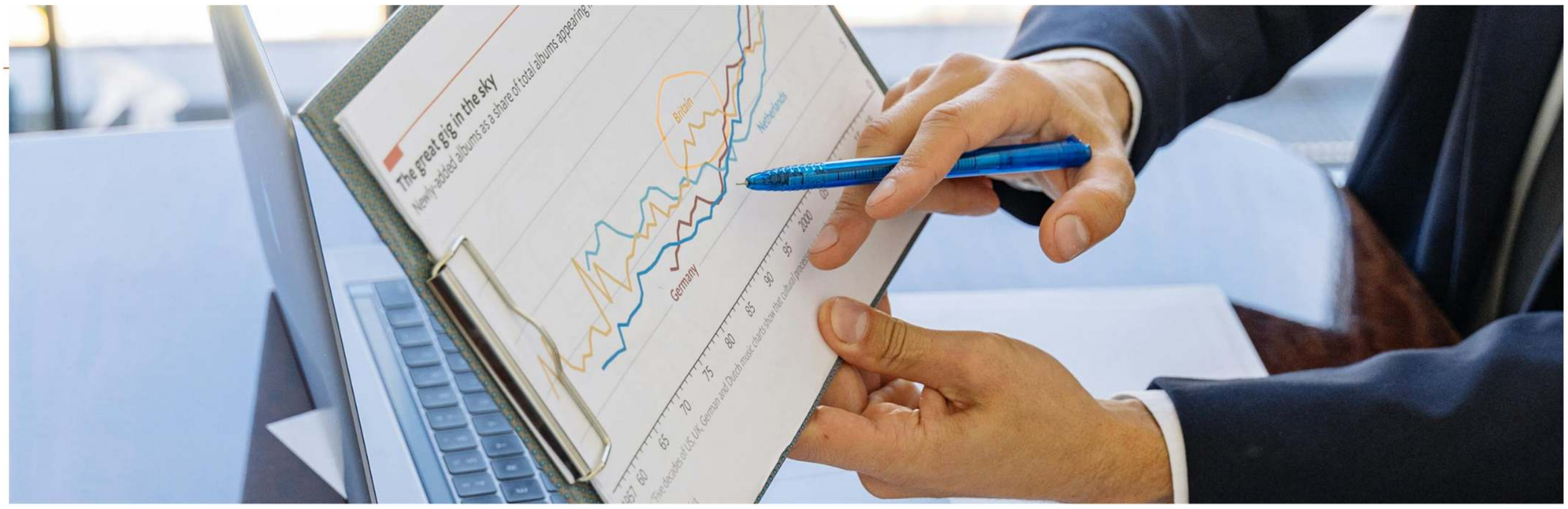
Customer Churn

Prediction & Analysis
by Guldanika Osmonova

REDER TELECOM

Table of Contents

Introduction



In the highly competitive telecommunications industry, customer churn represents a significant challenge for companies striving to maintain their market position and profitability. Churn, defined as the percentage of customers who discontinue their services or switch to competitors, poses a critical threat to the sustainability of businesses like Reder Telecom. The increasing rate of customer churn at Reder Telecom underscores the urgency of addressing this issue to safeguard revenue streams and enhance customer retention.

The telecommunications sector is characterized by intense competition, with numerous providers offering similar services. This competitive landscape makes it increasingly difficult for companies to retain their existing customer base. Customers today demand personalized and high-quality services, and any failure to meet these expectations can lead to dissatisfaction and eventual churn. Additionally, pricing pressures from competitors further complicate the ability to sustain competitive pricing strategies, impacting profitability. Network quality and performance issues also play a pivotal role in influencing customer satisfaction, often leading to churn if not adequately addressed. Moreover, building and maintaining customer loyalty remains a formidable challenge in this dynamic environment.

Recognizing the critical importance of predicting customer churn, Reder Telecom has embarked on a project to develop a predictive model capable of accurately forecasting which customers are at risk of discontinuing their services. The primary objective of this project is to leverage historical customer data and apply advanced machine learning techniques to identify and analyze the key factors influencing churn behavior. By doing so, Reder Telecom aims to create a robust model that not only predicts churn but also provides actionable insights to enhance customer retention strategies.

The significance of this project extends beyond mere prediction. By reducing churn, Reder Telecom can achieve substantial cost savings, as retaining existing customers is more cost-effective than acquiring new ones. Furthermore, a reduction in churn can lead to increased revenue, as satisfied customers are more likely to remain loyal and potentially purchase additional services. Understanding customer behavior and preferences through data analytics enables Reder Telecom to improve its services, thereby enhancing customer satisfaction and loyalty. Additionally, by proactively retaining customers, Reder Telecom can gain a competitive advantage over its rivals, positioning itself as a leader in the industry.

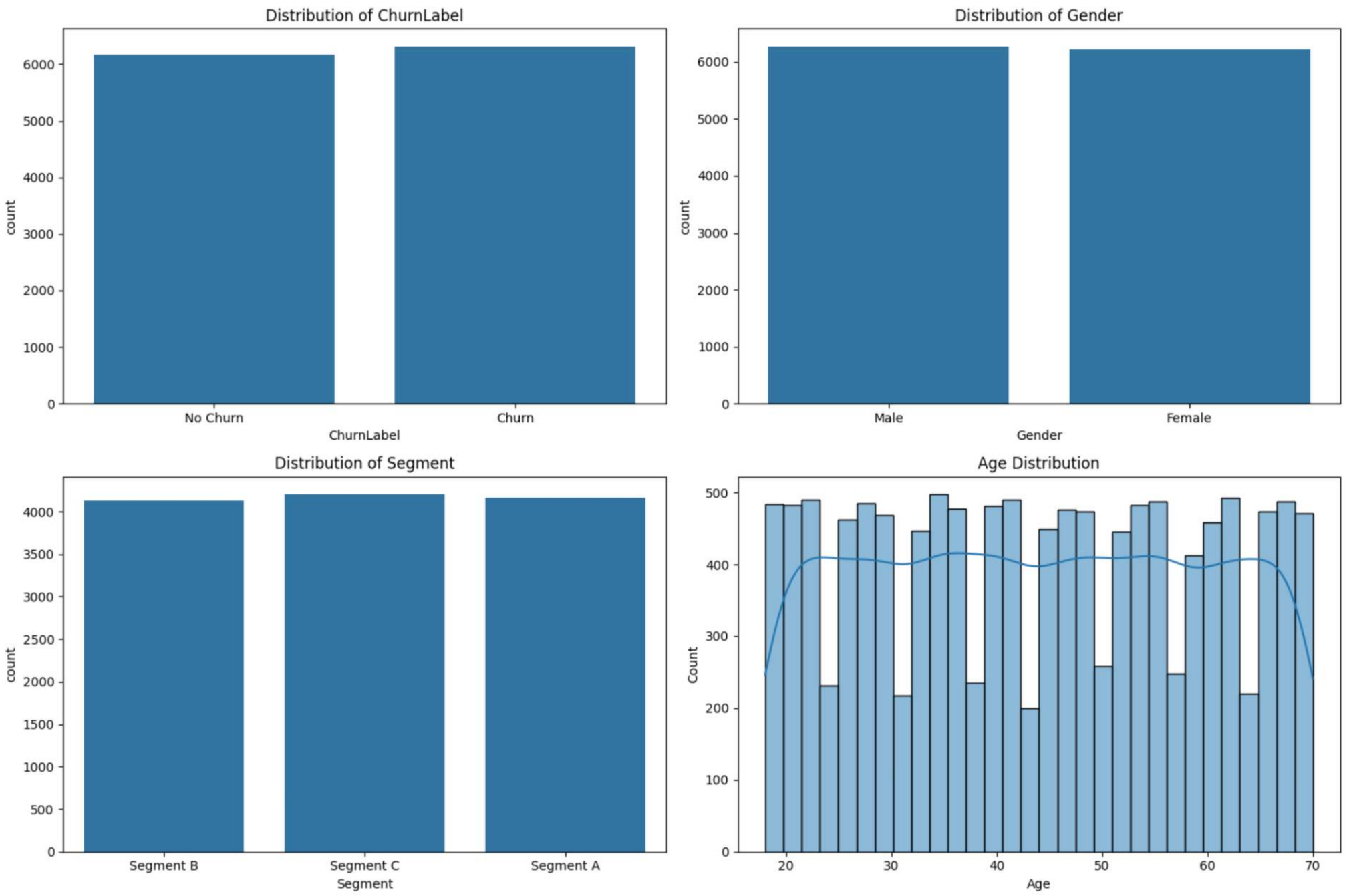
This report provides a comprehensive analysis of the customer churn prediction project, detailing the methodologies employed, data insights derived, and visualizations illustrating key patterns and correlations. The subsequent sections will delve into the data collection and exploration processes, data preprocessing and feature engineering techniques, model development and evaluation, and the insights gained from the analysis. Ultimately, the report will offer recommendations for Reder Telecom to reduce churn, improve customer satisfaction, and explore potential areas for further research and model enhancement.

Exploratory Data Analysis

The dataset comprises 12,483 entries and 21 columns, each representing a distinct aspect of customer interaction and engagement with Reder Telecom. Key variables include demographic information such as Customer ID, Age, Gender, and Location, as well as behavioral metrics like Purchase History, Subscription Details, Website Usage, Clickstream Data, Engagement Metrics, and Marketing Communication. Additionally, the dataset includes a Churn Label, a binary indicator of whether a customer has churned, which serves as the target variable for predictive modeling.

Initial data exploration involved a thorough examination of the dataset's structure and quality. This process included checking for missing values, duplicates, and inconsistencies. Notably, the dataset was found to be complete, with no missing values or duplicate entries, ensuring a robust foundation for analysis. The data types of each column were also verified, with a mix of numerical and categorical variables identified, necessitating appropriate preprocessing steps.

Summary statistics were generated to provide a high-level overview of the numerical features, revealing insights into the distribution and variability of key metrics such as Age, NPS, and various engagement measures. For instance, the average age of customers was found to be approximately 44 years, with a standard deviation indicating a diverse age range among the customer base. The Net Promoter Score (NPS), a critical measure of customer loyalty and satisfaction, exhibited a wide range, highlighting varying levels of customer engagement and satisfaction.



*A few things to note from these visuals,

- the dataset is almost evenly distributed between theChurnLabel=1 (churned) andChurnLabel=0 (not churned),
- the number ofMale andFemale gendered customers are almost equal,
- the distribution ofSegment variable is almost even across the three Segments (Segment A,Segment B andSegment C).,
- the number of customers of any particular Agein the dataset, seems to flunctuate between 200and `500

Data Preprocessing

Feature Engineering

From each of these unique values in `ServiceInteractions`, `PaymentHistory` and `ClickstreamData`, we will create new features from them.

```
[ ]  # From: ServiceInteractions
     # Extract: The number of service interactions made to customer service through Email, Call, and Chat.\n",

     for usit in unique_service_interaction_type:
       df[f'ServiceInteractions_{usit}'] = df['ServiceInteractions'].apply(lambda x: len([i for i in x if i['Type'] == usit]))
```

```
     # From: PaymentHistory
     # Extract:
     #   - PaymentHistoryNoOfLatePayments: The total number of late payments made by the user\n",
     #   - PaymentHistoryAvgNoOfLatePayments: The average number of late payments made by the user\n",

     df['PaymentHistoryNoOfLatePayments'] = df['PaymentHistory'].apply(lambda x: sum([i['Late_Payments'] for i in x]))
     df['PaymentHistoryAvgNoOfLatePayments'] = df['PaymentHistory'].apply(lambda x: np.mean([i['Late_Payments'] for i in x]))
```

```
[ ]  # From: ClickstreamData
     #Extract: The number of ClickstreamData Actions performed by the user on the website, these include 'Add to Cart', 'Search' and 'Click'.\n",

     for ucda in unique_clickstream_data_actions:
       df[f'ClickStreamData_{ucda}'] = df['ClickstreamData'].apply(lambda x: len([i for i in x if i['Action'] == ucda]))
```

Initially, we addressed the conversion of nested columns, which were imported as strings, into their appropriate data types using Python's `literal_eval` function. This conversion was necessary for columns such as `PurchaseHistory`, `SubscriptionDetails`, `WebsiteUsage`, `ClickstreamData`, `EngagementMetrics`, `Feedback`, and `MarketingCommunication`, which contained complex data structures like lists and dictionaries.

Subsequently, we extracted meaningful features from these complex structures. For instance, from the `SubscriptionDetails`, we calculated the `SubscriptionDuration` by computing the difference between the subscription start and end dates. Similarly, from `WebsiteUsage`, we derived `WebsitePageViews` and `WebsiteTimeSpent`, which represent the number of page views and the total time spent on the website, respectively. These features were crucial in capturing customer engagement levels.

We also focused on encoding categorical variables into numerical formats to facilitate their use in machine learning models. This step involved converting variables such as `Segment` and `EngagementMetricsFrequency` into numerical representations using appropriate encoding techniques.
Scaling and normalization of numerical features were performed to ensure that all features contributed equally to the model's performance. We employed `StandardScaler` from the `sklearn` library to standardize the features, which is particularly important for algorithms sensitive to the scale of input data.
The dataset was then split into training, validation, and test sets. This division was crucial for model training, hyperparameter tuning, and final evaluation. We allocated 80% of the data to the training set, while the remaining 20% was split into validation and test sets, with the test set being larger to provide a comprehensive evaluation of the model's performance.

Feature engineering also involved the removal of irrelevant features that did not contribute to the predictive power of the model. This step was guided by correlation analysis, which identified features with significant relationships to the target variable, `ChurnLabel`. For instance, features such as `PaymentHistoryNoOfLatePayments` and `ServiceInteractions` exhibited strong correlations with churn, indicating their importance in the model.

Through these preprocessing and feature engineering efforts, we ensured that the dataset was optimally structured for the subsequent model development phase, setting a solid foundation for accurate and reliable churn prediction. This preparation was pivotal in enhancing the model's ability to identify customers at risk of churning, thereby supporting Reder Telecom's strategic objectives.

Model Development

In the model development phase, we focused on constructing predictive models to accurately forecast customer churn for Reder Telecom. The primary objective was to identify customers at risk of discontinuing services, thereby enabling the company to implement targeted retention strategies. We employed two machine learning models: Logistic Regression and Decision Tree Classifier, chosen for their interpretability and effectiveness in classification tasks.

The dataset was divided into training, validation, and test sets to ensure robust model evaluation. The training set was used to fit the models, while the validation set helped in tuning model parameters and preventing overfitting. The test set provided an unbiased evaluation of the model's performance.

For model evaluation, we utilized four key metrics: accuracy, precision, recall, and F1 score. These metrics provided a comprehensive view of the model's ability to correctly classify churn and non-churn customers. The Decision Tree model demonstrated superior performance across most metrics on both validation and test datasets, with an accuracy score of 97.37% on the test set, compared to 96.91% for Logistic Regression. Precision and recall scores also favored the Decision Tree, indicating its effectiveness in identifying true churn cases.

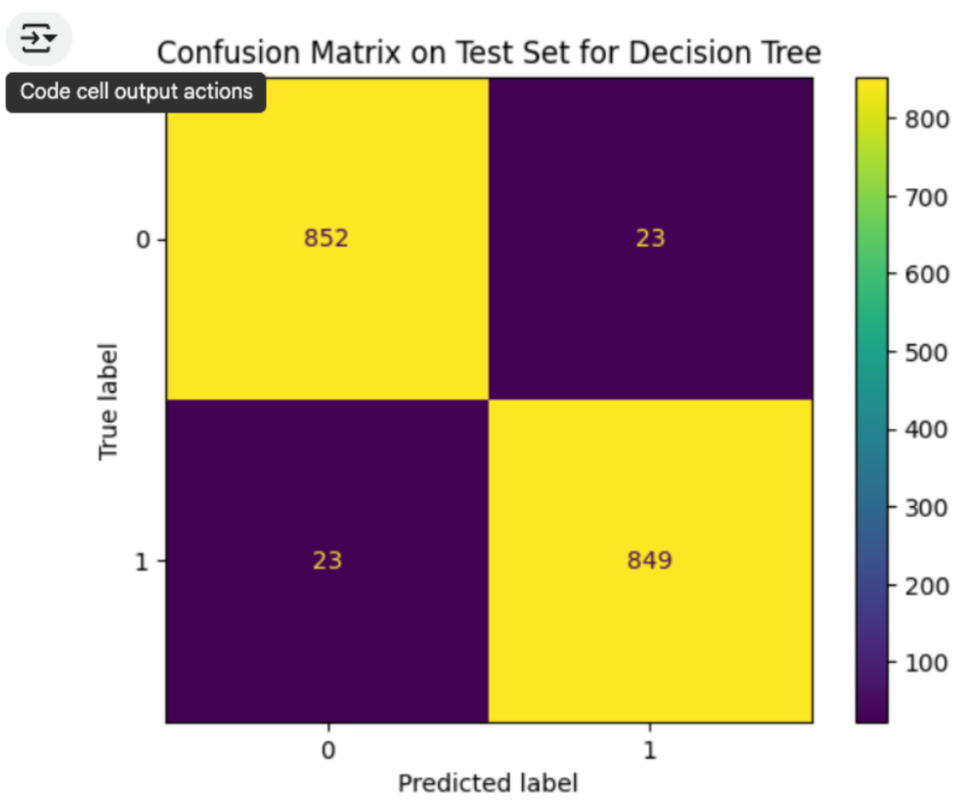
The confusion matrix further illustrated the models' performance, revealing that both models excelled in predicting non-churn customers. However, the Decision Tree model showed a higher capability in correctly identifying churned customers, which is critical for proactive retention efforts.

Model Selection

Two models were selected: Logistic Regression and Decision Tree.

Model Evaluation

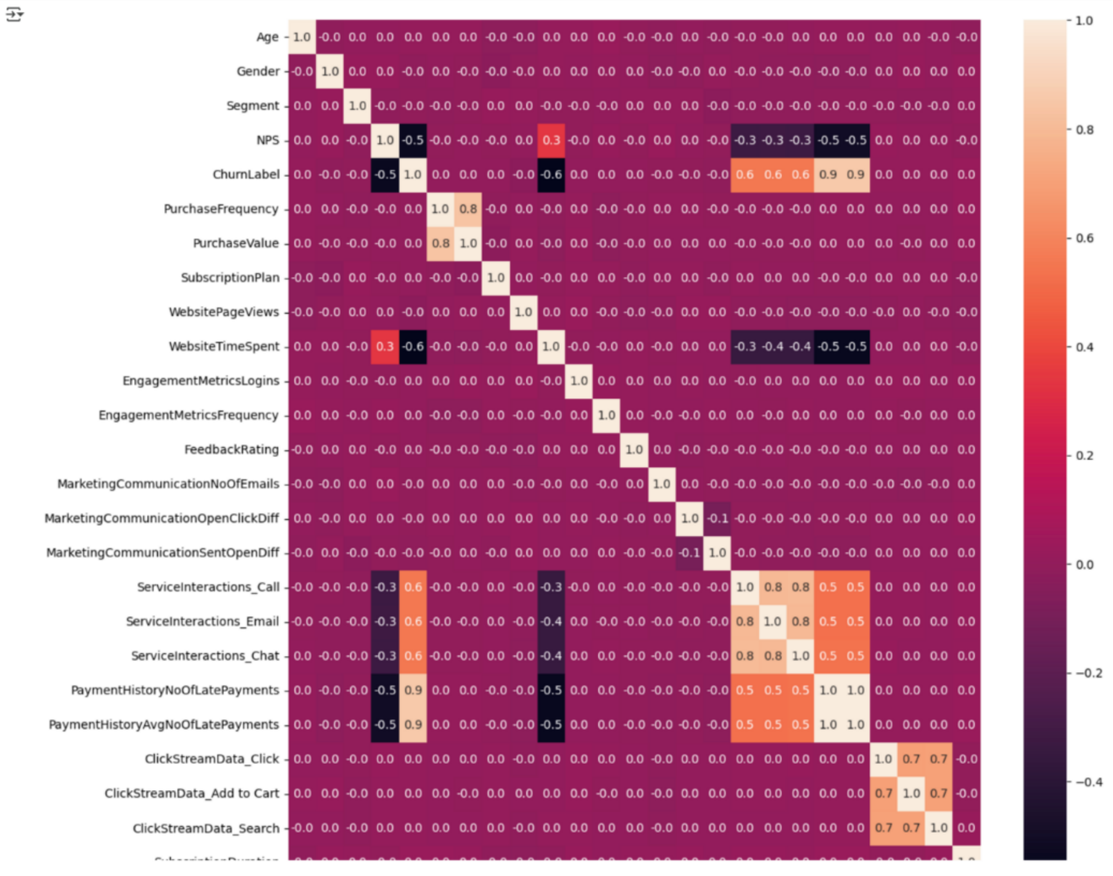
The Decision Tree model outperformed Logistic Regression.

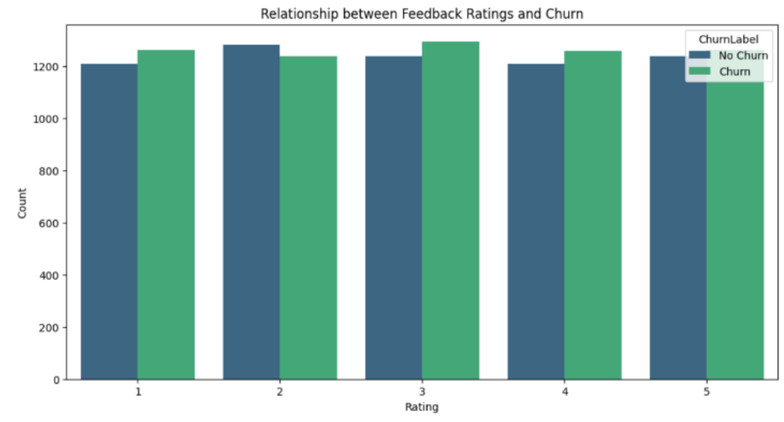


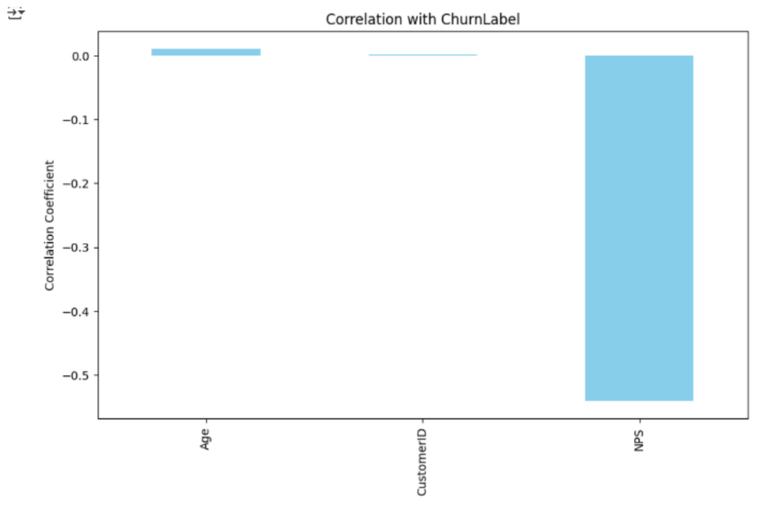
Data Insights

Data Visualization

The data analysis revealed interesting trends and relationships.







Firstly, the distribution of the target variable, ChurnLabel, indicates a balanced dataset, with nearly equal proportions of churned and non-churned customers. This balance is advantageous for model training, as it reduces the risk of bias towards either class. Visualizations of the demographic variables, such as Gender and Segment, reveal an even distribution across categories, suggesting that churn is not significantly skewed by these factors alone.

A correlation analysis highlights the Net Promoter Score (NPS) as a significant predictor of churn, with a strong negative correlation. This suggests that customers with lower NPS are more likely to churn, emphasizing the importance of customer satisfaction and loyalty in retention strategies. Additionally, the number of Service Interactions and Late Payments are positively correlated with churn, indicating that frequent service issues and payment delays are potential churn drivers.

Visualizations of the age distribution and monthly churn rates provide further insights. The age distribution shows fluctuations, but no specific age group is disproportionately affected by churn. The monthly churn rate analysis reveals consistent fluctuations over time, without clear seasonal patterns, suggesting that churn is influenced by factors other than time of year. The relationship between customer feedback ratings and churn was also examined. Lower feedback ratings are associated with higher churn rates, reinforcing the need for Reder Telecom to address customer concerns promptly and effectively to enhance satisfaction and reduce churn.

In summary, the data analysis underscores the critical role of customer satisfaction, service quality, and payment timeliness in influencing churn. These insights, supported by visualizations, provide a foundation for developing targeted strategies to improve customer retention. The subsequent section will delve into specific recommendations for Reder Telecom, leveraging these insights to formulate actionable strategies for reducing churn and enhancing customer loyalty.

Recommendations

Enhance Customer Service Quality

Given the strong correlation between frequent service interactions and churn, Reder Telecom should improve customer support efficiency through enhanced training, AI-driven support tools, and an advanced CRM system to proactively manage customer concerns.

Address Late Payments Proactively

Implementing flexible payment plans, automated reminders, and incentives for timely payments can help reduce the churn risk associated with payment delays.

Optimize Website Engagement

Enhancing the user experience through a more intuitive, personalized, and responsive website can increase customer satisfaction and retention.

Leverage Customer Feedback and NPS Monitoring

Regularly track the Net Promoter Score (NPS) and establish a structured feedback loop to improve service offerings based on customer input.

Deploy Targeted Retention Strategies

Utilize predictive analytics to identify at-risk customers early and implement personalized retention efforts, such as exclusive offers and tailored service packages.

Invest in Predictive Model Refinement

Continuously improve churn prediction models by incorporating new data sources and refining analytical techniques to stay ahead of evolving customer behavior trends.

By implementing these recommendations, Reder Telecom can effectively reduce customer churn, enhance customer satisfaction, and secure a stronger position in the competitive telecommunications market.