Predicting Fraud

Tim Hopp

2022-11-10

Predicting Fraud in Credit Card Transactions

Using a dataset consisting of ~**280,000** european credit card transactions made in two days to explore credit card fraud and to use a decision tree model to predict fraud in a testing dataset.

```
## [1] 284807      31
```

The Data is highly imbalanced as we can see **284,315** credit card transactions that are non-fraudelent and only **492** transaction that were fraud.

```
##
##      0      1
## 284315    492
```

This will require sampling methods that will balance the data out. Two baseline sample methods will be used: Up sampling and Down sampling.

Up sampling will increase the amount of fraudelent cases until they are balanced with the non-fraud cases. Down sampling will do the opposite. Up sampling will increase the data size but may lead to overfitting whereas down sampling will reduce the data size which could lead to other numerous problems. We will compare the Accuracy Under the Curve for both sampling methods to determine if one is prefferred to the other.
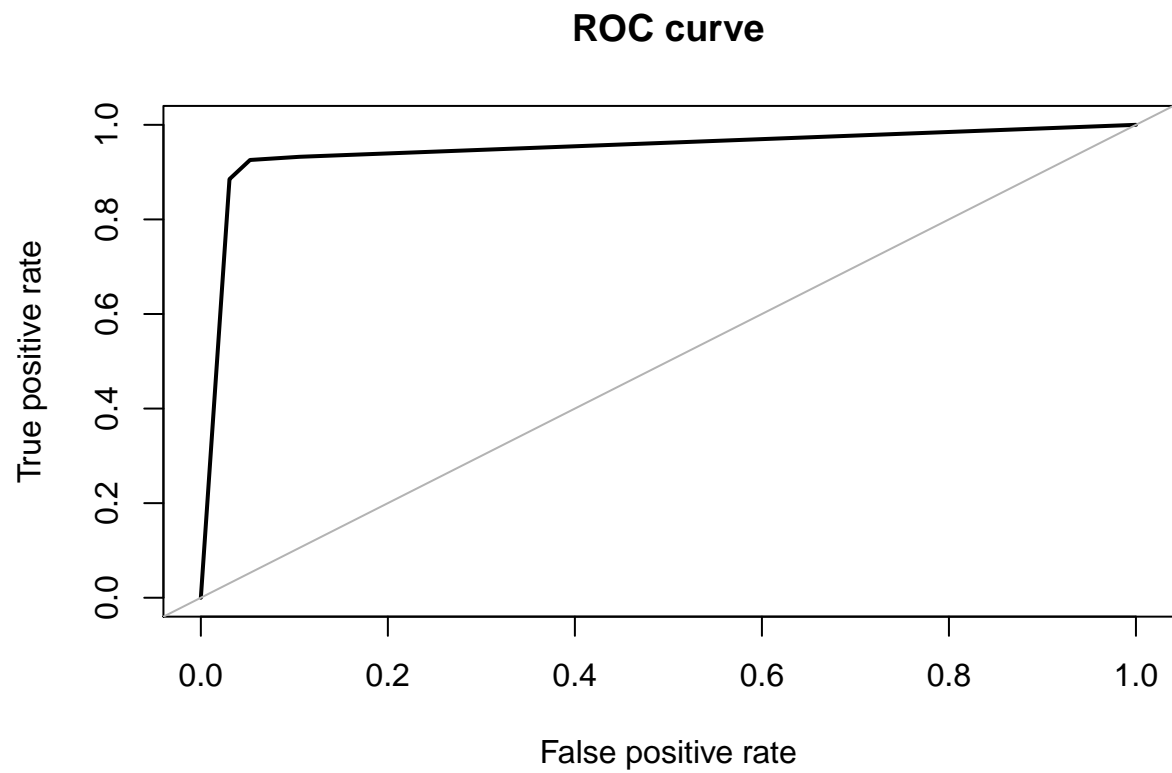
```
## [1] "Using the down sampling method we see a much smaller training data set"
```

```
##
## Not_Fraud      Fraud
##       344        344
```
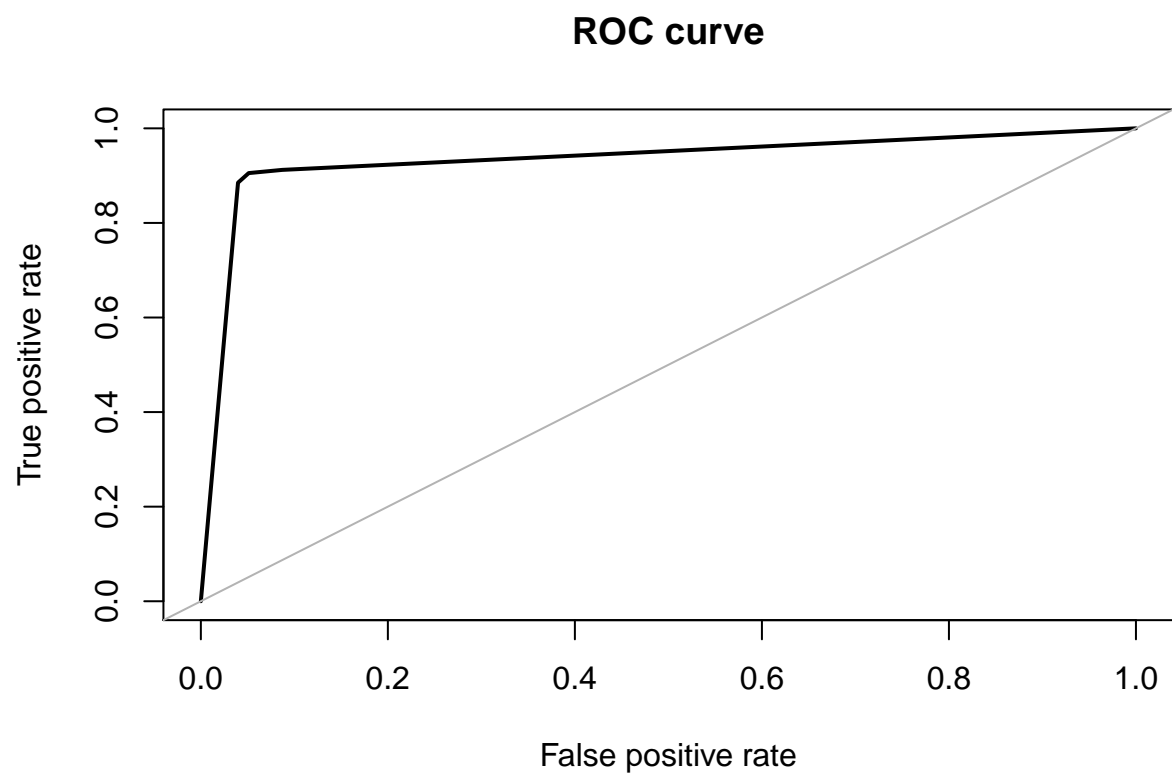
```
## [1] "Using up sampling method we see a much larger training data set"
```

```
##
## Not_Fraud      Fraud
##    199020     199020
```

Lastly we will fit both the up and down sample training sets to a Decision Tree model

ROC curve


```
## Area under the curve (AUC): 0.947
```

ROC curve

Area under the curve (AUC): 0.933

Figure 1: CreditCardFraud