Detecting Credit Card Fraud

Jae-Ho Lee

Abstract

Statistical learning techniques were implemented in order to detect fraudulent credit card activity based on various customer and transaction information.

Introduction

Credit card fraud is an act of theft that everyone experiences at some point – either by personal experience or indirectly through stories of an acquaintance. This may be of a familiar scenario: you're checking your bank account to track your spendings when all of a sudden you realize payments of hundreds of dollars have been made with your credit card, and you proceed to frantically call the credit card company. Although credit card fraud counts for only about 0.1 % of all card transactions[1], due to the relatively large consequences associated with the theft, the experience of having to deal with it is undesirable.

Statistical learning techniques have been applied to construct a model that would help identify instances of fraud early on, ideally in the midst of it happening. While a large portion of fraudulent transaction can be identified with the resulting model for this dataset, further data collection and analysis needs to be conducted in order to train a more robust the model and assess its effectiveness.

Methods

Data

The data was accessed via Kaggle. It contains information regarding credit card users' transaction over the span of two days in September, 2013. Most of the predictor information have been transformed with PCA due to confidentiality. Information that remains interpretable includes `Time`, `Amount`, and `Class`.

- `Time`: time elapsed (in seconds) between this transaction and the first transaction
- `Amount`: transaction amount
- `Class`: binary variable with values `fraud` and `genuine`

Let it be noted that the values of the response, `Class`, are highly imbalanced – with `fraud` accounting for 492 out of 284807, or 0.17 %, of total transaction.

[1] Wikipedia: Credit Card Fraud

Modeling

In order to predict fraudulent credit card transaction, the following modeling techniques were considered:

- Gradient boosting machine
- Boosted logistic regression
- Neural network

Two validation methods were utilized. First, models were validated through 5-fold cross validation. Then, subsampling conducted according to ROSE with 5-fold cross validation to address issue of imbalance in response variable. The True Positive Rate, or Sensitivity, will be the metric of choice for validation

Evaluation

5-fold cross validation

```
cv = trainControl(
  classProbs = TRUE,
  method = "cv",
  number = 5,
  summaryFunction = twoClassSummary
)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
set.seed(42)
gbm_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'gbm',
  metric = "Sens",
  trControl = cv,
  verbose = FALSE
)
```

```
set.seed(42)
logit_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'LogitBoost',
```

```
  metric = "Sens",
  trControl = cv
)
```

```
set.seed(42)
nnet_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'nnet',
  metric = "Sens",
  trControl = cv,
  trace = FALSE
)
```

Subsample for Imbalance

```
cv_ss = trainControl(
  classProbs = TRUE,
  method = "cv",
  number = 5,
  sampling = 'rose',
  summaryFunction = twoClassSummary
)
```

```
set.seed(42)
gbm_ss_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'gbm',
  metric = "Sens",
  trControl = cv_ss,
  verbose = FALSE
)
```

```
set.seed(42)
logit_ss_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'LogitBoost',
  metric = "Sens",
  trControl = cv_ss
)
```

```
set.seed(42)
nnet_ss_mod = train(
  Class ~ . - Time,
  data = cc_trn,
  method = 'nnet',
  metric = "Sens",
  trControl = cv_ss,
```

Table 1: **Boosted Logistic Regression**, 5-Fold CV with Subsampling for Imbalance

Metric	Value
True Positive Rate (%)	86.27
True Negative Rate (%)	1.28
Average Loss (€)	0.04
Maximum Loss (€)	1062.93
Total Loss (€)	5204.23

```
  trace = FALSE
)
```

Results

When only using 5-fold cross validation, the bootsted logistic regression heavily out-performed the other two models. Accounting for imbalance in the response, all three models turned out to perform fairly similarly. Nonetheless, the boosted logistic regression model best performed when validating for best True Positive prediction using 5-fold cv according to ROSE.

\begin{table}

\caption{5-Fold Cross Validation, %}

Model	True Positive Rate	True Negative Rate
Gradient Boosting Machine	49.72	0.04
Boosted Logistic Regression	73.40	0.01
Neural Network	64.84	0.02

\end{table}

\begin{table}

\caption{5-Fold CV with Subsampling for Imbalance, %}

Model	True Positive Rate	True Negative Rate
Gradient Boosting Machine	86.52	0.55
Boosted Logistic Regression	87.36	1.08
Neural Network	87.80	0.77

\end{table}

Discussion

Let us assume that the monetary consequence, or "loss", for a credit card company can be defined (refer to the Loss Calculation Table below).

Using to the selected boosted logistic regression model, 85 % of frauds can be detected. However, there are several limitations that make this model unjustifiable when making predictions on new datasets. First, the

Table 2: Loss Calculation Table

Actual	Predicted	Loss
Fraud	Genuine	0.5 x (Actual Amount)
Fraud	Fraud	0
Genuine	Genuine	0
Genuine	Fraud	1

data was collected from a very specific time frame – across two days in September, 2013. This makes the data liable to biased based on seasonal factors. Second, the location of the transactions made were all in Europe. Consequently, this use of this model to predict credit card fraud outside of this timeframe and in countries outside of Europe cannot be justified.

Fitting a model with data from more various timeframes and transactions from countries outside of Europe will allow the model to be more robust. Another direction for a more meaningful analysis would be to build a model that better predicts fraud that leads to larger "losses". Such model predictions would be more useful for a credit card company, as well as customers (since larger "losses" are associated with larger sums of fraudulent transaction).