Credit Card Fraud Detection Project

Ramin Ahmed

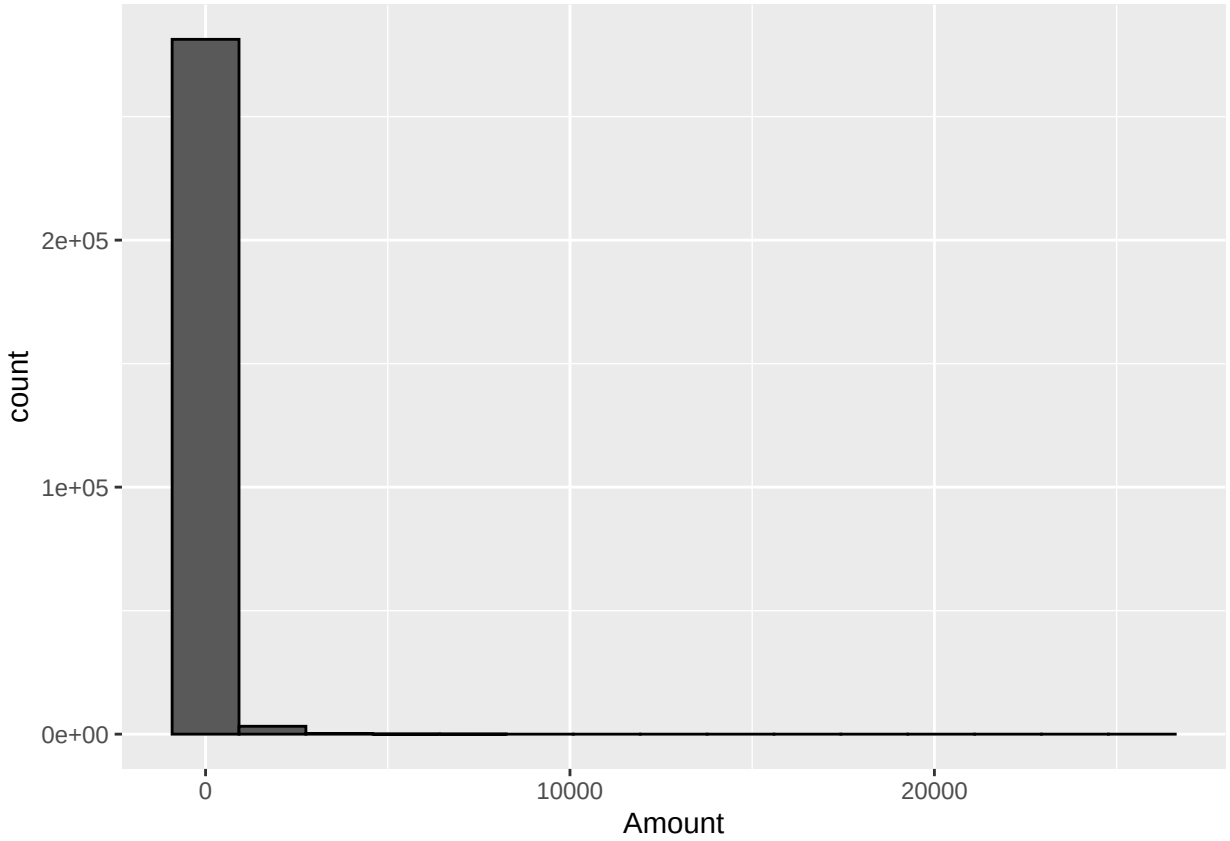
12/26/2019

Intoduction

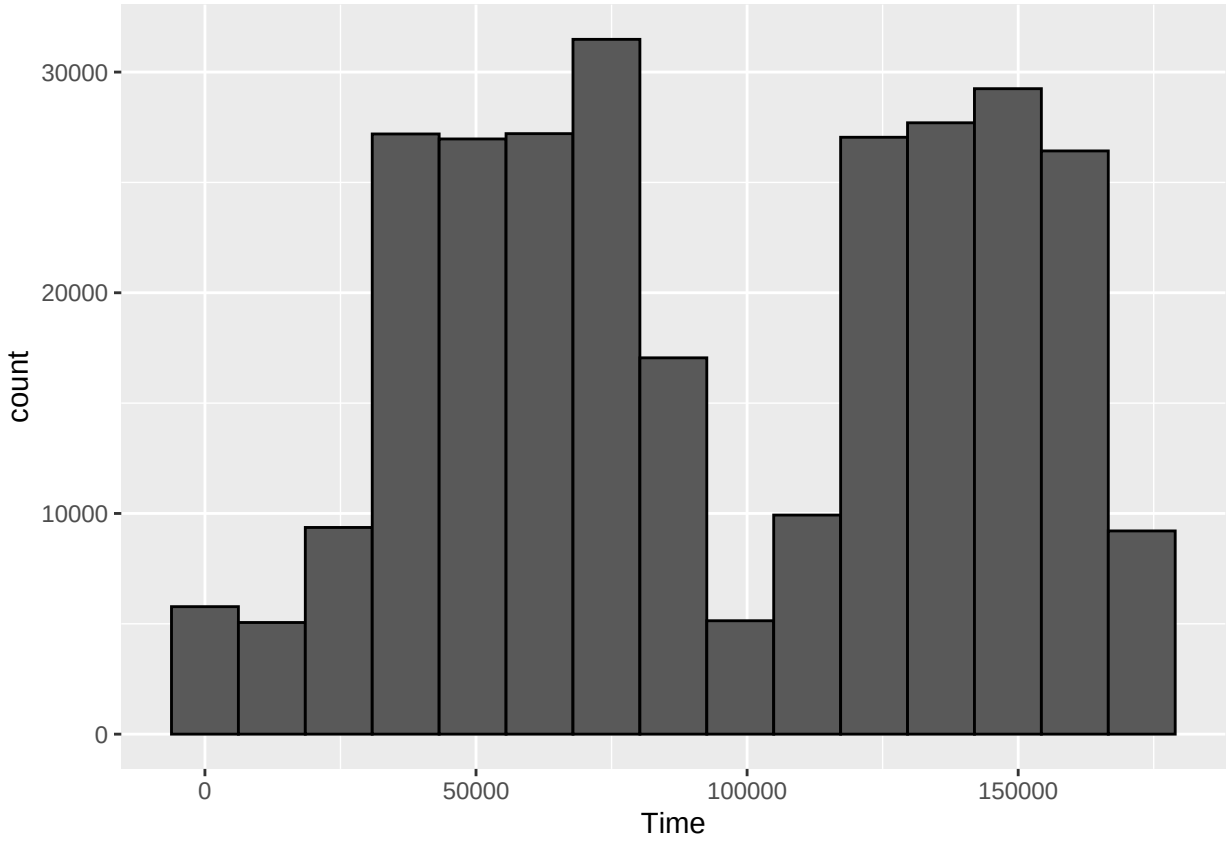
In this project, we create a credit card fraud detection system using credit card fraud detection data obtained from kaggle (data is also submitted in csv format). After doing some data wrangling, we see that the data contains 284807 observations with 31 features. All the features are anonymized and standardized except for Time and Amount feature. The Class column provides the status of the response where 1 corresponds to fraud and 0 as non-fraud. Here, I have to predict whether an entry, given the feauture values, are fraud or not. For this, we need to create training sets (as train) and test sets (as test) where the test set contains 30% data. The idea is to train our machine learning model on the train set and test the efficiency of our model on the test set.
Since the prevalence of the positive response (Class=1 is fraud) is very low, choosing machine learning model based on accuracy would not do any good. We, instead, use F1 score to test the efficiency of each model. We use various machine learning algorithms like logistic regression, linear discriminant analysis (LDA), quandratic discriminant analysis (QDA), naive bayes and anomaly detection. Given the nature of the problem, intuitively anomaly detection would be the prefect choice, but the analysis shows that this algorithm does not provide requisite F1 score and this is due to the underlying gaussian assumption. We found LDA to provide the highest F1 score and decided to go with it as the final recommendation.

Exploratory Data Analysis

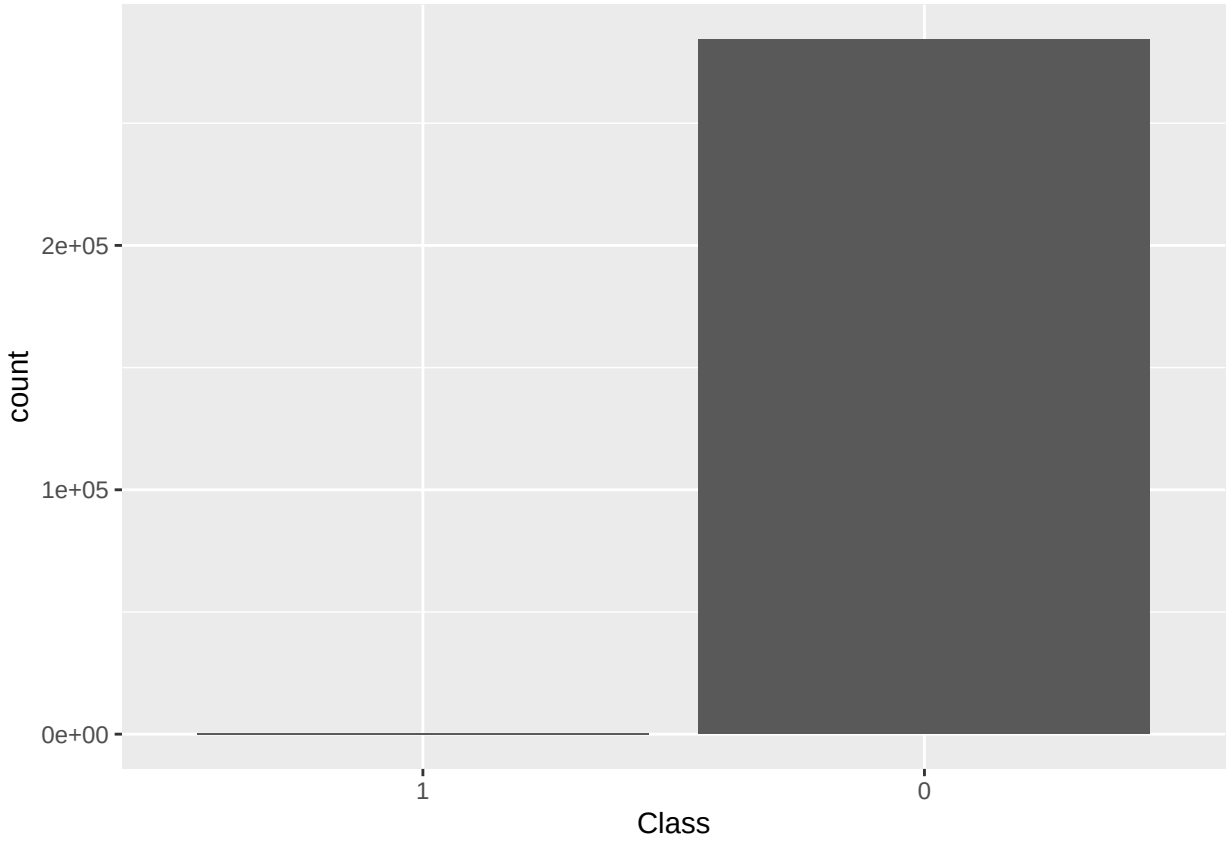
At first, I checked the structure of the data and performed requisite variable type conversion (converting Class variable from numeric to factor). Then when I checked the mean, standard deviation, minimum and maximum values for each feature, I found that except Time and Amount feauture, all the others are already standardized. I start investigating how these two variables are distributed.



Here, we see the distribution of the amount feature is right skewed with mean of \$88.3496193 and a maximum value of $\$2.569116 \times 10^4$ whereas the distribution of time feature is bimodal.



I also looked into the class distribution and determined the prevalence to be 0.1727486%. The following barchart helps to depict the prevalance issue.



Since the prevalence of the positive response (Class=1 is fraud) is very low, choosing machine learning model based on accuracy would not do any good. Because even predicting every observation as non-fraud activity will result in an accuracy of around 99%. We, instead, use F1 score to test the efficiency of each model. Since all the other features were already scaled, I scaled Time and Amount as well so that the algorithms do not unnecessarily put too much weight on them. I then create train sets and test sets where test set contains 30% of the data.

Algorithms

Logistic Regression

We start with logistic regression on the train set, predict for test data, derive the likelihood of the observations being fraud or non-fraud and convert into outputs by 0.5 cutt-off score. We observe an accuracy of 0.9992159 and an F1 score of 0.7351779.

Linear Discriminant Analysis (LDA)

We then explore with LDA on the train set, predict for test data, derive the likelihood of the observations being fraud or non-fraud and convert into outputs by 0.5 cutt-off score. We observe an accuracy of 0.9993212 and an F1 score of 0.7883212.

Quadratic Discriminant Analysis (QDA)

We then use QDA on the train set, predict for test data, derive the likelihood of the observations being fraud or non-fraud and convert into outputs by 0.5 cutt-off score. We observe an F1 score of 0.1060543.

Naive Bayes

We then analyze the problem with Naive Bayes on the train set, predict for test data, derive the likelihood of the observations being fraud or non-fraud and convert into outputs by 0.5 cutt-off score. We observe an F1 score of 0.4267631.

I wanted to use KNN and random forest, but for a data of such magnitude it becomes computationally very expensive. Even using PCA to reduce dimensions and then using these algorithms does not help. This is probably because besides making the data anonimized, I think the data already underwent some sort of dimensional reduction.

Anomaly Detection

Though not covered in this course, I really feel this algorithm becomes handy while analyzing such machine learning problems. For this algorithm, I, at first, created train set, validation set and test set. Here the test set contains 60% of the data where all of the observations where non-fraud. The validation set and test set contain 20% of the data each and each of them contain 50% of the fraud activities. In the training set, each feature is considered to be normally distributed with respective means and sandard deviations. Using this we calculate the the probability of each observation in validation set using the following equation:

$$p(x) = \prod_{j=1}^{n} p(x_j, \mu_j, \sigma_j^2)$$

where, x_j = value of feature j for an observation. μ_j = mean of feature j. σ_j^2 = variance of feature j. n = number of features. Once we get the probabilities, we use each the probability of each fraud entry as cut-off and select the cut-off which maximized the F1 score. Using this cutt-off, we perform our prediction on the test set. We observe an F1 score of 0.3625954 which is very poor. I think this algorithm is providing a poor result due to the gaussian assumption on these dimensionally reduced features.

Results

The follwoing table summarizes my analysis and its findings. It is evident the LDA provides the best F1 score.

```
##                method  F1_score
## 1 Logistic Regression 0.7351779
## 2                 LDA 0.7883212
## 3                 QDA 0.1060543
## 4         Naive Bayes 0.4267631
## 5   Anomaly Detection 0.3625954
```

Conclusion

In this project, I created a credit card fraud detection system. Through exploratory data analysis, I gained insights and explored the nature of the data and performed requitie transformation. I have analyzed the

system and tried to come up with a fraud detection system after exploring 5 different machine learning algorithms. Though anomaly detection algorithm intuitive seems promising in this scenario, LDA provided with the best result. I think making use of decision trees, random forest, KNN, nueral networks would have given probably a better detection system. But due to the computaional expensiveness of these algorithms and nature of data, those were not possible. So, as a future work, I recommend performing further analysis using these algorithms.