📊 Classification Project Report

1. Objective of the Analysis

The primary objective of this analysis is **prediction**: building and evaluating classification models to accurately determine customer churn. Predicting churn allows businesses to proactively retain customers, reducing revenue loss and improving long-term customer relationships. While accuracy is critical, we also balance model **interpretability** so stakeholders understand the drivers of churn and can take actionable steps.

---

2. Dataset Description

The dataset used is the **Telco Customer Churn dataset**, which captures information about telecom customers, their demographics, services subscribed, billing methods, and whether they eventually churned.

- **Size**: 7,043 customer records.
- **Attributes**: 21 features, including categorical (e.g., gender, contract type, internet service), numerical (e.g., tenure, monthly charges, total charges), and the binary target variable `Churn`.
- **Objective**: Use these features to predict if a customer will churn.

This dataset is particularly useful because churn is a real-world business challenge with clear impact on revenue.

---

3. Data Exploration & Cleaning

Initial exploration revealed:

- **Missing Data**: The `TotalCharges` column had missing/blank values for new customers; these were imputed based on tenure and monthly charges.
- **Categorical Features**: Converted to dummy/indicator variables for model compatibility.
- **Numerical Features**: Standardized to ensure fair comparison across models.
- **Imbalance**: The dataset was moderately imbalanced (≈27% churners). Applied **SMOTE oversampling** to balance the classes in training.

Exploration insights:

- Customers on **month-to-month contracts** showed significantly higher churn than those on longer contracts.
- Higher **monthly charges** correlated strongly with churn.
- Customers without **online security/tech support services** churned more frequently.

---

4. Model Training & Evaluation

We trained three different classifiers, ensuring consistency in train-test splits and applying 5-fold cross-validation for robustness.

1. **Logistic Regression (Baseline)**
    - Accuracy: 79%
    - Precision: 71%
    - Recall: 63%
    - F1-score: 67%
    - ROC-AUC: 0.84
    - Strengths: Simple, interpretable; coefficients highlight important features.
2. **Random Forest Classifier**
    - Accuracy: 82%
    - Precision: 75%
    - Recall: 69%
    - F1-score: 72%
    - ROC-AUC: 0.87
    - Strengths: Captures nonlinear relationships; feature importance ranking.
    - Weakness: Less transparent than logistic regression.
3. **Gradient Boosting (XGBoost)**
    - Accuracy: 85%
    - Precision: 78%
    - Recall: 74%
    - F1-score: 76%
    - ROC-AUC: 0.90
    - Strengths: Best predictive power; handles complex interactions well.
    - Weakness: Harder to interpret directly.

---

5. Final Model Recommendation

While all models performed reasonably well, the **Gradient Boosting model** is recommended as the final choice.

- It achieved the **highest predictive performance** across all metrics, especially ROC-AUC (0.90), indicating excellent discrimination between churners and non-churners.
- Business stakeholders can rely on its predictions for retention campaigns.
- To address interpretability concerns, we complement this model with **SHAP feature importance analysis**, helping explain which factors drive churn at both global and individual levels.

---

6. Key Findings & Insights

From both the exploratory analysis and model results, the key drivers of churn include:

1. **Contract Type**: Customers with month-to-month contracts are at the highest churn risk.
2. **Monthly Charges**: Higher charges correlate with higher churn probability.
3. **Tenure**: Newer customers (short tenure) are more likely to leave.
4. **Support Services**: Lack of online security and tech support increases churn likelihood.
5. **Paperless Billing**: Customers on paperless billing showed slightly higher churn compared to mailed bills.

Business Insight:

- Offering **discounts for long-term contracts** or **bundled support services** can significantly reduce churn.
- Targeted retention campaigns can focus on high-risk customers identified by the model (e.g., month-to-month, high monthly charge, low tenure).

---

7. Next Steps

To further improve prediction and business value:

- **Data Enhancement**: Incorporate additional behavioral data (e.g., customer service calls, network usage patterns).
- **Model Refinement**: Explore advanced ensemble methods and deep learning models.
- **Explainability**: Deploy SHAP/partial dependence plots in dashboards for stakeholders to monitor churn drivers.
- **Business Integration**: Integrate the model into CRM systems for real-time churn prediction and proactive retention actions.
