**Short Report**

**1. Objectives**

The goal was to build a machine learning model to predict customer churn based on behavior, demographics, and usage features. The aim was to analyze patterns, reduce churn, and provide actionable insights.

**2. EDA Observations**

`Support Calls` and `Payment Delay` showed high positive correlation with churn.

`Total Spend` and `Tenure` showed negative correlation with churn.

Class imbalance was moderate.

Outliers were present in `Payment Delay` and `Support Calls`, but retained.

Categorical variables such as `Subscription Type` and `Contract Length` showed clear distribution differences across churn.

**3. Approach (Modeling Pipeline)**

Data Cleaning: Removed nulls, encoded categorical variables.

Feature Engineering: Created `Calls\_per\_Month`, `Spend\_per\_Month`, and `Age\_Group`.

Scaling: Standardized all numerical features.

Feature Selection: Used LassoCV to reduce features.

Dimensionality Reduction: Applied PCA and LDA for insights.

Modeling: Trained Logistic Regression, Random Forest, Gradient Boosting.

Tuning: Optimized Random Forest using GridSearchCV.

---

**4. Comparison of Models & Final Model Explanation**

| Model | Accuracy | F1-Score |

| ----------------------- | -------- | -------- |

| Logistic Regression | 0.76 | 0.78 |

| Random Forest | 0.80 | 0.81 |

| Gradient Boosting | 0.82 | 0.83 |

| Tuned Random Forest | 0.84 | 0.85 |

Final model: Tuned Random Forest, selected based on best test set accuracy and F1-score.

**5. Key Findings**

Customers with frequent support calls and high payment delays are more likely to churn.

Customers with low tenure and low total spend also tend to churn.

Categorical features like `Contract Length` and `Subscription Type` were significant churn drivers.

**6. Strengths, Weaknesses, and Error Analysis**

**Strengths:**

Solid feature engineering and model tuning improved results.

Cross-validation added robustness.

**Weaknesses:**

Slight imbalance in churn labels may still bias predictions.

Outliers were not capped, which may affect models like Logistic Regression.

**Error Analysis:**

Most errors occurred near decision boundary — probability threshold tuning could improve results.

**7. Conclusion & Next Steps**

The tuned Random Forest model performed best and can be deployed for churn prediction. Future work can include:

SMOTE for balancing the dataset.

Trying advanced models like XGBoost or LightGBM.

Building a customer risk dashboard for business teams.