**A CLASSIFIER MODEL FOR PREDICTING WHETHER A CUSTOMER WILL SOON STOP DOING BUSINESS WITH SYRIATEL.**

# **Business Understanding**

SyriaTel is a telecommunications company and they want to reduce the amount of money spent on retaining customers who do not use their services for a long time. The features given in the dataset will give us more information as to whether a customer will be retained or they will churn.

SyriaTel is interested in retaining their customers as gaining new customers is expensive and they want to reduce the cost of retaining customers who stay for a short period in time. It will be detrimental to analyse the features provided to see if a pattern can be identified that will make the current customers stay with the company and come up with conclusive findings that the company can implement to retain their customers for longer periods of time.

# **Research Question**

Build a classifier model to predict whether a customer will ("soon") stop doing business with SyriaTel.

# **Objectives**

## **Main Objective**

To build a classifier model to predict whether a customer will ("soon") stop doing business with SyriaTel.

## **Specific Objectives**

* To determine which features determine if a customer will churn
* To increase the recall score to minimize false negatives.

# **Metric of Success**

The model will be considered a success if an accuracy score of 80% and a recall score of 70% is achieved.

# **Data Understanding**

## **Data Source**

Data is downloaded from Kaggle [website](https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset).

## **Data Description**

1. state: the state the customer lives in

2. account length: the number of days the customer has had an account

3. area code: the area code of the customer

4. phone number: the phone number of the customer

5. international plan: true if the customer has the international plan, otherwise false

6. voice mail plan: true if the customer has the voice mail plan, otherwise false

7. number vmail messages: the number of voicemails the customer has sent

8. total day minutes: total number of minutes the customer has been in calls during the day

9. total day calls: total number of calls the user has done during the day

10. total day charge: total amount of money the customer was charged by the Telecom company for calls during the day

11. total eve minutes: total number of minutes the customer has been in calls during the evening

12. total eve calls: total number of calls the customer has done during the evening

13. total eve charge: total amount of money the customer was charged by the Telecom company for calls during the evening

14. total night minutes: total number of minutes the customer has been in calls during the night

15. total night calls: total number of calls the customer has done during the night

16. total night charge: total amount of money the customer was charged by the Telecom company for calls during the night

17. total intl minutes: total number of minutes the user has been in international calls

18. total intl calls: total number of international calls the customer has done

19. total intl charge: total amount of money the customer was charged by the Telecom company for international calls

20. customer service calls: number of calls the customer has made to customer service

21. churn: true if the customer terminated their contract, otherwise false

# **Data Preparation**

## **Loading Libraries**

Load the libraries necessary for cleaning and analysis

## **Loading the Data**

Load the dataset from the CSV file. The name of the CSV file is “bigml\_59c28831336c6604c800002a.csv”.

The shape of the dataset is 3333 by 21 (3333 rows and 21 columns)

## **Cleaning the Data**

Data cleaning is the process of preparing data for analysis by removing information that is irrelevant or incorrect. This incorrect information may end up giving wrong predictions that may be detrimental for a business/company. Correct information is the lifeline of any good project.

The steps for the data cleaning process is:

1. Consistency - Ensure there are no duplicates in the data.

2. Uniformity - Ensure the data types for the datasets are accurate.

3. Completeness - Ensure the dataset has no missing values.

# 

# 

# **Modelling**

In model building the label to be used is churn since we want to predict if a customer will soon discontinue their services with SyriaTel. The features are:

1. state: the state the customer lives in
2. account length: the number of days the customer has had an account
3. area code: the area code of the customer
4. international plan: true if the customer has the international plan, otherwise false
5. voice mail plan: true if the customer has the voice mail plan, otherwise false
6. number vmail messages: the number of voicemails the customer has sent
7. total day calls: total number of calls the user has done during the day
8. total day charge: total amount of money the customer was charged by the Telecom company for calls during the day
9. total eve calls: total number of calls the customer has done during the evening
10. total eve charge: total amount of money the customer was charged by the Telecom company for calls during the evening
11. total night calls: total number of calls the customer has done during the night
12. total night charge: total amount of money the customer was charged by the Telecom company for calls during the night
13. total intl calls: total number of international calls the customer has done
14. total intl charge: total amount of money the customer was charged by the Telecom company for international calls
15. customer service calls: number of calls the customer has made to customer service

Baseline model was constructed logistic regression and the output displayed.

The aforementioned characteristics were used as features to predict whether a customer would churn. All the categorical columns were transformed into numerical columns and all the features were scaled before the model was fit.

The baseline model had poor metrics and thus other models were fitted. They included the DesionTree Classifier, K-Nearest Neighbor Classifier and the RandomForest Classifier.

Among the three models, the RandomForest Classifier has the best metrics hence hyperparameter tuning was applied to optmize it. Other more sophisticated models were added to the fray, Gradient Boost Classifier and AdaBoost Classifier.

The Gradient Boost Classifier was the best model overall since it had the highest recall score which was the best at correctly classifying the customers who were soon to discontinue their services with SyriaTel.

# 

# **Conclusion and Recommendation**

## **Conclusion**

* The initial goal of achieving 80% accuracy is achieved by all the models that were fitted.
* Gradient Boost Classifier model provides the highest recall score of all the models.
* Customer service calls, international plan, total day charges are the features with the greatest importance.

## 

## **Recommendations**

1. The number of calls the customer makes to customer service should be reduced to two. If the customer calls once, the customer service agent should do a follow up to ensure the customer's issue has been fully resolved to avoid making the customer call back again.

2. The customer service agents should be trained on how to treat, escalate and solve issues as soon as possible.

3. The phone charges are high hence a high churn rate, the business should consider having payment plans for different customers so that one is charged according to the service they consume.

4. The business should consider lowering or giving discounts for customers who have an international plan and also charge them per the service they use.