UNITED STATES MILITARY ACADEMY

PROJECT 4

MA477

SECTION A1

COL GROVER LAPORTE

By

CDT JOSEPH ZUCCARELLI, CO F1 '21

WEST POINT, NEW YORK

12 MAY 2021

_____ MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

_____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURES: _____

Machine Learning Models for Fraud Detection

CDT Joseph Zuccarelli

12 May 2021

Abstract

The purpose of this report is to use credit card transaction data in order to detect instances of fraud. Fraud detection is a critical issue for financial companies; therefore, many are developing machine learning algorithms so that customers are not charged for items that they did not purchase. The data set analyzed in this report contains credit card transactions during September 2013 made by European cardholders. Several popular machine learning techniques are employed to analyze this data–logistic regression, random forests, support vector machines, and artificial neural networks. Using these techniques, we created four models that classify a credit card transaction as fraudulent or genuine based on several variables associated with the transaction. We evaluate the performance of each model using Precision-Recall curves. Ultimately, the project outlined in this report is a rudimentary study for an introductory course in data science.

Introduction

Fraudulent financial behavior is a major problem with severe consequences for those in the finance industry, corporate organizations, and government. Fraud is defined as a deceptive act with the intent to acquire financial gain, and it is generally considered a state crime, yet in some cases it may fall under federal jurisdiction. Given the growing popularity of credit card transactions within the past decade, the rate of fraudulent credit card transactions is rising as well. Detecting fraudulent transactions manually is extremely time consuming and inefficient; therefore, financial organizations are investing resources to develop machine learning models that will handle this issue (Awoyemi et al., 2017). Therefore, in the following report we analyze credit card transaction data and use several popular machine learning methods to build models that predict whether or not a transaction is fraudulent or genuine with a fair degree of accuracy.

Data

The observational units under analysis in this study are credit card transactions made by European cardholders over the span of two days during September 2013. In total, the data set contains 30 predictors; however, due to confidentiality issues, many of these are unlabeled quantitative variables that are the result of a principal component analysis (PCA) transformation. We can assume that these variables represent the card's spending behavior. The only two labeled features that were not transformed with PCA are *Time* and *Amount. Time* represents the seconds elapsed between each transaction and the first transaction in the dataset. *Amount* represents the cost associated with the transaction. The response variable included in this data set is *Class*, which is a binary qualitative variable that indicates whether or not the transaction was fraudulent or genuine. In total, the data set includes 284,807 observations, in which 492 of those observations were fraudulent and 284,315 were genuine. Therefore, the data set is highly unbalanced, as the positive class (fraudulent transactions) only accounts for approximately 0.173% of the total observations.

Methods

We analyze the data set described in the previous section using four common machine learning methods: logistic regression, random forests, support vector machines (SVMs), and artificial neural networks (ANNs). Prior to using these methods to build models, we perform some exploratory analysis of the data set. Next, we divide it into a training (70%) and testing (30%) set. We then train each model using the training data set and evaluate its performance using the testing data set. Although all the machine learning methods highlighted above are useful for classification, each offers its own set of pros and cons.

Logistic regression is a robust and flexible method that is typically used for dichotomous classification prediction. This means that its goal is to find a decision boundary that separates one class from the other. When performing linear regression, the assumption is that the decision boundary is linear, meaning that it is a hyperplane in the high-dimensional feature space. The parameters included in a logistic regression model are roughly the weights for the model features. The learning algorithm tunes these weights in order to correctly classify the training observations (Gudivada et al., 2016).

Random forest is an ensemble classification technique that combines many decision tree predictors, in which each tree depends on the values of a random vector sampled independently. Note that all decision trees in the forest have the same distribution. The process of constructing each tree within the forest is as follows. First, we must assume that n is the number of training observations and p is the number of features in the data set. In order to determine the decision node at a tree, we choose $k << p$ as the number of variables to be selected. We then select a bootstrap sample from the training data set and use the remaining observations to estimate the model error. Therefore, we randomly choose k variables as a decision at a certain node in the tree and calculate the best split based on these k variables in the training data set (Dunham, 2009).

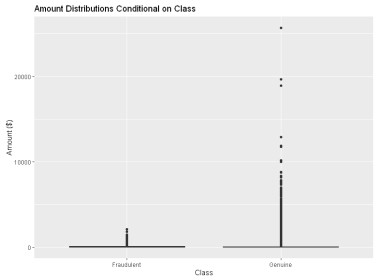
SVM is another common machine learning method used to classify binary variables, in which the observations of the two classes are linearly separable. Assuming that the classes are linearly separable, there exist hyperplanes that separate observations of the two classes–an infinite number of hyperplanes to be exact. The goal of the algorithm is to select the specific hyperplane that sits in the middle between the observations of the two classes. In mathematical terms, this means that the algorithm searches for the hyperplane that maximizes the minimum distance between that hyperplane and all the observations of the two classes. In other words, this optimal hyperplane lies equidistant from the observations of the two classes that are closest to it. This enables the model to classify new observations with a high-degree of accuracy (Gudivada et al., 2016).

ANNs are complex models that attempt to mimic the manner in which the human brain develops rules for classification. These models consist of three layers: an input layer, and output layer, and hidden layer(s). The input layer receives raw observations from the training data to perform pattern recognition. The hidden layer receives the raw observations from the input layer and processes them. The output layer classifies the observations processed by the input layer. During this process a training algorithm is used to learn the data set and modify the model based on the error rate between the observed and actual output. This results in considerably accurate classifications as outputs (Sairamya et al., 2019).

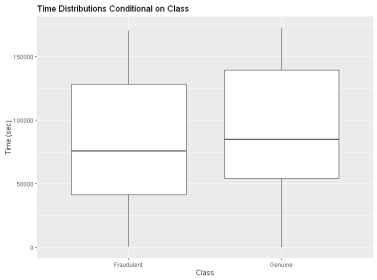
In order to evaluate the performance of each of these methods in building our models, we use a metric known as the area under the precision-recall curve (AUPRC). This metric is useful when dealing with an imbalanced data set such as ours in which one is most concerned with identifying the positive values. Precision-Recall (PR) curves highlight the tradeoff between precision and recall for different thresholds. Precision P is defined as the number of true positives T_p over the number of true positives T_p plus the number of false positives F_p. Recall R is defined as the number of true positives T_p over the number of true positives T_p plus the number of false negatives F_n. A high area under the PR curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). Average precision summarizes a PR curve as the weighted mean of precisions achieved at each threshold (Pedregosa et al., 2011).

Results

Figure 1 included below displays the relationship between the two labeled predictors in the data set, *Amount* and *Time*, and the response variable, *Class*. **Figure 1a** displays the amount distribution conditional on the class of the transaction (genuine or fraudulent). **Figure 1b** displays the time distribution conditional on the class of the transaction (genuine or fraudulent)



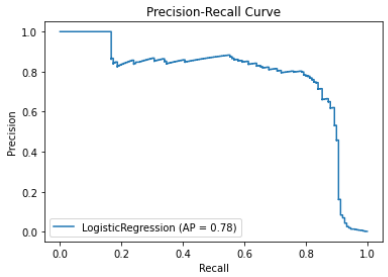
(a) Amount Distributions



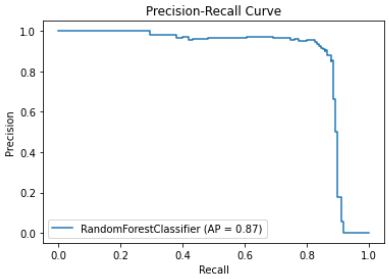
(b) Time Distributions

Figure 1: Predictor Distributions Conditional on Class

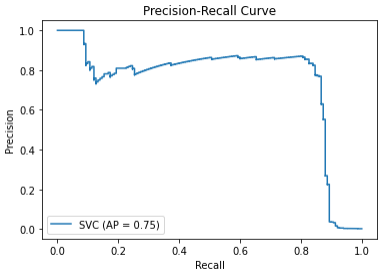
Figure 2 included below displays the PR curve for each model. The average precision score for the logistic regression model (**Figure 2a**) is 0.78. The average precision score for the random forest model (**Figure 2b**) is 0.87. The average precision score for the SVM model (**Figure 2c**) is 0.75. The average precision score for the ANN model (**Figure 2d**) is 0.78.



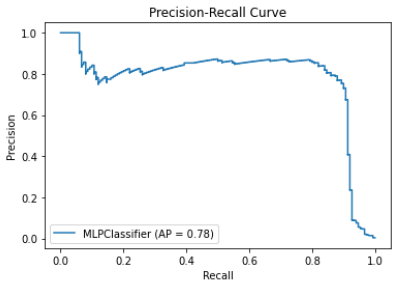
(a) Logistic Regression Model



(b) Random Forest Model



(c) SVM Model



(d) ANN Model

Figure 2: Precision-Recall Curves

Table 1 included on the following page displays the top five features in terms of importance from the random forest model. Remember that many of the features in the data set are unlabeled quantitative

variables that are the result of a principal component analysis (PCA) transformation.

Table 1: Feature Importances Random Forest Model

Feature	v14	v12	v17	v11	v10
Importance	0.14606	0.13657	0.11943	0.07362	0.07057

Discussion

Figure 1 tells us a bit about the relationship between the two labeled predictors in the data set, *Amount* and *Time*, and the response variable, *Class*. It appears that the genuine transactions typically involved larger amounts of cash then the fraudulent transactions. It also appears that on average the genuine transactions occurred a bit later during the period of observation than the fraudulent transactions. The time variable would perhaps be a bit more useful if it was the time of day as opposed to the time elapsed between the transaction and the first transaction in the data set. **Figure 2** indicates the performance of each of the models that we built. In terms of average precision score, the random forest model was the highest-performing model, followed by the ANN model, the logistic regression model, and the SVM model. **Table 1** highlights the most important features within the random forest model. This table would obviously be more informative if these features were labeled; however, this is not the case due to confidentiality issues. Overall, the analysis carried out in this report demonstrates that it is possible to classify credit card transactions as fraudulent or genuine with a high degree of precision and recall using several different machine learning methods. Although the models outlined in this report are certainly not worthy of use by high-end financial companies, they do highlight how machine learning can be used to increase the efficiency and effectiveness of what once were manual tasks like fraud detection.

References

[Awoyemi et al., 2017] Awoyemi, J. O., Adetunmbi, A. O., and Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9.

[Dunham, 2009] Dunham, K. (2009). Chapter 6 - phishing, smishing, and vishing. In *Mobile Malware Attacks and Defense*, pages 125–196. Syngress, Boston.

[Gudivada et al., 2016] Gudivada, V., Irfan, M., Fathi, E., and Rao, D. (2016). Chapter 5 - cognitive analytics: Going beyond big data analytics and machine learning. In Gudivada, V. N., Raghavan, V. V., Govindaraju, V., and Rao, C., editors, *Cognitive Computing: Theory and Applications*, volume 35 of *Handbook of Statistics*, pages 169–205. Elsevier.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Sairamya et al., 2019] Sairamya, N., Susmitha, L., Thomas George, S., and Subathra, M. (2019). Chapter 12 - hybrid approach for classification of electroencephalographic signals using time–frequency images with wavelets and texture features. In Hemanth, D. J., Gupta, D., and Emilia Balas, V., editors, *Intelligent Data Analysis for Biomedical Applications*, Intelligent Data-Centric Systems, pages 253–273. Academic Press.