TELECOM CHURN

PRESENTATION

The telecom industry in the Indian and Southeast Asian markets faces a significant challenge of customer churn, with an annual churn rate of 15-25%. Retaining high-value customers, who contribute 80% of the revenue, is crucial for telecom operators. The objective of this project is to analyze customer-level data from a leading telecom firm and build predictive models to identify high-value customers at risk of churn. By accurately predicting churn and understanding the main indicators, we aim to provide actionable insights and recommendations for effectiv customer retention strategies, thereby reducing revenue leakage and improving business performance.

Problem Statement

The telecom industry faces a significant challenge of customer churn, with an average annual churn rate of 15-25%. Retaining high-profitable customers has become the top priority for telecom operators, as acquiring new customers is considerably more expensive. To address this issue, we aim to predict customers at high risk of churn and identify the main indicators of churn using customer-level data from a leading telecom firm in the Indian and Southeast Asian market.

Defining churn is particularly complex in the prepaid model, where customers can simply stop using services without notice. In this project, we define churn as customers who have not utilized any services (incoming or outgoing calls, internet, etc.) over a specified period. We will focus on usage-based churn as it is more applicable to prepaid customers, who constitute the majority in this market.

Moreover, high-value customers, who contribute approximately 80% of the revenue, will be the primary focus of churn prediction. By reducing churn among high-value customers, we can significantly mitigate revenue leakage. Therefore, we will define high-value customers based on a specific metric and predict churn exclusively for this segment.

Our objective is to build predictive models that effectively identify customers at high risk of churn, enabling proactive retention strategies. Additionally, we aim to uncover key indicators of churn, providing insights into why customers choose to switch to other networks. Through this analysis, we will offer recommendations on managing customer churn and enhancing customer retention strategies.

By leveraging advanced analytics and machine learning techniques, we anticipate substantial improvements in churn prediction accuracy and actionable insights that can positively impact the telecom company's bottom line.

Objective

The objective of this casestudy is to predict churn among high-value customers in the telecom industry, specifically focusing on the Indian and Southeast Asian markets. By leveraging customer-level data spanning four consecutive months, we aim to build predictive models that accurately identify customers at high risk of churn. Additionally, we seek to identify the main indicators of churn, providing valuable insights into the factors that drive customer attrition. The ultimate goal is to enable the telecom company to proactively intervene and implement targeted retention strategies, thereby reducing churn, retaining high-value customers, and minimizing revenue loss. Through this project, we aim to contribute to the improvement of customer retention practices and the overall business performance of the telecom operator.

Methodology

01 Data Exploration

02 Data Understanding

03 Data Cleaning

04 Feature Engineering

EDA 05

Model Selection 06

Model Evaluation 07

Insights & Recom. 08

09 Conclusion

Data Exploration

The dataset contains customer-level information spanning four consecutive months (June, July, August, and September) in the telecom industry. In churn prediction, we assume that there are three phases of customer lifecycle : The ‘good’ phase [Month 6 & 7] The ‘action’ phase [Month 8] The ‘churn’ phase [Month 9] In this case, since we are working over a four-month window, the first two months are the ‘good’ phase, the third month is the ‘action’ phase, while the fourth month is the ‘churn’ phase. The data dictionary contains meanings of abbreviations. Some frequent ones are loc (local), IC (incoming), OG (outgoing), T2T (telecom operator to telecom operator), T2O (telecom operator to another operator), RECH (recharge) etc.

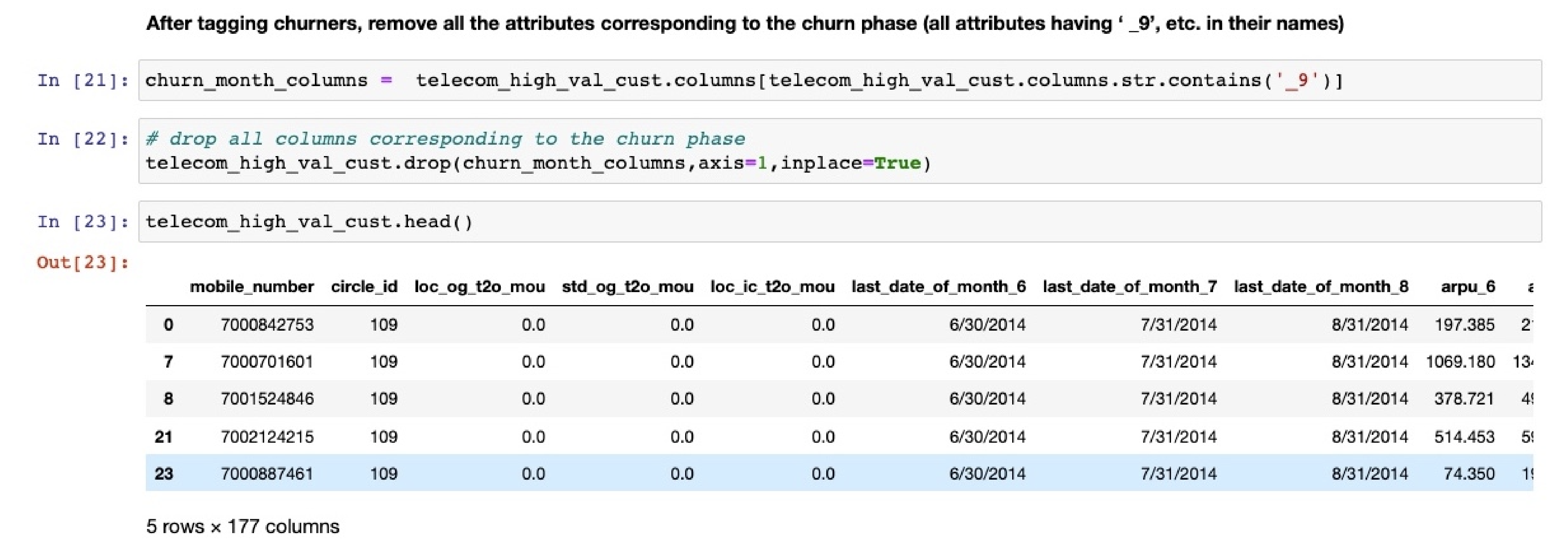
Data Shape: (99999, 226)

Data Prepration

1. **Filtering High-Value Customers**
   - To predict churn accurately, we will focus on high-value customers. We define high-value customers as those who have recharged with an amount greater than or equal to the 70th percentile of the average recharge amount in the first two months (the "good" phase).
   - Apply this filter to the dataset, which will result in approximately 30,000 rows representing high-value customers.

2. Tagging Churners and Removing Churn Phase Attributes

- To identify churned customers, we will tag them based on their behavior in the fourth month (the "churn" phase).
- Tag customers as churners (churn=1) if they have not made any incoming or outgoing calls and have not used mobile internet even once during this churn phase.
- Use the following attributes to tag churners: total_ic_mou_9, total_og_mou_9, vol_2g_mb_9, and vol_3g_mb_9.
- After tagging churners, remove all attributes corresponding to the churn phase by excluding attributes with names ending in '_9' or similar suffixes.



After tagging churners, remove all the attributes corresponding to the churn phase (all attributes having ' _9', etc. in their names)

```
In [21]: churn_month_columns =  telecom_high_val_cust.columns[telecom_high_val_cust.columns.str.contains('_9')]
```

```
In [22]: # drop all columns corresponding to the churn phase
         telecom_high_val_cust.drop(churn_month_columns,axis=1,inplace=True)
```

```
In [23]: telecom_high_val_cust.head()
```

Out[23]:

	mobile_number	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	arpu_6	a
0	7000842753	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	197.385	2
7	7000701601	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	1069.180	13
8	7001524846	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	378.721	4
21	7002124215	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	514.453	5
23	7000887461	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	74.350	1

5 rows × 177 columns

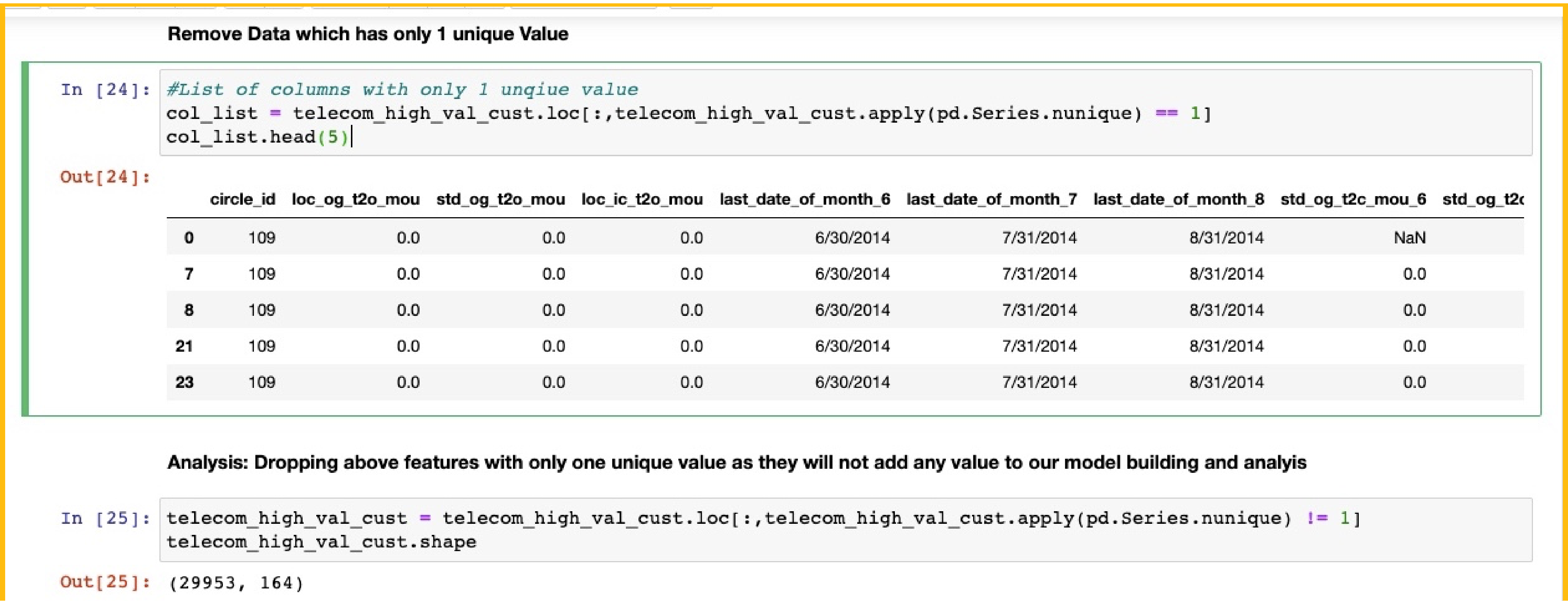
Data Cleaning

1.Removing Columns with Only 1 Unique Value

- Some columns in the dataset contain only one unique value and do not provide any useful information for our analysis and model building.
- We have identified the following columns with only one unique value: circle_id, loc_og_t2o_mou, std_og_t2o_mou, loc_ic_t2o_mou, last_date_of_month_6, last_date_of_month_7, last_date_of_month_8, std_og_t2c_mou_6, std_og_t2c_mou_7, std_og_t2c_mou_8, std_ic_t2o_mou_6, std_ic_t2o_mou_7, and std_ic_t2o_mou_8.
- We will drop these columns from the dataset as they do not add any value to our analysis.

2. Converting Date Columns to DateTime Format

- We have identified columns that represent dates, such as date_of_last_rech_6, date_of_last_rech_7, date_of_last_rech_8, date_of_last_rech_data_6, date_of_last_rech_data_7, and date_of_last_rech_data_8.
- To facilitate further analysis, we will convert these columns to the DateTime format using the pd.to_datetime() function.



Remove Data which has only 1 unique Value

```
In [24]: #List of columns with only 1 unqiue value
         col_list = telecom_high_val_cust.loc[:,telecom_high_val_cust.apply(pd.Series.nunique) == 1]
         col_list.head(5)
```

Out[24]:

	circle_id	loc_og_t2o_mou	std_og_t2o_mou	loc_ic_t2o_mou	last_date_of_month_6	last_date_of_month_7	last_date_of_month_8	std_og_t2c_mou_6	std_og_t2c
0	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	NaN	
7	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
8	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
21	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	
23	109	0.0	0.0	0.0	6/30/2014	7/31/2014	8/31/2014	0.0	

Analysis: Dropping above features with only one unique value as they will not add any value to our model building and analyis

```
In [25]: telecom_high_val_cust = telecom_high_val_cust.loc[:,telecom_high_val_cust.apply(pd.Series.nunique) != 1]
         telecom_high_val_cust.shape

Out[25]: (29953, 164)
```

Exploratory Data Analysis

1. **Recharge Amount Analysis:**

- We analyzed the recharge amount related variables such as total_rech_amt, max_rech_amt, av_rech_amt_data, and total_rech_amt_data.
- Plotted box plots to compare the distribution of recharge amounts for churned and non-churned customers in the 6th, 7th, and 8th months.
- Found a significant drop in total recharge amount, maximum recharge amount, and total recharge amount for data in the 8th month for churned customers.

2. 2G and 3G Usage Analysis:

- Explored the monthly 2G and 3G service schemes.
- Checked for missing values in these variables and found no missing values.
- Conducted analysis on 2G and 3G usage variables such as vol_2g_mb and vol_3g_mb.

3. Voice Call Minutes of Usage Analysis:

- Analyzed the minutes of usage for voice calls.
- Checked for missing values in voice call minutes of usage variables and filled the missing values with zeros.
- Examined the correlation between different voice call minutes of usage variables using a heatmap.
- Dropped highly correlated attributes related to outgoing and incoming minutes of usage.

4. Tenure Analysis:

- Derived the tenure of customers based on their age on network (aon) in months.
- Grouped customers into different tenure ranges and plotted a count plot to visualize churn behavior based on tenure.

5. Dropping Columns with High Missing Values:

- Identified columns with more than 30% missing values.
- Dropped columns related to dates and other variables with high missing percentages.

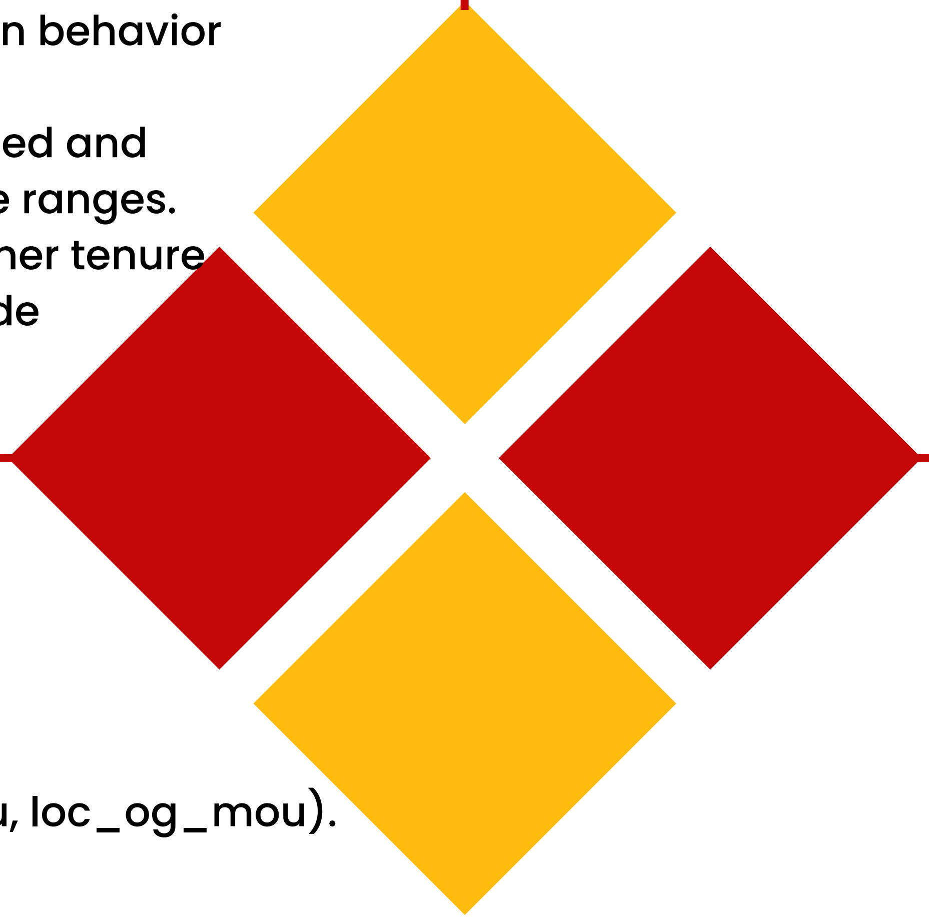
6. Replacing Missing Values:

- Replaced remaining missing values with zeros for all numeric variables.

7. Feature Engineering:

- Derived new features by combining variables from the 6th and 7th months.
- Dropped redundant columns and retained relevant features for further analysis.

Data Visualizatoin



Count Plot

Tenure analysis of customers

- The plot helps in understanding the churn behavior based on customer tenure.
- We can observe the distribution of churned and non-churned customers across different tenure ranges.
- This analysis provides insights into whether tenure has an impact on customer churn and can guide future strategies for customer retention.

Heatmap

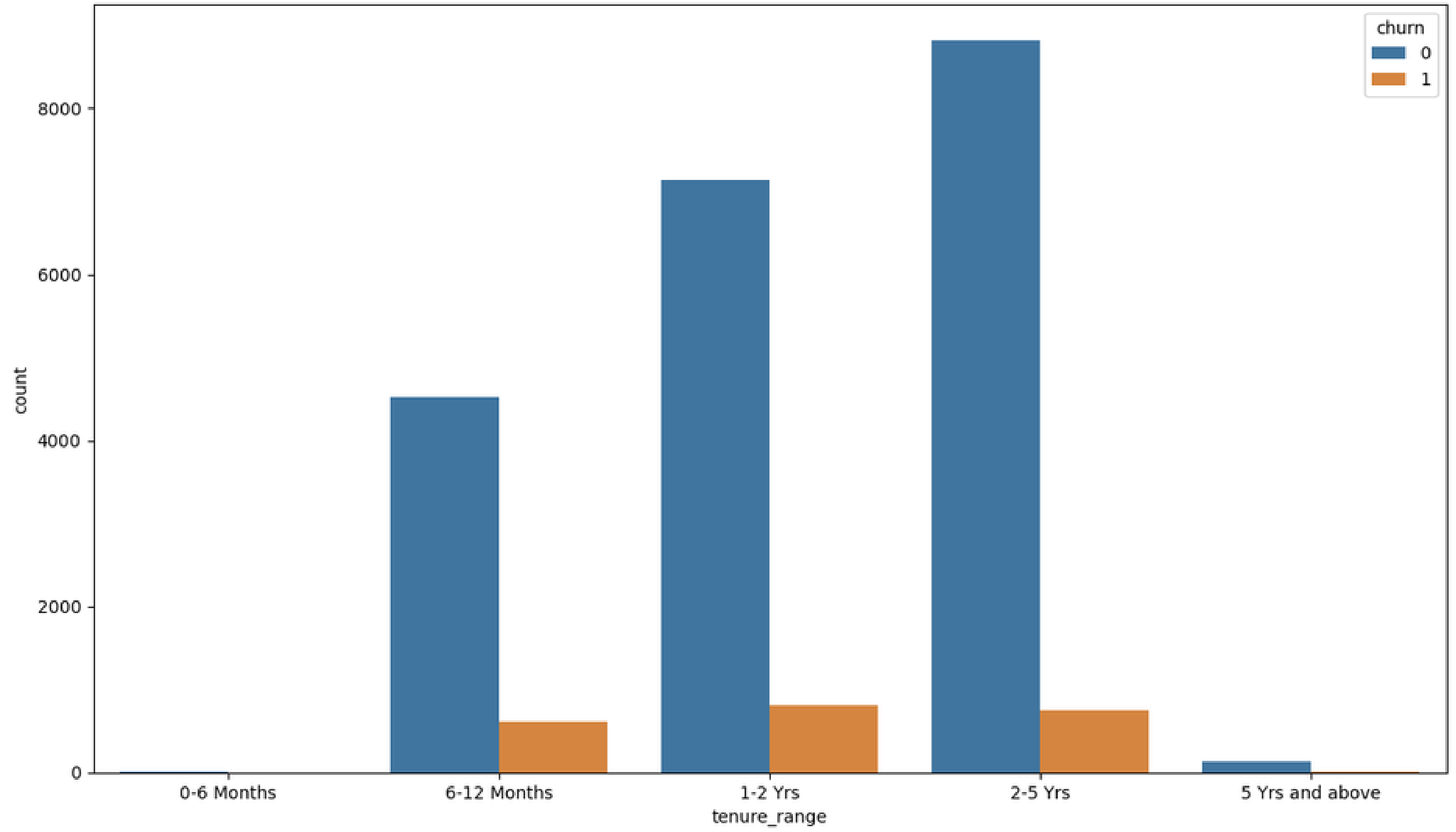
1. Outgoing Minutes of Usage (Mou_og_cols6):
   - Heat map analysis revealed highly correlated attributes.
   - Dropped correlated attributes (total_og_mou, std_og_mou, loc_og_mou).
   - Improved feature independence and reduced correlation.
2. Incoming Minutes of Usage (Mou_ic_cols6):
   - Heat map analysis identified highly correlated attributes.
   - Removed correlated attributes (total_ic_mou, std_ic_mou, loc_ic_mou).
   - Reduced correlation and improved feature independence.

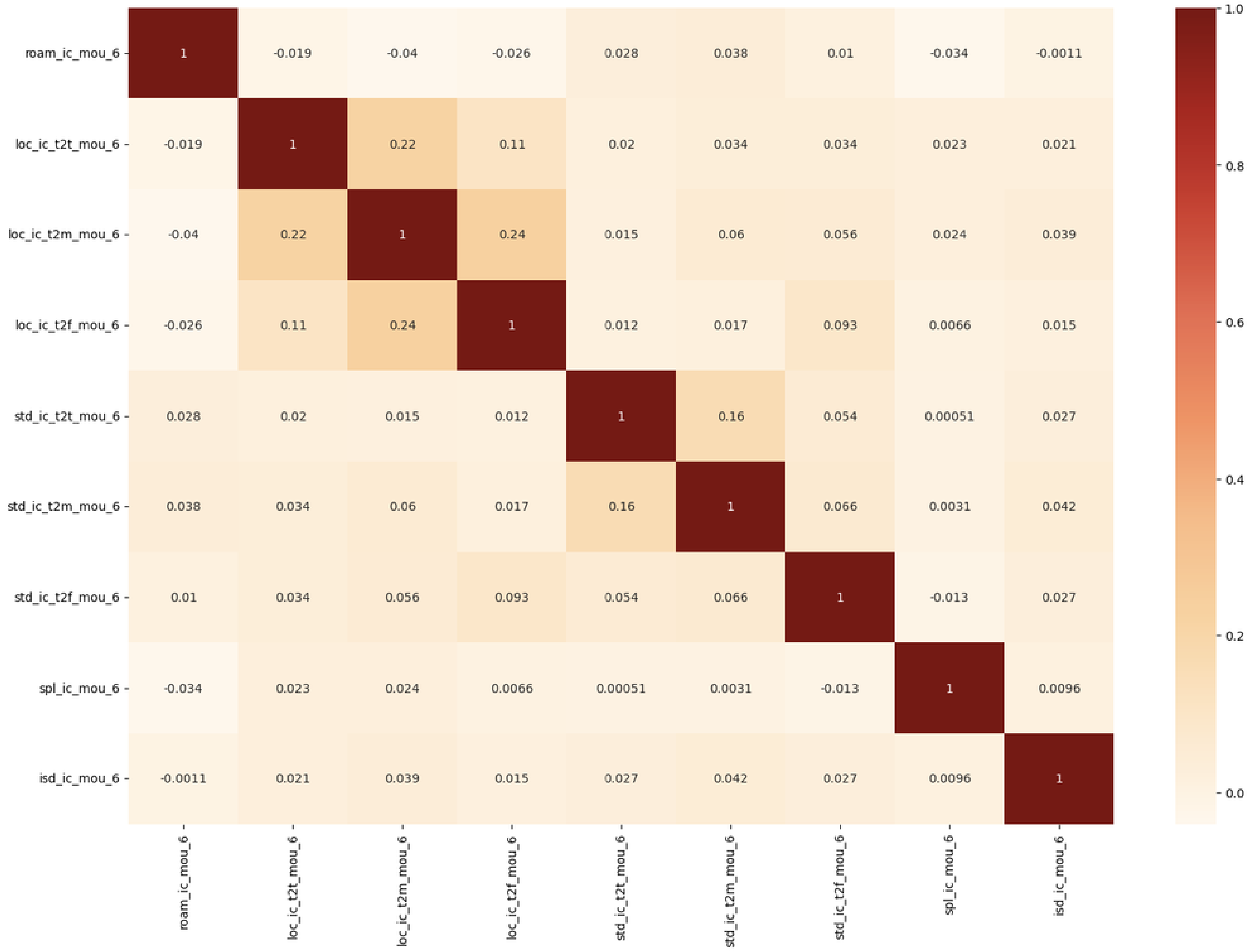
By dropping highly correlated attributes, we improved the dataset quality, reduced multicollinearity, and prepared the data for further analysis and modeling to identify predictors for customer churn.

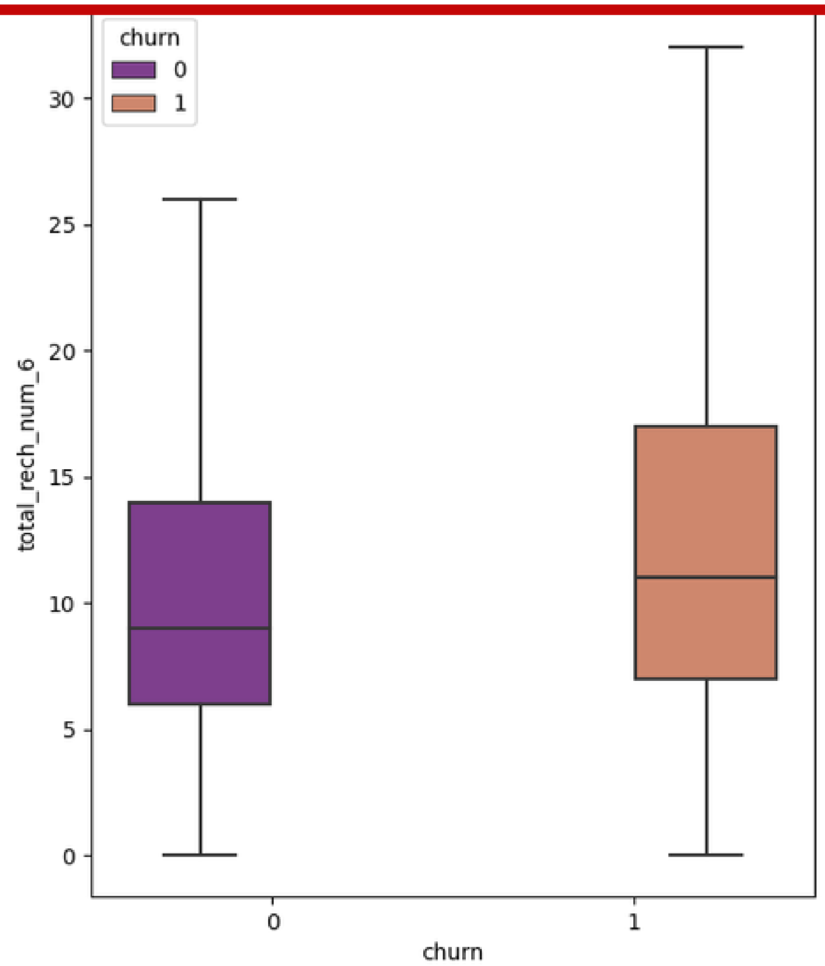
Bar Chart

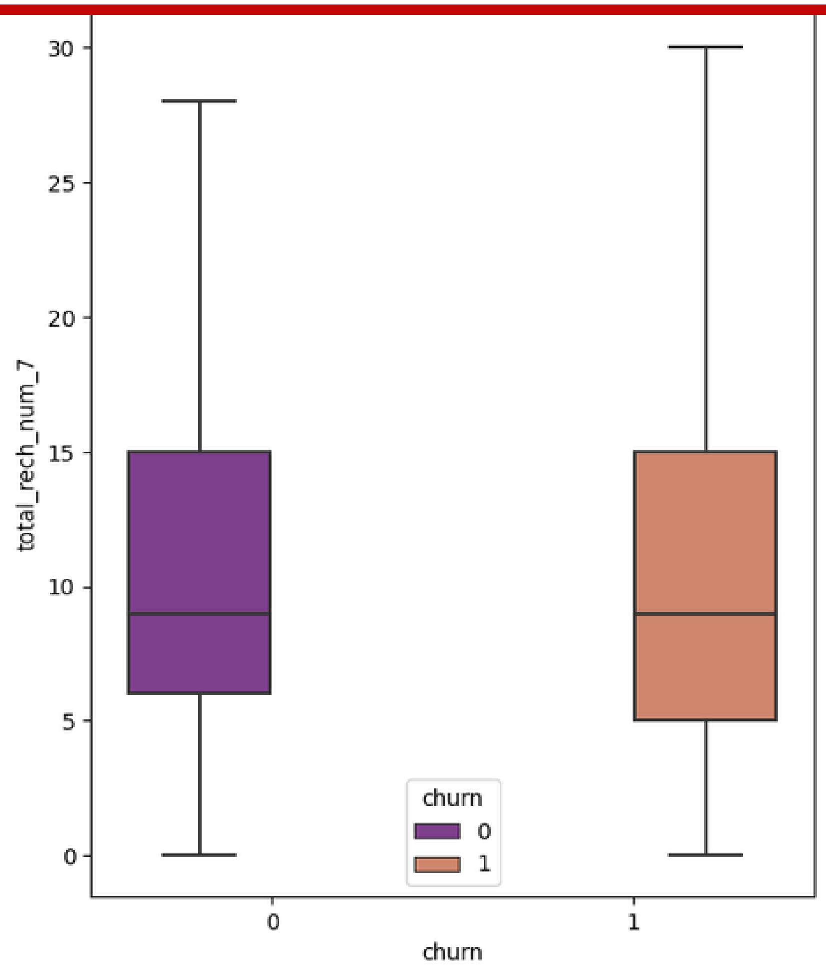
1. Recharge Amount:
   - Total recharge amount drops in the 8th month for churned customers.
   - Data recharge amount also decreases significantly in the 8th month for churned customers.
   - Maximum recharge amount for data shows a similar drop in the 8th month for churned customers.
2. Other Recharge Variables:
   - Some variables related to recharge show a high percentage of missing values.
   - Missing values in 'max_rech_data' variables indicate no recharge, so they are filled with 0.
3. Total Recharge Number:
   - There is a significant drop in the total recharge number in the 8th month for churned customers.
4. 2G and 3G Usage:
   - Monthly 2G/3G service schemes are available without missing values.
5. Minutes of Usage - Voice Calls:
   - Missing values for minutes of usage columns are relatively low (maximum of 3.91%).
   - Missing values indicate no usage for those specific call types and are filled with zeros.

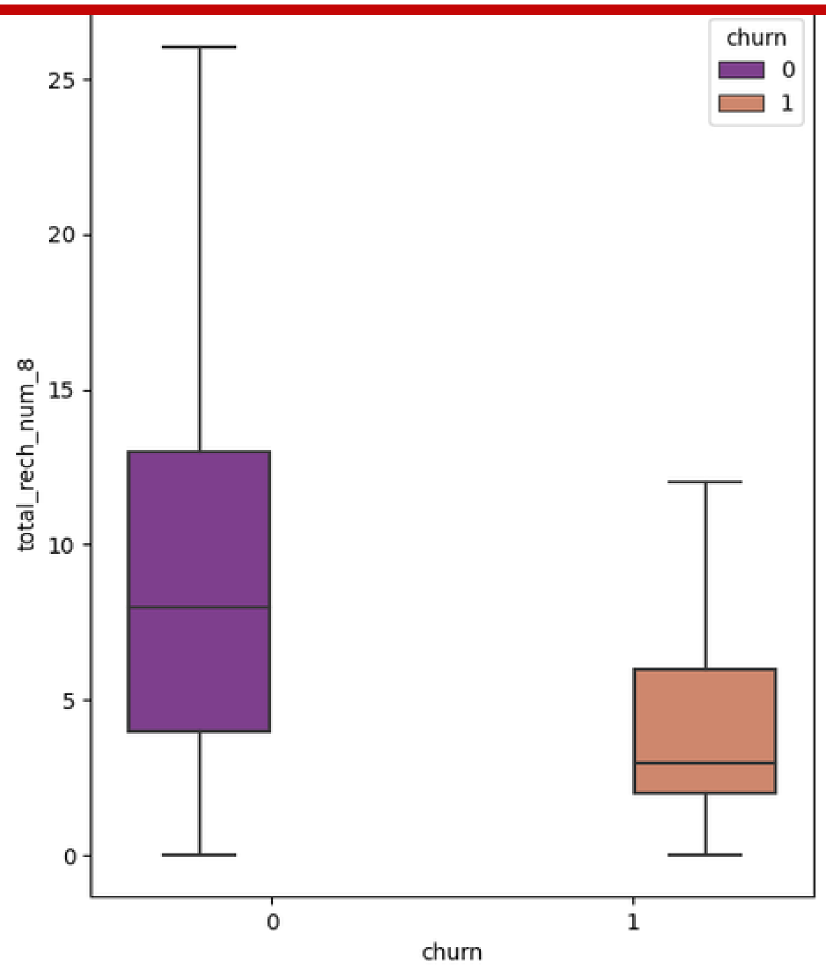
Graphs & Charts











Model Selection

1. **Splitting Data:**
   - The dataset was split into training and test sets using the train_test_split function from the sklearn.model_selection module.
   - The feature variables were stored in the X variable, excluding the 'churn' and 'mobile_number' columns.
   - The response variable (churn) was stored in the y variable.

2. **Principal Component Analysis (PCA):**
   - PCA was performed on the training set to reduce the dimensionality of the data.
   - The PCA components were calculated using the fit_transform method on the training set.
   - The same PCA transformation was applied to the test set using the transform method.

Model Evaluation

1. **Logistic Regression:**

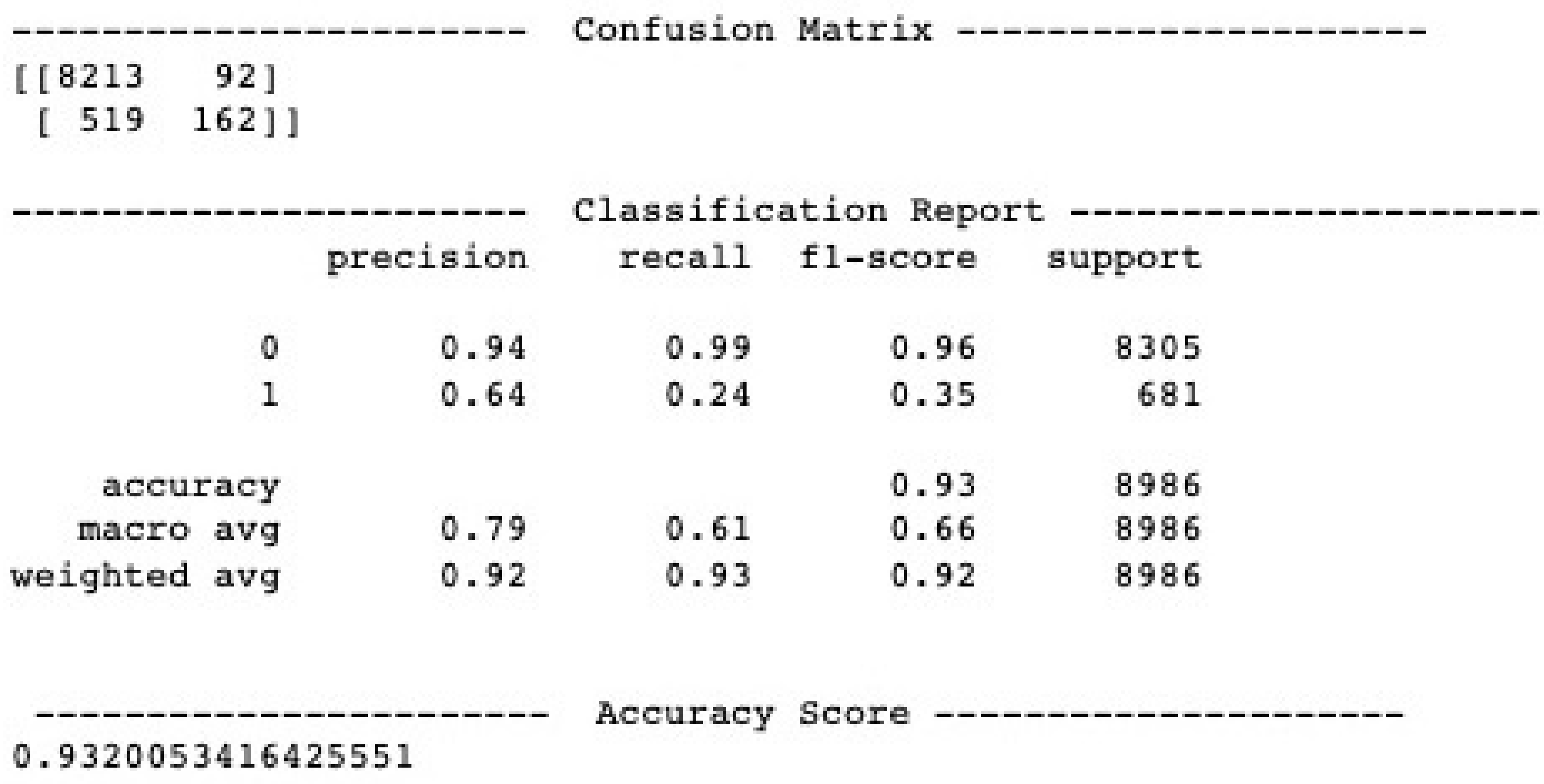
    - Logistic Regression was applied to the transformed data.
    - The model was trained using the fit method on the training set.
    - Predictions were made on the test set using the predict method.
    - Classification report, confusion matrix, and accuracy score were computed using the respective functions from sklearn.metrics.

2. Random Forest:

- Random Forest Classifier was applied to the original feature variables.
- The model was trained using the fit method on the training set.
- Predictions were made on the test set using the predict method.
- Classification report, confusion matrix, and accuracy score were computed using the respective functions from sklearn.metrics.

Model Evaluation

3. **Parameter Tuning:**
    - Hyperparameter tuning was performed on the Random Forest model.
    - GridSearchCV was used to tune the hyperparameters (max_depth, min_samples_leaf, min_samples_split).
    - Scree plots were generated to visualize the model performance with different parameter values.
    - The optimal hyperparameter values were selected based on the plots and used to train the final Random Forest model.
4. **Analysis:**
    - The models were evaluated based on their accuracy, precision, recall, and F1-score.
    - The Random Forest model achieved higher accuracy compared to Logistic Regression.
    - Hyperparameter tuning helped improve the performance of the Random Forest model.
    - Important features identified through feature importance analysis can provide insights into the factors influencing churn and guide future strategies for customer retention.



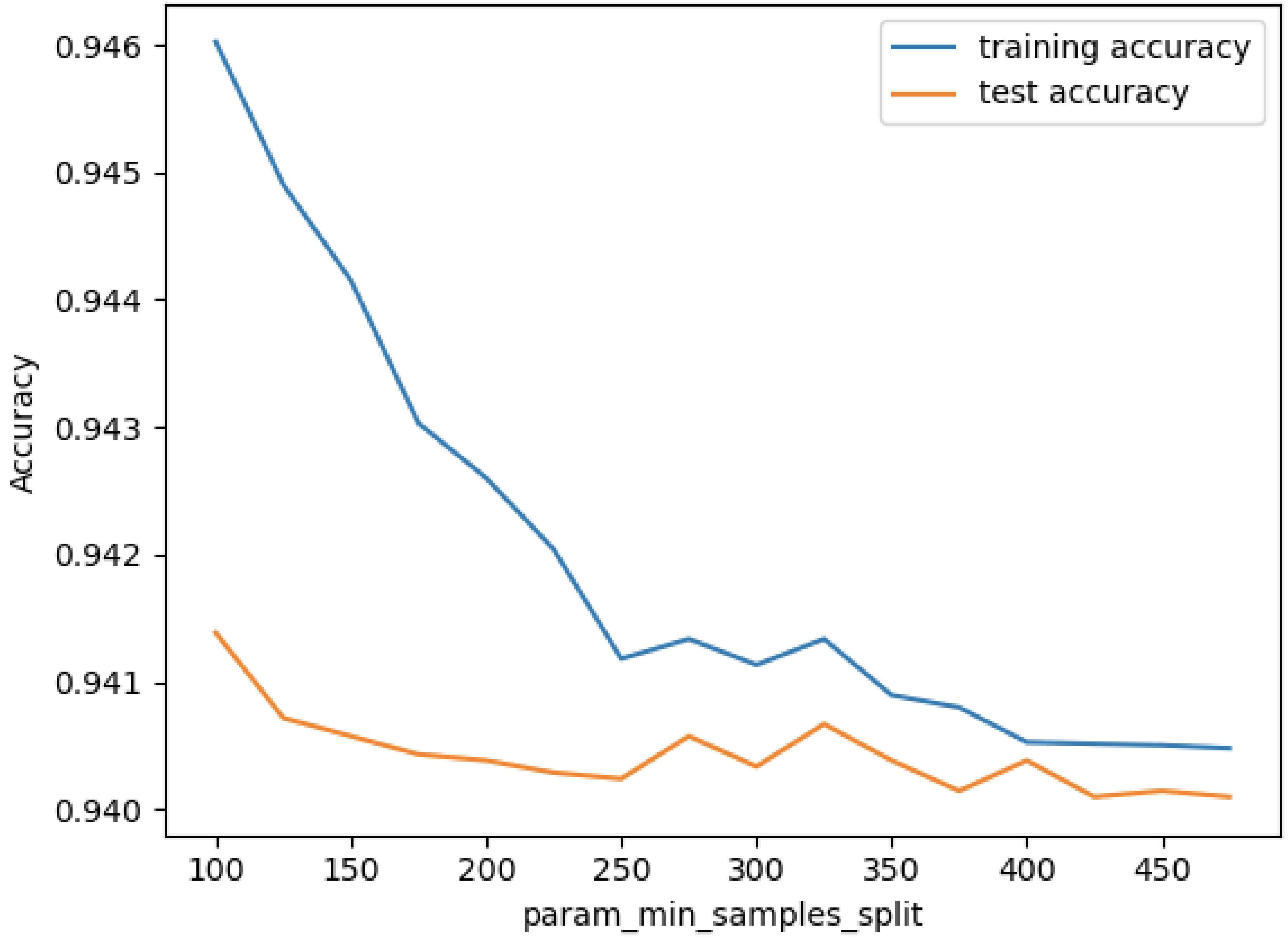
```
-----------------------  Confusion Matrix ---------------------
[[8213   92]
 [ 519  162]]

-----------------------  Classification Report ---------------------
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      8305
           1       0.64      0.24      0.35       681

    accuracy                           0.93      8986
   macro avg       0.79      0.61      0.66      8986
weighted avg       0.92      0.93      0.92      8986


 -----------------------  Accuracy Score ---------------------
0.9320053416425551
```



Insights & Analysis

- The models were evaluated based on their accuracy, precision, recall, and F1-score.
- The Random Forest model achieved higher accuracy compared to Logistic Regression.
- Hyperparameter tuning helped improve the performance of the Random Forest model.
- Important features identified through feature importance analysis can provide insights into the factors influencing churn and guide future strategies for customer retention.

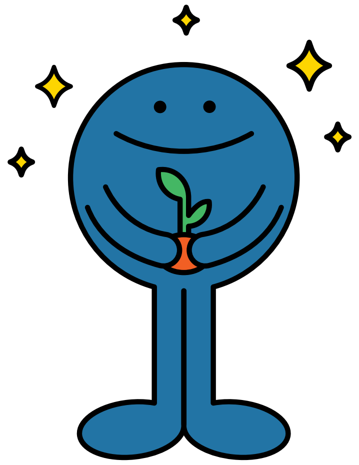
Conclusion

- Churn of high-value customers is relatively low, but the absence of new high-value customer acquisitions in the last 6 months is concerning. The company should focus on attracting new high-value customers to ensure sustainable growth and revenue generation.
- Customers with less than 4 years of tenure are more likely to churn. This segment requires special attention from the company. Rolling out new schemes like loyalty programs or personalized offers can help retain customers in this group and improve their loyalty.
- Average revenue per user (ARPU) is the most important feature in determining churn prediction. It indicates that customers who generate higher revenue are less likely to churn. The company should strive to increase the ARPU by providing value-added services, upselling/cross-selling, and improving customer experience.
- Incoming and outgoing calls on roaming in the 8th month are strong indicators of churn behavior. Customers who exhibit significant roaming activity in their calls are more likely to churn. The company should analyze the reasons behind this behavior and take proactive measures to address the concerns of these customers, such as offering attractive roaming packages or improving network coverage in specific areas.

In conclusion, the company should focus on acquiring new high-value customers, implementing loyalty programs for customers with shorter tenure, increasing average revenue per user, and addressing the churn behavior associated with roaming calls. These strategies can help reduce churn rate, improve customer retention, and drive business growth in the telecom industry.

Meet Our Team

Gopiganapathy Kabaleeswaran



Shruthi Sunil

Soundari Govindan

Thank You