Report on Orange_churn Task:

Task Overview: This analysis aims to predict customer churn in the telecommunications sector by examining attributes such as account details and usage patterns. Insights gained will help businesses develop strategies to enhance customer retention and improve satisfaction.

Task Outline :

- Data Preparation and Exploration
- Data Analysis and Visualization
- Machine Learning Model Development
- Model Optimization and Evaluation

1- Data Preparation and Exploration

- Churn is the target variable indicating whether a customer has left the service.

- Account Length, International Plan, Voice Mail Plan are the Features related to customer service usage.

- The training dataset consists of **2666 rows** and **20 columns**.

- Column names were reviewed, revealing a mix of categorical and numerical features.

- categorical variables are State, International plan, and Voice mail plan.

- The value counts of the "Churn" variable were analyzed, revealing that approximately **85.4%** of customers did not churn, while **14.6%** did.

- Normalizing these counts emphasized the imbalance in the dataset.

- The columns (State and Area code) were identified as irrelevant for the analysis and dropped from the dataset.

2- Data Analysis and Visualization

- A heatmap was generated to visualize the correlation matrix of the training dataset. This matrix provides insights into how features are related to one another, particularly in relation to the target variable, "Churn."

- Variables such as "Total day minutes," "Total day charge," and "Total eve minutes" exhibited strong positive correlations with each other, indicating that higher usage leads to increased charges.

- The "Churn" variable showed moderate negative correlations with "Total day minutes" and "Total day charge," suggesting that customers who use more services are less likely to churn.

- Features like "Number vmail messages" and "Total intl charge" displayed minimal correlation with "Churn," suggesting these factors may not significantly influence customer retention.

3- Machine Learning Model Development & Model Optimization and Evaluation

- The dataset was split into features (X) and the target variable (y), with "Churn" as the target.

- Categorical features were transformed into numerical format using one-hot encoding to prepare the data for modeling.

- To address the imbalance in the churned and non-churned classes, the SMOTEENN technique was applied. This method combines oversampling of the minority class and undersampling of the majority class.

- The data was divided into training, validation, and testing sets

- The XGBoost classifier was chosen due to its effectiveness in handling binary classification tasks.

- A confusion matrix was generated to visualize the model's predictions against actual values in the test set.

- The evaluation of the model's performance was conducted using several metrics ,Balanced Accuracy Score: Measures the model's ability to classify both classes correctly "0.910", ROC AUC Score: Evaluates the model's ability to distinguish between churned and non-churned customers" 0.910", Accuracy Score: Overall correctness of the model" 0.963".

- The best parameters identified from the tuning were: Max Depth: 5 ,Learning Rate: 0.1 , Gamma: 0 ,Scale Pos Weight: 1

- The refined model displayed improved performance with the following results: Balanced Accuracy: "0.923"(to be filled with actual metric) , ROC AUC: "0.923", Accuracy: "0.971"