Churn Prediction Task Report

1. Introduction The objective of this project is to build a predictive model to identify customers likely to churn, leveraging the Orange Telecom's Churn Dataset. By analyzing customer data and developing a machine learning model, the aim is to assist the company in implementing effective customer retention strategies. The churn-80 dataset is used for training and cross-validation, while the churn-20 dataset is reserved for final testing.

---

2. Dataset Description

- **Context**: The dataset includes customer activity data and a churn label indicating whether a customer canceled their subscription.
- **Datasets**:
    - **churn-80**: Contains 80% of the data for training.
    - **churn-20**: Contains 20% of the data for final evaluation.
- **Features**: The dataset includes various features, such as the customer's State of residency and total day/eve/night minutes.

---

3. Approach

3.1 Data Exploration

- **Missing Values**: Checked for missing values.
- **Class Imbalance**: Analyzed the distribution of churn vs. non-churn customers.
- **Feature Correlation**: Generated a correlation matrix to identify relationships between features and the target variable.
- **Duplicates**: Checked for duplicates.

3.2 Data Preprocessing

- **Encoding**: Applied one-hot encoding to categorical features.
- **Feature Selection**: Retained features with high predictive power, identified through feature importance and correlation analysis.

3.3 Model Training

- **Algorithms**: Choose Tree-Based Algorithms such as Random Forest and XGBoost since there's no linear relationship between the features and the target.
- **Hyperparameter Tuning**: Conducted grid search and cross-validation to optimize model performance.

3.4 Model Evaluation

- **Metrics**: Evaluated models using accuracy, precision, recall and F1-score
- **Visualization**: Generated a confusion matrix and XGBoost Tree Visualization

---

4. Results

- **Model Performance**:
    - Best model: XGBoost
    - Accuracy: 0.96
    - macro avg Precision: 0.93
    - macro avg Recall: 0.88
    - macro avg F1-score: 0.91
- **Key Observations**:
    - Due to class imbalance Recall for "True" is only 78%

---

5. Conclusion This project demonstrates how machine learning can be applied to predict customer churn effectively. The selected model achieved high performance on the test set, with 0.96 accuracy. Insights from the analysis can guide Orange Telecom’s customer retention programs.
