Fraud Detection

Group 19

01 Objective and Dataset

02 Methodology

03 Results

04 Next Steps

Objective and Data Set

• Objective

Predict which transaction is likely to be fraud, based on transaction information and personally identity information.

• Data Set

- The datasets from Vesta Corporation have two parts: ‘Identity’ and ‘Transaction’.
- In the joined dataset, there are 600K observations and 433 features (393 features of transaction and 41 features of identity).
- https://www.kaggle.com/c/ieee-fraud-detection/data

Note:

1. Most features, including transaction time, are anonymous.
2. Not all transactions observations have corresponding identity information;
3. The primary key, ‘TransactionID’, is unique.

Methodology

1 EDA

- Missing values
- Imbalanced data

2 Data Processing

- Missing values
- Imbalanced data

3 Feature Selection

- Filter Method
  - Fisher Criteria
- Wrapped Method
  - Forward Stepwise
- Embedded Method
  - LASSO

4 Model

- Logistic regression
- SVM
- Random Forest
- XGBoost
- Light GBM

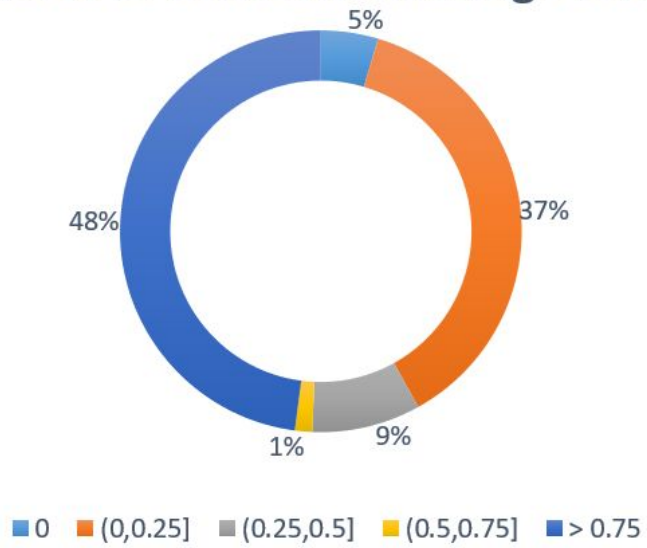
5 Evaluation

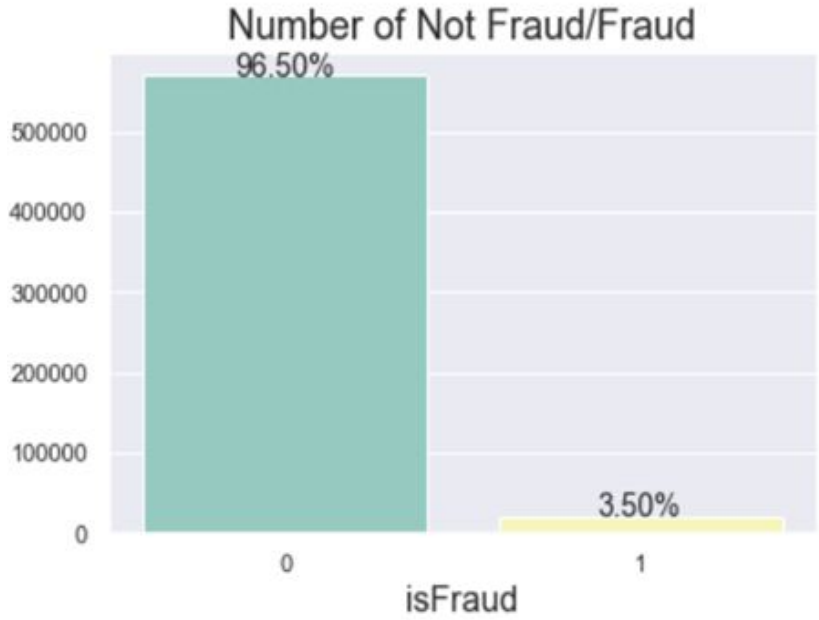
- AUC
- Precision/Recall
- F1-score

6 Prediction

Test our model on the test data set

Features of Different Missing Percentage





Methodology

1 EDA

- Missing values
- Imbalanced data

2 Data Processing

- Missing values

1. Remove the feature
2. Filled with
   Mean/Mode/Median
3. Linear Regression
4. Multiple Imputation

3 Feature Selection

- Filter Method
  - Fisher Criteria
- Wrapped Method
  - Forward Stepwise
- Embedded Method
  - LASSO

4 Model

- Logistic regression
- SVM
- Random Forest
- XGBoost
- Light GBM

5 Evaluation

- AUC
- Precision/Recall
- F1-score

6 Prediction

Test our model on the test data set

Methodology

1 EDA

- Missing values
- Imbalanced data

2 Data Processing

- Imbalanced data
  1. Up sampling
  2. Down sampling
  3. SMOTE

3 Feature Selection

- Filter Method
  - Fisher Criteria
- Wrapped Method
  - Forward Stepwise
- Embedded Method
  - LASSO

4 Model

- Logistic regression
- SVM
- Random Forest
- XGBoost
- Light GBM

5 Evaluation

- AUC
- Precision/Recall
- F1-score

6 Prediction

Test our model on the test data set

Preliminary Results

	Precision	Recall	F-1 Score	AUC
Dummy Classifier (Baseline)	0.0322	0.0325	0.0323	0.49
Logistic	0.0555	0.6594	0.1023	0.58
Random Forest	**0.4455**	0.4649	**0.4556**	0.88
SVM	0.2017	0.6148	0.3037	0.76
XGBoost	0.2531	**0.8646**	0.3916	**0.89**
LGBM	0.2644	0.6830	0.3813	0.85

Next Steps

- **Missing Value & Imbalance Data**
- **Feature Selection**
- **Tune Model Parameter：Grid Search**

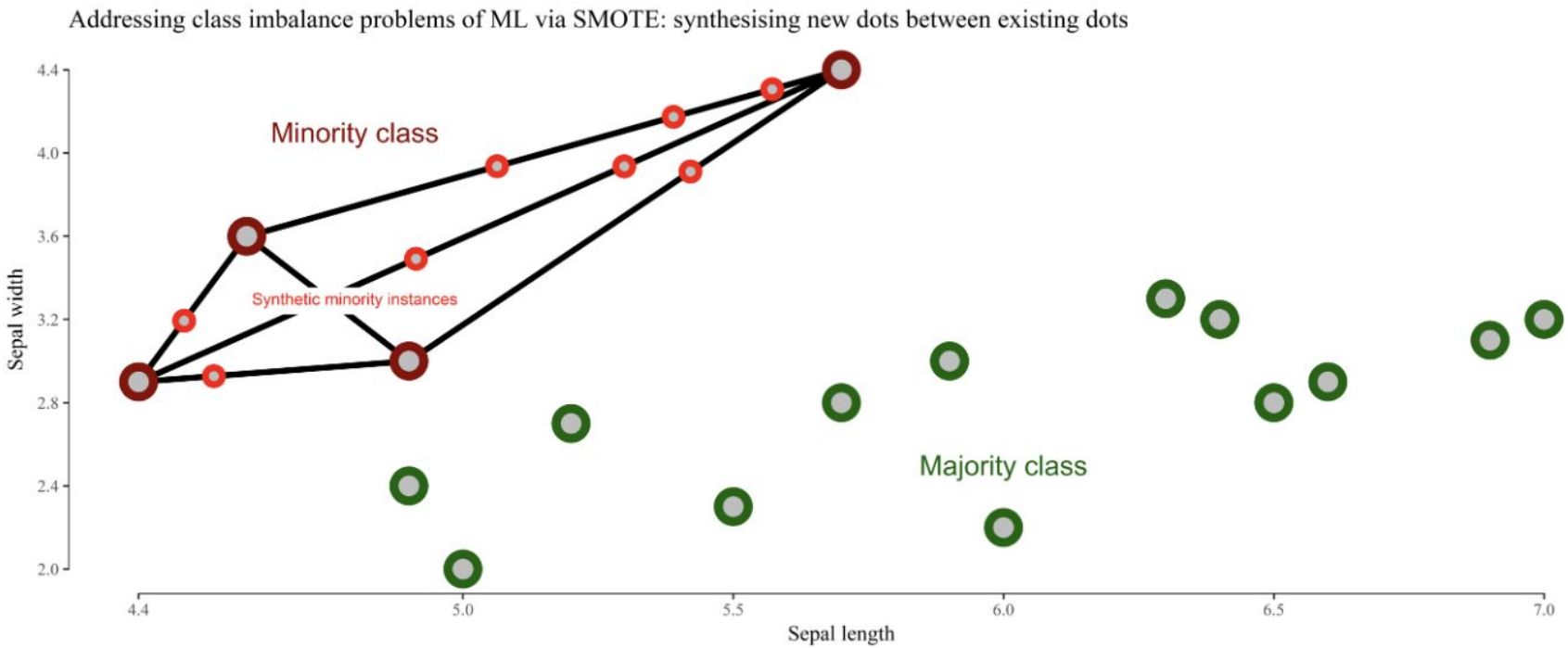
Q&A

Next Steps

- **Missing Value & Imbalance Data**
- **Feature Selection**
- **Tune Model Parameter：Grid Search**

Q&A

SMOTE



http://rikunert.com/SMOTE_explained