Fraud Detection Project

# Can you identify fraudulent transactions?

You’ve just been invited to interview with the machine learning team at a large financial institution. Their job is to find fraud, waste, and abuse in their payment stream. You have been presented with a sample dataset of transactions and a holdout set, your job is to walk through your process of exploring the data, building model(s), and evaluating them. Your target variable is called EVENT\_LABEL and contains a label "legit" or "fraud". There is a data dictionary at the end of this project.

## Models to be Evaluated.

You will train and evaluate three different models.

* **Logistic Regression**: Serves as a baseline for performance comparison.
* **Random Forest**: An ensemble method known for its robustness and ability to handle complex data structures.
* **Gradient Boosting Machine (GBM) OR XGBoost**: Advanced ensemble techniques known for their predictive power.

## Methodology / Approach

Here you want to document your methodology, here is the standard approach, ,

**Data Exploration and Preprocessing**

1. **Exploratory Data Analysis (EDA)**: Conduct an initial analysis to understand the data's characteristics, including distribution of the target variable, missing values, and potential outliers.
2. **Data Preprocessing**: Address missing values, encode categorical variables, and prepare the data for modeling.

**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.
2. **Feature Importance Analysis**: Determine the most influential features in predicting fraudulent transactions.

Your write up should include 3-4 insights, 2-3 recommendations AND address the following:

* The Firm believes that email domain and billing postal code is an important predictor, your write up should discuss why or why not.
* Finally, the firm wants to operate at a 5% false positive rate, how can they do that? based on your best performing model what is the rule with predicted probability of fraud threshold that will give them a 5% false positive rate, what is the recall and precision at that threshold? what does it mean to operate at 5% false positive rate?
* Firm would like to understand in “plain language” the difference between Random Forest and Gradient Boosting; and how they can operate at a 5% False positive rate. For example, what does it mean to operate at 5% false positive rate?

**Insights and Recommendations**

1. **Model Comparison**: Compare the models based on their performance and feature importance scores to identify the most effective model.
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.

**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.

**Final Deliverables**

* **Report**: A comprehensive write-up detailing the analysis, model development and evaluation process, key insights, and operational recommendations for maintaining a 5% false positive rate.
* **Predictions on Holdout Set**: Apply the best-performing model to the **fraud\_holdout.csv** dataset and prepare a submission file with predictions.

You have been provided two datasets:

* fraud\_training.csv – use this one to train and evaluate your model
* fraud\_holdout.csv – use this one to score your hold out set submit your predictions to Kaggle (just include ID and P\_EVENT\_LABEL in your CSV)

Good luck!!

Appendix

|  |  |
| --- | --- |
| Variable | Description |
| EVENT\_ID | Transaction Identifier |
| account\_age\_days | number of days since the account was created |
| transaction\_amt | the USD $ value of the transaction |
| transaction\_adj\_amt | the adjustment USD $ value to the transaction |
| historic\_velocity | measure of the historic USD $ amount used to purchase goods and services |
| ip\_address | ip address of transactor |
| user\_agent | user agent of the transactor |
| email\_domain | email domain of the transactor |
| phone\_number | phone number of the transactor |
| billing\_city | billing city name |
| billing\_postal | billing postal code |
| billing\_state | billing state code |
| card\_bin | first 6 digits of the credit card (determines the card type, issuing bank, debit/credit/prepaid) |
| currency | original currency code |
| cvv | Card Verification Value - the 3 digit number on back fo your card |
| signature\_image | code for the signature |
| transaction\_type | code for the transaction type |
| transaction\_env | code for the transaction environment |
| EVENT\_TIMESTAMP | timestamp when the transaction occurred |
| applicant\_name | name of the transactor - ignore |
| billing\_address | billing address of the card holder |
| merchant\_id | merchant identifier |
| locale | browser locale |
| tranaction\_initiate | code for type of transaction initiation |
| days\_since\_last\_logon | days since last transaction initiated |
| inital\_amount | amount of first transaction USD $ |
| EVENT\_LABEL | TARGET fraud / legit |