MACHINELEARNING

Telecom Customer Churn
Prediction and Segmentation

Introduction

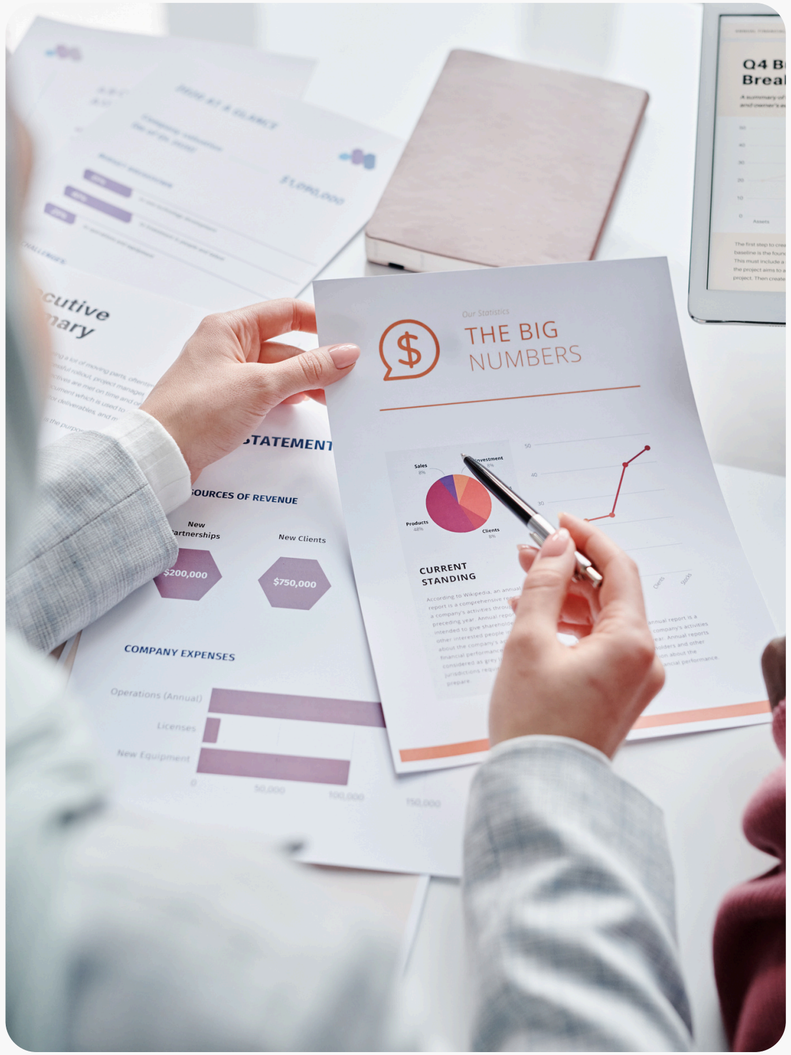
Our project aims to analyze customer behavior in the telecom industry to predict and reduce customer churn. Churn is a significant issue for telecom companies, as losing customers impacts revenue and market share. By implementing machine learning techniques, we intend to develop predictive models that help identify at-risk customers and propose strategies to improve customer retention.

Problem Solution

Dataset Overview:

We are using a telecom dataset that includes customer details like demographics, service usage, contract types, and payment methods. This data will help us understand customer behavior, predict who might churn, and group customers into segments. Before applying machine learning models, we'll clean and prepare the data to ensure accurate results

Expected Outcomes



Improved Customer Experience:

- Gain insights into customer satisfaction, leading to enhanced service offerings and reduced churn

Churn Prediction

- Identify customers likely to leave the service, allowing for targeted retention strategies.

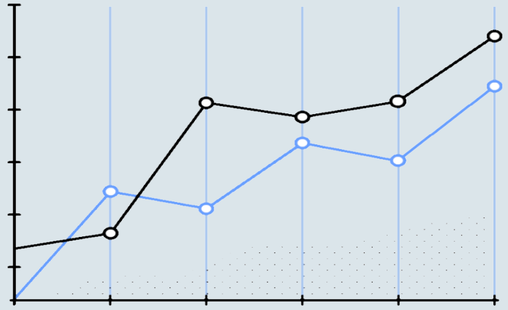
Operational Insights:

- Understand service usage patterns that correlate with churn, enabling better resource allocation and service enhancements.

Tools & Techniques

Data Preprocessing

- Used Pandas to load the dataset efficiently from a CSV file.
- Utilized Pandas for handling missing values, encoding categorical features, and ensuring data quality for analysis.



Visualization Tools:

- We used Python libraries like Matplotlib, Seaborn and Plotly Express to create visualizations that help us explore the data and present insights clearly.

Machine Learning Models

- RandomForestClassifier
- Logistic Regression
- DecisionTreeClassifier
- XGBoost Classifier
- SVC
- GaussianNB

Data Preprocessing

1.Handling Missing Values and duplicated data:

"We checked the dataset for duplicates and removed them. We also looked for any missing values and handled them by either removing incomplete records or filling in the gaps."

2. Outlier Detection:

We examined the data for outliers, which could distort model performance, and decided whether to remove or transform these values to improve accuracy.

3. Categorical Encoding:

We used two methods to convert categorical data:

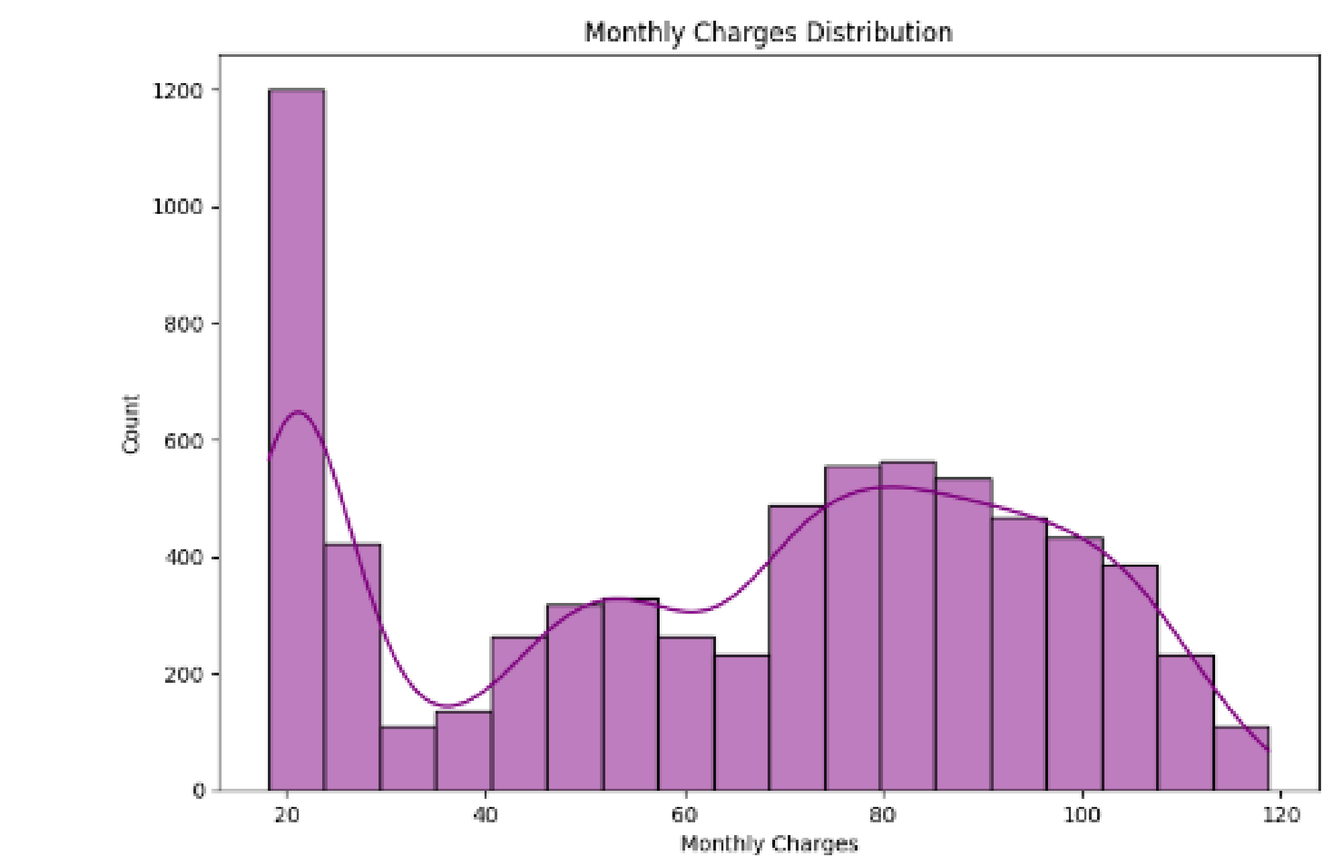
- Label Encoding for: 'PhoneService', 'gender', 'Partner', 'Dependents', 'PaperlessBilling', and 'Churn'.
- One-Hot Encoding for: 'PaymentMethod', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', and 'Contract'.

4. Feature Engineering:

- We created New feature ‘Eligibleforservice’ .This feature identifies customers eligible to subscribe to our service based on their internet service types.

Data Visualization

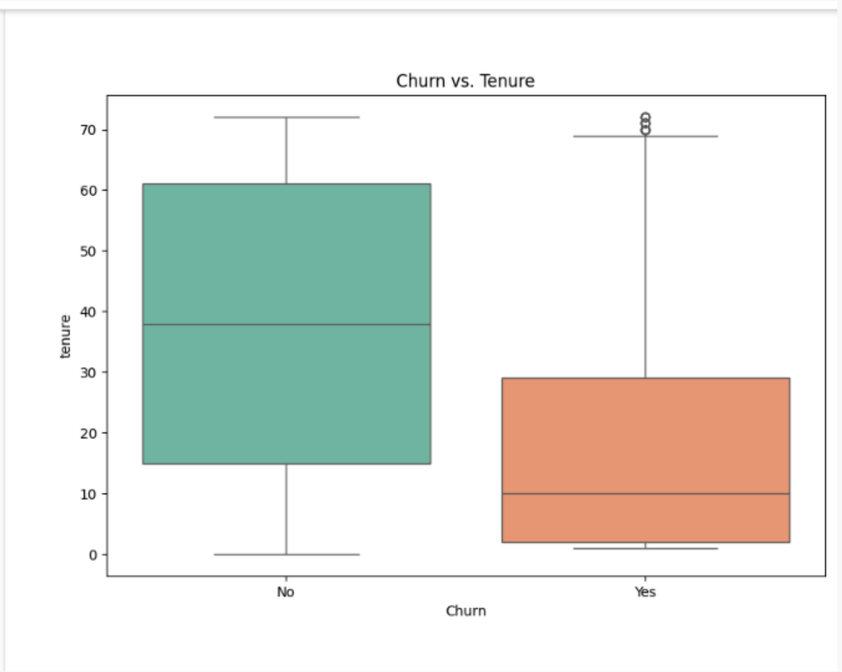
Monthly Charges Distribution (Histogram)

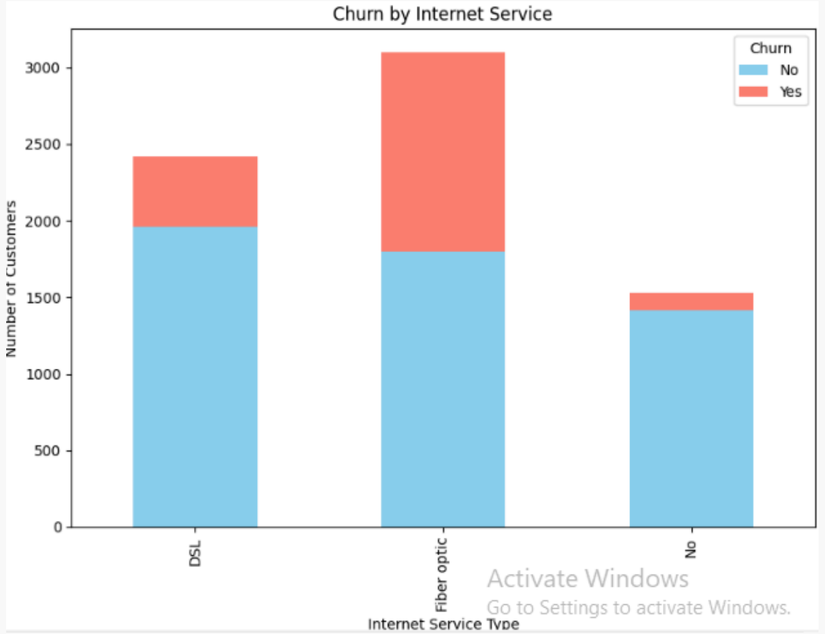


This histogram illustrates the distribution of monthly charges among customers. The inclusion of a kernel density estimate (KDE) curve helps smooth the distribution and highlight the underlying trends. The chart helps us identify any skewness in the charges and potential outliers.

Data Visualization

Churn Analysis and Correlations





The boxplot compares customer tenure between those who churned and those who didn't, revealing any potential tenure-related churn patterns. The bar plot, comparing churn rates by Internet service type, highlights whether specific services lead to higher churn rates, offering actionable insights for targeted service improvements.

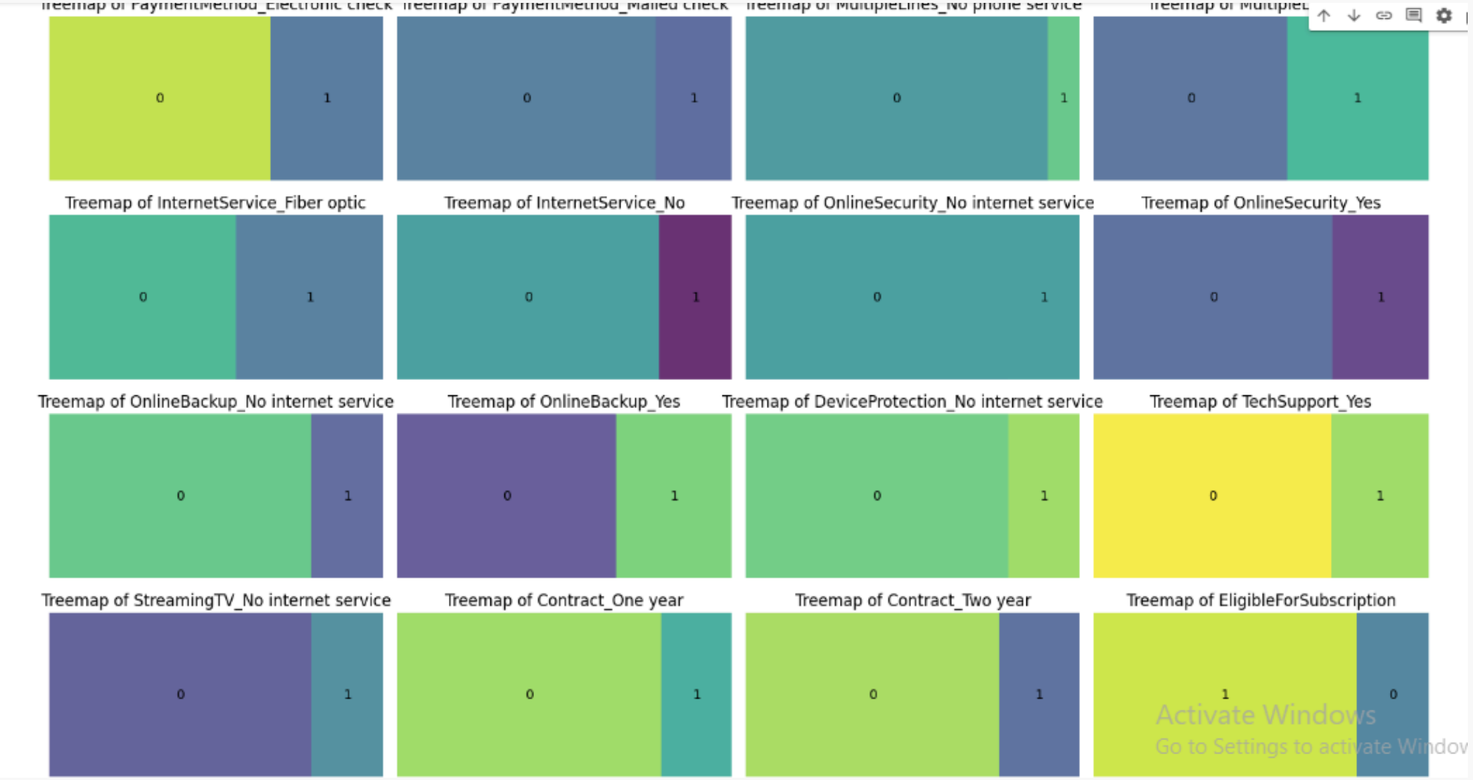
Data Visualization

Scatter Matrix of Customer Features and Churn Correlations

We selected key features like tenure, monthly charges, and Internet service types, including one-hot encoded columns. A scatter matrix plot was created to show relationships between these features, with churn status highlighted by color. This helps us spot potential patterns before modeling

Data Visualization

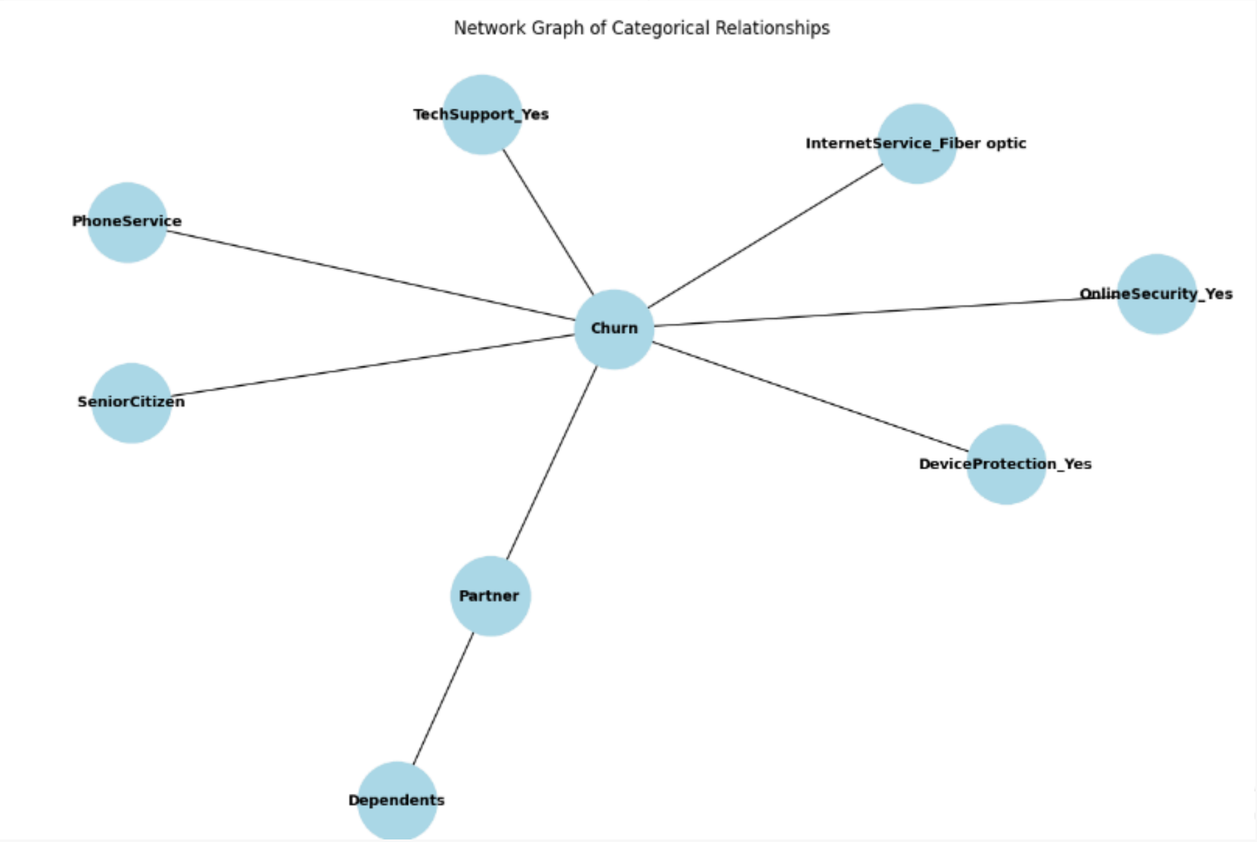
Treemaps of Categorical Feature Distributions



We created treemaps for all categorical variables to visualize the distribution of each category. Each treemap shows the proportion of different values within a variable, helping to quickly identify dominant categories. This was done for multiple features in the dataset, offering an overview of their composition

Data Visualization

Network Graph of Categorical Relationships

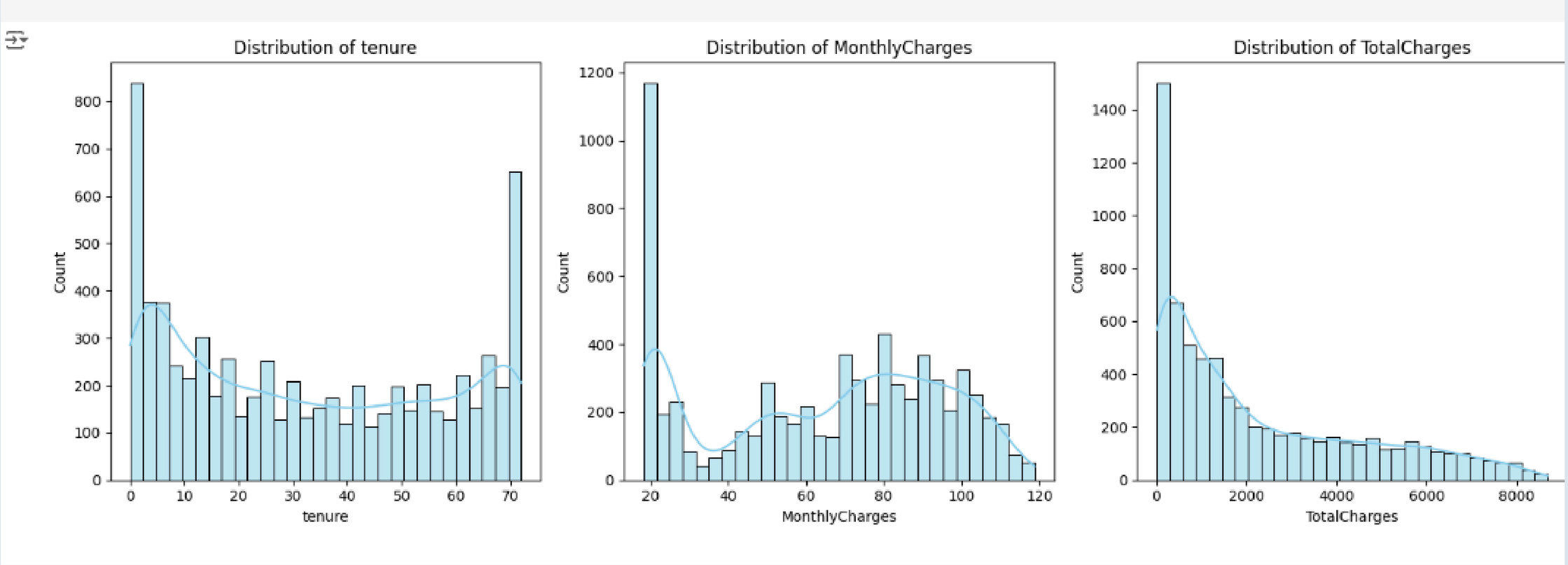


We constructed a network graph to display relationships between key categorical features, such as Partner status, Senior Citizen, and various services. The connections highlight how these features relate to customer churn, providing a clear visual representation of categorical dependencies and interactions in the dataset

Data Exploratory analysis (EDA)

1. Univariate Analysis

purpose: Explore the distribution of individual variables.

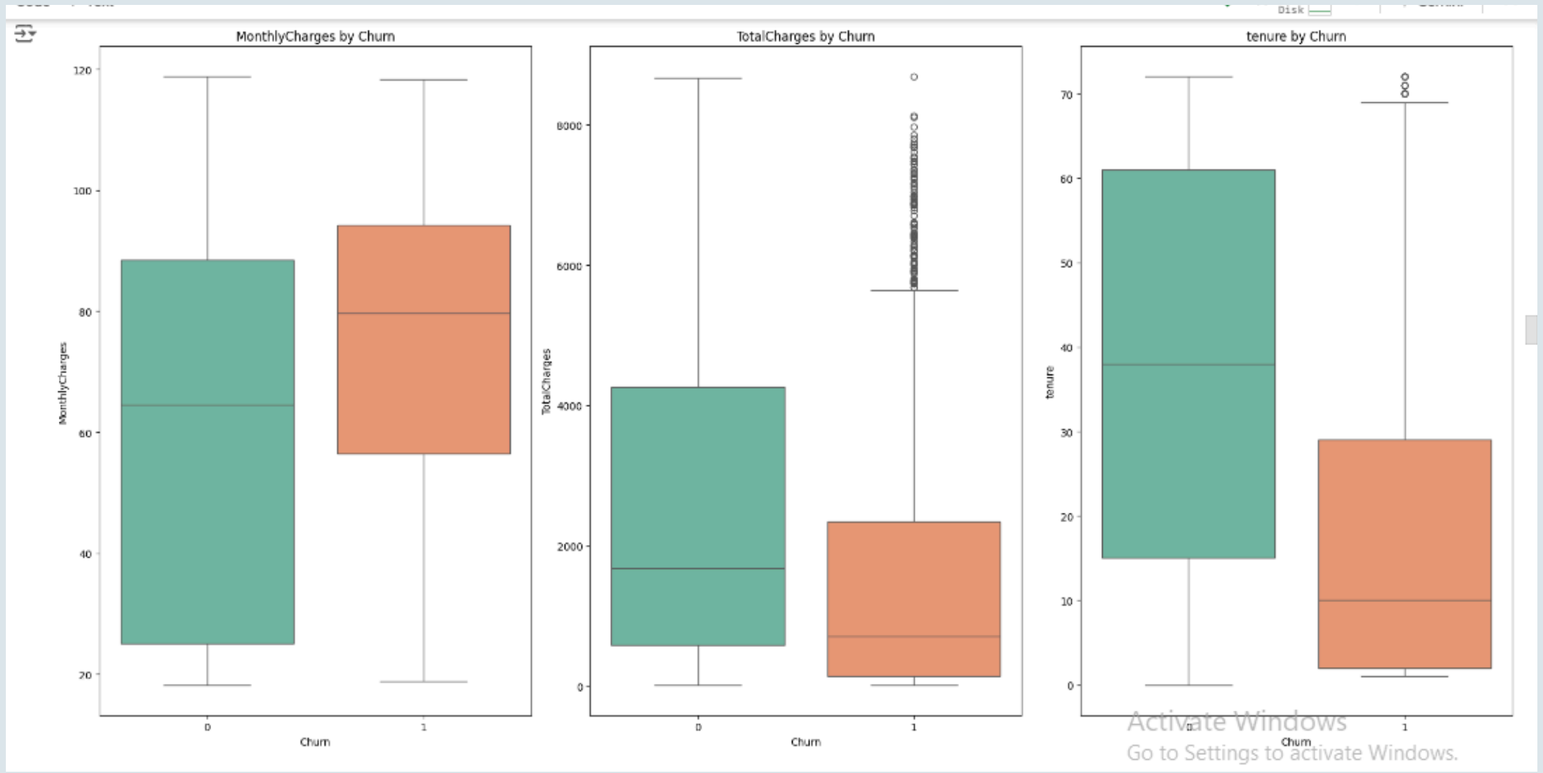


Insights: Provided an understanding of each feature's distribution, such as customer tenure, monthly charges, and churn rates.

Data Exploratory analysis (EDA)

2. Bivariate Analysis

Purpose: Understand relationships between two variables.

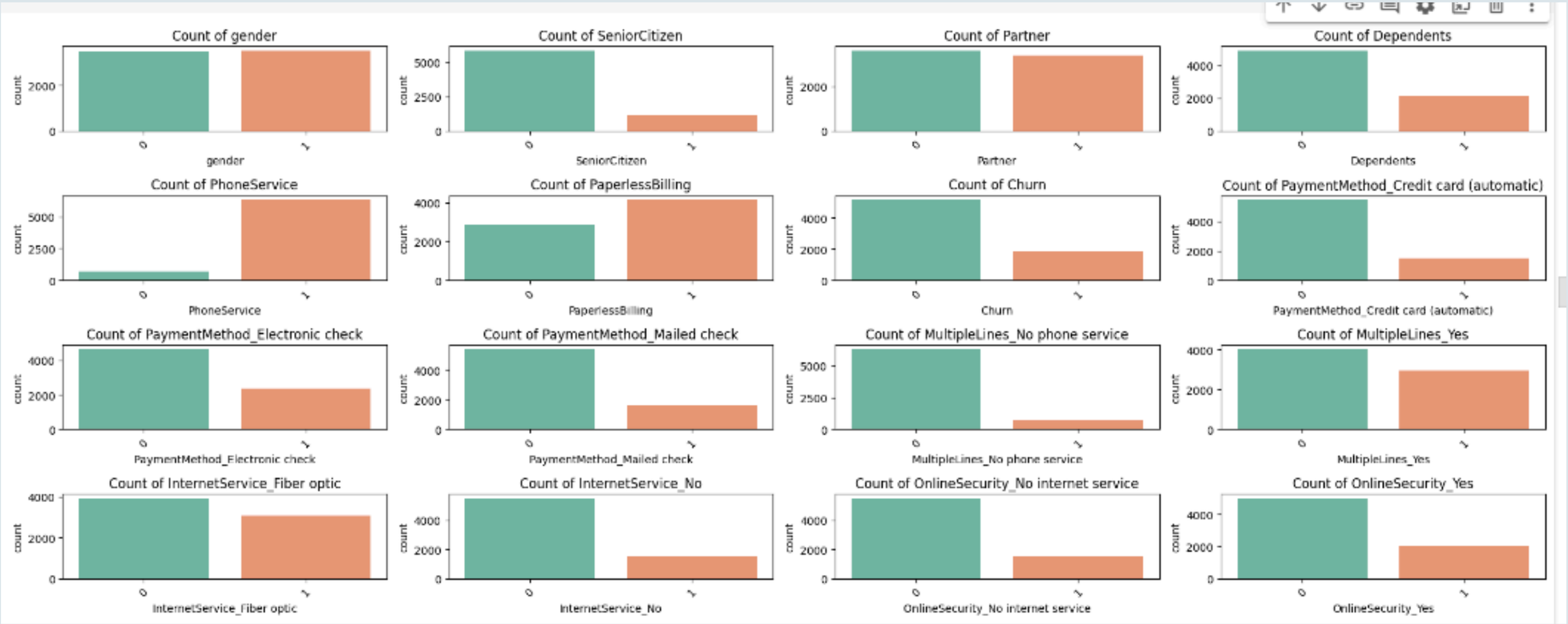


Insights: Helped uncover key patterns, like customers with shorter tenure were more likely to churn.

Data Exploratory analysis (EDA)

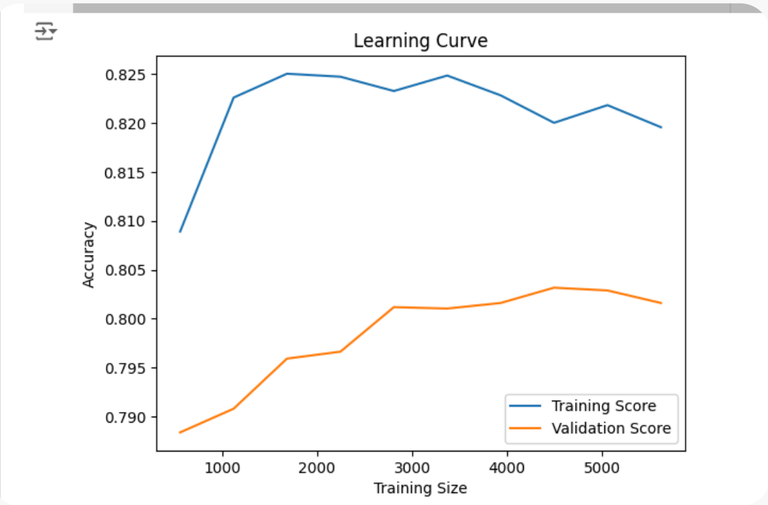
3. Categorical Analysis

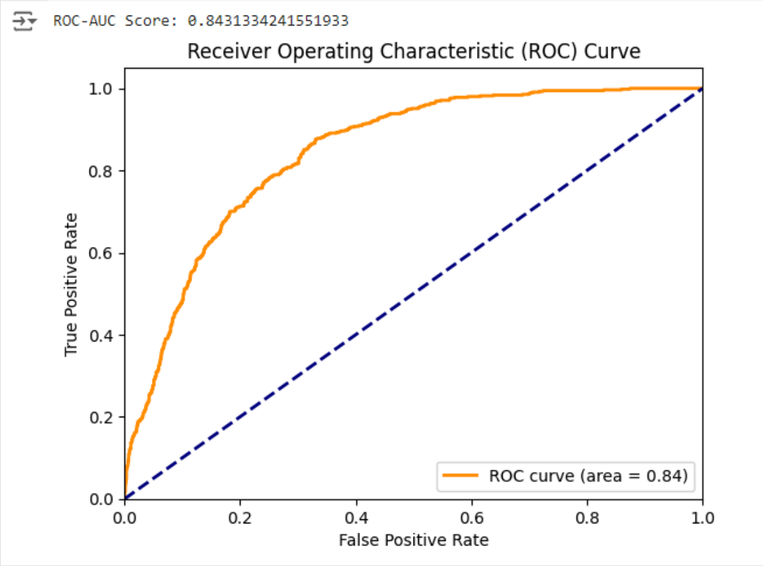
- Purpose: Analyze relationships among categorical variables.



- Insights: Identified strong patterns, such as customers with month-to-month contracts showing higher churn rates.

Machine Learning Models



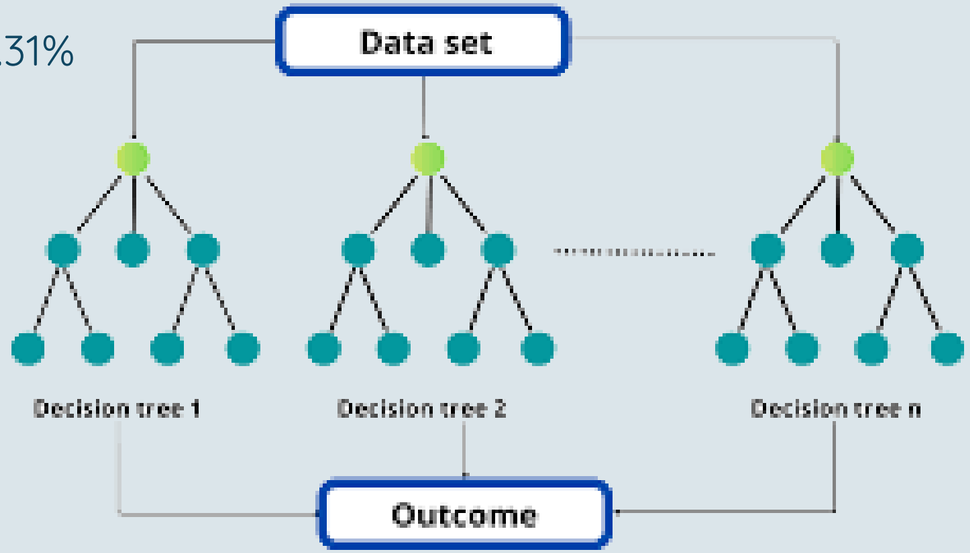


1. RandomForestClassifier

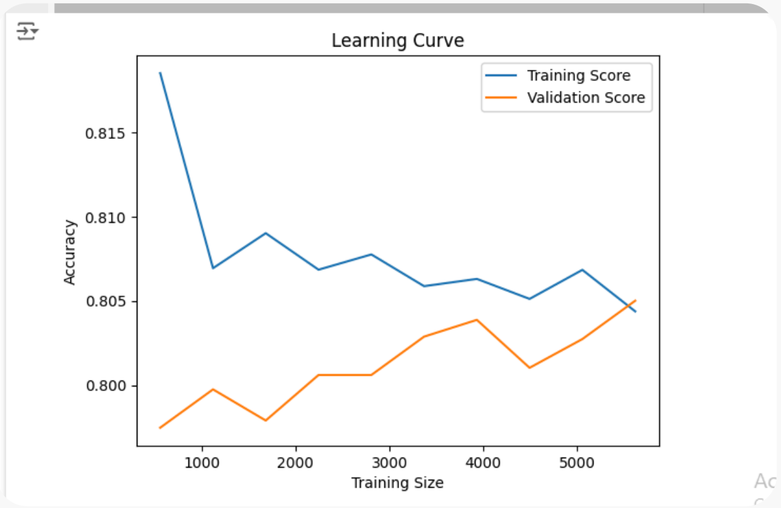
- **Description**: Ensemble of decision trees
- **Strength**: Robust and reduces overfitting

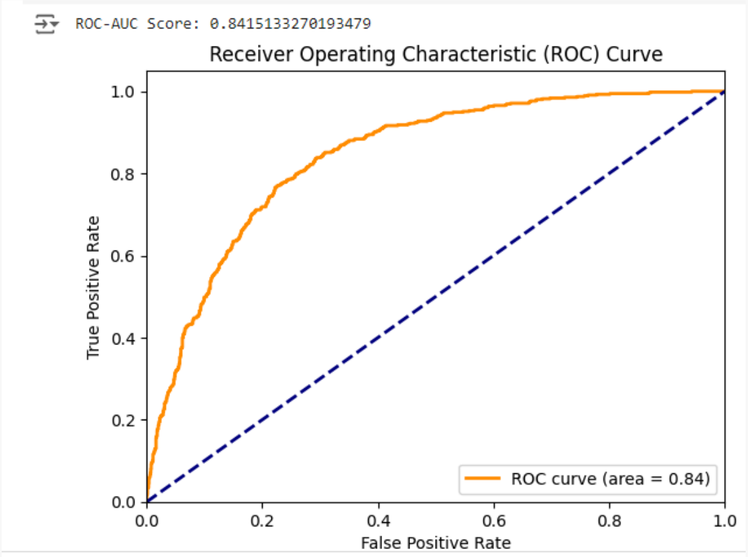
Performance Metrics

- Accuracy: 79%
- Precision: 80%
- Recall: 79%
- F1 Score: 79%
- ROC-AUC: 84.31%



Machine Learning Models



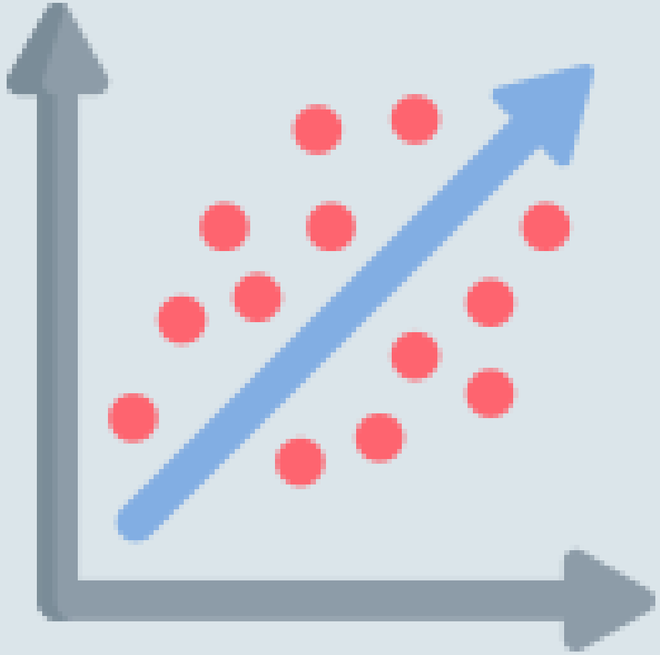


2. Logistic Regression

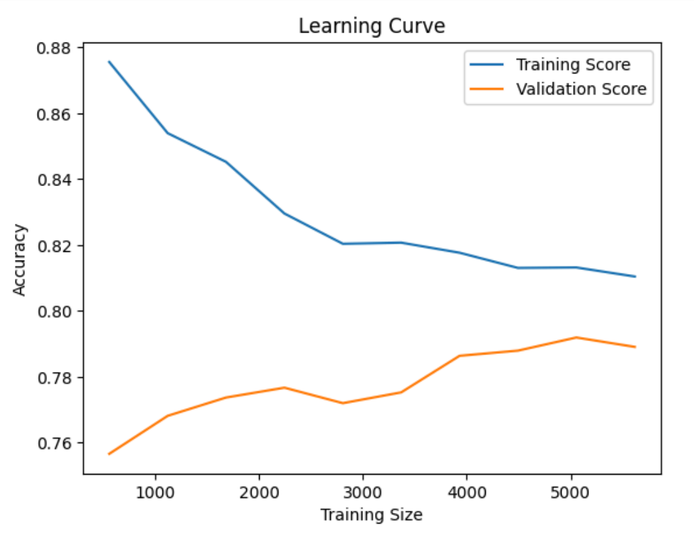
- **Description**: Used for binary classification
- **Strength**: Simple and interpretable model.

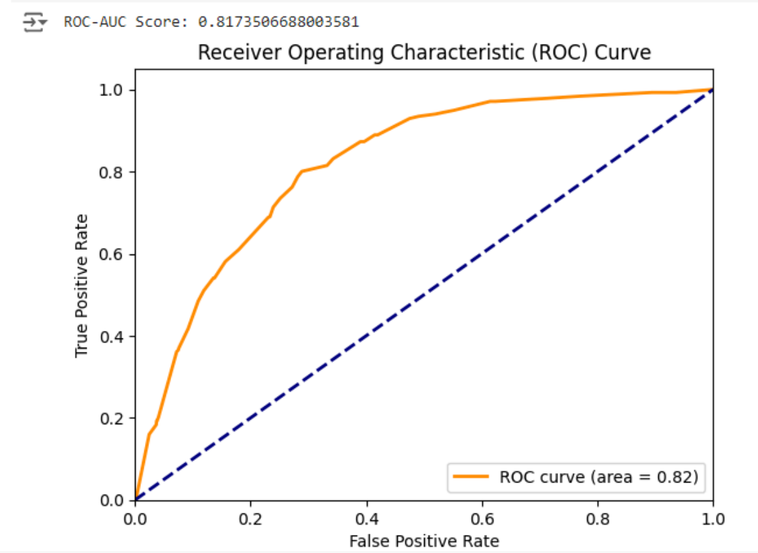
Performance Metrics

- Accuracy: 79%
- Precision: 80%)
- Recall: 79%
- F1 Score: 79%
- ROC-AUC: 84.15%
-



Machine Learning Models



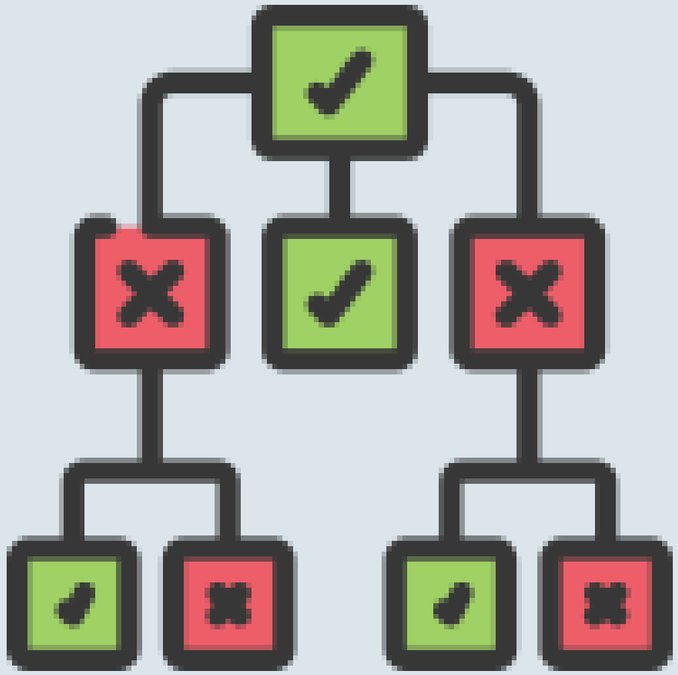


3. Decision Tree Classifier

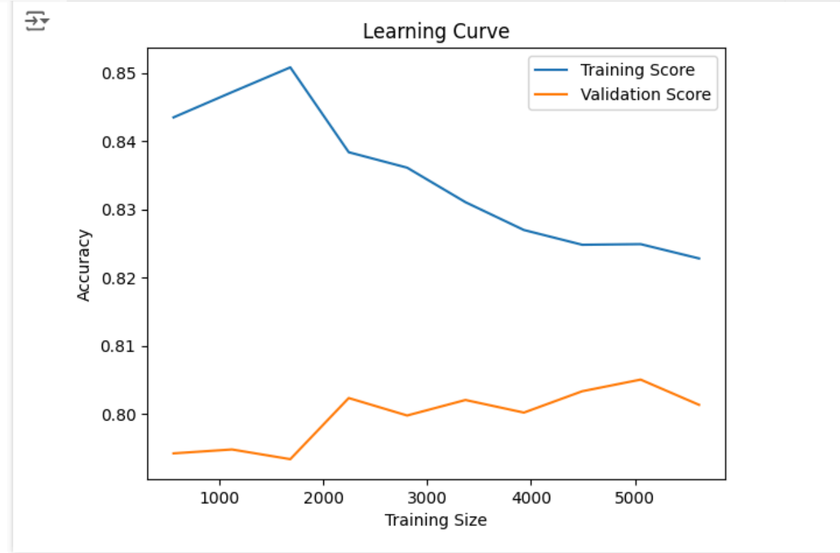
- **Description**: Tree-based model for classification and regression
- **Strength**: Easy to visualize and understand

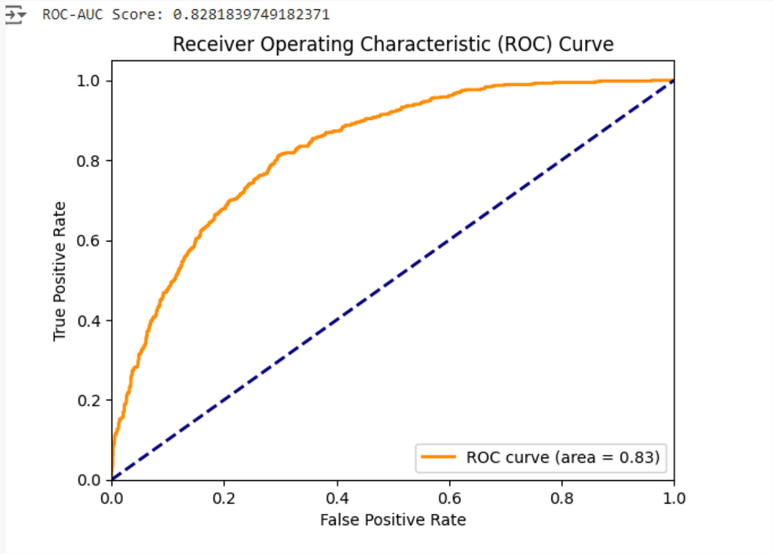
Performance Metrics

- Accuracy: 79%
- Precision: 80%
- Recall: 79%
- F1 Score: 79%
- ROC-AUC: 81.74%



Machine Learning Models





4. XGBoost Classifier

- **Description**: Used for binary classification
- **Strength**: Simple and interpretable model.

Performance Metrics

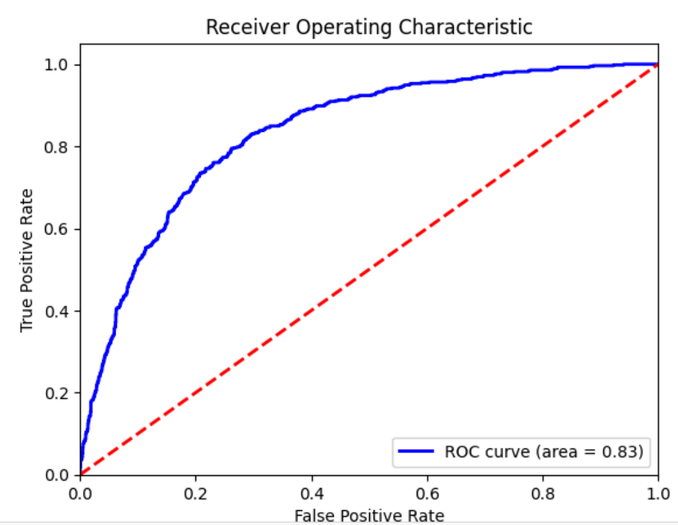
- Accuracy: 78%
- Precision: 79%
- Recall: 78%
- F1 Score: 78%
- ROC-AUC: 82.82%

XGBOOSt Algorthim

Machine Learning Models

5. SVC

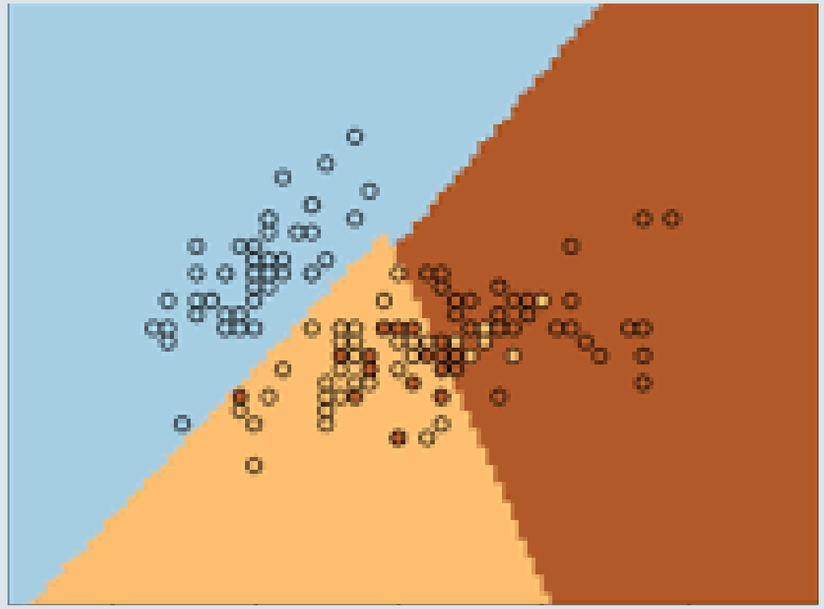
```
SVC Training accuracy: 0.8261
SVC Test accuracy: 0.7869
SVC Test Set Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.83      0.85      1556
           1       0.58      0.67      0.62       551

    accuracy                           0.79      2107
   macro avg       0.73      0.75      0.74      2107
weighted avg       0.80      0.79      0.79      2107

SVC ROC-AUC: 0.8346
```



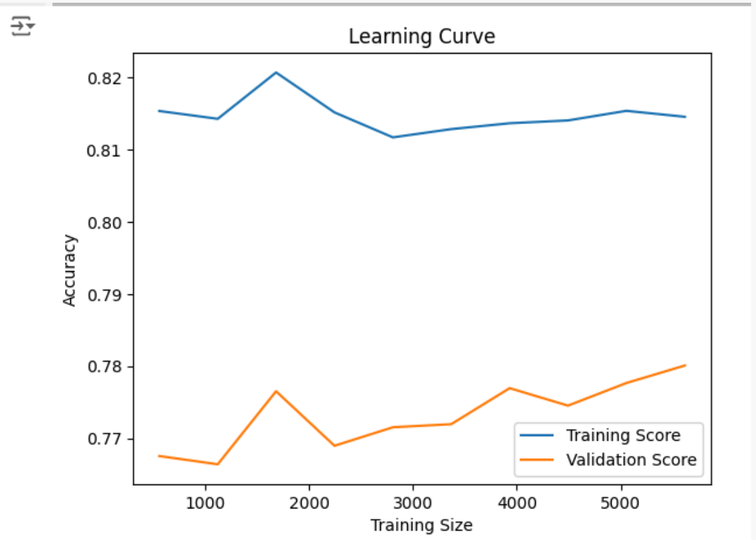
- **Description**: Effective in high-dimensional spaces
- **Strength**: Good for non-linear boundaries

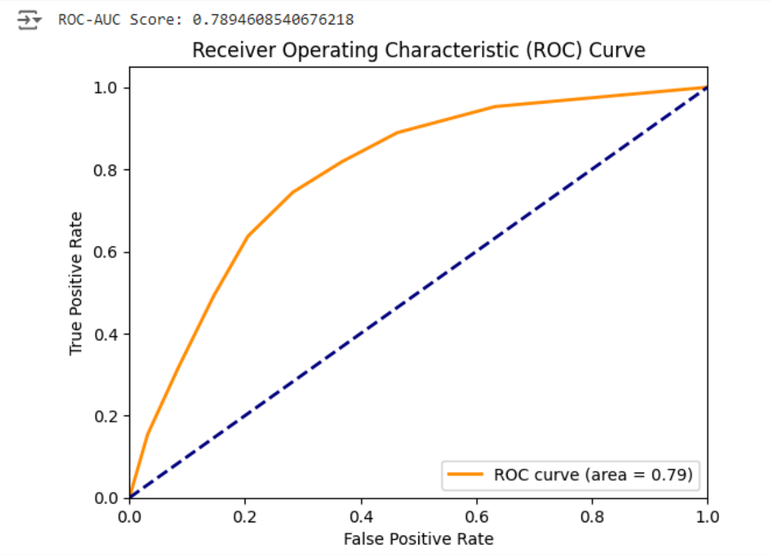
Performance Metrics

- Accuracy: 79%
- Precision: 80%
- Recall: 79%
- F1 Score: 79%
- ROC-AUC: 83.46%
-



Machine Learning Models



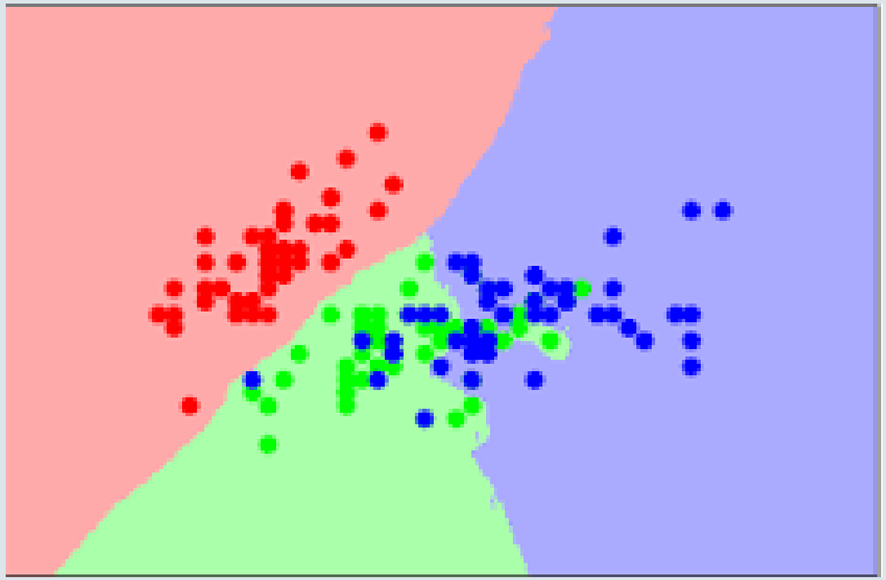


6. KNeighborsClassifier

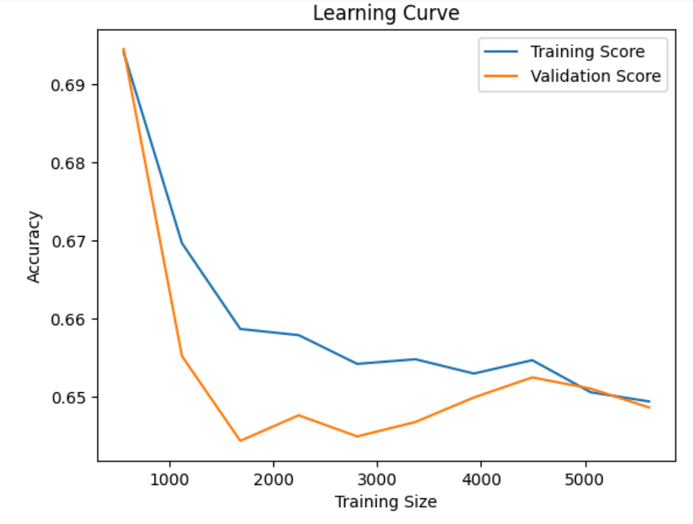
- **Description**: Instance-based learning
- **Strength**: Simple and intuitive

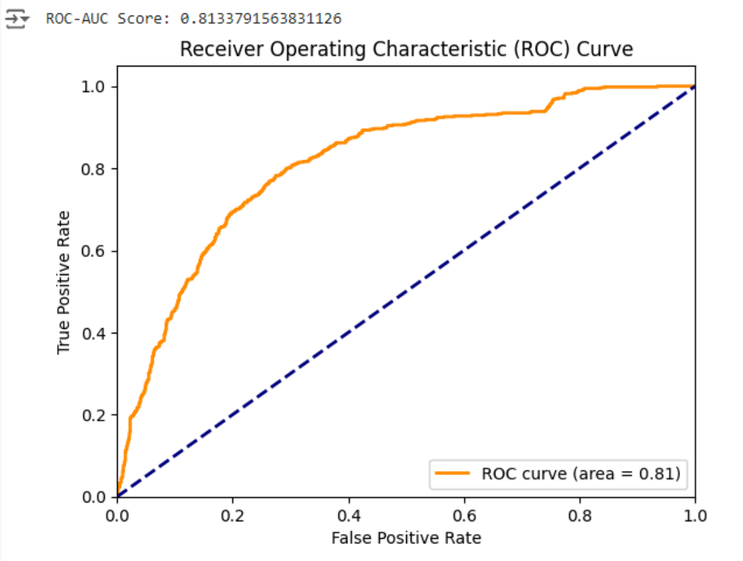
Performance Metrics

- Accuracy: 75%
- Precision: 77%
- Recall: 75%
- F1 Score: 76%
- ROC-AUC: 78.95%



Machine Learning Models





7.GaussianNB

- **Description**:Probabilistic model based on Bayes' theorem
- **Strength**: Fast and efficient for large datasets

Performance Metrics

- Accuracy: 69%
- Precision: 80%
- Recall: 69%
- F1 Score: 71%
- ROC-AUC: 83.46%

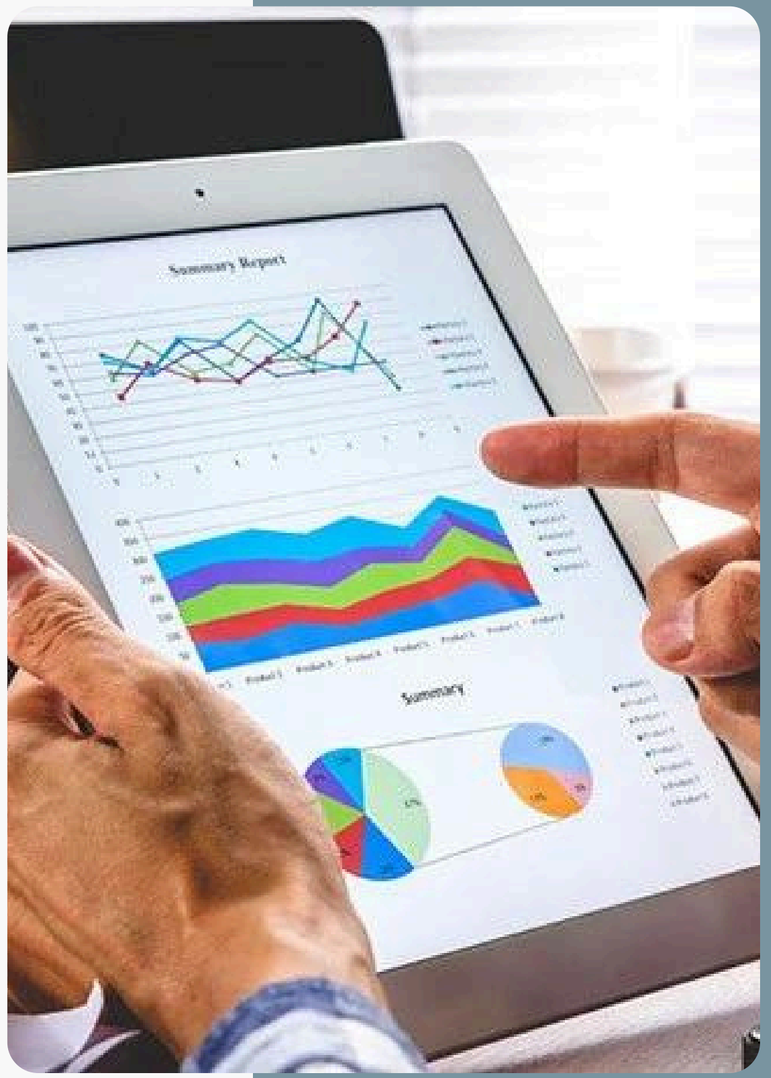
Gaussian NB

Unsupervised Learning

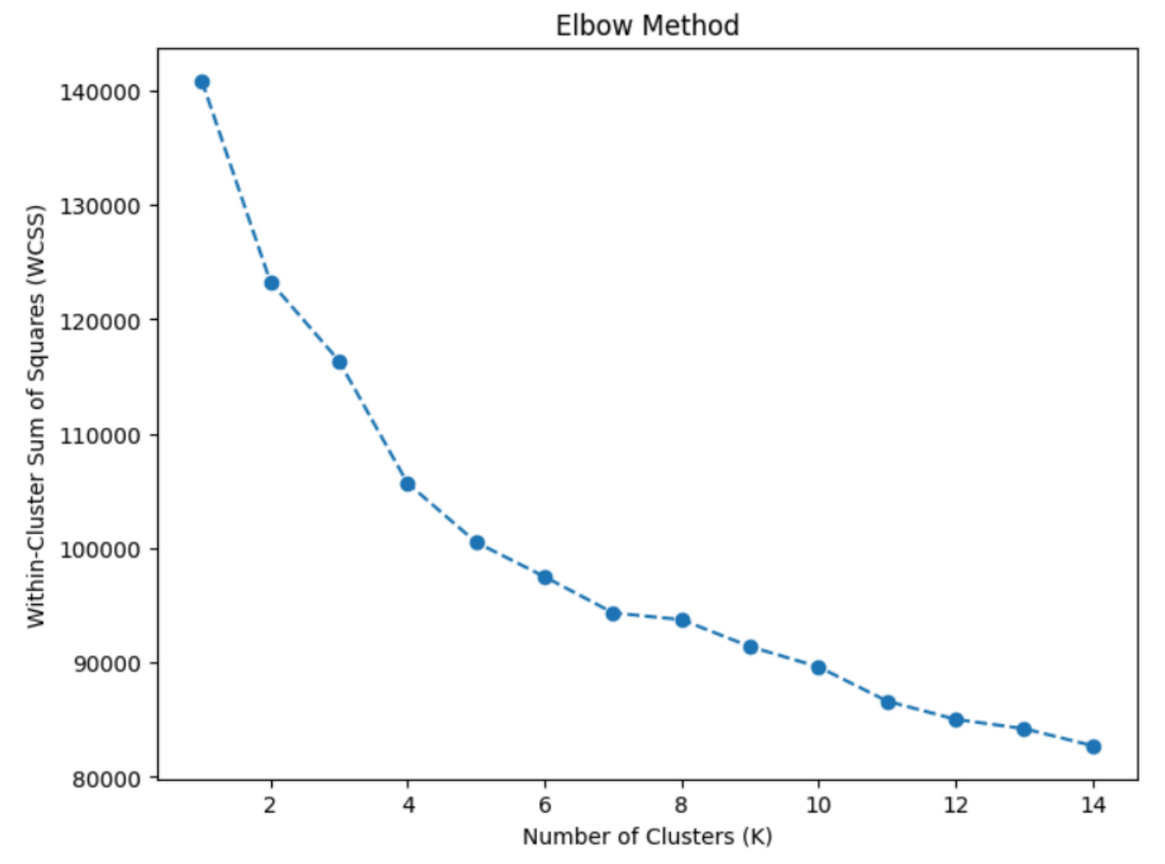
- **Objective:**

The goal of this project is to segment telecom customers based on their usage patterns and demographic characteristics. By grouping customers into distinct clusters, we can gain deeper insights into their behaviors, identify potential churn risks, and create targeted marketing strategies.

- **Key Techniques:**
  - Clustering: We employ K-Means and Agglomerative Clustering algorithms to group customers based on their similarity.
  - Dimensionality Reduction: Principal Component Analysis (PCA) is used to reduce the complexity of the data and visualize the clusters in a two-dimensional space.
  - Performance Evaluation: The Silhouette Score is utilized to assess the quality of the clustering, ensuring that the clusters are well-defined and meaningful.



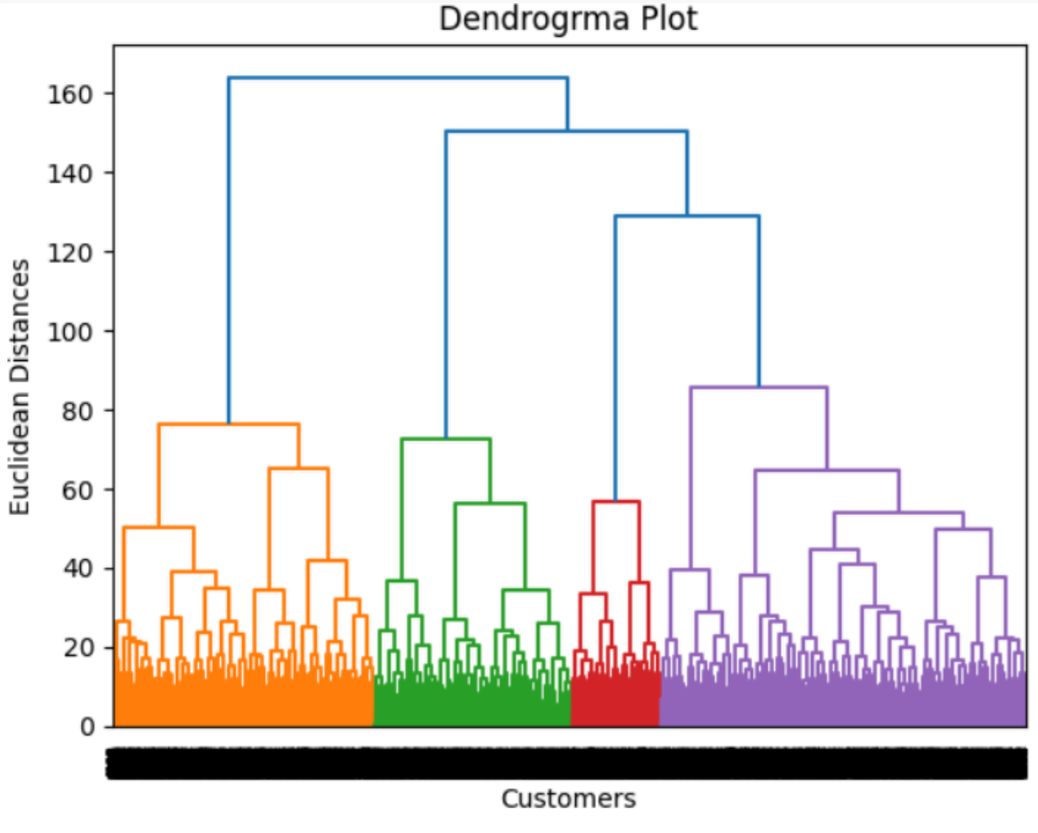
Machine Learning Models



1-K-Means Clustering

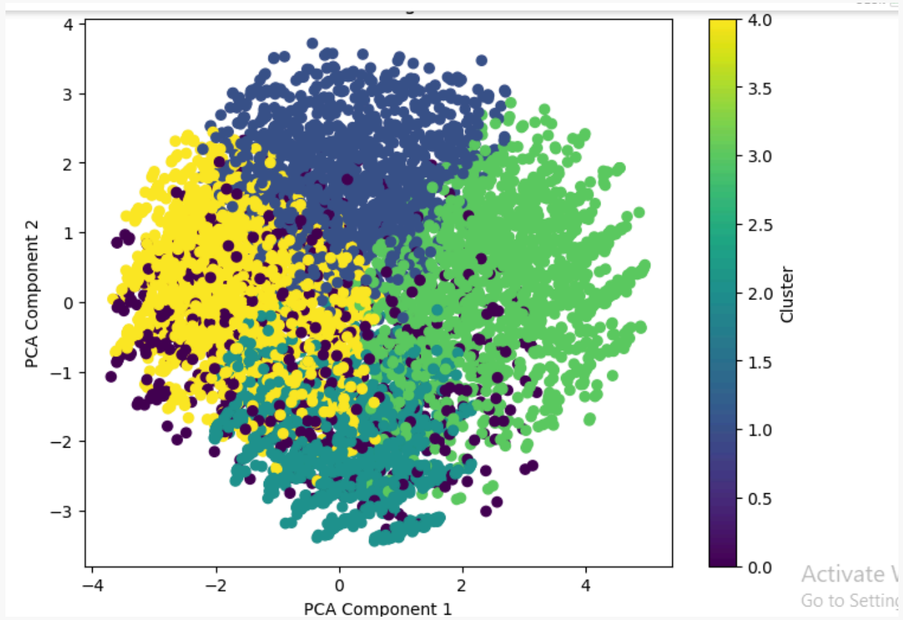
- **Objective:** Group customers into clusters based on their behavior.
- **Elbow Method:** Determine the optimal number of clusters (K) by plotting Within-Cluster Sum of Squares (WCSS).
- **Result:** Selected K = 5 as optimal
- **Silhouette Score:**
  - **Purpose:** To measure the quality of the clustering by assessing how well each data point fits within its assigned cluster and how distinct each cluster is from others.
  - **Result:** 0.13

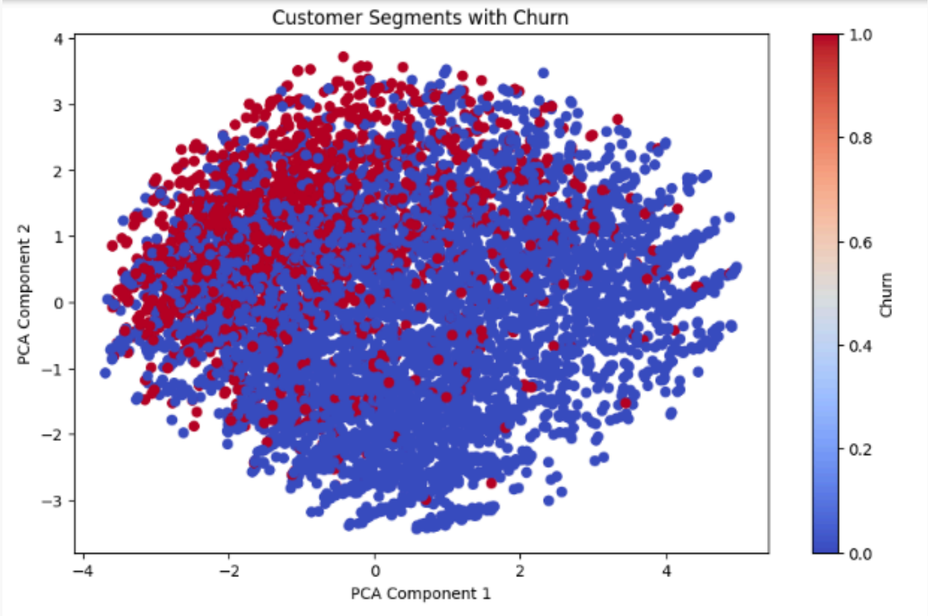
Machine Learning Models



2-Hierarchical Clustering

- **Objective:** Compare clustering methods.
- **Method:** Performed Agglomerative Clustering and plotted a Dendrogram to visualize cluster formation.
- **Silhouette Score:** 0.11





Principal Component Analysis (PCA)

1. Purpose: Reduce data dimensionality to visualize customer segments.

2. Approach:

- Used PCA with 2 components to project high-dimensional data onto a 2D space.
- Visualized the clusters and customer churn across principal components.

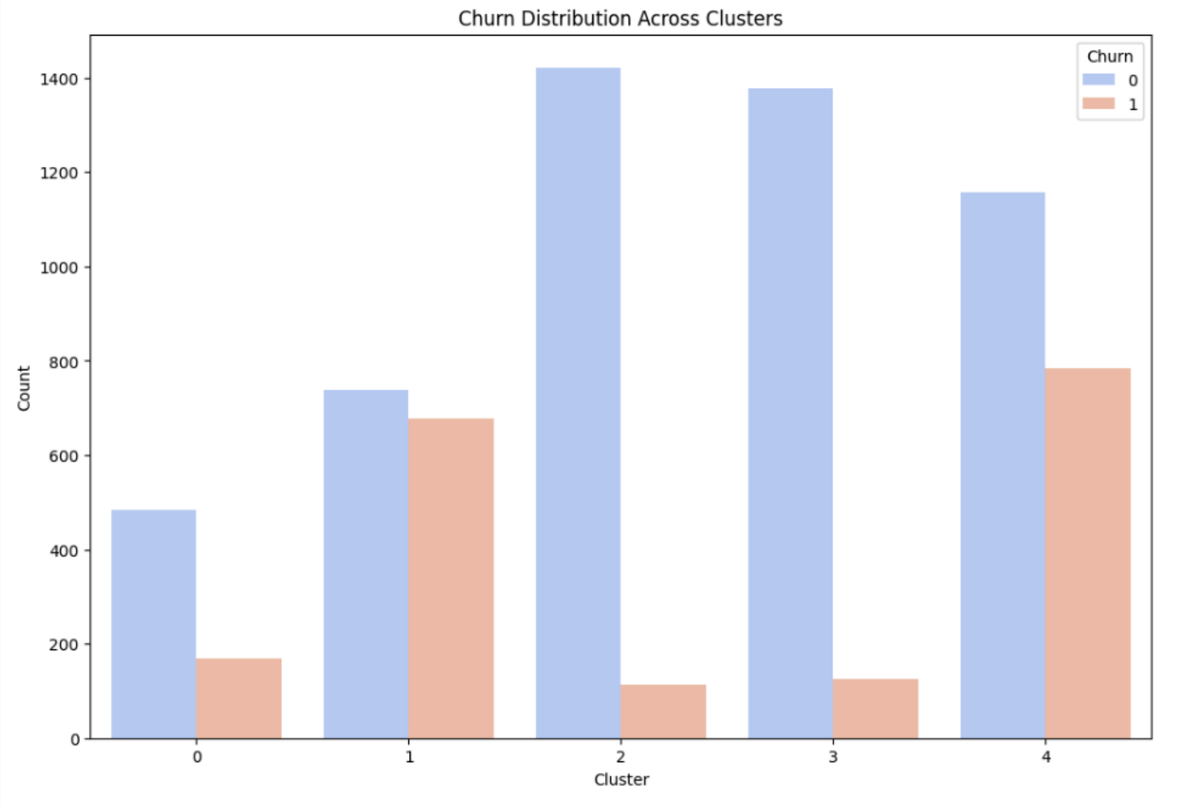
3. Visualization:

- Scatter plots display customer segments and clusters based on the principal components.
- Component 1 and Component 2 serve as axes, highlighting key variations in the data.
- components.

Churn Distribution Across Clusters

- **understand Churn Behavior**: Analyze the distribution of churn across different customer clusters to gain insights into which clusters have higher or lower churn rates.

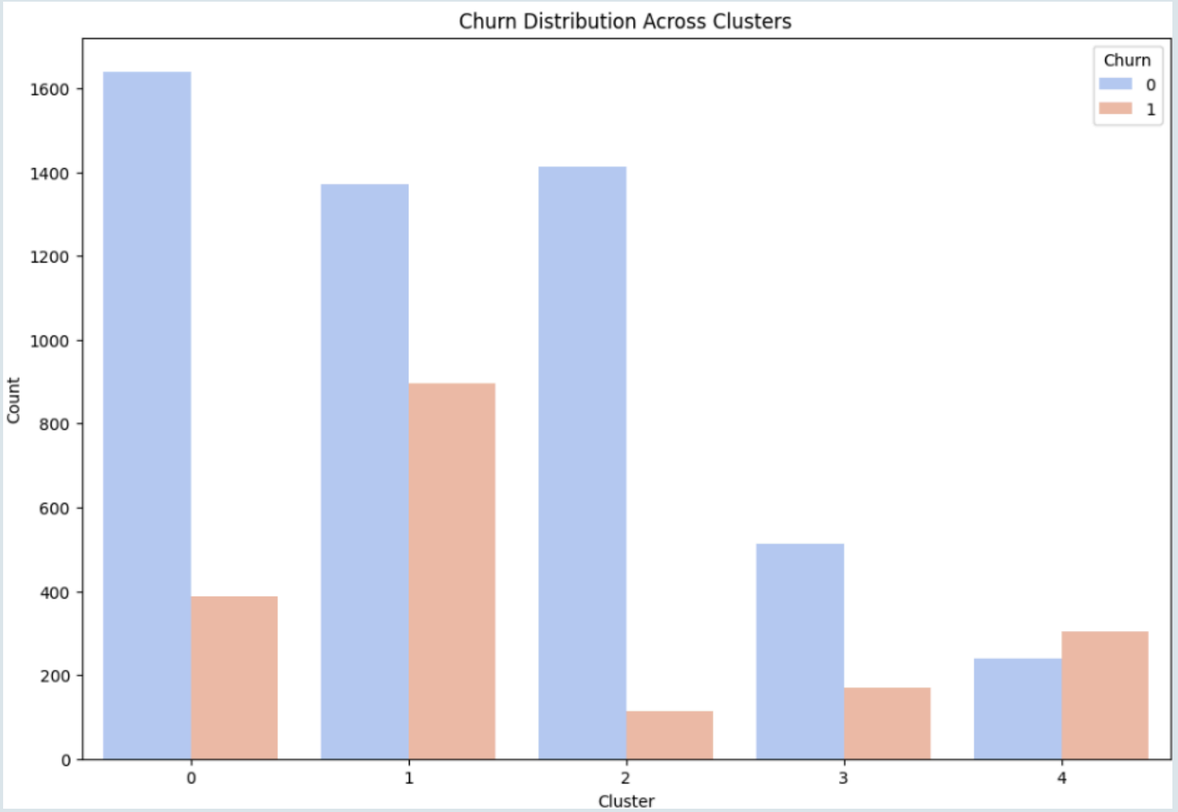
kmeans



Insights:

- **Cluster Insights:** By analyzing the distribution, we can identify which clusters have a higher proportion of customers who churn.

Hierarchical



Conclusion

This project successfully analyzed customer behavior in the telecom industry, predicting churn and segmenting customers into meaningful groups. By implementing supervised learning models, we accurately identified customers likely to churn, allowing businesses to target retention efforts effectively. The unsupervised clustering revealed distinct customer segments, helping companies tailor their services and offerings

Thank you