Project Report: Classification Analysis on Customer Churn Data

---

1. Main Objective

The objective of this analysis is to **predict customer churn** using classification models and to **identify the key drivers** behind customer retention. The project is focused on both prediction and interpretation to help stakeholders understand the characteristics of customers who are likely to leave and to support strategic decision-making for customer retention.

---

2. Data Description

We used the **Telco Customer Churn dataset**, which contains information about a telecom company's customers, including demographic data, account information, and services used.

- **Source**: Public dataset from IBM Sample Data Sets
- **Total Records**: 7,043 customers
- **Target Variable**: Churn (Yes/No)
- **Main Features**: tenure, MonthlyCharges, TotalCharges, Contract, InternetService, PaymentMethod, etc.

Goal: To build classification models that can accurately predict if a customer will churn based on the available features.

---

3. Data Exploration & Cleaning

- **Missing Values**: Found in TotalCharges, filled using median imputation.
- **Data Types**: Converted TotalCharges from object to numeric.
- **Encoding**: One-hot encoding applied to categorical features such as Contract, InternetService, and PaymentMethod.
- **Feature Engineering**: Derived MonthlyAverageCharge = TotalCharges / tenure for customers with tenure > 0.
- **Class Balance**: Slight imbalance observed (~26% churn rate). SMOTE applied for balancing during training.

---

4. Modeling Summary

We trained and evaluated the following classifiers using an 80/20 train-test split:

a. Logistic Regression (Baseline Model)

- **Accuracy**: 80%
- **F1-score**: 0.58
- **Interpretability**: High
- **Pros**: Easy to explain to stakeholders
- **Cons**: Lower performance on minority class

b. Random Forest Classifier

- **Accuracy**: 85%
- **F1-score**: 0.66
- **Feature Importance**: Provided clear insight into key drivers (e.g., contract type, tenure)
- **Pros**: Good accuracy and interpretability
- **Cons**: Slightly slower training time

c. XGBoost Classifier

- **Accuracy**: 87%
- **F1-score**: 0.70
- **Pros**: Highest accuracy and robust to outliers
- **Cons**: Less interpretable, requires tuning

---

5. Model Recommendation

The **XGBoost Classifier** is recommended as the final model due to its superior performance in terms of accuracy and F1-score. While it is less interpretable, it provides strong predictive power. For interpretation, we supplemented

XGBoost with SHAP (SHapley Additive exPlanations) to visualize and explain feature impact.

---

6. Key Findings & Insights

- **Contract Type** is the strongest predictor. Customers on month-to-month contracts are more likely to churn.
- **Tenure** is inversely correlated with churn — longer-tenured customers are less likely to churn.
- **Monthly Charges**: Higher charges correlate with increased likelihood of churn.
- **Paperless Billing** and **Electronic Payment Methods** are associated with higher churn risk.

These insights can help the marketing team develop targeted campaigns for customer retention.

---

7. Next Steps

- **Feature Enhancement**: Incorporate additional behavioral data (e.g., service usage frequency).
- **Model Improvement**: Explore stacking models and deep learning techniques.
- **Business Action**: Use insights to design customer loyalty programs and improve contract offerings.
- **Monitoring**: Deploy the model and monitor predictions over time to ensure stability.

Appendix (Optional)

- Python notebook with data preprocessing, model training, evaluation metrics, and SHAP visualizations.