Final Milestone

Adil Khan

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

Oct, 22, 2022

Topic

Detect Credit Card Fraud

Abstract

A machine learning model that can accurately detect whether a credit card transaction seems to be fraudulent. I'll start with exploratory data analysis and utilise machine learning techniques to comprehend and process this data. I will develop, train, and test a number of models to identify which transactions are fraudulent by examining the data set and recognising trends. My final goal is to test and choose the model that is the most accurate for this collection of data. The Machine Learning Group [http://mlg.ulb.ac.be/] and Worldline are the sources of the information used. There are both fraudulent and legitimate transactions included in this dataset of card transactions. There are 31 features and about 285,000 rows of data. For security reasons, 28 of these features are unavailable to the general public; nevertheless, the three I'm interested in are Time, Amount, and Class (fraudulent or non-fraudulent). We have a project opportunity for a classification using this data format. A process called classification uses data that has been divided up into a number of classifications. Such a project aims to determine what class or category a new data item will belong to. In light of the rising automation and electronicization of our environment, it is crucial to examine this data. We must continue to be vigilant in our search for strategies to weed out fraudulent information since credit card theft is on the rise. The created model(s) may prove to be of great use in today's increasingly computerised environment, where millions of transactions—many of them fraudulent—take place every day.

Initial Exploratory Analysis

The data set I utilised is a compilation of both legitimate and erroneous credit card transaction data. In the collection, there are 284,807 transactions. We must make do with what we have because a lot of our columns are kept anonymous for security reasons. The only two columns we are aware of are amount and transactions. The remaining columns, which are unknown, have already been scaled, though. Initial research reveals that the data set's mean is 88.35. We don't discover any Null values, which is good since it eliminates the need for us to figure out how to account for such values. It's interesting to note that the majority of transactions are legitimate, with fraudulent ones occuring just 0.17% of the time. Our largest transaction has a value of $25,691.16, however when combined with our mean of 88.35, it is simple to surmise that there is a substantial skew to the right, with most transactions having a value far lower than our maximum transaction (Figure 1).

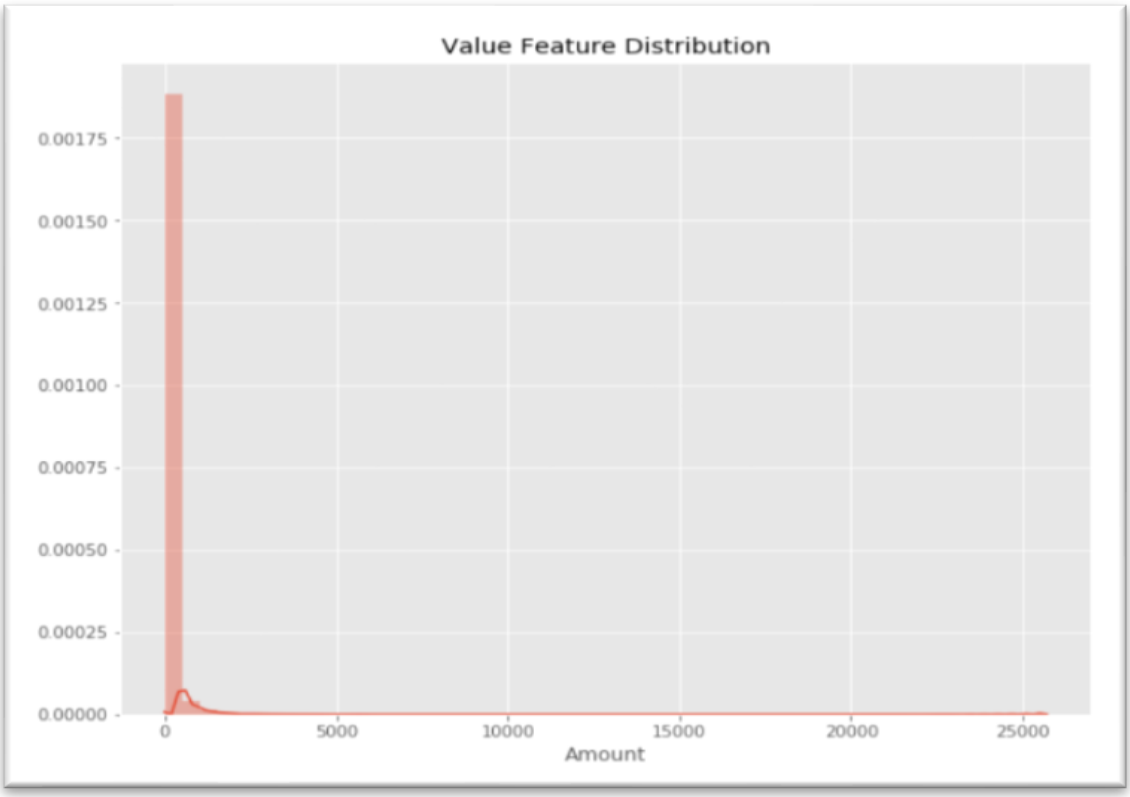

Figure 1: Heavy right skew

I initially started using some exploratory data analysis approaches to get acquainted with this data collection. Exploratory data analysis is a crucial initial step in exploring your data to better comprehend it and spot any patterns that may be forming to guide further data analysis. This method is useful since it enables us to clean up our data before devoting too much effort to analysis.

The duration of our time variable is measured in seconds since the very first transaction in the data collection. We can infer that our data set was gathered over a period of about 48 hours or two days because we have data for roughly 175,000 seconds. We can quickly see something interesting by picturing this time period. We observe that the number of transactions significantly decreases just before the halfway point of the data collection, then quickly increases shortly after. We can reasonably assume that this decline in activity takes place at nighttime even though we don't know the exact time of day.

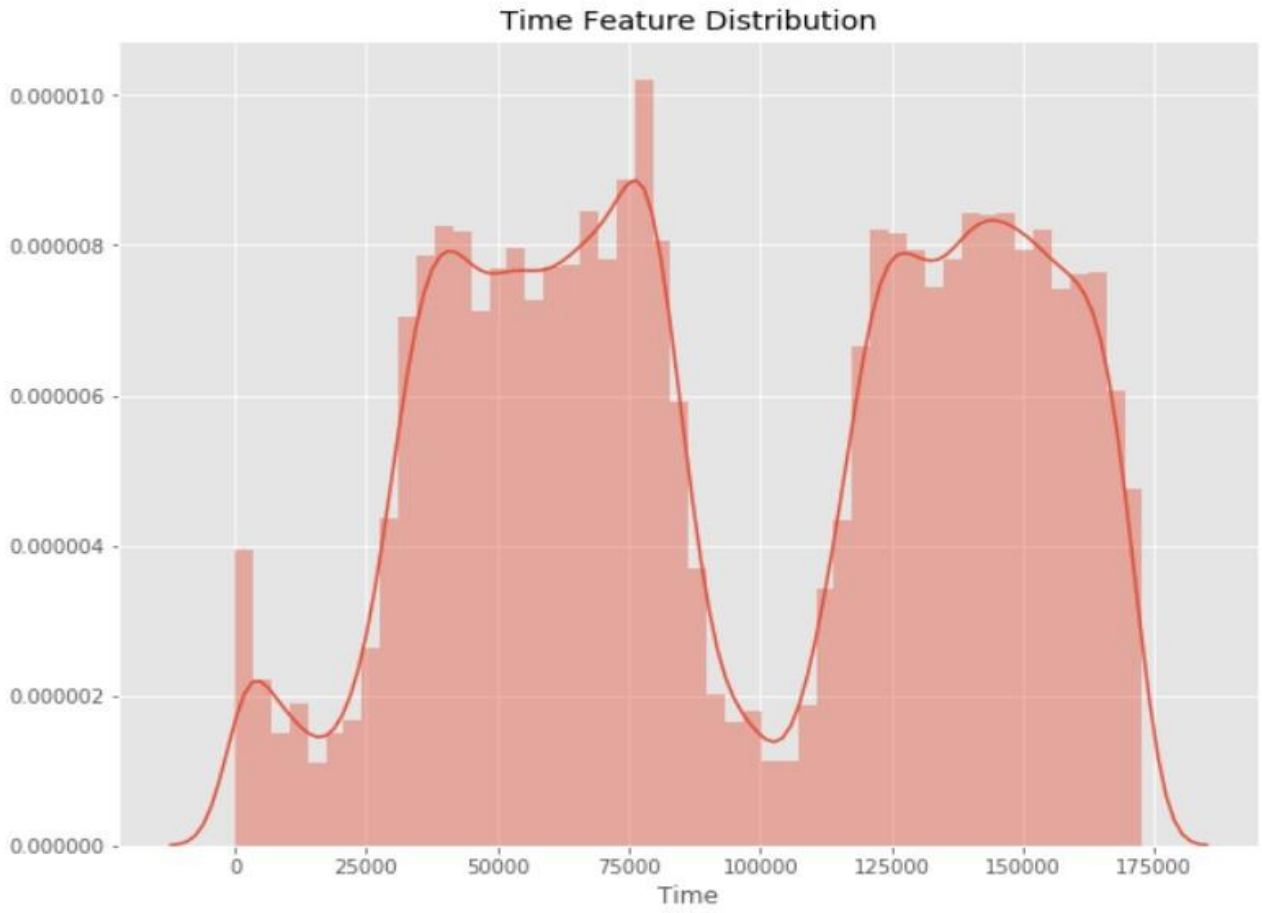

Figure 2: Distribution of time

Data Preparation

Our early research has revealed that our anonymised columns have been scaled for us. We must scale our Time and Amount features as well before we can fully analyse our data set. We run the risk of our machine learning algorithms not performing well without scaling these features. We can use StandardScaler from sklearn to standardise these attributes (see Appendix A for details on specific methods used throughout). By using this tool, the data is transformed so that the distribution's mean value is 0 and its standard deviation is 1. This enables us to create machine learning algorithms by giving our features a common scale.

Remember that the bulk of the transactions in our original data set were not fraudulent, creating a significant imbalance. We face the danger of overfitting with our algorithms if we use this data set as-is without correcting for this imbalance because it is likely assumed that the majority of transactions are legitimate. Before moving on with our study, we want to account for this imbalance since we are searching for a model that will identify patterns rather than make assumptions about outcomes. To correct the significant transaction type imbalance seen in the original data set, we must establish a training data set. Instead of assuming incorrectly that "most" provided transactions will be fraudulent, we will be able to develop a model that can recognise fraudulent transactions and classify them appropriately in this way. In order to achieve this, we may employ the random under-sampling approach, which effectively eliminates data in order to provide a training data set with a more evenly distributed distribution of transaction type (fraudulent or non-fraudulent). This will make it necessary for our systems to recognise fraudulent transactions when they occur.

The amount of fraudulent transactions in the original data set can first be counted in order to establish the balanced training set that will be utilised. We will have a 50/50 combination of transaction classes if we randomly choose the same number of legitimate transactions from the data set. Next, we want to combine the two "lists" and shuffle them in order to get a random class type. Since we chose the same number of non-fraudulent transactions as there were fraudulent ones (435), our new data set now comprises 870 total transactions. It is simple to notice how much more balanced the class distribution is now by comparing it to the distribution in the original dataset and the newly produced subsample.

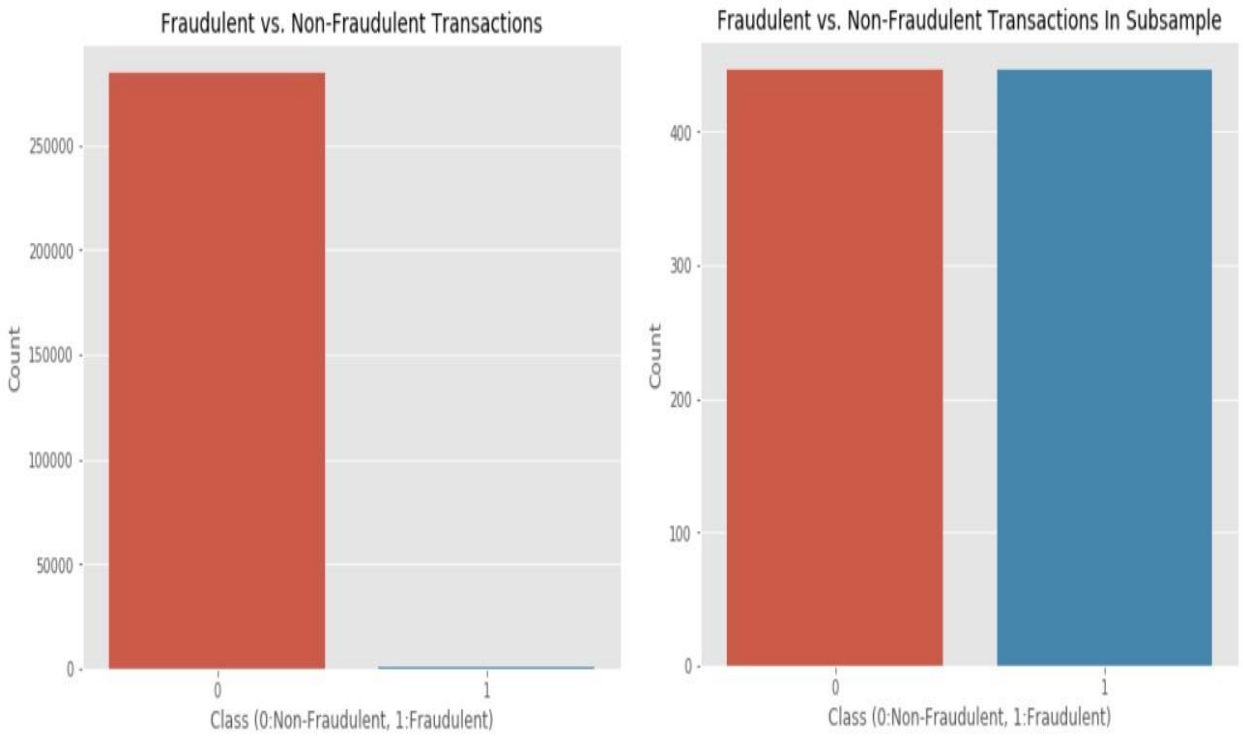

Figure 3: Class Distribution in original vs. subsample

Instead of utilising accuracy measures to evaluate performance, we will instead utilise the Receiver Operating Characteristics-Area Under the Curve (ROC-AUC) performance metric. For evaluating the success of classification models, which is what we will be doing, the ROC-AUC is a very helpful statistic. The AUC indicates the level or measure of separability, whereas the ROC is a probability curve. 2019 (Narkhede). The result of this measuring technique is a number between 0 and 1, where 1 represents a perfect score and 0 represents the reverse. The model's capacity for prediction improves with increasing score. By examining the AUC curve, we may even see this performance in graphic form. For this kind of data set with an uneven categorization, a confusion matrix is not a useful tool for evaluating accuracy.

Finding and dealing with any potential outliers is a crucial stage in the analysis of data. The way you handle outliers may have a big impact on the data you have and ultimately the accuracy of your model's predictions. I decided to look at attributes that had a correlation of at least 0.5 with the class variable in order to remove outliers from this data set (fraudulent or non-fraudulent). Before we go ahead and actually eliminate any outliers, we can first visualise the positive and negative associations with various characteristics. We utilise box plots to visualise our data and determine our interquartile range (IQR). Box plots are an effective tool for this as they make it simple to view the 25th and 75th percentiles and spot any extreme outliers. When defining our standards for eliminating outliers, we need to be fairly cautious. A lower threshold will eliminate more outliers, but we might only want to exclude the most extreme outliers in order to keep the majority of our data and reduce the risk to the accuracy of our model in both directions. Normally, if a data point is outside of 1.5 x IQR, it is regarded as an outlier; however, with our data set, using this criterion would drastically limit the amount of training data we haveWe will choose to only eliminate points that are outside of 2.5 x IQR because of this. Our subsample is reduced from 870 transactions to 612 transactions using these parameters.

We may apply a dimensionality reduction approach to project higher dimensional distributions into lower dimension representations as we are now unable to see our classes in many dimensions. Our data are organised in a high-dimensional space, and the t-distributed stochastic neighbour embedding (t-SNE) technique provides a way to visualise this arrangement. In both the high-dimensional and low-dimensional spaces, this method computes a similarity measure between pairs of instances. It then makes an effort to maximise the two measures by calculating the difference between anticipated and expected valuesThe transactions in our data set that were fraudulent and those that weren't can be effectively clustered using T-SNE. We may see the clusters produced by t-SNE by plotting our data on a two-dimensional plane and using a scatterplot.

Training Algorithms

We now need to go forward and train and test our classification algorithms after properly preparing and cleaning our data set to make it usable. To achieve this, we first divide the data into two sections using an 80/20 train-test split. In order to prevent overfitting for our fairly small sample, I utilised the k-fold cross-validation approach for resampling. The data set is shuffled, divided into k groups, worked on separately, and the model's ability is then summarised using a sample of

model assessment scores. The various categorization algorithms we have at our disposal may then be examined to see how well they would work with the data we currently have. We can choose the best algorithm for our model by weighing the performance of several algorithms and quickly viewing them. Several widely used classification techniques were side-by-side evaluated in this study: Logistic Regression, Linear Discriminant Analysis, K Nearest Neighbors (KNN), Classification Trees, Support Vector, and Random Forest. This snapshot's results are displayed in Figure 4.

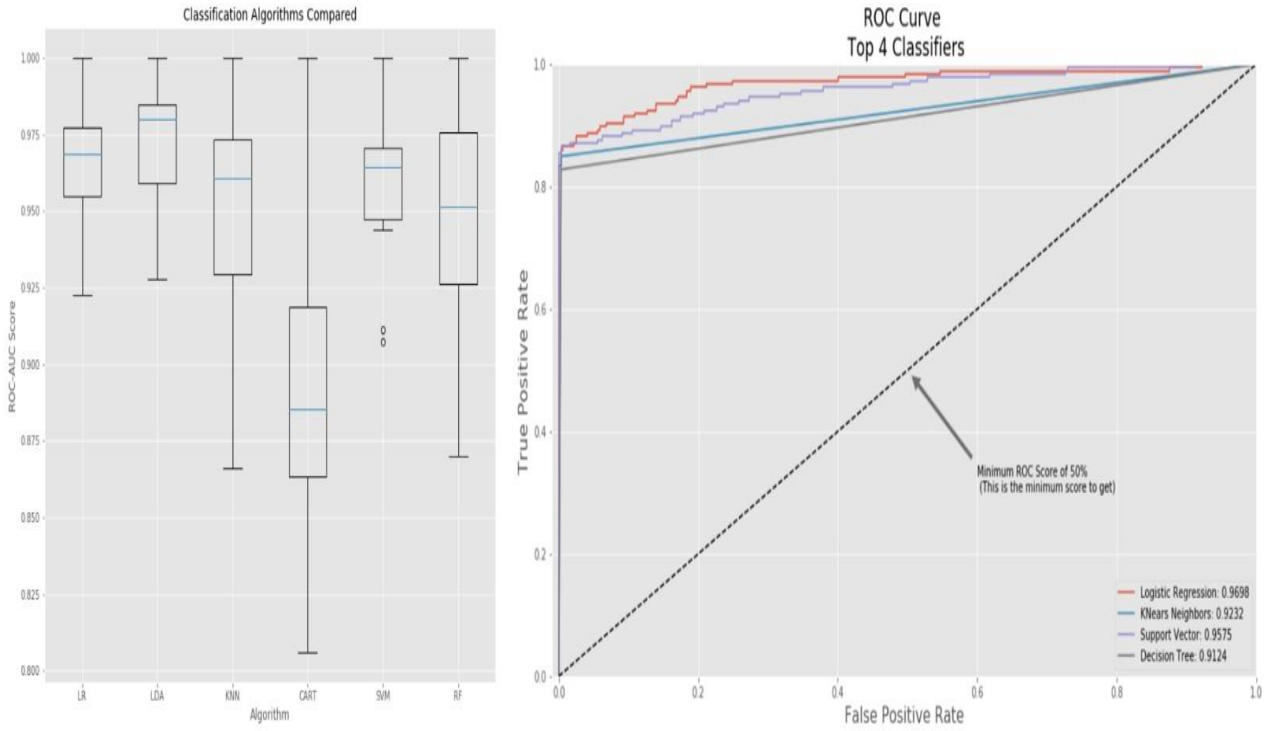

Figure 4: Classification algorithms compared

We can observe that several of the examined algorithms provide results that are comparable, but some perform better than the others. For our data, the Logistic Regression technique appears to be more accurate than the competing classifiers. We then use the trained version of our model to generate predictions. The outcomes from the use of logistic regression are more than satisfactory. While the 1 class (fraudulent transactions) has 97% precision, the 0 class (transactions without fraud) is predicted with 94% precision and 99% recall. This indicates that the system only misses 3% of transactions that are fraudulent. More training data may be provided to further enhance this.

We were able to apply logistic regression to create an accurate model for forecasting fraudulent credit card transactions by investigating, cleaning, and organising our data.

Appendix

Main libraries used:

- Matplotlib: Visualization with Python

- Scipy: Python-based ecosystem of probability distributions and statistical functions.
- Numpy: Core library for computing with Python.
- Pandas: Open-source data analysis and manipulation tool.
- Seaborn: Python data visualization library based on Matplotlib.

References:

The domain from which the data is derived is from the organization Worldline and the Machine Learning Group [http://mlg.ulb.ac.be/].

1. Albon, C. (2018). Machine learning with Python cookbook: practical solutions from preprocessing to deep learning. Sebastopol, CA: OReilly Media.
    a. Alternatively, this textbook presents algorithmic method for data analysis and deep learning using the Python language.
2. Machine Learning Group. (2018, March 23). Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/mlg-ulb/creditcardfraud/home
    a. This is my data source and general description of our data set, housed within Kaggle.
3. Machine Learning Group. (n.d.). DEFEATFRAUD: Assessment and validation of deep feature engineering and learning solutions for fraud detection. Retrieved from https://mlg.ulb.ac.be/wordpress/portfolio_page/defeatfraud-assessment-and-validation-of-deep-feature-engineering-and-learning-solutions-for-fraud-detection/
    a. This project description outlines the overall goals of the Machine Learning Group as they attempt to develop new and improve upon existing mechanisms for detecting credit card fraud transactions using machine learning algorithms.
4. Couronne, R., Probst, P., & Boulesteix, A.-L. (2018, July 17). Random forest versus logistic regression: a large-scale benchmark experiment. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5
    a. This research article discusses the differences, benefits and disadvantages to using random forest versus logistic regression algorithms for regression and classification.
5. DataFlair Team. (2019, October 11). 11 Top Machine Learning Algorithms used by Data Scientists. Retrieved from https://data-flair.training/blogs/machine-learning-algorithms/
    a. This article lays out some of the most commonly used machine learning algorithms for data analysis and model development. The article begins by discussing supervised learning algorithms, and moves into unsupervised algorithms, to cover a breadth of available options.
6. DataFlair Team. (2020, February 19). Project in R - Uber Data Analysis Project. Retrieved from https://data-flair.training/blogs/r-data-science-project-uber-data-analysis/
    a. This is an unrelated project which analyzes Uber pickup data for New York City. The assessment of the data and general work through are helpful as they provide a

sort of vague wireframe for how to address a data set with the intent to create predictive models.

7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning: with applications in R. New York: Springer.
    a. This textbook helps to review many statistical methods for data analysis using R, including logistic regression and random forests.
8. Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Hoboken, NJ: Wiley.
    a. A textbook which offers a guide to data visualization, from the exploratory step through the project presentation step.
9. Nadim, A. H., Sayem, I. M., Mutsuddy, A., & Chowdhury, M. S. (2020, February 13). Retrieved from https://ieeexplore.ieee.org/document/8995753
    a. A secondary article discussed the use of machine learning algorithms to address credit card fraud, investigating the use of various regression algorithms in the process.
10. Puh, M., & Brkić, L. (2019, July 11). Detecting Credit Card Fraud Using Selected Machine Learning Algorithms. Retrieved from https://ieeexplore.ieee.org/document/8757212
    a. This article discusses the growth in interest for applying machine learning techniques to the mission of detection fraudulent credit card transactions and points out the challenges that can arise with these attempts.