Report : Customer Churn Prediction

Name- Chaitanya Jindal [[chaitanya.jindal2002@gmail.com](mailto:chaitanya.jindal2002@gmail.com)]

# Data Collection

The dataset was obtained from a CSV file named "customer\_churn\_large\_dataset.csv." This dataset consists of various customer-related attributes, including demographic information, subscription details, and whether a customer has churned or not.

# Data Preprocessing

* **Handling Missing Values:** A crucial initial step in data preprocessing is checking for missing values. Fortunately, in this dataset, there were no missing values detected, indicating that it was clean and complete.
* **Duplicate Rows:** Duplicate rows were checked for and none were found in the dataset, implying that each row represented a unique customer.
* **Removing Irrelevant Columns:** The "CustomerID" and "Name" columns were dropped as they do not contribute to the predictive power of the model.

# Feature Engineering

* **Label Encoding:** To handle categorical features, label encoding was applied to "Gender" and "Location" columns. Label encoding converts categorical values into numeric form, which can be used by machine learning algorithms.
* **Min-Max Scaling:** Numerical features, such as "Age," "Subscription\_Length\_Months," "Monthly\_Bill," and "Total\_Usage\_GB," were scaled using Min-Max scaling. This technique transforms the features to a common range of values [0, 1], which can help improve model performance.

# Data Splitting

The dataset was divided into training and testing sets using a 70-30 split ratio. The training data was used to train the models, while the testing data was reserved for evaluating their performance.

# Model Building & Training

Five different classification models were employed:

1. **Logistic Regression**
2. **AdaBoost Classifier**
3. **Random Forest Classifier**
4. **Decision Tree Classifier**
5. **K-Nearest Neighbors (KNN) Classifier**

Each model was trained on the training data and subsequently evaluated using the testing data.

# Accuracy Comparison

The primary metric used to assess the models' performance was accuracy. However, the accuracy results for all five models were consistently around 50%, which suggests that the models struggled to differentiate between customers who churned and those who did not.

**Note on Accuracy :**It's essential to note that achieving an accuracy of 50% may indicate that the models are performing no better than random chance in predicting customer churn. This could be due to several factors:

1. **Imbalanced Data:** If the dataset has a significant class imbalance (e.g., a large number of non-churning customers and a small number of churning customers), the models might perform poorly.
2. **Feature Selection:** The selection of features can significantly impact model performance. It's possible that more relevant features or advanced feature engineering techniques could improve predictive accuracy.
3. **Hyperparameter Tuning:** Model hyperparameters play a crucial role in their performance. A more exhaustive search for optimal hyperparameters could potentially yield better results.
4. **Data Quality:** The quality and relevance of the data itself can affect model performance. Ensuring that the dataset captures essential customer behavior and churn factors is vital.

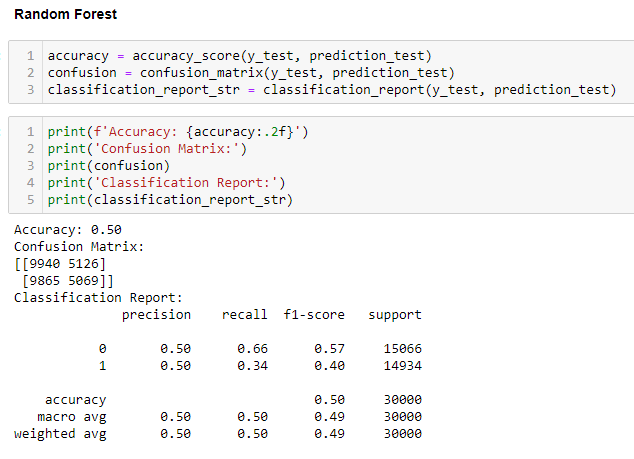
# Conclusion

In conclusion, while the initial models achieved an accuracy rate of 50%, further investigation and refinement are needed to build a more accurate churn prediction model. This might involve exploring different algorithms, addressing data imbalances, and fine-tuning model parameters to achieve a more reliable prediction of customer churn.

# Model Performance Metrics

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated