thomasamuti-cpu / **customer_project**

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Watch 0

This github repository contains my submission for the Moringa Phase 3 project on predicting whether a customer will soon stop doing business with SyriaTel, a telecommunications company. This is a binary classification problem.

☆ **0** stars **0** forks **0** watching Branches Activity Tags

Public repository

1 Branch 0 Tags Go to file t Go to file Add file + Code ···

AmutiMombo Commits History added e412c02 · now

Figures	Data Evaluation, Recommendations,…	52 minutes ago
Pdfs	Commits History added	now
.gitignore	Data Evaluation, Recommendations,…	52 minutes ago
ChurnGuard.pptx	README added	11 minutes ago
README.md	README added	11 minutes ago
customer.csv	First Commit - Python notebook wit…	2 days ago
customer.ipynb	Data Evaluation, Recommendations,…	52 minutes ago

README

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.

It was sourced from Kaggle (https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset?resource=download).

Each row represents a single customer and contains 21 feature columns and 3333 customer records.

- The features are generally on:

1. Customer usage – How customers used the services.

2. Service plans – Types of services customers subscribed to.

3. Customer support – Number of customer service calls.

4. Administrative features – Such as state and duration which customers had accounts.

5. 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.

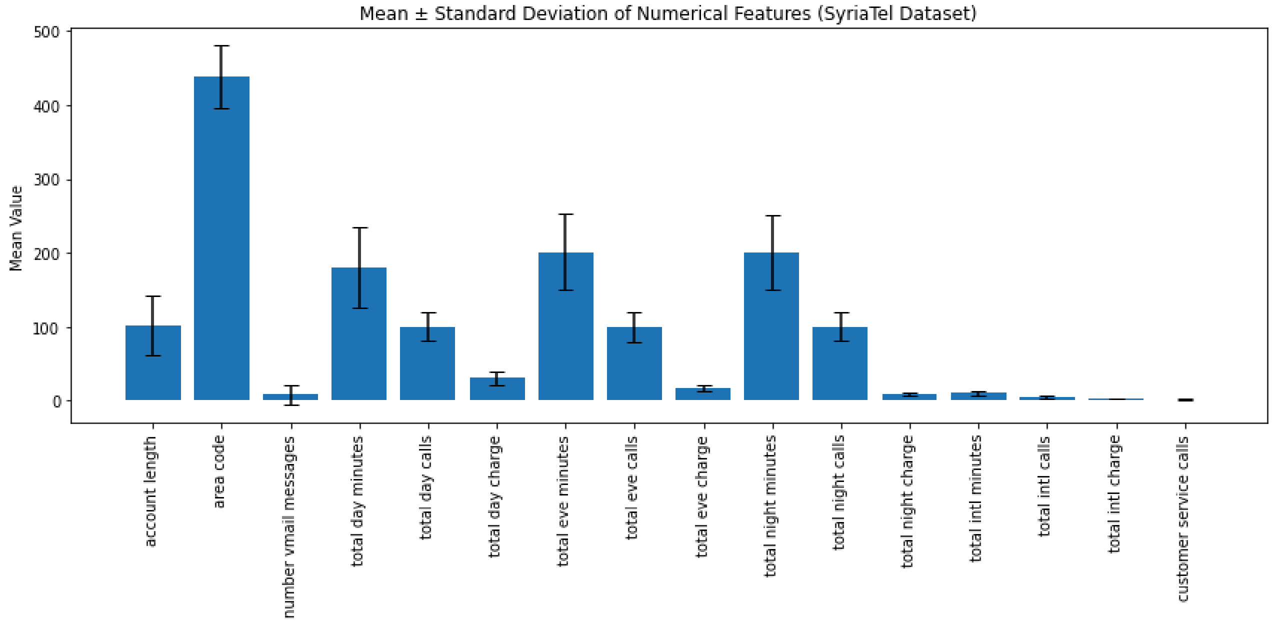
2. Checked for missing and duplicated data:

- There were none.

Results

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

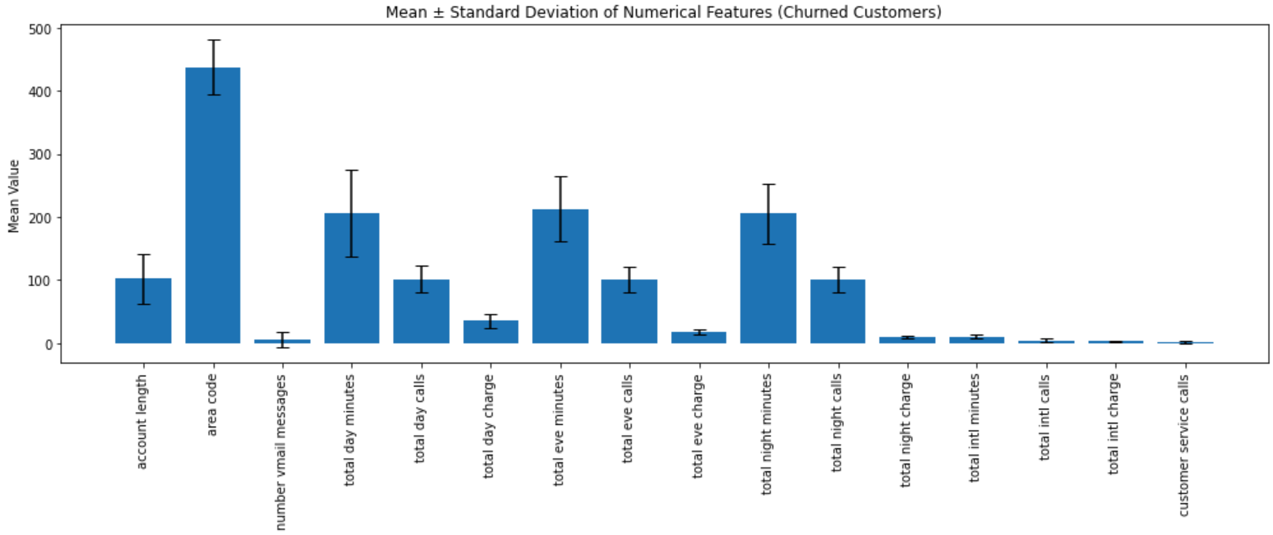
The mean values of the numerical features of customers irrespective of whether they churned or not is as shown:



Summary findings of the general categorical features irrespective of whether a customer churned or not is as shown:

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

The mean values of the numerical features of customers who churned was as shown:

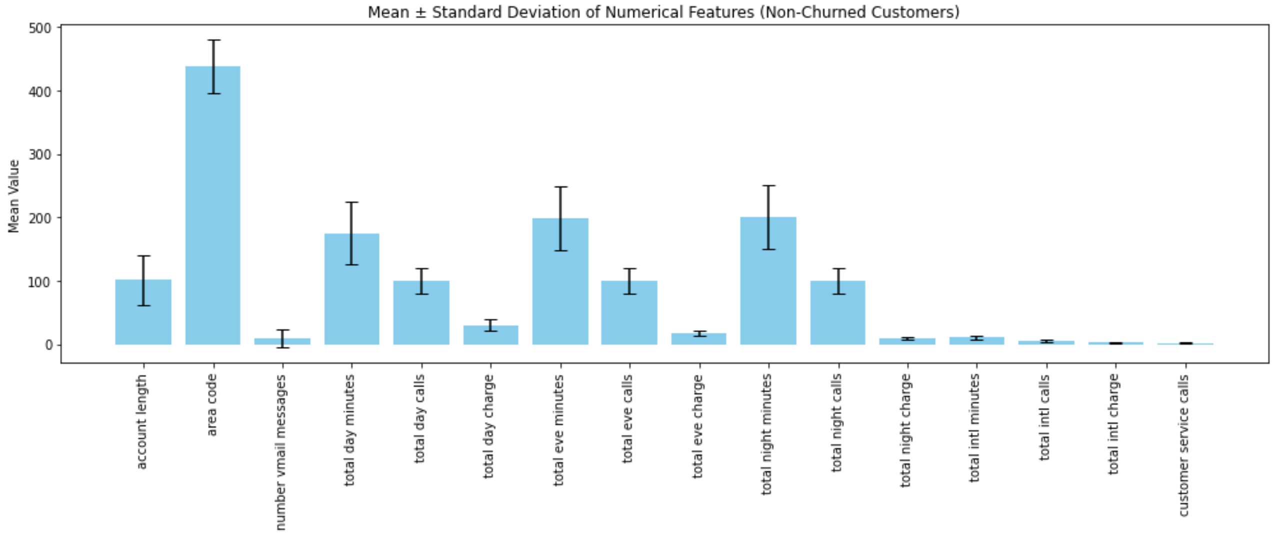


Objective 1: Potential revenue lost by those who churned:

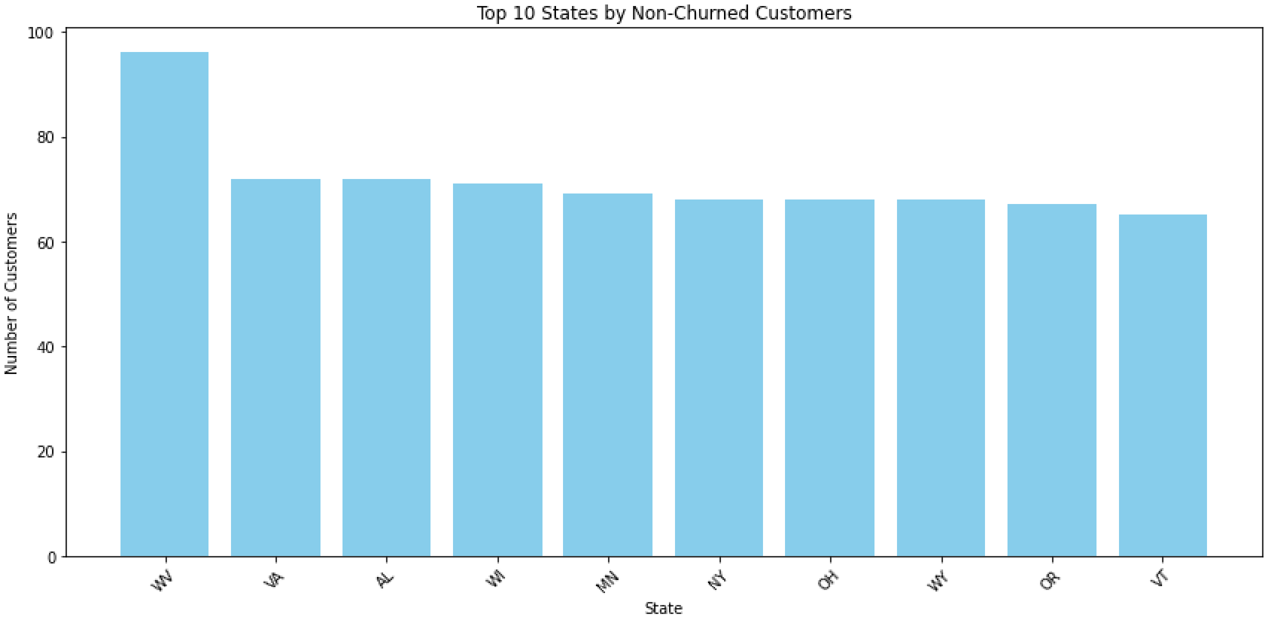
1. Total day charge = 35.17 +/- 9.25
2. Total evening charge = 18.05 +/- 4.31
3. Total night charge = 9.23 +/- 2.27
4. Total international charge = 2.88 +/- 0.75

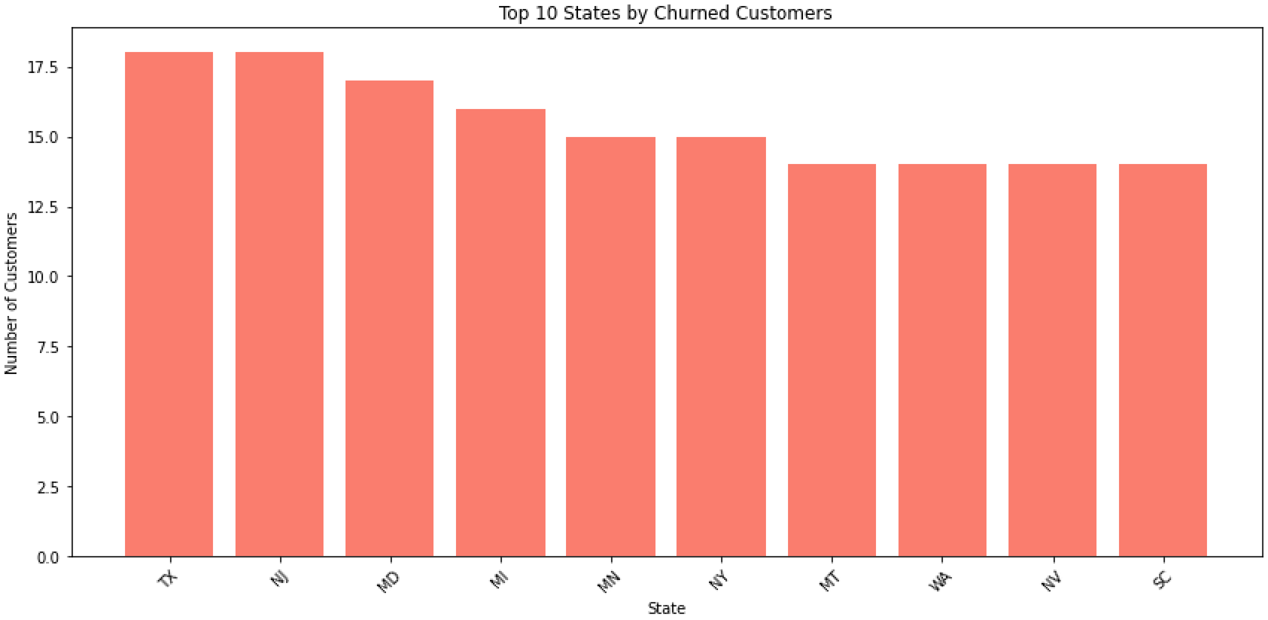
Therefore the total charges per customer = 65.33 U.S dollars. For a total of 483 customers over 9 months duration, the total loss was 31554.39 dollars.

The mean values of the numerical features of customers who did not churn was as shown:

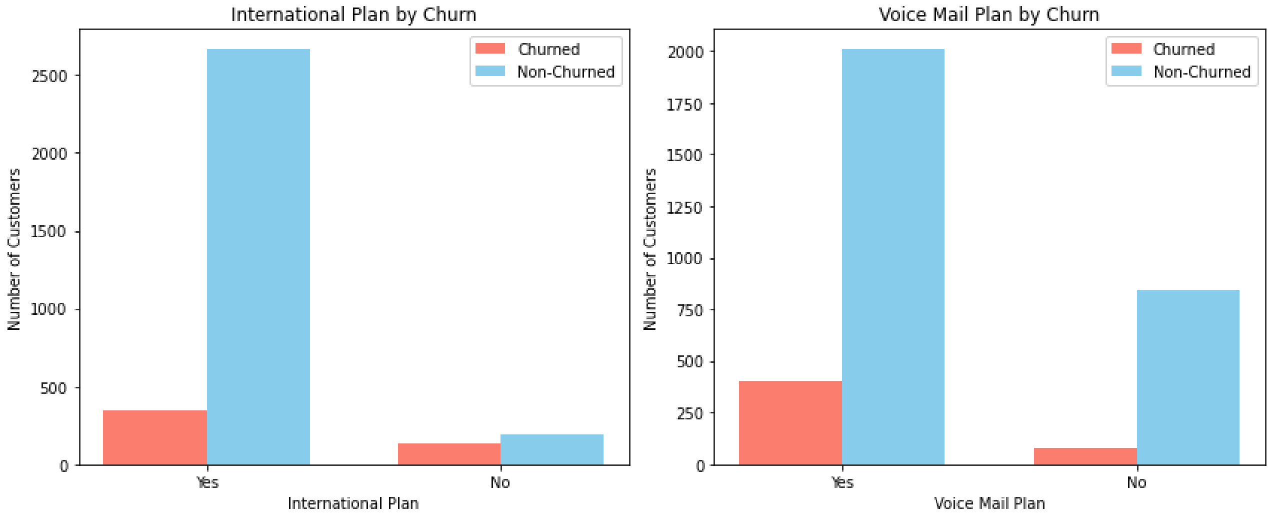


The distribution of states based on whether customers churned or not was as shown:





The differences in 'international plan' and 'voice mail plan' between the churn and non churn groups was demonstrated as shown:



Objective 2: As pertain model accuracy in predicting training and testing sets:

1. Logistic Regression - 0.86 (Testing), 0.87 (Training)

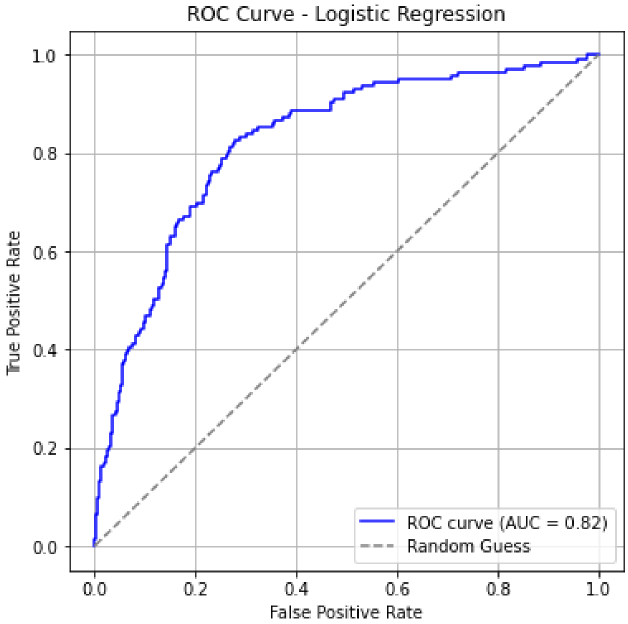
- Less accurate possibly because it missed out on capturing non-linear patterns Generalized well and did not over-fit since the accuracy scores for training and testing datasets were almost the same.

2. Decision Classifier - 0.91 (Testing), 0.87 (Training)

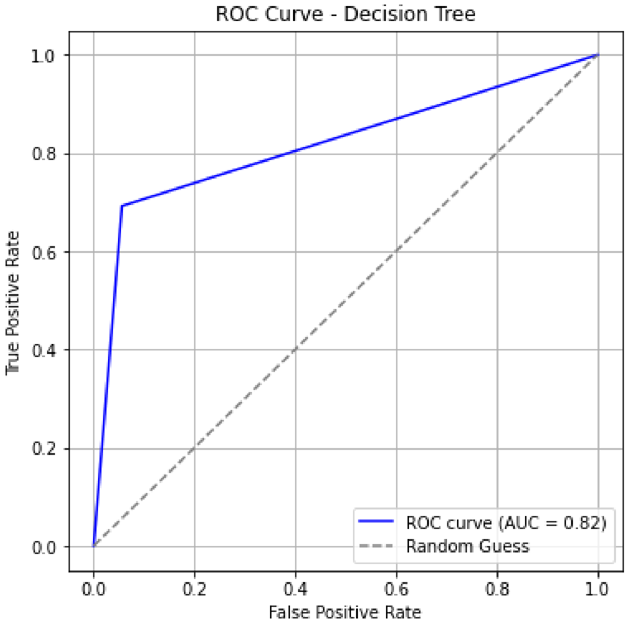
- More accurate for testing data than Logistic Regression. Testing accuracy was higher than the training accuracy. This could have been due to random variation in the split.

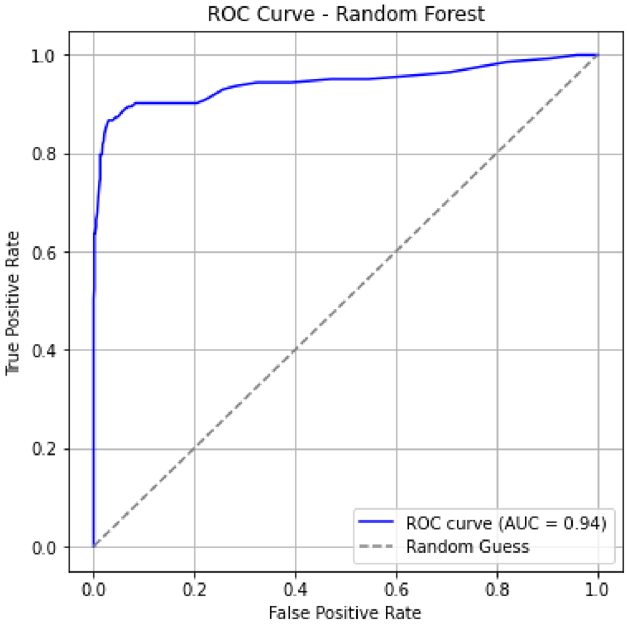
3. Random Forest Classifier - 0.94 (Testing), 0.87 (Training)

- Best generalization. Has a higher testing accuracy than training accuracy indicating possible data split randomness.



The ROC/AUC of the models are as shown:





Objective 3: The shared key features that best predicted customer churn based on the 2 best performing models (Decision Classifier and Random Forest Classifier) 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

1. Limited time frame:

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%