Fraud Detection

# Can you identify transactions fraud?

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 the 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.

Your task is to build and compare 3 models (Random Forest, Decision Tree and a Logistic Regression). Your report should compare their performance to find the most effective model at predicting EVENT\_LABEL. What features where most useful. Be sure to follow the example executive summary from project 1. You’ll need 3-4 insights and 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 6% false positive rate, based on your best performing model what is the rule with score threshold that will give them a 6% false positive rate, what is the recall and precision at that threshold?

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)

In addition to your model report (executive summary and detailed analysis) the Firm would like to understand in “plain language” how Random Forest works and how they can operate at a 6% False positive rate. For example what does it mean to operate at 6% false positive rate.

* Executive Summary is worth 25 points
* Model report is worth 20
* Kaggle benchmark is up to 5

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 |