PREDICTING CUSTOMER CHURN IN THE TELECOMMUNICATIONS INDUSTRY

A DATA-DRIVEN APPROACH TO IMPROVE CUSTOMER RETENTION

INTRODUCTION

- This project aims to develop a classification model that will predict customer churn for syriatel, a telecommunications company. I have chosen to follow the CRISP-DM method to complete this project. It will involve six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The project purposes to provide insights into the patterns and factors influencing customer churn, and also develop a predictive model to assist in reducing customer attrition.

BUSINESS UNDERSTANDING

Syriatel is the major stakeholder of this project.

Main objectives:

- Reduce customer churn.
- Predicting customer churn will enable proactive measures for customer retention.
- The project aims to identify patterns leading to customer churn.
- Recommendations will be provided to mitigate customer churn.
- The focus is on maximizing customer retention and profit for syriatel.

DATA UNDERSTANDING

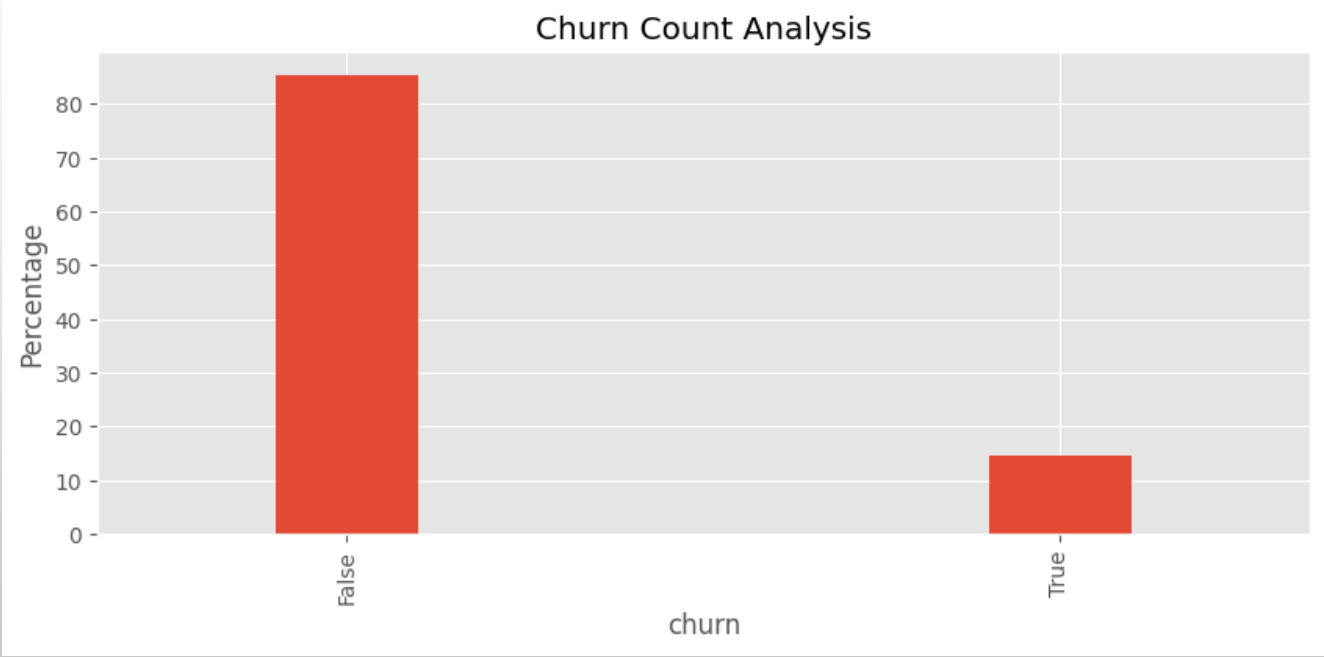
- The dataset is available at kaggle.Com for download through the link
- Https://www.Kaggle.Com/becksddf/churn-in-telecoms-dataset
- There are 3333 rows and 21 columns in our dataset
- We have both categorical and numerical variables in our dataset
- Categorical variables: 'state', 'phone number', 'international plan', 'voice mail plan', 'churn'
- Numerical variables: 'account length', 'area code', 'number vmail messages',
- 'total day minutes', 'total day calls', 'total day charge',
- 'total eve minutes', 'total eve calls', 'total eve charge',
- 'total night minutes', 'total night calls', 'total night charge',
- 'total intl minutes', 'total intl calls', 'total intl charge',
- 'customer service calls'

DATA EXPLORATION

- We identified that our dataset had zero null values and no duplicates. That's great!
- We also visualized some categorical variables so that we can see their effect on 'CHURN' which in this case is/will be our dependent variable(predicted).
- We also plotted distribution plots for the continuous variable to have an idea of how they are distributed.
- A correlation matrix was visualized in a heatmap so that it can be easy to interprate.

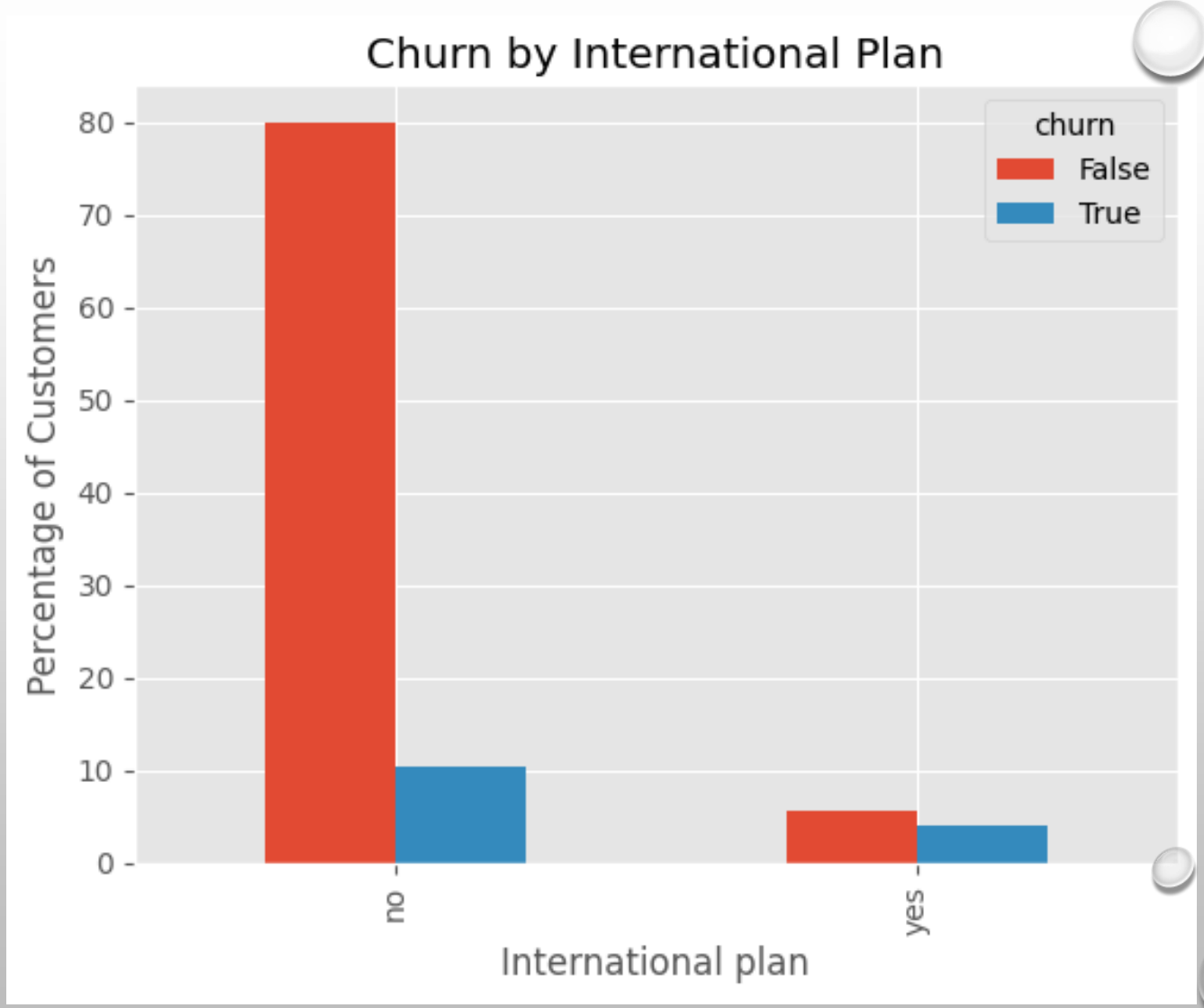
DATA VISUALIZATION

- The analysis of the churn variable reveals that 85.51% of customers do not churn, while 14.49% of customers churn from the company.



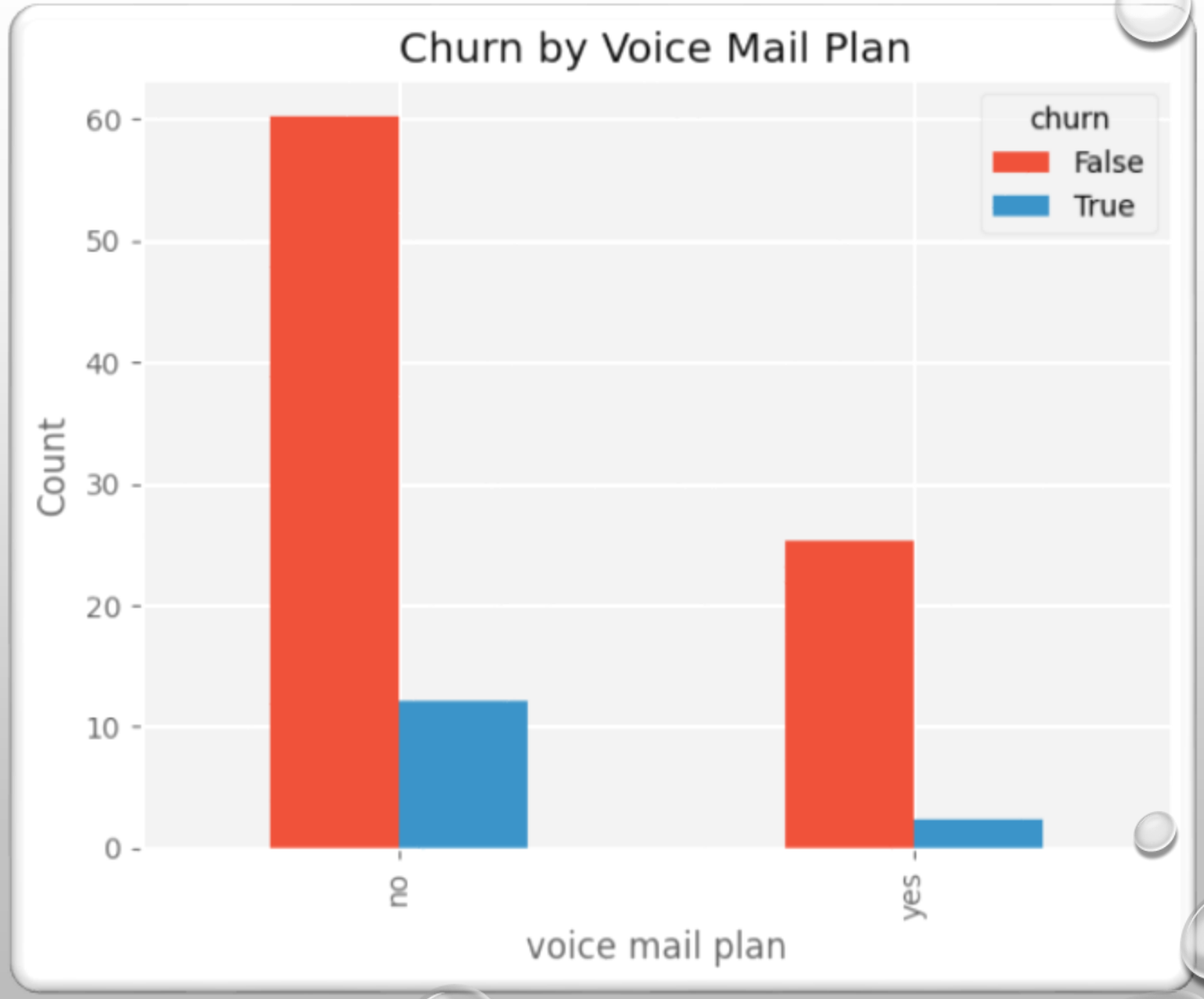
DATA VISUALIZATION

- Based on the bar graph above, it is evident that customers without an international plan have a higher percentage in both the 'false' and 'true' categories compared to customers with an international plan. This suggests that having an international plan may be associated with a lower likelihood of churn.



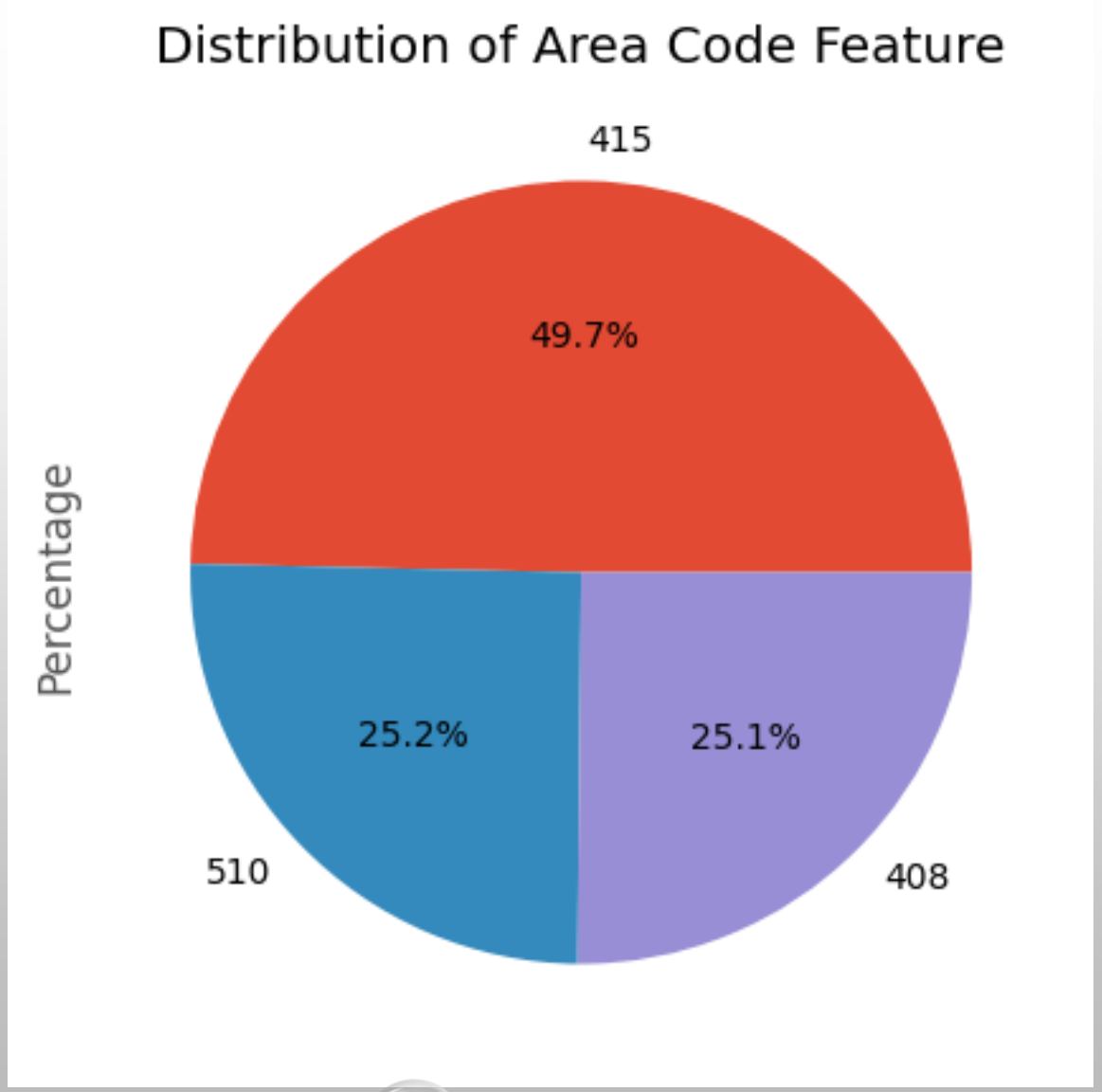
DATA VISUALIZATION

- From the graph above, it can be observed that the majority of customers who do not have a voice mail plan are in the 'false' category, while a smaller proportion is in the 'true' category. In addition, customers with a voice mail plan have a higher count in the 'false' category compared to the 'true' category. This may suggest that having a voice mail plan may have some influence on reducing churn

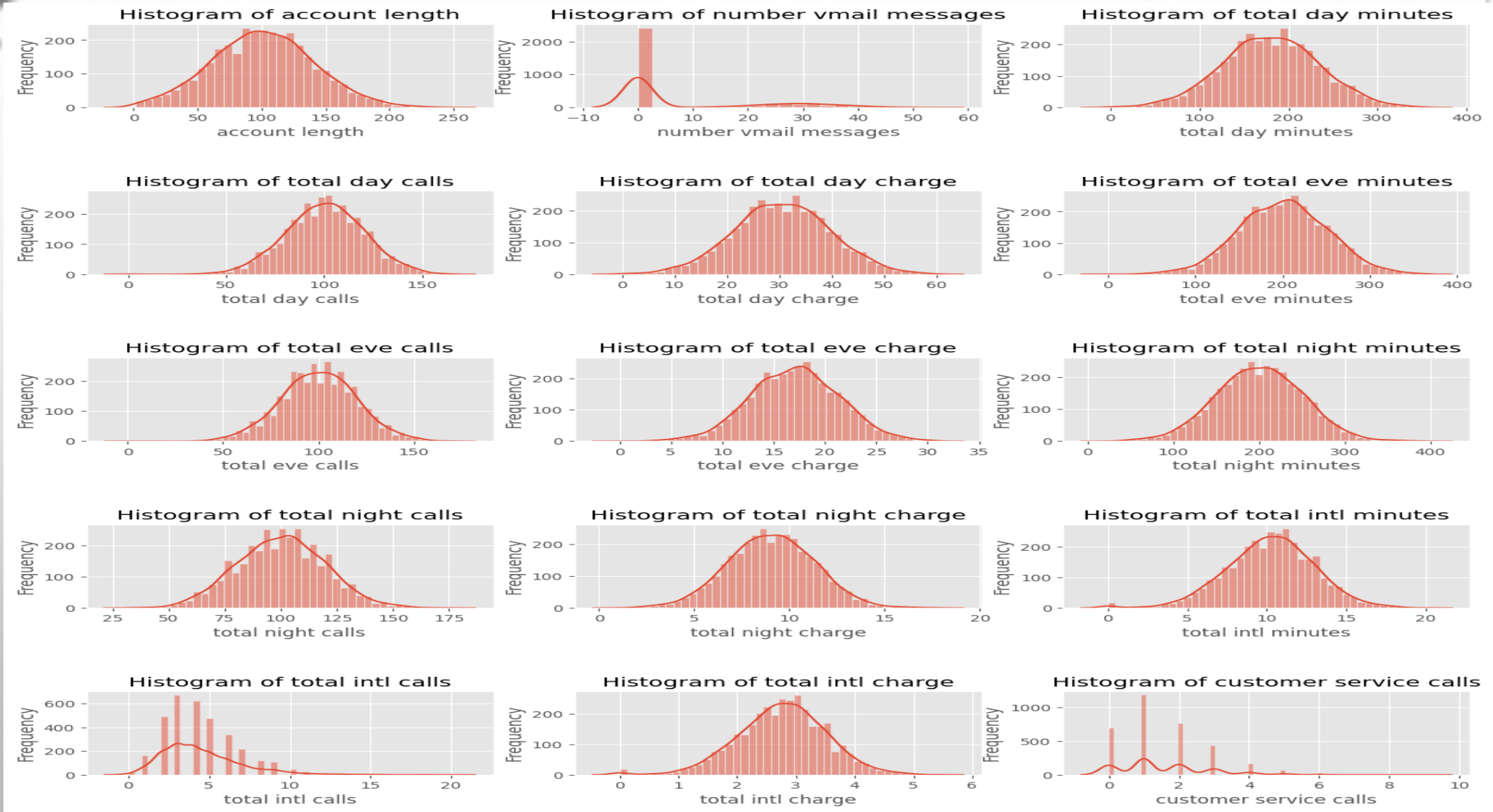


DATA VISUALIZATION

- Based on the distribution of area codes, the pie chart shows that 49.7% of customers come from area code 415, which is approximately half of the total number of customers. The remaining customers are evenly distributed between area codes 510 and 408, with each accounting for about 25% of the customer base. This clearly shows that Syriatel telecommunications has a huge fanbase at area code 415, thus they should capitalize on that area mostly.



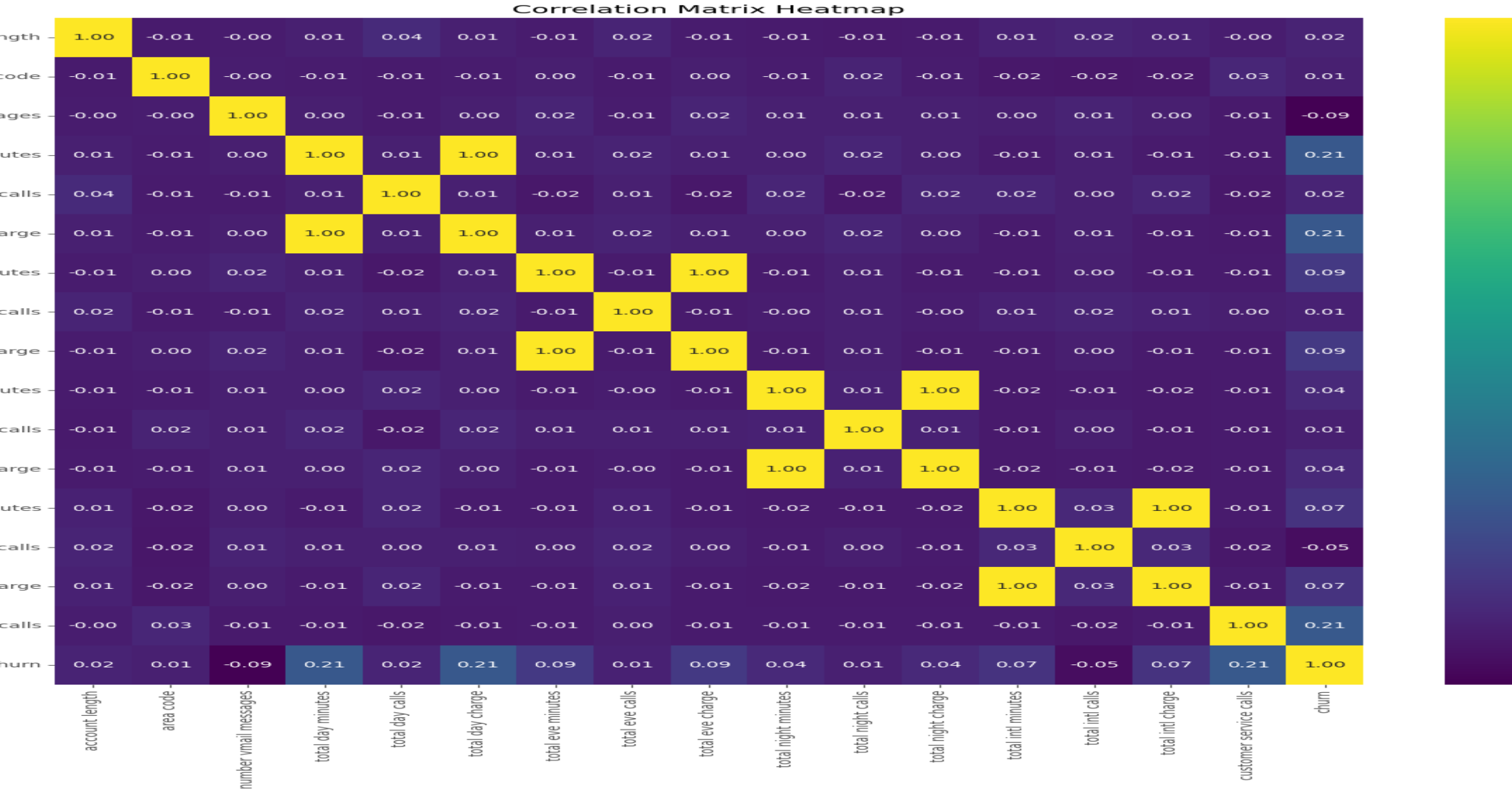
DISTRIBUTION PLOTS FOR NUMERIC VARIABLES



NUMERIC VARIABLES DIST EXPLANATION

- The numerical variables exhibit diverse distributions and ranges, indicating variations in customer behavior and call patterns. While some variables follow approximately normal distributions, others display skewed distributions. This suggests that the variables may require different handling approaches based on their distributions for further analysis and modeling.

CORRELATION MATRIX PLOT



CORRELATION MATRIX EXPLANATION

- From the above correlation matrix, we can observe that most of the variables are not strongly correlated. However, there ares some variables that exhibit a perfect correlation. This makes sense since some variables are directly correlated. For example, tel charges are directly correlated to the time spent on call.

DATA PROCESSING

To prepare our data for modeling, we:

- Remove outliers from the dataset
- Address multicollinearity by removing highly correlated features
- Perform scaling on the numerical variables
- Handle class imbalance using the SMOTE method

MODELING

- In the modeling step, we trained and evaluated different machine learning models on our dataset to make predictions for the target variable. This involved selecting appropriate algorithms, tuning their parameters, and assessing their performance using various evaluation metrics. The goal is to find the model that best captures the patterns and relationships in the data and provides accurate predictions.
- The three different models are Logistic Regression, Random Forest, and Support Vector Machine (SVM), are trained and evaluated using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and ROC AUC score.

MODEL 1: LOGISTIC REGRESSION

- After training the model, the results were as follows
- The accuracy of 0.783102 means that the model predicted approximately 78.31% of the cases correctly.
- The precision of 0.367150 indicates that around 36.71% of the predicted positive cases were actually true positives.
- The recall of 0.649573 suggests that the model captured approximately 64.95% of the actual positive cases.
- The F1 score of 0.469136 represents the balance between precision and recall, with at least 46.91% of values correctly predicted.
- The ROC AUC score of 0.727893 indicates a moderate level of discrimination power in distinguishing between the two classes.

	Model Evaluation Metric	Model Evaluation Score
0	Accuracy	0.783102
1	Precision	0.367150
2	Recall	0.649573
3	F1 Score	0.469136
4	ROC AUC Score	0.727893

MODEL 2: RANDOM FOREST

- The accuracy of 0.918033 means that the model predicted approximately 91.80% of the cases correctly.
- The precision of 0.728070 indicates that around 72.81% of the predicted positive cases were actually true positives.
- The recall of 0.709402 suggests that the model captured approximately 70.94% of the actual positive cases.
- The F1 score of 0.718615 represents the balance between precision and recall, with at least 71.86% of values correctly predicted.
- The ROC AUC score of 0.831772 indicates a moderate level of discrimination power in distinguishing between the two classes.

	Model Evaluation Metric	Model Evaluation Score
0	Accuracy	0.918033
1	Precision	0.728070
2	Recall	0.709402
3	F1 Score	0.718615
4	ROC AUC Score	0.831772

MODEL 3: SUPPORT VECTOR CLASSIFIER

- The SVC model achieved an accuracy of 0.823455, indicating that approximately 82.35% of the predictions made by the model were correct.
- The precision score of 0.427673 suggests that around 42.77% of the predicted positive cases were actually true positive cases.
- The recall score of 0.581197 indicates that the model captured approximately 58.12% of the actual positive cases.
- The F1 score of 0.492754 represents the balance between precision and recall, with a higher value indicating a better trade-off between the two.
- Lastly, the ROC AUC score of 0.723291 suggests that the model has a moderate level of discrimination power in distinguishing between the two classes.

	Model Evaluation Metric	Model Evaluation Score
0	Accuracy	0.823455
1	Precision	0.427673
2	Recall	0.581197
3	F1 Score	0.492754
4	ROC AUC Score	0.723291

RECEIVER OPERATOR CHARACTERISTIC (ROC) CURVE

CONCLUSION

- Based on the above ROC curves, the random forest model demonstrates the best performance among the three models. It exhibits the highest area under the curve (AUC), indicating a better ability to distinguish between the positive and negative classes. The random forest model's ROC curve is closer to the top-left corner, which suggests a higher true positive rate and a lower false positive rate. This indicates that the random forest model is more effective in predicting the target variable compared to the other models.