**Classification**

Predicting Customer Churn

## **Introduction**

Churn is a measure of how many customers stop using a product or service. A high churn rate can negatively impact monthly recurring revenue and can also indicate dissatisfaction with a product or service. In its most simplistic form, the churn rate is the percentage of total customers that stop using/paying over a period.

A company may want a machine learning model to predict customer churn for several reasons:

1. **Retention:** By predicting which customers are likely to churn, a company can take proactive measures to prevent them from leaving. This may involve offering incentives, improving customer service, or addressing specific pain points.
2. **Cost reduction:** Acquiring new customers can be expensive, so it is often more cost-effective for a company to retain existing customers. Predicting customer churn can help a company identify at-risk customers and allocate resources accordingly.
3. **Competitive advantage:** Companies able to predict and prevent customer churn may have a competitive advantage over those that do not**.** By retaining more customers, a company can increase revenue and market share.
4. **Customer satisfaction:** Predicting and preventing churn can lead to increased customer satisfaction. By addressing customer concerns and improving the customer experience, a company can build stronger relationships with its customers.

Overall, predicting customer churn can help a company increase retention, reduce costs, gain a competitive advantage, and improve customer satisfaction hence the need for this project.

## **Aims:**

We aimed to find the likelihood of a customer leaving the organization, the key indicators of churn as well as the retention strategies that can be implemented to avert this problem.

**Methodology:**

This work was done based on the CRISP-DM algorithm and this article shall be focused on explaining these steps which involves the following;

* Business understanding.
* Data understanding.
* Data preparation.
* Modeling.
* Evaluation.
* Deployment.

This work was done using Jupyter notebook in Google Collaboratory in Python programming language.

The last step in this process, which is deployment, will be carried out in the coming weeks therefore we will not talk about it in this write-up.

## **Business understanding:**

The data for the analysis was provided with no background information on the dataset.

**Data Understanding:**

The data for the project was provided in a csv format. The following describes the columns present in the data.

**Gender** -- Whether the customer is a male or a female

**SeniorCitizen** -- Whether a customer is a senior citizen or not

**Partner** -- Whether the customer has a partner or not (Yes, No)

**Dependents** -- Whether the customer has dependents or not (Yes, No)

**Tenure** -- Number of months the customer has stayed with the company

**Phone Service** -- Whether the customer has a phone service or not (Yes, No)

**MultipleLines** -- Whether the customer has multiple lines or not

**InternetService** -- Customer's internet service provider (DSL, Fiber Optic, No)

**OnlineSecurity** -- Whether the customer has online security or not (Yes, No, No Internet)

**OnlineBackup** -- Whether the customer has online backup or not (Yes, No, No Internet)

**DeviceProtection** -- Whether the customer has device protection or not (Yes, No, No internet service)

**TechSupport** -- Whether the customer has tech support or not (Yes, No, No internet)

**StreamingTV** -- Whether the customer has streaming TV or not (Yes, No, No internet service)

**StreamingMovies** -- Whether the customer has streaming movies or not (Yes, No, No Internet service)

**Contract** -- The contract term of the customer (Month-to-Month, One year, Two year)

**PaperlessBilling** -- Whether the customer has paperless billing or not (Yes, No)

**Payment Method** -- The customer's payment method (Electronic check, mailed check, Bank transfer(automatic), Credit card(automatic))

**MonthlyCharges** -- The amount charged to the customer monthly

**TotalCharges** -- The total amount charged to the customer

**Churn** -- Whether the customer churned or not (Yes or No)

We asked the following questions and were set to answer them:

**Questions**

1. Do senior citizens have a higher churn rate than others?
2. Do customers with dependents have higher church rates?
3. Does age and gender contribute to the churn rate?
4. Is there a relationship between tenure and churn rate?
5. Does the contract term affect the churn rate?

### **Hypothesis**

**Null Hypothesis**: The contract term does not affect attrition

**Alternate hypothesis**: The contract term affects attrition.

We imported libraries and packages then we read the data into pandas using the read csv method then loaded the data into pandas data frame.

We called the following methods:

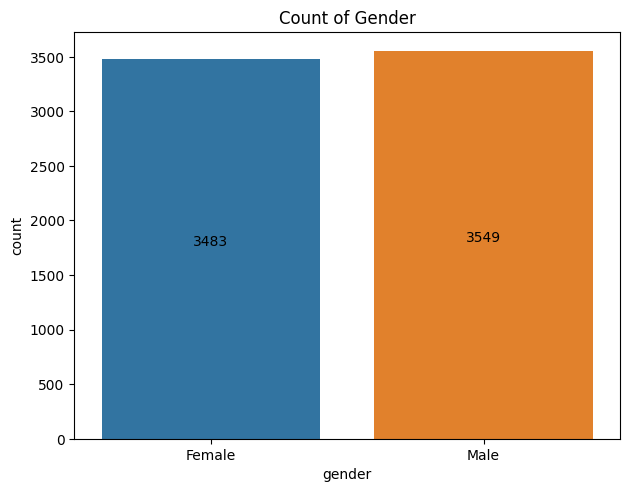
1. Head method –to take a quick look at what is inside the data set
2. Shape method –to see the number of rows and columns in the data set
3. Info method –to see the number of non-null values in each column and the data type
4. Is null method –to check if there are null values in the data set
5. Is duplicated method –to see if there are duplicated rows in the data set
6. Unique method –to check the number of unique values in each column
7. Drop NA method –we discovered there were 11 null values in the dataset, and they were a small number, so we dropped them
8. We dropped the ID column because it is not useful in our work
9. Describe method –to get quick statistics about the numeric variables present in the dataset.
10. We discovered that the Total Charges column’s data type was object, so we changed it to numeric.

### **Univariate and Bivariate Analysis**

We did univariate and bivariate analysis on the dataset, and we found out that:

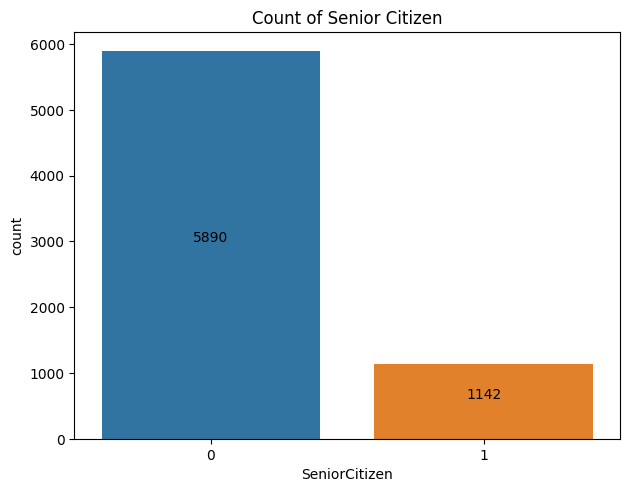
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **tenure** | 32.421 | 24.5452 | 1.0 | 9.0 | 29.0 | 55.0 | 72.0 |
| **MonthlyCharges** | 64.798 | 30.0859 | 18.25 | 35.5875 | 70.35 | 89.8625 | 118.75 |
| **TotalCharges** | 2283.30 | 2266.7713 | 18.8 | 401.45 | 1397.475 | 3794.7375 |  |

1. **Gender:** as seen from the graph below, there were 3483 females and 3549 males.

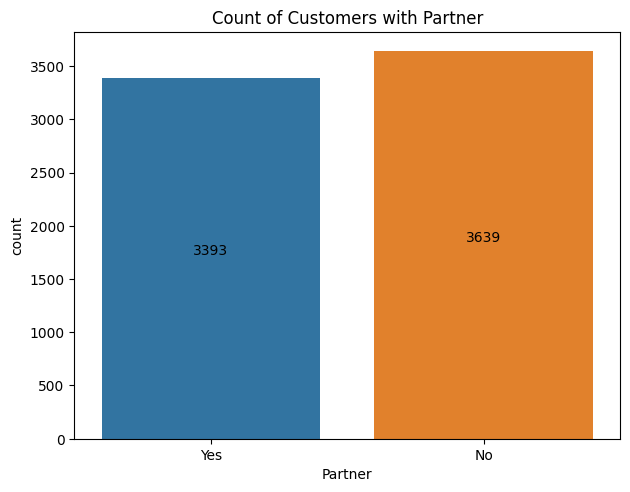


|  |  |  |
| --- | --- | --- |
| **gender** | **Churn** | **count** |
| Female | No | 2549 |
| Female | Yes | 939 |
| Male | No | 2625 |
| Male | Yes | 930 |

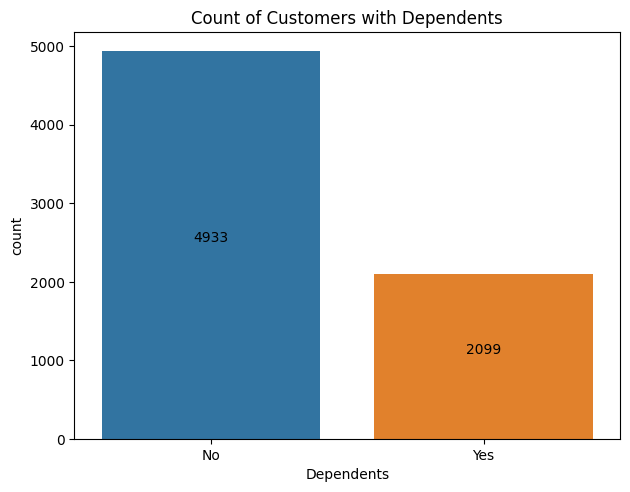
1. **Senior Citizen:** 1142 of the total customers are labeled senior citizens



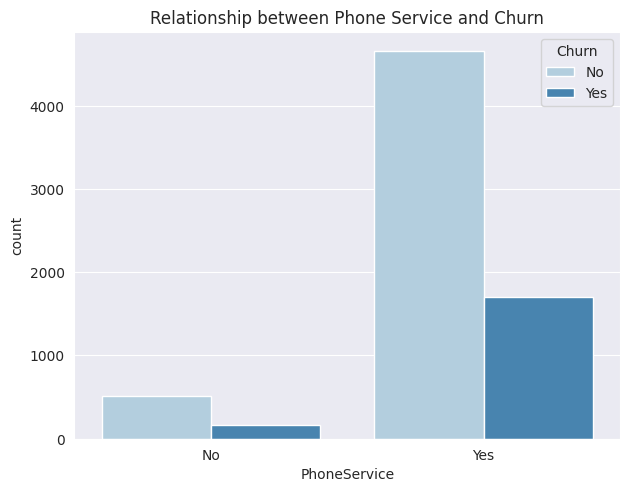
**3. Partner:**



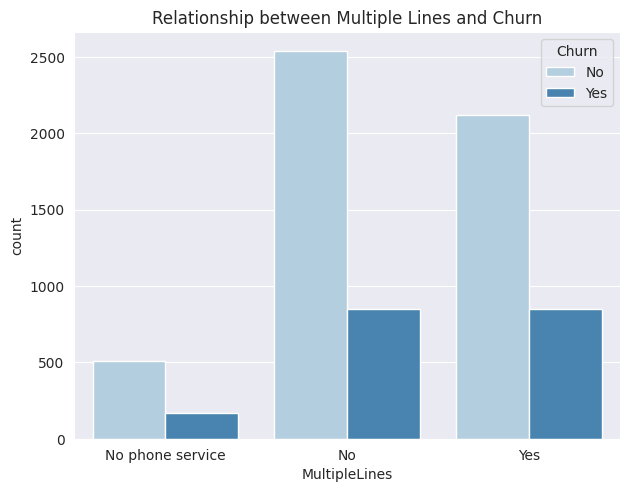
1. **Dependents:**



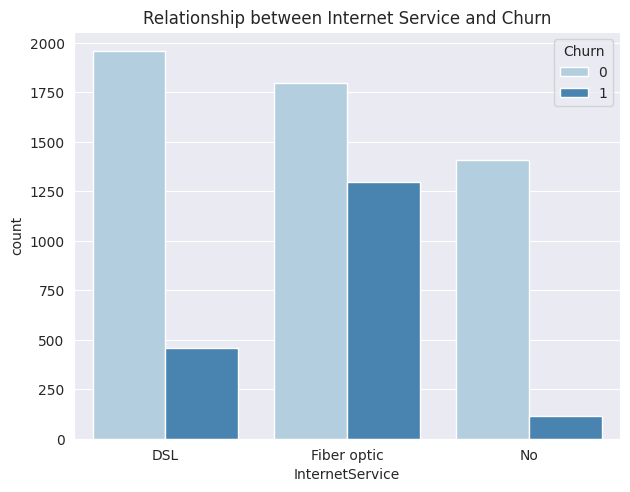
1. **Phone Service:** As seen from the chart below, more customers who had phone services left the company than those who did not.



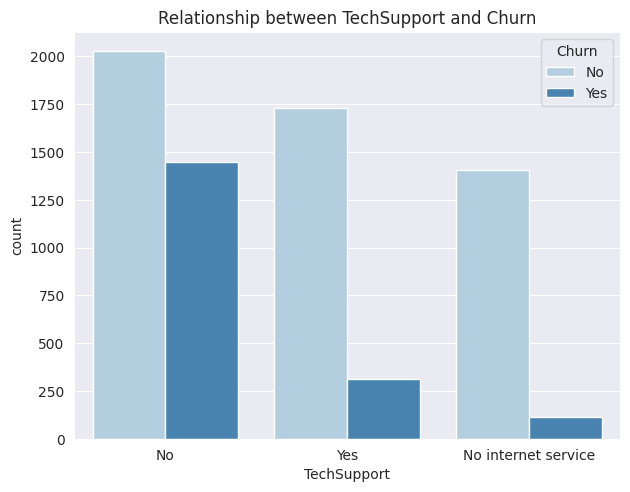
1. **Multiple Lines**



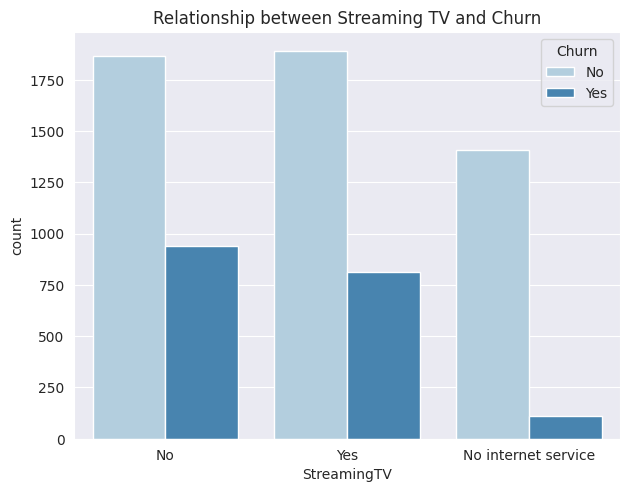
1. **Internet Service:** From the chart below, we can see that a lot of customers who have internet services are leaving the company, especially those who use fiber optic network.

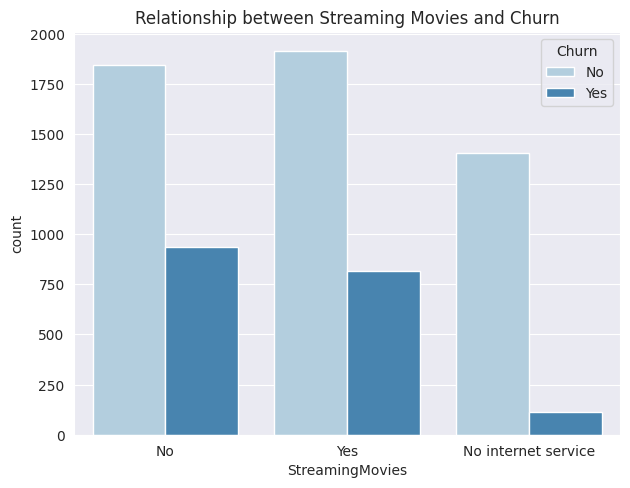


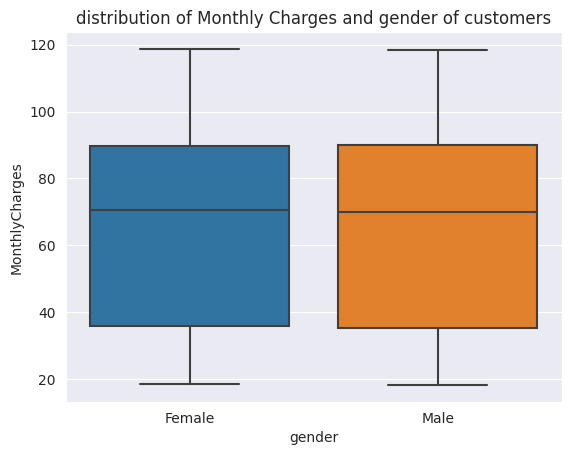
1. **Tech Support:** customers without tech support left the company more, either they did not intend to stay long with the company, or they felt neglected.



1. **Streaming TV**



1. **Streaming Movies** 
2. **Monthly Charges:** The spending habits of both genders on the network are the same



### **Hypothesis Testing**

**Null Hypothesis**: The contract term does not affect attrition

**Alternate hypothesis**: The contract term affects attrition

We used the one-way ANOVA test to determine if there were significant differences among the means of the independent groups. It assessed whether the variation between group means was larger than the variation within each group.

We got a p-value of zero (0). We saw that there is compelling evidence to reject the null hypothesis. This indicates significant differences among the group means being compared (Churn and Contract). We therefore conclude that the contract term affects attrition, the null hypothesis is rejected.

## **Data preparation.**

There were columns in the dataset that were non-numeric, so we did feature encoding using one hot encoder.

This is a classification project therefore, the data contained both the labels and the independent data in one dataset. We separated the dataset into x and y, the y subset which is the labels is a slice of the dataset containing only the churn column while the rest of the subset is the rest of the dataset with the churn column dropped.

We scaled the data set with standard scaler, and we did resampling also.

We split the dataset into train, test and validation sets using the scikit learn train test split function.

## **Modeling**

We trained the following models:

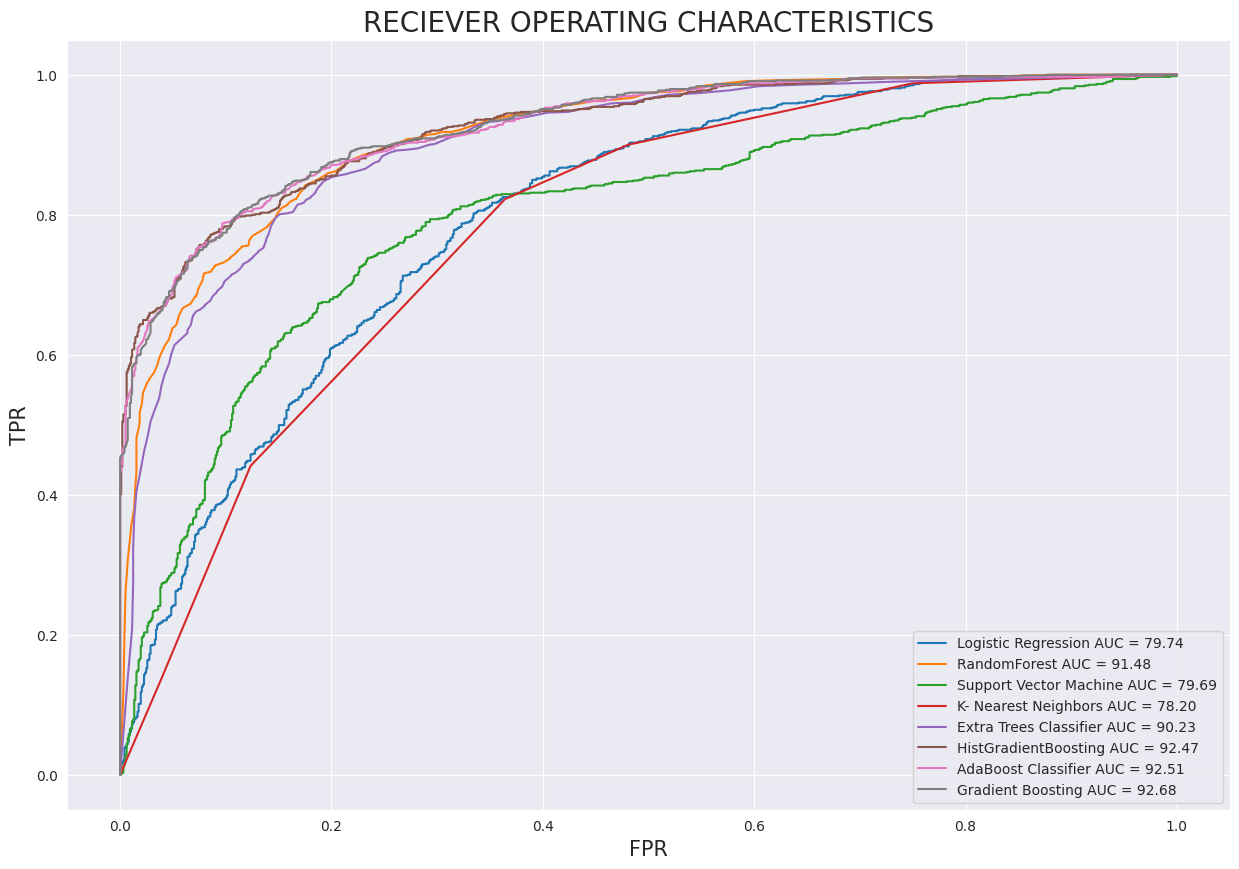
1. Logistic Regression
2. Random Forest Classifier
3. Support Vector Machine
4. K- Nearest Neighbors
5. Extra Trees Classifier
6. Histogram-based Gradient Boosting Classifier
7. Ada Boost Classifier
8. Gradient Boosting Classifier

## **Evaluation**

We trained the first eight models with default hyperparameters and got the following results for their metrics evaluation as shown in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy Score** | **Precision Score** | **Recall Score** | **F1 Score** |
| Gradient Boosting | 83.96 | 82.78 | 85.9 | 84.31 |
| AdaBoost | 83.91 | 83.69 | 84.37 | 84.03 |
| Histogram-based Gradient Boosting | 83.29 | 82.81 | 84.17 | 83.49 |
| Random Forest | 82.88 | 83.63 | 81.92 | 82.77 |
| Extra Trees Classifier | 82.21 | 82.18 | 82.43 | 82.3 |
| Support Vector Machine | 74.47 | 76.28 | 71.3 | 73.71 |
| K- Nearest Neighbors | 72.94 | 69.46 | 82.23 | 75.3 |
| Logistic Regression | 72.07 | 71.4 | 73.95 | 72 |

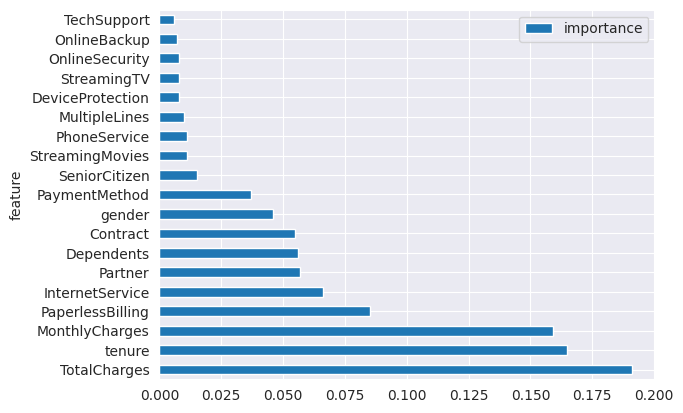
From the chart below we saw that Gradient boosting had the best score of 92.68% while K-Nearest Neighbor was the worst performer with 78.20%.



* For Accuracy Score, Gradient boosting is the best performer with a score of 83.91% while Logistic Regression is the worst performer with a score of 72.07%.
* Precision Score, Ada Boost is the best performer with 83.69% score while K-Nearest Neighbors performed worst with a score of 69.46%.
* For Recall Score, the best performer is Gradient Boosting with 85.90% score and Support Vector Machine was the worst performer with 71.30%.
* For f1 Score, Gradient Boosting performed the best with 84.31% score and Logistic Regression performed the worst with a score of 72.23%
* Because our data is unbalanced, the F1 score and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) were used to evaluate the performance of the various models.
* Therefore, Gradient Boosting was the best model among the considered models

### **Feature Importance**

From the chart below we saw that not all the features were contributing very much to the results we got from the training and evaluation of our models, so we dropped some of the less key features.



### **Hyperparameter Tuning**

From research we did about the gradient boosting model, we started our hyperparameter tuning with the following parameters:

1. N estimators
2. Max depth
3. Learning rate

We passed initial arguments to these parameters and searched for the best parameters using grid search cv. The optimum values returned were:

1. N estimators –100
2. Max depth –5
3. Learning rate –0.1

We retrained the model this time using the optimum values from the hyperparameter tuning and reevaluate the performance metrices, we got the following results:

1. Accuracy Score: 85.97%
2. Precision Score: 86.35%
3. Recall Score: 85.89%
4. F1 Score: 86.12%
5. ROC AUC Score: 93.68%

The result above shows that the performance metrices have increased because the initial value of F1 score was 83.96 but it got to 86.12 after hyperparameter tuning. Also, the ROC AUC score was 92.68% but increased to 93.68% after hyperparameter tuning.

## **Conclusion**

We discovered that the best performing model for this project is gradient boosting classifier, it performed better with N estimators of 100, max depth of 5 and learning rate of 0.1. We also believe that a better performance can still be achieved by adding more values to the hyperparameters.

There is evidence that customers are leaving the company because of high charges. The charges are mostly coming from internet services and the most affected customers are those who have partners and dependents both genders inclusive.