**DATA REPORT**

**SyriaTel customer churn**

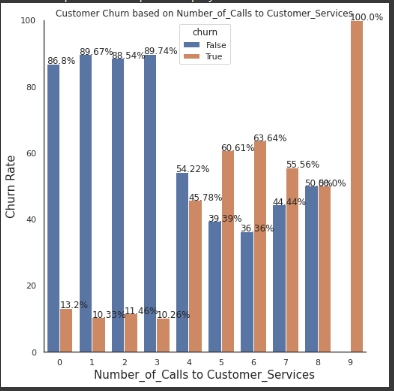
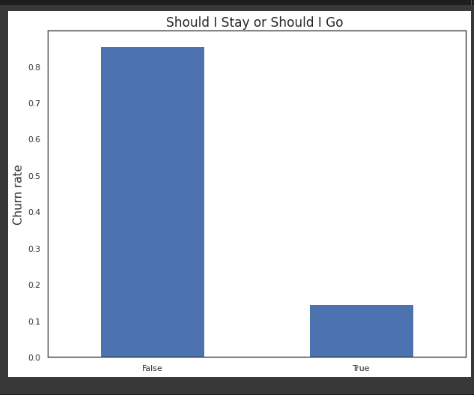
**1.INTRODUCTION**

SyriaTel is a Telecommunication company that provides communication and data services. In today’s economy every company is trying its best to increase their profit by reducing it loss. In the case of SyriaTel company it has been experiencing a drop on its return income. One of the cause they have identify is customer churn(customer leaving or stop using the services provided by the company.

This [project](https://github.com/Davidkamiti/phase-3-project/blob/main/Phase_3_project.ipynb) is aimed at Building a classifier to predict whether a customer will stop doing business with SyriaTel and the causes of customer churning.

**2. Methods used to solve the problem**

The method used to solve is l started by finding if there was missing data in our dataset which l found there was none. Proceeded to visualizing the relationship of churn(target variable) with other variables without normalization and after normalization. Encoded categorical variables and then scaled the numerical variables. Then splitted the data into training and testing sets.

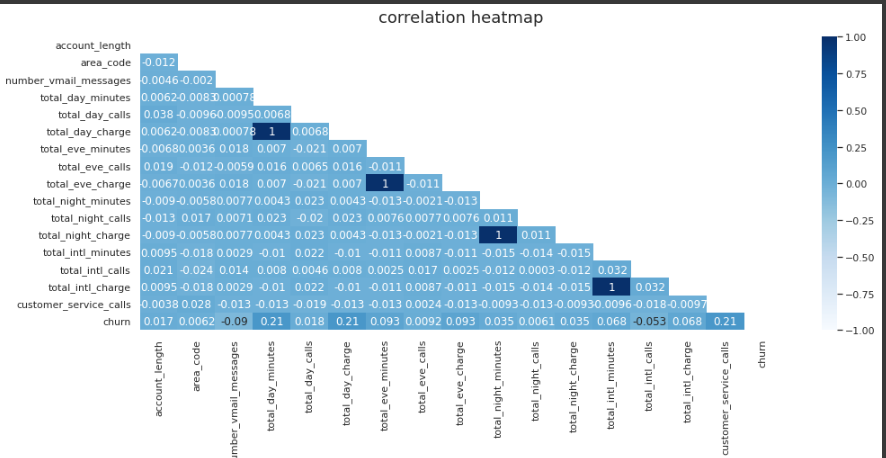


The graphs show the number of customers who stayed and those who left (false=stay and true=left) and the other shows number of churn rate vs number of calls where customers who called more than 4 times left or stopped using the company services respectively.

**3. Dataset and Features**

The [data](https://github.com/Davidkamiti/phase-3-project/blob/main/bigml_59c28831336c6604c800002a.csv) used here has 3333 rows and 21 columns.

Dropped the phone\_number column as it more personal for every individual and it had no effect on customer churning.



This are the feature and their relationship to each other.

**4. Models**

Different models were used like logistic regression model, SMOTE model, Decision Tree, GridSearch, Random Forest, XGBoost model. Also evaluated the model performance using recall, precision and confusion matrix. In all the model, XGBoost is the one that performed better compared to others.

The results for Xgboost test set were:

Accuracy = 0.9448

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation | precision | recall | F1score | support |  |
| 0 | 0.95 | 0.98 | 0.97 | 713 |  |
| 1 | 0.88 | 0.72 | 0.79 | 121 |  |

The classification report provides more insights into the performance of the model. The precision score of the 0 class is 0.95, which means that 95% of the instances the model predicted as 0 are actually 0. The recall score of 0.98 means that the model was able to identify 98% of the actual 0 instances. For the 1 class, the precision score is 0.88 and the recall score is 0.72.

This are the Accuracy and evaluation results after tuning the XGBoost model to obtain best hyperparameter.

To get more result on the other models you can try [notebook](https://github.com/Davidkamiti/phase-3-project/blob/main/Phase_3_project.ipynb)

**5. conclusion and recommendation**

the best performing model(xgboost) can be used for deployment as performed better on predicting the unseen data.

The other model can also be tuned to improve their performance

This project has provided the model that can be used to predict the customer churn in SyriaTel company.

Found out that customer that call more than four times are the ones that mostly left the company and some state also experienced large number of customers that churned.

The model was not balanced in the target variable.