**NAME OF THE PROJECT**

Customer Churn Analysis

**ACKNOWLEDGMENT:**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped me and guided me in completion project.

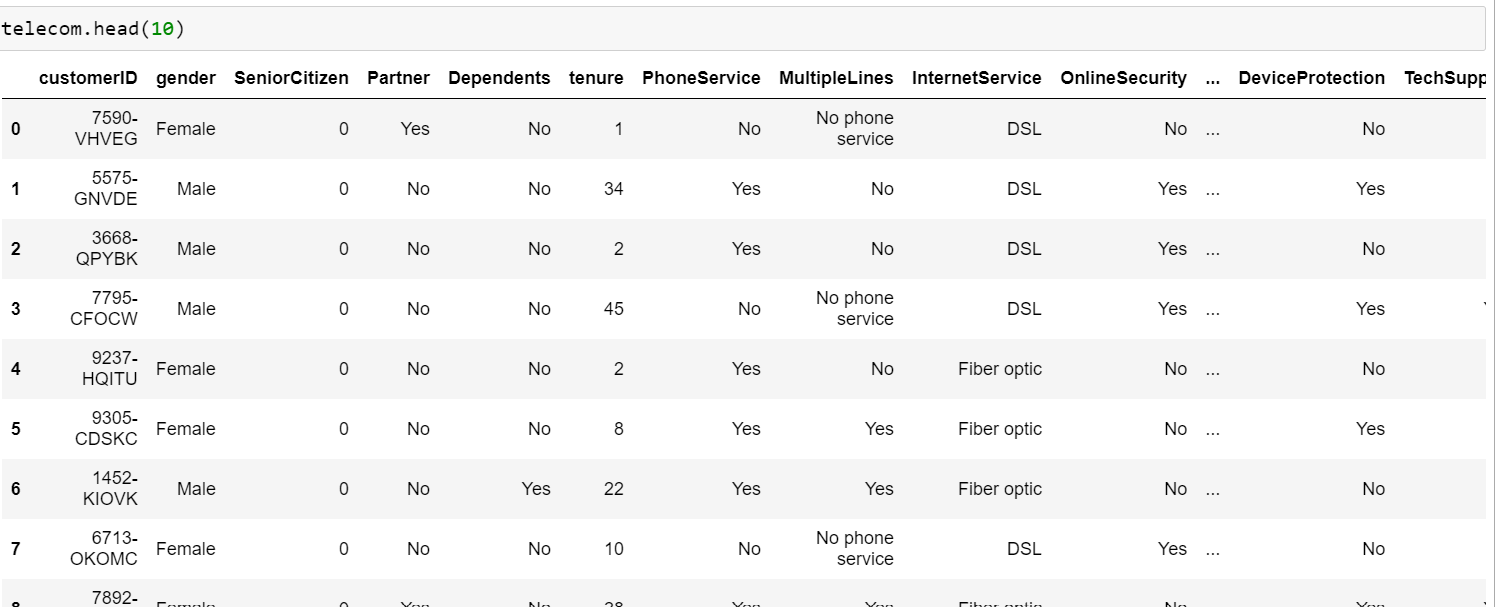
**INTRODUCTION**

1. Problem statement

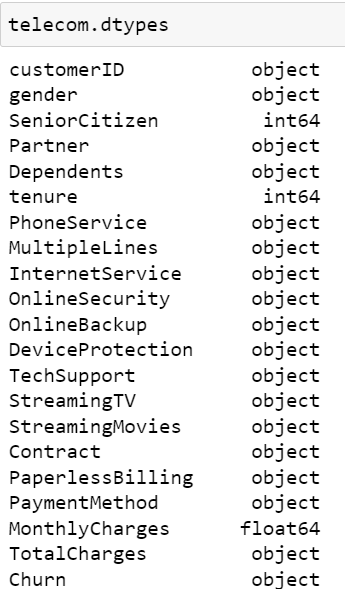
Customer churn analysis is the process of using data to understand why your customers have stopped using your product or service. Analyzing your churn doesn’t **only** mean [knowing what your churn rate is](https://baremetrics.com/academy/churn). It’s about figuring out why customers are churning at the rate they are, and how to fix the problem.

Explore Dataset:

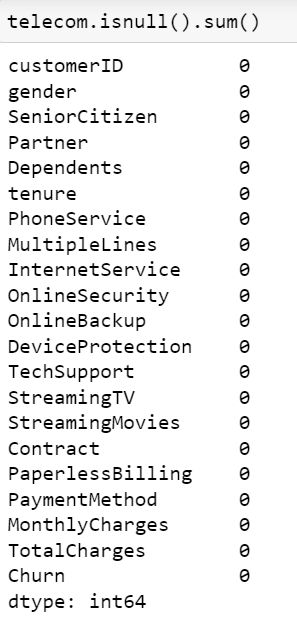




**The dataset has 21 columns:**

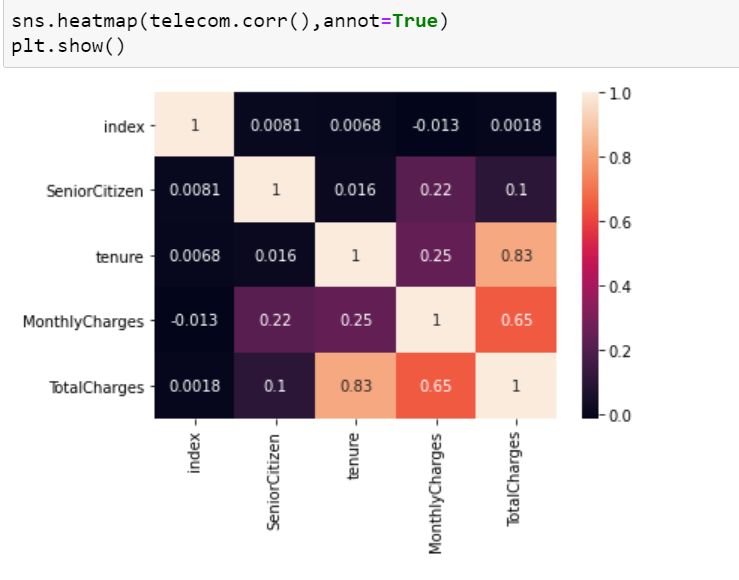


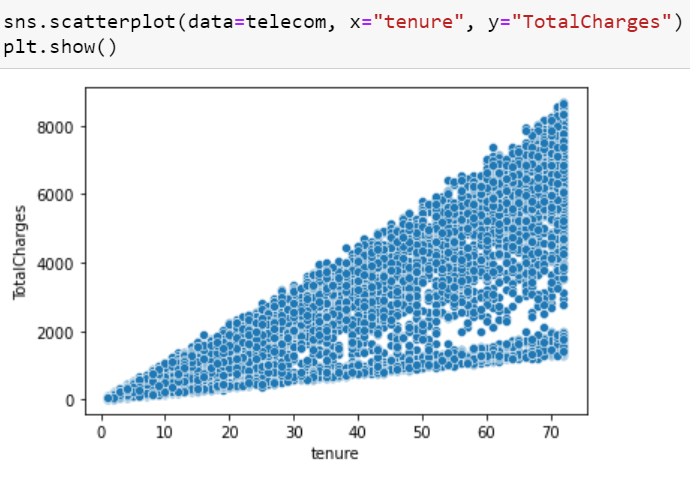
**Data Pre-Processing**

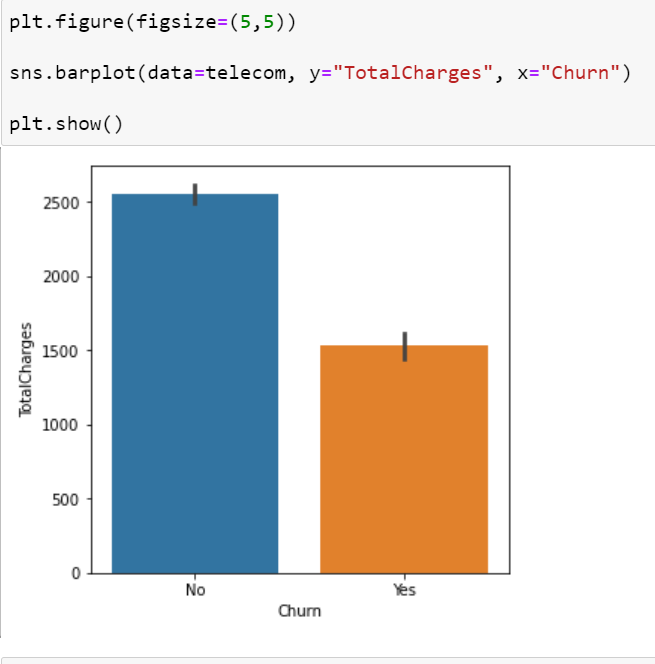


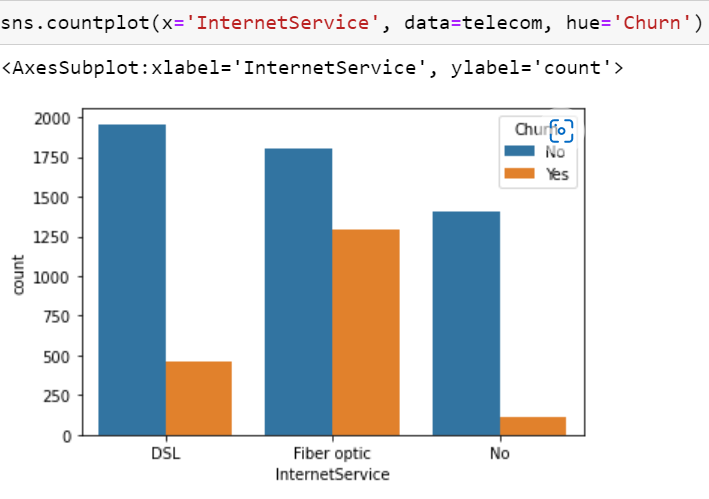
## Exploratory Data Analysis

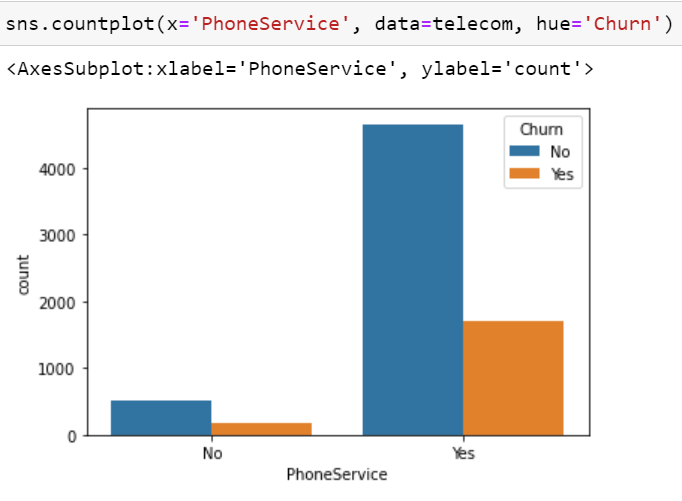
As the purpose of this experiment is to identify patterns that can yield to customer’s churn, I will be focusing mainly on the churn portion of the dataset for the exploratory analysis. The customer lifespan (in months) is represented by the feature Tenure and customer churn by the feature Churn, which is the target variable of this experiment. The bar chart below can provide right away a good insight on how churn is distributed across the customer lifespan.



****



****

****

## Summary of EDA

* High-end device users have churned less as compared to budget phone users.
* Churned users have half the bills paid and amount transacted compared to retained. Incentivize or push to double the # of payments and amount paid in the first seven days to reduce churn.
* Users need to be nudged to complete the first payment, ensuring retention.
* Ensure more than four payments to reduce churn.
* Referred users with 0 First payment churned higher than organic users.

## Models

In this experiment, I applied four different ML algorithms to analyze and compare the Recall score obtained by each of them. Those are listed below:

[Support vector machines (SVMs)](https://scikit-learn.org/stable/modules/svm.html)

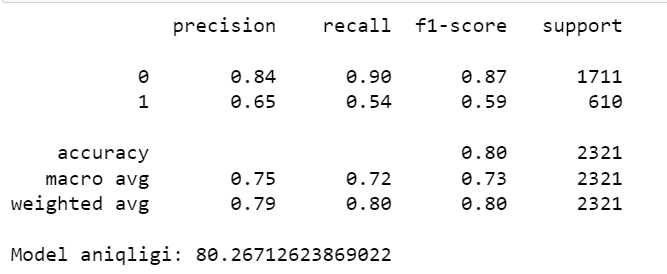
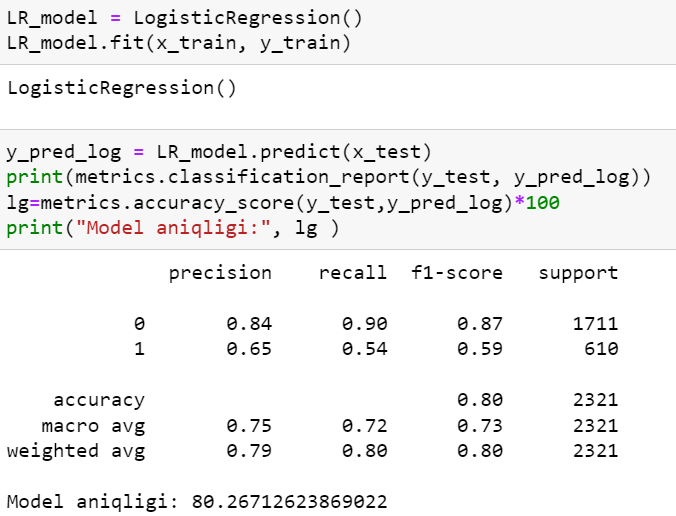
[Logistic Regression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

[XGBoost](https://xgboost.readthedocs.io/en/latest/)

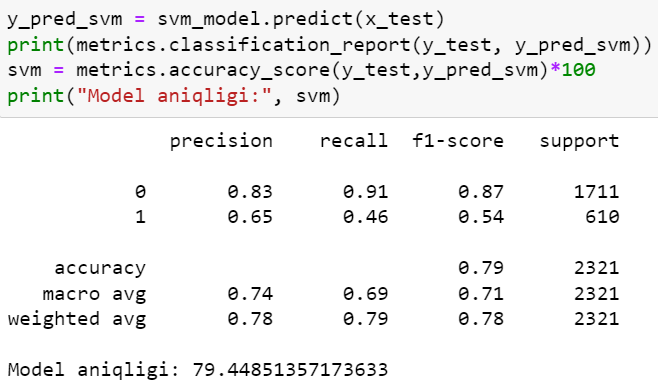
Decision Tree

Random Forest

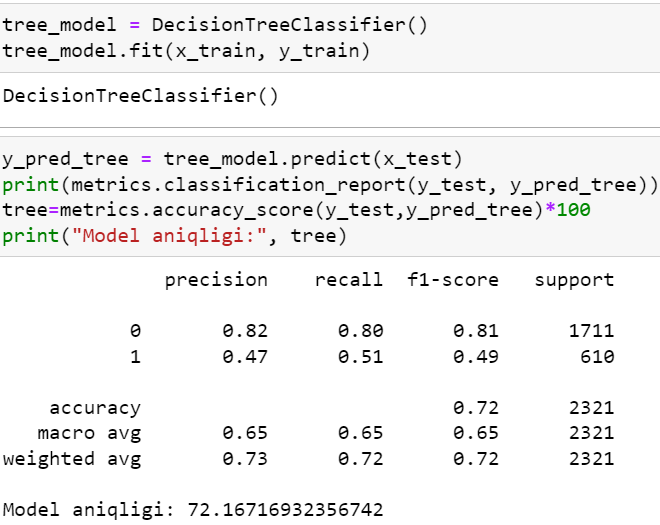
To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were Logistic Regression (0.80 Recall score). But there is still room for optimization.



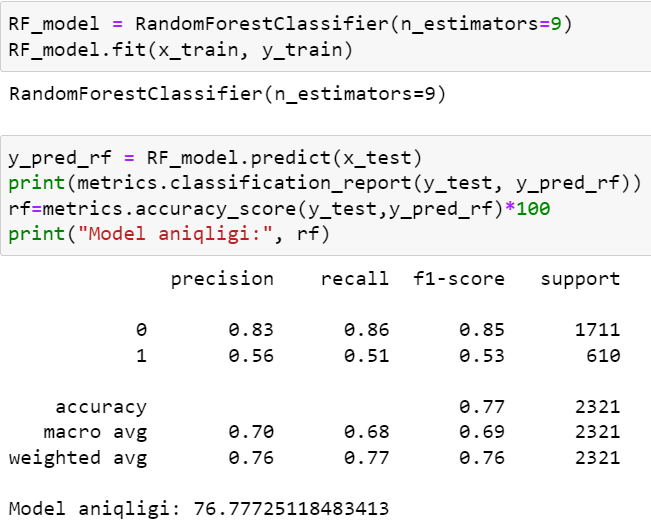
To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were SVM (0.79 Recall score). But there is still room for optimization.



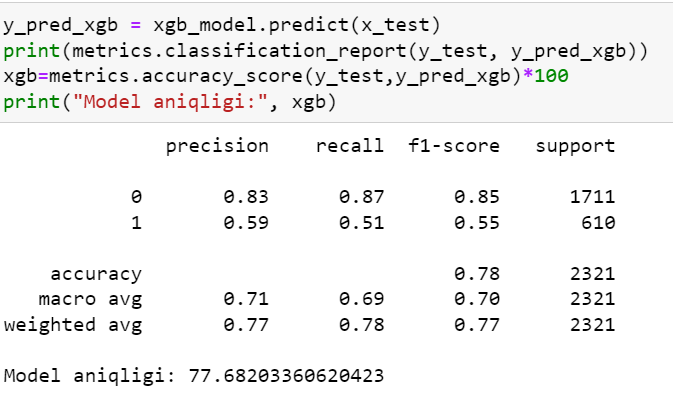
To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were DecisionTreeClassifier(0.72 Recall score). But there is still room for optimization.



To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were RandomForestClassifier(0.72 Recall score). But there is still room for optimization



To compare their performances, as a first step, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance. The top-performers were XGB Boost(0.77 Recall score). But there is still room for optimization.



To compare their performances to all model, I applied [cross-validation](https://scikit-learn.org/stable/modules/cross_validation.html) method which is a technique that partitions the data into subsets, training the data on a subset and use the other subset to evaluate the model’s performance.

Conclusion

No algorithm will predict churn with 100% accuracy. There will always be a trade-off between precision and recall. That's why it's important to test and understand the strengths and weaknessesof each classifier and get the best out of each.

If the goal is to engage and reach out to the customers to prevent them from churning, it's acceptable to engage with those who are mistakenly tagged as ‘not churned,’ as it does not cause any negative impact. It could potentially make them even happier with the service. This is the kind of model that can add value from day one if proper action is taken out of meaningful information it produces.

Submitted by:

Chandan Kumar