Fraud Detection

Weikai Mao

Problem

- Type: Binary classification problem.
- Target: Predict the probability that a transaction is fraudulent.

Preprocessing

- Handle missing values.
- Encode categorical variables.
- PCA to reduce dimension.

Modeling

Logistic regression; Random forest; Gradient boosting; XGBoost.

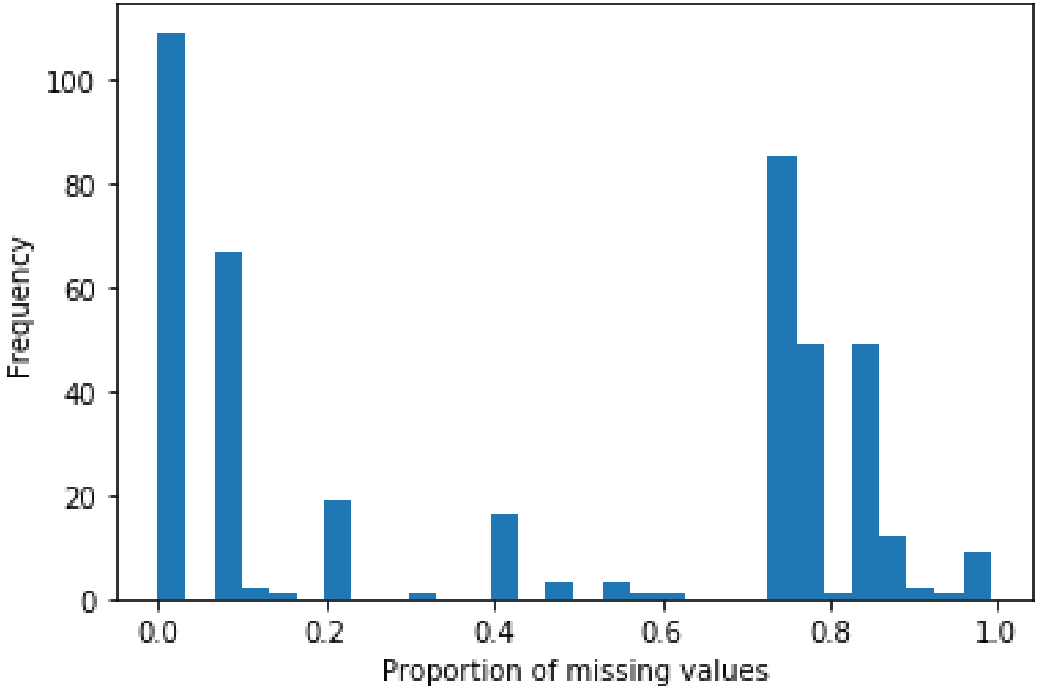
Data Preprocessing

Data Overview

- There are 433 predictors.
- Training set has 590540 observations, and test set has 506691.

Handle missing values

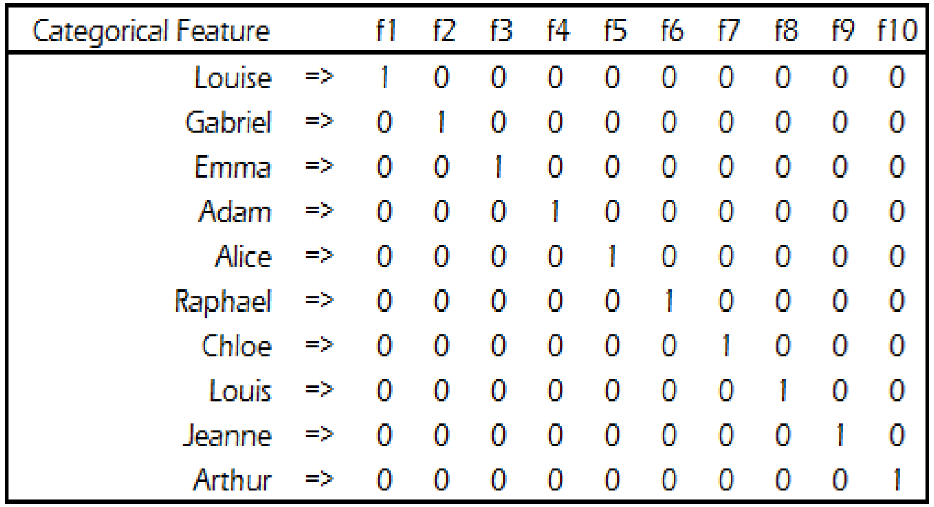
- Drop the variables with high proportion (**70%**) of missing values.
- Fill the missing values in **categorical** variables with their **mode**.
- Fill the missing values in **numerical** variables with their **mean**.



Data Preprocessing

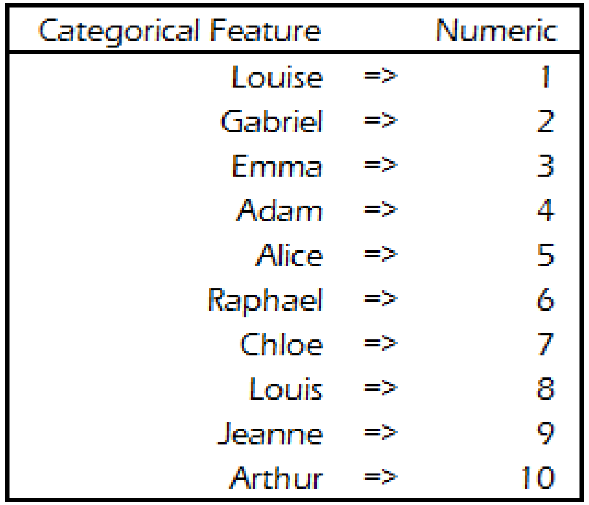
Encode categorical variables

- One-Hot Encoding



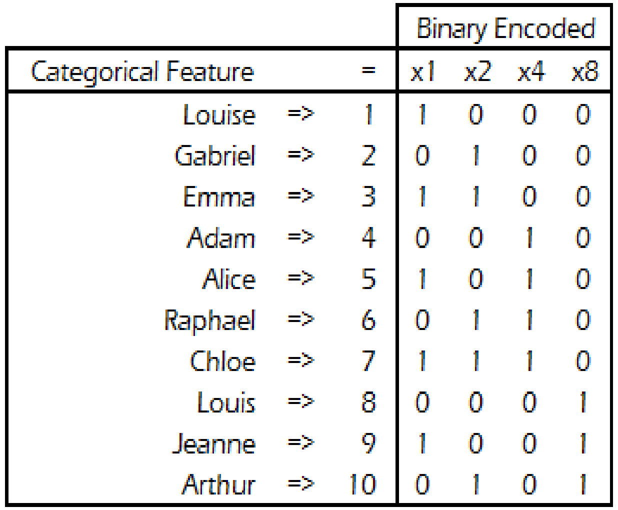
Categorical Feature		f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
Louise	=>	1	0	0	0	0	0	0	0	0	0
Gabriel	=>	0	1	0	0	0	0	0	0	0	0
Emma	=>	0	0	1	0	0	0	0	0	0	0
Adam	=>	0	0	0	1	0	0	0	0	0	0
Alice	=>	0	0	0	0	1	0	0	0	0	0
Raphael	=>	0	0	0	0	0	1	0	0	0	0
Chloe	=>	0	0	0	0	0	0	1	0	0	0
Louis	=>	0	0	0	0	0	0	0	1	0	0
Jeanne	=>	0	0	0	0	0	0	0	0	1	0
Arthur	=>	0	0	0	0	0	0	0	0	0	1

- Numeric Encoding



Categorical Feature		Numeric
Louise	=>	1
Gabriel	=>	2
Emma	=>	3
Adam	=>	4
Alice	=>	5
Raphael	=>	6
Chloe	=>	7
Louis	=>	8
Jeanne	=>	9
Arthur	=>	10

- Binary Encoding



			Binary Encoded			
Categorical Feature		=	x1	x2	x4	x8
Louise	=>	1	1	0	0	0
Gabriel	=>	2	0	1	0	0
Emma	=>	3	1	1	0	0
Adam	=>	4	0	0	1	0
Alice	=>	5	1	0	1	0
Raphael	=>	6	0	1	1	0
Chloe	=>	7	1	1	1	0
Louis	=>	8	0	0	0	1
Jeanne	=>	9	1	0	0	1
Arthur	=>	10	0	1	0	1

Prediction performance after encoding: Binary > Numeric > One-hot.

Tables: https://medium.com/data-design/visiting-categorical-features-and-encoding-in-decision-trees-53400fa65931

Modeling

Logistic regression

- The prediction score is 0.713617.

Bagging trees

- 100 trees.
- The prediction score is 0.896448.

Random forest

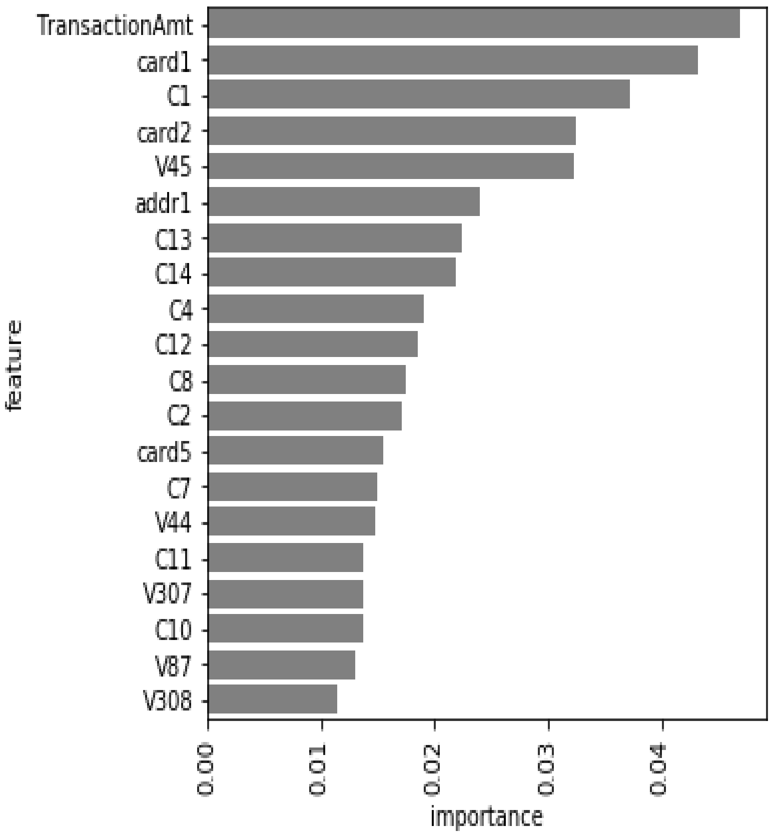
- 100 trees and 50 features.
- The prediction score is 0.895868.

max_features	n_estimators	PCA	prediction score
223	100	99%	0.871838
190	100	100%	0.892415
100	100	100%	0.894553
50	100	100%	0.895868
223	100	100%	0.896448
90	200	100%	0.898140
100	200	100%	0.899070
50	200	100%	0.900798
15	1000	100%	0.904874

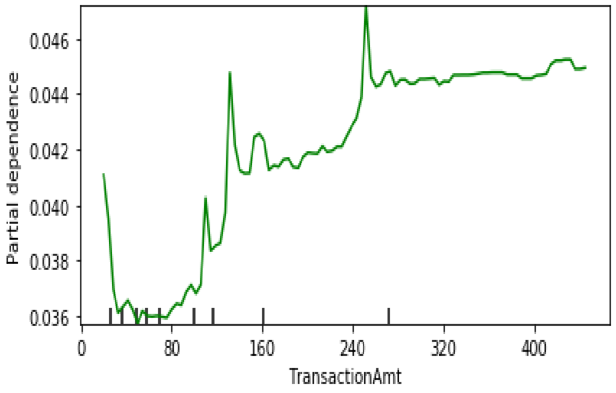
The key parameter is the number of trees (n_estimators).

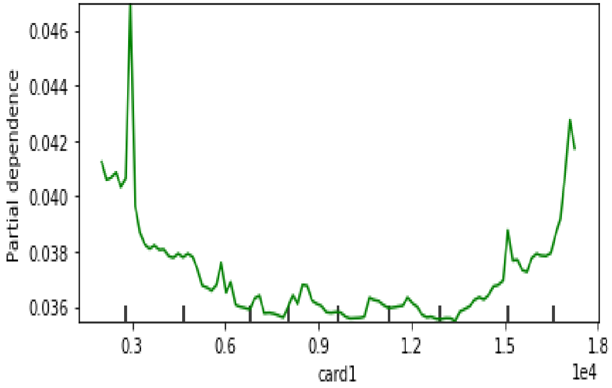
Modeling

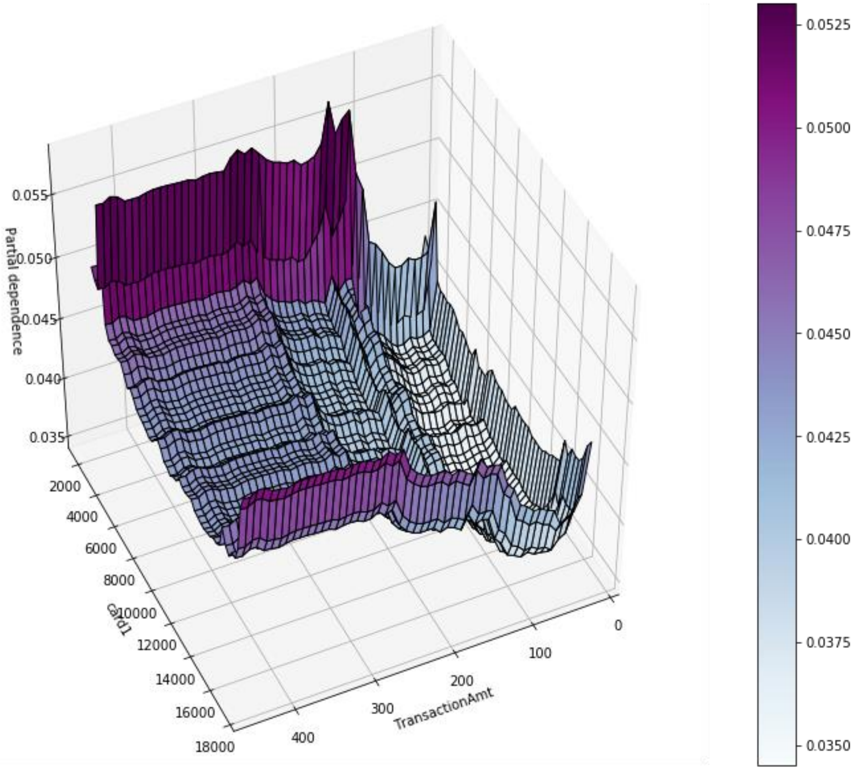
Importances



Partial Dependence Plots







Modeling

Use grid search to do parameters tuning.

Models	Parameters	Prediction Scores
Logistic regression	**-**	**0.871838**
Random forest	**max_features=15, min_samples_leaf=1, n_estimators=1000**	**0.904874**
Gradient boosting	**max_depth=10, min_samples_leaf=0.001, learning_rate=0.1, n_estimators=100**	**0.919523**
XGBoost	**Same as above**	**0.931355**