Fraud Detection and ML

Version 1.0

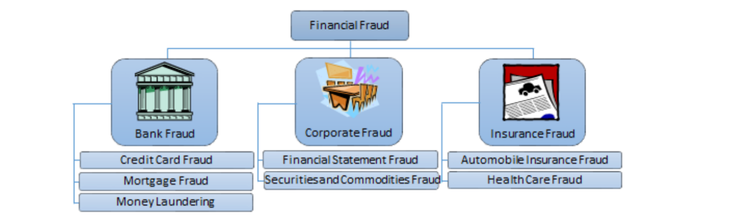
Dr. Ye Luo
HKU Business School

Jan 2023

Outline of Today

- Fraud Detection as a whole.
- Decision Tree.
- Model ensemble and Random Forest.
- Real examples.

Fraud


Financial Fraud
Bank Fraud
Corporate Fraud
Insurance Fraud
Credit Card Fraud
Mortgage Fraud
Money Laundering
Financial Statement Fraud
Securities and Commodities Fraud
Automobile Insurance Fraud
Health Care Fraud

Fraud is becoming a bigger and bigger issue

- "Fraud detection has been one of the major challenges for most organizations particularly those in banking, finance, retail, and e-commerce. This goes without saying that any fraud negatively affects an organization's bottom line, its reputation and deter future prospects and current customers alike to transact with it."
- "More often than not, for any fraud detected, the organization ends up paying for the losses. Additionally, it takes the good customers away from them while attracting more fraudsters."

How does modern organizations deal with fraud?

- Historical data are needed for the analysis.
- Help with the feature generation process.
- Classic machine learning or advanced machine learning algorithms.

Fraud Detection is not straight forward

- Changing fraud patterns over time. Fraudsters are always in the lookout to find new and innovative ways to get around the system.
- Class Imbalance. Practically only a small percentage of customers have fraudulent intentions. Consequently, there's an imbalance in the classification of fraud detection models (that usually classify transactions as either fraudulent or non-fraudulent) which makes it harder to build them.
- Model Interpretation. The model typically give a score indicating whether the transaction is likely to be fraudulent or not, without any explanations.
- Feature construction. Manually construct the features can be time and financially costly.

How much fraud do we have?

- Experts predict online credit card fraud to soar at a whopping 32 Billion USD in 2020.
- Fraud is larger than most of the blue-chip stock's profit, such as Coca-Cola (2 Billion), J.P. Morgan (23.5 Billion).
- Fraud is usually less than 1 percent. For many fintech firms such as Square, the fraud can be controller within 0.1 percent using modern data science. The break even line of business is usually around 0.5% to 1% of fraud. Ant Finance: fraud rate is under 0.2 percent. Large banks: less than 1%.

What kinds of fraud are there?

- Insurance claims.
- Medical claims and health care.
- E-commerce.
- Banking and credit card payments.
- Preventing loan application fraud.
- Money laundering.

Insurance Fraud

- Fake claims: NLP can help to detect fake and falsified claims. There are many hidden clues in these textual datasets. The rule-based engines don't catch the suspicious correlations in textual data, and fraud analysts can easily miss important evidence in boring investigation files.
- Duplicate claims and overstating repair cost
- Simple facts in the data:
    - Fraudulent claims are more likely not reported to police.
    - Old vehicles are more likely to be involved in fraud.
    - Eighty percent of accidents that happen during holidays involves fraud.
    - Scams are more likely to involve third parties (car dealers, repairers) than legitimate claims.

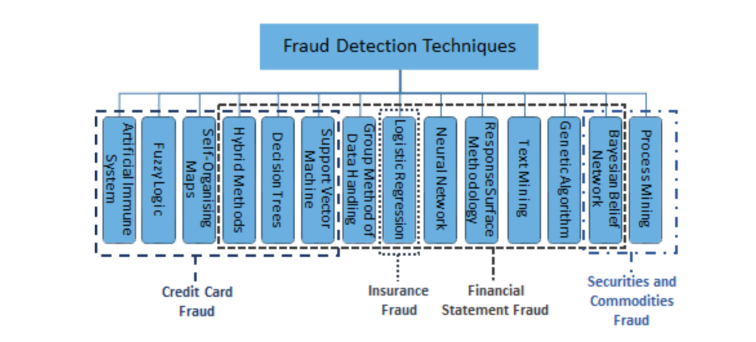
Banking and credit cards

- Data credibility assessment. Verify human identity via public sources (credit report by central banks) and transactions.
- Duplicate transactions from the merchant
- Account theft and unusual transactions.

Methodologies

- 1st Generation: Rule Based - expert system.
- 2nd Generation: Logistic regressions.
- 3rd Generation: Random Forest and Networks
- Future generation: DL?

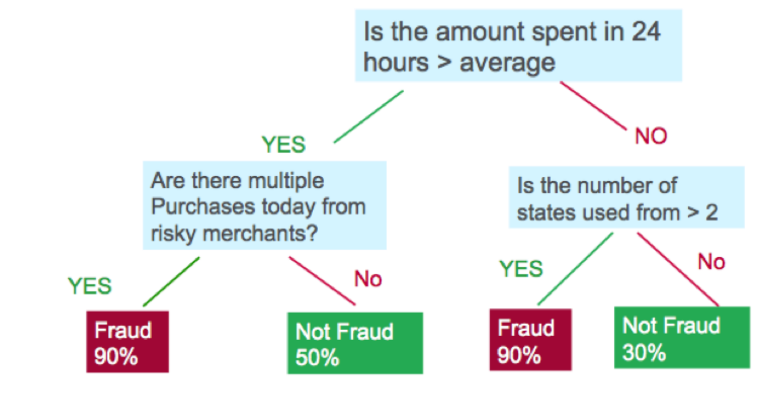
Fraud



Fraud Detection Techniques
Artificial Immune System
Fuzzy Logic
Self-Organising Maps
Hybrid Methods
Decision Trees
Support Vector Machine
Group Method of Data Handling
Logistic Regression
Neural Network
Response Surface Methodology
Text Mining
Genetic Algorithm
Bayesian Belief Network
Process Mining
Credit Card Fraud
Insurance Fraud
Financial Statement Fraud
Securities and Commodities Fraud

Decision Tree

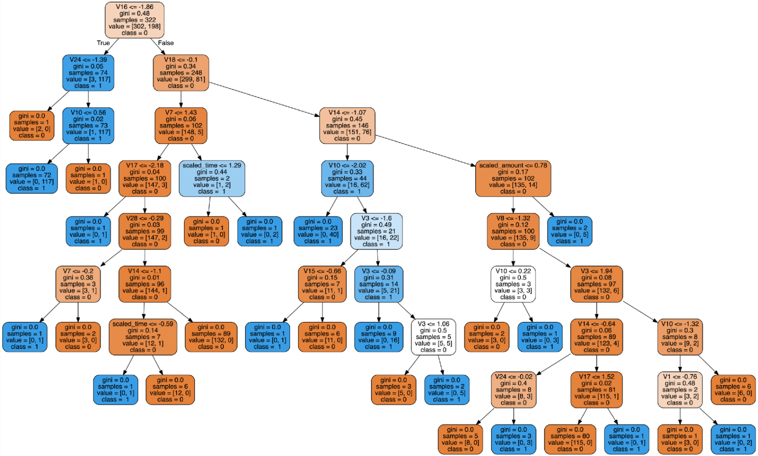
- Decision variables $x_1, ..., x_k$.
- Decision to be made: y.
- A decision is a mapping from $x_1, ..., x_k$ to y.
- For a tree, we use $1(x_j < \gamma)$ or $1(x_j \geq \gamma)$ as base functions(nodes or leaves), where γ is to be learned.
- Then the decision follows a tree structure, from top node to leaves.


Is the amount spent in 24 hours > average
YES
NO
Are there multiple Purchases today from risky merchants?
Is the number of states used from > 2
YES
No
YES
No
Fraud 90%
Not Fraud 50%
Fraud 90%
Not Fraud 30%

What is good about DT?

- Decision tress often mimic the human level thinking so its so simple to understand the data and make some good interpretations.
- Decision trees actually make you see the logic for the data to interpret(not like black box algorithms like SVM,NN,etc..)

Complex trees



How to train(grow) a tree: Evaluation at Every Leaf

- At each leaf, let

$$p_s := \frac{\text{Number of samples in s-class in the leaf}}{\text{Number of samples in the leaf}},$$

  $s = 1, 2, ..., K$.

- $K = 2$ for binary response variable - default risk prediction, fraud detection.
- Ideal case: Some $p_s = 1$, others $= 0$: pure class.
- The (Gini) Impurity measure: $\text{Gini} := 1 - \sum_{s=1}^{K} p_s^2$.

Splitting the Node

- At each leaf, propose a splitting rule at j^{th} variable x_j, $j = 1, 2, ..., p$.
- Split the node by cut-off rule $x_j < \gamma_j$ and $x_j \geq \gamma_j$.
- The splitting should improve the Gini index at the root.
- Find the maximum improvement over all possible splits (leaf, x_j and γ).

How to train(grow) a tree: a simple greedy algorithm

(0) Given a loss (Gini) function and an existing tree(you start with null).

(1) Given a tree, find a decision variable (with the best γ) that reduces the loss function the most, if attach to a leaf. (even simpler for dummy variables.)

(2) Construct a new tree based on step (1).

(3) Go back to (1).

- Greedy algorithms tend to be stumble and easy to overfit.

Formal Algorithm

Standard TIDIT algorithm: Random Forest For Binary Classification.

Input: a sequence of n samples (x_1, y_1),...,(x_n, y_n), $y_i \in \{0, 1\}$, $i = 1, 2, ..., n$.
For B = 1,2, ..., M do:

1. Sample with replacement and obtain $(x_1^B, y_1^B), ..., (x_n^B, y_n^B)$.
2. Fit a decision tree $h^B(x)$ to the bootstrapped data.

3. Construct a voting rule $h(x) = 1(\frac{1}{M}\sum_{B=1}^{M} h^B(x) > c)$. When $c = 0.5$, such rule is the majority voting rule.

How to train(grow) a tree: a simple ϵ greedy algorithm

(0) Given a loss (Gini) function and an existing tree(you start with null).

(1) Given a tree, find a decision variable (with the best γ) that reduces the loss function the most, if attach to a leaf. (even simpler for dummy variables.)

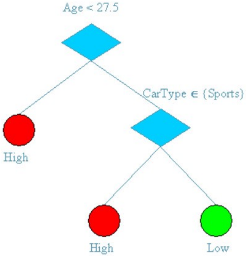
(2) Construct a new tree based on step (1) with probability $1 - \epsilon$. With probability ϵ, randomly pick a variable and a leaf, and split.

(3) Stop if some criterion is hit. Otherwise, go back to (1).

Grow a tree

Tid	Age	Car Type	Class
0	23	Family	High
1	17	Sports	High
2	43	Sports	High
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

Numeric Categorical



1) Age $<$ 27.5 $\Rightarrow$ High

2) Age $>=$ 27.5 and
CarType $=$ Sports $\Rightarrow$ High

3) Age $>=$ 27.5 and
CarType $\neq$ Sports $\Rightarrow$ High

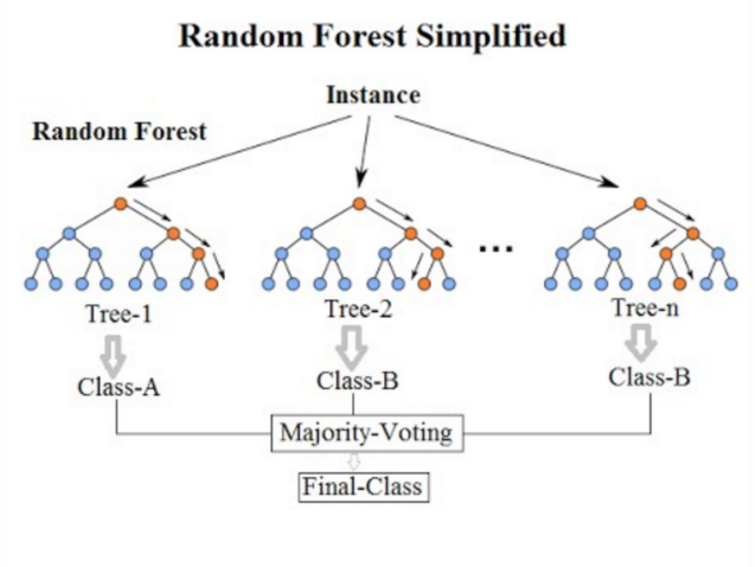
Problems with DT

- Single decision maker, easy to lead to wrong decisions if overfit.
- Need pruning to reduce overfit.
- Does not work well enough in practice.

Ensemble the models

- An ensembler is a method that combines multiple models together.
- A linear regression can be viewed as a linear ensembler of different variables.
- In practice, we would like to construct multiple models instead of a simple model.
- Random Forest instead of Decision Tree.

Random Forest(it is really a forest!)



How to create a forest?

- Main idea: you need different samples. Different trees will be generated from different samples.
- Let's randomly sample data from the data set(with or without repetition).
- Each tree will be constructed based on random samples.
- Trees vote on the decisions by offering their predictions - "majority vote"!
- Improved much of robustness!

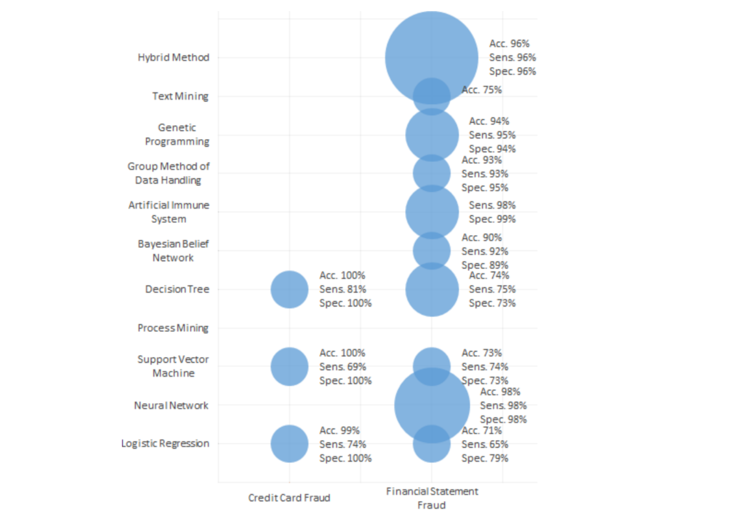
Some useful features

- IP address, phone number, national id - can identify a lot of information from national id and phone number, etc.
- Satellite/Tower data, geographic locations, clustering effects.
- Social network, firm ownership network, etc.

Undersampling: one more thing before going to practice

- Unbalanced sample creates bad decisions for detecting fraud - it will try to not to make type 2 error: judge good people as bad people.
- Therefore, one simple strategy is reweighting and undersampling: give larger weights to the bad points when training a model.
- Simple strategy(reweighting): weights on good point = number of bad points/number of good points, weights on bad point = 1.
- Undersampling: you sample the same amount of good and bad points when constructing the forest.
- Python method: sklearn.ensemble.RandomForestClassifier.

Survey on methods



Survey on methods: type 1 rate

Table 2. Sensitivity results for fraud detection practices

Research	Fraud Investigated	Method Investigated	Sensitivity
[3]	Credit card transaction fraud from a real world example	Logistic model (regression) Support vector machines Random forests	24.6-74.0% 43.0-68.7% 42.3-81.2%
[12]	Financial statement fraud from a selection of Greek manufacturing firms	Decision trees Neural networks Bayesian belief networks	75.0% 82.5% 91.7%
[19]	Financial statement fraud with financial items from a selection of public Chinese companies	Support vector machine Genetic programming Neural network (feed forward) Group method of data handling Logistic model (regression) Neural network (probabilistic)	55.43-73.60% 85.64-95.09% 67.24-80.21% 87.44-93.46% 62.91-65.23% 87.53-98.09%
[7]	Financial statement fraud with managerial statements	Text mining with singular validation decomposition vector	95.65%
[4]	Financial statement fraud with financial items from a selection of public Chinese companies	CDA CART Neural network (exhaustive pruning)	61.96% 72.40% 80.83%
[16]	Credit card fraud using legitimate customer transaction history as well as generic fraud transactions	Bayesian learning with Dempster-Shafer combination	71-83%
[9]	Financial statement fraud from Accounting and Auditing Enforcement Releases by the Securities and Exchange Commission	Genetic algorithm	13-27%
[25]	Transactional fraud in automated bank machines and point of sale from a financial institution	Coevolution artificial immune system Standard evolution artificial immune system	97.688-98.266% 92.486-95.376%

Survey on methods: type 2 rate

Table 3. Specificity results for fraud detection practices

Research	Fraud Investigated	Method Investigated	Specificity
[3]	Credit card transaction fraud from a real world example	Logistic model (regression) Support vector machines Random forests	96.7-99.8% 95.7-99.8% 97.9-99.8%
[12]	Financial statement fraud from a selection of Greek manufacturing firms	Decision trees Neural networks Bayesian belief networks	72.5% 77.5% 88.9%
[19]	Financial statement fraud with financial items from a selection of public Chinese companies	Support vector machine Genetic programming Neural network (feed forward) Group method of data handling Logistic model (regression) Neural network (probabilistic)	70.41-73.41% 89.27-94.14% 75.32-78.77% 88.34-95.18% 70.66-78.88% 94.07-98.09%
[7]	Financial statement fraud with managerial statements	Text mining with singular validation decomposition vector	95.65%
[4]	Financial statement fraud with financial items from a selection of public Chinese companies	CDA CART Neural network (exhaustive pruning)	80.77% 72.36% 73.45%
[9]	Financial statement fraud from Accounting and Auditing Enforcement Releases by the Securities and Exchange Commission	Genetic algorithm	98%-100%
[25]	Transactional fraud in automated bank machines and point of sale from a financial institution	Coevolution artificial immune system Standard evolution artificial immune system	95.862-97.122% 99.311%

A study on RF based fraud detection

- Liu et. al., International Journal of Economics and Finance, 2015.
- Data from China Stock market and Accounting Research (CSMAR).
- Try to look at listed companies involving manipulation of profits.

A study on RF based fraud detection

- Assumption: Company commits fraud in different years and its annual report meets that fraud samples selection and annual report from the non-fraud years meets the non-fraud samples.
- ST, *ST companies are excluded.
- Only involves manufacturing companies from 1998 to 2014.
- Create a balanced sample: 138 fraud companies and 160 non-fraud.

A study on RF based fraud detection

- Variables include: Debt to equity market, current asset ratio, fixed assets ratio.
- (incomes) Accounts receivable and income ratio. Inventory and income ratio, mobile asset turnover, fixed assets and income ratio
- (growth) Price earnings ratio, sale ratio, book value.
- (Profitability) return on invested capital, Long term capital gains, operating margin, return on assets, etc.

Results

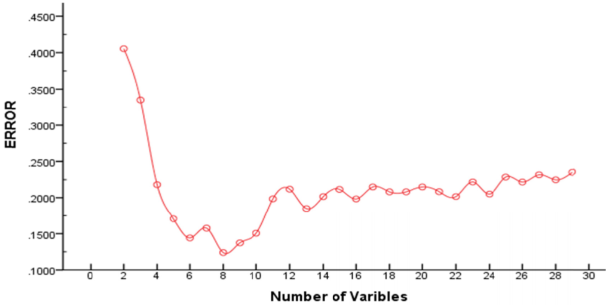
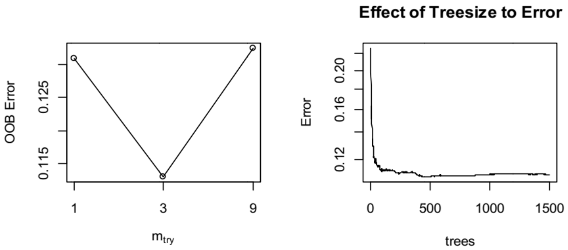


Figure 1. RF five-fold across-test

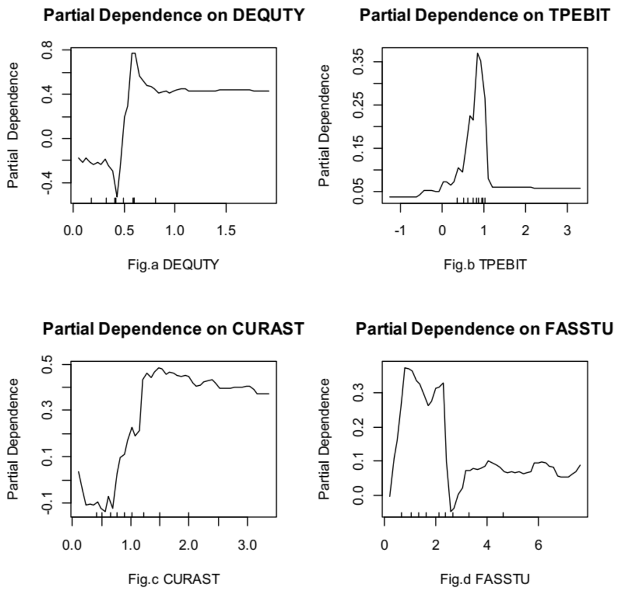
Results

Table 2. Variables in models

Variables for random forest	Variables for other models
TPEBIT	LOTCAG
PS	MANEXP
CURAST	PE
FASSTU	ACRESAL
TINEAR	CURAST
DEQUTY	WORCAP
CURASS	LTDCAP
FIXASS	



Results


Partial Dependence on DEQUTY
Partial Dependence
-0.4
0.0
0.4
0.8
0.0
0.5
1.0
1.5
Fig.a DEQUTY
Partial Dependence on TPEBIT
Partial Dependence
0.05
0.15
0.25
0.35
-1
0
1
2
3
Fig.b TPEBIT
Partial Dependence on CURAST
Partial Dependence
-0.1
0.1
0.3
0.5
0.0
1.0
2.0
3.0
Fig.c CURAST
Partial Dependence on FASSTU
Partial Dependence
0.0
0.1
0.2
0.3
0
2
4
6
Fig.d FASSTU

2d-Results

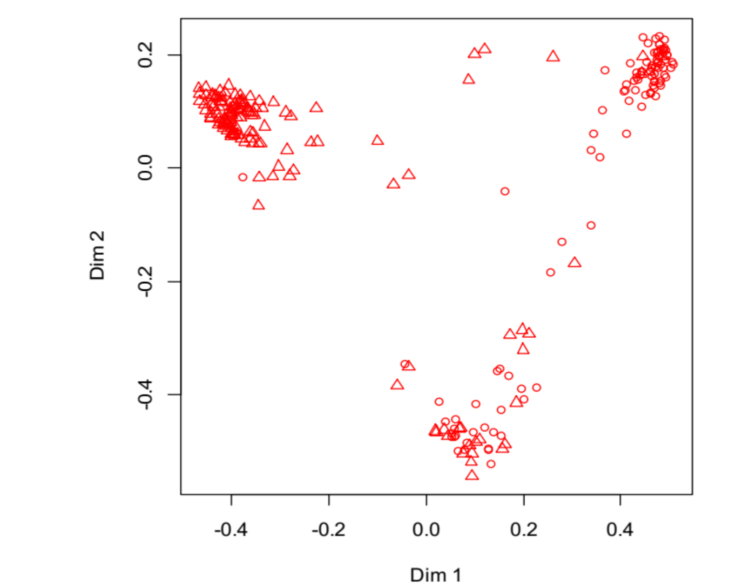


Table 5. Results of model test

Models	Fraud (%)	Nfraud (%)	Total (%)
Logistic	42.03	23.18	32.18
KNN	88.41	87.50	87.92
DT	77.54	83.13	80.54
SVM	66.67	80.00	73.83

Cross-validation

Table 6. Five-fold cross-validation results

Models	Nfraud (%)	Fraud (%)	Total (%)
Logistic	37.50	49.01	42.91
KNN	59.00	63.19	60.11
DT	68.13	64.62	66.43
SVM	81.88	78.13	80.18
RF	85.16	90.71	88.00