Telecom Customer Churn Prediction

Analytical and Predictive Modeling

Sudeep Fullel

Student Id: 23140750

Bachelor of Science with Honours Computer and Data Science

(Faculty of Computing, Engineering, and the Built Environment)

Birmingham City University (BCU)

Sunway college Kathmandu, Nepal

Sudeep.Fullel@mail.bcu.ac.uk

Abstract

Customer churn prediction is crucial for telecommunications companies. This project tackled this challenge by developing a machine learning model to predict customer churn. Beginning with thorough dataset exploration and exploratory data analysis (EDA), we gained crucial insights into customer behavior and identified potential churn indicators. This informed subsequent feature engineering efforts, where we created new variables to enhance model performance. Various classification algorithms were evaluated, with the Random Forest model ultimately selected for its superior predictive accuracy. SHAP (SHapley Additive exPlanations) analysis provided valuable interpretability, revealing key factors driving churn predictions. Finally, the model was deployed as an interactive web application using Streamlit, facilitating real-time churn risk assessment. This work provides a practical and interpretable churn prediction solution for the telecommunications industry. Future research will focus on incorporating richer features, exploring advanced modeling techniques, and integrating real-time data for enhanced prediction and personalized interventions.

Contents

Introduction

In the highly competitive telecommunications industry, customer retention is a critical factor in sustaining profitability and growth. Customer churn, defined as the rate at which customers discontinue their subscription to a service, directly impacts revenue and market share. For telecom companies, retaining customers is often more cost-effective than acquiring new ones, making churn prediction a strategic focus area. By identifying customers at risk of churning, telecom providers can proactively implement retention strategies, offer personalized services, and improve overall customer satisfaction.

The primary goal of this project is to develop a predictive model capable of identifying customers at risk of churning based on their historical usage data, demographics, and account information. Using machine learning techniques, we aim to discover patterns in customer behavior and identify key features that signal a customer's likelihood to leave. These insights are intended to help telecom companies implement targeted retention strategies and allocate resources more efficiently.

Through this project, we aim to create a reliable, interpretable model that offers telecom companies actionable insights to improve customer retention strategies.

Problem Definition and Dataset Analysis

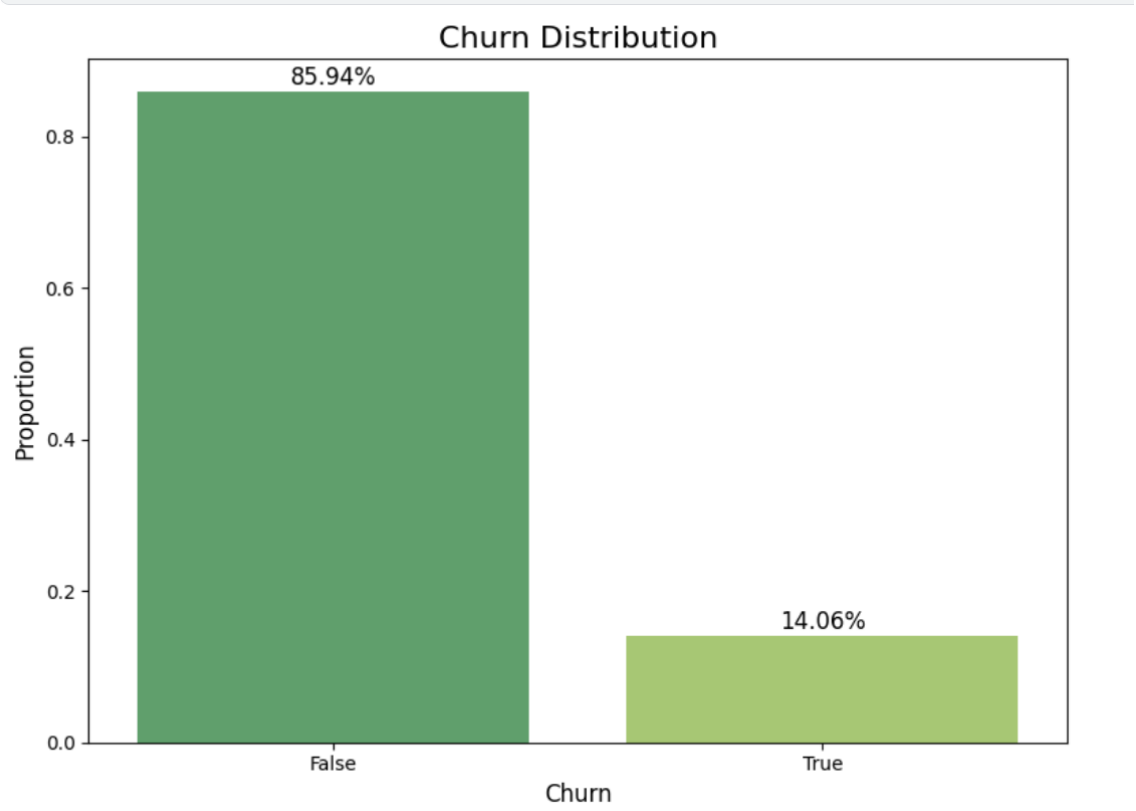
2.1 Problem Definition

The problem involves predicting whether a telecom customer will churn based on their account activities and usage patterns. A binary classification model is built where the target variable, churn, indicates whether a customer is likely to leave.

2.2 Dataset Overview

The Churn in Telecoms dataset, sourced from Kaggle, provides historical data on customer accounts in a telecommunications company, including demographics, service usage, and whether a customer ultimately left the service (churned). The dataset includes multiple numeric, categorical, and binary features that together offer a comprehensive profile of each customer, enabling us to predict churn patterns based on historical trends.

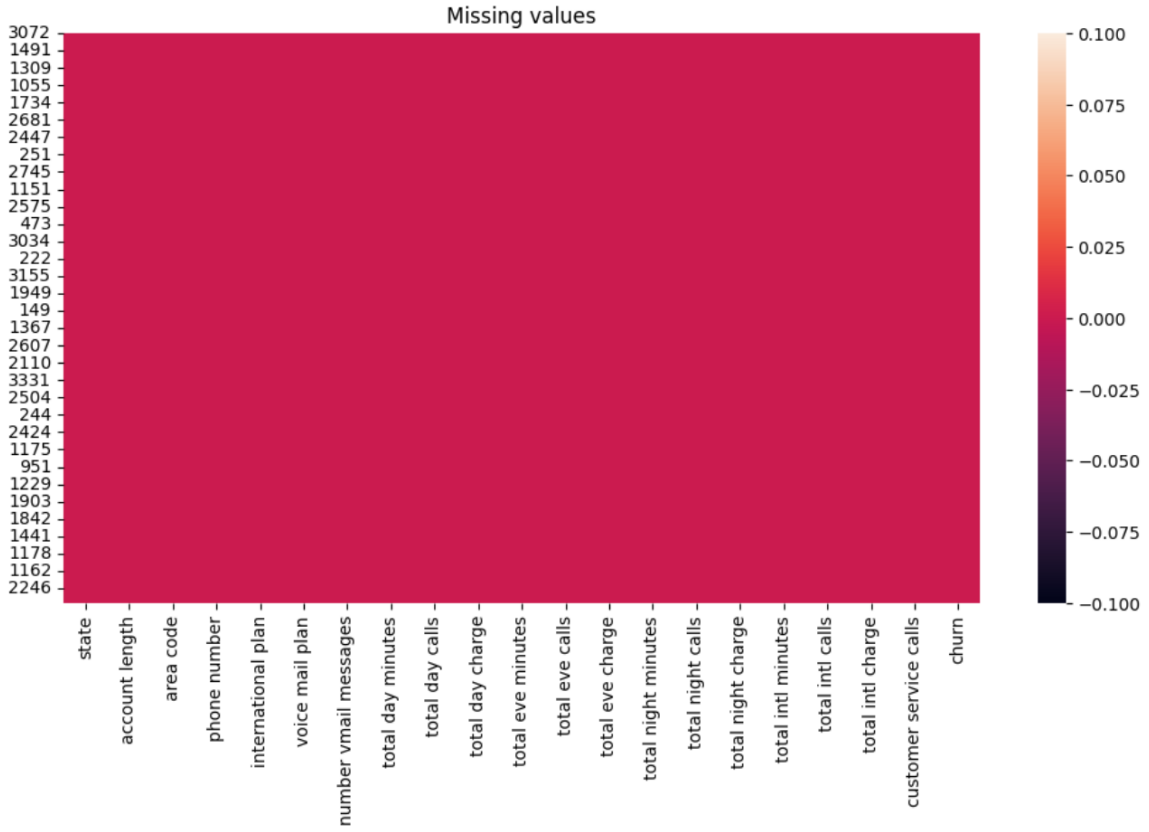
The target variable (churn) has a high imbalance, with approximately 15% of customers classified as churned.



2.3 Exploratory Data Analysis

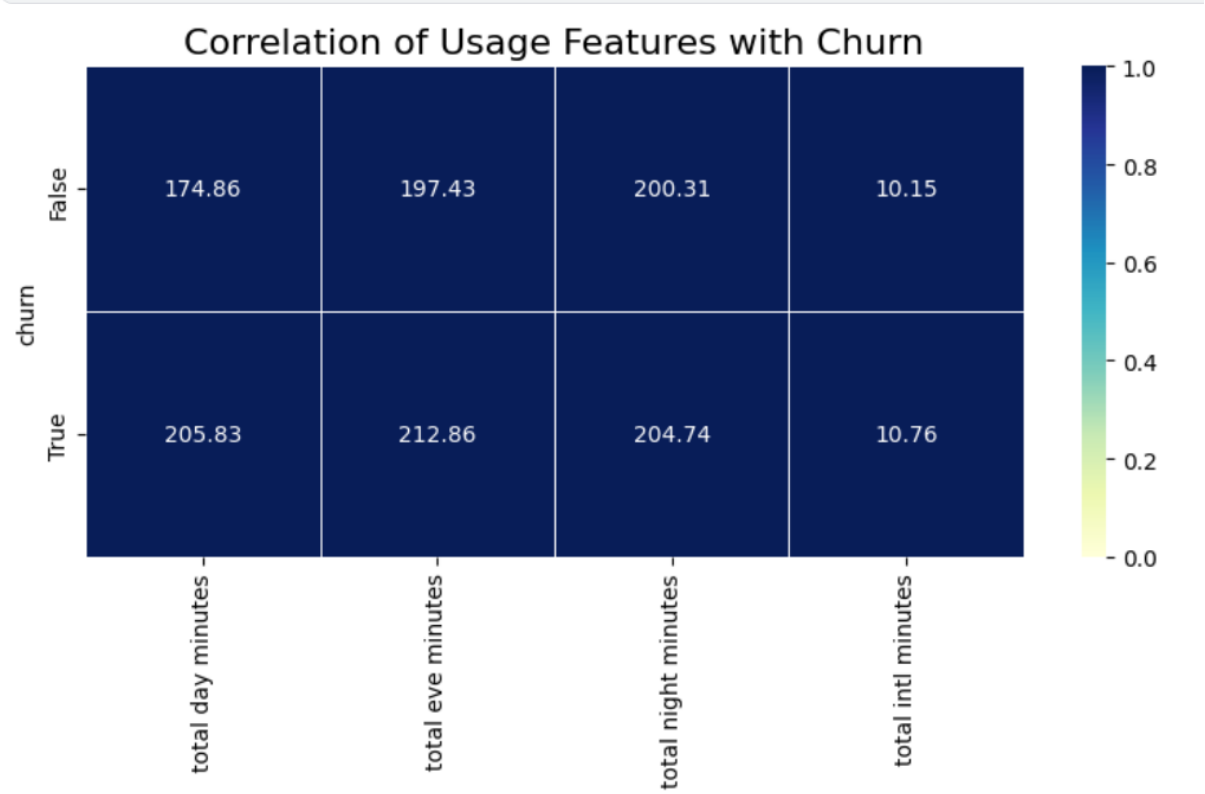
In this section, we examine key variables in the dataset and explore relationships that might affect customer churn. We address several exploratory questions using visual analysis to understand the influence of various factors on churn behavior. This analysis includes examining customer usage patterns, customer service interactions, voice mail and international plans, total charges, and account tenure.

An examination of missing values reveals that the dataset is complete, with no missing entries across any features:



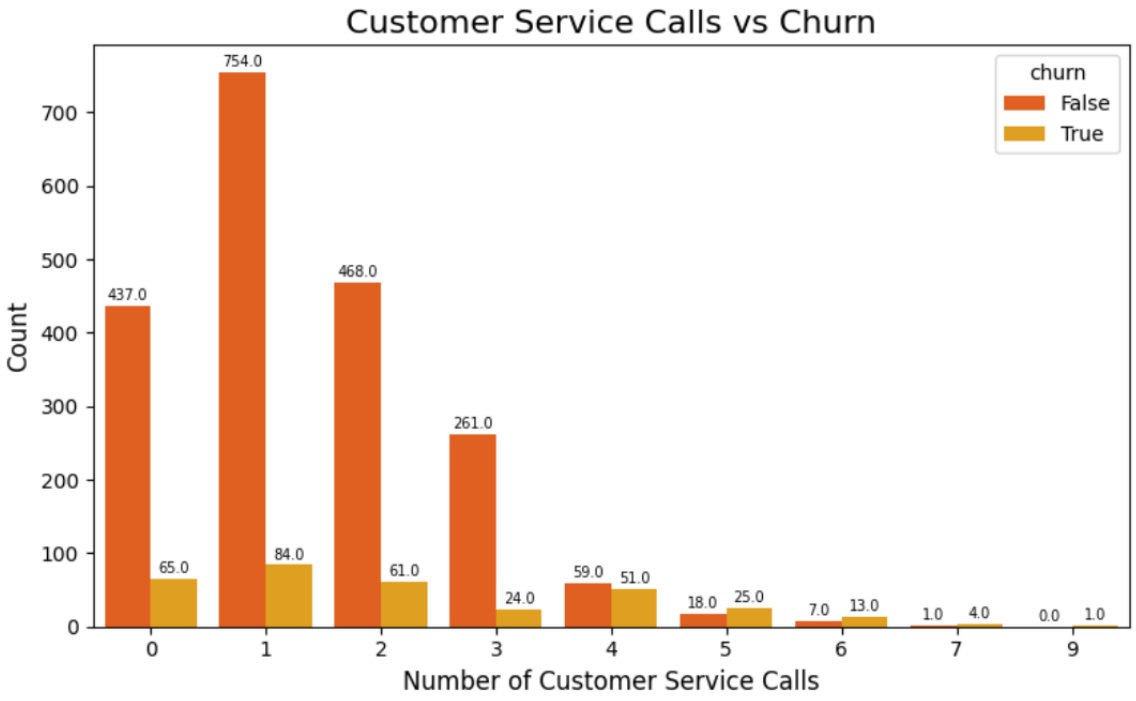
a) Customer Usage and Churn

The correlation heatmap shows the average usage statistics for churned and non-churned customers. Churned customers generally show higher usage in terms of total day, evening, and night minutes compared to non-churned customers. This trend suggests that customers with higher usage might have a higher propensity to churn, possibly due to costs associated with high usage or dissatisfaction with service relative to their needs.



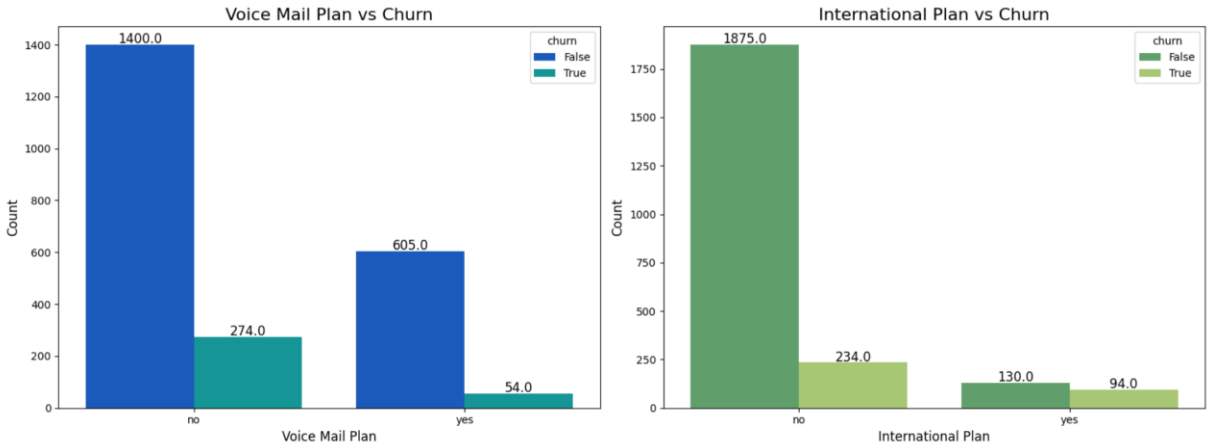
b) Customer Service Calls and Churn

The bar chart illustrates the relationship between the number of customer service calls and churn. A significant pattern is observed where customers who make more frequent calls to customer service are more likely to churn. Notably, customers with zero, one or two calls to customer service have lower churn rates, while those with three or more calls show increased churn rates.



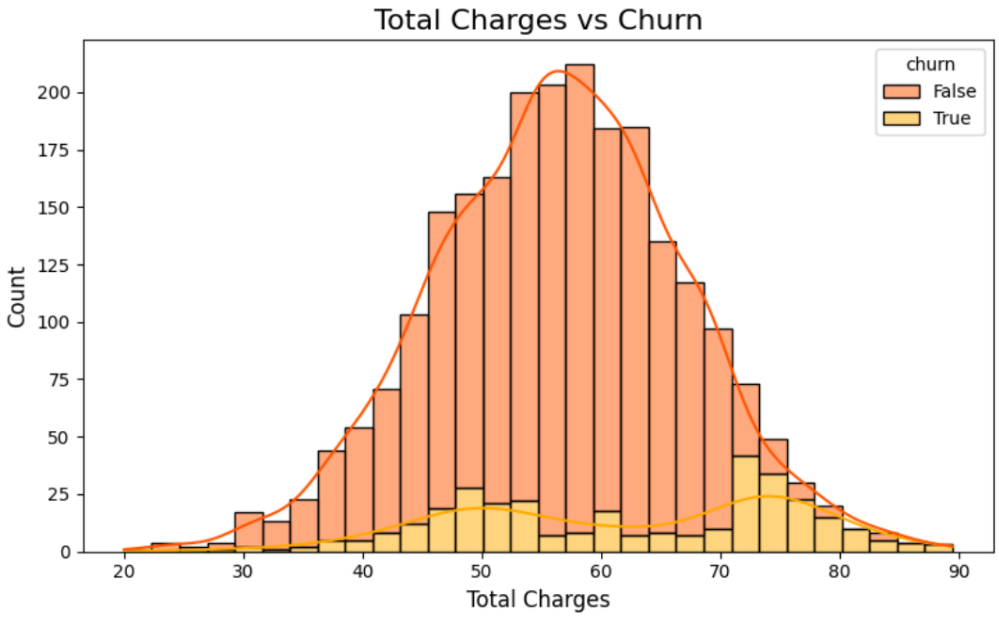
c) Voice Mail and International Plans' Impact on Churn

The side-by-side bar charts depict the relationship between churn and subscription to voice mail and international plans. For the voice mail plan, a substantial majority of customers who do not churn do not have a voice mail plan, while a smaller proportion of churned customers have subscribed to this plan. Conversely, customers with an international plan exhibit slightly higher churn rates, suggesting that international plan subscribers might have specific needs or expectations that, if unmet, could lead to churn.



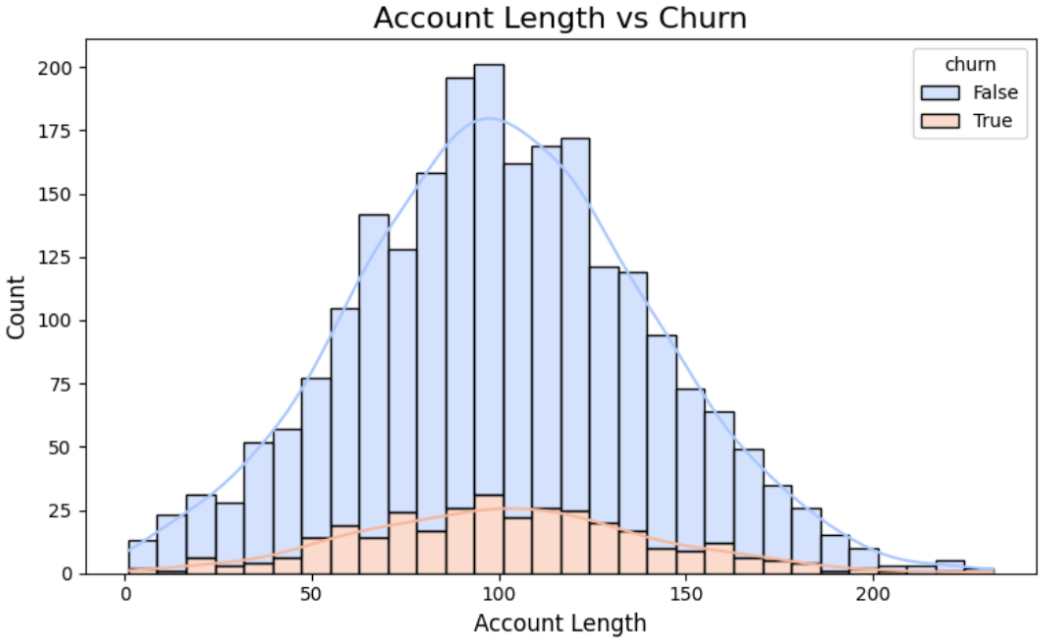
d) Total Charges and Churn:

The "Total Charges vs Churn" histogram seems to indicate a correlation between higher total charges and a lower churn rate. Customers with higher total charges appear less likely to churn. This could be because they are more invested in the service or are more satisfied overall. There appears to be some churn happening even at lower total charge levels though



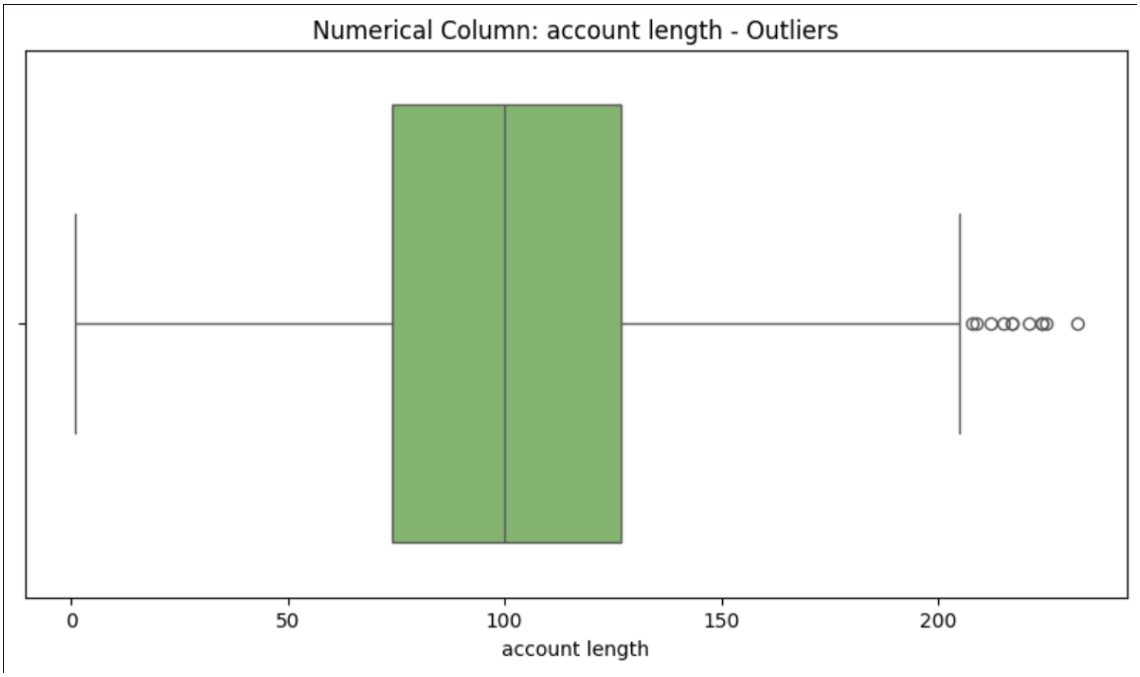
e) Account Length and Churn:

The "Account Length vs Churn" histogram suggests there isn't a strong, direct relationship between account length and churn. The distribution of account lengths for churned customers appears quite like the overall distribution. This might imply that customers churn relatively consistently across different account lengths.



2.4 Outlier Detection

Outliers were detected using visual and statistical analysis. Thresholds were set, for instance, filtering customers with over eight service calls as potential outliers in dissatisfaction.



Feature Engineering and Selection

Feature engineering is an essential step in preparing data for machine learning models. This process involves creating new features that can improve the predictive power of the model and encoding categorical variables. In this project, several features were engineered, and redundant columns were removed to enhance the model's performance.

3.1 Creating Aggregated Features

To capture the overall usage, charges, and call frequency of each customer, three new features were created:

a) Total Charge: Sum of the charges from day, evening, night, and international calls. This feature represents the overall charges incurred by the customer.
b) Total Usage: Sum of the usage minutes across day, evening, night, and international calls. This feature gives a measure of the total time customers spend on calls.
c) Total Calls: Sum of the call counts across day, evening, night, and international periods. This feature captures the total number of calls a customer made.

3.2 Dropping Redundant Columns

Since the individual charge, minutes, and call columns have been aggregated into the new features, the original columns were dropped to reduce dimensionality and potential multicollinearity.

```
# Create the new columns for total charge, total usage, and total calls
train_df['total charge'] = train_df['total day charge'] + train_df['total eve charge'] + train_df['total night charge'] + train_df['total intl charge']
train_df['total_usage'] = train_df['total day minutes'] + train_df['total eve minutes'] + train_df['total night minutes'] + train_df['total intl minutes']
train_df['total_calls'] = train_df['total day calls'] + train_df['total eve calls'] + train_df['total night calls'] + train_df['total intl calls']
```

Additionally, the phone number and state columns were removed, as these features are not expected to be predictive of churn and could introduce unnecessary noise.

```
# Drop the 'state' and 'phone number' column
train_df = train_df.drop(["phone number","state"], axis=1)
```

The dataset was then prepared for modeling by separating the target variable (churn) and encoding it using label encoding to convert it into a numerical format. To handle categorical features within the dataset, a custom function was applied to label-encode these columns consistently across the training and test sets. These features of engineering steps streamline the data and make it ready for model training, improving the likelihood of uncovering meaningful patterns related to customer churn.

Model Development and Evaluation

4.1 Model Selection

This project explored several classification models to predict customer churn. The chosen algorithms include Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, Decision Tree, XGBoost, LightGBM, Extra Trees, and CatBoost. These models were selected for their proven effectiveness in classification tasks and their ability to handle a mix of numerical and categorical features.

4.2 Model Training and Hyperparameter Tuning

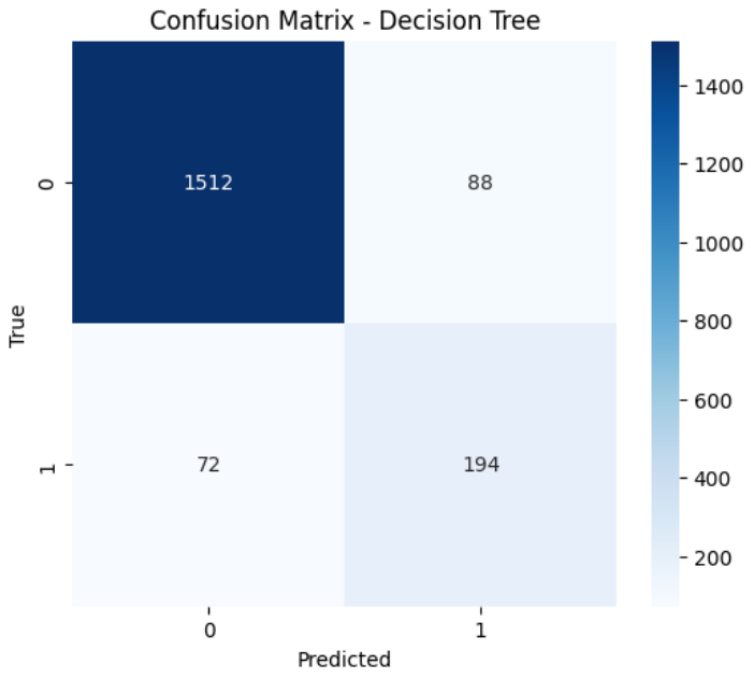
Before training, the dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain the class distribution across both sets. This ensures a reliable evaluation of the models' ability to generalize to unseen data.

Hyperparameter tuning was crucial for optimizing each model's performance. Optimal hyperparameters for each algorithm were determined through Optuna. The best-performing hyperparameter configurations were subsequently used for model training and evaluation.

4.3 Evaluation Metrics

A robust model evaluation framework was implemented, employing 5-fold stratified cross-validation on the training set. This approach provides a more stable estimate of model performance compared to a single train/test split. The following metrics were used to assess each model:

- Accuracy: The overall proportion of correctly classified instances.
- Precision: Out of all instances predicted as churn, the proportion that churned.
- Recall (Sensitivity): Out of all instances that churned, the proportion correctly identified by the model.
- F1-Score: The harmonic means of precision and recall, balancing both metrics.
- ROC-AUC: Area Under the Receiver Operating Characteristic Curve, measuring the model's ability to distinguish between churning and non-churning customers.




```
Decision Tree - Accuracy: 0.9185, Precision: 0.9165, Recall: 0.9143, F1-Score: 0.9153, ROC-AUC: 0.8347
Evaluating model: XGBoost
```

In addition to cross-validation metrics, confusion matrices were generated for each model to visualize the distribution of true positives, true negatives, false positives, and false negatives. This allows for a more granular understanding of model performance and potential misclassifications.

The results of the model evaluation are summarized in Table 1 (training set performance) and Table 2 (test set performance). Based on the test set results, the top-performing models were Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost, all exhibiting high accuracy, precision, recall, F1-score, and ROC-AUC values. While other models showed promising performance on the training set, some, such as the Decision Tree, indicated potential overfitting, highlighting the importance of evaluating performance on unseen data.

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
0	Logistic Regression	0.863340	0.834049	0.863344	0.834961	0.803369	[[1557, 43], [212, 54]]
1	Random Forest	0.949618	0.952192	0.950161	0.945829	0.918340	[[1598, 2], [91, 175]]
2	Gradient Boosting	0.945869	0.943160	0.944802	0.941627	0.926107	[[1580, 20], [83, 183]]
3	AdaBoost	0.902456	0.893895	0.902465	0.891374	0.904992	[[1564, 36], [146, 120]]
4	Decision Tree	0.918532	0.916541	0.914255	0.915291	0.834657	[[1512, 88], [72, 194]]
5	XGBoost	0.939437	0.936917	0.939443	0.936339	0.919309	[[1572, 28], [85, 181]]
6	LightGBM	0.942113	0.940159	0.942122	0.938732	0.919619	[[1578, 22], [86, 180]]
7	Extra Trees	0.929241	0.927079	0.929796	0.923863	0.916527	[[1578, 22], [109, 157]]
8	CatBoost	0.938900	0.936270	0.938907	0.936086	0.914483	[[1569, 31], [83, 183]]

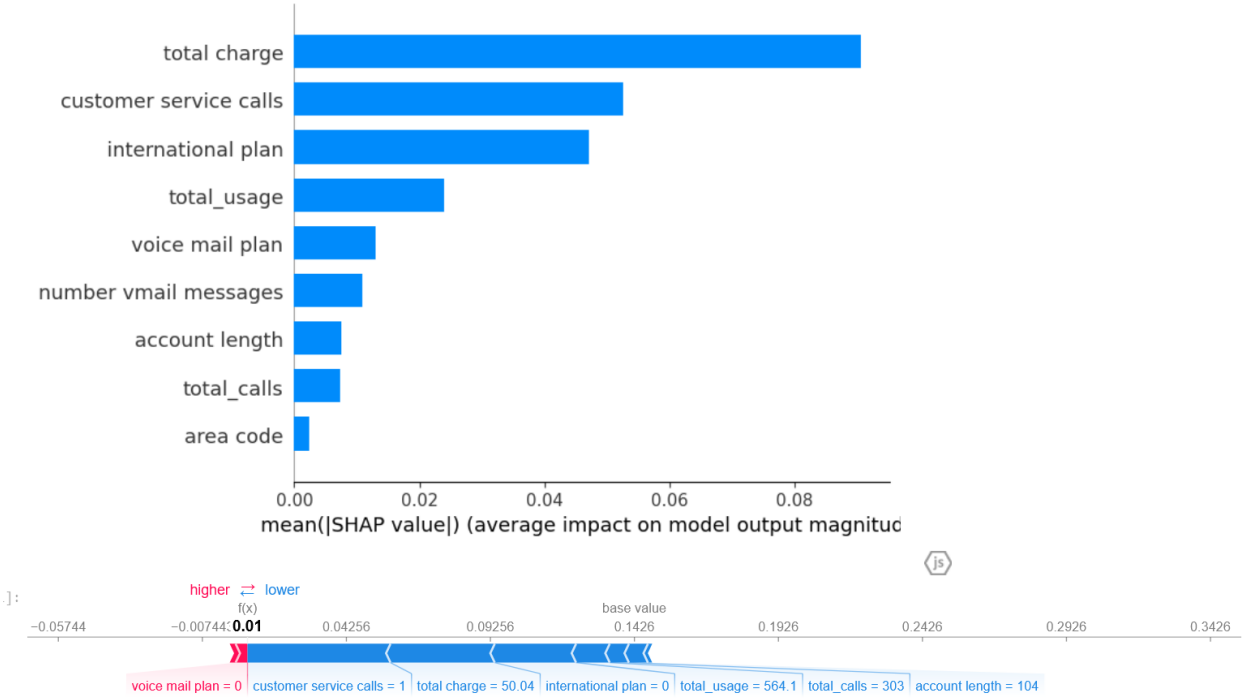
Figure 1: Train data evaluation

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Confusion Matrix
0	Logistic Regression	0.853	0.822826	0.853	0.819653	0.793335	[[824, 21], [126, 29]]
1	Random Forest	0.938	0.941565	0.939	0.933124	0.904829	[[843, 2], [59, 96]]
2	Gradient Boosting	0.926	0.926614	0.929	0.923892	0.897034	[[832, 13], [58, 97]]
3	AdaBoost	0.893	0.884164	0.893	0.880289	0.844890	[[825, 20], [87, 68]]
4	Decision Tree	0.902	0.910053	0.908	0.908938	0.833155	[[795, 50], [42, 113]]
5	XGBoost	0.937	0.935021	0.937	0.933406	0.895839	[[832, 13], [50, 105]]
6	LightGBM	0.937	0.937825	0.937	0.931462	0.829212	[[840, 5], [58, 97]]
7	Extra Trees	0.917	0.912125	0.916	0.908441	0.895553	[[830, 15], [69, 86]]
8	CatBoost	0.933	0.930281	0.933	0.929651	0.886620	[[828, 17], [50, 105]]

Figure 2: Test data evaluation

Model Interpretation and Explainability

This section leverages SHAP (SHapley Additive exPlanations) values to interpret the Random Forest model, which demonstrated strong predictive performance. SHAP values quantify the contribution of each feature to a specific prediction, allowing us to understand feature importance and their directional impact.



5.1 Global Feature Importance (refering to the bar plot image):

The SHAP summary plot provides a global view of feature importance. total_charge emerges as the most influential factor driving churn predictions, followed by customer_service_calls and international_plan. This suggests that higher total charges tend to reduce churn, while more customer service calls and having an international plan are associated with increased churn risk. Other influential factors include total_usage, voice_mail_plan, and the number of voicemail messages. Features like account_length, total_calls, and area_code appear to have less impact.

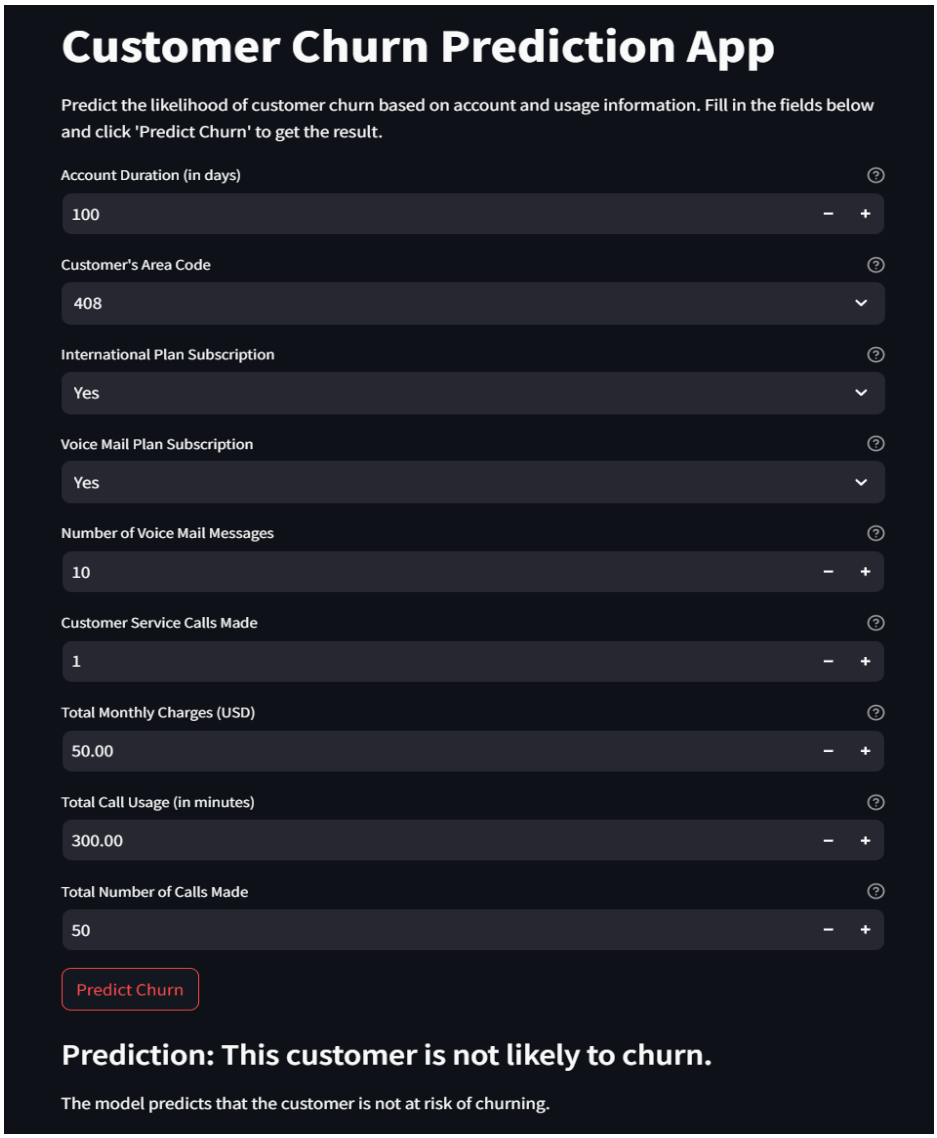
5.2 Local Explainability (refering to the force plot image):

SHAP force plots provide local explanations for individual predictions. Figure Y illustrates the prediction for a single customer. The force plot reveals how each feature value pushes the prediction towards churn (red) or non-churn (blue). For example, this customer's lack of a voice mail plan (voice_mail_plan = 0) significantly reduces the predicted churn probability, while their one customer service call pushes the prediction slightly towards churn. The total_charge of 50.04 contributes negatively to the churn prediction, suggesting that it's lower than average for churning customers.

Deployment Strategy and Considerations

To make the churn prediction model readily accessible and usable, it has been deployed as an interactive web application using Streamlit. This deployment strategy allows stakeholders to easily interact with the model and obtain predictions without requiring technical expertise or direct access to the model's code.

The application presents a user-friendly interface where users can input customer information, including account duration, area code, international and voicemail plan subscriptions, number of voicemail messages, customer service calls made, total charges, total usage, and total calls. Clear descriptions and appropriate input validation are provided for each field to guide users and ensure data quality.



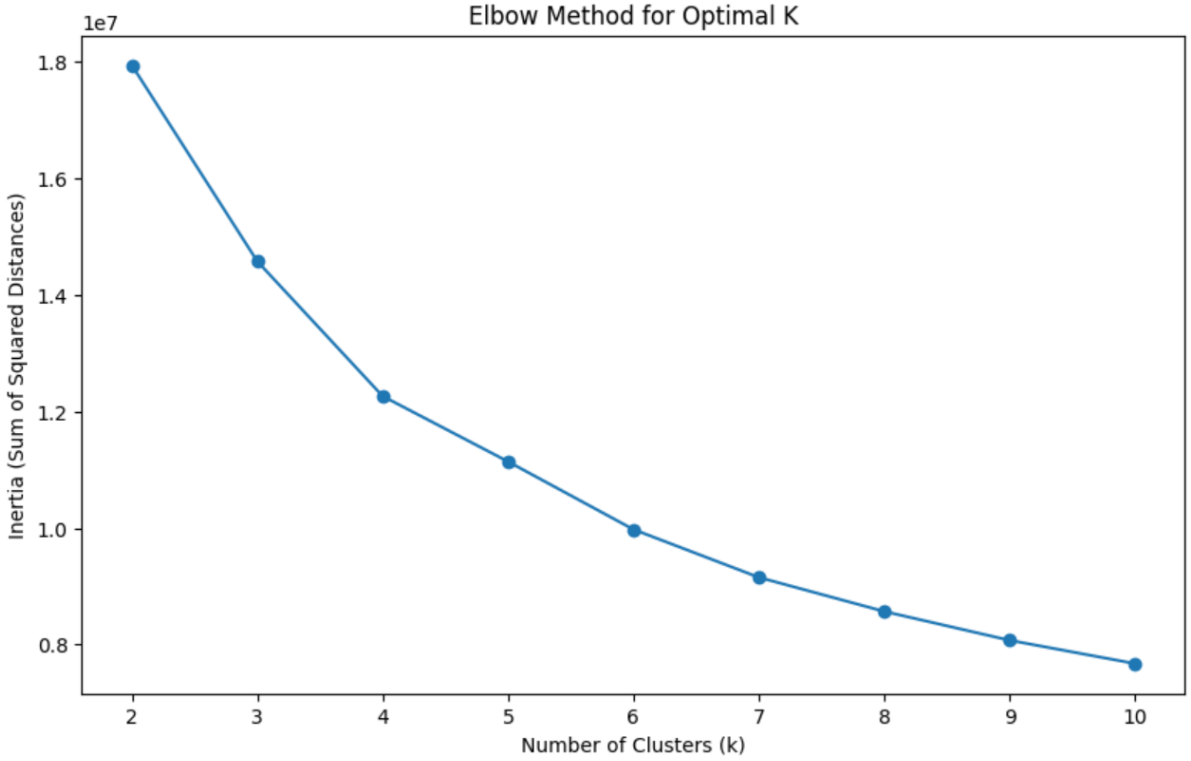
Upon clicking the "Predict Churn" button, the application utilizes the trained Random Forest model to generate a prediction, indicating whether the customer is likely to churn or not. The prediction is presented with a clear and concise explanation, aiding in immediate interpretation.

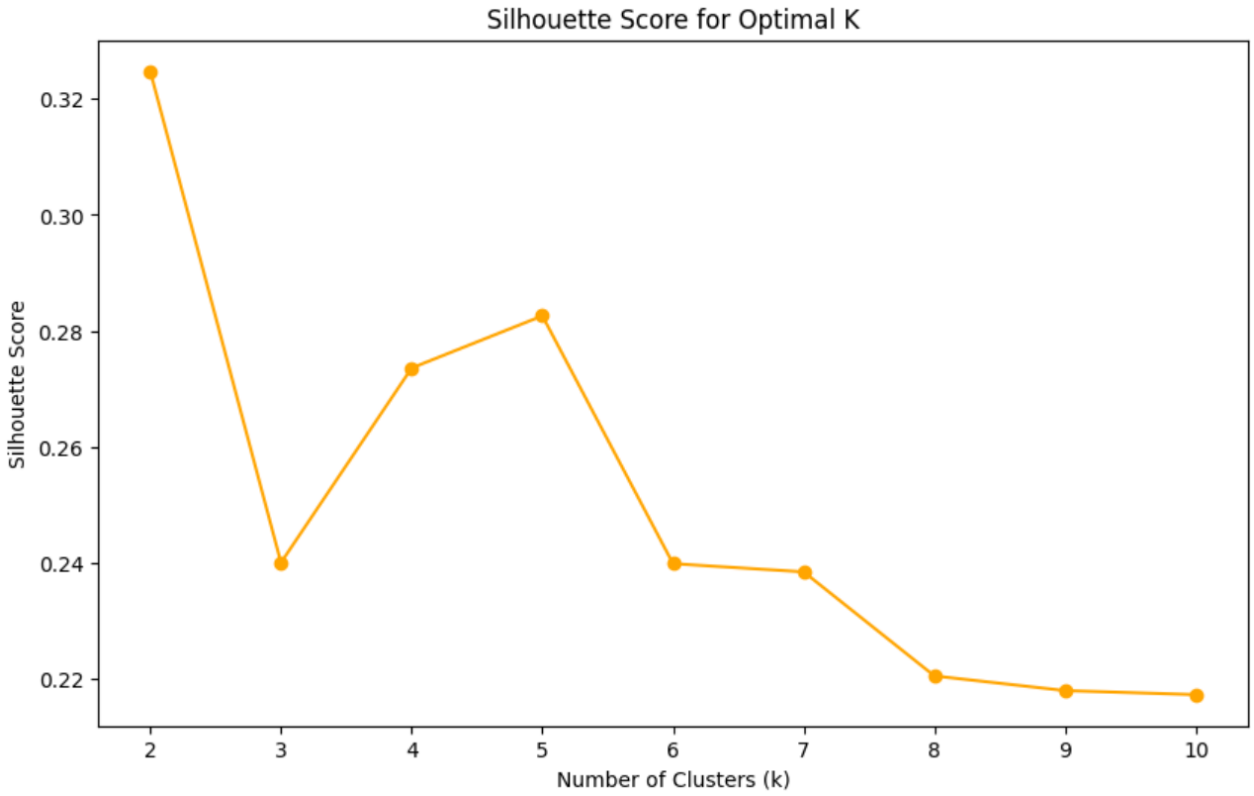
This Streamlit application provides a practical and efficient way to leverage the churn prediction model for real-time decision-making. Its accessibility and ease of use empower stakeholders to proactively identify at-risk customers and implement targeted retention strategies. Future enhancements could include integrating the application with existing CRM systems or developing a batch prediction capability for analyzing larger customer datasets.

Unsupervised Learning

In addition to churn prediction, this project explored unsupervised learning to segment customers based on their usage patterns and account characteristics. K-Means clustering was employed to group similar customers together, potentially revealing distinct customer segments with varying behaviors and needs.

The Elbow method and the Silhouette score were used to determine the optimal number of clusters (k) for K-Means. The Elbow method looks for a "kink" in the plot of inertia (sum of squared distances within clusters) versus k, suggesting a point of diminishing returns in adding more clusters. The Silhouette score measures how similar a data point is to its own cluster compared to other clusters. A higher Silhouette score indicates better-defined clusters.

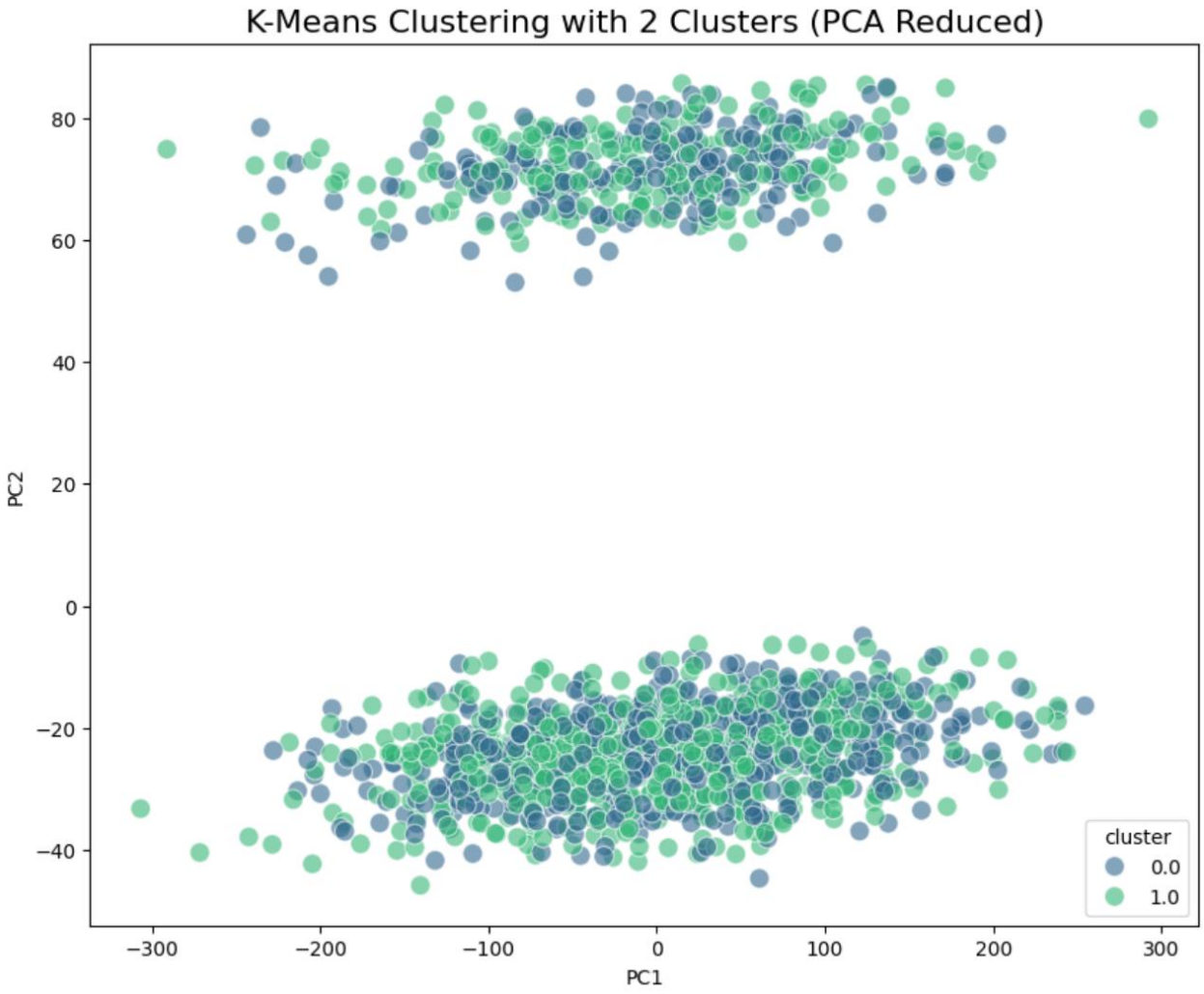




Based on the Silhouette score, which provides a more robust measure of cluster quality, an optimal k value of 2 was selected. The Elbow method plot also supports this choice by indicating that after around 2 clusters there isn't as large of a difference in SSE change.

7.1 K-Means Clustering

K-Means clustering was performed on the training data using the optimal k value. To visualize the clusters, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data to two principal components. The resulting clusters were then plotted using a scatter plot, where each point represents a customer and the color represents the assigned cluster.



Conclusion and Future Work

This project successfully built and deployed a machine learning model to predict customer churn, leveraging a Random Forest algorithm and SHAP analysis for interpretation. The model achieved high predictive accuracy and identified key churn drivers, including total charges, customer service calls, and international plan subscriptions. The Streamlit web application provides an accessible platform for real-time churn prediction.

Future enhancements include exploring richer features, experimenting with advanced algorithms, integrating real-time data, enhancing explainability within the application, conducting A/B testing of interventions, performing cost-benefit analysis, and calibrating model probabilities. These improvements will further refine the model and maximize its impact on customer retention strategies.

Appendix

GitHub Repository: https://github.com/12sudeep/telecom-churn-analysis-and-prediction