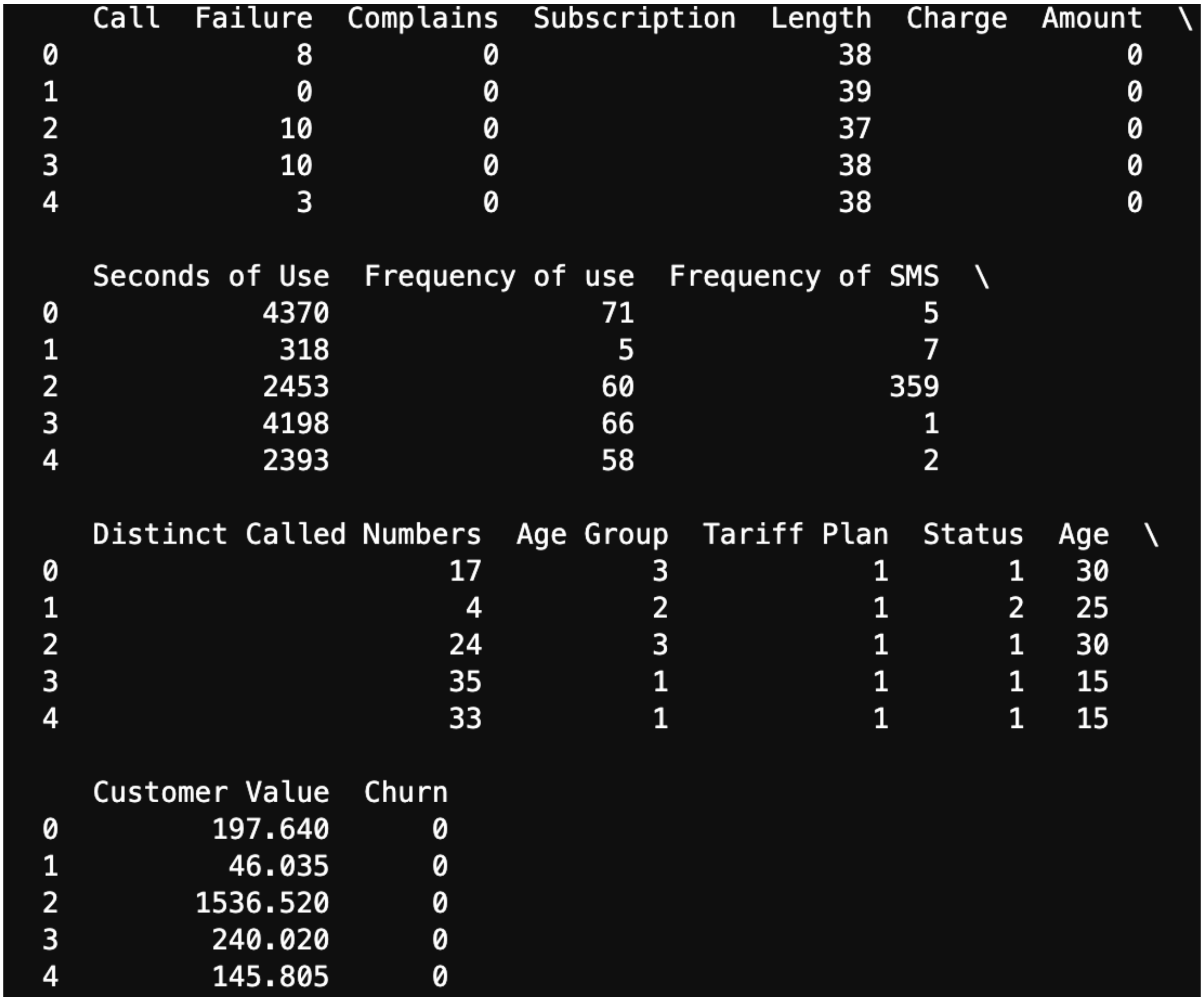
EXPLORATORY DATA ANALYSIS

3,150 customer records, each represented by 14 variables, make up the dataset used in this investigation. Customer behavior, use trends, service interaction, and demographic traits that are pertinent to churn prediction are all reflected in these aspects. The dataset has no missing values, according to a breakdown of its structure, guaranteeing a clear basis for preprocessing and modeling. The dataset is numerical and ideal for machine learning applications because all of its attributes are either of the integer or float data types. With 3,150 records for each characteristic, the dataset is comprehensive and devoid of missing information. This promotes more reliable modeling and reduces the need for imputation.



```
   Call  Failure  Complains  Subscription  Length  Charge  Amount  \
0              8          0                    38               0
1              0          0                    39               0
2             10          0                    37               0
3             10          0                    38               0
4              3          0                    38               0

   Seconds of Use  Frequency of use  Frequency of SMS  \
0            4370                71                 5
1             318                 5                 7
2            2453                60               359
3            4198                66                 1
4            2393                58                 2

   Distinct Called Numbers  Age Group  Tariff Plan  Status  Age  \
0                       17          3            1       1   30
1                        4          2            1       2   25
2                       24          3            1       1   30
3                       35          1            1       1   15
4                       33          1            1       1   15

   Customer Value  Churn
0         197.640      0
1          46.035      0
2        1536.520      0
3         240.020      0
4         145.805      0
```

Fig.1: Sample of Dataset (First Five Records)

These early records shed light on the distribution and organization of the data. For example, even if the majority of the customers in this sample have no complaints and no charges, there is a noticeable variation in their usage and value indicators, indicating considerable diversity in consumer behavior.

The following patterns were revealed by a first examination:

- Some users may experience up to ten call failures, with values ranging from low to quite high.
- The majority of consumers have extremely few or no complaints, which may be a sign of either underreporting or great service.
- The comparatively long subscription length (37–39 months for these cases) suggests that customers often remain subscribers for long stretches of time before discontinuing.
- The wide range of customer values (for example, 46.035 to 1536.52) might significantly affect the chance of turnover.
- The consumer base appears to be demographically diversified based on age groups and age values.
- The focus of the classification modeling will be the binary Churn variable.

Feature Distribution Analysis

Histograms were used to visualize univariate distributions in order to better comprehend the properties of each feature. This revealed information on the data's dispersion, modality, and skewness across all independent and dependent variables.

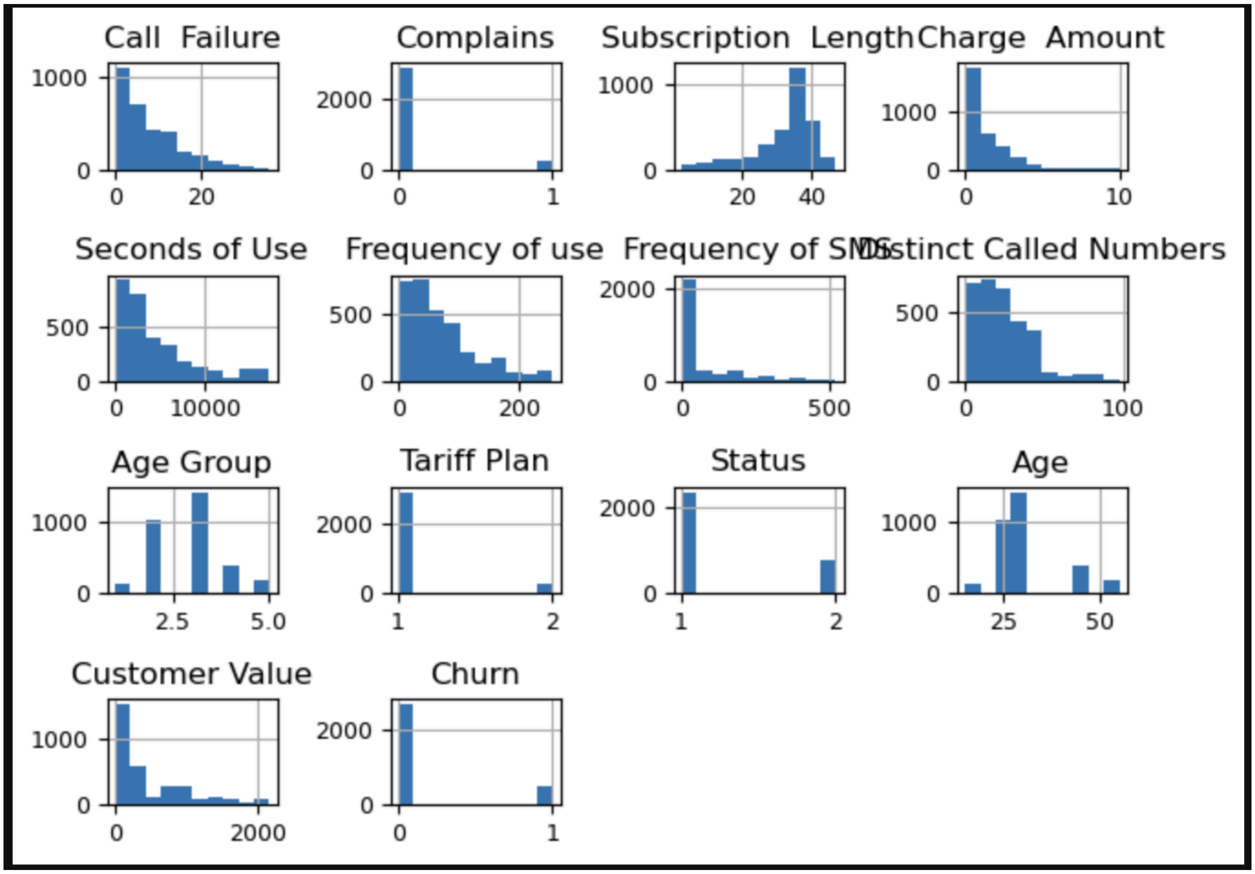


Fig.2: Univariate distributions of the Dataset

The plots above draw attention to a number of important findings:

- There are a few high-usage or high-value outliers among the majority of customers, as seen by the right-skewed distributions of Call Failure, Seconds of Use, Frequency of Use, Charge Amount, Distinct Called Numbers, and Customer Value.
- The majority of clients have zero complaints, subscribe to a single tariff plan, and belong to a dominating status group, making complaints, tariff plan, and status highly unbalanced categorical variables.
- With a concentration of clients with longer-term subscriptions (usually between 35 and 40 months), Subscription Length exhibits a considerable right skew.
- SMS frequency is highly skewed, with a tiny minority of users sending a disproportionately high number of messages and the vast majority sending very few.

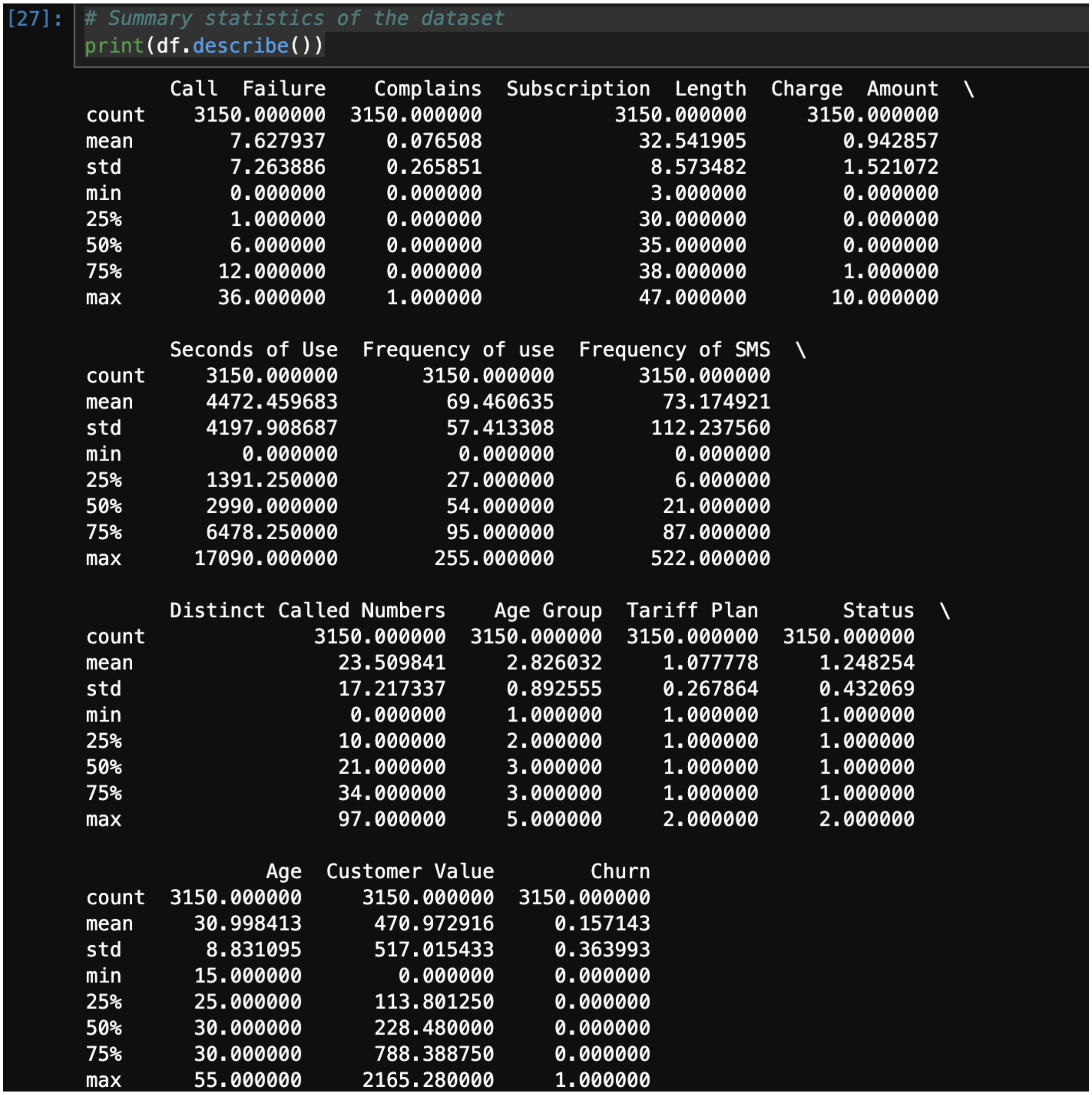
- The majority of people are in their early 20s to late 30s, with older age groups seeing a discernible fall. This is reflected in the Age Group categorical variable, where Group 3 has the highest population.
- The target variable, the churn variable, is wildly unbalanced. While churners (Churn = 1) make up a significantly lower percentage of the sample, the majority of consumers are non-churners (Churn = 0). Model performance is anticipated to be impacted by this class imbalance, which will be rectified with the proper methods.

Important hints on the underlying patterns in the data are provided by these distributions. For instance, the heavy-tailed nature of usage-based features implies that a small percentage of users are strong users, while the majority just utilize basic services. A Pareto-like distribution of revenue generation, in which a small fraction of customers account for a big share of value, may also be reflected in the skewness of customer value.

Comprehending these feature attributes was essential for directing the preprocessing pipeline (such as transformation and scaling) as well as the choice of appropriate models and assessment measures.

Descriptive Statistics

Numerous user behaviors and service usage patterns are revealed by a preliminary statistical analysis of the variables.



```
[27]: # Summary statistics of the dataset
      print(df.describe())
```

```
       Call  Failure    Complains  Subscription  Length  Charge  Amount  \
count    3150.000000  3150.000000           3150.000000     3150.000000   
mean        7.627937     0.076508             32.541905        0.942857   
std         7.263886     0.265851              8.573482        1.521072   
min         0.000000     0.000000              3.000000        0.000000   
25%         1.000000     0.000000             30.000000        0.000000   
50%         6.000000     0.000000             35.000000        0.000000   
75%        12.000000     0.000000             38.000000        1.000000   
max        36.000000     1.000000             47.000000       10.000000   

       Seconds of Use  Frequency of use  Frequency of SMS  \
count     3150.000000       3150.000000       3150.000000   
mean      4472.459683         69.460635         73.174921   
std       4197.908687         57.413308        112.237560   
min          0.000000          0.000000          0.000000   
25%       1391.250000         27.000000          6.000000   
50%       2990.000000         54.000000         21.000000   
75%       6478.250000         95.000000         87.000000   
max      17090.000000        255.000000        522.000000   

       Distinct Called Numbers    Age Group  Tariff Plan       Status  \
count              3150.000000  3150.000000  3150.000000  3150.000000   
mean                 23.509841     2.826032     1.077778     1.248254   
std                  17.217337     0.892555     0.267864     0.432069   
min                   0.000000     1.000000     1.000000     1.000000   
25%                  10.000000     2.000000     1.000000     1.000000   
50%                  21.000000     3.000000     1.000000     1.000000   
75%                  34.000000     3.000000     1.000000     1.000000   
max                  97.000000     5.000000     2.000000     2.000000   

               Age  Customer Value        Churn  
count  3150.000000     3150.000000  3150.000000  
mean     30.998413      470.972916     0.157143  
std       8.831095      517.015433     0.363993  
min      15.000000        0.000000     0.000000  
25%      25.000000      113.801250     0.000000  
50%      30.000000      228.480000     0.000000  
75%      30.000000      788.388750     0.000000  
max      55.000000     2165.280000     1.000000  
```

Fig.3: Descriptive statistics

With a standard deviation of 8.57 and a mean subscription duration of almost 32.5 months, the data indicates substantial variability in client tenure. With an average of 7.63 failures and a maximum of 36, call failure events exhibit a positively skewed distribution, suggesting that although the majority of consumers had comparatively few call failures, some had noticeably more.

75% of consumers are charged one or less, while a tiny percentage are charged ten or more, indicating a significant right-skewed charge amount. High standard deviations (4,198 and 517.02, respectively) are also seen for Seconds of Use and Customer Value, indicating significant variance in usage intensity and revenue contribution. The top 25% of customers have values over 788.39, while the median customer value is 228.48.

Communication styles also differ greatly. For example, a significant percentage of consumers send very few texts, as seen by the frequency of SMS, which varies from 0 to 522 with a median of only 21 messages. With a mean of 23.5 and a range of 0 to 97, Distinct Called Numbers is another characteristic that exhibits significant variability, indicating variations in users' social connection or contact variety.

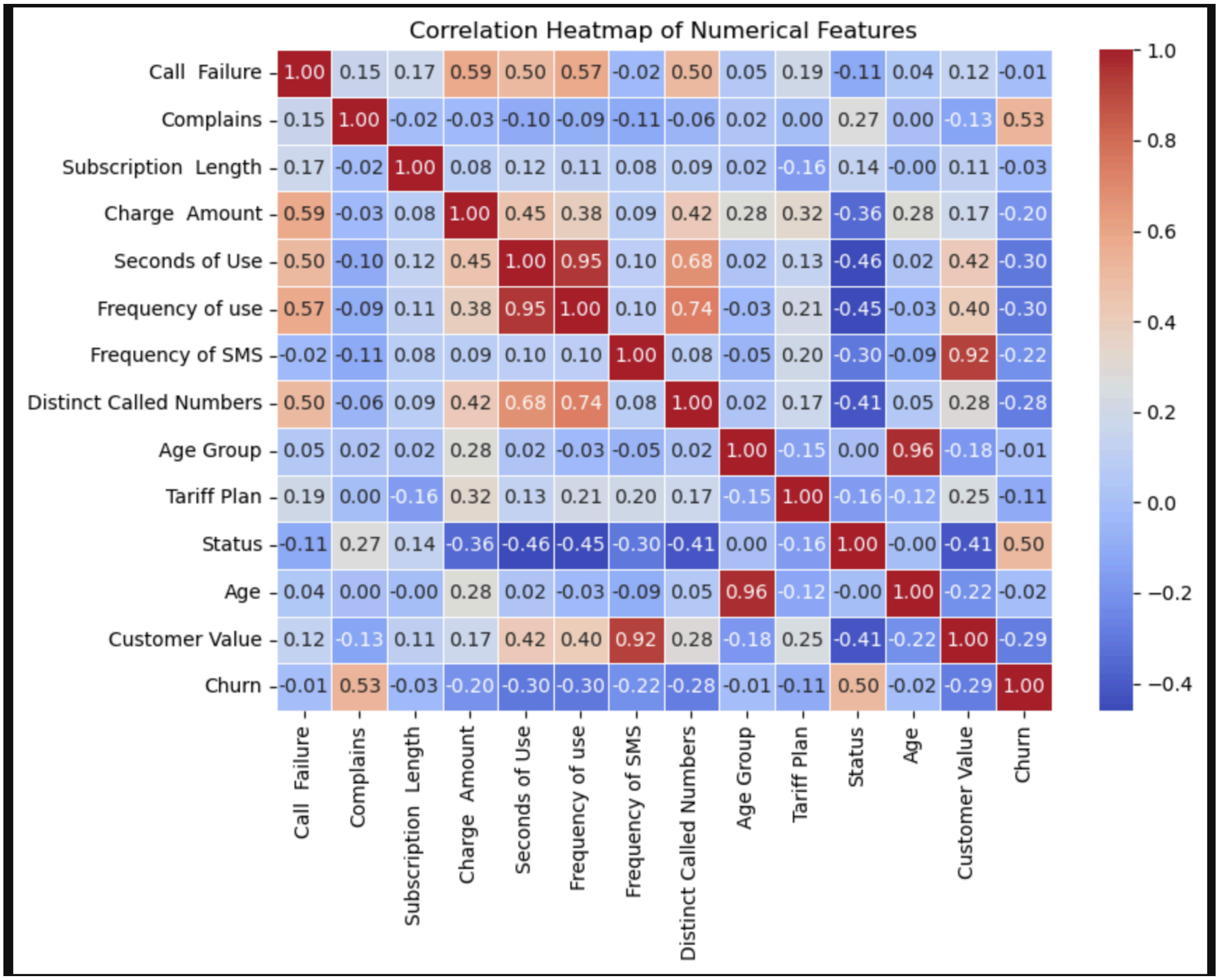
Demographic features like Age range from 15 to 55, with a mean of 31. The median closely matches the mean, indicating a strong central tendency in the distribution. The majority of users are in the 25–35 age range.

There is an unequal distribution of binary and category variables including Status, Tariff Plan, and Complaints. For instance, 92% of consumers are on the same pricing plan, while just 7.65% of customers complained. There is a class imbalance that has to be addressed in the modeling stage, since only 15.7% of the customers are classified as churners with regard to the Churn goal variable.

These descriptive statistics indicate possible predictive capability in aspects including usage patterns, customer value, and service experience (e.g., complaints, call failures), in addition to highlighting the skewness and variation across important behavioral markers.

Correlation Analysis

To investigate linear correlations between the dataset's numerical properties, a Pearson correlation heatmap was created (see Figure 4).



	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
Call Failure	1.00	0.15	0.17	0.59	0.50	0.57	-0.02	0.50	0.05	0.19	-0.11	0.04	0.12	-0.01
Complains	0.15	1.00	-0.02	-0.03	-0.10	-0.09	-0.11	-0.06	0.02	0.00	0.27	0.00	-0.13	0.53
Subscription Length	0.17	-0.02	1.00	0.08	0.12	0.11	0.08	0.09	0.02	-0.16	0.14	-0.00	0.11	-0.03
Charge Amount	0.59	-0.03	0.08	1.00	0.45	0.38	0.09	0.42	0.28	0.32	-0.36	0.28	0.17	-0.20
Seconds of Use	0.50	-0.10	0.12	0.45	1.00	0.95	0.10	0.68	0.02	0.13	-0.46	0.02	0.42	-0.30
Frequency of use	0.57	-0.09	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	0.21	-0.45	-0.03	0.40	-0.30
Frequency of SMS	-0.02	-0.11	0.08	0.09	0.10	0.10	1.00	0.08	-0.05	0.20	-0.30	-0.09	0.92	-0.22
Distinct Called Numbers	0.50	-0.06	0.09	0.42	0.68	0.74	0.08	1.00	0.02	0.17	-0.41	0.05	0.28	-0.28
Age Group	0.05	0.02	0.02	0.28	0.02	-0.03	-0.05	0.02	1.00	-0.15	0.00	0.96	-0.18	-0.01
Tariff Plan	0.19	0.00	-0.16	0.32	0.13	0.21	0.20	0.17	-0.15	1.00	-0.16	-0.12	0.25	-0.11
Status	-0.11	0.27	0.14	-0.36	-0.46	-0.45	-0.30	-0.41	0.00	-0.16	1.00	-0.00	-0.41	0.50
Age	0.04	0.00	-0.00	0.28	0.02	-0.03	-0.09	0.05	0.96	-0.12	-0.00	1.00	-0.22	-0.02
Customer Value	0.12	-0.13	0.11	0.17	0.42	0.40	0.92	0.28	-0.18	0.25	-0.41	-0.22	1.00	-0.29
Churn	-0.01	0.53	-0.03	-0.20	-0.30	-0.30	-0.22	-0.28	-0.01	-0.11	0.50	-0.02	-0.29	1.00

Fig.4: Correlation Heatmap of Numerical Features

The following significant trends surfaced from this analysis:

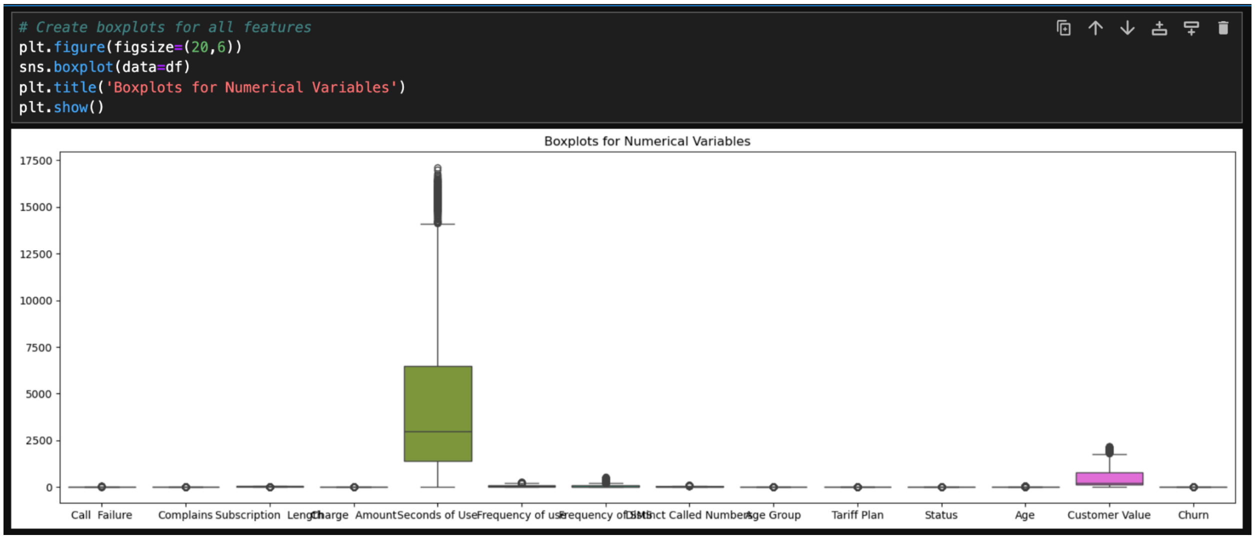
- Significantly positive associations were discovered between:
  - Longer call durations are inherently associated with higher usage frequency, according to the relationship between seconds of use and frequency of use (r = 0.95).
  - According to Customer Value, Frequency of Use (r = 0.92), and Seconds of Use (r = 0.92), more engaged consumers (measured by call volume) significantly increase revenue.
  - Age Group and Age (r = 0.96), as would be predicted given their conceptual overlap.

- Churn's relationships with the majority of factors were only modest to moderate:
    - Complaints had the largest positive association (r = 0.53), confirming the notion that unhappy consumers are more likely to go.
    - Churn and Seconds of Use (r = -0.30) and Churn and Frequency of Use (r = -0.30) showed a somewhat negative connection, suggesting that consumers who use the service less frequently are more likely to churn.
    - Also, there was a small negative association between churn and customer value (r = -0.29), indicating that high-value customers are somewhat less likely to go.
- Service interruptions might not have a direct effect on customer retention in this sample, as seen by the weak correlation between call failures and churn (r = -0.01).
- It's interesting to note that there was a moderately positive association between Status and Churn (r = 0.50), suggesting that a customer's account or service status may influence their propensity to leave.

All things considered, these results point to a collection of characteristics, including Complaints, Frequency of Use, Customer Value, and Status, that show more significant correlations with turnover and might be helpful predictors in the next modeling phase.

Box Plot Analysis

Boxplots were created for every characteristic in order to investigate the distribution of numerical variables in greater detail and spot any possible outliers (see Figure 5).



```python
# Create boxplots for all features
plt.figure(figsize=(20,6))
sns.boxplot(data=df)
plt.title('Boxplots for Numerical Variables')
plt.show()
```

Fig.5: Boxplots of Numerical Variables

Several insights emerged:

- There were several outliers on the higher end of the spectrum, and Seconds of Use and Customer Value showed considerable fluctuation. This is consistent with the previous correlation study that found these characteristics to be important predictors of user value and engagement.
- Although there were comparatively fewer severe outliers, there was still significant variation in the frequency of use of SMS.
- The majority of other parameters, including Tariff Plan, Call Failure, Complaints, and Subscription Length, showed low median values and were closely distributed, indicating category traits or consistent usage.
- In the box plot, Status, Age Group, and Churn were shown as binary or categorical variables with distinct groups and little to no dispersion.
- While Charge Amount and Distinct Called Numbers had several high-end outliers that suggested sporadic surges in user activity or invoicing, they also displayed moderate variability.

All things considered, the boxplots assisted in highlighting possible skewness and the existence of outliers, especially in variables pertaining to invoicing and consumption. Particularly if models are sensitive to scale or distributional assumptions, these features might need to be taken into account while preprocessing the data and developing the model.

Churn Rate by Complaint Status

A grouped bar chart (see Figure 6) was used to analyze the association between customer complaints and churn. It shows the churn status for customers with and without complaints (0 = retained, 1 = churned).

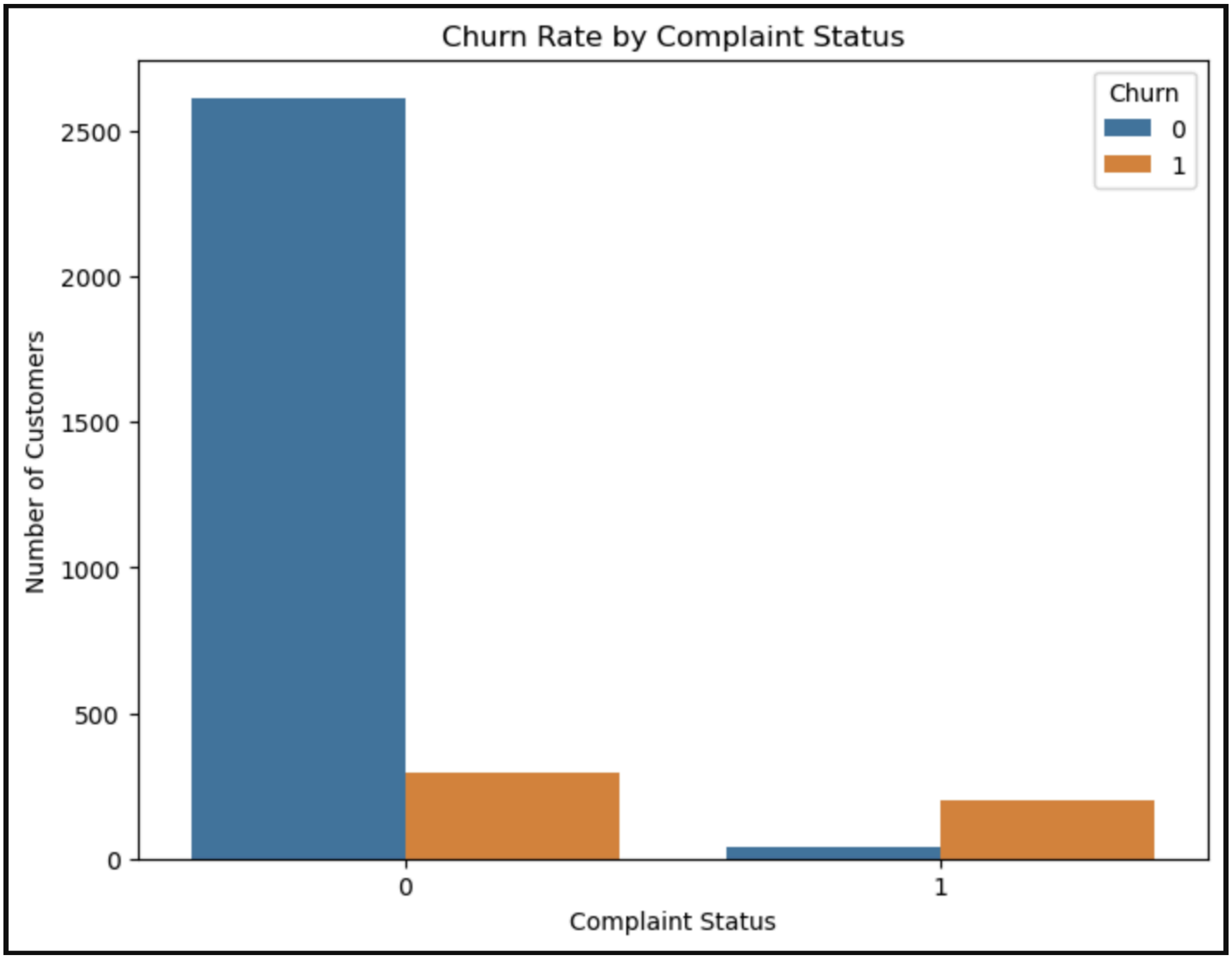


Fig.6: Churn Rate by Complaint Status

As can be seen, the turnover rate was significantly greater for consumers who filed complaints (Complaints = 1) than for those who did not (Complaints = 0). Specifically:

- Churn rate among customers who complained: 83%
- Churn rate among customers who did not complain: 10%

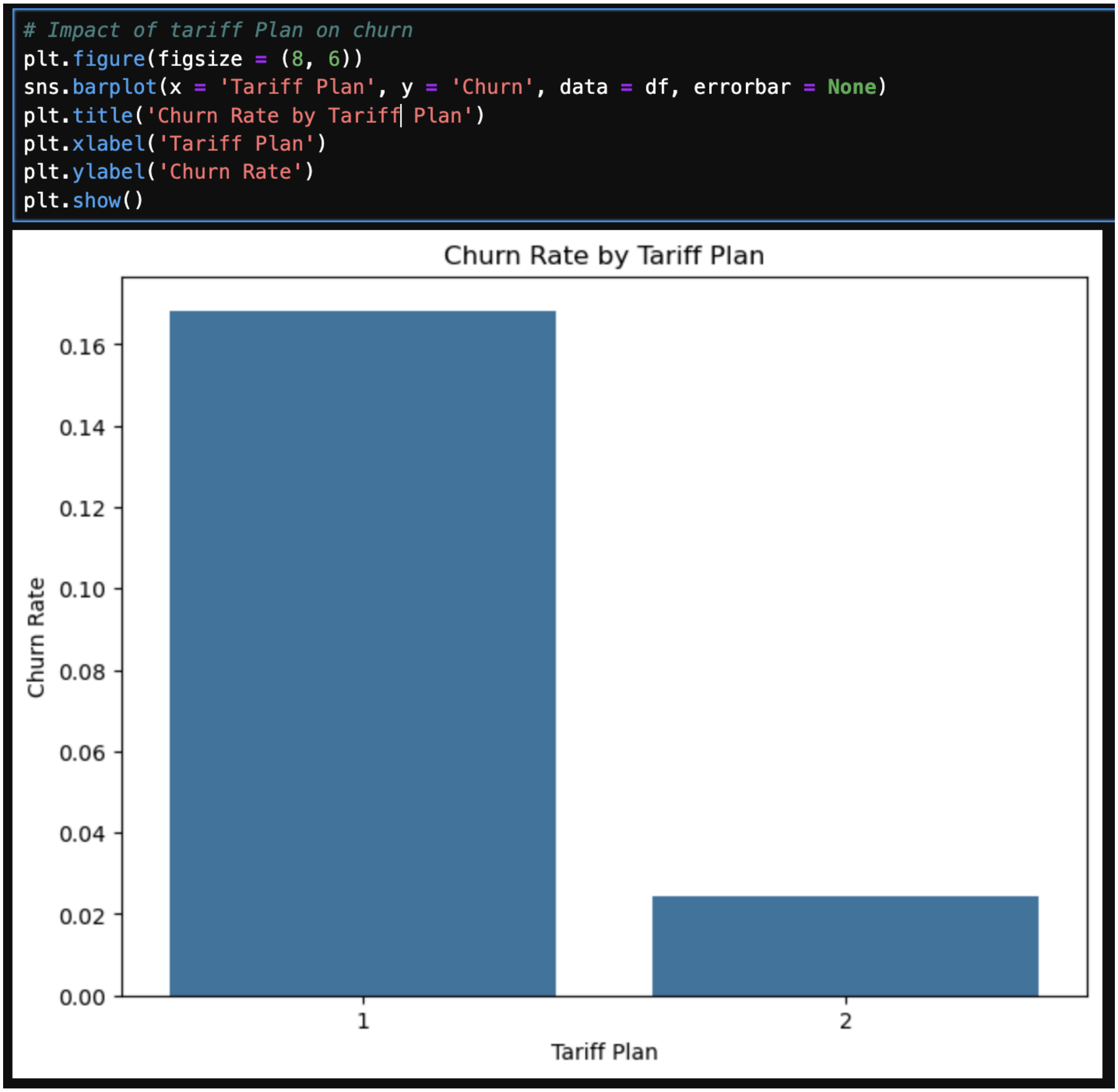
Using a contingency table, a chi-square test of independence was performed to determine if the observed difference was statistically significant. The results were as follows:

- Chi-square statistic: 886.21
- p-value: < 0.001

We reject the null hypothesis and come to the conclusion that there is a substantial correlation between attrition and customer complaints since the p-value is significantly lower than the traditional cutoff point of 0.05. This result supports the idea that complaint behavior is a reliable predictor of churn risk, since it is consistent with the correlation coefficient of 0.53 seen in the previous heatmap (Section 4.4).

These findings show the potential benefits of real-time complaint tracking for client retention tactics and prediction churn models.

Churn Rate by Tariff Plan



```
# Impact of tariff Plan on churn
plt.figure(figsize = (8, 6))
sns.barplot(x = 'Tariff Plan', y = 'Churn', data = df, errorbar = None)
plt.title('Churn Rate by Tariff Plan')
plt.xlabel('Tariff Plan')
plt.ylabel('Churn Rate')
plt.show()
```

Fig.7: Churn Rate by Tariff Plan

A bar plot showing the average churn rate for each tariff plan was created in order to determine whether customer attrition differs among them (Figure 7).

- Tariff Plan 1 exhibits a churn rate of approximately 17%.
- Tariff Plan 2 shows a markedly lower churn rate of around 2.5%.

Customers under Tariff Plan 1 are far more likely to churn than those under Tariff Plan 2, according to this striking disparity. The discrepancy could be explained by elements like each plan's cost, service restrictions, or level of customer satisfaction.

Whether this difference is statistically significant might be verified by additional statistical testing (such as a chi-square test). However, the visual data alone suggests that choosing a tariff plan is probably a component in customer attrition and should be taken into account in retention and predictive modeling methods.

4.1.7 Churn Rate by Charge Amount

The churn rate was computed for each distinct charge amount in order to examine the relationship between billing activity and customer churn.

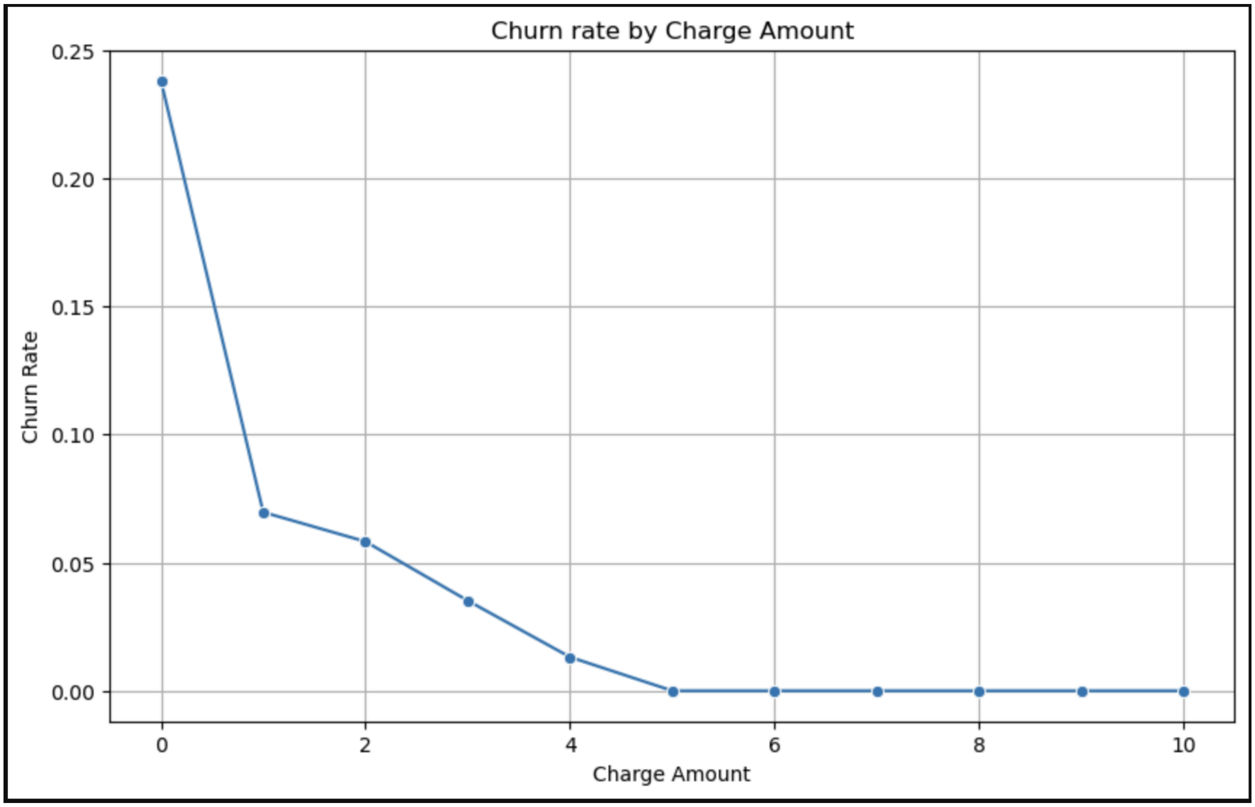


Fig.8: Churn Rate by Charge Amount

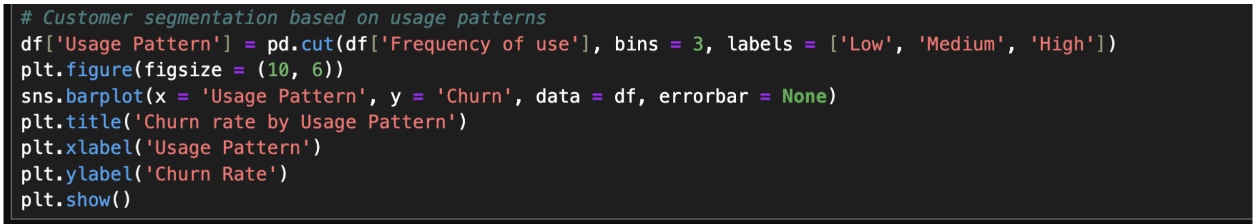
Charge amount and churn rate have a significant negative connection, as seen in the resulting image (Figure 8). In particular:

- The highest churn rate (around 24%) is seen among customers who have no charges.
- As charge quantities increase, the churn rate gradually decreases, falling below 10% at 1 and approaching 0% by 5 and above.
- After a while, the trend flattens out, indicating a threshold effect where retention is stabilized by sufficiently high engagement (as shown by the charge amount)..

Interpretation:

According to this tendency, consumers who are not being billed because they are not using the service or for other reasons are far more likely to leave. On the other hand, churn is far less likely to occur among users who actively utilize the service and pay expenditures. This realization highlights the significance of early identification and re-engagement tactics for clients that have little to no billing activity.

Churn by Usage Pattern



```
# Customer segmentation based on usage patterns
df['Usage Pattern'] = pd.cut(df['Frequency of use'], bins = 3, labels = ['Low', 'Medium', 'High'])
plt.figure(figsize = (10, 6))
sns.barplot(x = 'Usage Pattern', y = 'Churn', data = df, errorbar = None)
plt.title('Churn rate by Usage Pattern')
plt.xlabel('Usage Pattern')
plt.ylabel('Churn Rate')
plt.show()
```

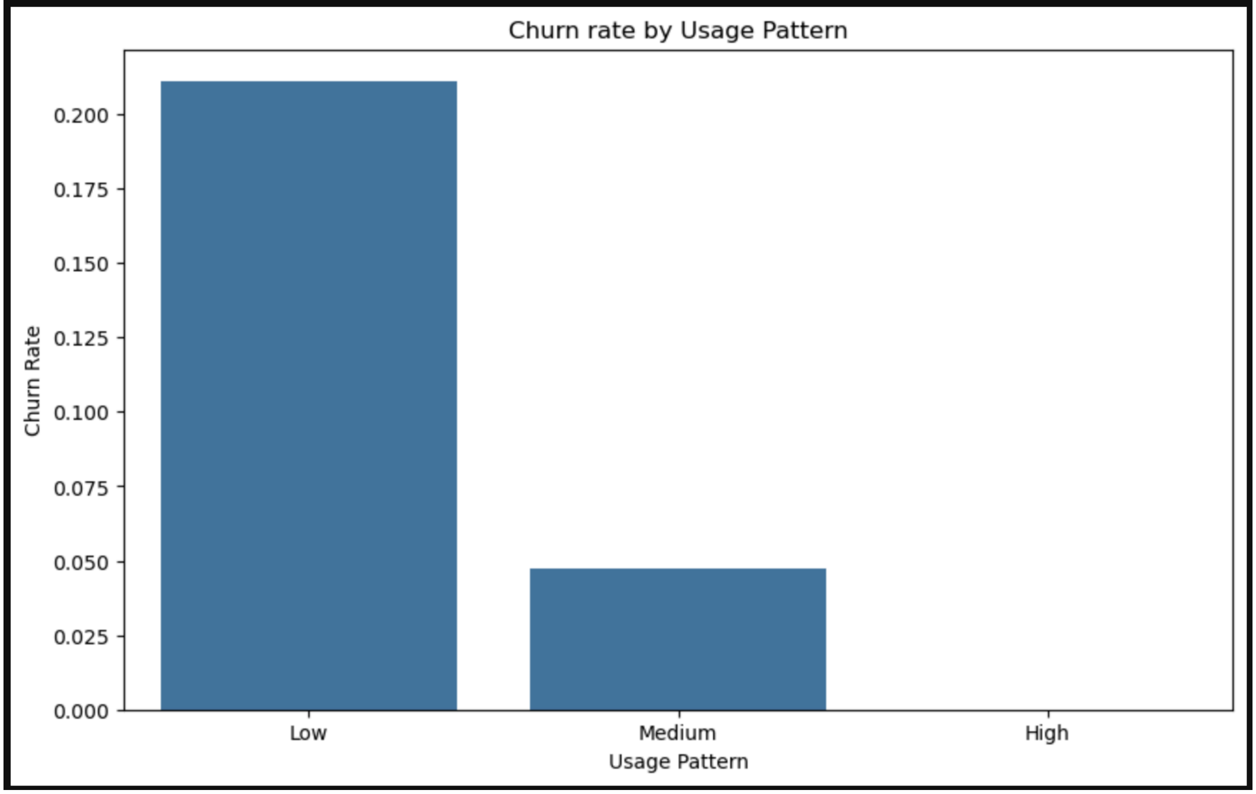
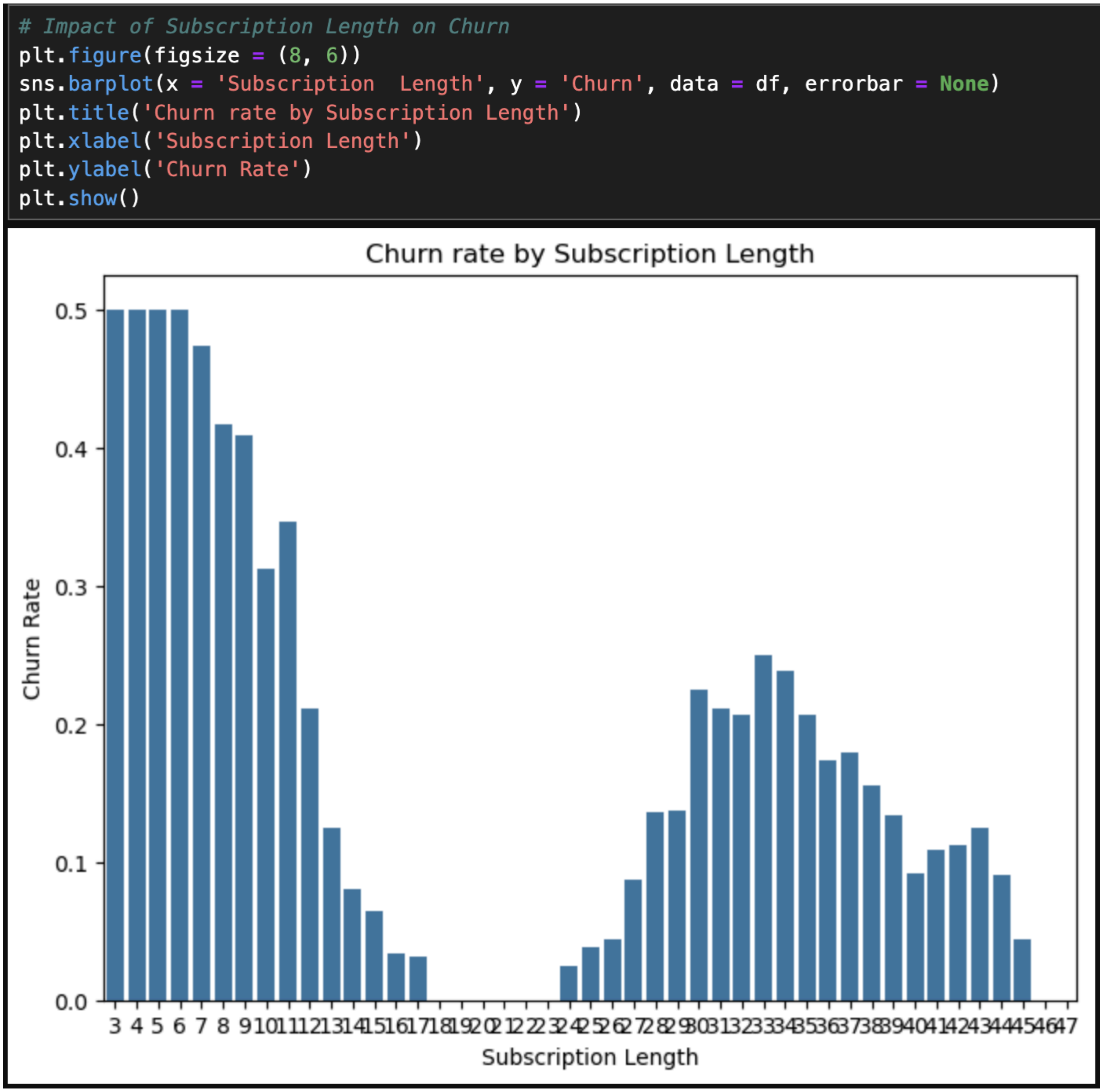


Fig.9: Churn Rate by Usage Pattern

User engagement and churn likelihood have a significant negative connection, according to the analysis of churn across usage patterns, which are divided into low, medium, and high frequency of use (Figure 9). Consumers in the medium and high usage categories had far lower turnover rates (nearly 0% in high usage) than those in the low usage patterns, who have churn rates of over 20%. This is consistent with earlier study on customer behavior, which found that infrequent service interaction frequently indicates disengagement or discontent and raises the likelihood of churn (Buckinx & Van den Poel, 2005). This emphasizes how crucial it is to encourage continuous user engagement and product consumption as a way to reduce attrition.

Subscription Length and Churn Dynamics

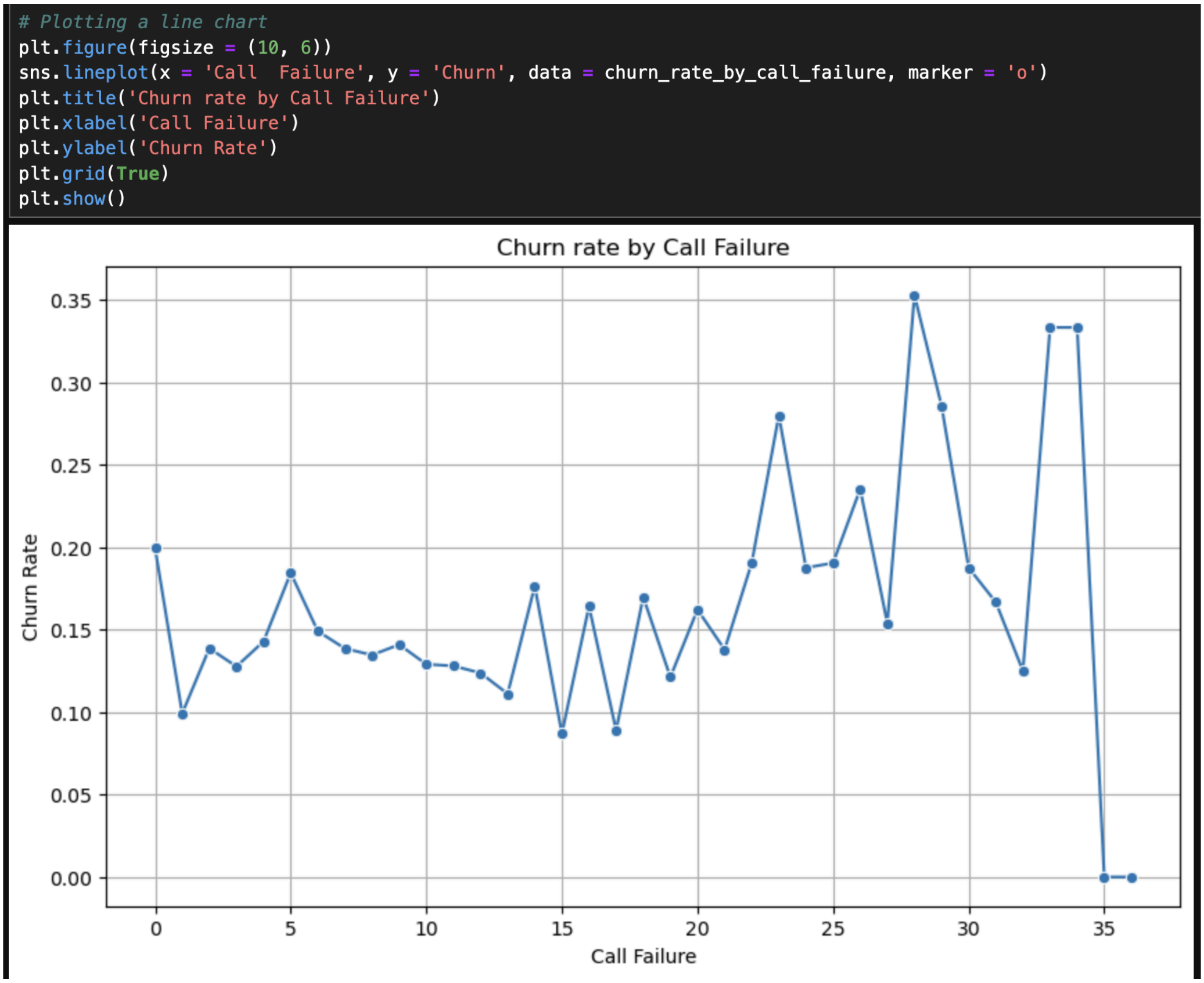


```
# Impact of Subscription Length on Churn
plt.figure(figsize = (8, 6))
sns.barplot(x = 'Subscription  Length', y = 'Churn', data = df, errorbar = None)
plt.title('Churn rate by Subscription Length')
plt.xlabel('Subscription Length')
plt.ylabel('Churn Rate')
plt.show()
```

Fig.10: Churn Rate by Complaint Status

Another important component of churn behavior is subscription length (Figure 10). Shorter-term (less than ten months) customers have very high churn rates, which peak at about 50% and progressively decrease as subscription duration rises. This lends credence to the tenure effect idea, which holds that loyalty and higher switching costs over time promote consumer commitment (Verhoef, 2003). It's interesting to note that there is a secondary high in the churn rate around the 30- to 35-month period, which might indicate that loyal customers may eventually approach saturation or become the focus of rival products. The necessity of focused retention efforts at various phases of the customer lifetime is shown by this nonlinear connection.

Churn Rate by Call Failure

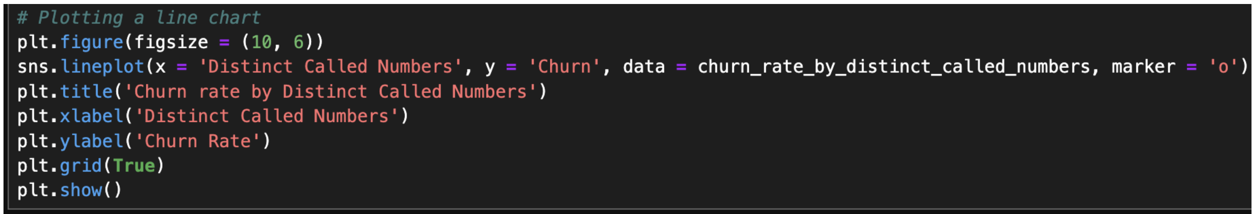


```
# Plotting a line chart
plt.figure(figsize = (10, 6))
sns.lineplot(x = 'Call  Failure', y = 'Churn', data = churn_rate_by_call_failure, marker = 'o')
plt.title('Churn rate by Call Failure')
plt.xlabel('Call Failure')
plt.ylabel('Churn Rate')
plt.grid(True)
plt.show()
```

Fig.11: Churn Rate by Call Failure

The impact of call failures on churn is examined in Figure 11. At lower call failure levels, attrition varies initially, but after 20 failures, a distinct increasing trend becomes apparent, with churn rates rising beyond 30%. Significantly, the increase in failures from 28 to 35 shows that technical problems are a direct cause of customer discontent, a conclusion supported by research on service quality (Ahn et al., 2006). Thus, churn may be directly decreased by lowering call failures through proactive network maintenance and customer assistance.

Churn Rate by Distinct Called Numbers



```python
# Plotting a line chart
plt.figure(figsize = (10, 6))
sns.lineplot(x = 'Distinct Called Numbers', y = 'Churn', data = churn_rate_by_distinct_called_numbers, marker = 'o')
plt.title('Churn rate by Distinct Called Numbers')
plt.xlabel('Distinct Called Numbers')
plt.ylabel('Churn Rate')
plt.grid(True)
plt.show()
```

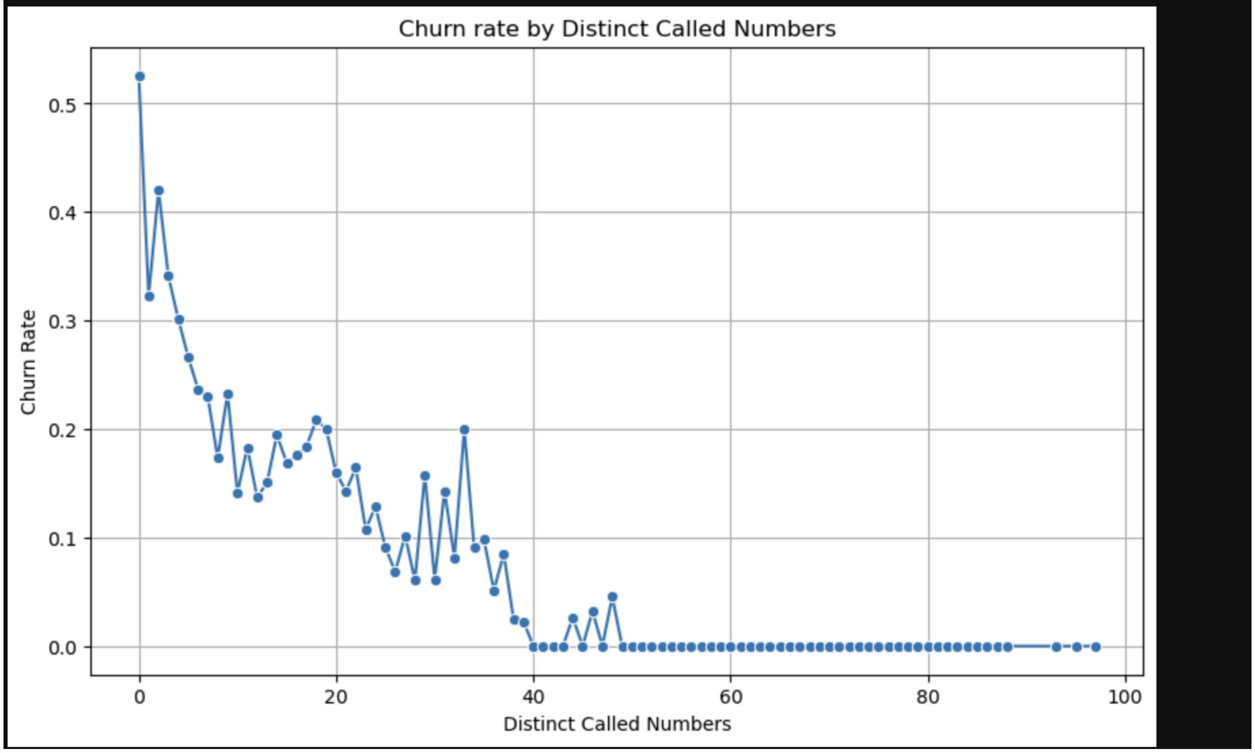
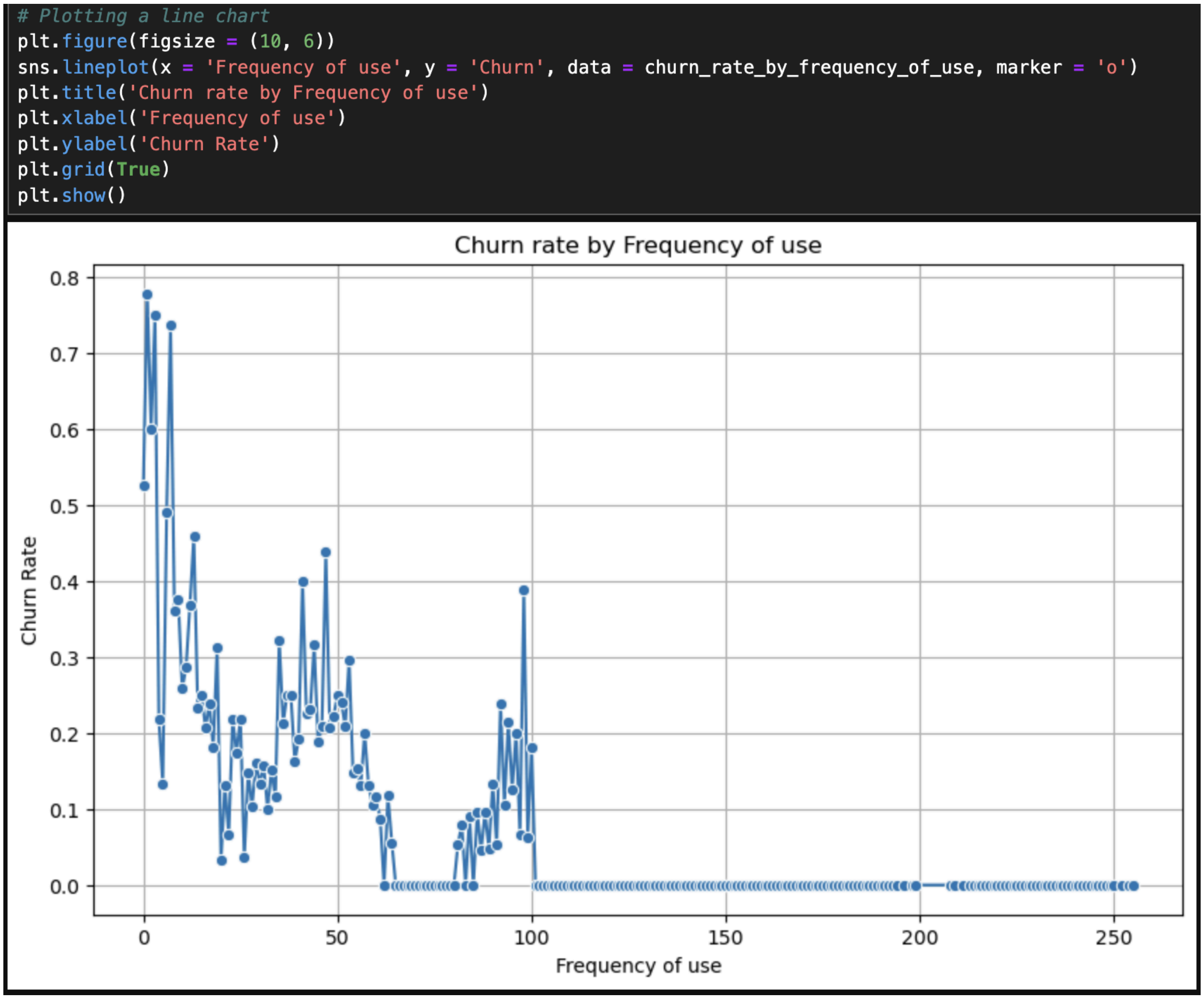


Fig.12: Churn Rate by Distinct Called Numbers

The churn rate and the number of unique phone numbers are clearly correlated negatively, as seen in Figure 12. The attrition rate was significantly higher for customers who phoned fewer unique numbers; customers who contacted just one to five separate lines had churn rates ranging from 30% to 50%. As the number of different contacts rises, this value progressively decreases, with churn stabilizing close to zero at 40 distinct numbers.

This pattern implies that churn propensity is negatively correlated with more involvement, as evidenced by a wider communication network. Stronger network externalities, which are produced by frequent and varied connections, lower the chance of service cessation, according to research in the literature currently under publication (Verhoef, 2003). Wider consumers are probably more integrated into the telecom service and, as a result, incur higher switching costs.

Churn Rate by Frequency of Use



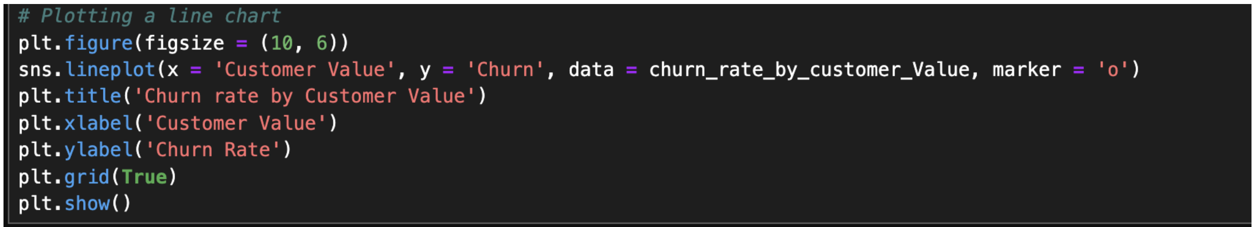
```
# Plotting a line chart
plt.figure(figsize = (10, 6))
sns.lineplot(x = 'Frequency of use', y = 'Churn', data = churn_rate_by_frequency_of_use, marker = 'o')
plt.title('Churn rate by Frequency of use')
plt.xlabel('Frequency of use')
plt.ylabel('Churn Rate')
plt.grid(True)
plt.show()
```

Fig.13: Churn Rate by Frequency of Use

A similar trend may be seen in the link between usage frequency and churn, as seen in Figure 13. Users with fewer phone encounters have far higher churn rates; for individuals with less than ten uses, these rates frequently surpass 70%. However, the churn rate drastically decreases as frequency rises. Although there are occasional oscillations between 40 and 100 encounters, the overall decreasing trend is consistent. Notably, for consumers with more than 100 interactions, turnover is almost nonexistent.

This is an example of behavioral inertia, which lowers the risk of churn by making recurrent service use routine (Bolton et al., 2000). More qualitative research is necessary since the volatility shown in the mid-frequency brackets may indicate situational users or those going through transitions (such as moving or traveling).

Churn Rate by Customer Value



```
# Plotting a line chart
plt.figure(figsize = (10, 6))
sns.lineplot(x = 'Customer Value', y = 'Churn', data = churn_rate_by_customer_Value, marker = 'o')
plt.title('Churn rate by Customer Value')
plt.xlabel('Customer Value')
plt.ylabel('Churn Rate')
plt.grid(True)
plt.show()
```

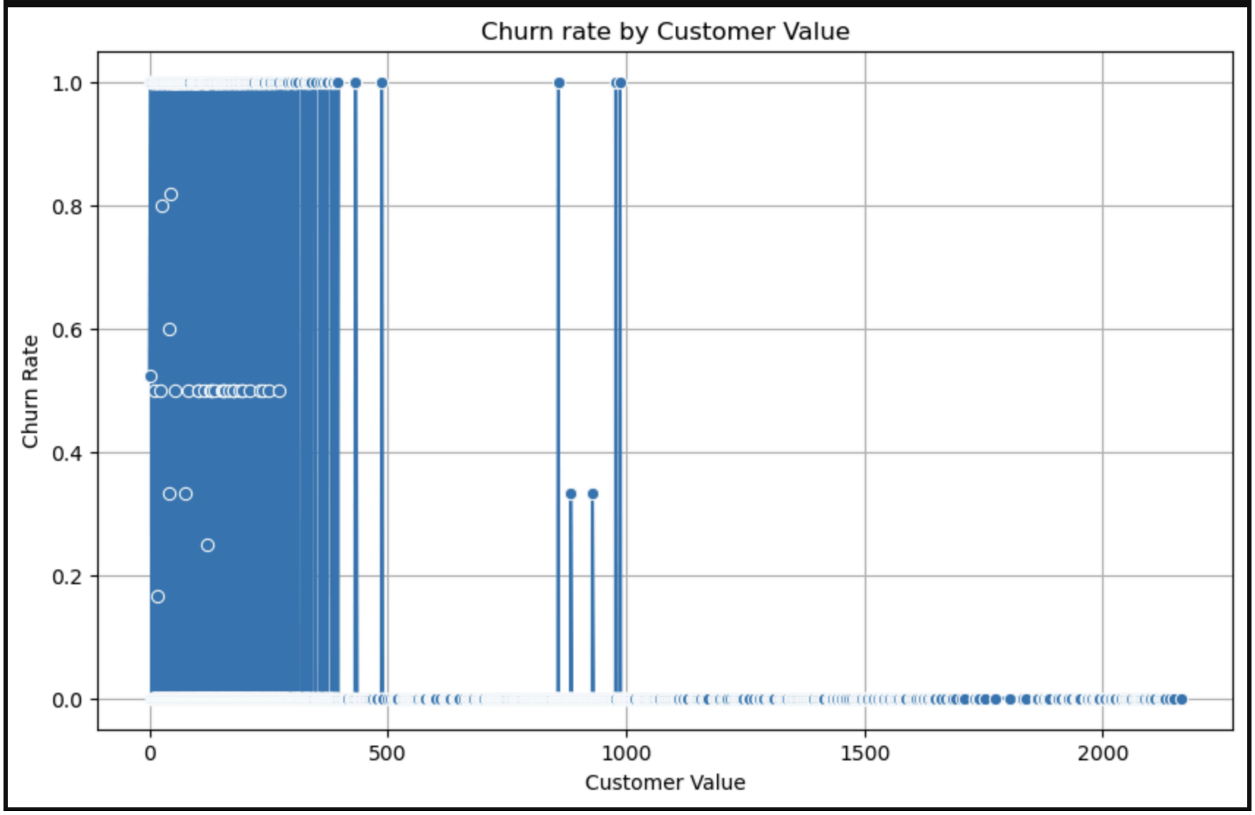


Fig.14: Churn Rate by Frequency of Use

Customer value, which is probably related to revenue or ARPU, seems to be a good indicator of turnover, as seen in Figure 14. Low-monetary-value customers have far higher turnover rates, which often fall between 50 and 60 percent. High-value clients (more than 500 units) exhibit extremely low churn rates, which almost approach 0%.

The significance of customer lifetime value (CLV) as a predictor for retention strategy is shown by this classification (Reinartz and Kumar, 2002). With the use of behavioral segmentation and predictive modeling, telecom providers may be able to proactively retain high-value clients by detecting and pursuing those who are at risk early on.

EDA Conclusion

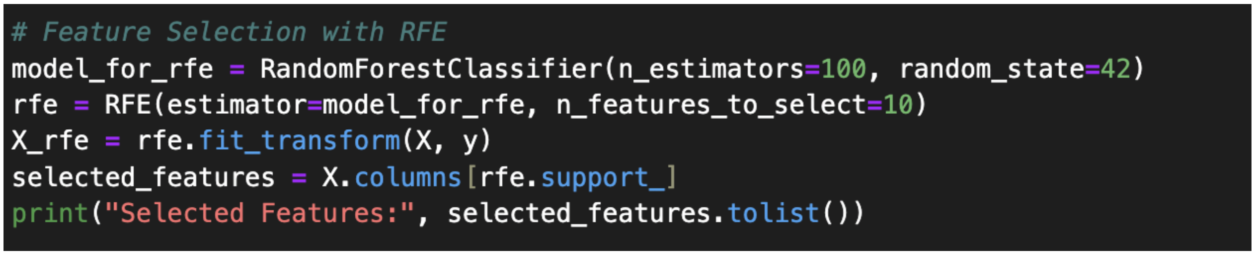
Understanding the dataset's structure, distribution, and possible causes of customer attrition was made possible by the exploratory data analysis. While summary data showed significant diversity across important numerical variables including call failures, use frequency, and customer value, preliminary tests showed a somewhat unbalanced class distribution that required subsequent resampling procedures. These insights were further enhanced by visualizations, which revealed non-linear correlations and patterns that would not be immediately clear from numerical summaries alone. According to existing churn research (e.g., Buckinx & Van den Poel, 2005; Ahn et al., 2006), churn was disproportionately greater among customers with shorter subscription durations, frequent call failures, and lower engagement levels. Furthermore, significant segmentation cues with possible predictive value were brought to light by category breakdowns of customer status and complaints.

Overall, the EDA verified the inclusion of behavioral and service quality indicators as important churn factors and provided guidance for feature selection and modeling technique. The results highlight how crucial multidimensional feature analysis that takes into account tenure, consumption trends, and service dependability is for precisely identifying clients who are at danger. The model creation and assessment procedures in the upcoming study phases will be directly influenced by these discoveries.

Feature Selection using Recursive Feature Elimination (RFE)

A crucial first step in creating predictive models is selecting features effectively. This is especially true for churn prediction, where high-dimensional data might mask trends and raise the possibility of overfitting (Guyon and Elisseeff, 2003). Recursive Feature Elimination (RFE) was used using a Random Forest classifier as the basis estimator to determine the most significant predictors of customer turnover.

Until the ideal subset of features is obtained, RFE recursively eliminates the least significant features depending on the model's performance. When feature interactions are complicated and non-linear, as is frequently the case in customer churn scenarios, this model-driven method is very helpful (Saeys, Inza, and Larrañaga, 2007). Because of its robustness, capacity to capture non-linear correlations, and integrated feature significance metrics, the Random Forest model was selected as the estimator for RFE. This makes it an excellent choice for heterogeneous tabular data (Breiman, 2001).



```
# Feature Selection with RFE
model_for_rfe = RandomForestClassifier(n_estimators=100, random_state=42)
rfe = RFE(estimator=model_for_rfe, n_features_to_select=10)
X_rfe = rfe.fit_transform(X, y)
selected_features = X.columns[rfe.support_]
print("Selected Features:", selected_features.tolist())
```

Out of the 13 available predictors, the RFE process was set up to choose the top 10 most important attributes. The chosen attributes were:

- Call Failure
- Complains
- Subscription Length
- Seconds of Use
- Frequency of Use
- Frequency of SMS
- Distinct Called Numbers
- Status
- Age
- Customer Value

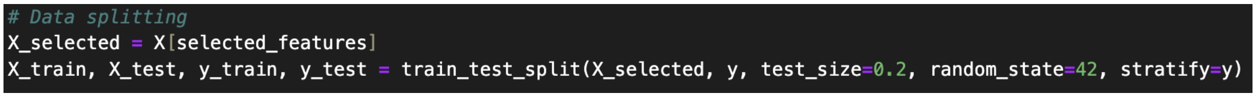
These characteristics were thought to be the most useful in differentiating between consumers who have left and those who have remained. For example, churn in telecommunications has long been linked to service quality, which is directly related to aspects like complaints and call failure (Ahn, Han, and Lee, 2006). Similar to this, usage-related metrics such as Customer Value, Frequency of Use, and Seconds of Use provide information on monetary value and engagement, two aspects that have a big impact on churn behavior (Verbeke et al., 2012).

Notably, account and demographic attributes like Age, Subscription Length, and Status (e.g., active/inactive) were also kept, suggesting that churn risk is influenced by customer profiles and lifecycle stages. In addition to increasing computing efficiency and model interpretability, the elimination of less informative variables is critical for real-world deployment, where responsiveness and explainability are critical.

By concentrating model training on the most pertinent aspects of the data and eliminating noise and redundancy, the feature selection stage improved the learning process.

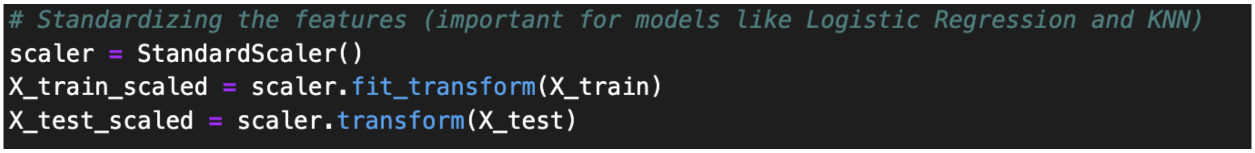
Data Splitting, Standardization, and Resampling

The dataset was ready for model training and assessment after the top 10 features were chosen using Recursive Feature Elimination (RFE). The dataset was divided into training and test sets to make model training easier and evaluate generalization performance. To maintain the percentage of churned vs non-churned clients in both categories, a stratified sample technique was used. Due to class imbalance, this is especially crucial in churn prediction tasks, where the minority class, that is churned customers, is frequently underrepresented (He and Garcia, 2009). As is customary in supervised learning studies, an 80:20 train-test split was employed (Kuhn and Johnson, 2013).



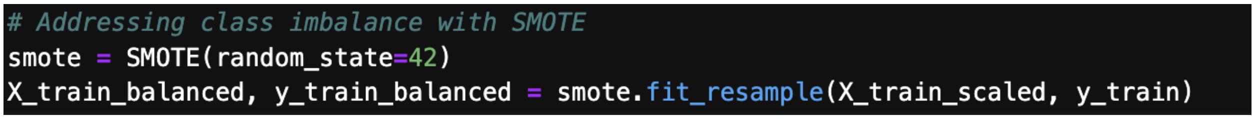
```python
# Data splitting
X_selected = X[selected_features]
X_train, X_test, y_train, y_test = train_test_split(X_selected, y, test_size=0.2, random_state=42, stratify=y)
```

The sklearn.preprocessing module's StandardScaler was used to normalize the features prior to model fitting. Each feature is transformed to have a standard deviation of one and a mean of zero. Since linear models like Logistic Regression and distance-based algorithms like k-Nearest Neighbors are sensitive to the size of the input variables, standardization is an essential step (Han, Kamber, and Pei, 2012). Features with higher magnitudes may disproportionately affect the model without scaling, resulting in less-than-ideal performance.



```python
# Standardizing the features (important for models like Logistic Regression and KNN)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

An oversampling strategy was used to balance the training data because of the dataset's class imbalance. For this, the Synthetic Minority Oversampling Technique (SMOTE) was used. By interpolating between existing samples, SMOTE creates synthetic instances of the minority class (churned consumers), enhancing the classifier's capacity to learn decision limits that are impartial toward the majority class (Chawla et al., 2002).



```
# Addressing class imbalance with SMOTE
smote = SMOTE(random_state=42)
X_train_balanced, y_train_balanced = smote.fit_resample(X_train_scaled, y_train)
```

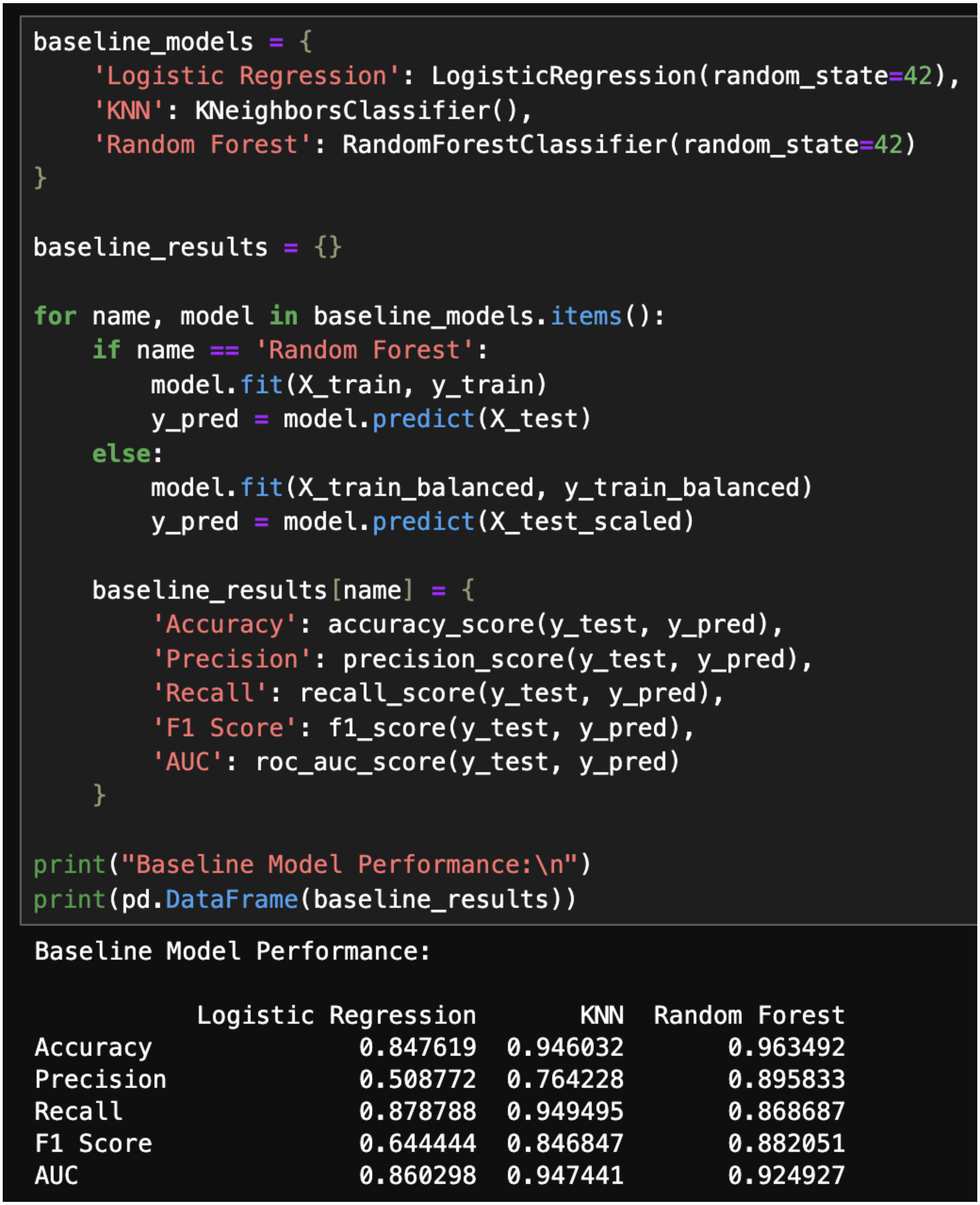
When the evaluation metrics of interest include recall, F1-score, and ROC-AUC metrics that are sensitive to class imbalance, this resampling step makes sure that the learning algorithms are exposed to a balanced representation of both classes during training (Fernández et al., 2018).

Model Performance

Three classification algorithms (Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression) were used to evaluate how well various machine learning approaches predicted customer attrition. To create a fair baseline comparison and ascertain the additional benefit of model improvement, the model was evaluated both before and after hyperparameter tweaking.

Baseline Performance

In order to assure repeatability, random_state=42 was added to the baseline models, which were otherwise trained with default hyperparameters. K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression were the three models that were assessed. Preprocessing stages were implemented selectively since various algorithms had different assumptions and sensitivities. In particular, because tree-based models are insensitive to feature scaling, the Random Forest classifier was trained on the original, unscaled feature set, whereas Logistic Regression and KNN were trained on data that had been both standardized and balanced using the Synthetic Minority Oversampling Technique (SMOTE) (Lemaitre et al., 2017).



```
baseline_models = {
    'Logistic Regression': LogisticRegression(random_state=42),
    'KNN': KNeighborsClassifier(),
    'Random Forest': RandomForestClassifier(random_state=42)
}

baseline_results = {}

for name, model in baseline_models.items():
    if name == 'Random Forest':
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
    else:
        model.fit(X_train_balanced, y_train_balanced)
        y_pred = model.predict(X_test_scaled)

    baseline_results[name] = {
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred),
        'Recall': recall_score(y_test, y_pred),
        'F1 Score': f1_score(y_test, y_pred),
        'AUC': roc_auc_score(y_test, y_pred)
    }

print("Baseline Model Performance:\n")
print(pd.DataFrame(baseline_results))
```

```
Baseline Model Performance:

           Logistic Regression       KNN  Random Forest
Accuracy              0.847619  0.946032       0.963492
Precision             0.508772  0.764228       0.895833
Recall                0.878788  0.949495       0.868687
F1 Score              0.644444  0.846847       0.882051
AUC                   0.860298  0.947441       0.924927
```

Metric	Logistic Regression	KNN	Random Forest
Accuracy	0.8476	0.9460	0.9635
Precision	0.5088	0.7642	0.8958
Recall	0.8788	0.9495	0.8687
F1 Score	0.6444	0.8468	0.8821
AUC	0.8603	0.9474	0.9249

Table 1: Baseline Model Performance

The Random Forest model performed the best overall out of the three. With a high precision of 0.8958, an F1 score of 0.8821, and an accuracy of 96.35%, this model struck the optimum balance between preventing false positives and accurately forecasting customer attrition. The ensemble aspect of Random Forest, which minimizes overfitting and captures intricate non-linear interactions between features, is responsible for this impressive performance (Breiman, 2001). Furthermore, the model maintains a healthy trade-off between recall and precision, as seen by its better F1 score, a crucial parameter for unbalanced classification issues.

KNN produced more false positives than Random Forest, as seen by its significantly lower accuracy (76.42%) while having the best recall (94.95%). High recall indicates that the majority of churners were accurately recognized in a churn prediction scenario, which is crucial from the perspective of customer retention. False positives, or the mistaken classification of loyal customers as churners, can result in needless intervention attempts and higher operating expenses with no real gain. As a result, even though KNN could be useful for early warning systems, its practical applicability might be limited unless it is further refined.

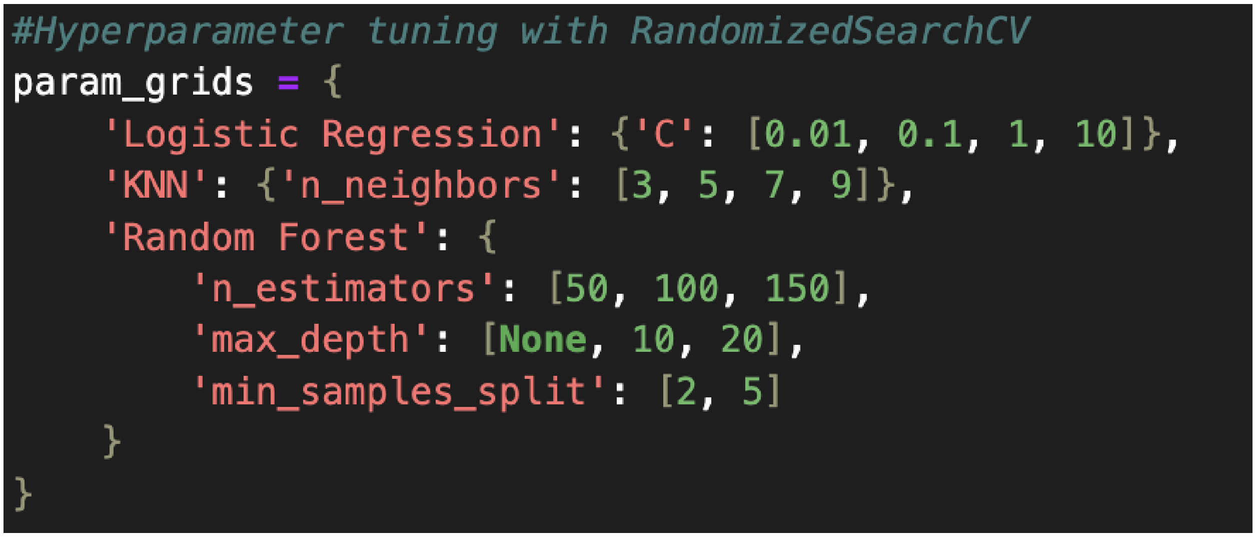
Despite being an often used baseline because of its ease of use and interpretability, logistic regression scored poorly in comparison to the other models. It had the lowest precision score (50.88%), indicating that over half of the clients it had anticipated leaving were kept. The resulting F1 score of 0.6444 confirms its weak discriminative power in this situation, despite its relatively strong recall (87.88%). This outcome is not wholly surprising because logistic regression makes the assumption that factors are independent and have linear connections, which may not be true in real-world churn datasets with complicated temporal and behavioral patterns (Burez & Van den Poel, 2009).

These findings were further supported by the AUC (Area Under the Curve) values. KNN surpassed Logistic Regression (0.8603) by achieving the greatest AUC (0.9474), followed closely by Random Forest (0.9249). For targeted retention tactics that give priority to high-risk clients, the high AUC values show that both KNN and Random Forest are good at rating churn probability.

In conclusion, Random Forest proved to be the most well-rounded and successful baseline classifier for churn prediction, even though all three models had potential. It is an excellent candidate for more tweaking and implementation because to its high accuracy, competitive recall, and robustness to irrelevant or noisy characteristics. Even at the expense of occasional false positives, KNN's comparatively high recall performance could support its application in situations where reducing the chance of missing possible churners is the main business objective. Without substantial feature transformation or dimensionality reduction, logistic regression seems less appropriate to capture the intricate feature interactions present in customer turnover behavior, despite its value for interpretability and preliminary modeling.

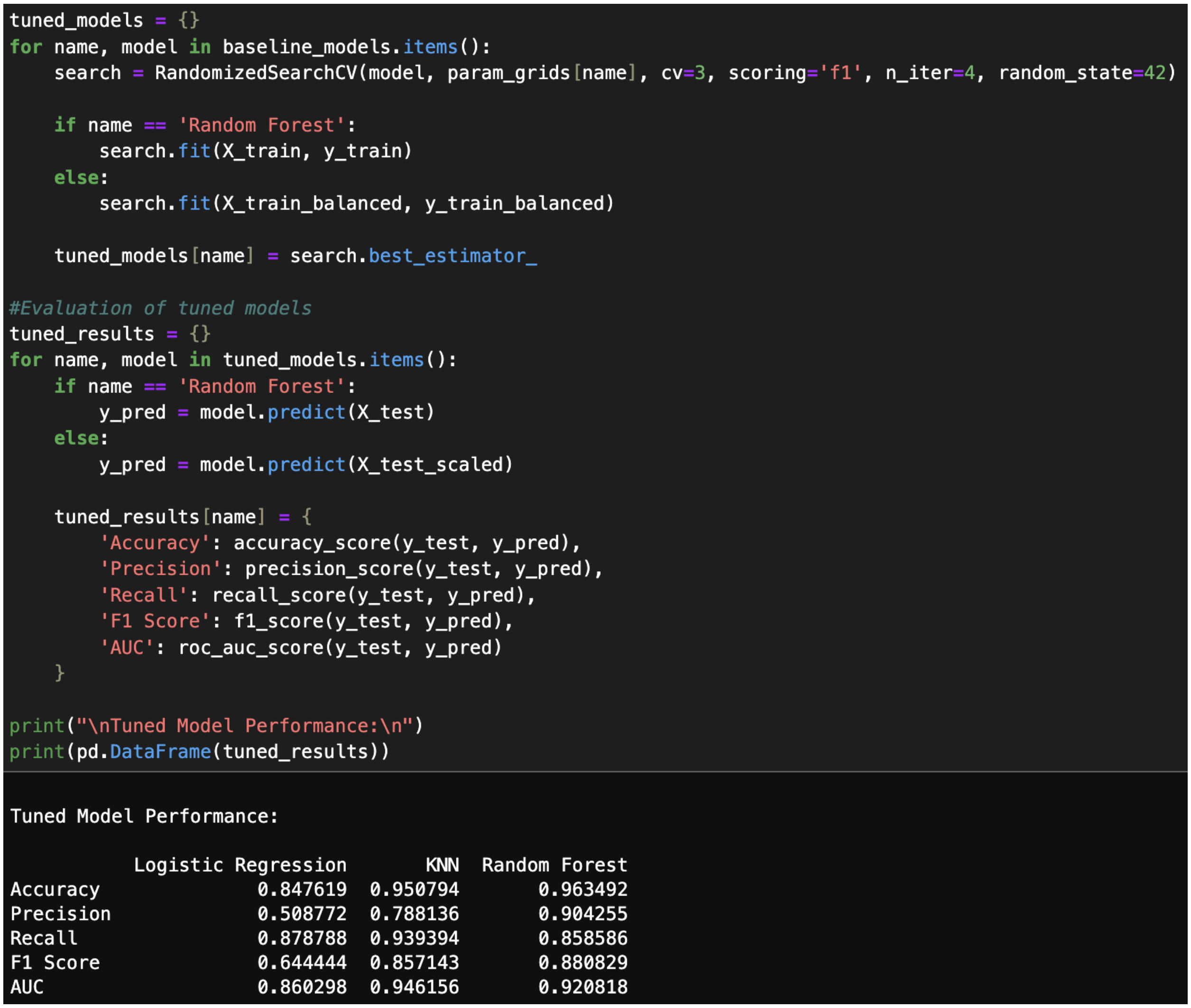
Tuned Model Performance

Hyperparameter adjustment was done to improve prediction performance after baseline models were evaluated. For this, the scikit-learn RandomizedSearchCV technique was employed.



```
#Hyperparameter tuning with RandomizedSearchCV
param_grids = {
    'Logistic Regression': {'C': [0.01, 0.1, 1, 10]},
    'KNN': {'n_neighbors': [3, 5, 7, 9]},
    'Random Forest': {
        'n_estimators': [50, 100, 150],
        'max_depth': [None, 10, 20],
        'min_samples_split': [2, 5]
    }
}
```

This method uses k-fold cross-validation, in this example 3-fold, and F1 score as the optimization criteria to assess model performance after conducting a randomized search across predetermined hyperparameter distributions. RandomizedSearchCV provides a more computationally efficient approach than an exhaustive grid search, but it is still capable of finding performant parameter combinations, especially when time or computer resources are limited (Bergstra and Bengio, 2012).



```python
tuned_models = {}
for name, model in baseline_models.items():
    search = RandomizedSearchCV(model, param_grids[name], cv=3, scoring='f1', n_iter=4, random_state=42)

    if name == 'Random Forest':
        search.fit(X_train, y_train)
    else:
        search.fit(X_train_balanced, y_train_balanced)

    tuned_models[name] = search.best_estimator_

#Evaluation of tuned models
tuned_results = {}
for name, model in tuned_models.items():
    if name == 'Random Forest':
        y_pred = model.predict(X_test)
    else:
        y_pred = model.predict(X_test_scaled)

    tuned_results[name] = {
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred),
        'Recall': recall_score(y_test, y_pred),
        'F1 Score': f1_score(y_test, y_pred),
        'AUC': roc_auc_score(y_test, y_pred)
    }

print("\nTuned Model Performance:\n")
print(pd.DataFrame(tuned_results))
```

```
Tuned Model Performance:

           Logistic Regression       KNN  Random Forest
Accuracy              0.847619  0.950794       0.963492
Precision             0.508772  0.788136       0.904255
Recall                0.878788  0.939394       0.858586
F1 Score              0.644444  0.857143       0.880829
AUC                   0.860298  0.946156       0.920818
```

Metric	Logistic Regression	KNN	Random Forest
Accuracy	0.8476	0.9508	0.9635
Precision	0.5088	0.7881	0.9043
Recall	0.8788	0.9394	0.8586

F1 Score	0.6444	0.8571	0.8808
AUC	0.8603	0.9462	0.9208

Table 2: Tuned Model Performance

The K-Nearest Neighbors (KNN) model showed the most improvement. The model's F1 score increased from its baseline of 0.8468 to 0.8571 after tweaking. The increased accuracy (from 0.7642 to 0.7881) was the main driver of this improvement, indicating that the modified KNN model produced fewer false positive predictions, which is crucial in minimizing needless retention interventions (Burez & Van den Poel, 2009). KNN's comparatively good performance here suggests that neighborhood-like patterns in the feature space, including comparable use or complaint behaviors among small clusters of customers, may have an impact on customer turnover despite its susceptibility to data scaling and local patterns.

Tuning produced slight gains for the Random Forest classifier. While recall somewhat declined (from 0.8687 to 0.8586), precision increased from 0.8958 to 0.9043. The model maintained its balanced classification performance after tweaking, as evidenced by the total F1 score, which was almost constant between 0.8821 and 0.8808. This finding supports the idea that tree-based ensemble approaches are naturally resilient and frequently perform well even with default hyperparameters, especially considering the baseline Random Forest model's already good performance (Breiman, 2001; Hastie, Tibshirani, and Friedman, 2009). Perhaps because the model currently captures the most important decision limits, the little adjustments imply declining results from more tuning.

On the other hand, after hyperparameter adjustment, Logistic Regression did not improve on any of the assessment measures. Two main reasons for this could be that class balancing (using SMOTE) and feature standardization may have already maximized the model's representational power in the baseline phase, and second, logistic regression is a linear model that might not adequately capture intricate non-linear interactions among features without additional transformations or interaction terms. This implies that logistic regression is probably inadequate for the complex job of churn prediction in this dataset, even though it is still a suitable baseline because of its interpretability and low computing cost.

In addition, all models demonstrated strong discriminative performance when comparing Area Under the Curve (AUC) values; Random Forest achieved 0.9208, KNN achieved 0.9462, and Logistic Regression achieved 0.8603. The models' ability to differentiate between churners and non-churners across a variety of categorization thresholds is validated by these values.

Comparative Insights

The Random Forest model continuously produced the greatest overall performance in both baseline and adjusted assessments, demonstrating its applicability for churn prediction tasks. In accordance with findings from Cutler et al. (2007) and Breiman (2001), its high accuracy and F1 score were a result of its capacity to handle heterogeneous data types and predict complicated, non-linear connections.

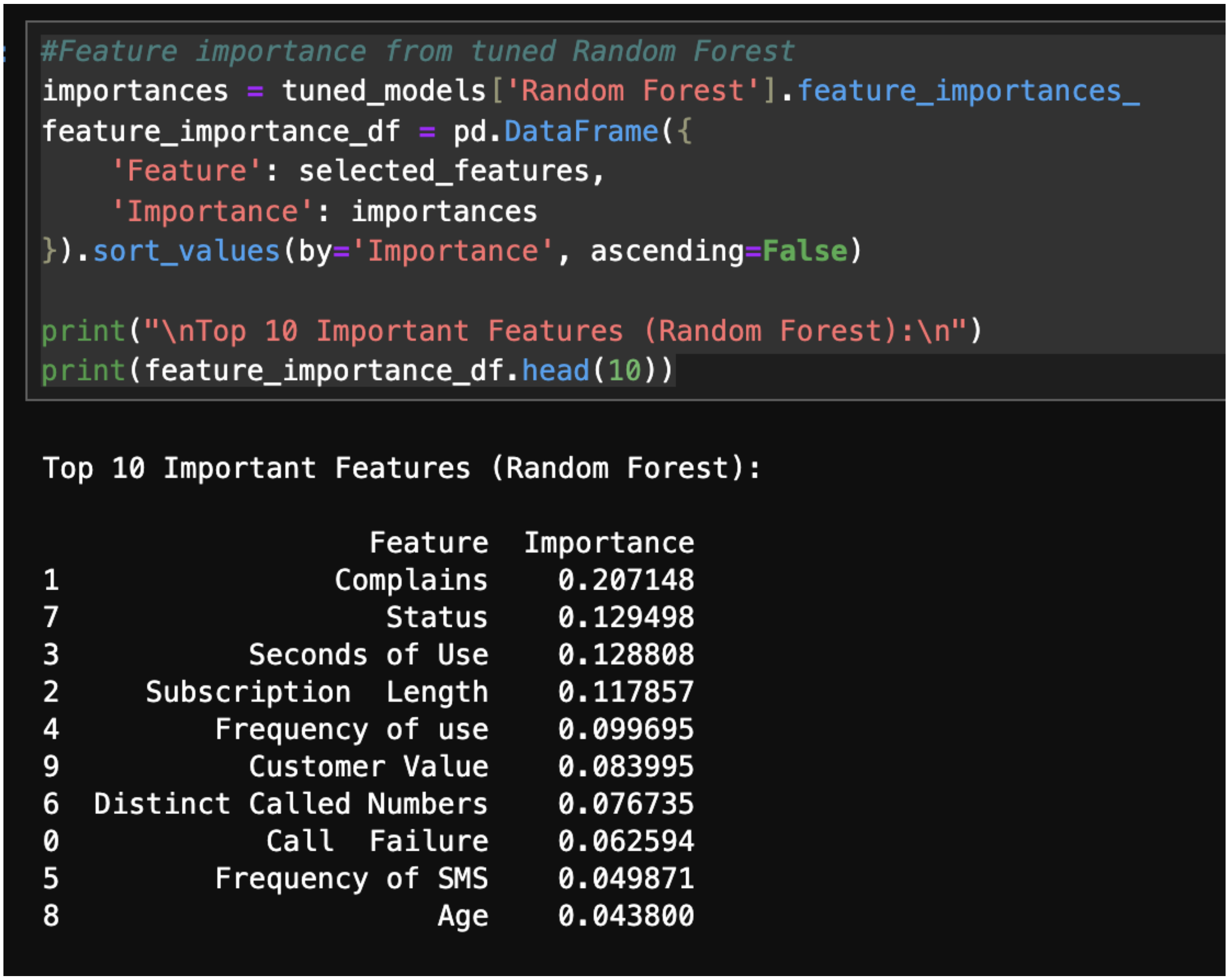
KNN also performed competitively, especially in recall. However, rigorous preprocessing, including SMOTE and standardization, was necessary because of its sensitivity to feature scaling and class imbalance. KNN demonstrated its promise when backed by robust data preparation, as seen by its enhanced precision when tweaked.

Despite being straightforward and easy to understand, logistic regression fared poorly in terms of accuracy and F1 score. In accordance with previous research, its linear assumptions probably hampered its ability to represent churn behavior (Verbeke et al., 2012). Even so, it is still useful in situations when quickness and interpretability are more important than accuracy.

Overall, the findings highlight the need of ensemble models for reliable churn prediction, especially Random Forest. They also stress how crucial it is to compare various algorithms since they each have unique advantages based on the demands of the company and the properties of the data.

Feature Importance Analysis (Random Forest)

The factors most important in forecasting customer turnover are highlighted by the feature importance analysis, which is obtained from the modified Random Forest model using Gini impurity-based measurements. The most important predictor was "complaints," indicating that customer discontent is a powerful turnover indicator. "Status," "Seconds of Use," and "Subscription Length" came next, suggesting that service tenure and engagement length are also important factors in customer retention. "Frequency of Use," "Customer Value," and "Distinct Called Numbers," which represented behavioral usage patterns, were further significant aspects.



```
#Feature importance from tuned Random Forest
importances = tuned_models['Random Forest'].feature_importances_
feature_importance_df = pd.DataFrame({
    'Feature': selected_features,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)

print("\nTop 10 Important Features (Random Forest):\n")
print(feature_importance_df.head(10))
```

```
Top 10 Important Features (Random Forest):

                   Feature  Importance
1                Complains    0.207148
7                   Status    0.129498
3           Seconds of Use    0.128808
2     Subscription  Length    0.117857
4         Frequency of use    0.099695
9           Customer Value    0.083995
6  Distinct Called Numbers    0.076735
0            Call  Failure    0.062594
5         Frequency of SMS    0.049871
8                      Age    0.043800
```

Though to a lesser extent, less significant characteristics like "Age" and "SMS Frequency" nevertheless influence the overall model. These findings offer practical advice for focused churn reduction tactics, such giving complaints or decreasing usage patterns priority.