Project 2 Model Template

Finding Fraud Faster

< your name here >

# Executive Summary

## Analysis

## Recommendations

## METHODOLOGY

**Data Exploration and Preprocessing**

1. **Exploratory Data Analysis (EDA) & FEATURE SCREENING**: Conduct an initial analysis to understand the data's characteristics, including distribution of the target variable, missing values, and potential outliers. PUT SOME CHARTS AND TABLES in this section!
2. **Data Preprocessing**: Address missing values, encode categorical variables, and standardize numerical features to prepare the data for modeling. PUT A TABLE of DATA TRANSFORMATIONS in this section.

**Model Development**

1. **Model Training**: Develop models using Logistic Regression, Random Forest, and GBM/XGBoost on the training data.
2. **Parameter Tuning**: Optimize model parameters to enhance performance.
3. **Feature Selection**: Identify and retain the most informative features for the models.

**Model Evaluation**

1. **Performance Metrics**: Evaluate models using accuracy, AUC-ROC, precision, recall, and F1-score. PUT A TABLE HERE
2. **Feature Importance Analysis**: Determine the most influential features in predicting fraudulent transactions. MAKE put a SET OF Charts HERE with Top N variables.
3. **Model FPR/TPR/Threshold Table – Make a table**

|  |  |  |
| --- | --- | --- |
| **Target False Positive Rate** | **True Positive Rate (TPR)** | **Prob Threshold** |
| 1% | 50.7% | 0.935 |
| 2% | 67.3% | 0.892 |
| 3% | 68.0% | 0.878 |
| 4% | 77.3% | 0.789 |
| 5% | 80.7% | 0.743 |
| 6% | 88.0% | 0.634 |

**Insights and Recommendations**

1. **Model Comparison**: Compare the models based on their performance and feature importance scores to identify the most effective model.
   1. **Table of ROC**-AUC, Precision, Recall, and F1 on TEST set.
   2. **ROC Charts for each model on Test Set.**
2. **Feature Evaluation**: Discuss the importance of email domain and billing postal code as predictors of fraud.
3. **Operational Strategy at 5% FPR**: Propose a strategy to achieve and maintain a 5% false positive rate, detailing its implications on recall and precision.

*HINT Example: IF predicted\_proba >= 0.743 THEN fraud. This rule will catch 80% of all frauds, while incorrectly classifying 5% of legitimate transactions as fraud. Additionally the precision at 0.743 and above is XX%*

**Plain Language Explanations**

1. **Random Forest vs. GBM/XGBoost**: Explain the differences between Random Forest and Gradient Boosting models in simple terms, focusing on their approach to building trees and correcting errors.
2. **Understanding 5% False Positive Rate**: Describe what operating at a 5% false positive rate means in practical terms, emphasizing its impact on customer experience and fraud detection accuracy.