Project 1: Detect Credit Card Fraud

# **Lara Clasen Spring Term 2020 LLClasen.github.io**

# Which Domain?

*What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.*

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.
   1. 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>
   1. 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/>
   1. 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>
   1. 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/>
   1. 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/>
   1. 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.
   1. 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.
   1. 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>
   1. 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>
    1. 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.

# Which Data?

*What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.*

The data being used is a Card Transactions dataset that consists of a combination of fraudulent and non-fraudulent transactions. There are nearly 285,000 rows of data and 31 features. 28 of these features are nondescript for the public for security reasons, but the three I am interested in are Time, Amount, and Class (fraudulent or non-fraudulent). The data set itself can be found at [https://www.kaggle.com/mlg-ulb/creditcardfraud/home].

# Research Questions? Benefits? Why analyze these data?

*How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.*

Because our data represents just two days’ worth of transactions, we are able to narrow our approach a bit. We can expect to see generally normal distribution when we look at our Time feature, and we can make several assumptions thanks to this. I want to first identify whether that normal Time distribution does in fact exist, as well as review the credit card transaction Amounts and the number of fraud/non-fraud transactions that exist in the original data set. By finding descriptive statistics on these variables we will be able to form relevant research questions, such as: At what time of day do we see the highest fraudulent activity? What is the average transaction amount of a fraudulent transaction?

These data are important to analyze as our world becomes increasingly automated and electronic. With credit card fraud on the rise, we must continue to be diligent about finding ways to cut through the falsehoods.

# What Method?

*What methods will you be using? What will those methods provide in terms of analysis? How is this useful?*

Beginning with exploratory data analysis, I will use machine learning algorithms to understand and work through these data. I plan to begin with logistic regression which will attempt to assess how incorrect a prediction is, and random forests to predict whether a transaction is indeed fraudulent. I will then assess models for accuracy, recall and precision. It may also be worth looking at the confusion matrix for this particular data set to better visualize the effects of each model.

# Potential Issues?

*What challenges do you anticipate having? What could cause this project to go off schedule?*

I believe that I have all of the necessary tools to complete this analysis using skills learned in previous courses. That being said, I have been unexpectedly derailed in prior projects due to differences in data characteristics, previously unused methods, etc., that I have not yet been tasked with. It will likely be a bit of a learning curve to work on a new project on my own, but at this point I do not expect any major setbacks.

# Concluding Remarks

*Tie it all together. Think of this section as your final report’s abstract.*

For my first project I plan to analyze a data set containing fraudulent and non-fraudulent credit card transactions, with the intent of building model to best predict when a transaction represents fraudulent behavior. This type of model could be immensely useful in today’s increasingly electronic world, where millions of transactions occur daily, many of them fraudulent. By exploring the data set and understanding trends, I will build and train multiple models to predict which transactions are fraudulent, ultimately aiming to test and select the most accurate model for this set of data.