```
# Customer Churn Prediction Project

## Overview

This project aims to predict customer churn for Lloyds Banking Group using advanced data analysis and machine learning techniques. The workflow includes data integration, exploratory data analysis (EDA), feature engineering, model training, evaluation, and business recommendations.

---

## Project Structure

```
├── Customer_Churn_Data_Large.xlsx   # Raw data (multiple sheets)
├── processed_customer_churn.csv     # Cleaned and feature-engineered data
├── customer_churn_eda.py            # EDA and preprocessing script
├── customer_churn_model.py          # Model training and evaluation script
├── plots/                           # Visualizations from EDA and modeling
│   ├── distribution_*.png
│   ├── correlation_matrix.png
│   ├── confusion_matrix.png
│   ├── roc_curve.png
│   └── feature_importances.png
├── requirements.txt                 # Python dependencies
└── README.md                        # Project documentation
```

---

1. Data Sources
```

-
- - **Transaction_History:** Customer purchases, amounts, and product categories.
- - **Customer_Service:** Service interactions, types, and resolution status.
- - **Online_Activity:** Login frequency, last login, and service usage.
- - **Churn_Status:** Target variable (churned or not).
-
- All sheets are merged on `CustomerID` to create a unified dataset.
-
- ---
-
- ## 2. Exploratory Data Analysis (EDA)
-
- - **Statistical Summaries:** Descriptive statistics for all features.
- - **Visualizations:** Histograms, bar plots, and correlation heatmaps to understand feature distributions and relationships.
- - **Feature Engineering:** Aggregated transaction, service, and activity data per customer.
-
- See the `plots/` directory for all generated visualizations.
-
- ---
-
- ## 3. Data Cleaning & Preprocessing
-
- - **Missing Values:** Imputed using median (numerical) or mode (categorical).
- - **Outliers:** Detected and capped or transformed as needed.
- - **Encoding:** Categorical variables one-hot encoded.
- - **Scaling:** Numerical features standardized.
-
- The final dataset is saved as `processed_customer_churn.csv`.

```
•
•  ---
•
•  ## 4. Model Development
•
•  - **Algorithm:** Random Forest Classifier (chosen for balance of
   accuracy and interpretability).
•  - **Training:** Performed with cross-validation and
   hyperparameter tuning (`GridSearchCV`).
•  - **Evaluation Metrics:** Precision, recall, F1 score, ROC-AUC,
   and confusion matrix.
•  - **Feature Importance:** Identified key drivers of churn.
•
•  All model results and plots are saved in the project directory.
•
•  ---
•
•  ## 5. Business Recommendations
•
•  - Use the model to identify at-risk customers and target them
   with retention strategies.
•  - Focus on the most important features (see
   `feature_importances.png`) for actionable insights.
•  - Regularly retrain the model with new data to maintain accuracy.
•  - Consider further improvements: advanced algorithms (e.g.,
   XGBoost), more feature engineering, or ensemble methods.
•
•  ---
•
•  ## 6. How to Run
•
•  1. **Install dependencies:**
•     ```bash
•     pip install -r requirements.txt
•     ```
•
```

2. **Run EDA and preprocessing:**

    ```bash
    python customer_churn_eda.py
    ```

3. **Train and evaluate the model:**

    ```bash
    python customer_churn_model.py
    ```

---

7. Results

- **Confusion Matrix:** `confusion_matrix.png`
- **ROC Curve:** `roc_curve.png`
- **Feature Importances:** `feature_importances.png`
- **Classification Report:** Printed in terminal after running the model script.

---

8. Contributors

- Your Name (Project Lead)

---

9. License

This project is for educational and internal use at Lloyds Banking Group.