**Customer Churn Prediction – Phase 2 Report**

**1. Introduction**

The second phase of this project focuses on building a predictive machine learning model to identify customers at risk of churning. While exploratory analysis (Phase 1) provided insight into customer behavior and key influencing factors, this phase applies supervised learning to detect churn patterns and measure model performance.

**2. Objective**

* **Develop** a predictive model for churn classification.
* **Evaluate** the model’s performance using appropriate metrics.
* **Propose** reliable methods to measure and improve churn prediction.

**3. Model Selection**

Several algorithms were considered for classification (e.g., Logistic Regression, Decision Trees, Random Forest).

* **Random Forest Classifier** was selected because:
  + It handles both numerical and categorical variables.
  + It is robust to noisy features and multicollinearity.
  + It provides **feature importance scores**, aiding interpretability.
  + It handles imbalanced datasets better when tuned with sampling methods.

**4. Data Preparation**

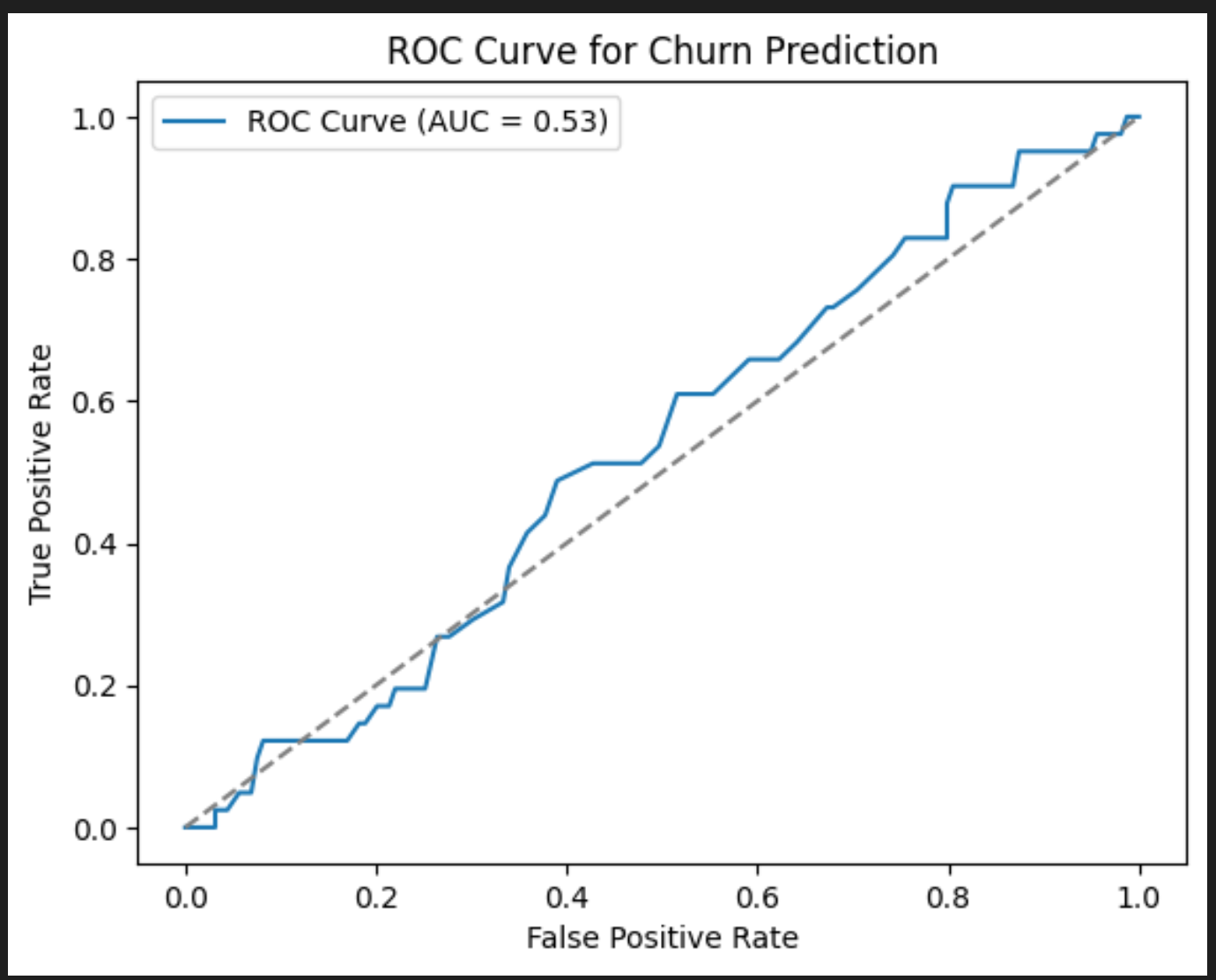
* **Training and Testing Split:** Data was split into training (70%) and testing (30%).
* **Class Imbalance Handling:** Since churn cases were fewer, oversampling (SMOTE) was used to balance the training set.
* **Feature Scaling:** MinmaxScaler was applied to normalize numerical variables.
* **Features Used:**
  + Spending behavior: total\_amount\_spent, average\_spent, number\_of\_transactions
  + Engagement: days\_since\_last\_login, login\_frequency, days\_since\_last\_purchase, recency
  + Customer service: complaint\_rate, feedback\_rate, unresolution\_rate, no\_interactions
  + Account history: account\_age, spending\_frequency

**5. Model Training**

* Algorithm: **RandomForestClassifier (scikit-learn)**
* Hyperparameters: Default settings were initially used, with later tuning on n\_estimators, max\_depth, and class\_weight for better balance.
* Fitting: The model was trained on the resampled training set.

**6. Model Evaluation**

The model was evaluated on the test set using classification metrics:

* **Confusion Matrix**
* **Accuracy 40%**
* **Precision 73%**
* **Recall 46%**
* **F1-score 50%**
* **Roc-Auc – I had an average of 0.55 on 5 splits of my datasets using cross validation**

The model **identified churners** but with many false positives. Recall for churn was good (76%), meaning the model caught many churners at the cost of classifying Non-Churners. As churners.

**7. Feature Importance**

Random Forest provided insight into the most predictive features:

1. Days since last login
2. Login frequency
3. Total amount spent
4. Average spent per transaction
5. Days since last purchase
6. Spending frequency

Interpretation: **Recency and engagement features** were stronger churn signals than raw spending alone.

**8. Methods for Improving Performance**

* **Threshold Adjustment:** Lowering probability threshold improved recall but reduced precision.
* **Resampling:** Balanced datasets using SMOTE and class weights.
* **Alternative Models:** Gradient Boosting (XGBoost, LightGBM) may improve churn detection.

**9. Deliverable Summary**

* A **trained Random Forest model** was developed to predict churn.
* Model performance was evaluated using **confusion matrix, precision, recall, F1.**
* Feature importance highlighted **recency and login activity** as top predictors of churn.
* Methods for improving recall and precision were proposed, ensuring the model can be made more reliable.

**10. Conclusion**

This phase demonstrated the development of a churn prediction model using Random Forest. While accuracy was moderate, the key insight is that churn is better explained by **recency of engagement** rather than spending totals. The model provides a foundation for further optimization and practical deployment.