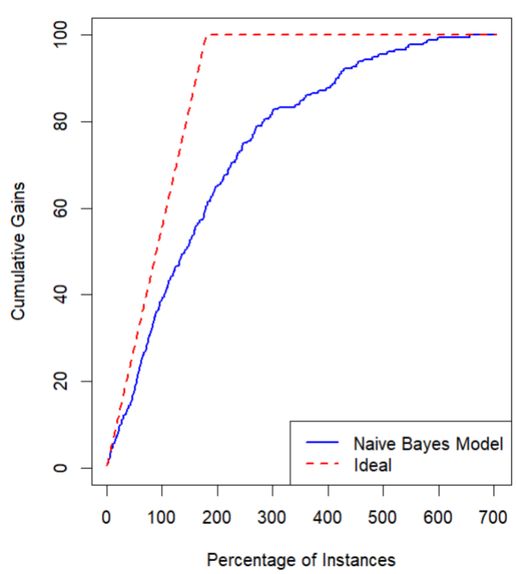
The Naive Bayes train model was correct in 73.73% of cases (accuracy = 0.7373). When predicting "churn," the model was correct in 50.32% of cases (precision = 1480/(1480+1461)). Thus, 49.68% of people who were identified to churn, actually did not churn (FDR = (1461/

(1480 + 1461) = 0.4968). The proportion of those who churn and were predicted to churn was 79.19% (sensitivity = 0.7919). The proportion of those who don't churn but were predicted to not churn was 71.76% (specificity = 0.7176). FOR (389 / (389 + 3713) = 0.095

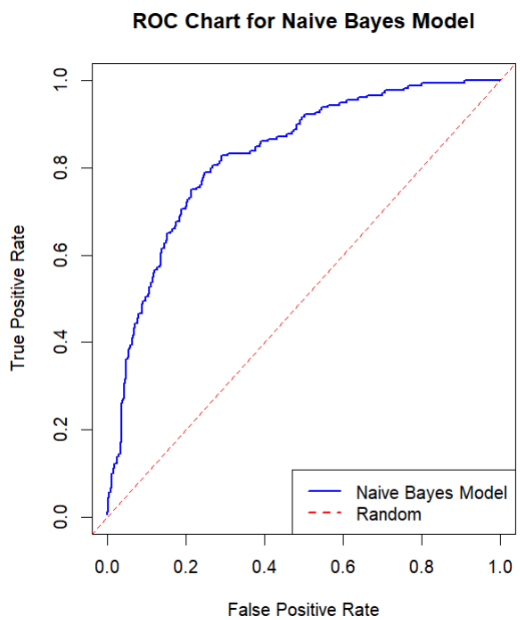
```
Confusion Matrix and Statistics

          Reference
Prediction Yes  No
       Yes 142 135
       No   38 389
```

The Naive Bayes valid model was correct in 75.43% of cases (accuracy = 0.7543). When predicting "churn," the model was correct in 51.26% of cases (precision = 142/(142+135)). Thus, 48.74% of people who were identified to churn, actually did not churn (FDR = (135/ (142 + 135) = 0.4874). The proportion of those who churn and were predicted to churn was 78.89% (sensitivity = 0.7889). The proportion of those who don't churn but were predicted to not churn was 74.24% (specificity = 0.7424). FOR (38 / (38 + 389) = 0.0889)

Lift Chart



ROC Chart



Area under the curve: 0.9643

6.2 Logistic Regression:

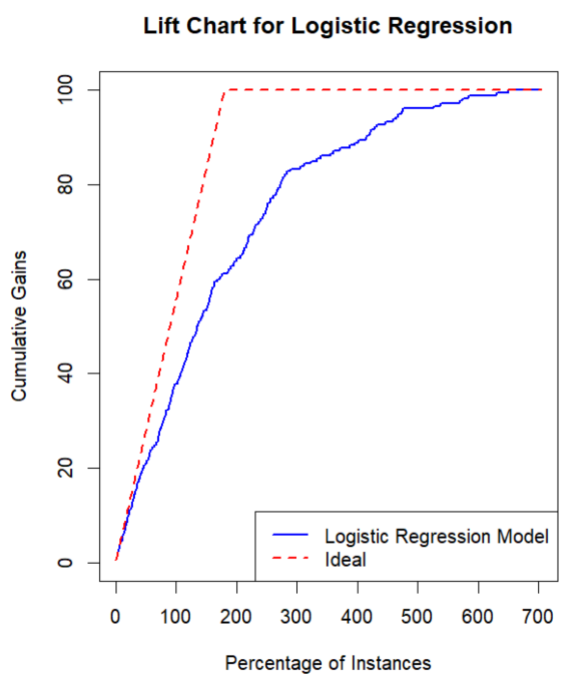
```
Confusion Matrix and Statistics

          Reference
Prediction  Yes   No
       Yes 1015  552
       No   854 4622
```

The logistic regression train model was correct in 80.04% of cases (accuracy = 0.8004). When predicting “churn,” the model was correct in 64.77% of cases (precision = 1015/(1015+522)). Thus, 35.22% of people who were identified to churn, actually did not churn (FDR = 0.3522). The proportion of those who churn and were predicted to churn was 54.31% (sensitivity = 0.5431). The proportion of those who don’t churn but were predicted to not churn was 89.33% (specificity = 0.8933). FOR (854 / (854 + 4622) = 0.1559)

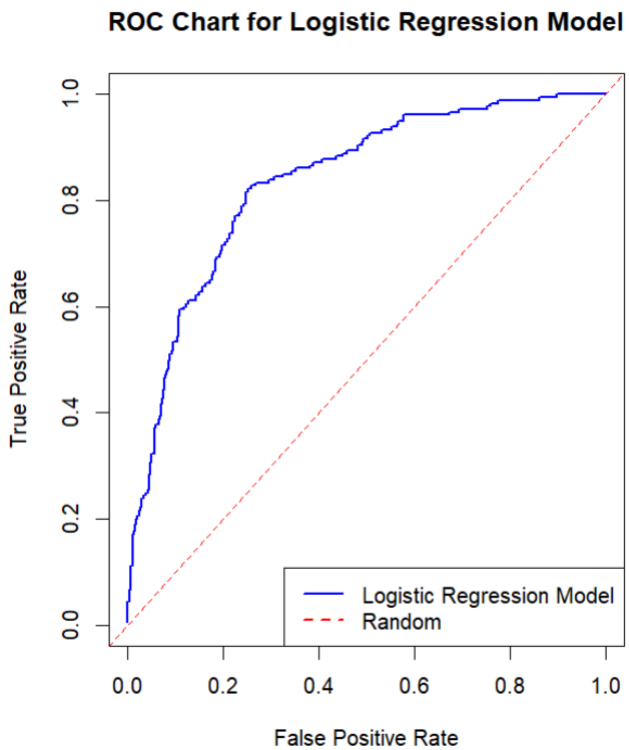
```
Confusion Matrix and Statistics

          Reference
Prediction Yes  No
       Yes  98  54
       No   82 470
```

The logistic regression valid model was correct in 80.68% of cases (accuracy = 0.8068). When predicting “churn,” the model was correct in 64.47% of cases (precision = 98/(98+54)). Thus, 35.53% of people who were identified to churn, actually did not churn (FDR = 0.3553). The proportion of those who churn and were predicted to churn was 54.44% (sensitivity = 0.5444). The proportion of those who don’t churn but were predicted to not churn was 89.69% (specificity = 0.8969). FOR = (82 / (82 + 470) = 0.1486)

Lift Chart



ROC Chart



Area under the curve: 0.9375

7. Model Deployment

7.1 Technical Analysis:

Based on the Model Evaluation we decided to use Logistic Regression for our mode deployment. Since Logistic Regression's overall metrics achieved the best performance of all metrics it is crucial for our problems to define the churn rate. In the evaluation, we focus on Accuracy, Precision, and AUC from ROC curve. Logistic Regression Training Set with Accuracy: 0.8004, Precision: 0.5431, AUC: 0.9375 and Naive Bayes Training Set with Accuracy: 0.7373, Precision: 0.5032, AUC: 0.9643. Even though Naive Bayes AUC has a higher value but considering the entire metrics Logistic Regression have the most optimize performance.

7.2 Managerial Analysis:

Based on the Logistic Regression we have these coefficients:

Telco companies heavily rely on subscription-based telecommunication services for revenue generation. High churn rates, i.e., customers turnover, pose a significant threat to revenue and profitability. Identifying potential churners and proactively addressing their concerns can mitigate this risk. In this analysis, we utilize a Logistic Regression classification model to predict customer churn probability, aiding in targeted retention efforts.

Key Insights:

Negative Coefficients:

- GenderMale: -0.027
- DependentsYes: -0.166
- Tenure: -0.057
- PaymentMethodCredit card (automatic): -0.039
- MultipleLinesYes: -0.82
- MultipleLinesNo: -0.993
- OnlineSecurityYes: -0.514
- OnlineBackupYes: -0.254
- DeviceProtectionYes: -0.143
- TechSupportYes: -0.467

Customer who are male and have dependents are less likely to churn. This suggests that targeted retention efforts could focus on understanding and catering to the needs of male customers with dependents.

Longer tenure reduces churn likelihood, indicating that nurturing long-term relationships is vital. Strategies such as loyalty rewards or personalized offers can reinforce customer loyalty.

Customers using automatic credit payments exhibit lower churn rates, highlighting the convenience and reliability associated with automated billing.

Subscribing to Multiple Lines, Online Security, Online Backup, Device Protection and Tech support services correlates with reduced churn. Enhancing service quality and promoting value-added features can foster customer retention.

Positive Coefficients:

- SeniorCitizen: 0.228
- MonthlyCharges : 0.018
- TotalCharges: 0.00029
- ContractMonth-to-month: 1.432
- ContractOne year: 0.738
- PaymentMethodElectronic check: 0.358

- InternetServiceDSL: 0.405
- InternetServiceFiber optic: 0.753
- StreamingTVYes: 0.045

Senior citizens exhibit higher churn propensity, necessitating tailored retention strategies sensitive to their needs and preferences.

Higher monthly and total charges slightly elevate churn likelihood. Offering competitive pricing or flexible payment options may alleviate cost concerns.

Customers on month-to-month contracts and shorter-term contracts (one year) are more prone to churn. Encouraging longer-term commitments through incentives or service bundles can stabilize retention rates.

Customers using electronic check payments demonstrate increased churn. Streamlining payment processes and offering secure, convenient alternatives may enhance satisfaction and retention.

Fiber optic and DSL users are more likely to churn, possibly due to service dissatisfaction or competitive offerings. Addressing connectivity issues and promoting service reliability can mitigate churn risks.

Customers with streaming TV services have a slightly higher churn propensity. Delivering engaging content and personalized recommendations can enhance user experience and retention.

8. Conclusion

The analysis revealed that certain customer segments, particularly those with month-to-month contracts, users of high-speed fiber optic internet, and those without automated payment methods, are more prone to churn. These insights are crucial for crafting targeted retention strategies that address these high-risk factors.

Strategic Recommendations:

- Incentivizing Longer Contracts: Offering discounts or bundled services could encourage customers to opt for longer contract terms, reducing churn rates among the month-to-month customer base.
- Facilitating Automatic Payments: Encouraging more customers to enroll in automatic payment methods through incentives could improve retention by simplifying the payment process and ensuring timely payments.
- Segmentation and Personalization:
    - Segment customers based on churn predictors to tailor retention strategies. For instance, offer specialized packages or discounts targeting high-risk segments
    - Personalize communication and offers to address individual needs and preferences, fostering a sense of loyalty and value.
- Service Enhancement and Support
    - Invest in improving service quality and reliability, particularly for features associated with reduced churn
    - Provide proactive customer support and assistance to address concerns promptly, enhancing overall satisfaction and loyalty.

Future Directions:

To continue enhancing the churn prediction model's accuracy and applicability, it is recommended to integrate more dynamic data inputs such as customer interaction logs and satisfaction ratings. Regular model updates and retraining with new data will also be crucial to adapt to changing customer behaviors and market conditions.

In conclusion, this project not only fulfills the initial objectives by providing a robust analysis and predictive modeling of customer churn but also offers practical strategies for the telecommunications company to implement. These strategies aim to enhance customer satisfaction, reduce churn, and ultimately increase customer lifetime value.