



Predicting Customer Churn in Telecom Industry Using Data Science and Machine Learning

Supervised by/ Eng. Sherif Said
Group:ALX_AIS3_M1d

Team Members



Mina Michel



Omar Abdel
elrazek



Muhammad
Ibrahim



Nirvana Nagy




Sarah Ayman





Agenda

1. Story Telling
 2. Exploratory Data Analysis (EDA)
 - Data Cleaning
 - Data visualization and Dashboards
 3. Preprocessing
 4. Model building and prediction
 5. Model evaluation and tuning
 6. Deployment and interpretation
 7. Conclusion
- 

Target Audience

**Telecom
Executives and
Decision
Makers**

**Marketing and
Customer
Retention Team**

**Data Science
and Analytics
Team**

**Customer
Support and
Service Team**

**Investors and
Stakeholders**

Story Telling



Story Telling

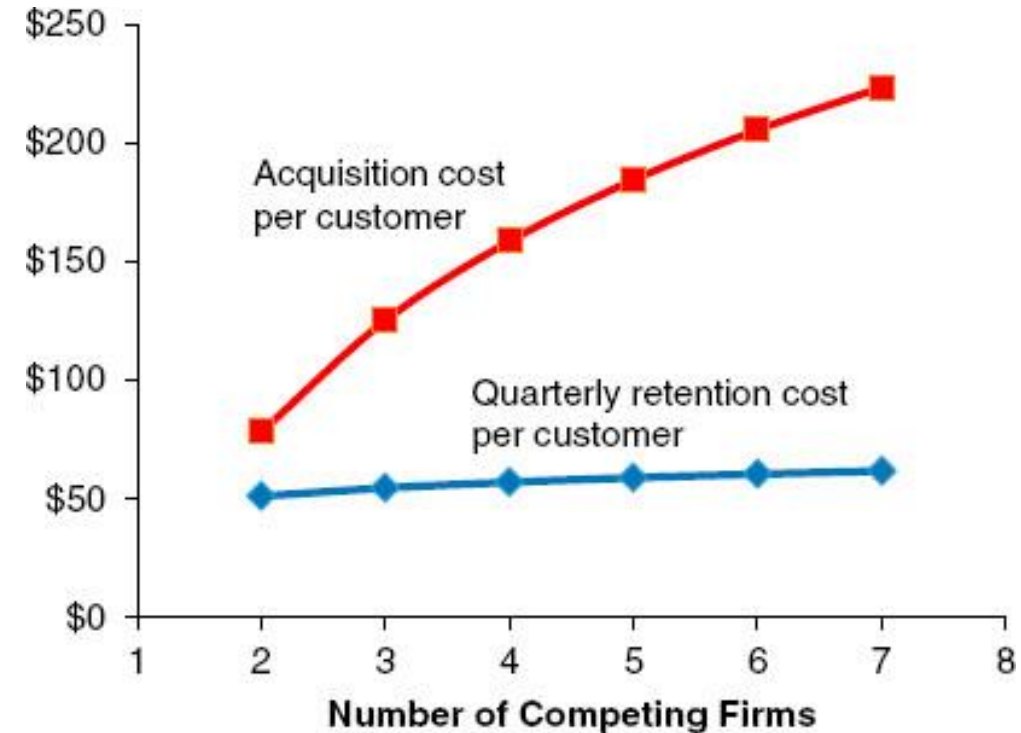


Problems

1.The Cost of Customer Acquisition

In today's fast-paced telecom industry, **customer acquisition is expensive and time-consuming.**

- The cost of attracting new customers (Customer Acquisition Cost, or CAC) is far higher than retaining current ones.
- Your company invests heavily in **marketing, advertising, and sales**, which quickly add up.
- Loyal customers** don't need the same level of attention, and they already trust your brand.
- They also refer others through **word-of-mouth**, reducing marketing costs.



2. Building Relationships with New Customers

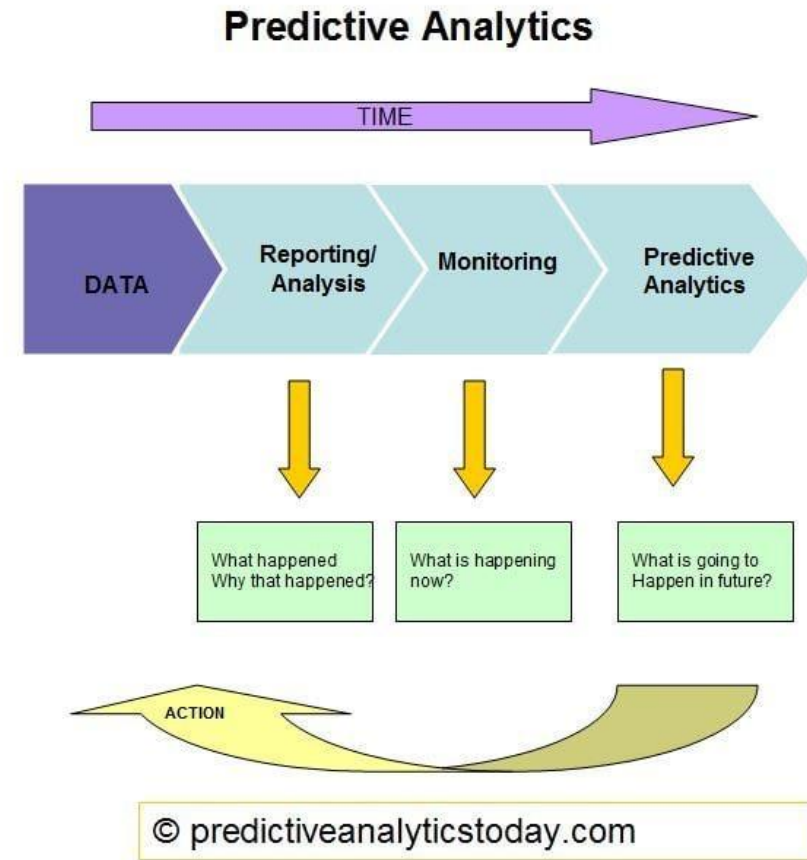
- New customers require **education**, such as free trials and product information, which increases CAC.
- **Trust and credibility** take time to build. New customers need multiple touchpoints (advertisements, testimonials) before they commit.
- The sales cycle for acquiring new customers is much longer compared to keeping existing customers.



Insights

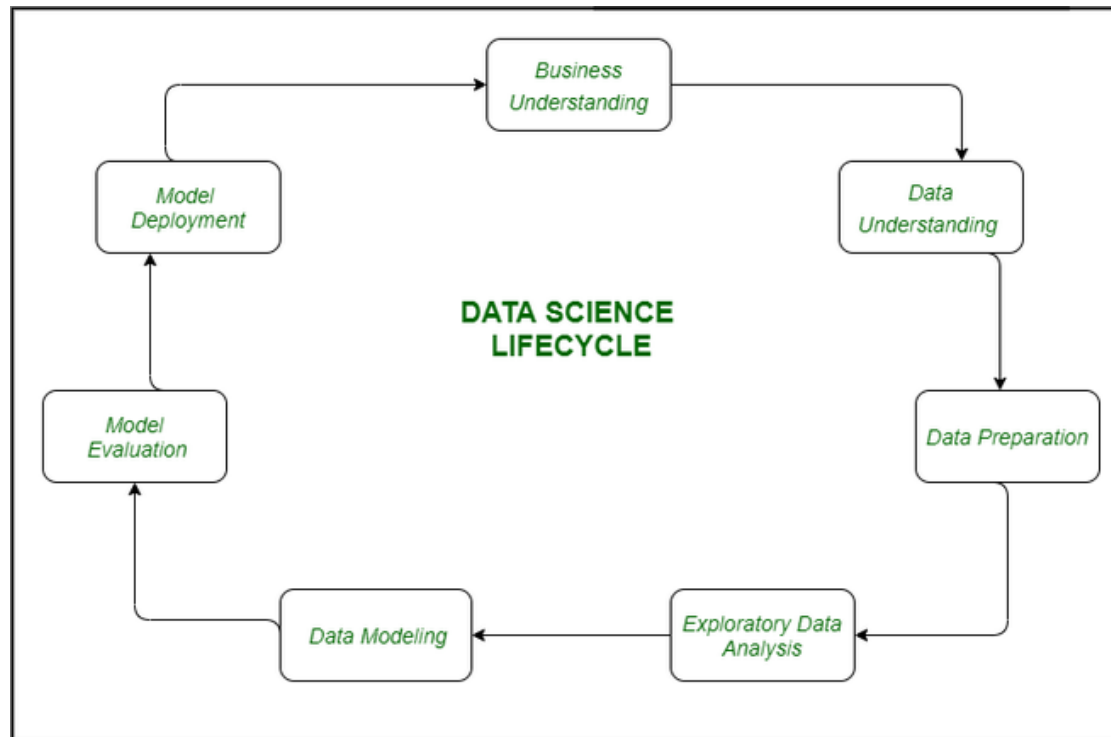
1. The Power of Predicting Churn

- Predictive churn models allow companies to take **proactive steps** to keep customers.
- Given the **high cost of acquiring customers**, knowing which customers are likely to leave is invaluable.
- This is where **machine learning** comes in—enabling data-driven insights into customer behavior.



2. Data-Driven Approach and Methodology

- Data Exploration and Cleaning:** We clean the data to ensure model accuracy.
- Exploratory Data Analysis (EDA):** Uncover the factors contributing to churn, such as usage patterns and complaints.
- Model Building:** We test several machine learning models (Logistic Regression, Decision Trees, Random Forest) to find the best predictor.



3. Real-Time Dashboards

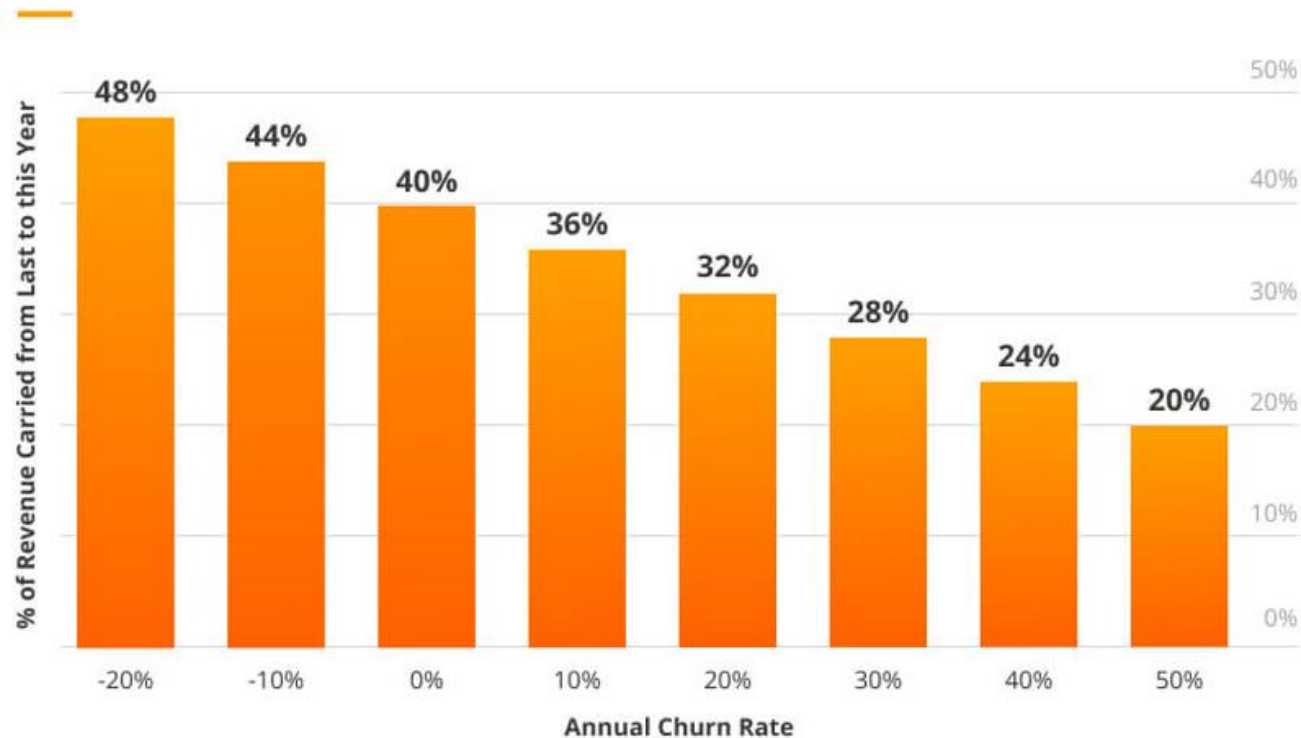
- We create **interactive dashboards** for stakeholders to track churn predictions and monitor retention trends in real-time.
- These dashboards provide **actionable insights** on why customers are likely to leave (service issues, pricing, etc.).



Impacts

1.Reducing Churn Rates

- **Lower churn rates** mean fewer customers leaving, directly impacting revenue.
- A **10% reduction in churn** can have a significant financial impact, especially given the high cost of acquiring new customers.



2. Improving Customer Satisfaction

- Proactively addressing customer concerns leads to **higher satisfaction**, making them more likely to stay loyal and even recommend your brand.
- Satisfied customers become advocates, generating **organic growth** through word-of-mouth.



EDA

- EDA helps uncover **patterns, relationships, and trends** in the data.
- Critical for understanding which **features** influence customer churn.
- Sets the **foundation** for building predictive models.



1. Load and Explore the Dataset

```
# Load the dataset
import pandas as pd
data = pd.read_csv('Customer Churn.csv')
data.head()
```

	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn
0	8	0	38	0	4370	71	5	17	3	1	1	30	197.640	0
1	0	0	39	0	318	5	7	4	2	1	2	25	46.035	0
2	10	0	37	0	2453	60	359	24	3	1	1	30	1536.520	0
3	10	0	38	0	4198	66	1	35	1	1	1	15	240.020	0
4	3	0	38	0	2393	58	2	33	1	1	1	15	145.805	0

1. Load and Explore the Dataset

- The dataset contains **3150** rows and **14** columns.
- No missing values.
- Most columns are **integers** except for Customer Value, is a float.
- The **Churn** column is the **target** variable for **prediction**.

```
data.info() # Check for data types and missing values
data.describe() # Summary statistics
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3150 entries, 0 to 3149
Data columns (total 14 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Call Failure                          3150 non-null   int64
1   Complains                             3150 non-null   int64
2   Subscription Length                   3150 non-null   int64
3   Charge Amount                         3150 non-null   int64
4   Seconds of Use                        3150 non-null   int64
5   Frequency of use                      3150 non-null   int64
6   Frequency of SMS                      3150 non-null   int64
7   Distinct Called Numbers               3150 non-null   int64
8   Age Group                            3150 non-null   int64
9   Tariff Plan                           3150 non-null   int64
10  Status                               3150 non-null   int64
11  Age                                   3150 non-null   int64
12  Customer Value                        3150 non-null   float64
13  Churn                                3150 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 344.7 KB
```


1. Load and Explore the Dataset

	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group
count	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	7.627937	0.076508	32.541905	0.942857	4472.459683	69.460635	73.174921	23.509841	2.826032
std	7.263886	0.265851	8.573482	1.521072	4197.908687	57.413308	112.237560	17.217337	0.892555
min	0.000000	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	1.000000	0.000000	30.000000	0.000000	1391.250000	27.000000	6.000000	10.000000	2.000000
50%	6.000000	0.000000	35.000000	0.000000	2990.000000	54.000000	21.000000	21.000000	3.000000
75%	12.000000	0.000000	38.000000	1.000000	6478.250000	95.000000	87.000000	34.000000	3.000000
max	36.000000	1.000000	47.000000	10.000000	17090.000000	255.000000	522.000000	97.000000	5.000000

2. Data Cleaning and Preparation

```
# Check for missing values
print("Missing values in each column:\n", data.isnull().sum())

# Check for duplicate rows
duplicate_rows = data.duplicated().sum()
print(f"\nNumber of duplicate rows: {duplicate_rows}")
```

Missing values in each column:

Call Failure	0
Complains	0
Subscription Length	0
Charge Amount	0
Seconds of Use	0
Frequency of use	0
Frequency of SMS	0
Distinct Called Numbers	0
Age Group	0
Tariff Plan	0
Status	0
Age	0
Customer Value	0
Churn	0
dtype: int64	

Number of duplicate rows: 300

➤ The data has **no missing values** but has **300 duplicate rows**.

2. Data Cleaning and Preparation

```
# Remove duplicate rows
data_cleaned = data.drop_duplicates()

# Verify that duplicates are removed
print(f'Number of rows after removing duplicates: {data_cleaned.shape[0]}')
```

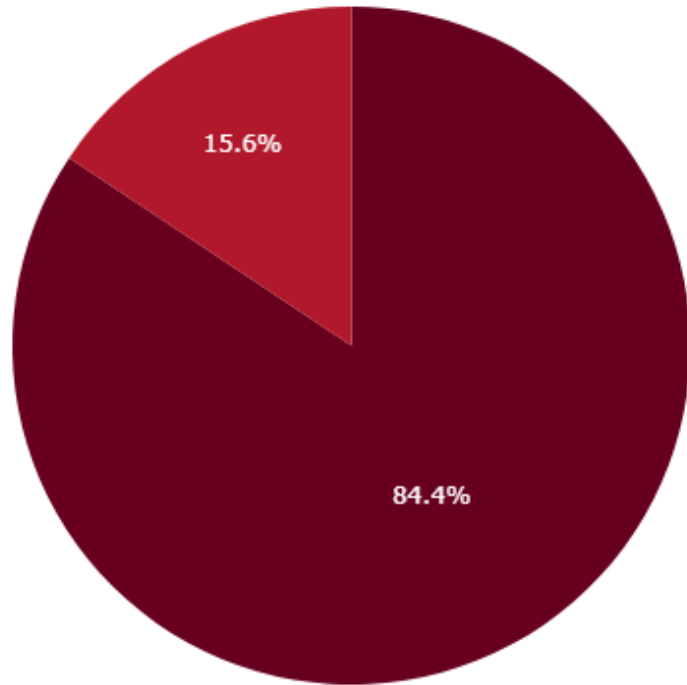
Number of rows after removing duplicates: 2850

```
# Clean column names (remove extra spaces and strip them)
data_cleaned.columns = data_cleaned.columns.str.replace(' ', ' ').str.strip()

# Confirm cleaned column names
print("Cleaned column names:", data_cleaned.columns)
```

Cleaned column names: Index(['Call Failure', 'Complains', 'Subscription Length', 'Charge Amount', 'Seconds of Use', 'Frequency of use', 'Frequency of SMS', 'Distinct Called Numbers', 'Age Group', 'Tariff Plan', 'Status', 'Age', 'Customer Value', 'Churn'], dtype='object')

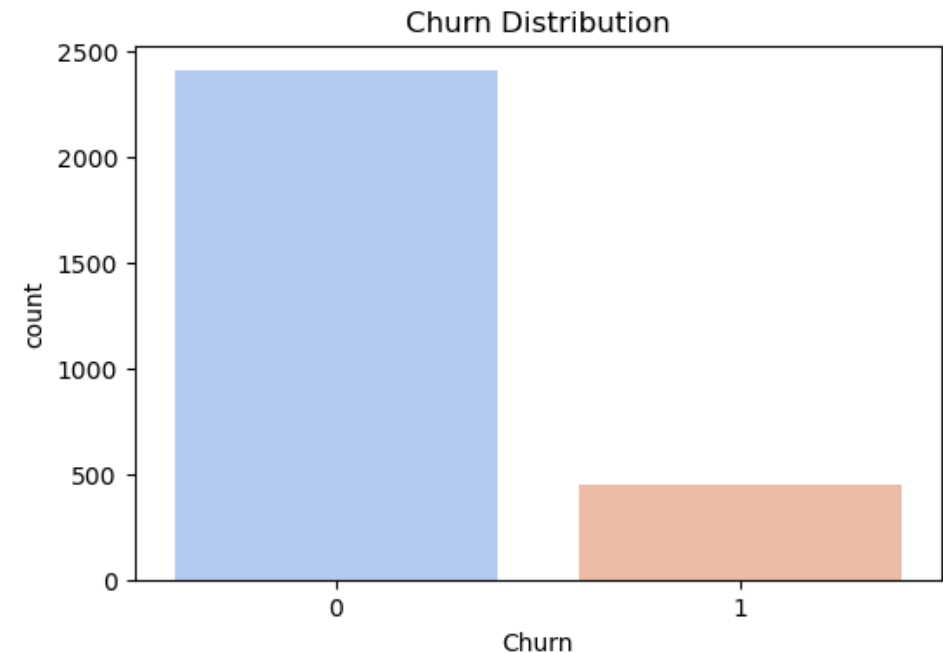
Churn Rate Distribution:



- Class **imbalance** observed, which could affect model performance.

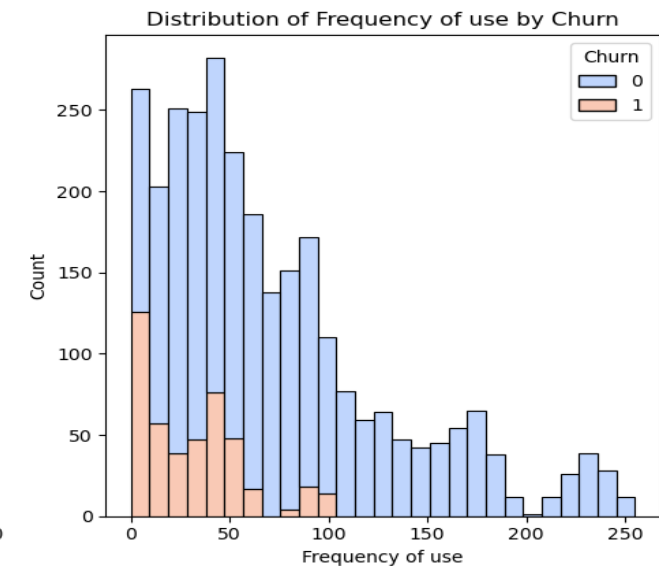
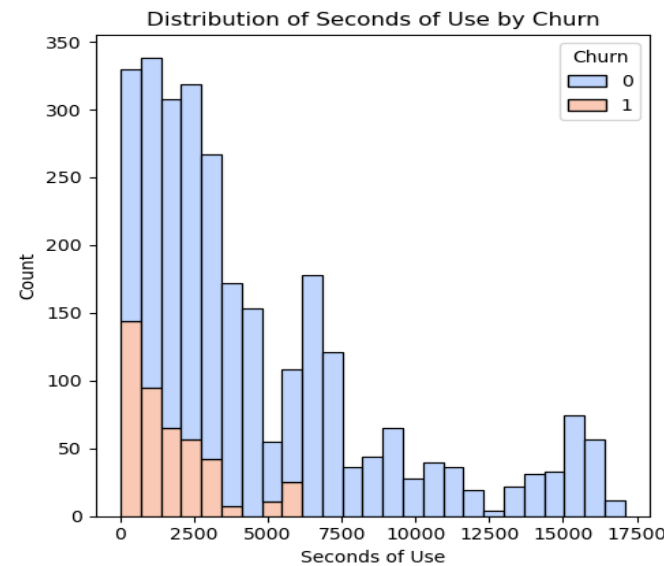
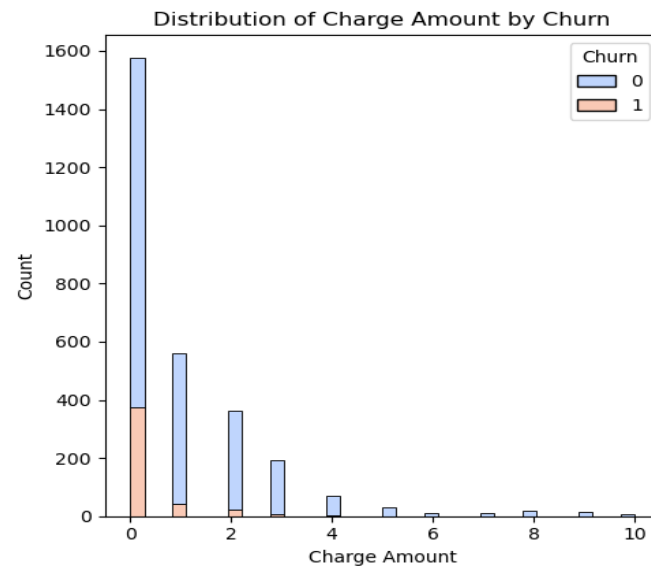
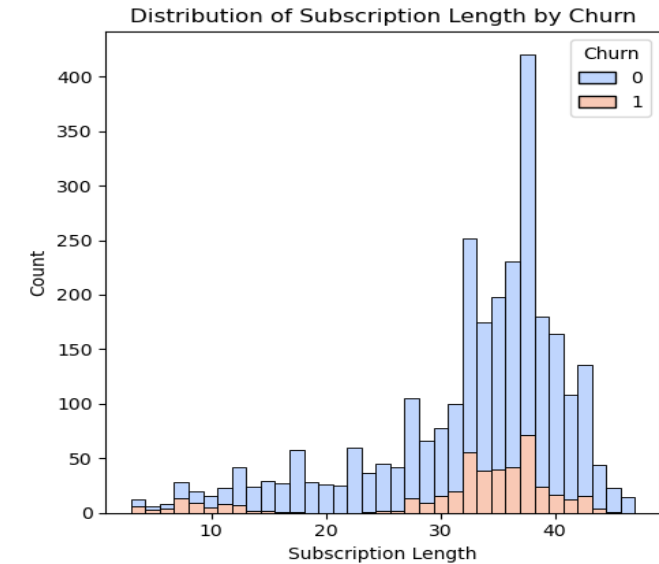
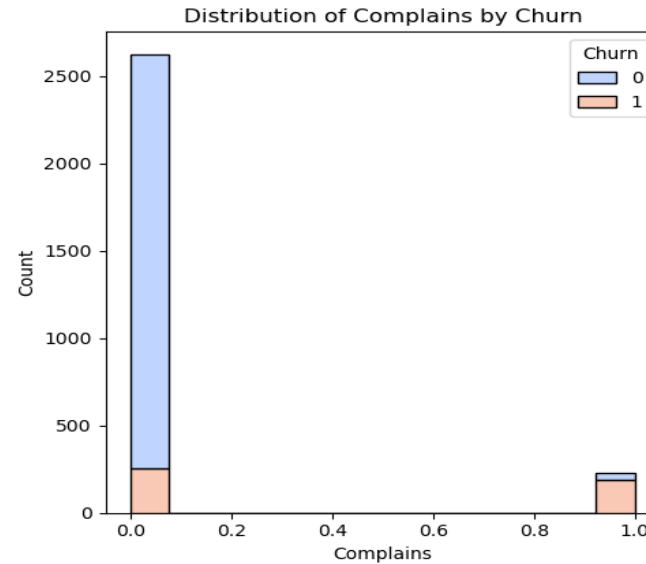
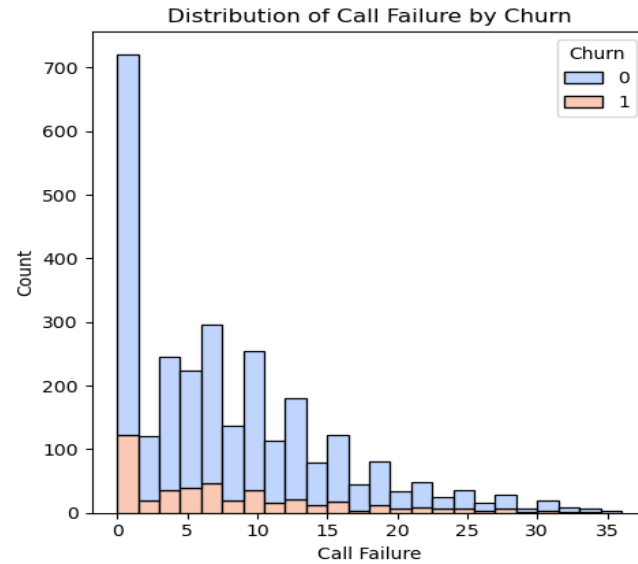
■ No Churn
■ Churn

- **84.4%** of customers did not churn, while **15.6%** did churn.

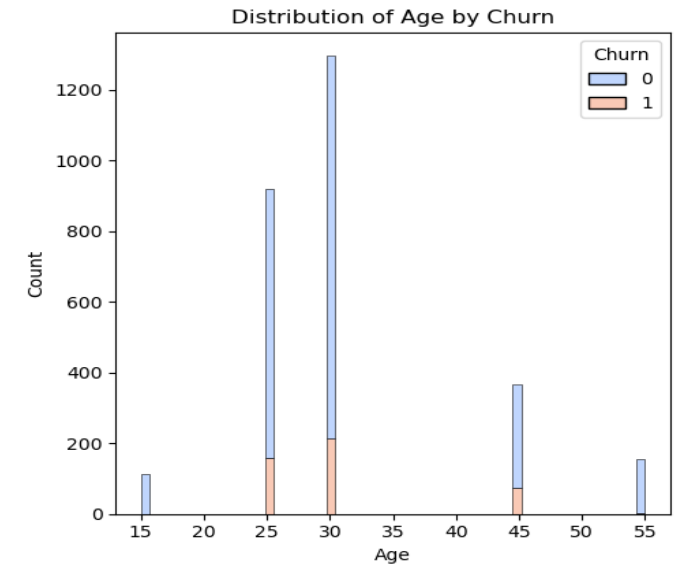
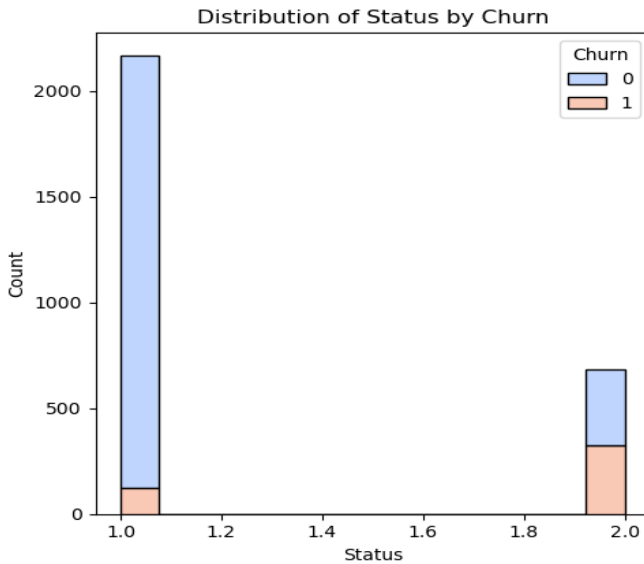
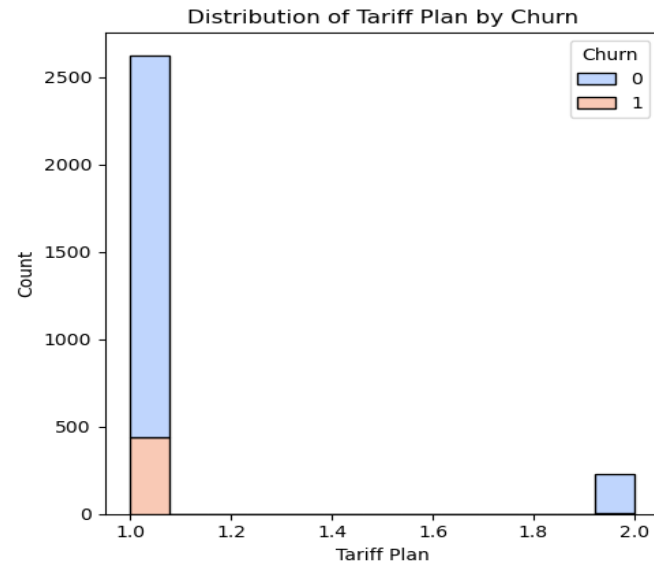
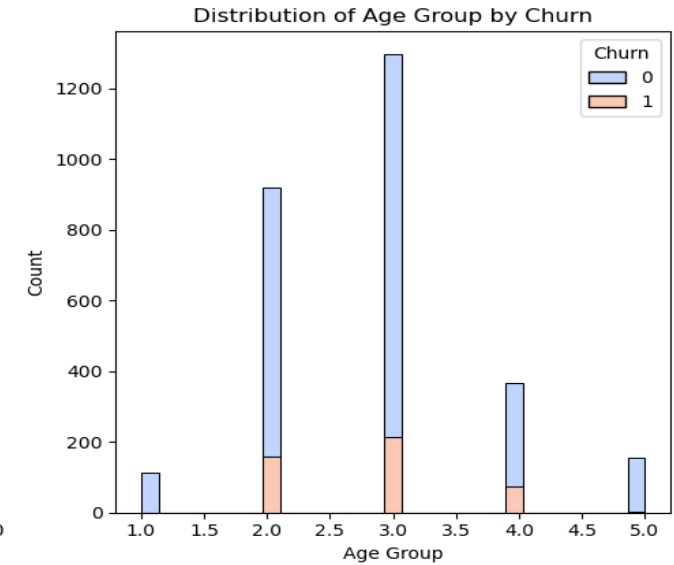
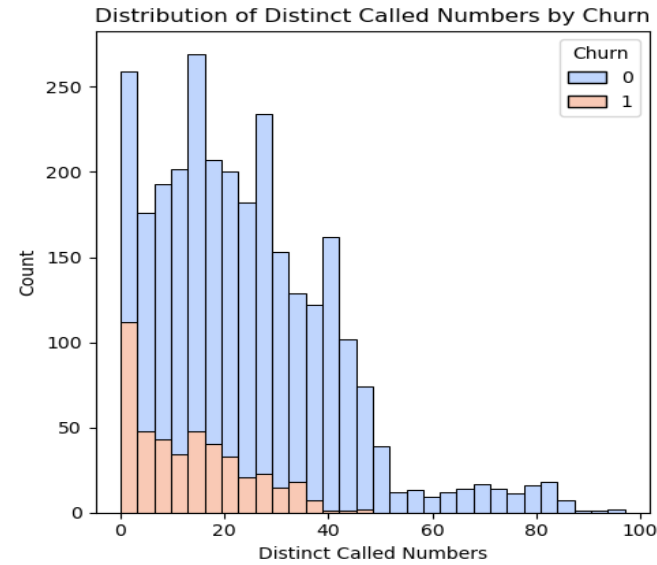
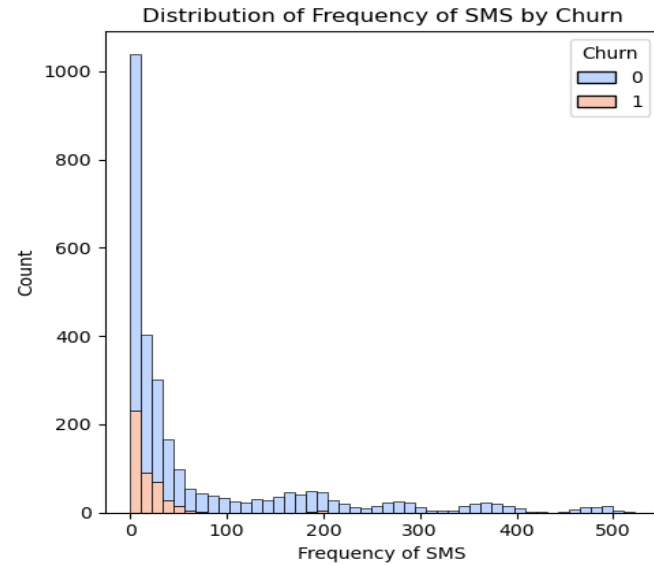


Number of Non-Churned Customers (0): **2404**
Number of Churned Customers (1): **446**

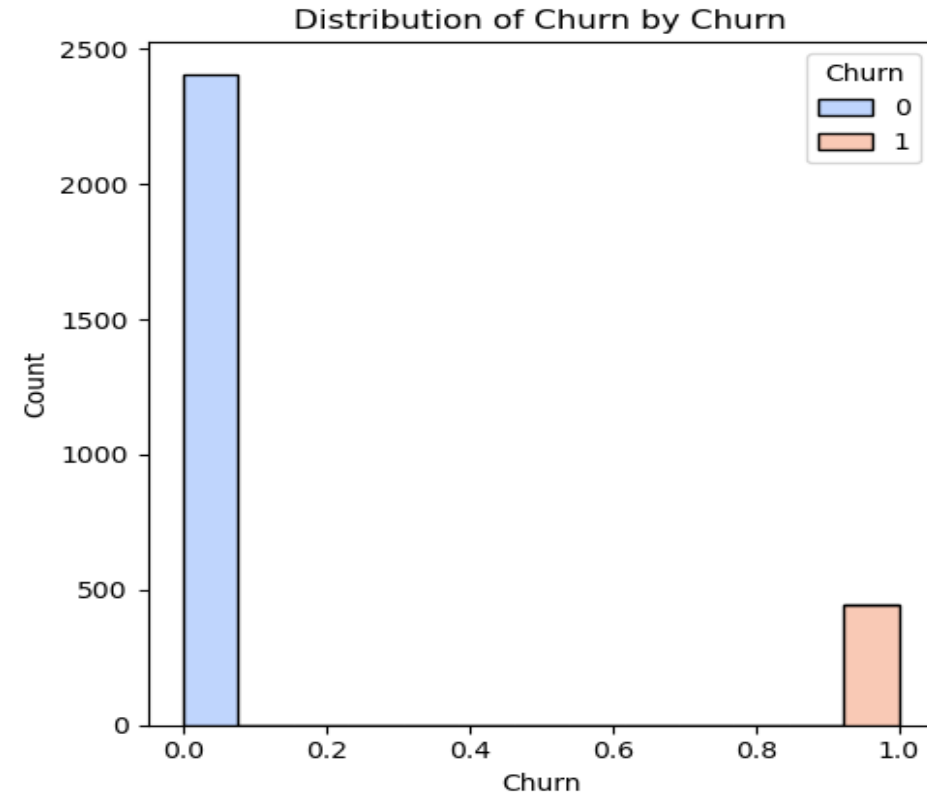
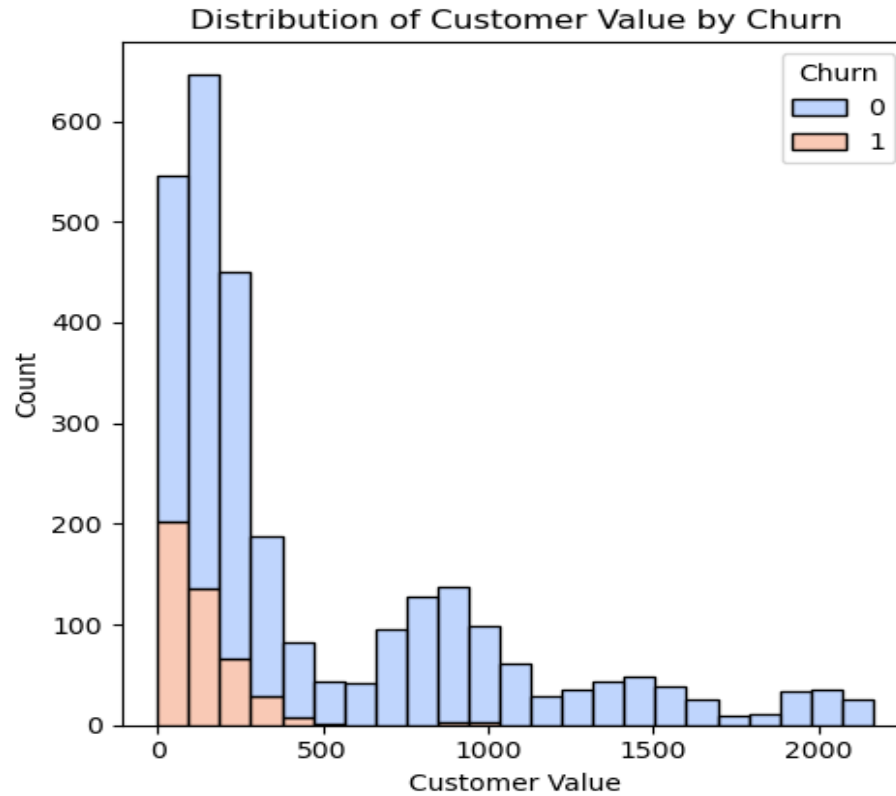
Histograms for Distributions



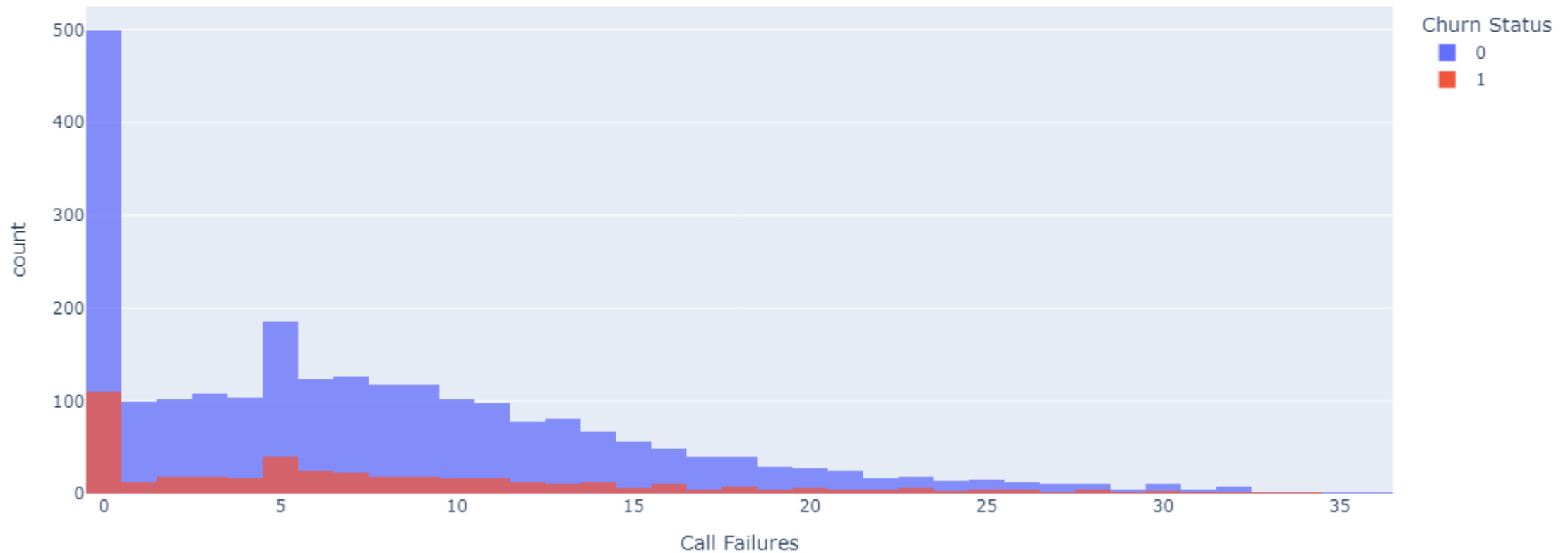
Histograms for Distributions



Histograms for Distributions

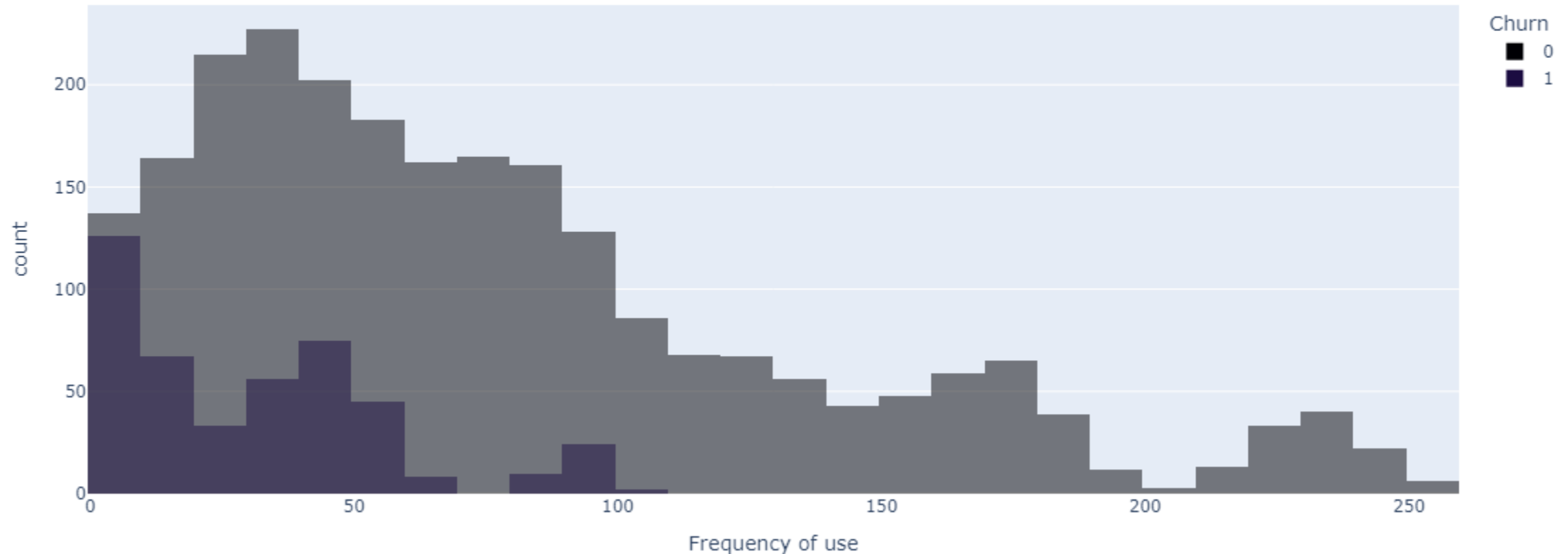


Call Failure Distribution by Churn Status



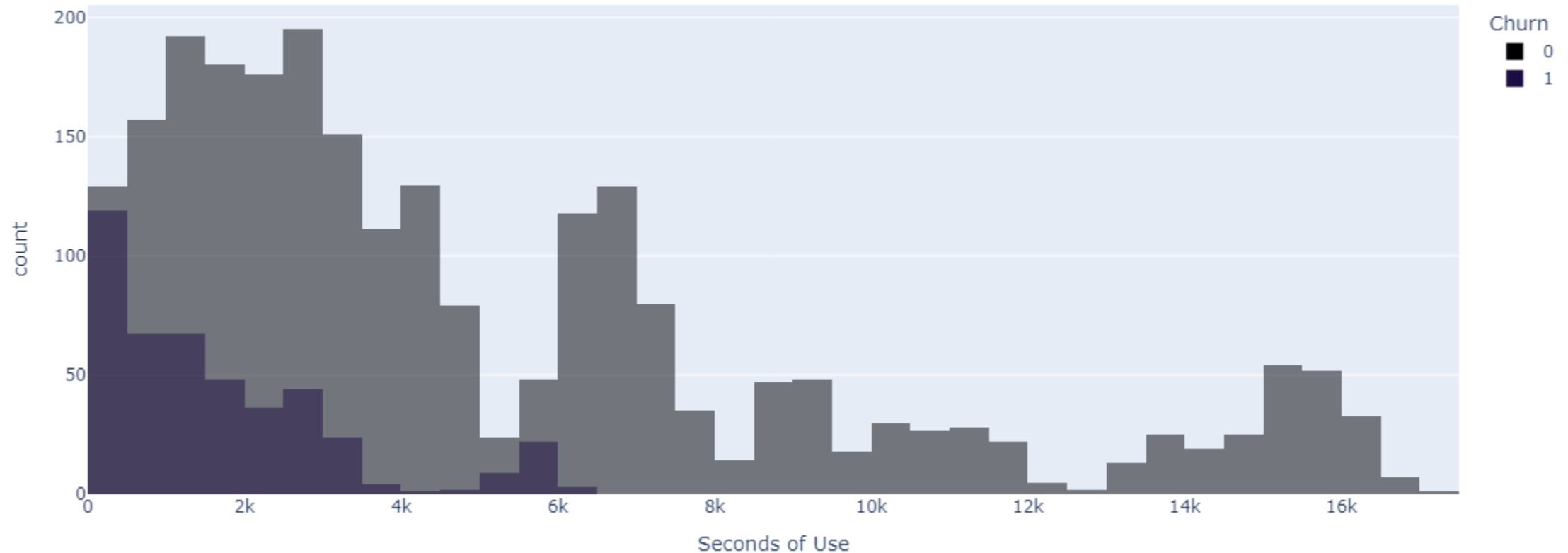
- A higher number of call failures correlates with increased churn rates.

Frequency of Use by Churn Status



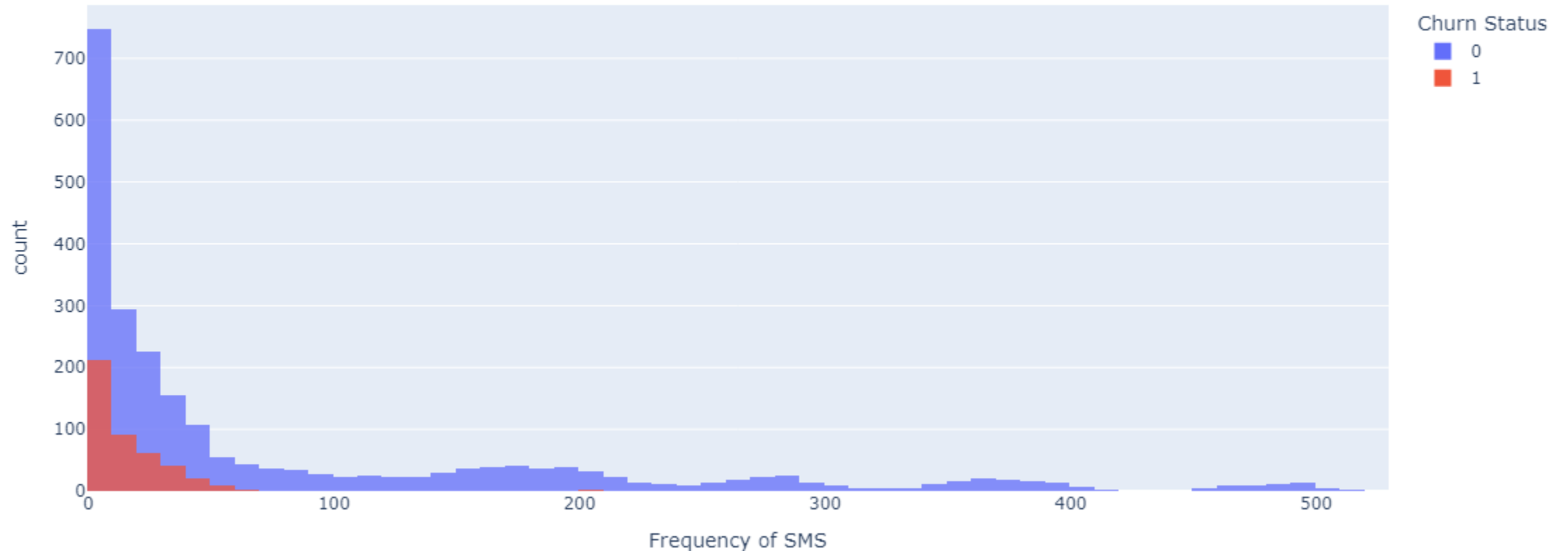
- Customers with *fewer service interactions* are more likely to **churn**. *Higher* frequency users are generally **retained**.

Seconds of Use by Churn Status



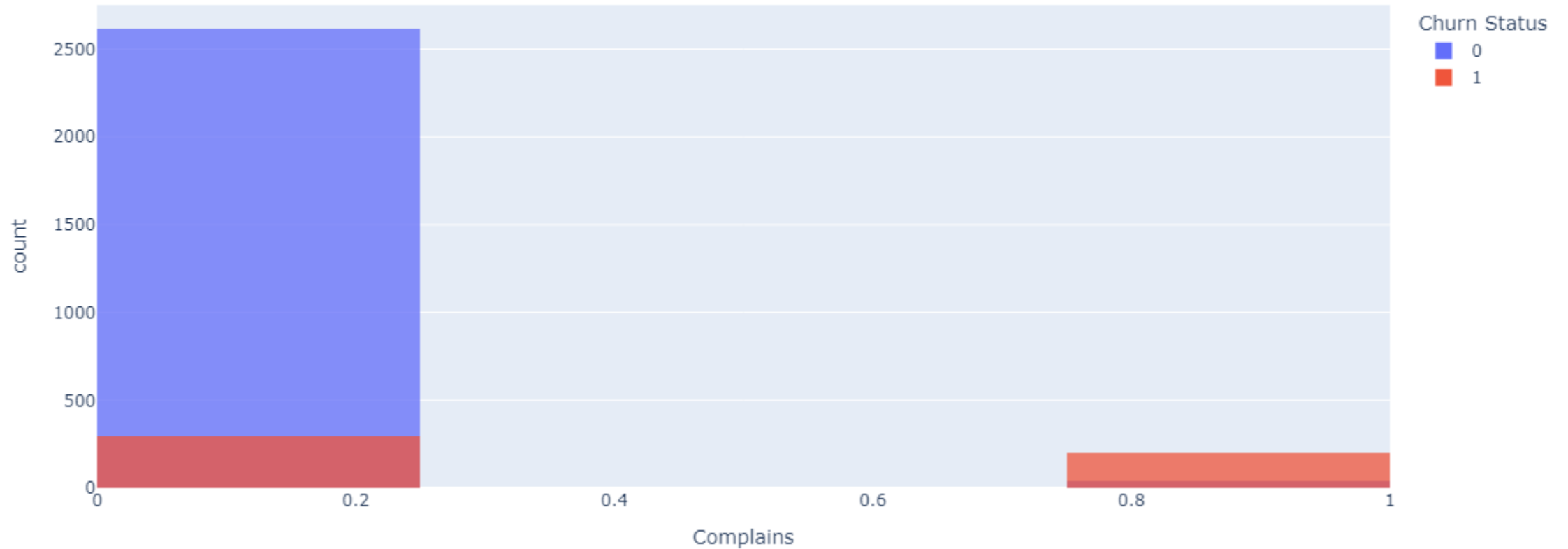
➤ **Churners** generally spend *less time on the service*. *Longer engagement* correlates with **retention**.

Frequency of SMS Distribution by Churn Status



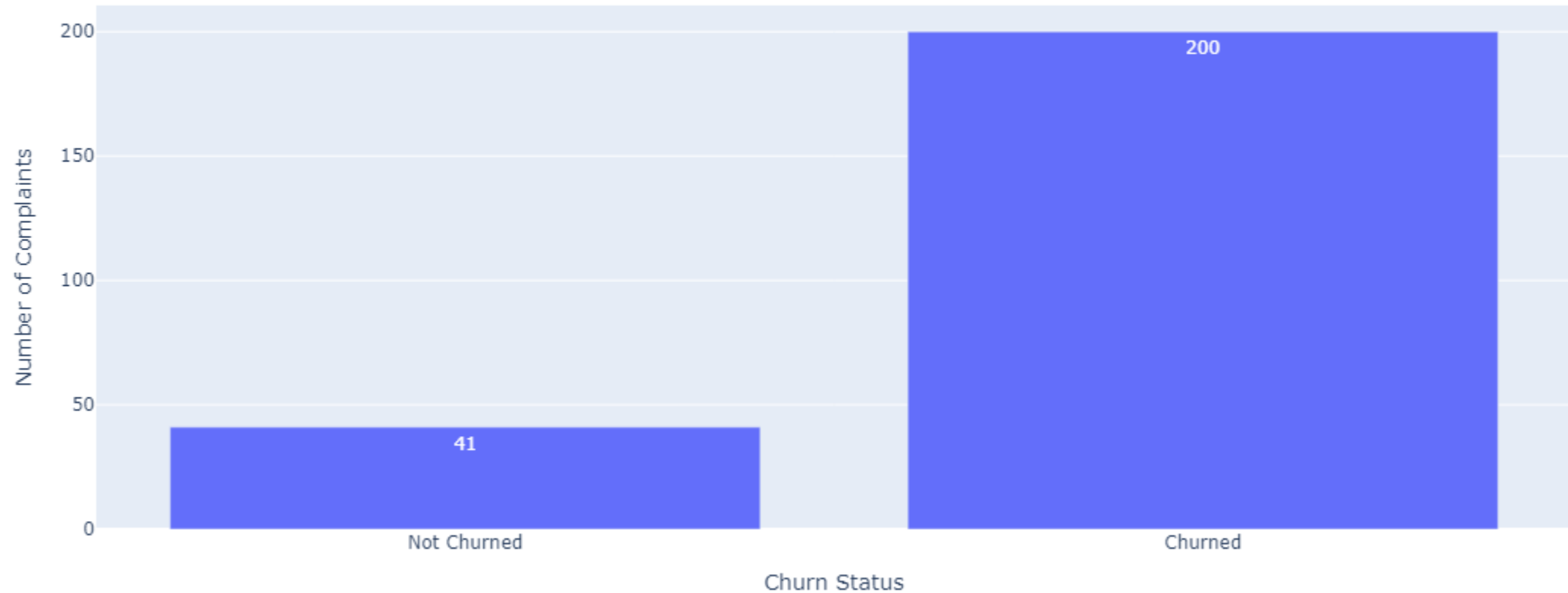
- Customers with *fewer service interactions* are more likely to **churn**. *Higher* frequency users are generally **retained**.

Complains Distribution by Churn Status



➤ Customers who register more complaints are more likely to churn.

Complains Distribution by Churn Status



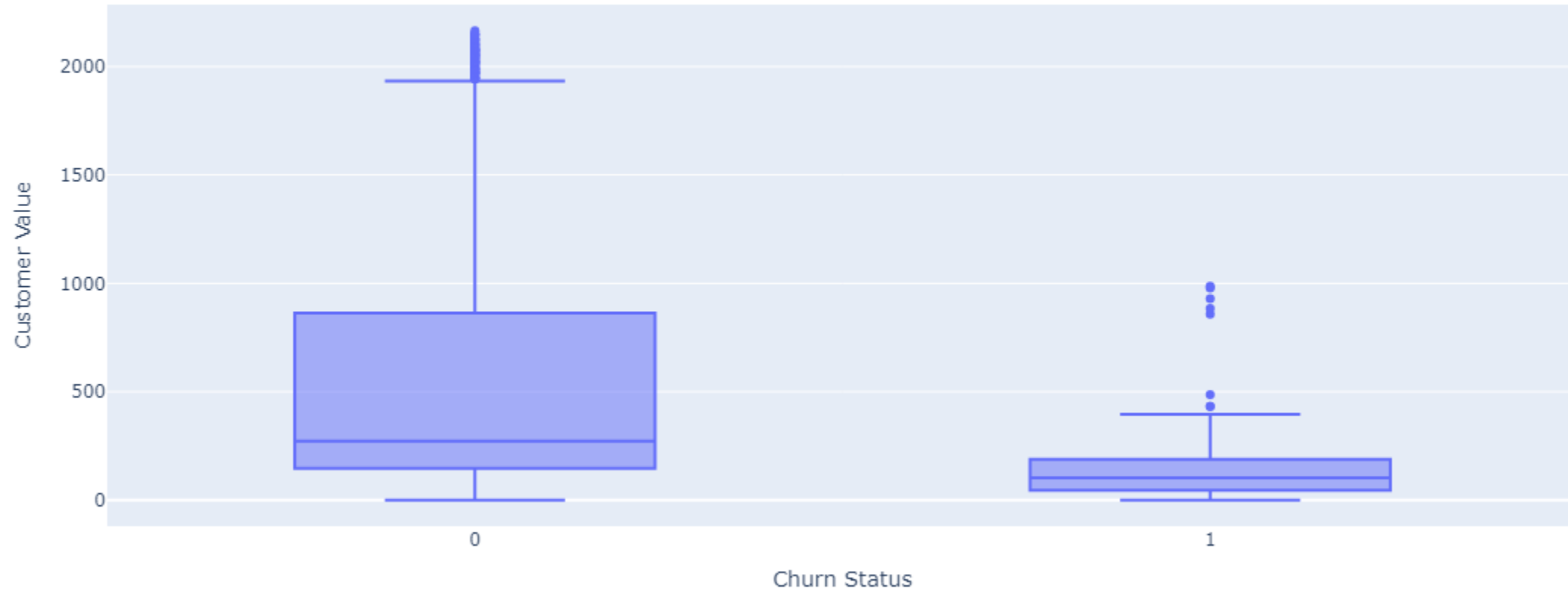
➤ Customers who register more complaints are more likely to churn.

Subscription Length Distribution by Churn Status



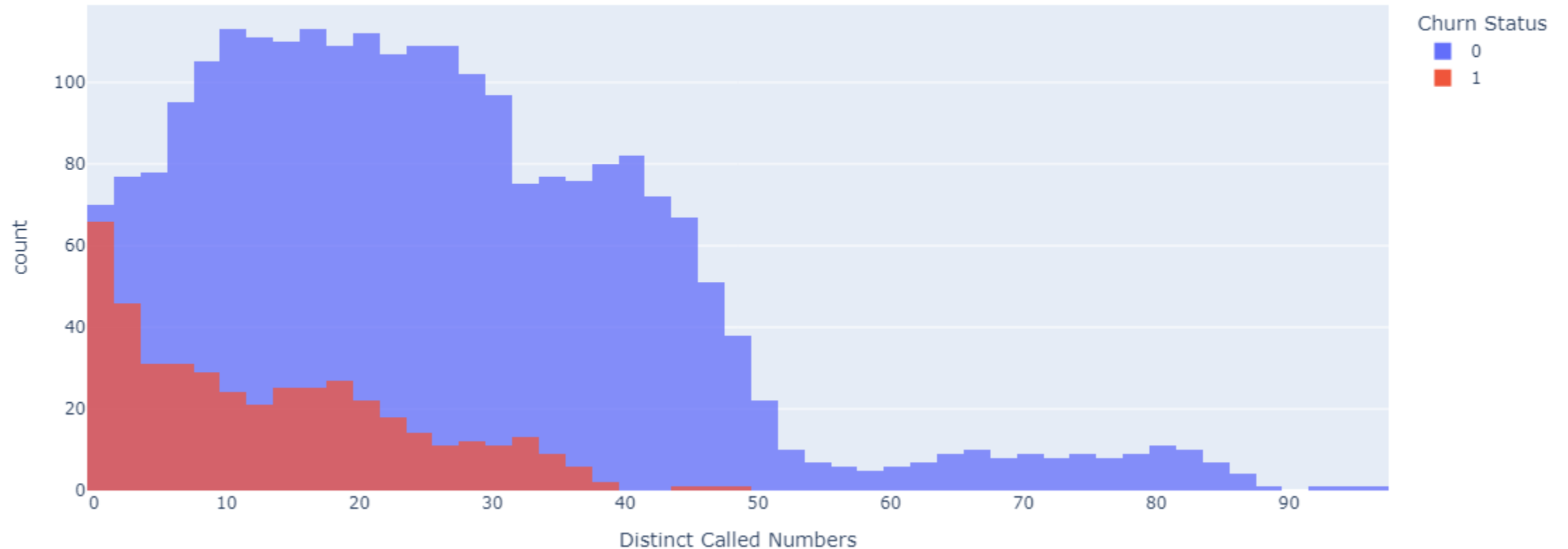
➤ Longer subscription lengths are associated with lower churn rates.

Customer Value Distribution by Churn Status



- Non-churned customers have a higher overall customer value distribution compared to churned customers.

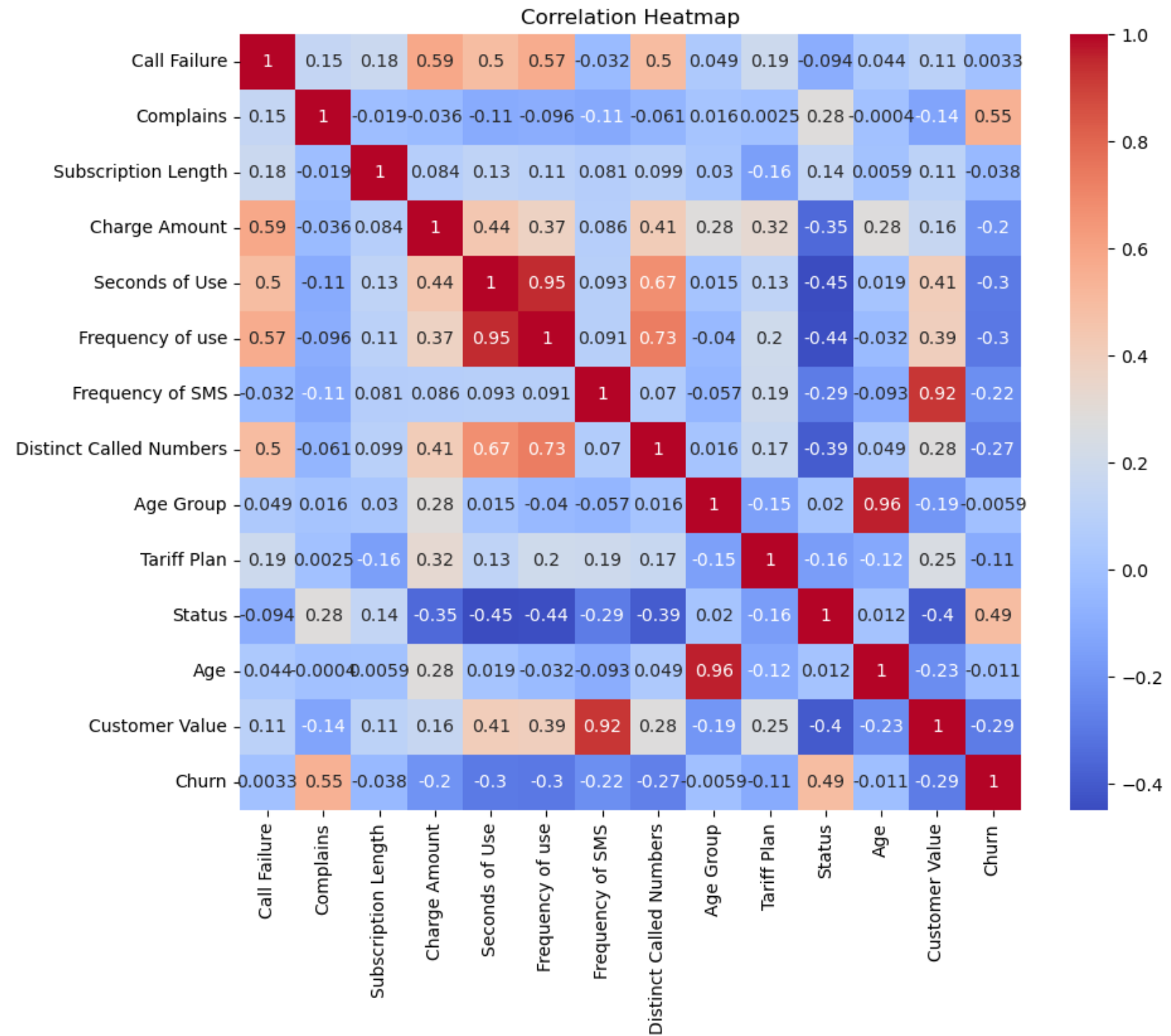
Distinct Called Numbers Distribution by Churn Status



➤ Customers who engage with a broader range of contacts are less likely to churn.



This heatmap visualizes the correlation between various features in the dataset and can guide feature selection for predictive models by highlighting the most relevant features.



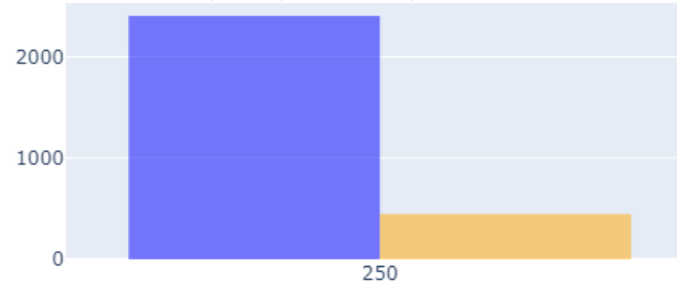
Interactive Dashboard Using Plotly

This Dashboard visually highlights key metrics differentiating churned (in yellow) and non-churned (in blue) customers.

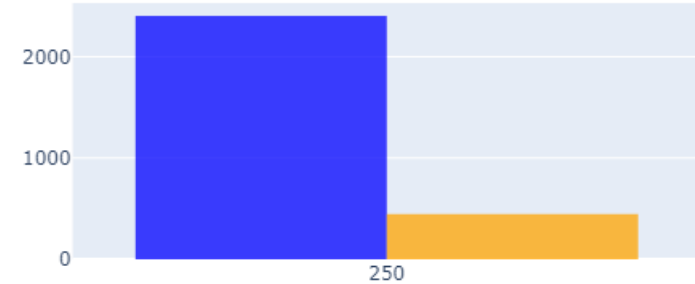
Customer Churn Dashboard

All Customers ▼

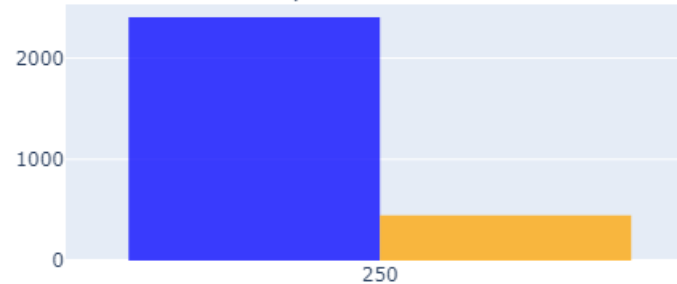
Frequency of Use by Churn Status



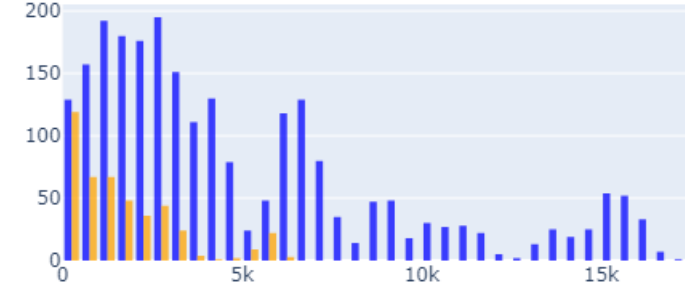
Call Failure Distribution



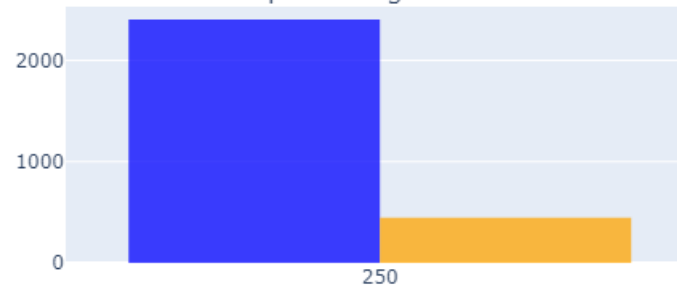
Complains Distribution



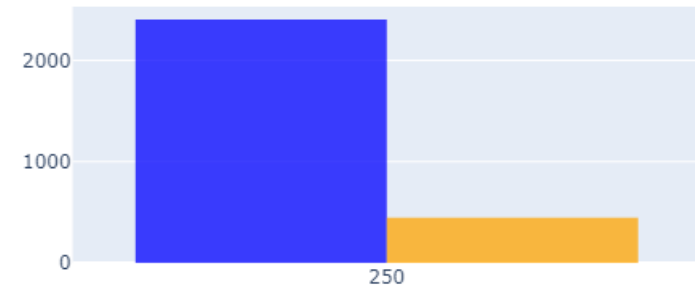
Seconds of Use



Subscription Length Distribution



Distinct Called Numbers Distribution














Summary of Findings

- Strong correlation between **usage frequency** and **churn** likelihood. *Low engagement* (frequency/seconds of use) *strongly correlates* with **churn**.
- EDA provided **actionable insights** for the predictive model.
- Insights were used to guide **key recommendations**.

Descriptive statistics

	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age
count	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000	2850.000000
mean	7.802456	0.080702	32.452982	0.974737	4534.243158	70.484912	73.789825	23.870526	2.835088	1.080351	1.240000	31.077193
std	7.326172	0.272424	8.723075	1.550618	4199.712303	57.401512	112.062397	17.193929	0.893503	0.271883	0.427158	8.861934
min	0.000000	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	15.000000
25%	1.000000	0.000000	29.000000	0.000000	1458.750000	28.000000	7.000000	11.000000	2.000000	1.000000	1.000000	25.000000
50%	6.000000	0.000000	35.000000	0.000000	3041.000000	54.500000	22.000000	21.000000	3.000000	1.000000	1.000000	30.000000
75%	12.000000	0.000000	38.000000	2.000000	6500.000000	96.000000	88.000000	34.000000	3.000000	1.000000	1.000000	30.000000
max	36.000000	1.000000	47.000000	10.000000	17090.000000	255.000000	522.000000	97.000000	5.000000	2.000000	2.000000	55.000000

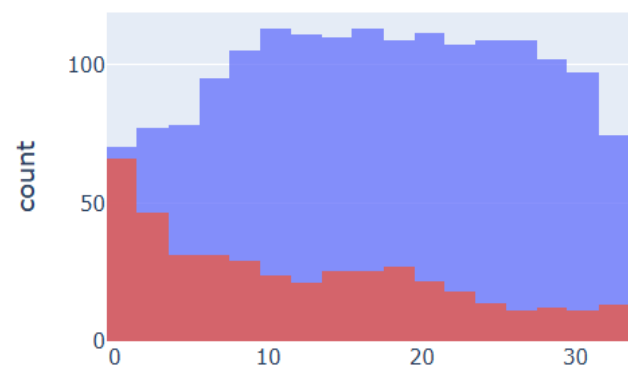
- 
- 
- 
- Call Failure:** On average, users experience 7.80 call failures, with a maximum of 36.
 - Complains:** The average complaint rate is 0.08, most customers don't complain (75% have 0 complaints), with very few having 1 complaint.
 - Subscription Length:** Customers have been subscribed for an average of 32.45 months, with a maximum of 47 months.
 - Charge Amount:** Average charge amount is low at 0.97, with a maximum of 10.
 - Seconds of Use:** Users spend an average of 4534 seconds using the service, with some using it as much as 17,090 seconds.
 - Frequency of Use:** The average frequency of use is 70.48, though it ranges up to 255.
- 
- 

- 
- 
- 
- **Distinct Called Numbers:** Users call an average of 23.87 distinct numbers.
 - **Age Group:** The average falls around 2.84, likely representing customers in the 2-3 age group category.
 - **Tariff Plan:** Most customers fall under the 1st tariff plan (mean = 1.08).
 - **Status:** The average status is 1.24, indicating most customers are active.
 - **Age:** The average age of users is 31.08, with a maximum of 55 years.
 - **Customer Value:** The average lifetime value of customers is 474.99, with a maximum of 2165.28.
 - **Churn:** The average churn rate is 0.156, meaning 15.6% of customers have churned.
- 
- 
- 

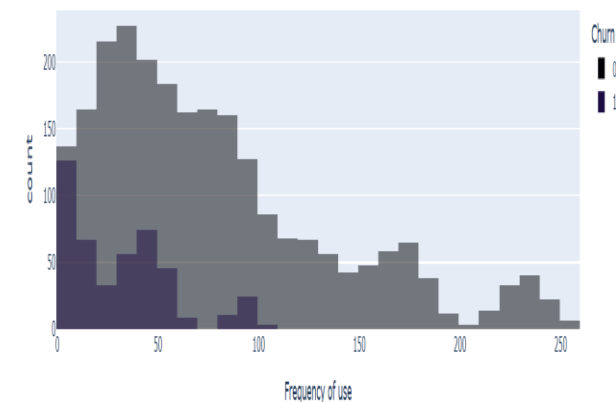
The visualizations provide the following insights:

1. **Churn Distribution:** As expected, the dataset is imbalanced, with the majority of customers not churning (about 84%).
1. **Key Feature Distributions:**
 1. **Call Failure:** Customers with more call failures tend to churn more frequently, suggesting a potential link between poor service quality and churn.
 2. **Complains:** A small portion of customers raise complaints, but those who do are more likely to churn.
 3. **Subscription Length:** Customers with shorter subscription lengths appear to churn more frequently.
 4. **Seconds of Use:** Those with lower usage seem more prone to churn.
 5. **Frequency of Use and SMS:** Lower usage and fewer SMS seem to correlate with higher churn rates.
 6. **Distinct Called Numbers:** Customers with fewer distinct numbers called are more likely to churn.
 7. **Age and Age Group:** No strong pattern, but younger customers might churn slightly more.
 8. **Customer Value:** Lower customer value tends to be associated with higher churn rates

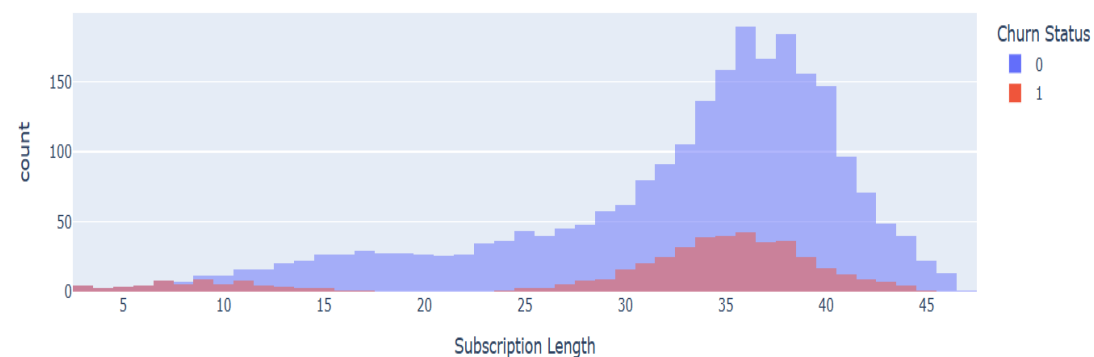
Distinct Called Numbers Distribution by Churn Status



Frequency of Use by Churn Status



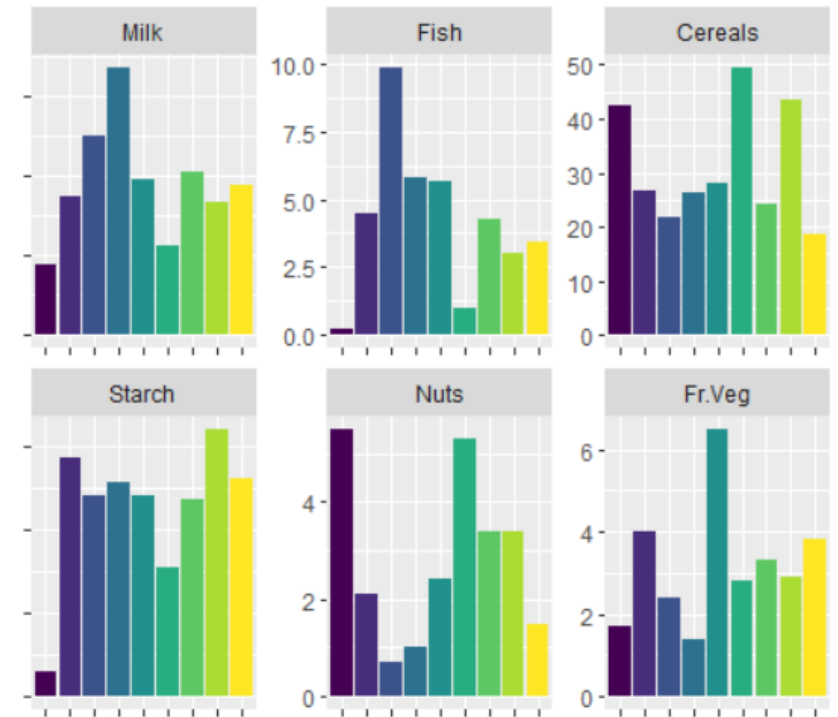
Subscription Length Distribution by Churn Status



Recommendations



- **Engagement Programs:** Develop targeted programs to increase the frequency and duration of use.
- **Technical Improvements:** Focus on minimizing call failures through better infrastructure and customer service.
- **Complaint Resolution:** Establish robust complaint management systems to improve customer satisfaction.
- **Incentives for Long-Term Plans:** Consider introducing loyalty programs or discounts for longer subscription commitments.
- **Cross-Functional Marketing:** Use marketing campaigns to promote the use of diverse services, ensuring customers are aware of all available options.

Clustering Analysis and Visualizations





Data Segmentation and Analysis Process

- The dataset was divided into churned and non-churned customers.
 - Each group was divided into four clusters according to the features defining each cluster.
 - Features classification: Average of use, Subscription Length, Status and Customer Value.
 - Based on the number of features associated, it was the best to divide each group into four cluster (Cluster 0 : Cluster 3).
- 
- 

Key Findings and Insights

- Based on analysis of correlation of the features with churned customers, it was found that **Call Failure, Complains, Subscription Length and Customer Value** are the factors that were found to have the strongest correlation with customer.
- Based on these features these are the key values found:
 - Group 0: **High Call Failure, Low Complains, Moderate Usage, Moderate Value**
 - Group 1: **Low Call Failure, Very Low Complains, Low Usage, Low Value**
 - Group 2: **Moderate Call Failure, High Complains, High Usage, High Value**
 - Group 3: **Highest Among All Groups, Moderate Complains, High Usage, Moderate High Value**
- ❑ Insights and Findings:
 - High call failures appear to be a primary issue. Addressing network or service quality problems could help prevent similar long-term customers from churning.
 - Implementing proactive churn prevention strategies, such as contacting long-term customers showing signs of dissatisfaction (e.g., increased call failures), could improve retention.

Non-Churned Customers Analysis

- Based on these features these are the key values found:

All Customers have respectively Good Subscription Length (Around 30 Months).

- Group 0: **Few or Low Call Failure and Complains.**
- Group 1: **Moderate Call Failure and High Value Customers.**
- Group 2: **Lowest Call Failure among the groups and No Complains.**
- Group 3: **High Call Failure and Low Complains.**
- ❑ Insights and Findings:

- **The goal here is trying to keep these customers as long as possible:**

For group 0: We can just keep the high quality of service.

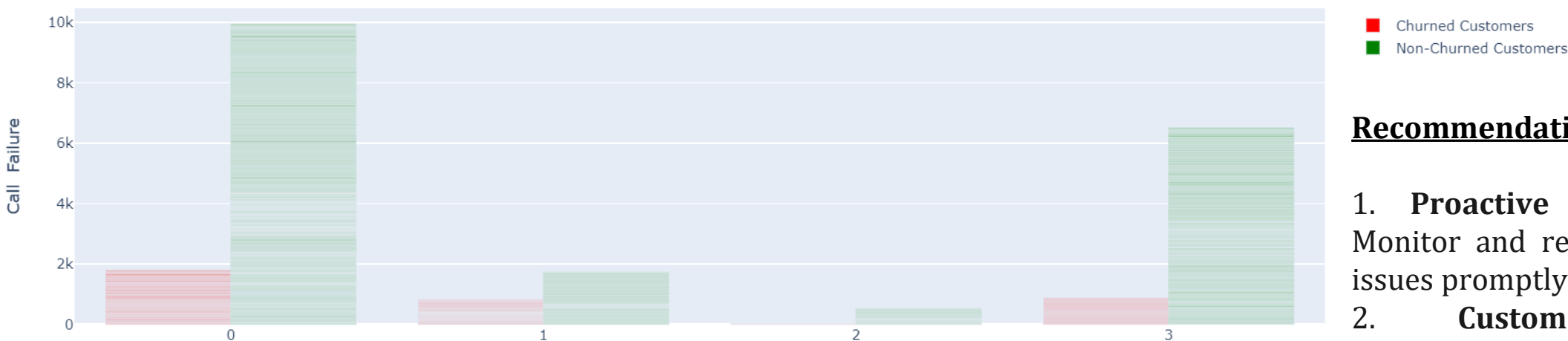
For group 1: This group is defined by high SMS usage Frequency and long Subscription Length, that's why they are high-valued so, we need to make the call failure for them even less even if their usage was mostly SMSs.

For group 2: This group is the priority to keep their subscription through personalized offers and continued high service quality.

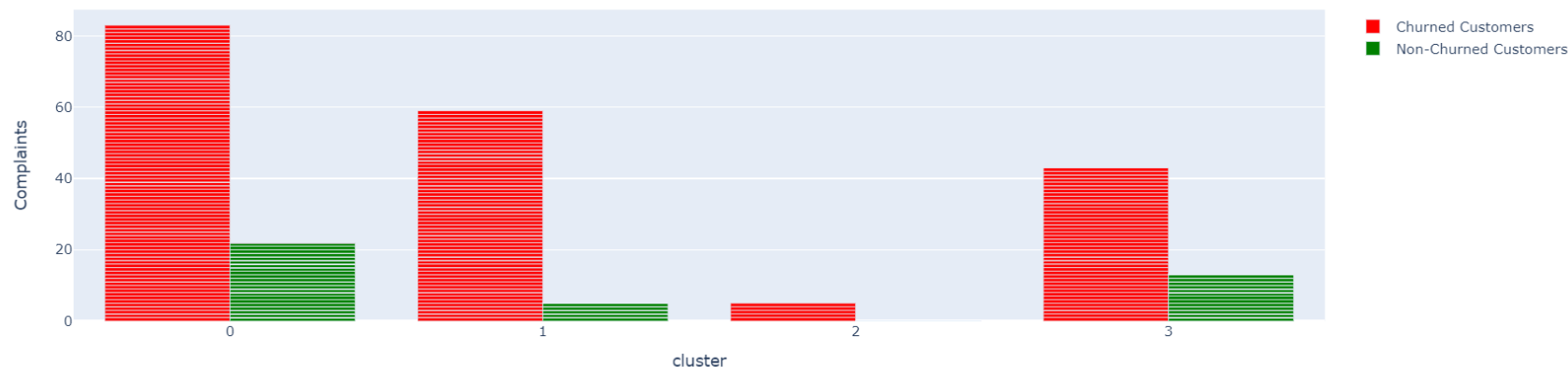
For group 3: They face potential issues with service quality that need addressing. Improving call reliability can significantly enhance customer satisfaction and retention in this segment.

Insights Visualizations

Call Failure Rate by Cluster



Complaints by Cluster

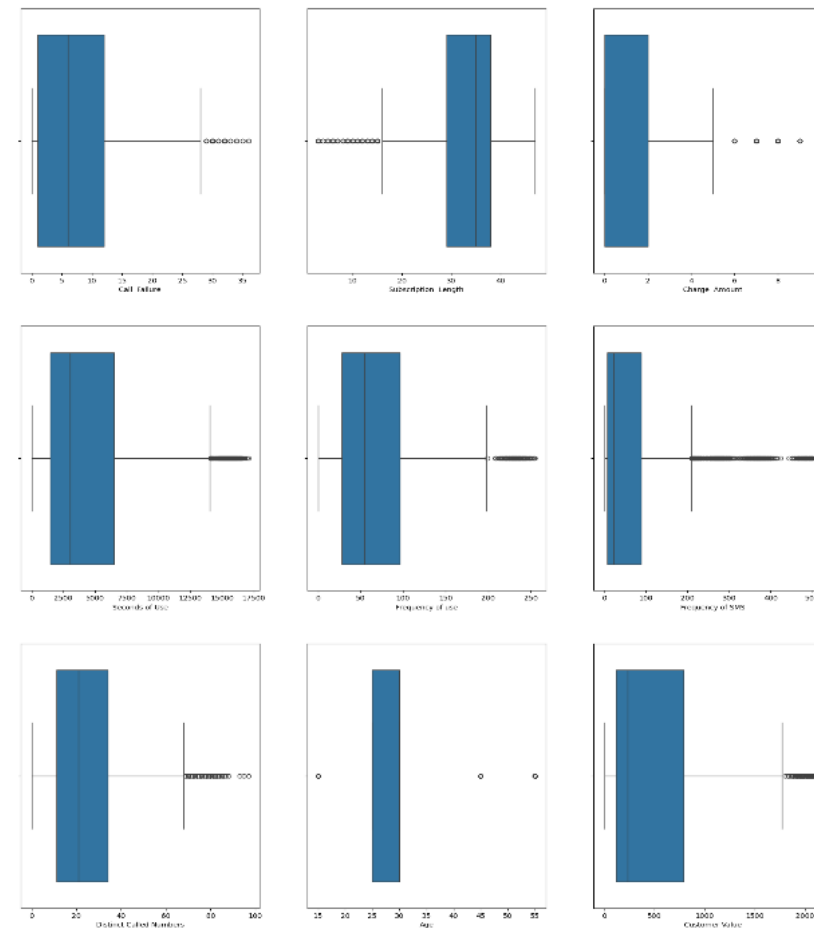
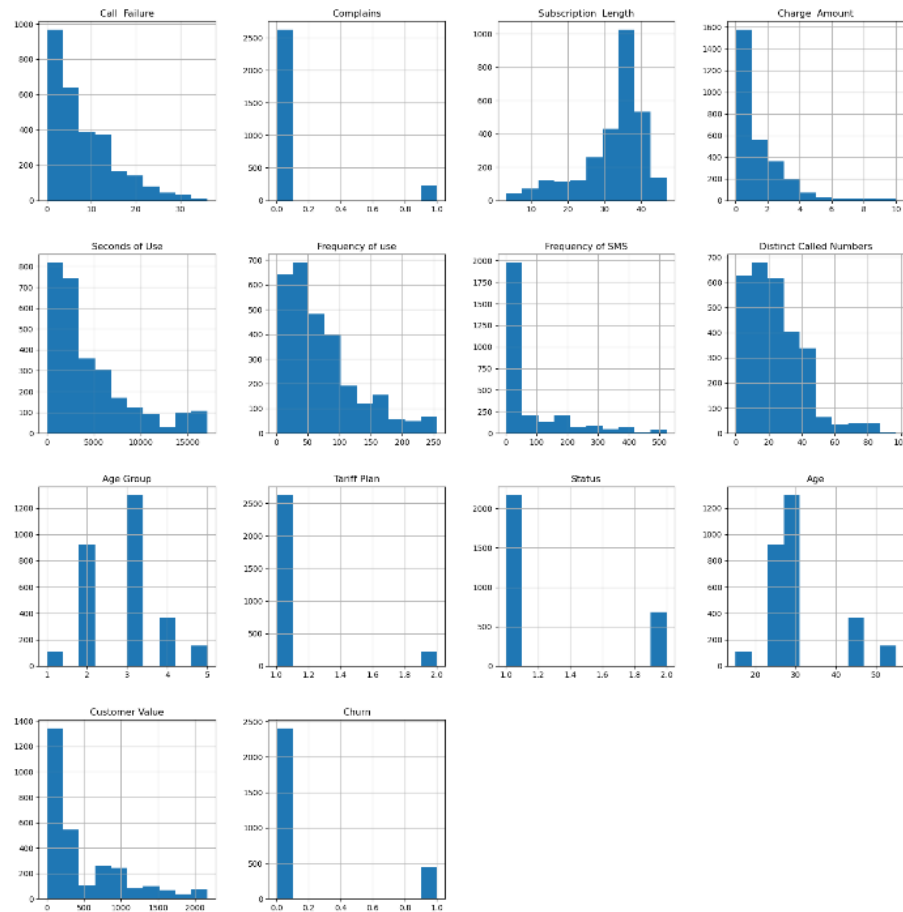


Recommendations and Next Steps:

- Proactive Issue Resolution:** Monitor and resolve service quality issues promptly.
- Customer Engagement Campaigns:** Develop initiatives to encourage usage, especially for low-usage customers.
- Tailored Retention Strategies:** Implement loyalty programs and premium support for high-value customers.
- Quality of Service Improvements:** Address network and technical issues to reduce churn rates.

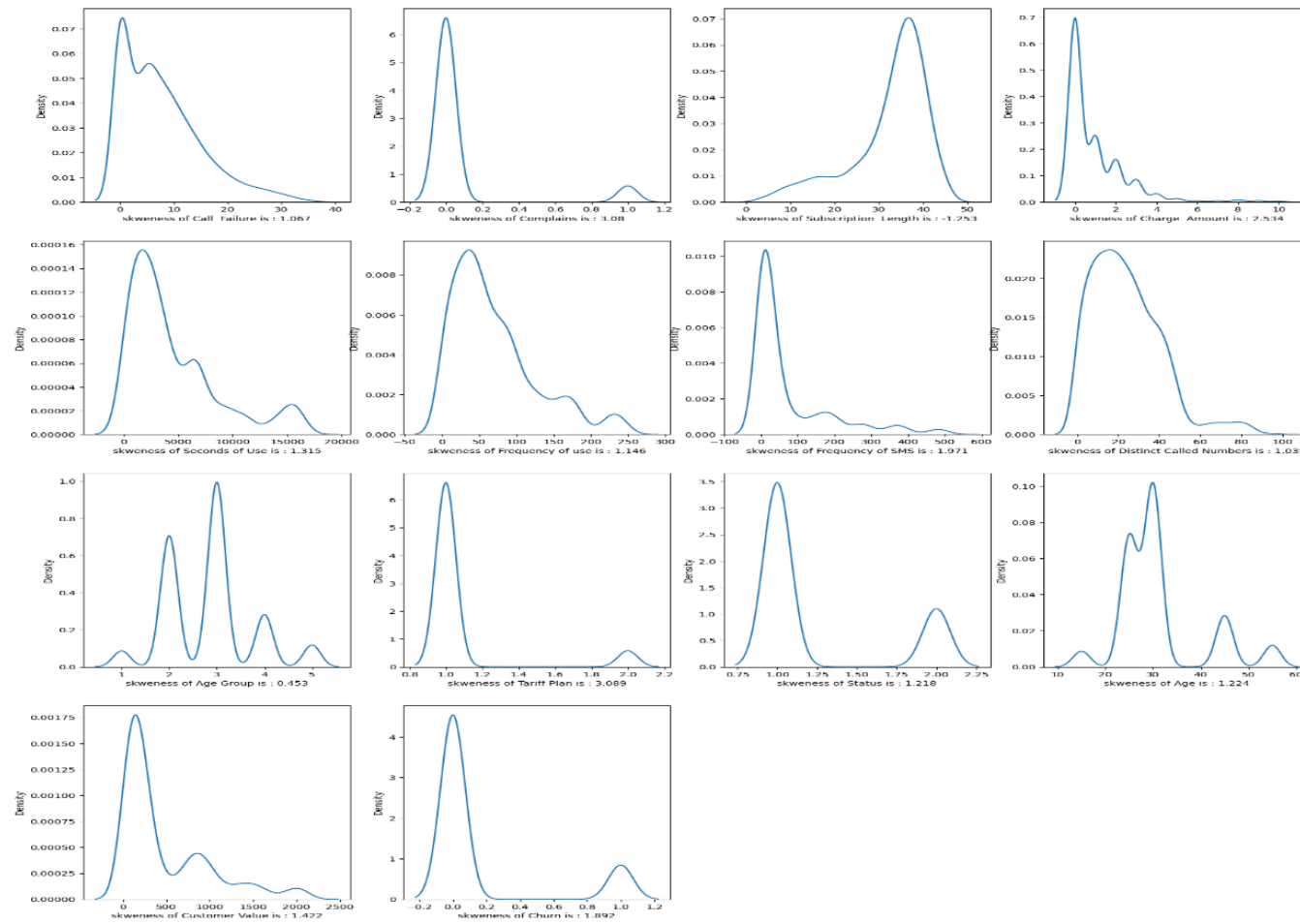
Outlier Detection

- Detecting outliers helps determine if they are due to data entry or measurement errors, or if they represent valid extreme values.
- Handling outliers is critical to ensure accurate model performance.



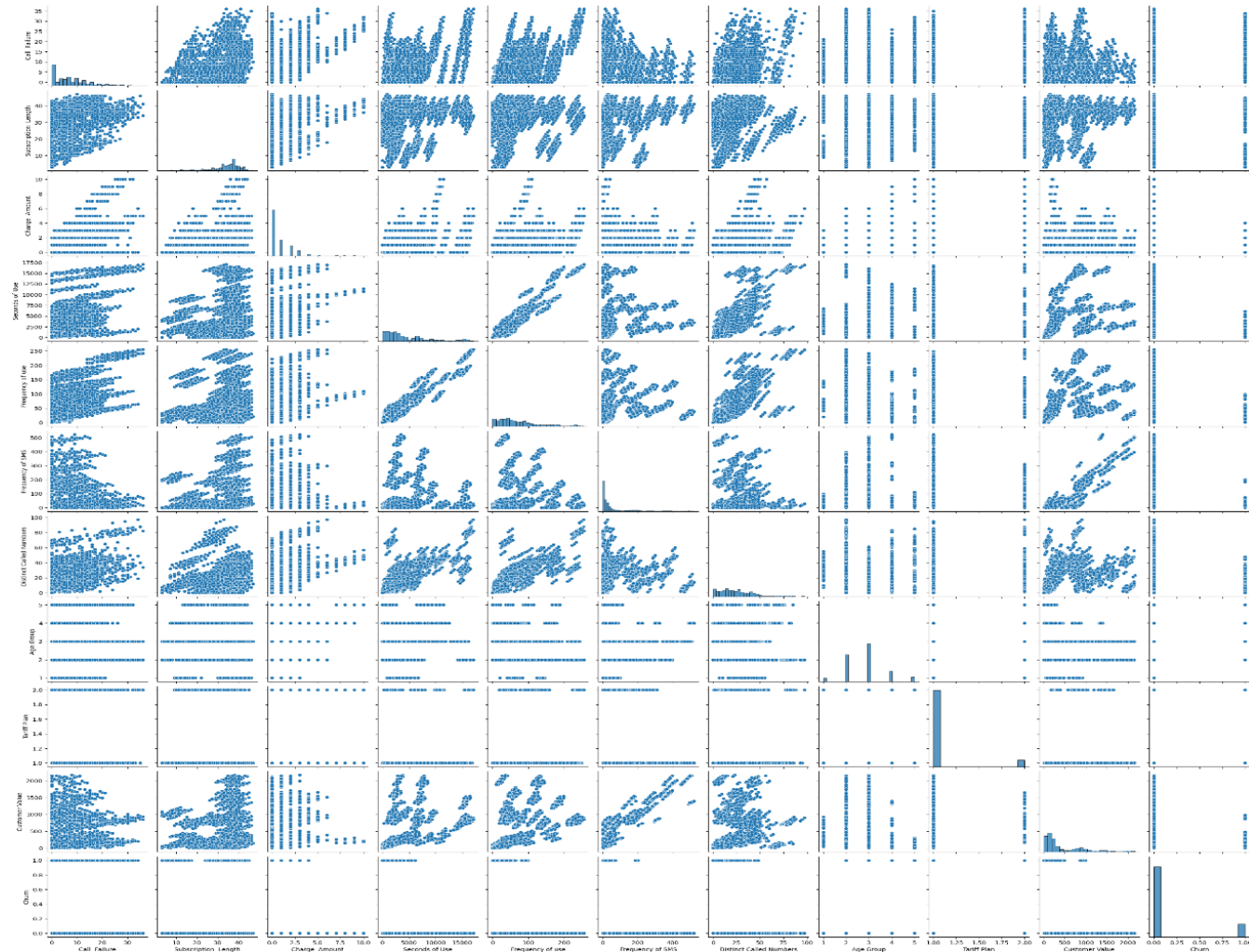
Feature Skewness

- Understanding the skewness of features allows us to decide whether transformation techniques are applicable.
- Skewness in discrete values can affect model accuracy.



Feature Relationships

- Use scatter plots to capture relationships between features and detect multicollinearity.
- Managing multicollinearity is essential to prevent overfitting in the model.



Feature Engineering

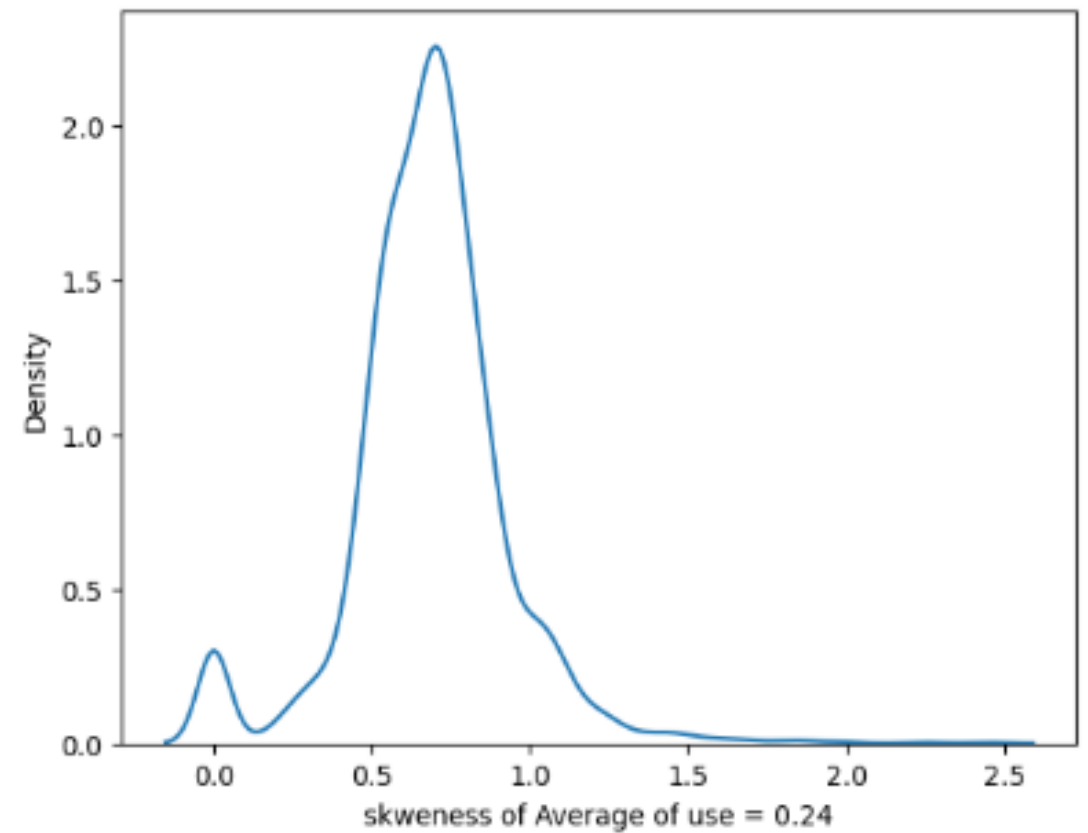
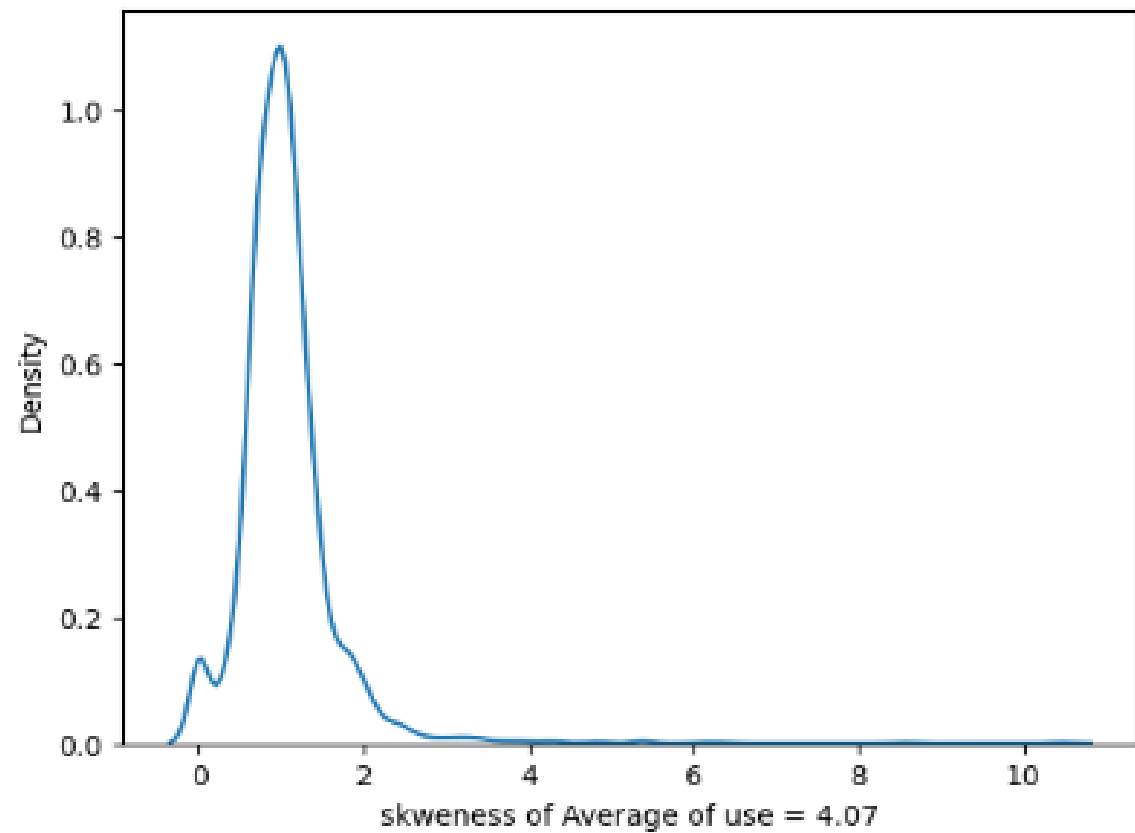
- Combine related features to create new, meaningful inputs for the model.
- Use techniques like log transformations to reduce skewness.

we'll combine the two features (Frequency of use, Seconds of Use) to overcome multicollinearity as they implies each other Seconds of Use are the total calls seconds and Frequency of use are the total calls of one customer we'll get new column named **Average Call Duration** and its formula = **Minutes of Use / Frequency of Use**

```
In [22]: data_eng['Average of use'] = (data_eng['Seconds of Use']/60)/data_eng['Frequency of use']  
data_eng.drop(['Frequency of use','Seconds of Use'],axis=1,inplace=True)  
data_eng.head()
```

```
Out[22]:
```

	Call Failure	Complains	Subscription Length	Charge Amount	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Customer Value	Churn	Average of use
0	8	0	38	0	5	17	3	1	1	197.640	0	1.025822
1	0	0	39	0	7	4	2	1	2	46.035	0	1.060000
2	10	0	37	0	359	24	3	1	1	1536.520	0	0.681389
3	10	0	38	0	1	35	1	1	1	240.020	0	1.060101
4	3	0	38	0	2	33	1	1	1	145.805	0	0.687644



Important Features

- Use mutual information to determine the most important features that affect churn.
- Prioritize these features during model building for improved predictions.

Mutual Information and Feature Selection

Mutual Information (MI) is a measure of the mutual dependence between two variables. It quantifies how much knowing the value of one variable reduces uncertainty about the other. When used for feature selection, MI helps to determine which features have the most predictive power regarding the target variable.

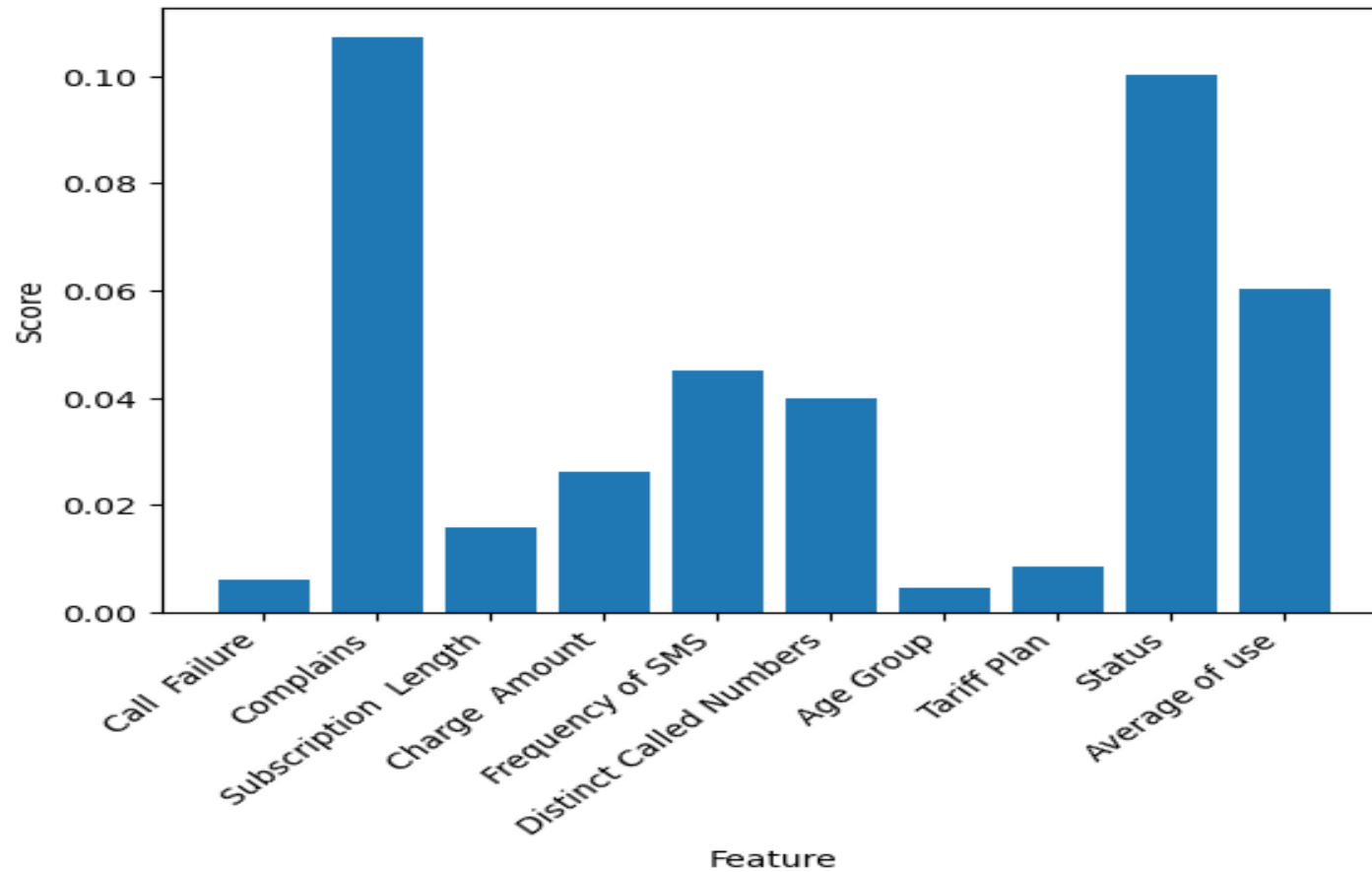
Formula for Mutual Information:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

Where:

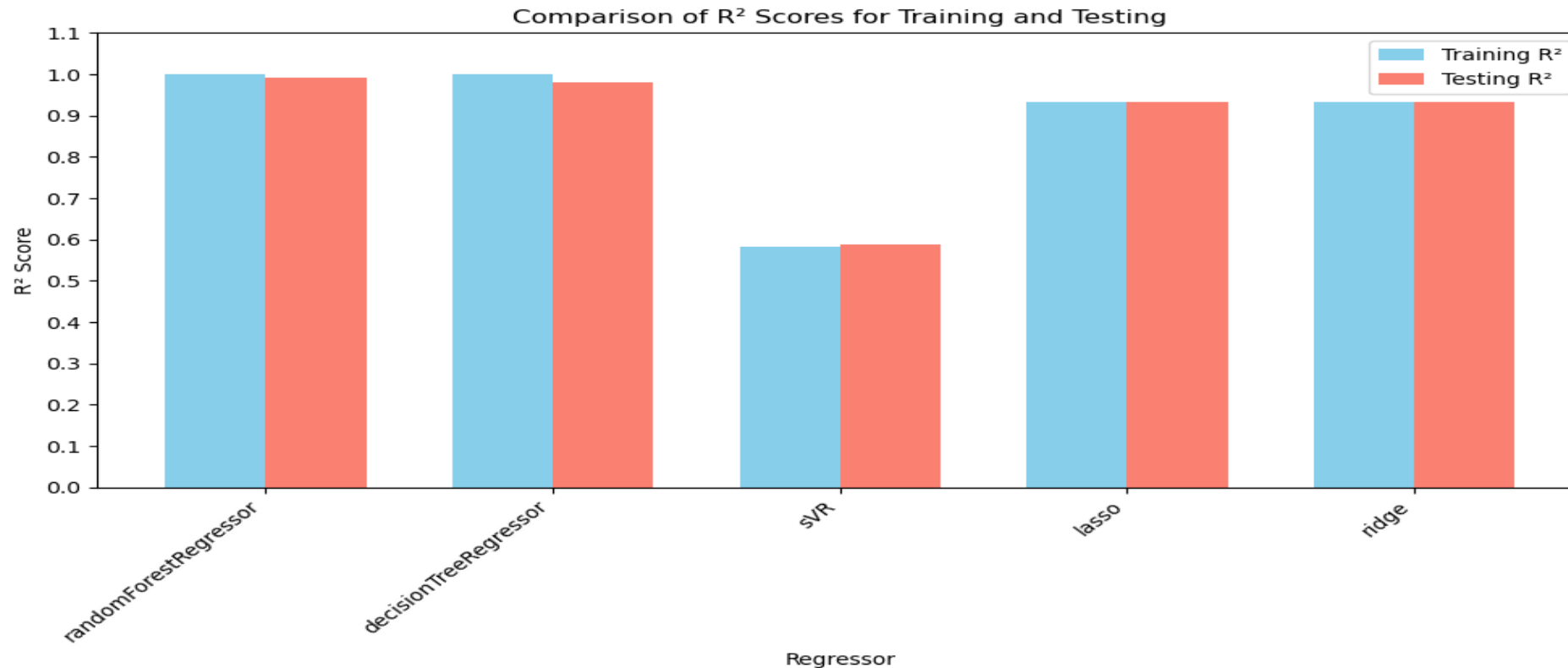
- (X) : Feature variable.
- (Y) : Target variable.
- $(p(x, y))$: Joint probability distribution of (X) and (Y) .
- $(p(x))$: Marginal probability of (X) .
- $(p(y))$: Marginal probability of (Y) .

most important features

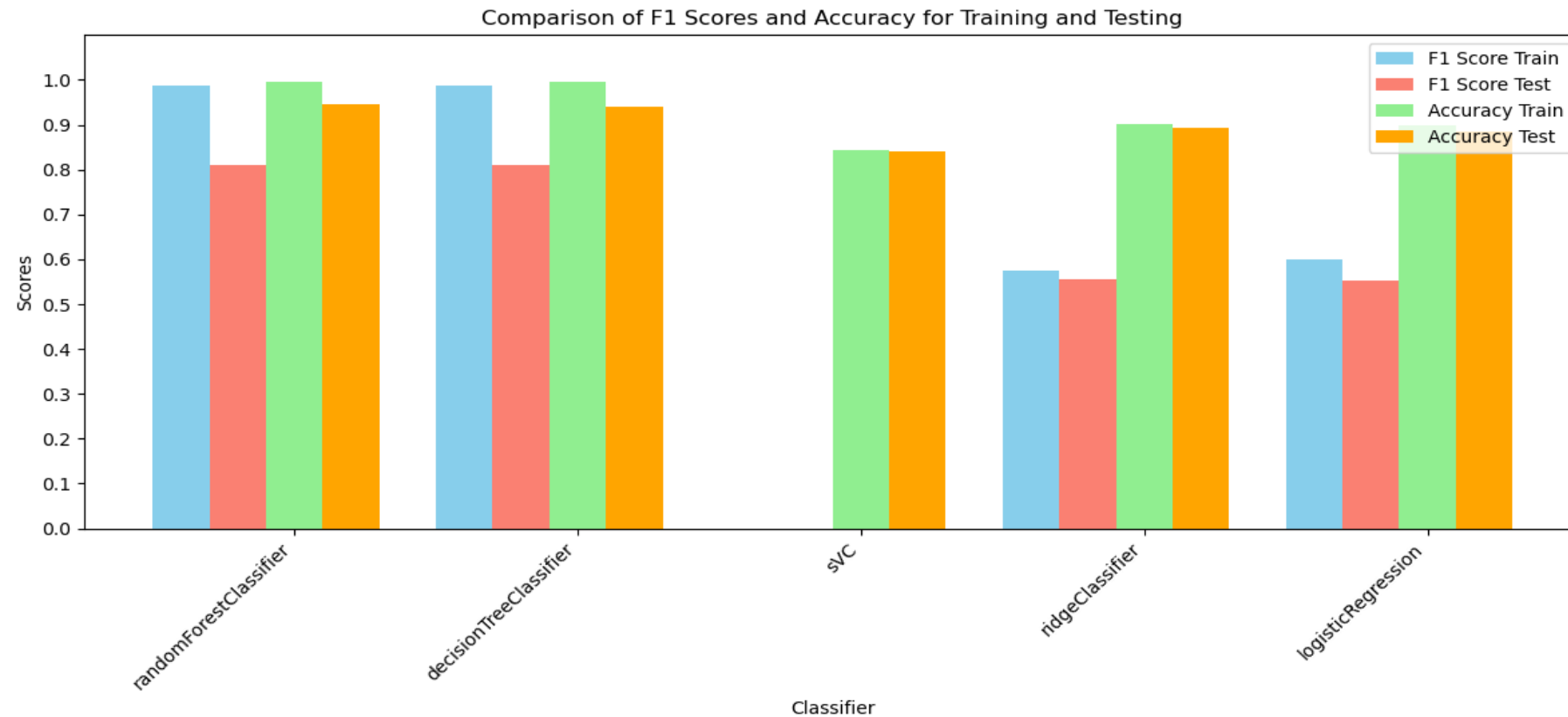


Model Selection

- Compare R^2 scores for regressors to evaluate customer scoring.

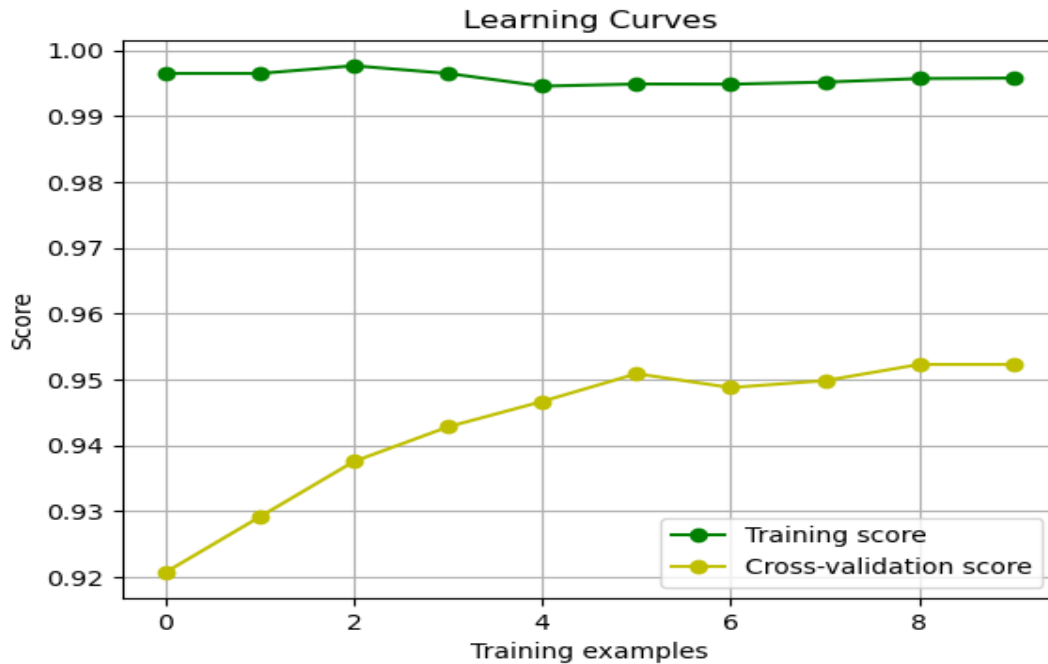


- Analyze F1 and accuracy scores for classifiers to predict churn.

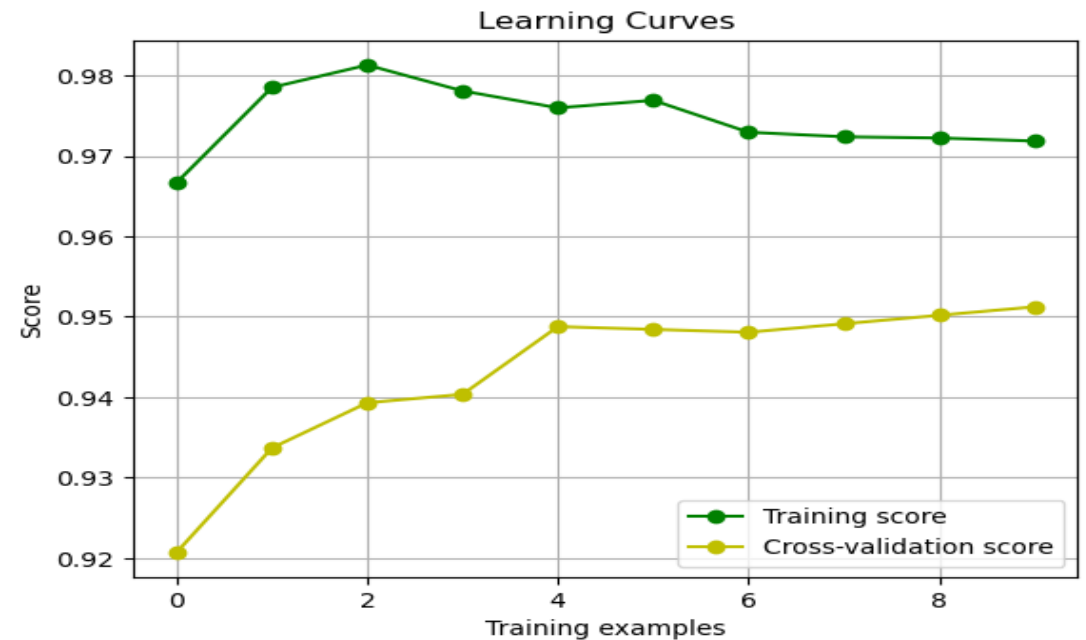


Best Performing Model

- The Random Forest model produced the best results after tuning and training.
- Learning curves show its superior fit to the data.



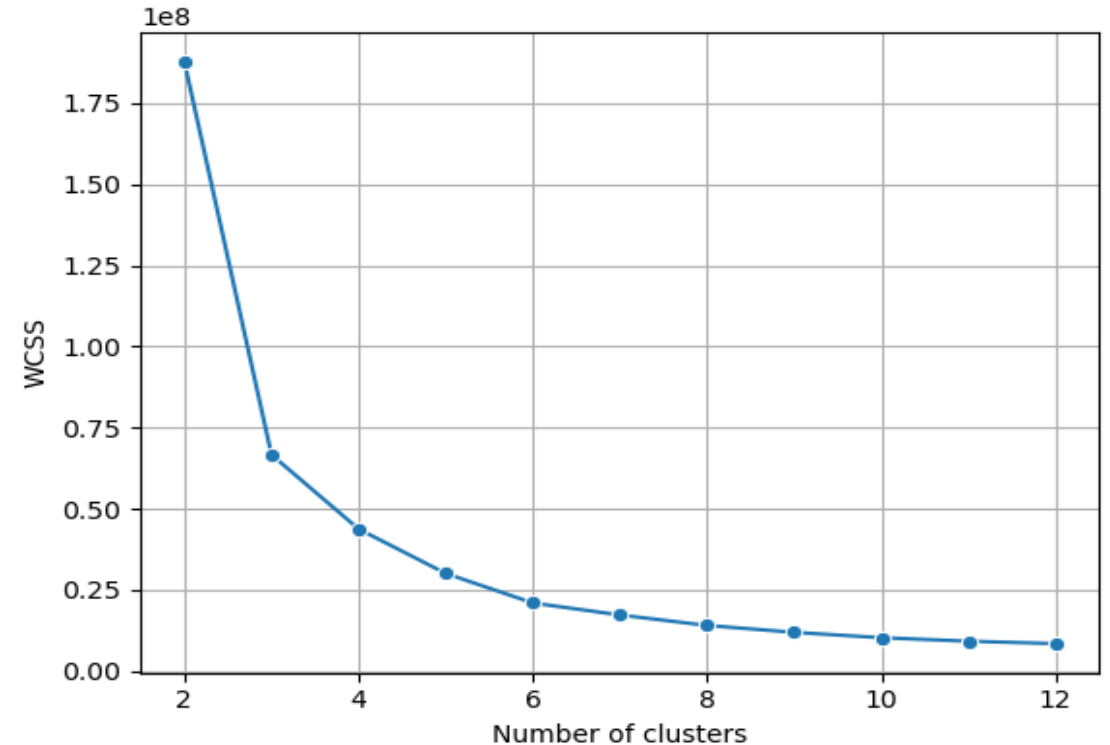
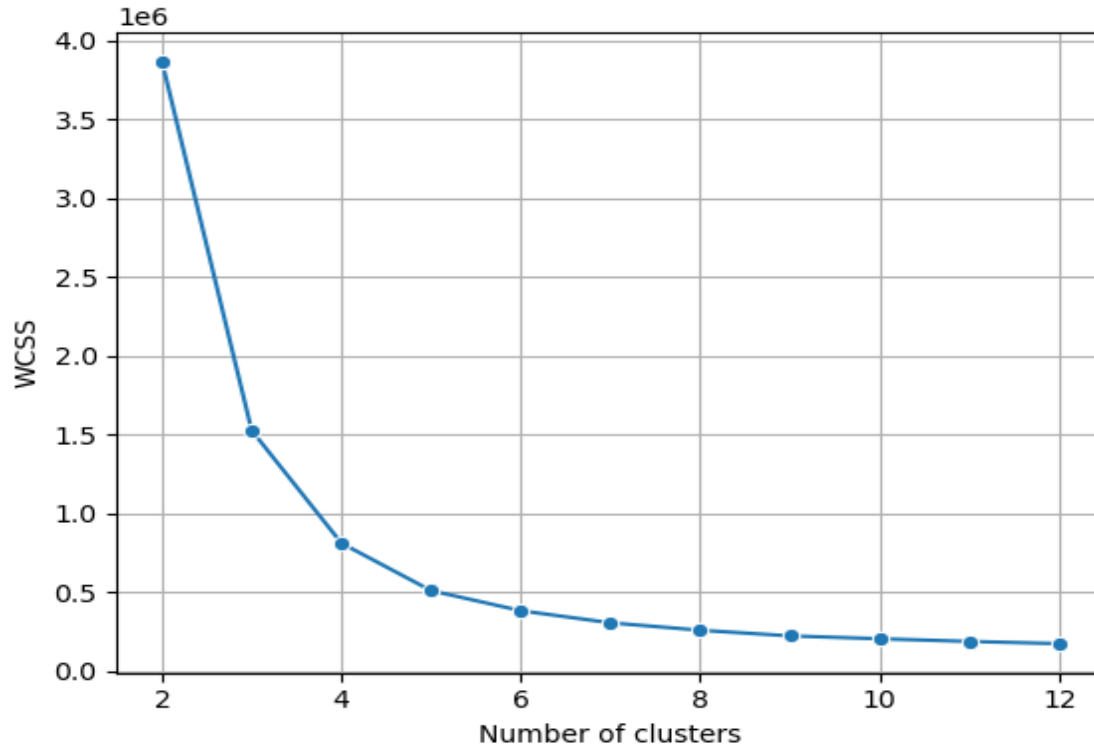
Random forest learning curve (best model fitted our data)



After fine tuning and training
The testing :F1 Score is :0.82
The test accuracy is :0.94

Customer Segmentation

- Unsupervised learning was used to segment customers into groups with distinct characteristics.
- Segmentation was applied to both churned and non-churned customers for deeper insights.



In [85]:

```
Churned = data_eng[data_eng['Churn']==1].copy()  
NotChurned = data_eng[data_eng['Churn']==0].copy()
```

we can apply elbow curve based on inertia WCSS

Inertia (WCSS)

$$\text{Inertia} = \sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

(n) is the total number of data points.

(x_i) represents each data point.

(μ_j) represents the centroid of the cluster to which (x_i) belongs.

(C) represents the set of all cluster centroids.



Deployment as a web application.

Customer Churn Prediction

Call failure:

Complains:

Subscription length:

Charge amount:

Frequency of sms:

Distinct called numbers:

Age group:

Tariff plan:

Status:

Average of use:

PREDICT

Predicted Customer Value: 1287.17

Churn Status: **Not Churned**

Customer Churn Prediction

Call failure:

Complains:

Subscription length:

Charge amount:

Frequency of sms:

Distinct called numbers:

Age group:

Tariff plan:

Status:

Average of use:

PREDICT

Predicted Customer Value: 62.09

Churn Status: **Churned**



Thank You

Feeling gratitude and not expressing it is like
wrapping a present and not giving it.