ECE/CS/ME 539 Introduction to Artificial Neural Networks

Project Progress Report

Detection of Credit Card Fraud with Artificial Neural Networks and Multiple Classifier Systems

Team #: 20

Team members:

Matthew Laluzerne: Mechanical Engineering Graduate

Rui Feng: Electrical and Computer Engineering Undergraduate

Ran Zhao: Statistics Graduate

Date of submission: 4/8/23

Abstract

For this progress report, we have finished writing the Data preprocessing code and building the classification model. Currently, we are trying to make improvements to our classification model. The classifiers used were KNN, DNN, NBC, a Decision Tree, and Logistic Regression. The classifier with the best performance was the Decision Tree

Introduction

Ever since the appearance of credit cards in 1950, the number of people preferring credit cards for payments has been rising every year. A report from Federal Reserve Bank of San Francisco shows that 51% of retail transactions are done using credit or debit cards, a lot higher than cash which has only 26% [1]. The rapid growth in the number of credit card transactions has led to a severe problem: the rise of fraudulent activities [2]. Many people have been in a credit card fraud case. 151 million American card holders have been the fraud victims. That's 65% of card holders and the number is still rising compared to 58% last year [1]. Credit Card Fraud greatly harms our economy. The Nilson Report, which monitors the payment industry, released a forecast in December 2022, indicating that $165.1 billion will be lost due to card fraud over the next 10 years [3].

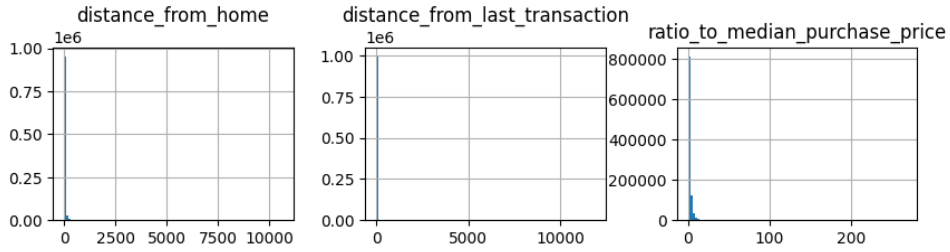
The poor protection against credit card fraud has driven us to develop a solution better than keeping a good habit when using credit cards. Machine learning has proved to be effective at anomaly detection, which means it is powerful for detecting unusual patterns such as fraud detection. As credit card fraud has become a main part of fraudulent activities, it has drawn great attention from machine-learning areas. The following techniques have been applied for fraud detection: Decision Tree, Neural networks, Logistic Regression, Genetic algorithms, clustering techniques, and outlier detection [4]. Currently, the study shows that, among all the techniques in machine learning, KNN, Logistic Regression Classifier, Random Forest, and Bayes have the highest accuracy rates in the field [5]. Furthermore, more research has focused on how to improve the performance of these techniques. To deal with imbalance data problem, a variety of data augmentation techniques have been compared and a new data augmentation model, K-CGAN, has been proposed [6]. We would like to find further optimization probability based on the materials we learn from this course.

Methods

The dataset we plan to use for the project is about Credit Card Fraud from Kaggle. It contains 1,000,000 samples and 7 features in total with a size of 76.28 MB. This data was collected from consumer transactions with some features withheld and others processed into new features to avoid giving away sensitive consumer data. The data is labeled with the fraud column as either fraudulent or non-fraudulent. This creates a binary classification problem to label transactions as fraudulent or not fraudulent. The data set contains the following columns:

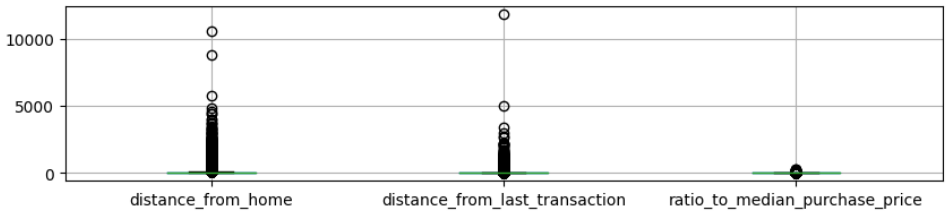
1. distance_from_home: the distance from home where the transaction happened.
2. distance_from_last_transaction: the distance from last transaction happened.

3. ratio_to_median_purchase_price: Ratio of purchased price transaction to median purchase price.
4. repeat_retailer: Is the transaction happened from same retailer.
5. used_chip: Is the transaction through chip (credit card).
6. used_pin_number: Is the transaction happened by using PIN number.
7. online_order: Is the transaction an online order.
8. fraud: Is the transaction fraudulent.



Among these 8 variables, the first three are numerical while the others are binary. For the five binary variables, the proportion of zeros are 0.118464, 0.649601, 0.899392, 0.349448, 0.912597, respectively. For the three numerical variables, the following are the histograms and boxplots of them.

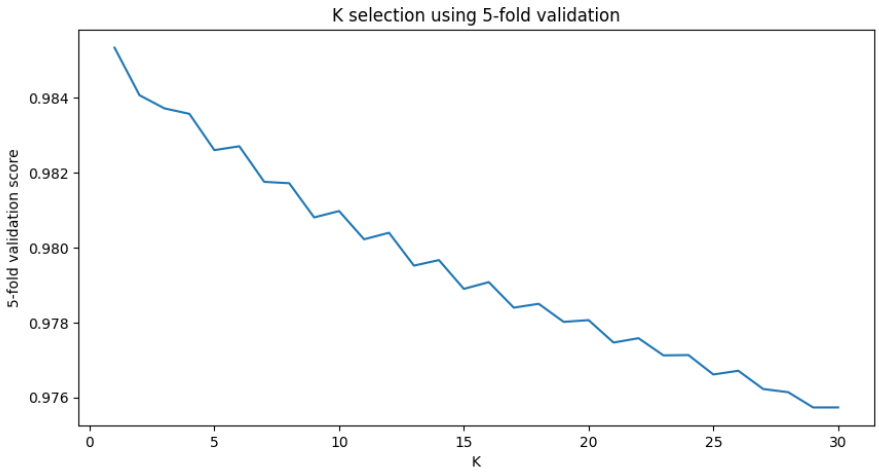
Upon closer inspection of the data set, 91.26% of transactions are genuine and only 8.74% are fraudulent. This means that accuracy is not a reasonable choice to evaluate the performance of the model because guessing that all transactions are legitimate would give a very high accuracy score. The metric that will be used to evaluate the model performance is recall. Maximizing recall will minimize the number of false negatives. This is desirable because a false negative means failing to detect fraud, which could have disastrous consequences for consumers and credit card companies. A false positive would not be as damaging because the credit card company can call the consumer to verify the transaction.



For preprocessing, the data was split into testing and training data sets at an 80 percent training data ratio with the sklearn module [7]. Finally, each feature was scaled to a standard normal distribution with the sklearn standard scaler [7].

The first model that was trained was a decision tree using the sklearn decision tree classifier [7]. The sklearn package was also used to develop a logistic regression classifier, and a naïve bayesian classifier. Standard parameters were used for these models, besides changing increasing the number of iterations on the logistic regression so that it would be able to find a

solution. The sklearn was also used to train a KNN classifier with k=1 selected by 5-fold cross validation. Finally, a deep neural network was created with tensorflow [8] and keras [9].



Results

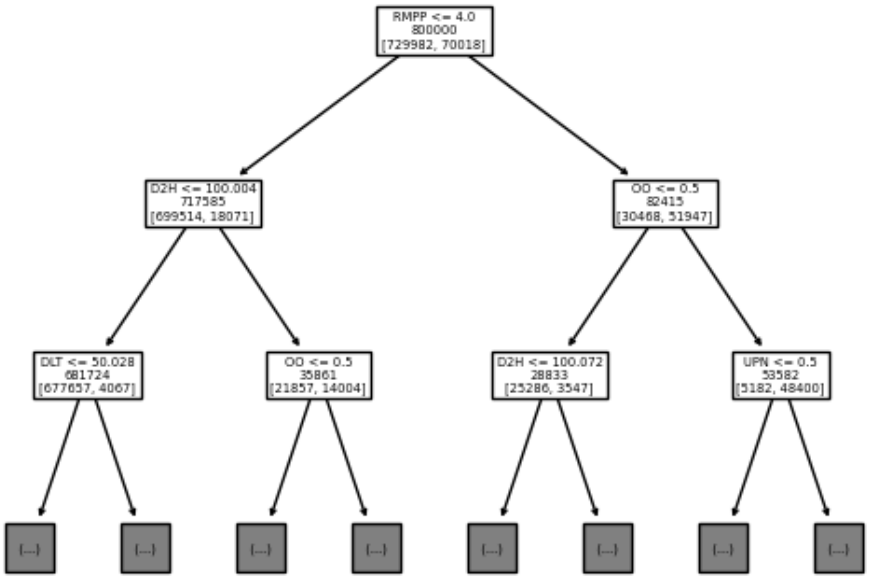
The recall results of each of the classifiers are shown below in the following table. The recall results are from running the model on the testing data.

Model	Accuracy	Recall
Decision Tree	**0.99998**	**0.99999**
K Nearest Neighbors	**0.98613**	**0.93762**
Deep Neural Network	**0.98044**	**0.97170**
Naïve Bayesian Classifier	**0.95074**	**0.98469**
Logistic Regression	**0.959145**	**0.99304**

As expected, the decision tree was the best performing classifier by far. The number of misclassifications was only 4. The confusion matrix for the decision tree is shown below.

True/Predicted	-	+
-	182613	2
+	2	17383

The decision tree had the following structure, which has been abbreviated to only show the first 2 out of 7 layers of the tree.



Discussion

The reason decision tree is the best classifier for this dataset is that it handles imbalanced data very well. Other classifiers, like KNN, treat all data points equally. Since the dataset we are using is highly imbalanced which has only 8.74% fraudulent data. Even if we try to down sample the fraudulent data points to handle imbalanced data, this could lead to underfitting and worse performance of our model because a substantial portion of the non-fraudulent data is discarded. Moreover, the dataset contains categorical and continuous data. Since decision tree classifier performs well on both type of data and Naïve Bayesian Classifier perform especially well on categorical dataset since it assumes data are independent. Overall, the decision tree classifier has the property suitable for the dataset we are using.

Currently, more work is needed to optimize the other models to get performance more comparable to the decision tree. On the DNN classifier, changing the structure and hyperparameters has the potential to increase the performance of the classifier. Another optimization that could be made is better preprocessing of the data including down sampling the data so that there is an equal portion of non fraudulent and fraudulent transactions in the training data. These issues will be addressed in the final version of the report.

References

[1] “2023 Credit Card Fraud Report” Accessed: Mar. 3,2023.

[2] S. Benson Edwin Raj and A. Annie Portia, "Analysis on credit card fraud detection methods," 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), Tirunelveli, India, 2011, pp. 152-156, doi: 10.1109/ICCCET.2011.5762457.

[3] ‘Credit card fraud statistics” by John Egan. Accessed: Mar. 2, 2023.

[4] Delamaire, L., Abdou, H., & Pointon, J. (2009). Credit card fraud and detection techniques: a review. Banks and Bank systems, 4(2), 57-68.

[5] CARD, R. T. T. B. C. (2023) A Proposed Framework of Dimensionality Reduction Techniques to Boost Credit Card Fraud Classification.

[6] Strelcenia E, Prakoonwit S. Improving Classification Performance in Credit Card Fraud Detection by Using New Data Augmentation. AI. 2023; 4(1):172-198.

[7] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Apr. 08, 2023. [Online].

[8] M. Abadi et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems”

[9] F. Chollet, “Keras,” 2015.