SYRIATEL CHURN STUDY

CONTENT

- Business & Data Understanding
- Exploratory Data Analysis
- Modeling
- Conclusions & Recommendations
- Next Steps

BUSINESS UNDERSTANING

The marketing team in syriatel would like to understand churn trends help them become more competitive against competition. This will help to improve their customer acquistion and retention strategy.

OUR ROLE

Understanding the reasons behind customer churn

Build a prediction model to help proof the business against churn

DATA UNDERSTANDING

Total Minutes

Call duration in minutes – day, evening, night, international

Total Calls

Total number of calls – day evening, night international

Total Charge

Cost of phone call - day evening, night international

Customer Service Calls

Number of calls to customer service

State

Region customer is located in

Phone number

Customer’s phone number

International Plan

Whether or not customer has an international plan

Voicemail Plan

Whether or not customer has a voicemail plan

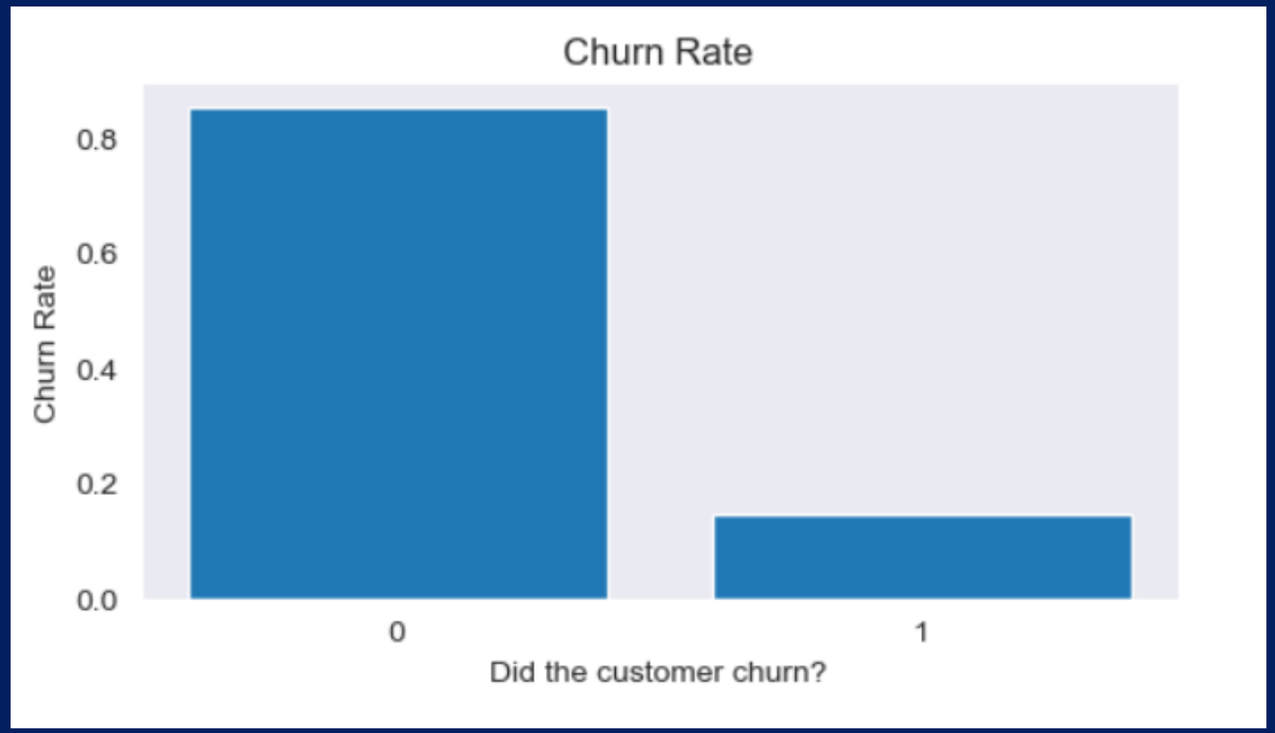
Churn

Whether or not customer has churned

DATA ANALYSIS

CHURN RATE

The data shows a churn rate of 14%, meaning that our target variable is imbalanced. We will therefore have to correct for the imbalances when modeling

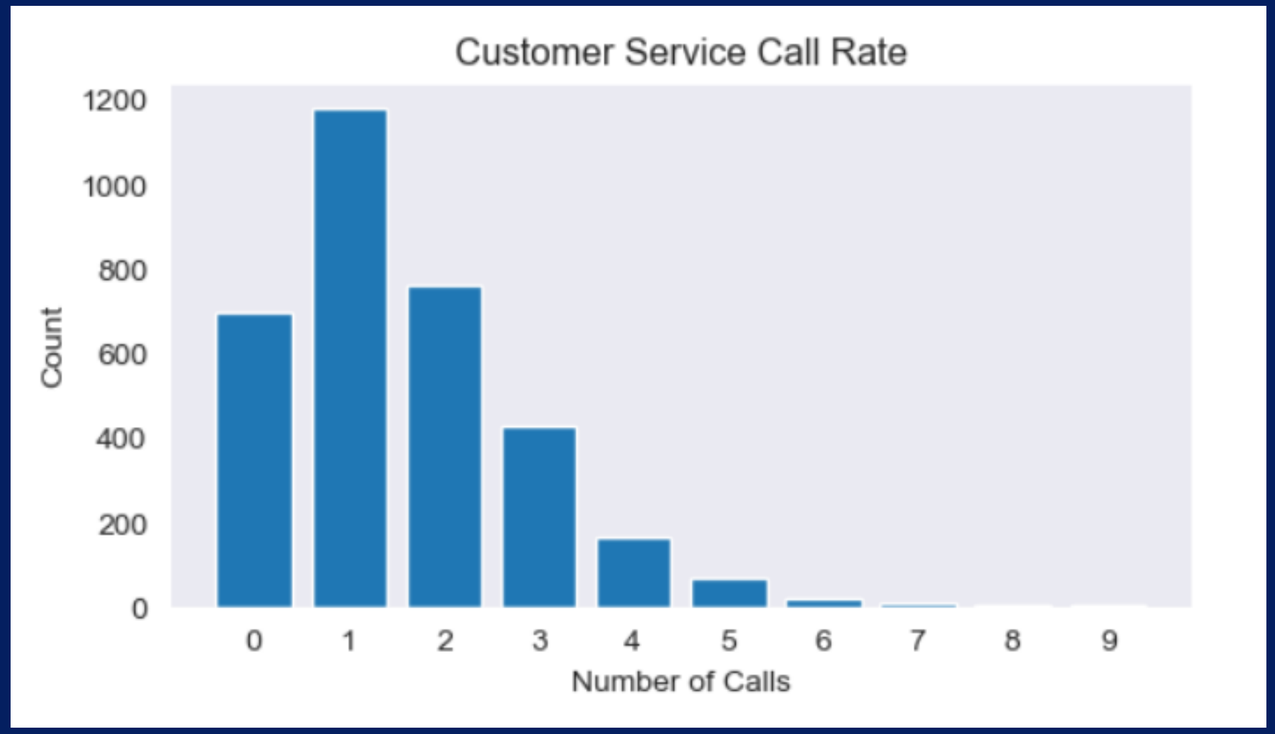


GEOGRAPHIC DISTRIBUTION

The data shows a churn rate of 14%, meaning that our target variable is imbalanced. We will therefore have to correct for the imbalances when modeling

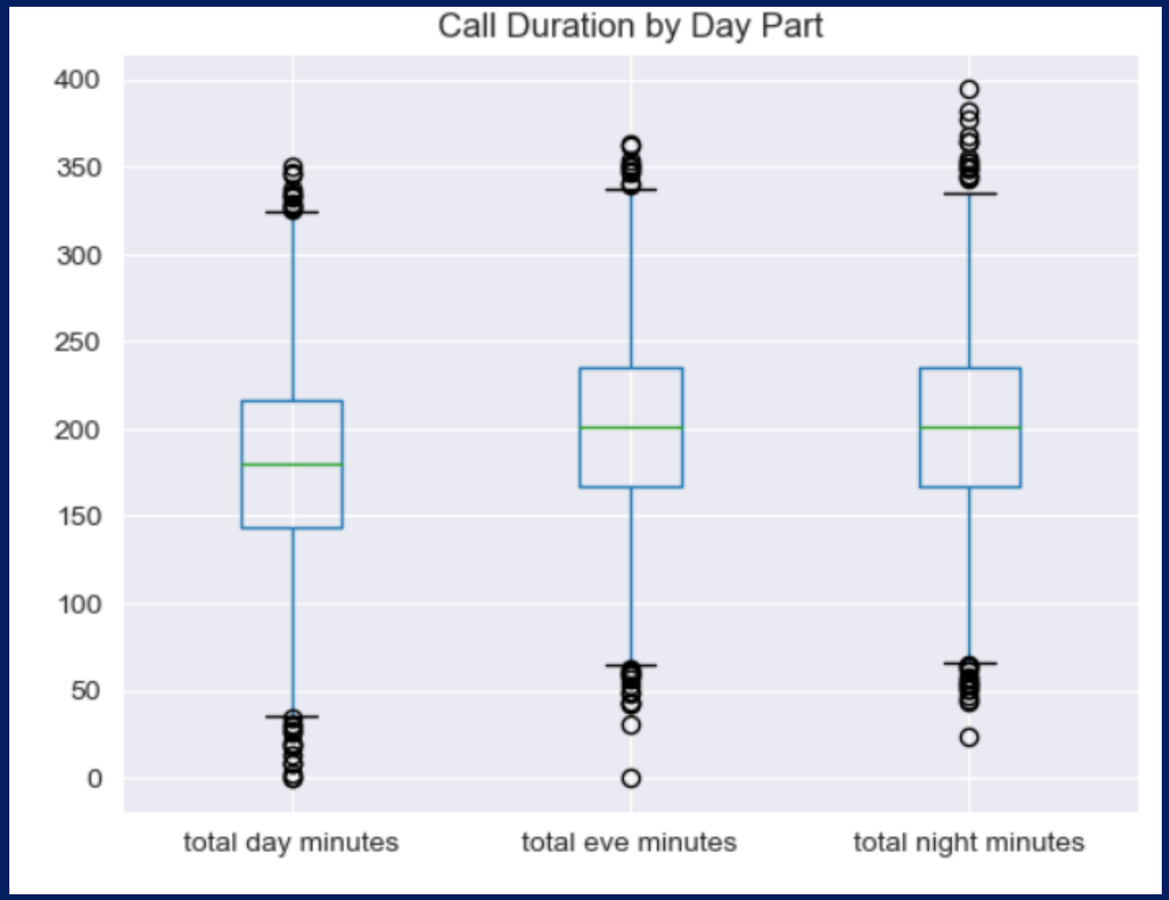
CUSTOMER SERVICE CALLS

Calls to customer service is binomially distributed with most people making 1 to 3 calls



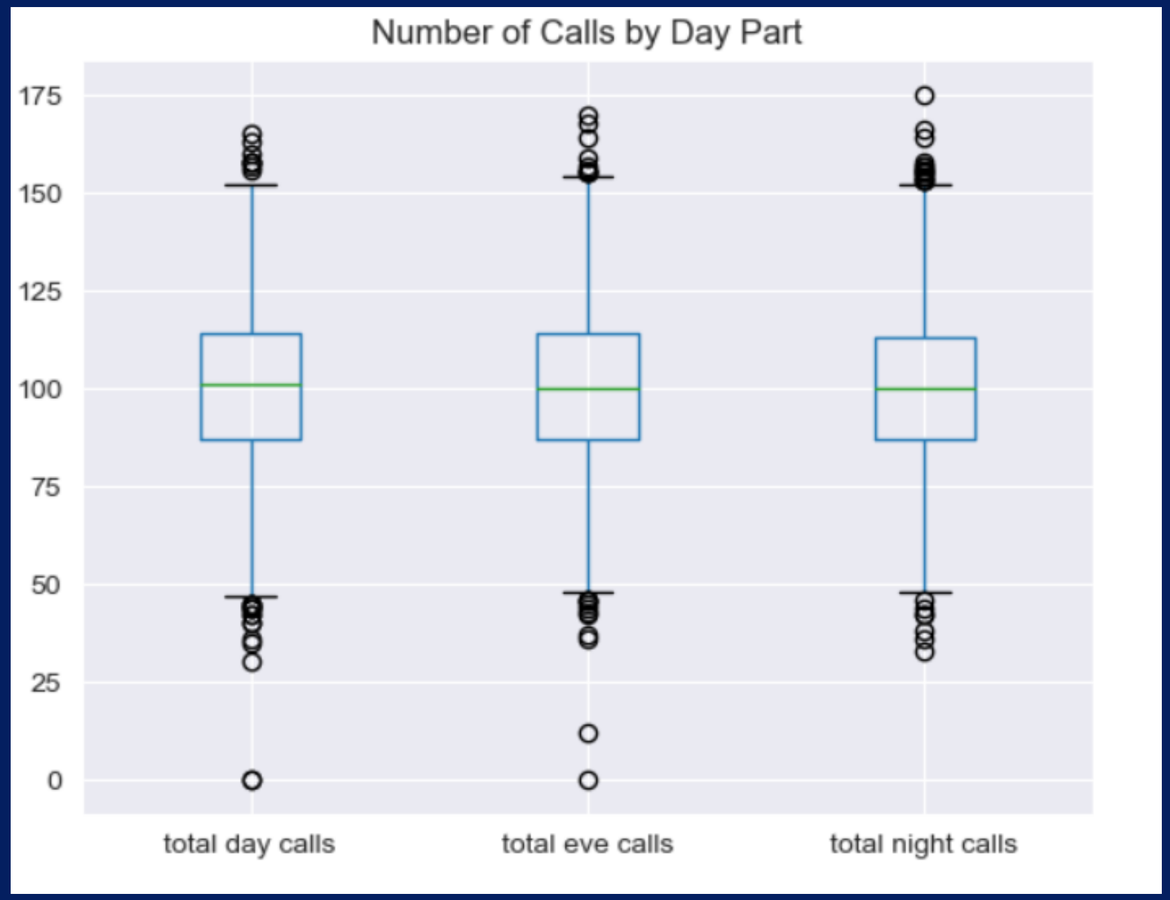
CALL DURATION

Call duration increases by day part, thus evening and night calls last longer than day calls



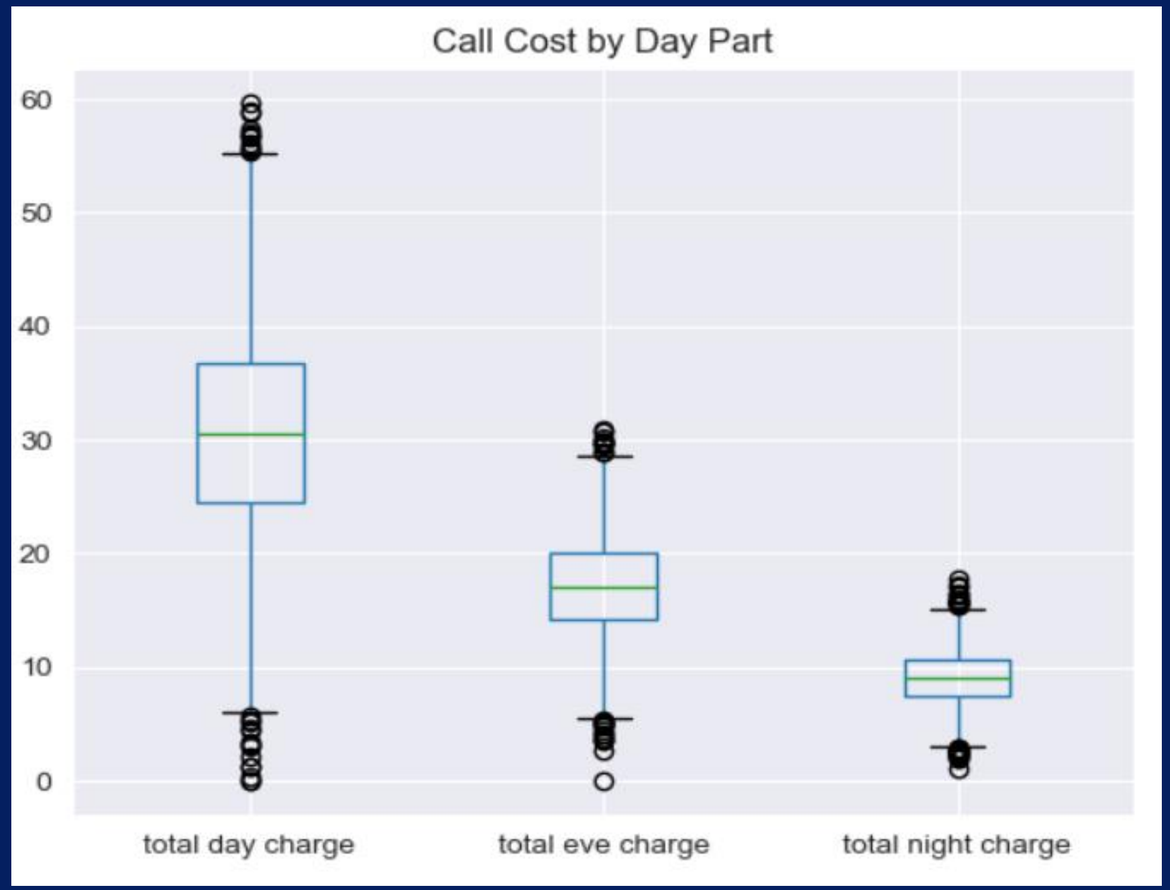
NUMBER OF CALLS

Number of calls made doesn't vary by day part



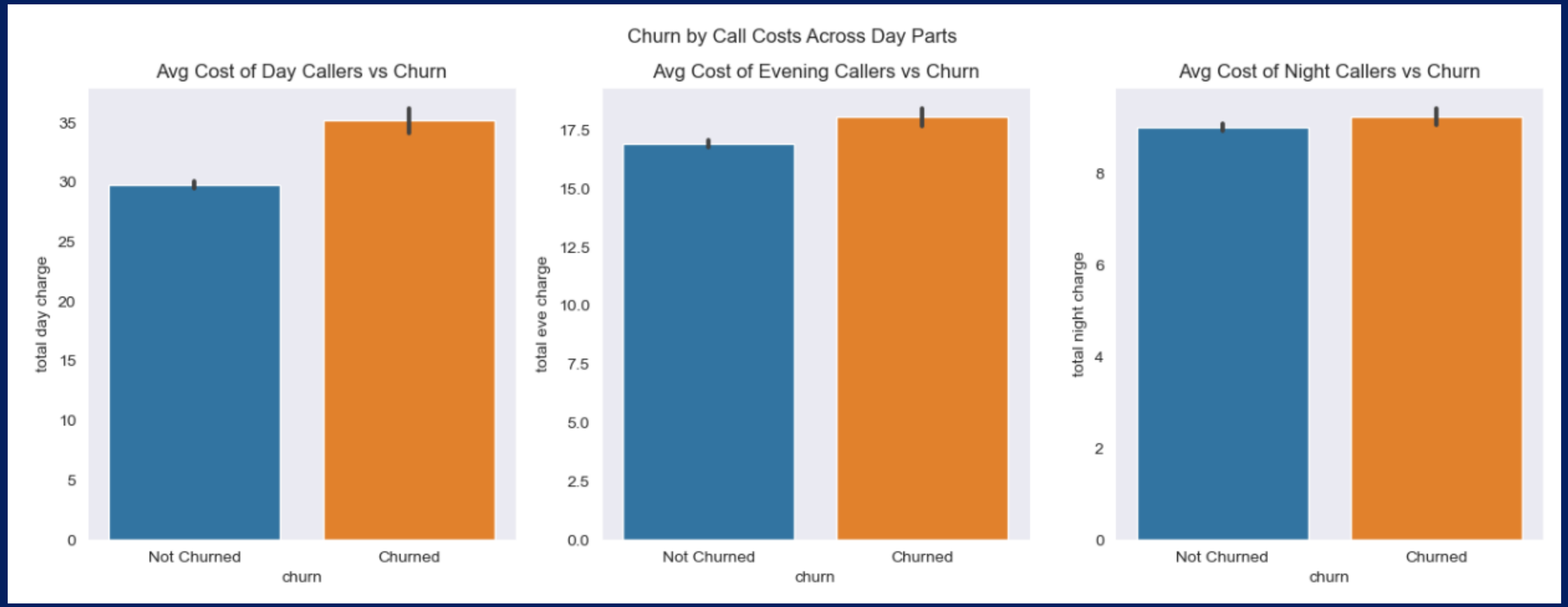
CALL COSTS

Costs tend to be cheaper in the evenings and at night, the reason why customers spend more time on the phone at night, than during the day. Also, there's a much bigger variance in call costs during the day



HOW EXPENSIVE IS CHURN FOR SYRIATEL?

Those who churn spend more on average than those who don't, especially those who call during the day. Syriatel is therefore losing customers who bring in more revenue through churn



MODELLING

success metrics

Precision

Quantifies the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives)

Recall

The ability of a model to correctly identify positive instances out of all actual positive instances in the dataset. High recall >> model has a good ability to identify positive cases when they occur. Low recall >> higher chance to incorrectly classify actual positive instances as negative

Accuracy

measures the overall correctness of the model's predictions by calculating the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances

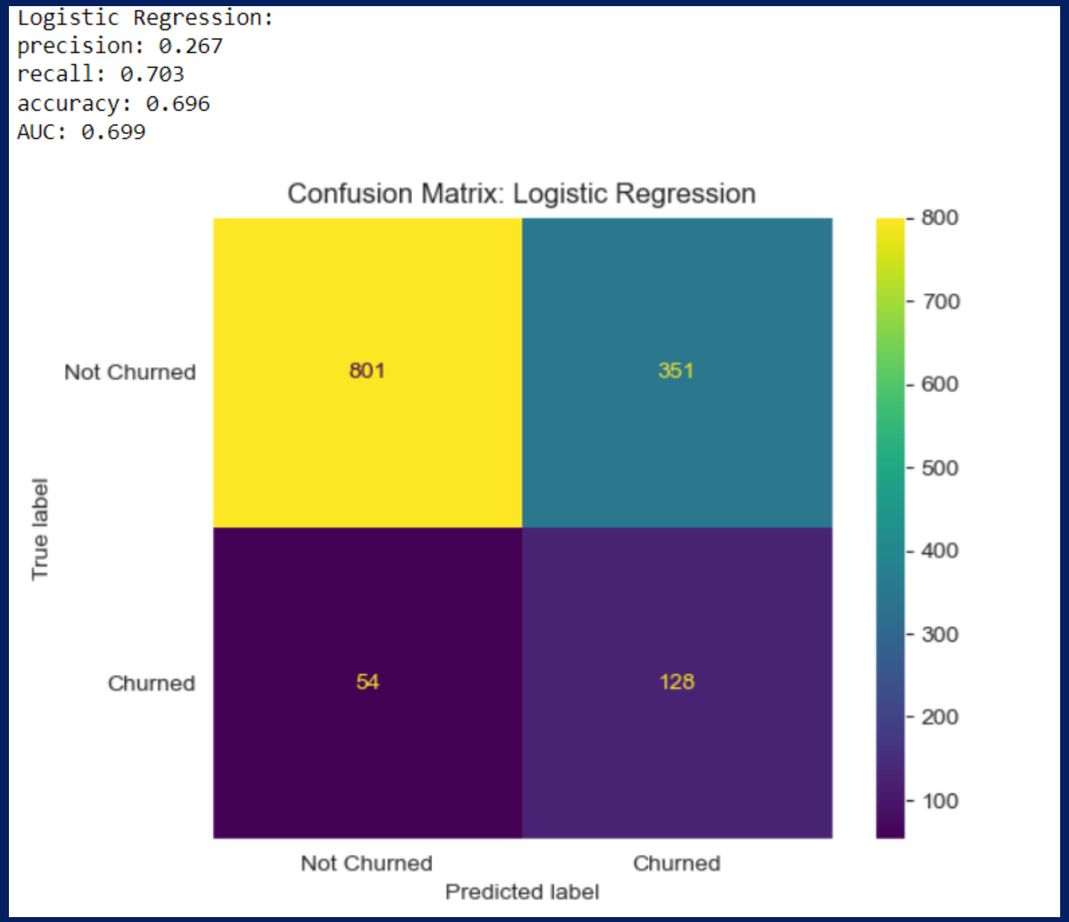
AUC

It is particularly useful when dealing with imbalanced datasets, where the class distribution is uneven. A high AUC score suggests that the model is effective at distinguishing between positive and negative instances, regardless of the threshold chosen for classification

BASELINE MODEL

We start with logistic regression as our baseline model. We will use a pipeline to streamline our work and balanced class weight to account for class imbalances

From the metrics, we see that our baseline model can be improved. We have a high number of false positives affecting the precision score.



ITERATED MODELS

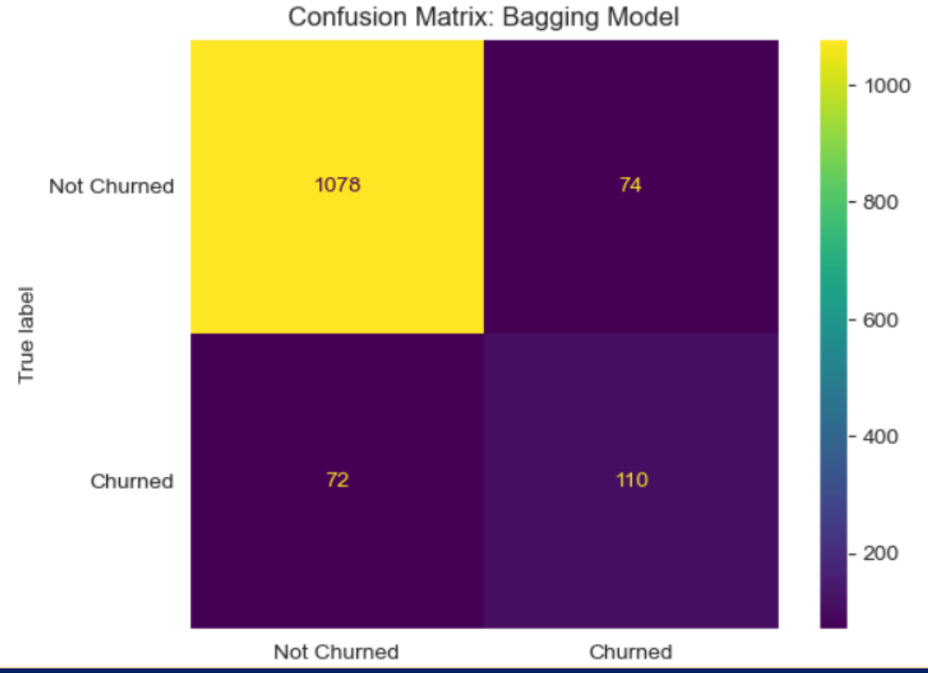
The comparison of the 4 models shows that the model labelled ‘Improved 2’- Ensemble Bagging Classifier, performs the best. The tuned hyperparameters of this classifier finds a good balance between precision and recall, and has a high AUC score

We therefore conclude that a Bagging Classifier model with 10 `max_features`, 40 `n_estimators`, with a DecisitionTree of `max_depth` None and `min_samples_split` 5 is the better model at predicting churn

Areas of further investigation include:
- trying other models like ensemble methods
- further tuning of the model
- applying dimensionality reduction to engineer correlated features

Metric	Baseline	Improved 1	Improved 2	Improved 3
precision	0.265	0.491	0.598	0.835
recall	0.714	0.582	0.604	0.582
accuracy	0.691	0.861	0.891	0.927
AUC	0.701	0.743	0.77	0.782

```
Ensemble Bagging Classifier:
precision: 0.598
recall: 0.604
accuracy: 0.891
AUC: 0.77
```



THANK YOU!