**Final Report: Customer Churn Prediction for T-Mobile**

**Objective**

**By:- Santosh Singh**

The objective of this analysis is to predict customer churn for **T-Mobile** based on customer features such as tenure, monthly charges, total charges, and other relevant variables. The goal is to evaluate and compare different machine learning models to identify the most accurate and reliable model for predicting churn, which will help T-Mobile take proactive measures in retaining customers.

**Data Overview**

The dataset contains the following key features:

* **Customer ID**: Unique identifier for each T-Mobile customer.
* **Tenure**: The number of months a customer has been with T-Mobile.
* **MonthlyCharges**: The amount a customer is charged monthly.
* **TotalCharges**: The total amount paid by the customer during their tenure with T-Mobile.
* **Churn**: Target variable indicating whether the customer has churned (1) or not (0).

**Data Preprocessing**

* **Duplicate Removal**: Any duplicate entries were removed from the dataset to ensure data integrity.
* **Missing Values**: The "TotalCharges" column contained missing values, which were handled by converting it to a numeric format and replacing missing values with the median.
* **Categorical Encoding**: Categorical variables, such as "gender", "Partner", "Dependents", were encoded using label encoding to convert them into numerical values.
* **Feature Scaling**: Numerical features (tenure, monthly charges, total charges) were scaled using StandardScaler to ensure uniformity in the data, which improves model performance.

**Exploratory Data Analysis (EDA)**

Exploratory analysis was conducted to understand the relationships between features and the churn outcome:

* **Churn Distribution**: The distribution of churned vs. non-churned customers was visualized using a count plot. T-Mobile's dataset had a slightly imbalanced churn distribution, with more non-churned customers.
* **Correlation Matrix**: A heatmap was generated to identify correlations between features. Notably, tenure showed a strong negative correlation with churn, implying that customers with longer tenure are less likely to churn.
* **Feature Distributions**: Histograms and KDE plots were created to visualize distributions for tenure, monthly charges, and total charges. Customers with lower monthly charges and longer tenure were less likely to churn.
* **Customer Behavior Insights**: Key insights include that customers with lower total charges and higher tenure are less likely to churn, whereas customers with higher monthly charges tend to show varied churn patterns.

**Model Building**

Three machine learning models were evaluated for predicting churn:

1. **Logistic Regression**
2. **Random Forest Classifier**
3. **Gradient Boosting Classifier**

Each model was trained using the preprocessed data, and the models' performance was evaluated using test data.

**Model Evaluation**

Performance metrics were used to evaluate each model:

* **Accuracy**: The proportion of correct predictions.
* **ROC AUC**: Measures how well the model distinguishes between churned and non-churned customers.
* **F1 Score**: Balances precision and recall, providing insight into the model's ability to correctly classify both churned and non-churned customers.

**Model Performance Summary**:

* **Logistic Regression**:
  + Accuracy: 80.5%
  + ROC AUC: 0.84
  + F1 Score: 0.78
* **Random Forest Classifier**:
  + Accuracy: 82.0%
  + ROC AUC: 0.86
  + F1 Score: 0.79
* **Gradient Boosting Classifier**:
  + Accuracy: 81.5%
  + ROC AUC: 0.85
  + F1 Score: 0.78

**Visualization**

Several visualizations were created to compare the models:

* **Confusion Matrices**: For each model, confusion matrices were plotted to show true positives, true negatives, false positives, and false negatives.
* **ROC Curves**: ROC curves were plotted to visualize the trade-off between true positive rate (TPR) and false positive rate (FPR).
* **Model Comparison Bar Plot**: A bar plot was created to compare the models based on accuracy, ROC AUC, and F1 score, with different colors to visually enhance clarity.

**Conclusion**

* **Best Model**: The **Random Forest** model outperformed the others in terms of accuracy, ROC AUC, and F1 score. It demonstrated a good balance between precision and recall, making it the most reliable model for predicting churn.
* **Logistic Regression**: Logistic Regression performed adequately but was slightly less accurate compared to Random Forest and Gradient Boosting.
* **Gradient Boosting**: The Gradient Boosting model had similar performance to Random Forest but showed slightly lower accuracy and ROC AUC.

**Future Work**

* **Hyperparameter Tuning**: Hyperparameters for Random Forest and Gradient Boosting models (such as n\_estimators, max\_depth) can be tuned to further improve performance.
* **Feature Engineering**: Including additional features like customer service interactions, payment history, or service usage could potentially improve model performance.
* **Deployment**: The Random Forest model can be deployed in a real-time churn prediction system, helping T-Mobile to take proactive actions to retain customers.

**Interactive Dashboard (Optional)**

An **interactive dashboard** could be created using **Streamlit** to allow stakeholders to explore the churn analysis results interactively:

* **Data Metrics**: Display summary statistics for the dataset.
* **Churn Distribution**: A bar chart showing the distribution of churned vs. non-churned customers.
* **Model Performance**: A table comparing accuracy, ROC AUC, precision, recall, and F1 score for each model.
* **ROC Curves**: Interactive visualization of ROC curves for each model.

This dashboard would allow stakeholders at T-Mobile to interact with the churn analysis and make data-driven decisions regarding customer retention strategies.