ChurnGuard: Predicting Customer Exit

Thomas Amuti

Highlights

- Project Overview
- Business Understanding
  - Business Problem and Questions
- Data Understanding
- Data Preparation
- Results
- Recommendations
- Limitations and Suggestions for further analysis

Project Overview

- Customer churn is a major concern for SyriaTel, a telecommunication company.
- There is however lack of data driven solutions to address this.
- Solutions would help SyriaTel to:
  1. Understand the potential revenue impact of customer churn.
  2. Identify at-risk customers early, enabling proactive retention strategies.
  3. Gain insights into customer behaviours and service factors that influence churn.

Business Questions

1. What is the estimated revenue at risk from customers who churn?

2. Can we accurately predict whether a customer will churn?

3. Which customer behaviours and service features are most strongly associated with churn?

Data Understanding

- The dataset used to find answers to the business questions:
  - ✓ Contains customer-level data from SyriaTel collected over 9 months.
  - ✓ Each row represents a single customer and contains 21 feature columns and 3333 customer records.
  - ✓ The features are generally on:
    - ❑ Customer usage – How customers used the services.
    - ❑ Service plans – Types of services customers subscribed to.
    - ❑ Customer support – Number of customer service calls.
    - ❑ Administrative features – Such as state and duration which customers had accounts.
    - ❑ Identifiers – Customer phone number.

Data Preparation – Used libraries

- We used several libraries to analyse the dataset:
  1. Pandas and numpy- to read our csv files and perform exploratory data analysis.
  2. Matplotlib - for visualizations.
  3. Sklearn – for model fitting and evaluation.
     - Logistic Regression – Baseline predictions
     - Decision Classifier – To capture complex patterns in dataset
     - Random Forest Classifier - To handle high dimensional data

Data Preparation – Data Cleaning

- Prior to data analysis we:
  1. Selected the needed columns.
     - All the features were retained other than phone number which is a unique identifier with no statistical significance as a predictor.
  2. Checked for missing and duplicated data:
     - There were none.

Results

- There were 483 customers who churned and 2850 who did not.

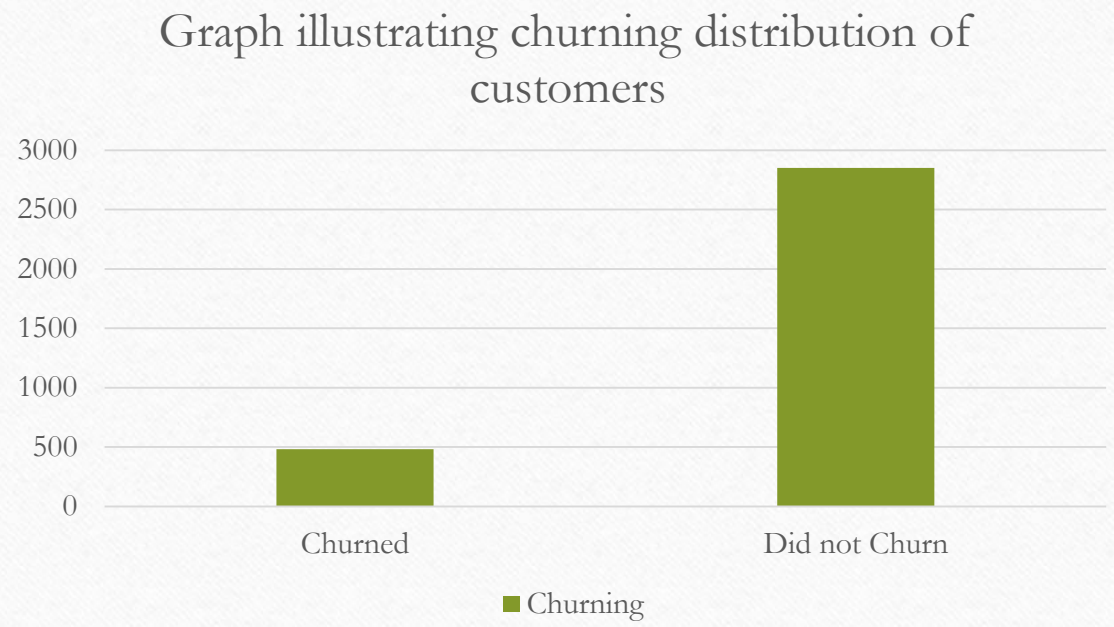
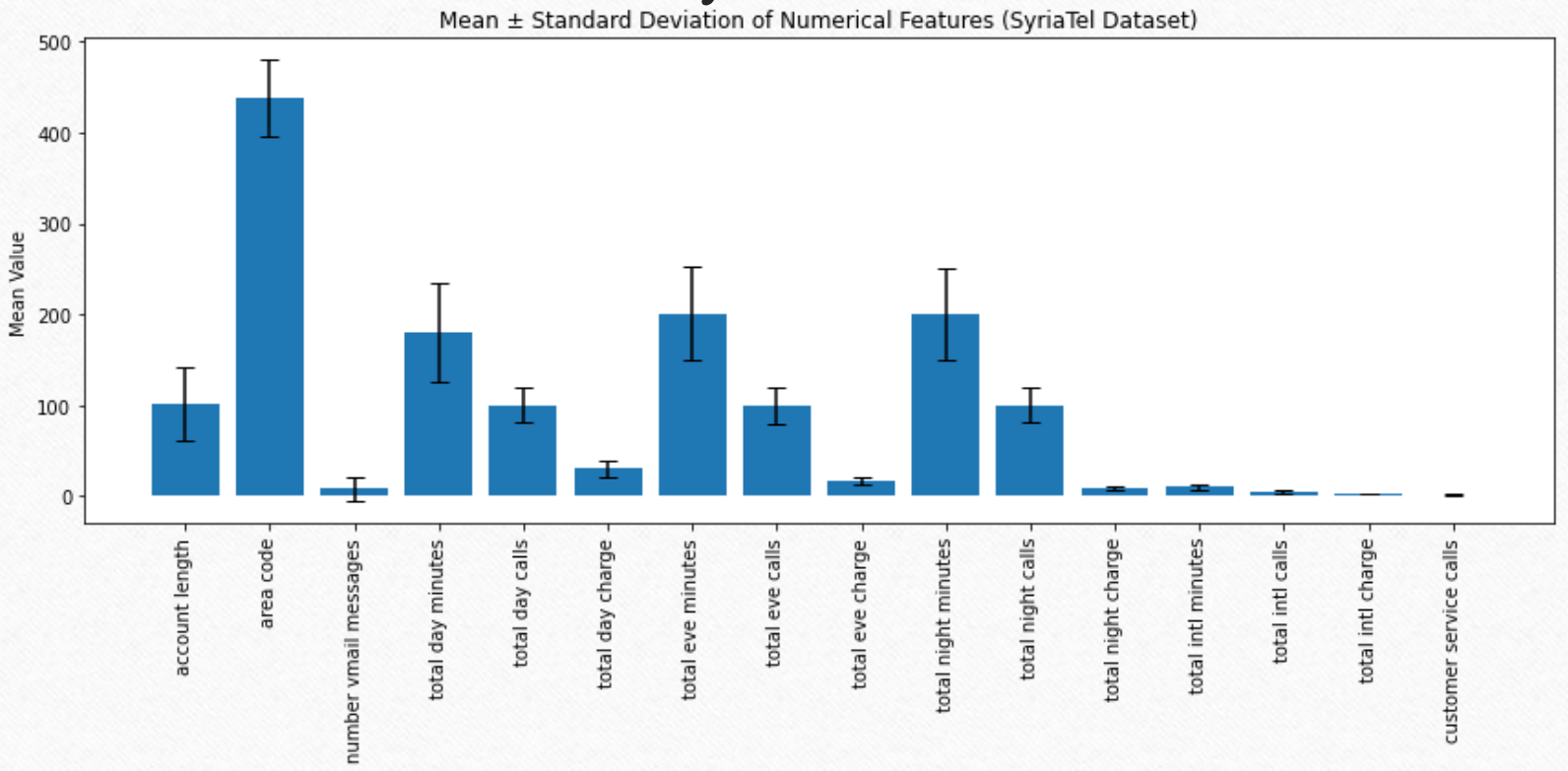


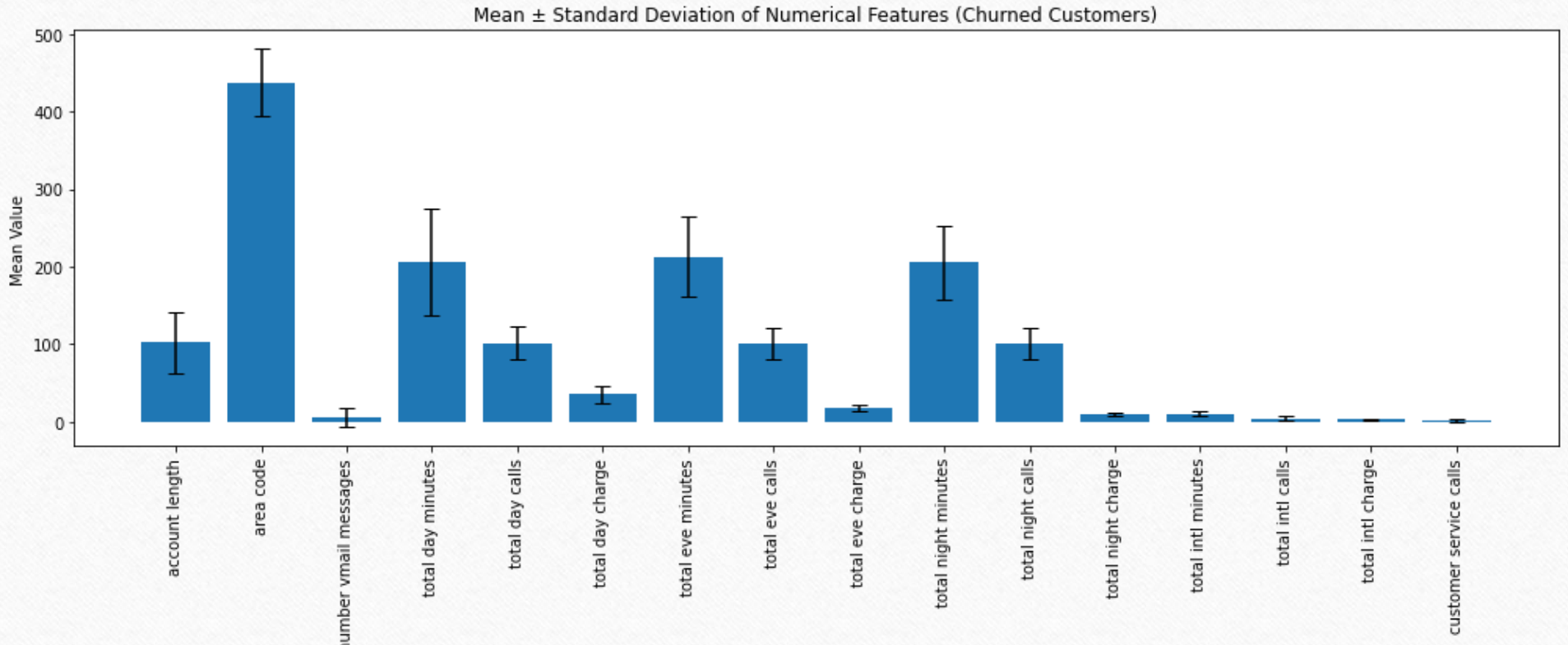
Figure illustrating the mean values of the numerical features of customers irrespective of whether they churned or not



Summary findings of the general categorical features irrespective of whether customer churned or not.

- The most common states were: WV, MN, NY, AL, WI, OH, OR, VA, WY, CT.
- Most customers did not have international plan (3010 versus 323).
- Most customers did not have a voice mail plan (2411 vs 922).

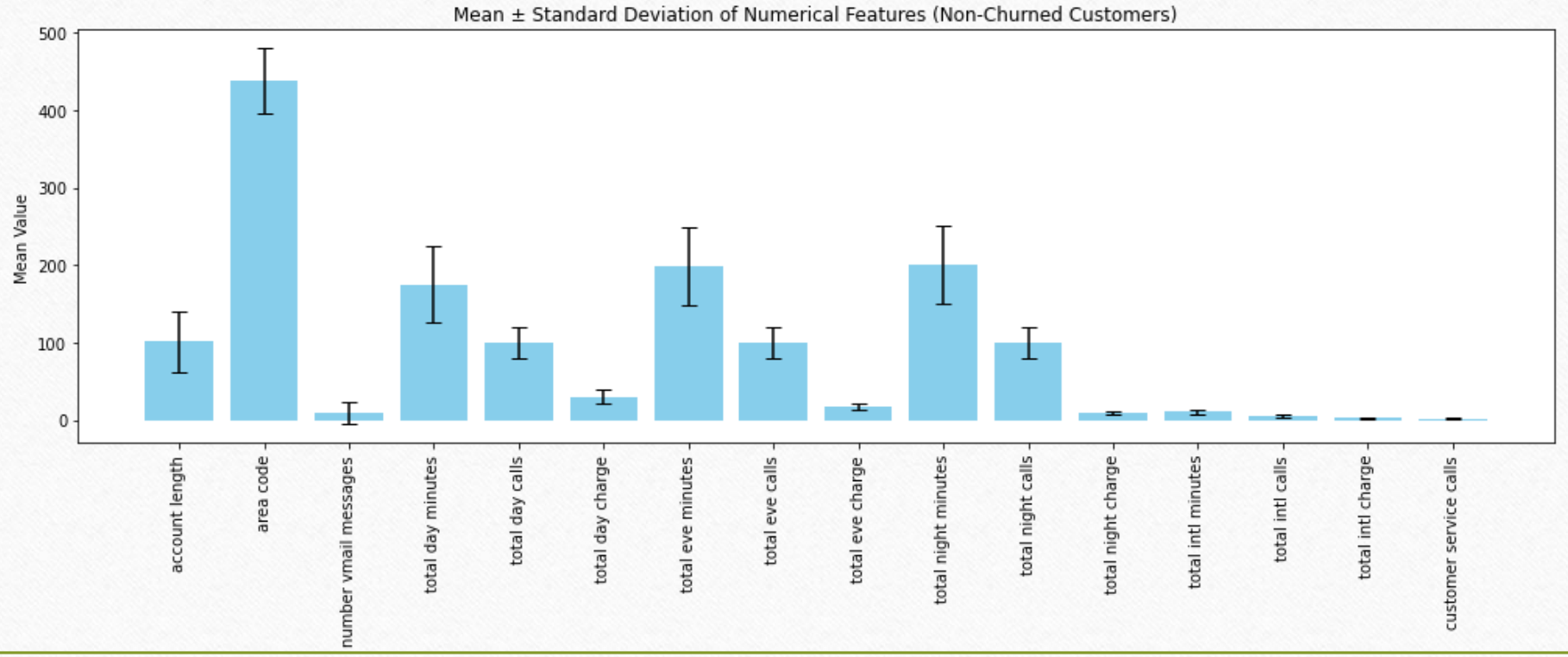
Figure illustrating the mean values of the numerical features of customers who churned



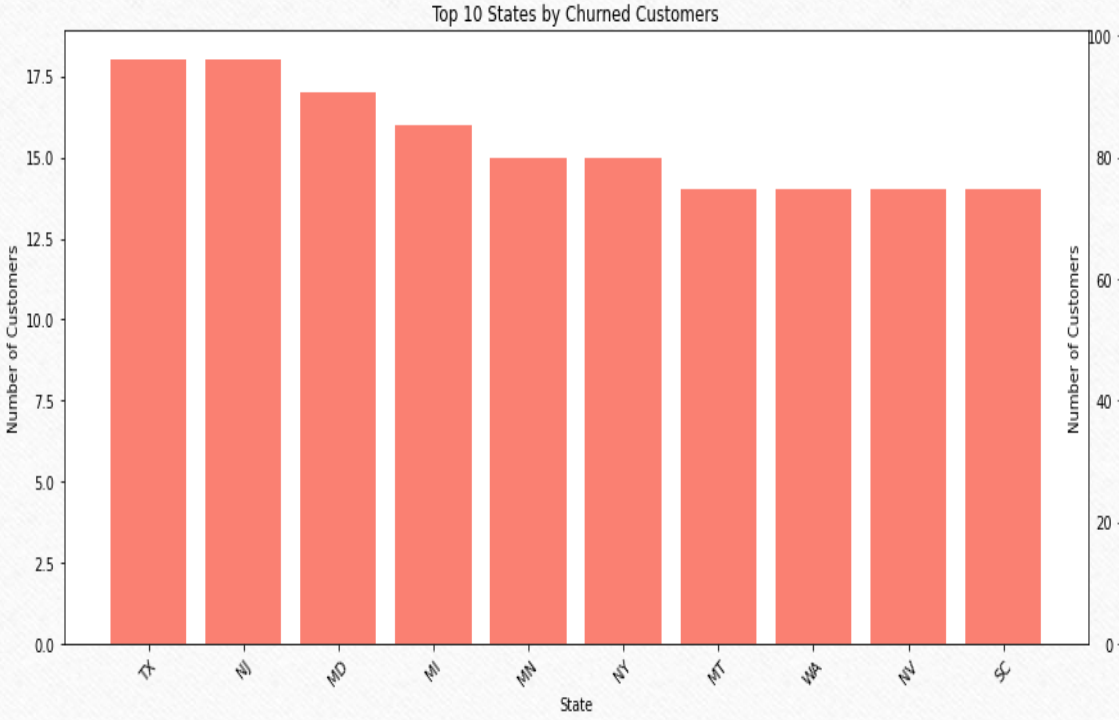
Estimated Revenue at Risk

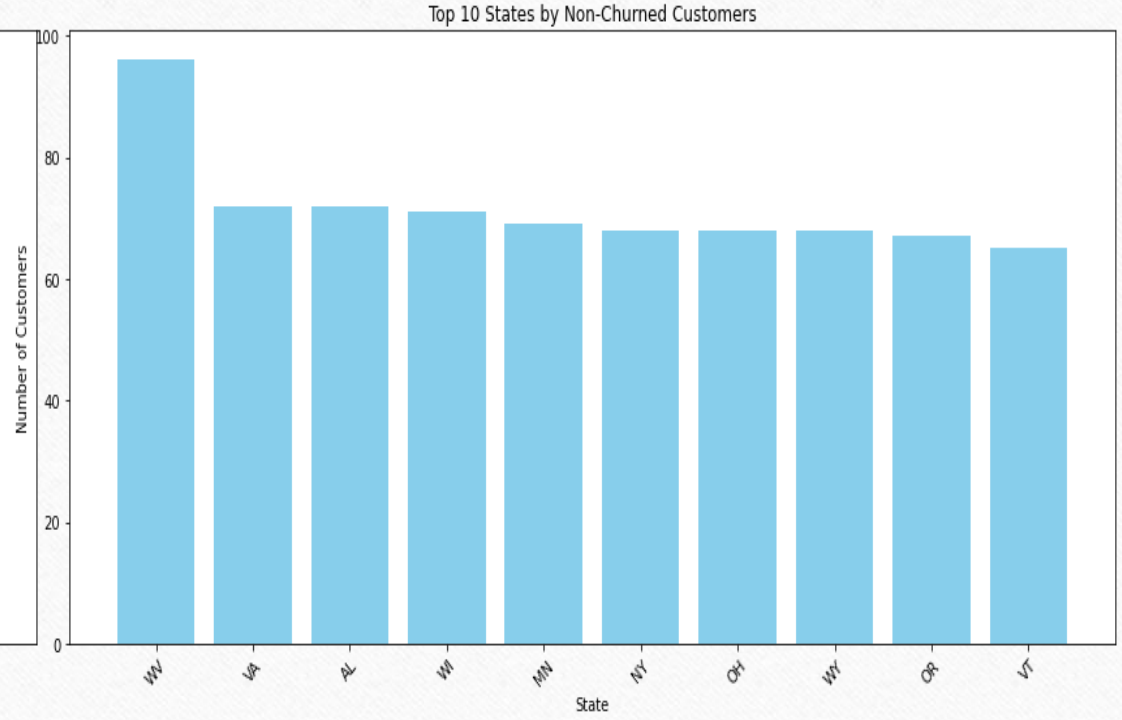
- Charges incurred by those who churned:
  1. Total day charge = 35.17 +/- 9.25
  2. Total evening charge = 18.05 +/- 4.31
  3. Total night charge = 9.23 +/- 2.27
  4. Total international charge = 2.88 +/- 0.75
- Total of charges = 65.33 U.S dollars.
- For a total of 483 customers over 9 months duration = 31554.39 dollars.

Figure illustrating the mean values of the numerical features of customers who did not churn

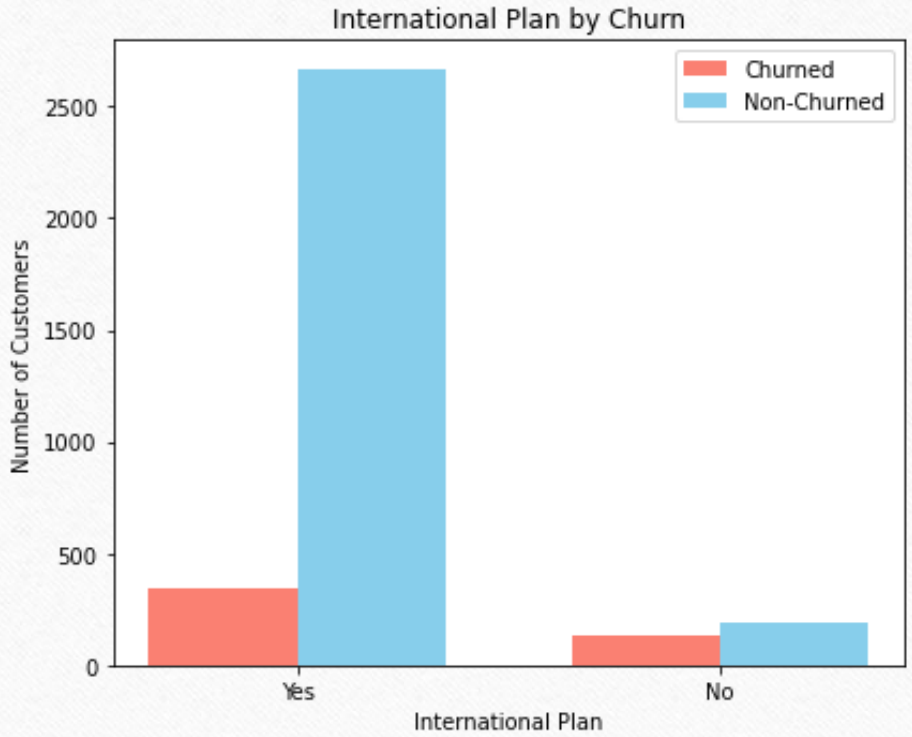


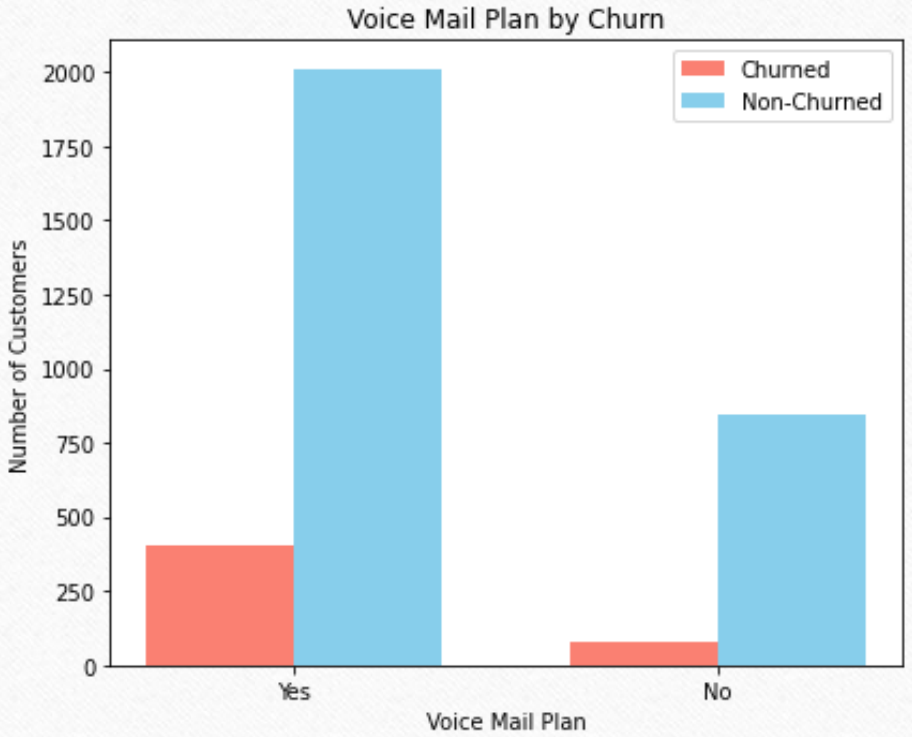
Figures illustrating the 10 most common states for both Churn vs Non-Churn groups





Figures illustrating the differences in 'international plan' and 'voice mail plan' between the churn and non churn groups.





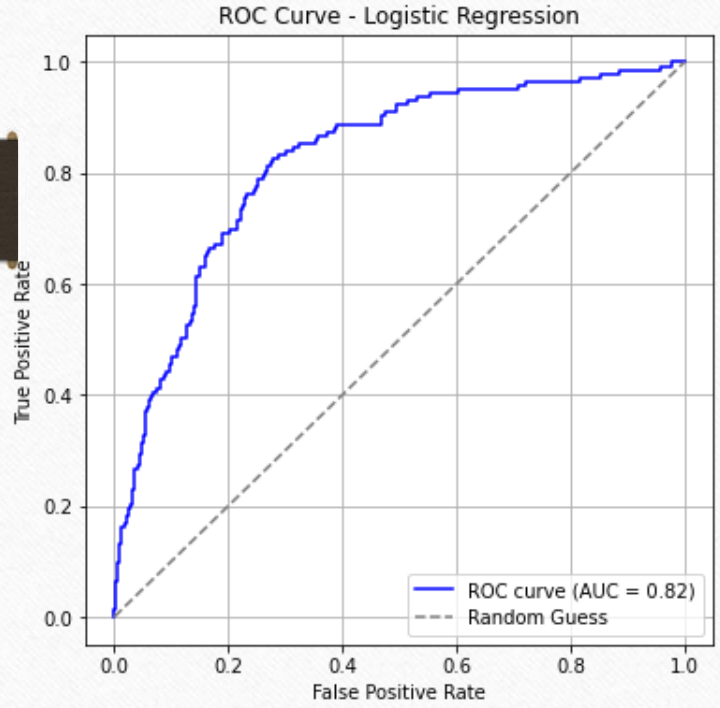
Model Results (Accuracy)

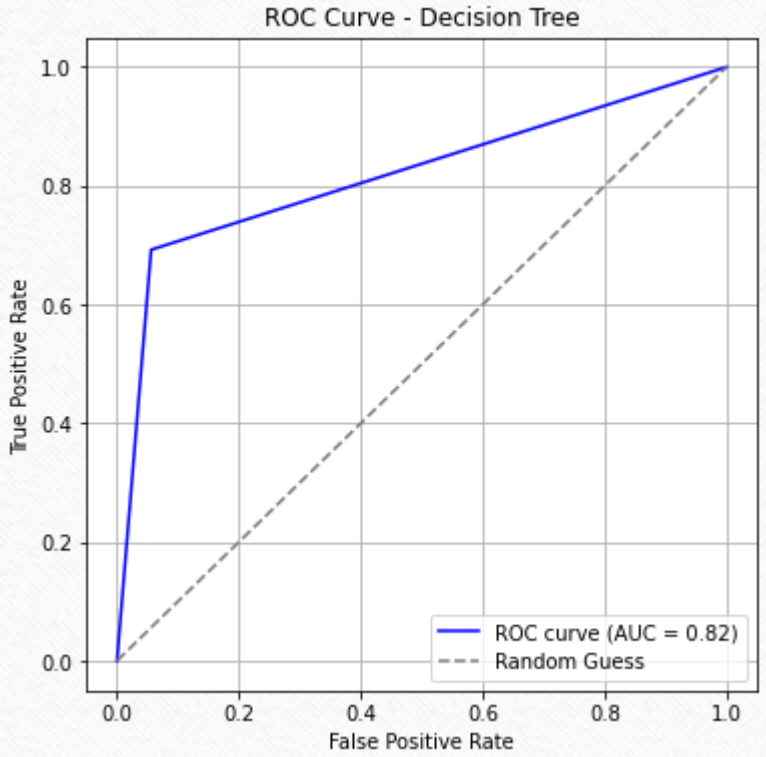
1. Logistic Regression - 0.86 (Testing), 0.87 (Training)
   - Less accurate possibly because it missed out on capturing non-linear patterns.
   - Generalized well and did not over-fit since the accuracy scores for training and testing datasets were almost the same.

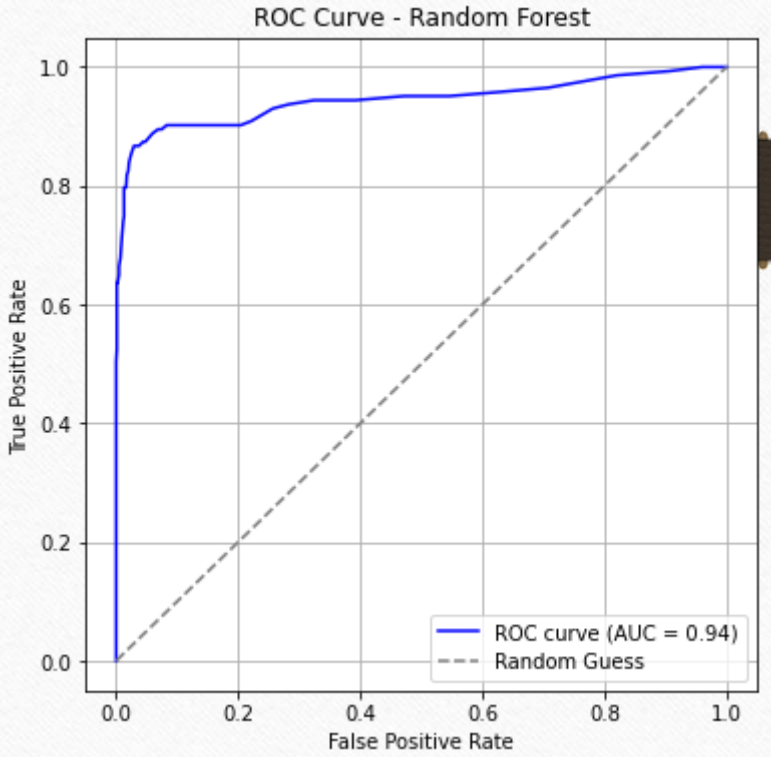
2. Decision Classifier - 0.91 (Testing), 0.87 (Training)
   - More accurate for testing data than Logistic Regression.
   - Testing accuracy was higher than the training accuracy. This could have been due to random variation in the split.

3. Random Forest Classifier - 0.94 (Testing), 0.87 (Training)
   - Best generalization.
   - Has a higher testing accuracy than training accuracy, indicating possible data split randomness.

Model Results – ROC/ AUC







Key features in predicting customer churn

- From the Two Best Models (Decision Tree vs Random Forest), key features were:
  - ✓ Total day charge, total day minutes, total eve charge, total eve minutes, total international charge, total international; calls, total night charge.
  - ✓ Customer service calls, international plan_yes.

Recommendations

- An approximate 16% of customers churned and 31554.39 dollars were lost.

- To avoid this, we recommend:
  - 1. Using the Random Forest Classifier method to predict customer churn.
  - 2. Focusing on the aforementioned key features.

Limitations of the dataset

- Limited time frame:
  - The dataset was collected over a short period, which may not capture seasonal or long-term usage patterns.
- Class imbalance:
  - The number of churned customers was much smaller than non-churned, which can affect model performance and bias predictions toward the majority class.
- Geographic bias:
  - States are included, but the dataset may not represent the full national population or capture regional variations accurately.

Suggestions for further analysis

- Adding more customer attributes:
  - Demographics: age, gender, income, occupation.
  - Subscription details: contract type, tenure, plan changes, add-ons.
  - Service quality metrics: dropped calls, network coverage, complaint history.
- Increasing dataset size and diversity:
  - Collect data across multiple time periods and regions to improve model generalization.
  - Balance the dataset better between churned and non-churned customers to reduce class imbalance.