

Phase 3 Project

By Lucy Munge


Phoenix_dspt04





سورية يقتل
SYRIATEL

SyriaTel Customer Churn Prediction



The core objective of this study is to predict churn in advance and pinpoint the primary factors that may influence customers to migrate to other telecom providers.

The questions this project seeks to answer include:

- What are the factors that are contributing to customer churning?
- What attributes do the customers who churn have?
- How can SyriaTel increase customer retention?

Project Overview



1

Business Problem

2

Data Loading and Understanding

3

Data Preparation

4

Modeling

5

Evaluation

6

Model of Choice: Decision
Trees

7

Recommendations and
Future Investigations

7

Conclusion

Project objective



The main objective for this project is to build predictive model with 85% accuracy that will help to:

- Gain insight on the factors that are contributing to customer churning
- Gain insights on the attributes the customers who churn have
- Increase customer retention

1. Business Problem

I have been tasked by SyriaTel, a Telecommunication company, to build a classification model that will predict whether a customer will soon stop doing business with Syria and their main interest is in reducing the amount of money that is lost because of customers who don't stick around very long.



2. Business Questions




The questions this project seeks to answer include:

- What are the factors that are contributing to customer churning?
- What attributes do the customers who churn have?
- How can SyriaTel increase customer retention?



2. Data Loading and Understanding

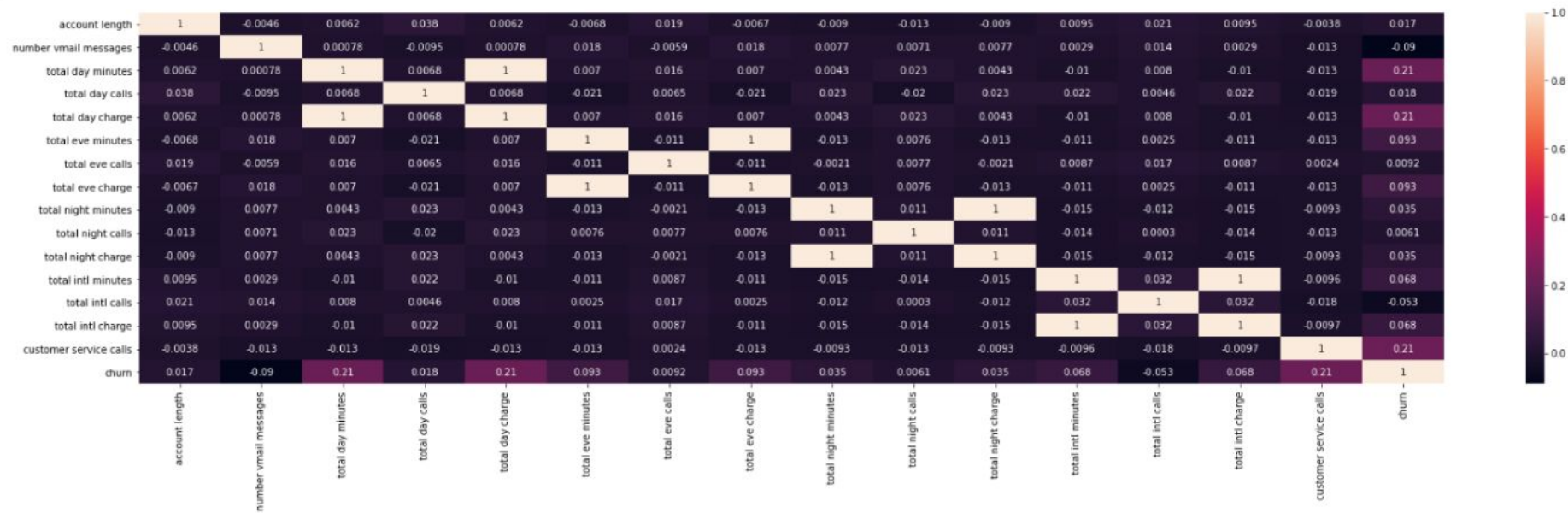


During our EDA, there are some findings that have come up and need to be addressed before getting into modeling. Below are the findings:

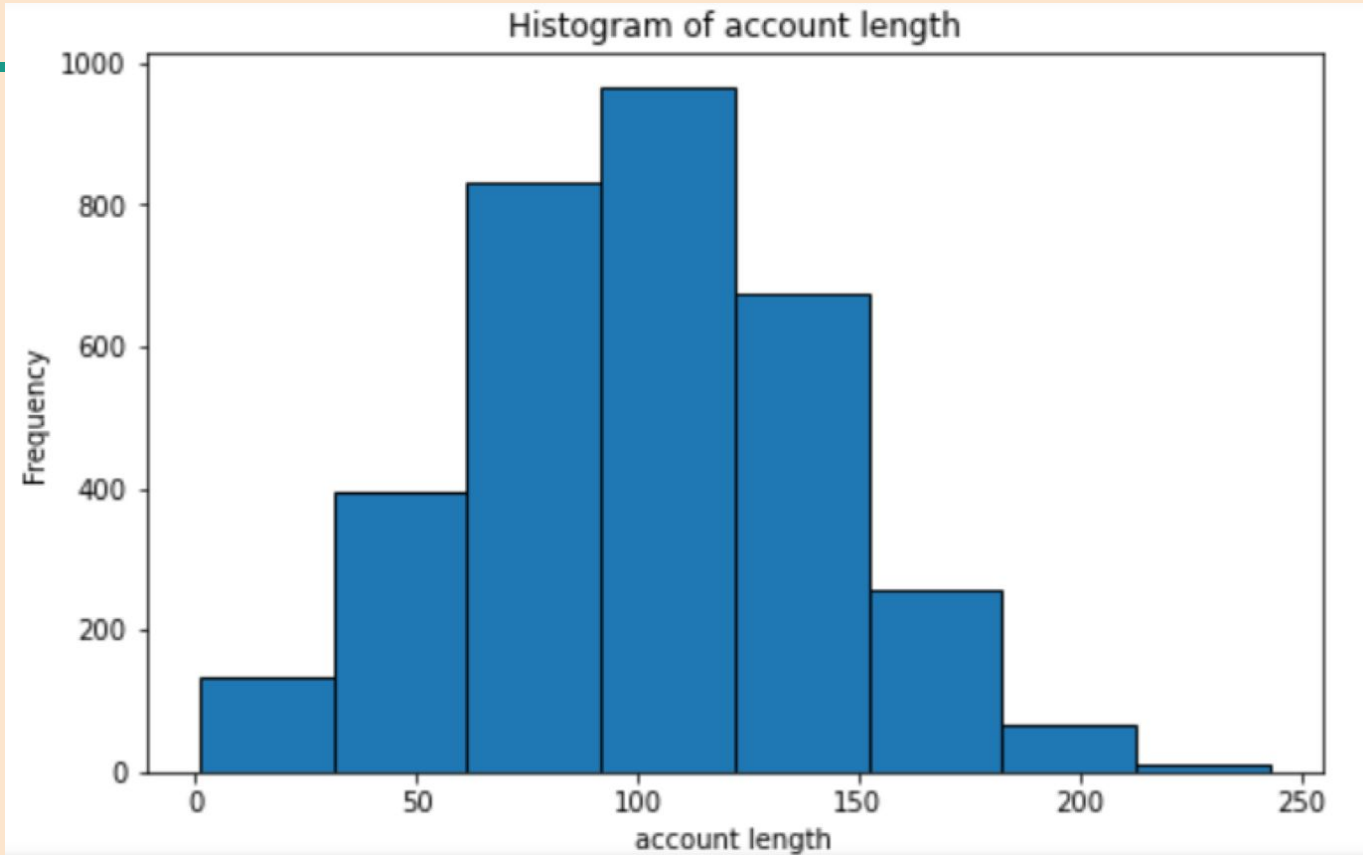
Finding 1: Converting data type - an important data type issue related to the 'area code' column. Although it is represented as an integer in the dataset, the values it contains are essentially placeholders or labels, not numerical values that carry mathematical significance.

Finding 2: High correlation - Multicollinearity - Our examination of the heatmap representation of the data revealed that several columns exhibit high levels of correlation with each other. This observation indicates the presence of multicollinearity, a condition where independent variables in our dataset are highly interrelated.


Heatmap to check how the columns are correlated



Histograms for numerical features to visualize their distribution

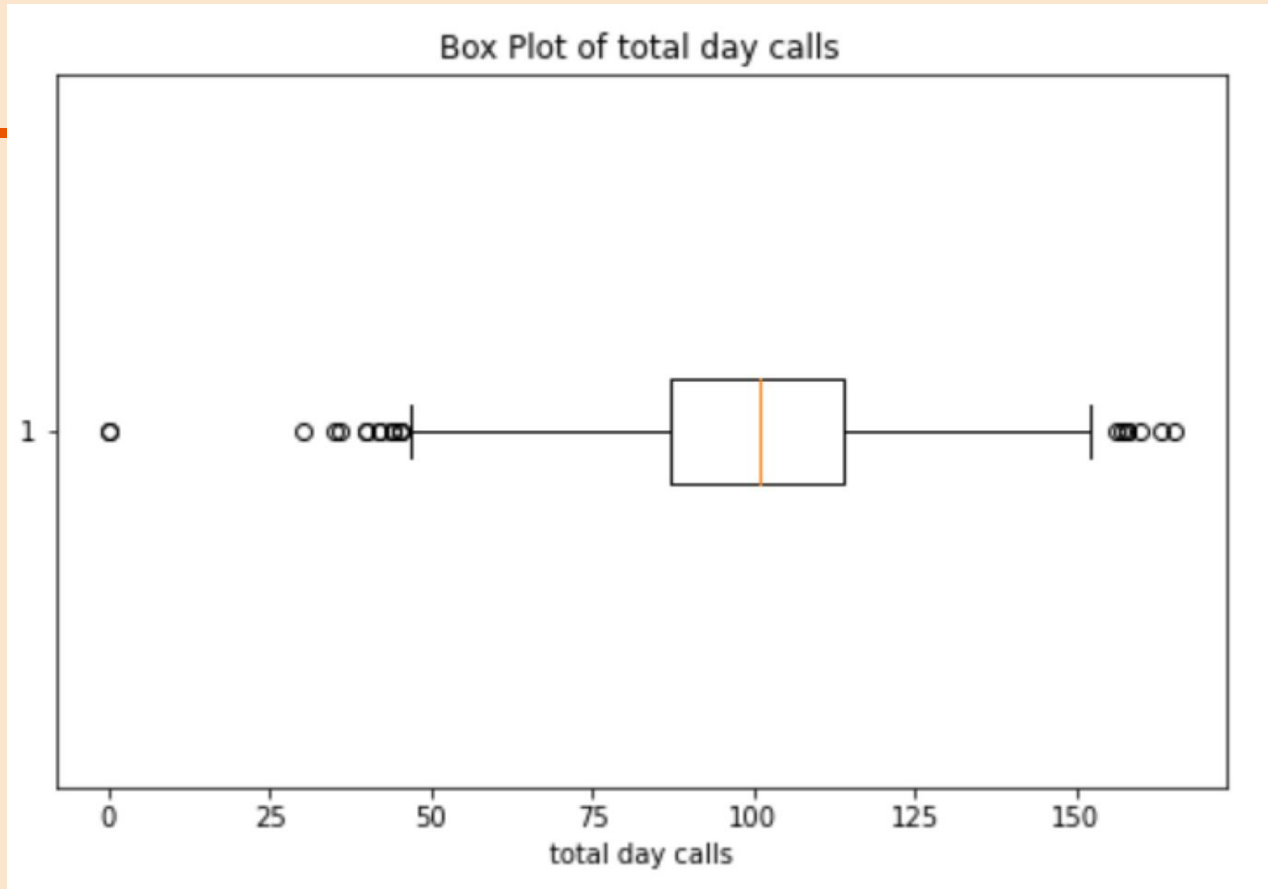


Findings:

A horizontal bar with a teal segment on the left and an orange segment on the right.

Finding 3: Outliers - we observed the presence of a significant number of outliers in our dataset, as indicated by the boxplots. These outliers have the potential to impact our modeling process. However, it is important to note that, in this case, these outliers are not anomalies that should be removed. Instead, they are a noteworthy aspect of our dataset that we should be aware of during our modeling process.

Box plots to identify outliers and visualize the spread of data

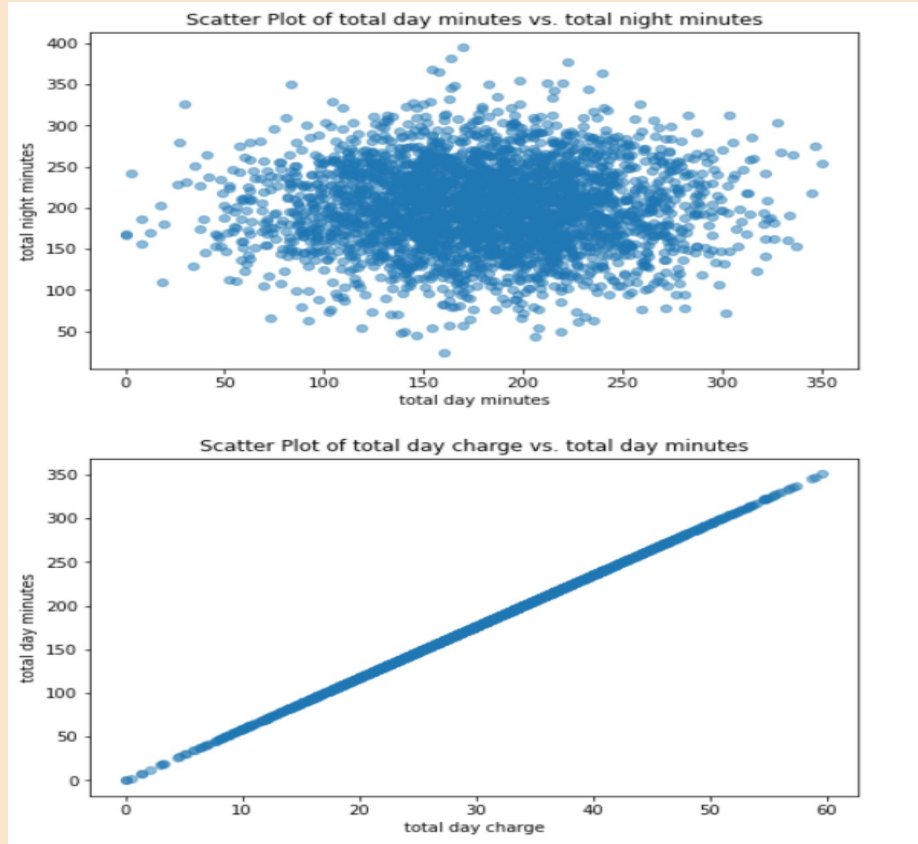




Finding 4: Scatter plots showing multicollinearity

Our analysis of the scatter plots has revealed the presence of features that exhibit a perfect correlation with each other. This perfect correlation is a clear indicator of multicollinearity, a condition where independent variables in our dataset are highly interrelated, possibly to the extent that they move in perfect synchronization. Multicollinearity poses a significant challenge to our modeling process and can lead to less reliable statistical inferences.

Finding 4: Scatter plots showing multicollinearity




3. Data Preparation



Below are some of the steps followed during the data preparation and some more findings:

- **step 1: Removing irrelevant columns** - Given that our dataset does not contain any missing values and duplicates, the next step is to streamline our data by removing columns that are not essential for our analysis or modeling.
- **Step 2. Feature Engineering** - In our dataset, we have identified that both our target variable and certain feature columns are categorical in nature. To effectively use this data in our modeling process, it is advisable to encode these categorical variables into a numerical format.

Finding 5: Encoding



Encoding categorical variables is a crucial step, as many machine learning algorithms require numerical inputs for model training. By performing appropriate encoding techniques, such as one-hot encoding or label encoding, we can convert the categorical values into a numerical representation that the model can understand.

Finding 6: Imbalanced data


The 'churn' column represents a binary outcome, where 'False' represents customers who did not churn (i.e., stayed), and 'True' represents customers who did churn (i.e., left). The 'churn_encoded' column is a numerical representation of 'churn,' where 0 typically corresponds to 'False,' and 1 corresponds to 'True.' The majority of the data (about 85.51%) falls into the 'False' or 0 category, while the minority of the data (about 14.49%) falls into the 'True' or 1 category. This indicates that the dataset might be imbalanced, with a higher proportion of non-churned customers.

Step 3. Choosing the Target and the Features



We've chosen the relevant features and the column churn as our target for our models.

4. Modeling



Develop a predictive model designed to anticipate whether a customer is on the verge of discontinuing their engagement with Syria. The primary objective is to curtail financial losses stemming from customers who have a short-lived association with the entity.

Model Used:

1. Logistic Regression
2. Decision Trees
3. KNN Classifier Model

5. Evaluation



- The logistic regression model shows a balanced performance with reasonably good accuracy, precision, recall, and F1 score. It captures positive cases effectively while maintaining precision. The ROC AUC score is also decent.
- The decision tree model exhibits high accuracy, especially for the majority class. However, it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class. The model is well-suited for imbalanced datasets.
- The KNN classifier model achieves decent accuracy and performance for the majority class but struggles with the minority class, similar to the decision tree model.

6. Model of Choice: Decision Trees



From the above models, We've chosen to use the Decision Trees Model.

- The decision tree model was evaluated with precision, recall, and F1 score for both Class 0 and Class 1.
- From the accuracy results above our model correctly predicts the class labels for the majority of instances in the test data. The precision metric is very important as it measures how accurate the model is at identifying the majority class which is the customers who don't churn.
- The downside of our model is that it has lower precision, recall, and F1 score for the minority class. This suggests that it may not perform as well on classifying the minority class, the same way it does the majority class. But this is majorly attributed to the fact that our data is also highly imbalanced. But compared to the other two models, this ones performs much better.
- Our model performs well in terms of precision and recall for both classes on the holdout test data, thus it can be deployed in a real-world scenario.

7. Recommendations and Future Investigations

- **Customer Service Calls Investigation:** Dig deeper to understand why some customers need to contact customer service frequently. This will help in finding ways to better assist them.
- **International Plan Churn Investigation:** Since some of the customers with international plans are leaving, it's essential to explore ways to retain these customers.
- **High Churn States Analysis:** Look into the states where many customers are leaving to identify any patterns or reasons for the high churn rates.
- **Incentives for High Bill Customers:** Find ways to encourage customers with high daily charges (over \$55) to stay with SyriaTel. This might involve offering extra benefits and perks. Currently, all of these high-bill customers are leaving, which is a concern.
- **Incentives for Customers who stay more than 6 months:** Find ways to encourage customers to stay with the company even longer, eg giving them loyalty points, offers, etc, as this will help in creating a form of loyalty.

Conclusion



In conclusion, the decision tree model appears to be providing reasonably good results, especially for the majority class, which is typical for imbalanced datasets and is the best option out of the three models built above.



Thank you!!!.