Telecom Churn Prediction Analysis by ChurnBusters Inc.

Business Understanding

- ChurnBusters Inc. is a leading telecommunications analytics firm specializing in the reduction of customer churn rates.
- In the telecom industry, "churning" denotes customers switching providers or canceling subscriptions.
- Our approach helps businesses identify churned customers based on time since their last interaction.

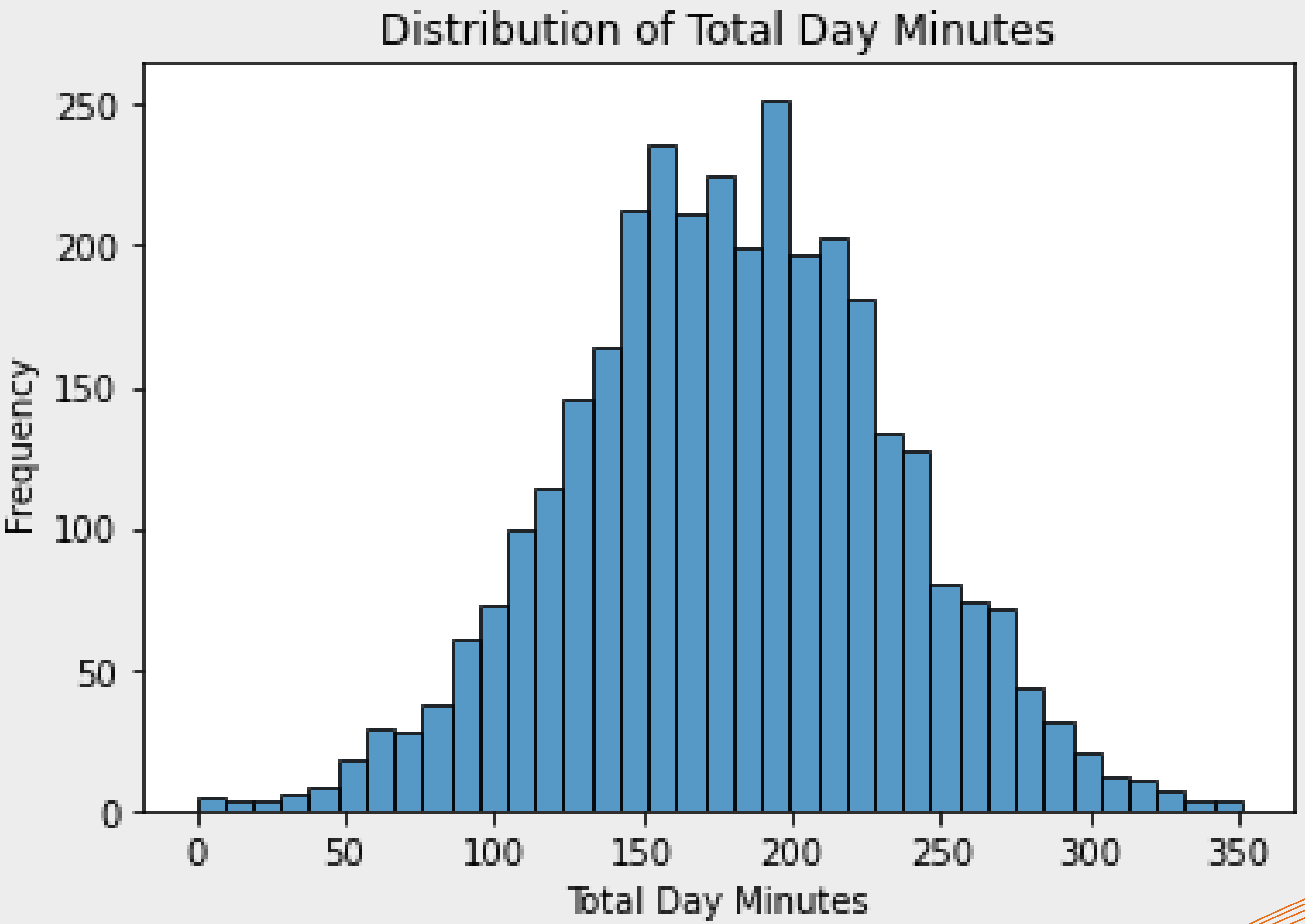
Data Understanding

- Our dataset offers an in-depth view of telecom customer information.

- Attributes include customer location, account length, plan status, call details, and more.

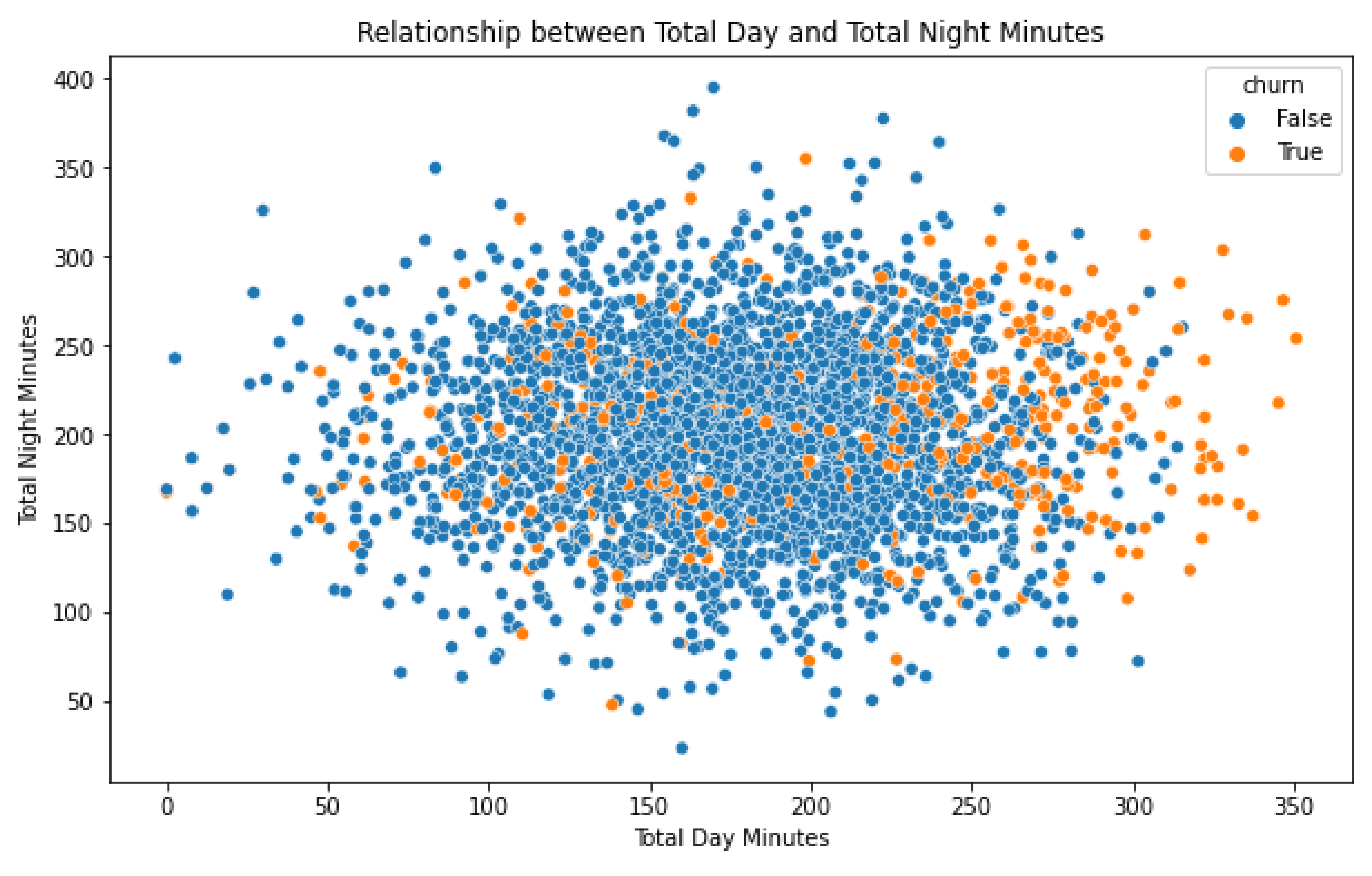
- The dataset is a valuable source for understanding customer behavior and churn factors.

EDA



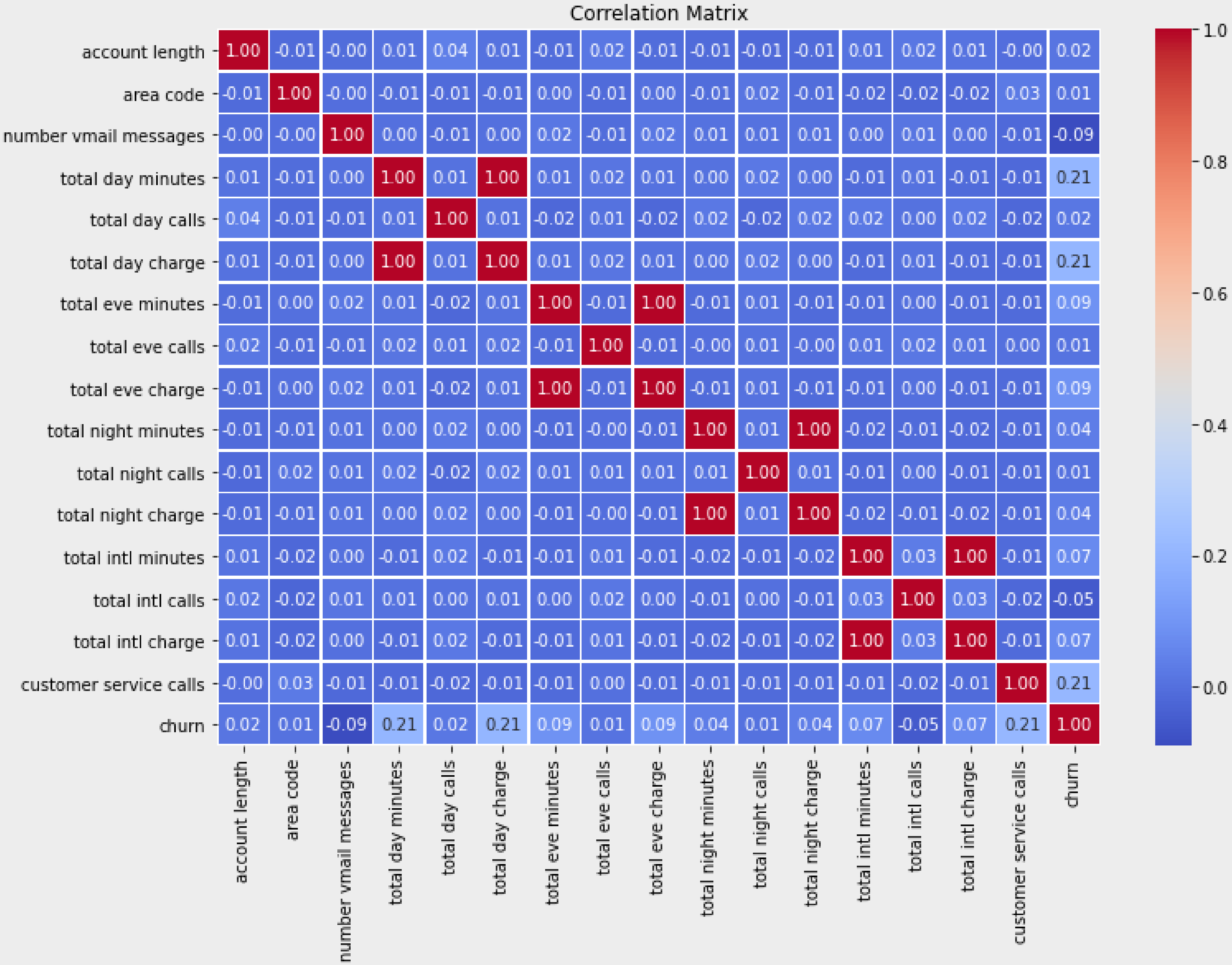
- Histogram displays "Total Day Minutes" distribution.
- X-axis: Daytime call minutes; Y-axis: Customer count.
- Majority: 175-200 minutes/day. Slight right-skew, implying potential outliers.

EDA



- Scatter plot explores "Total Day Minutes" vs. "Total Night Minutes."
- Blue dots (non-churn): Uniform distribution, weak relationship.
- Orange dots (churn): Scattered points, varied usage pattern.
- Indicates a weak link between day and night minutes, with churned customers using phones less.

EDA



- Heatmap visualizes variable correlations
- Dark red: Strong positive correlation; Dark blue: Strong negative correlation.
- Identifies high correlations: total day/eve/intl minutes & their respective charges.

DATA PRE-PROCESSING

Label Encoding for Ordinal Categorical Data:

Transforming ordinal categorical data in the columns `international plan`, `voice mail plan` and `churn` into numerical values.

One-Hot Encoding for Nominal Categorical Data:

Converting categorical data for the "state" column into binary values, creating separate columns for each category.

Splitting Data into Train and Test Sets:

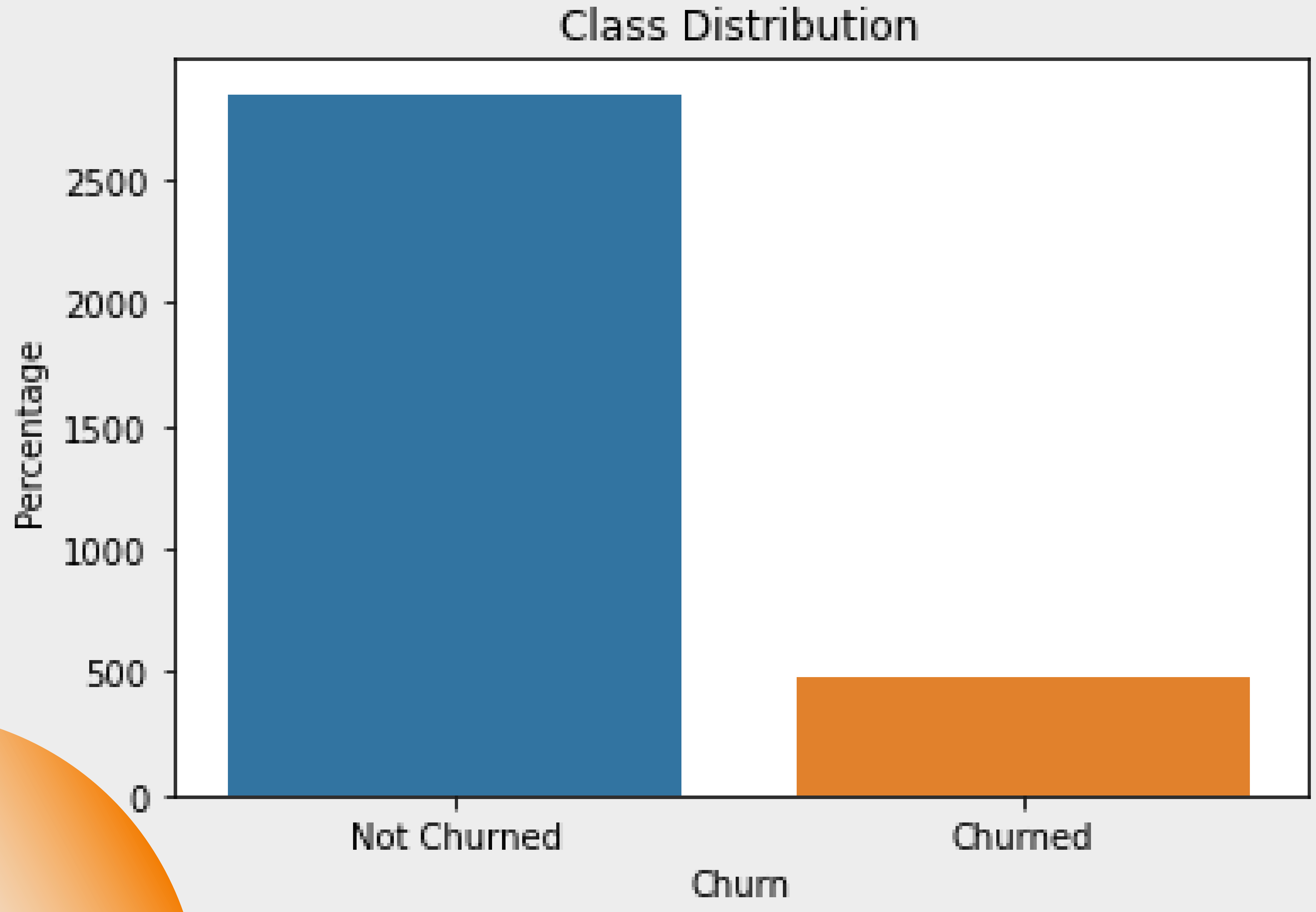
Separating the dataset into two parts – one for training the model and the other for evaluating its performance.

Feature Scaling for Numerical Data:

Standardizing numerical features to bring them to a common scale, aiding machine learning model performance.

DATA PRE-PROCESSING

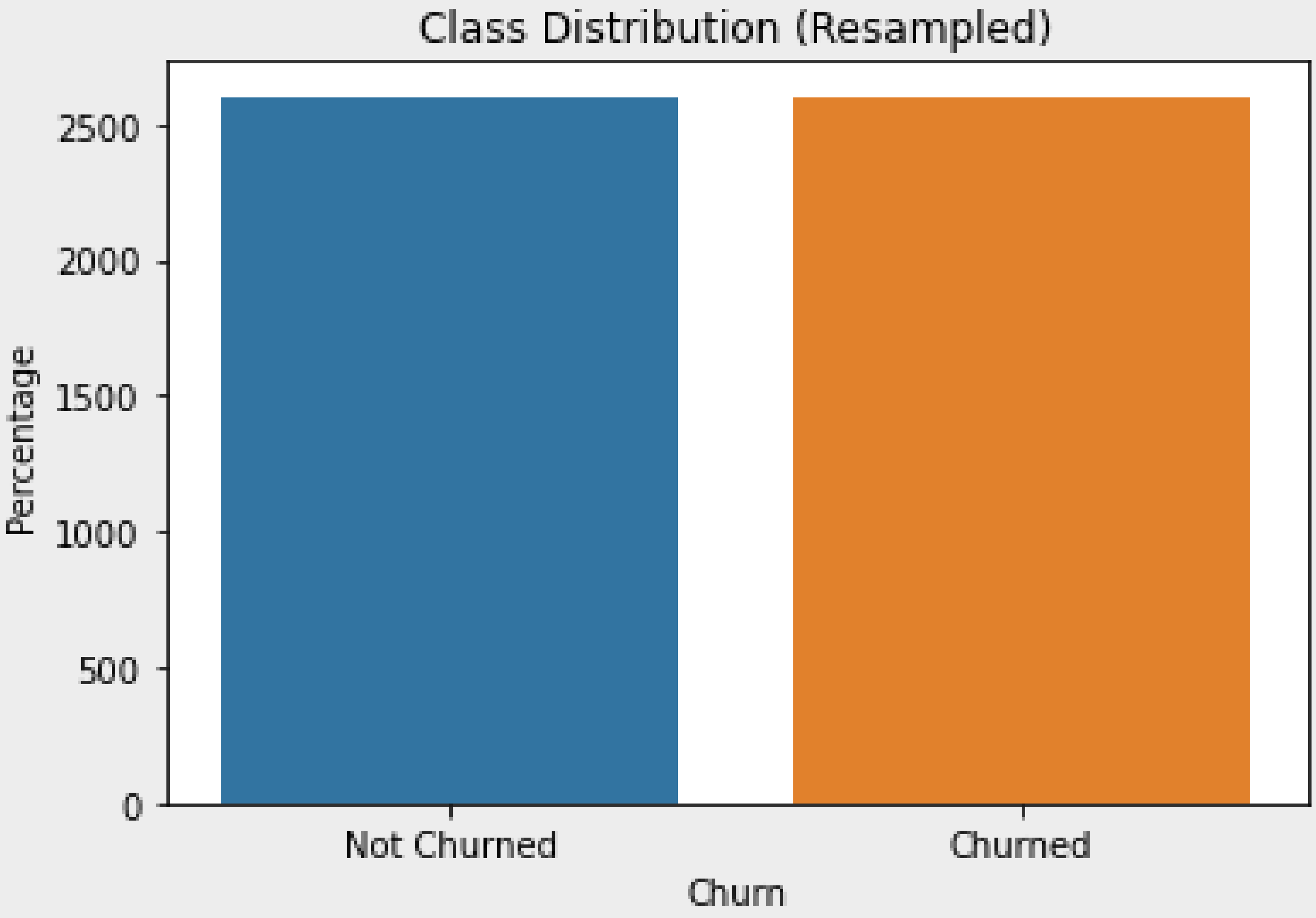
Check for Data Imbalance



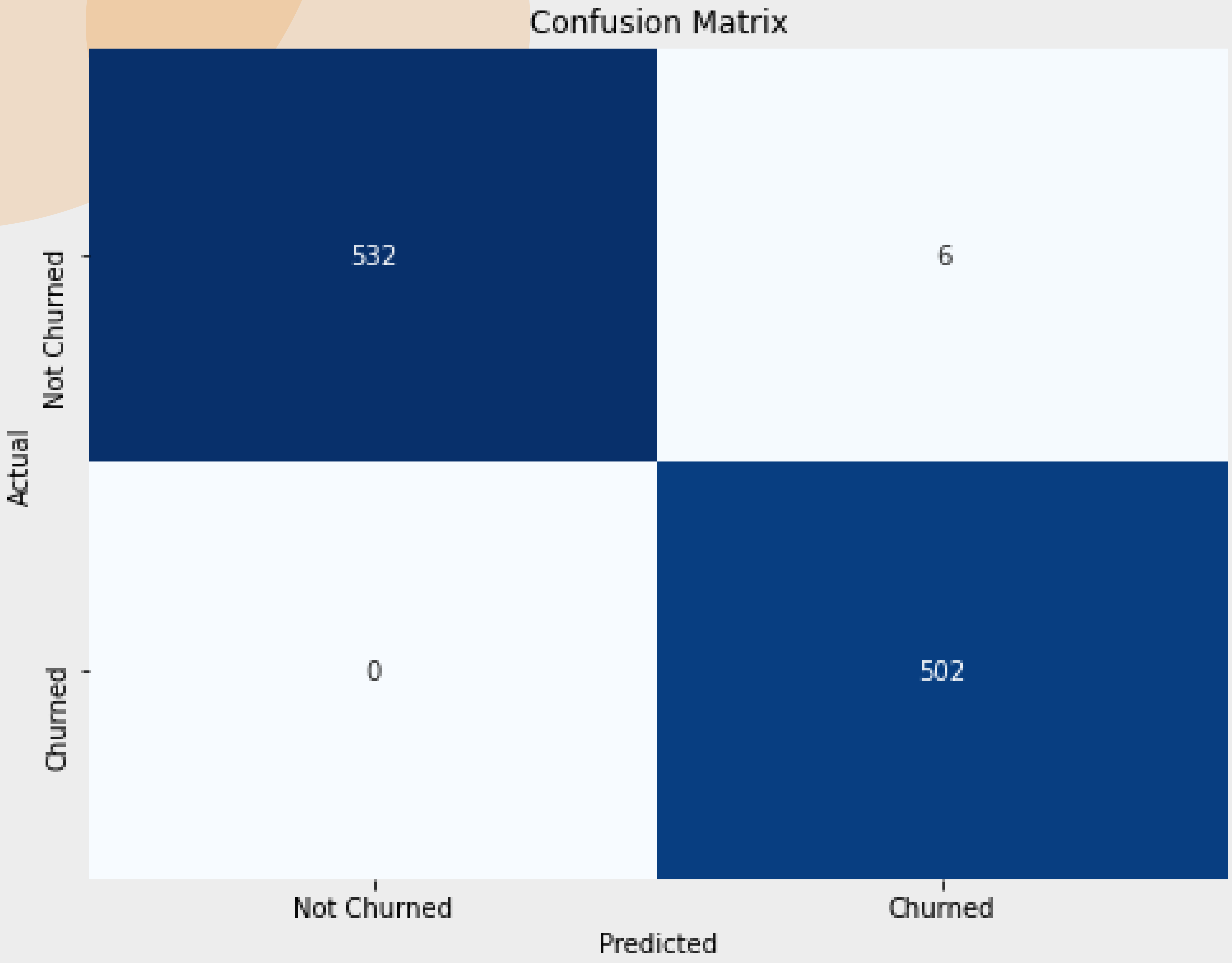
The figure reveals a class imbalance of "Not Churned" at around 2,850 (85.51%) instances, while "Churned" has 483 (14.49%) instances.

DATA PRE-PROCESSING

Solved imbalance through SMOTE



Model 1: Logistic Regression

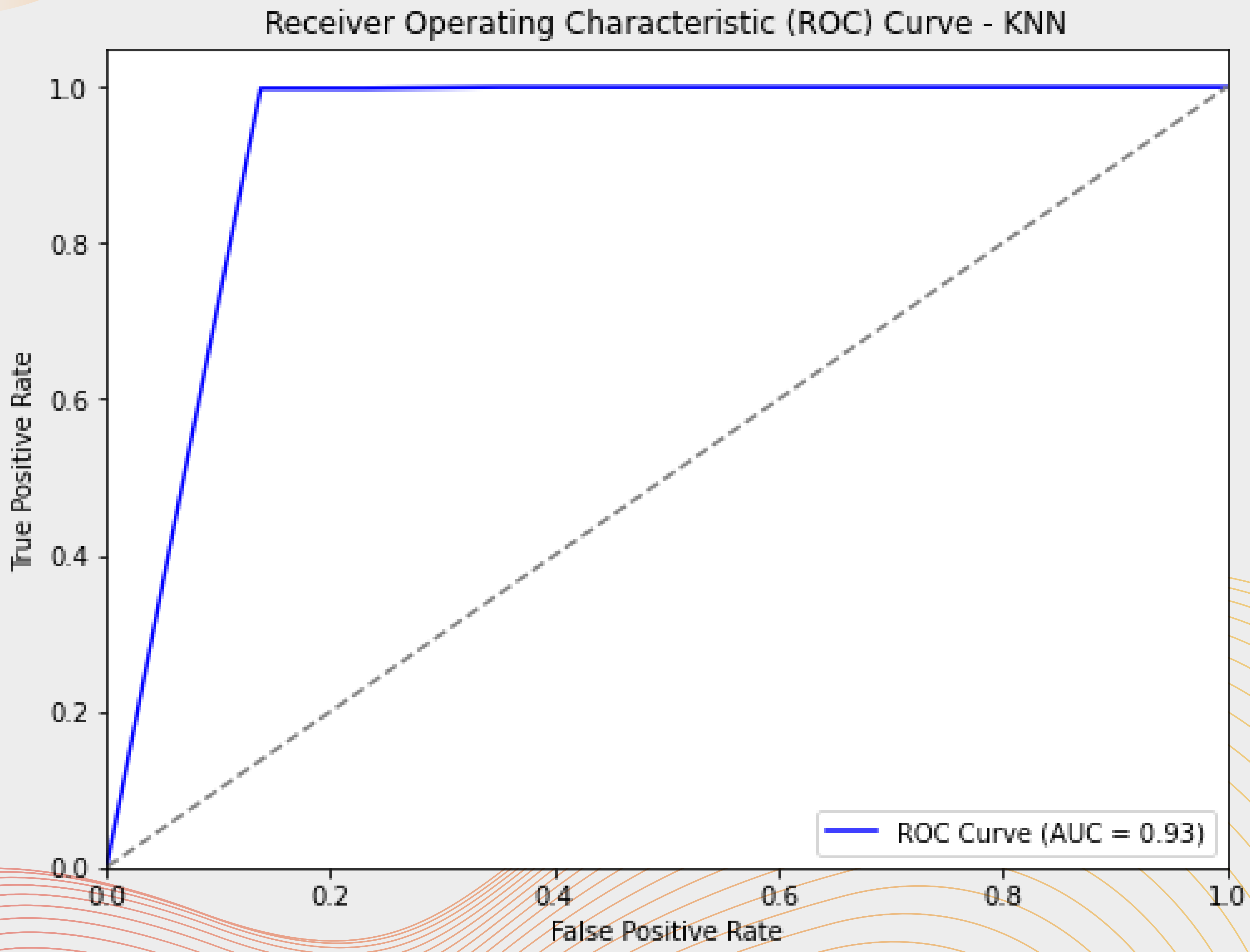


Exceptional model performance:

- **532** True Positives (TP): Correct "Churned" predictions.
- **0** False Negatives (FN): No "Churned" instances missed.
- **6** False Positives (FP): Incorrect "Churned" predictions.
- **502** True Negatives (TN): Correct "Not Churned" predictions.

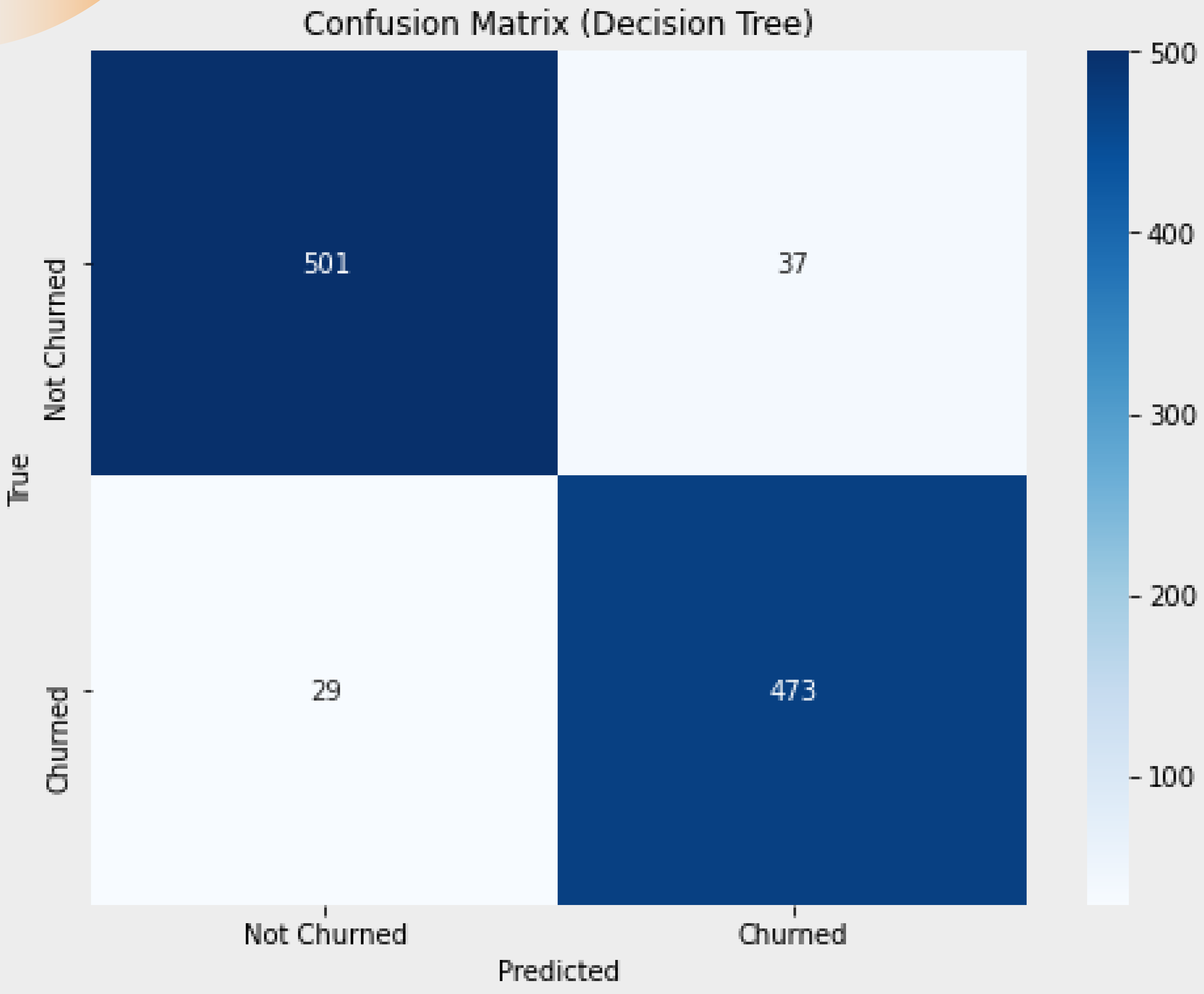
High accuracy score of **0.99.**

Model 2: K-Nearest Neighbors (KNN)



- Accuracy: 0.82, correctly classifying 82% of instances.
- ROC Curve: AUC = 0.93, demonstrating "Churned" vs. "Not Churned" discrimination.
- High "Not Churned" Precision, lower Recall, suggesting potential misses.
- Highlights a trade-off between precision and recall.

Model 3: K-Decision Tree Model



- 473 True Positives (TP): Accurately predicts "Churned" customers.
- 501 True Negatives (TN): Precisely identifies "Not Churned" customers.
- Minimal Errors: 37 False Positives (FP) and 29 False Negatives (FN).
- Balanced Performance: Emphasizes true positives and true negatives, minimizing prediction errors.
- High Accuracy: Achieves an impressive 93.65% overall accuracy.

KNN Tuning:

- Improved performance with an accuracy of approximately 85.96% from 81.83%.
- Achieves a balanced precision of 77.47% and a perfect recall of 100%.
- F1-score of about 87.30% harmonizes precision and recall.
- Overcomes imbalanced precision and recall observed in the original KNN model.

Decision Tree Tuning:

- Maintains high accuracy at approximately 93.65%.
- Demonstrates exceptional precision of around 93.95% for churn prediction.
- A recall of 92.83% effectively captures customers at risk of churning.
- F1-score of about 93.39% showcases a balance between accurate predictions and churn case identification.

Conclusion

The base Logistic Regression model demonstrates outstanding predictive capabilities with perfect scores but raises concerns about potential overfitting. After hyperparameter tuning, the KNN model shows slightly lower accuracy while maintaining high recall, while Decision Tree model offers a balanced performance with strong accuracy, precision, recall, and F1-score.

RECOMMENDATION

Based on our model evaluations, it is recommend that the tuned Decision Tree model be used for accurate predictions of both churned and not churned customers. It effectively captures churned customers while maintaining good precision, ensuring robust and reliable customer churn predictions.

Thank You

ChurnBusters. Inc