SpeakX Assignment

A Report on Churn Prediction

Name: Chigicherla Thatireddy

Reg no: 12107331

Aim: To develop a predictive model that can identify customers at risk of churning, enabling the company to take proactive measures to retain them.

Steps Involved in this process:

1.Understanding Data

- a. Connecting to data
- b. Viewing the data
- c. Summarizing the data

2.Data Preprocessing

- a. Data Cleaning
- b. Converting the Categorical into Numerical

3.Exploratory Data Analysis

4.Data Modeling

- a.Train test split
- b.Building a model
- c.Prediction

5.Model Evaluation

1. **Understanding the Data.**

These are the columns present in the data set:

- CustomerID: A unique identifier for each customer.
- Gender: Whether the customer is male or female.
- SeniorCitizen: Whether the customer is a senior citizen (typically 65 years or older).
- Partner: Whether the customer has a partner (e.g., spouse or significant other).
- Dependents: Whether the customer has dependents (e.g., children or other family members).
- Tenure: The length of time (in months) the customer has been with the company.
- PhoneService: Whether the customer subscribes to phone service.
- MultipleLines: Whether the customer has multiple phone lines.
- InternetService: The type of internet service subscribed by the customer (e.g., DSL, fiber optic).
- OnlineSecurity: Whether the customer has online security features.
- OnlineBackup: Whether the customer has online backup services.
- DeviceProtection: Whether the customer has device protection services.
- TechSupport: Whether the customer has technical support services.
- StreamingTV: Whether the customer subscribes to streaming TV services.
- StreamingMovies: Whether the customer subscribes to streaming movie services.
- Contract: The type of contract the customer has (e.g., month-to-month, one year, two years).
- PaperlessBilling: Whether the customer receives electronic bills instead of paper bills.
- PaymentMethod: The method used by the customer to make payments (e.g., electronic check, mailed check, bank transfer).
- MonthlyCharges: The amount charged to the customer each month.
- TotalCharges: The total amount charged to the customer over the entire tenure.
- Churn: Whether the customer has churned (i.e., stopped using the company's services).

There are 7043 entries of customer. There are 21 columns of which

18 are categorical or text data and 3 are numerical data.

2. Data Preprocessing:

- There are no missing values in the dataset.
- There are no duplicate values in the dataset.
- There are only three numerical columns and they are: Monthly Charges, Tenure, SeniorCitizen.
- There are no outliers for this.

a. Feature Engineering:

There are lot of categorical variable and these influence the churn a lot. So to make further use of these, it’s better to convert this into numerical values. This is done by using Binary Encoding.

3. Exploratory Data Analysis:

- Knowing the relation between the feature variables and the output variable is important.
- Knowing which features are important for the customer churn is also an important insight.

Insights:

1. 26.5% customers churn out
2. Customers who have monthly charges over 75 are more likely to churn out
3. Customers who have tenure less than 12 months are more likely to churn out
4. Customers who don’t have partners neither dependant are more likely to churn out
5. Customers who senior citizens are more likely to churn out
6. Customers who have monthly billing plans are more likely to churn out
7. Customers who have paperless bill are more likely to churn out
8. Customers with no techsupport, no device protection, no online backup, no online security are most likely to churn out
9. Customers who have fibreoptic internet service are most likely to churn out
10. Customers who have electronic check as payment mode are most likely to churn out
11. Gender doesn’t have any significant impact on churn in.

4. Model Building:

- To ensure that all features to have same scale normalizing the data is best. So used MinMaxScaler to Normalize the data.
- The dataset is divided into train and test dataset. Form the entire dataset 70% of the data is divided into train dataset and remaining 30% data is divided into Test dataset.
- X_Train, X_Test contains all the necessary feature variables.
- Y_Train , Y_Test contains the Output variable called "**Churn**".

A. Logistic Regression:

Logistic regression is suitable for churn prediction because it is designed for binary classification problems, where the goal is to predict one of two possible outcomes. Churn prediction typically involves predicting whether a customer will churn (leave) or not, which is a binary outcome (yes/no, 1/0).

- Used **sklearn** library for building the logistic Regresssion.
- Used X_Train, Y_Train for fitting/building the model.
- Predicted the Churn variable using the X_Test data which is unseen data for the model.

Evaluation:

For evaluating the model confusion matrix and Accuracy score has been used.

1. The **confusion matrix** is as shown bellow:

```
[[1382, 144]
 [ 278, 309]]
```

Interpreting the matrix:

- The model correctly predicted 1382 instances as positive (customers who churned and were predicted to churn).
- The model incorrectly predicted 144 instances as positive (customers who did not churn but were predicted to churn).
- The model incorrectly predicted 278 instances as negative (customers who churned but were predicted not to churn).
- The model correctly predicted 309 instances as negative (customers who did not churn and were predicted not to churn).

2. **Accuracy** of the model is 0.80.

B. Random Forest:

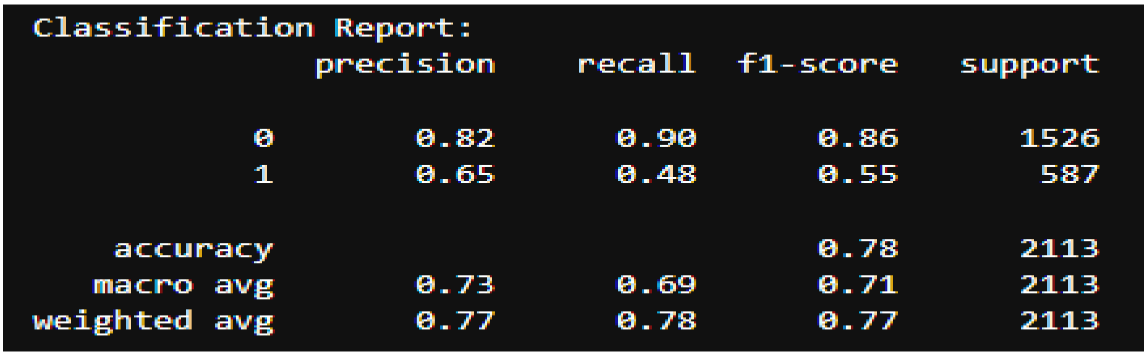
Random Forest is a powerful and versatile machine learning algorithm that is well-suited for churn prediction. Random Forest can effectively handle imbalanced datasets by adjusting class weights.

- Used **sklearn** library for building the Random Forest.
- Used X_Train, Y_Train for fitting/building the model.
- Predicted the Churn variable using the X_Test data which is unseen data for the model.

Evaluation:

- Accuracy of the model is 0.78

Classification report:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1526
           1       0.65      0.48      0.55       587

    accuracy                           0.78      2113
   macro avg       0.73      0.69      0.71      2113
weighted avg       0.77      0.78      0.77      2113
```

C. Gradient Boosting.

Gradient Boosting is an ensemble learning technique that combines the predictions of multiple weak learners (typically decision trees) sequentially. It builds trees iteratively, with each new tree focusing on the errors made by the previous ones. This iterative process allows Gradient Boosting to gradually improve prediction accuracy and capture complex patterns in the data, making it well-suited for churn prediction tasks where the relationships between features and churn may be nonlinear and intricate.

- Used **sklearn** library for building the Gradient Boositng.
- Used X_Train, Y_Train for fitting/building the model.
- Predicted the Churn variable using the X_Test data which is unseen data for the model.

Evaluation:

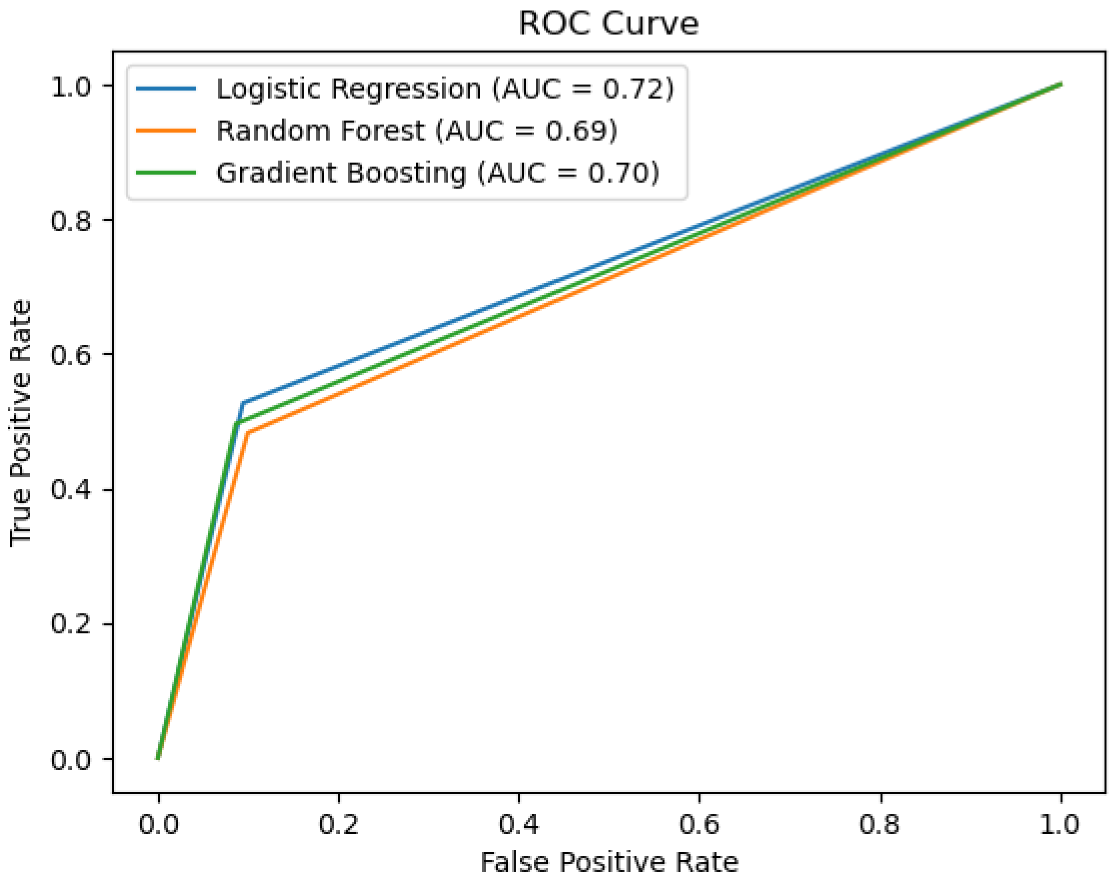
- Accuracy of the model is 0.80.
- Weighted Average of F1-Score is 0.79.

Confusion matrix is as follows:

- The model correctly predicted 1394 instances as positive (customers who churned and were predicted to churn).
- The model incorrectly predicted 132 instances as positive (customers who did not churn but were predicted to churn).
- The model incorrectly predicted 296 instances as negative (customers who churned but were predicted not to churn).
- The model correctly predicted 291 instances as negative (customers who did not churn and were predicted not to churn).

Finding the best Model:

- Used ROC curve to evaluate the models.



By Seeing the graph, Logistic regression is the best model for Churn prediction with AUC of 0.72, then Gradient Boosting with AUC of 0.70 and Random Forest with AUC as 0.69.