A Solution to Classification on Imbalanced Data

Credit Card Fraud Detection as an Example

J. He[1], J. Dong[1], Z. Jiang[1]

The Chinese University of Hong Kong, Shenzhen

[1]School of Science and Engineering

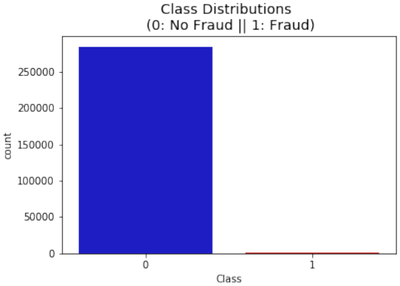
July 26, 2019

Table of Contents

Table of Contents

What's the Real World Problem?

- The observations in different categories can be **imbalanced**.
- The **cost** of false negative prediction and that of false positive prediction could be different
- Even we know the cost ratio of the two kinds of false prediction, it could be **dynamic** through the time.
- That means the conventional accuracy fail to give satisfying evaluation for the models



Correlation Heat Map

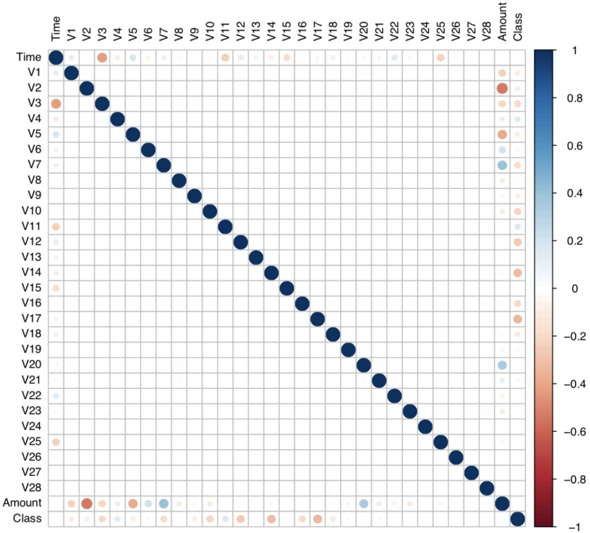


Table of Contents

The model selection algorithm

- Every model gives a response probability instead of a class so we can adjust the threshold.
- The weghted cost function for any classifier Θ.

$$L(\Theta) = \alpha \times FP + \beta \times FN$$

- Find the model with the lowest cost function.

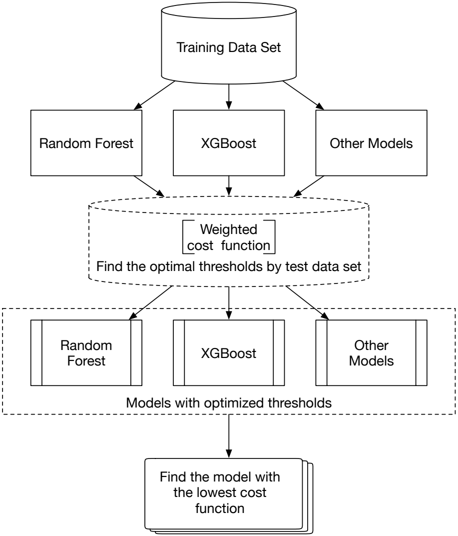
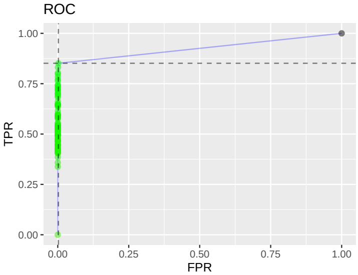
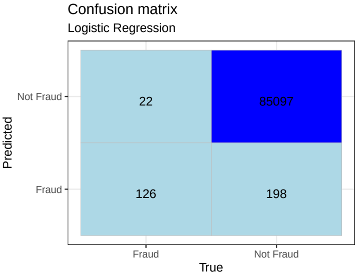
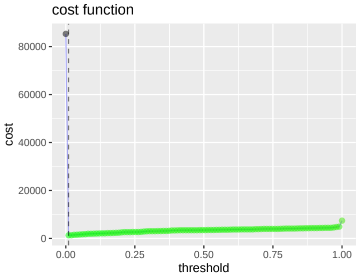


Table of Contents

Logistic Regression



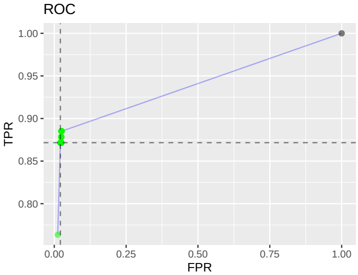


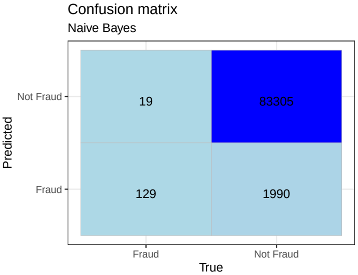


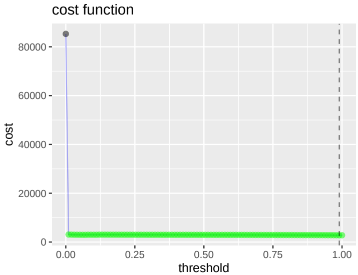
threshold at 0.01 − cost of FP = 1, cost of FN = 50

total cost = 1298

Naive Bayes



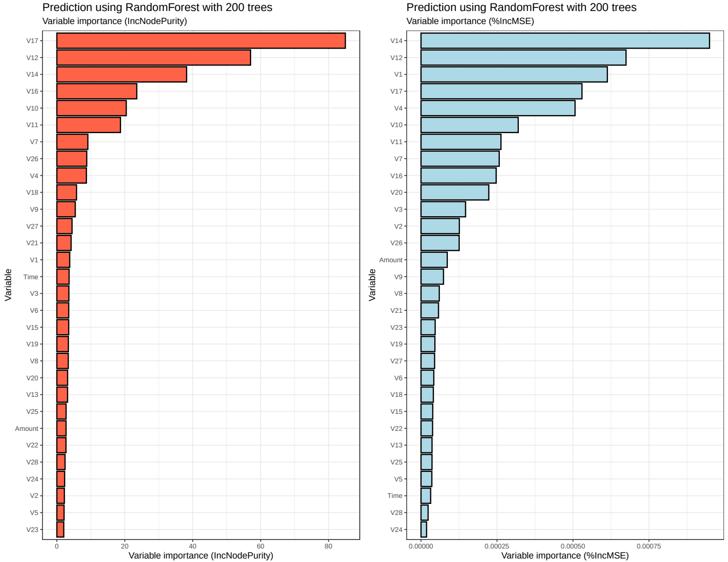




threshold at 0.99 − cost of FP = 1, cost of FN = 10
total cost = 2760

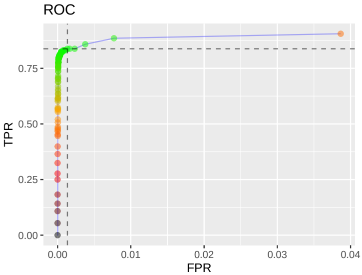
Random Forest: ntree = 200

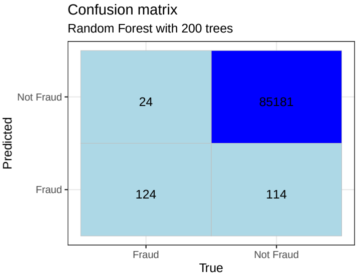
The importance map.

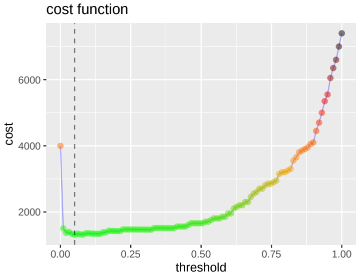


Random Forest: ntree = 200

The curves and confusion matrix







threshold at 0.05 − cost of FP = 1, cost of FN = 50

total cost = 1314

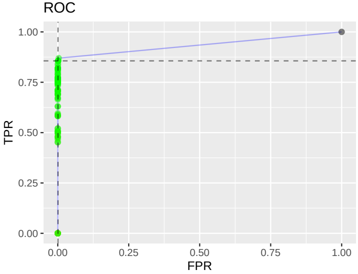
Boosting

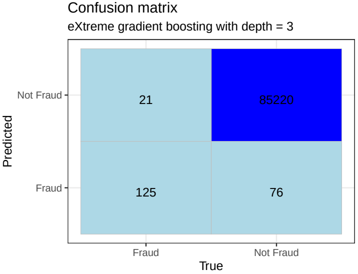
- Parameters adopted within each type of method:
  - Gradient boosting & Ada Boosting: # Trees = (500, 1000, 1500, 2000)
  - Depth: 3 and 7 (eXtreme gradient boosting)
- The best model selected: Extreme Gradient Boosting with depth = 3

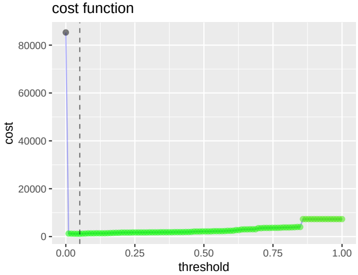
threshold	tpr	fpr	tp	fp	tn	fn	cost
0.0000000	1.0000000	1.0000000	146	85296	0	0	85296.00000
0.0101010	0.8698630	0.0034937	127	298	84998	19	1248.00000
0.0202020	0.8561644	0.0017234	125	147	85149	21	1197.00000
0.0303030	0.8561644	0.0013365	125	114	85182	21	1164.00000
0.0404040	0.8561644	0.0010903	125	93	85203	21	1143.00000
0.0505051	0.8561644	0.0008910	125	76	85220	21	1126.00000
0.0606061	0.8424658	0.0007386	123	63	85233	23	1213.00000
0.0707071	0.8356164	0.0006331	122	54	85242	24	1254.00000
0.0808081	0.8219178	0.0005510	120	47	85249	26	1347.00000
0.0909091	0.8219178	0.0005159	120	44	85252	26	1344.00000
0.1010101	0.8219178	0.0005159	120	44	85252	26	1344.00000
0.1111111	0.8150685	0.0004924	119	42	85254	27	1392.00000
0.1212121	0.8150685	0.0003048	119	26	85270	27	1376.00000

Boosting: Corresponding Curves

The best model Selected: eXtreme gradient boosting with depth $= 3$.







threshold at 0.05 − cost of FP = 1, cost of FN = 50
total cost = 1126

Model Comparison

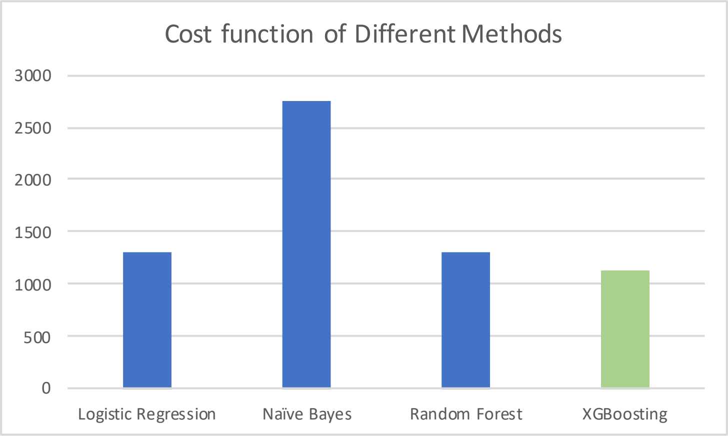


Table of Contents

Summary

- Models for extremely imbalanced data can not be directly evaluated by accuracy.
- XGBoosting and Logistic regression perform well.
- Advantages of our solution.
  - Utilizing flexible weighted cost function.
  - Only need to train one time.
- Disadvantages.
  - Binary classifiers are not suitable for this methodology.
  - The methodology does not improve the performance from data transformation.
- Further improvement.
  - One can try oversampling or undersampling to improve.
  - Cross validation can be implemented to prevent overfitting.