ECO 395 Project: Fraud Exploration and Detection

Patrick Chase

5/8/2021

Abstract

The problem I'd like to solve is the identification of credit card fraud for a leading payment service company, Vesta Corporation, who specializes in guaranteeing card-not-present (CNP) transactions. Vesta partnered with IEEE Computation Intelligence Society in hosting a competition on Kaggle with a real world data set. The goal was simple. Create the most effective classification model to identify whether or not a transaction is fraudulent. While the competition is closed, I think this particular problem and rich data set provide a real world opportunity to demonstrate the skills I've attained in this class.

Introduction

Fraud detection has been a persistent problem since the widespread the use of non-cash payment systems became popular in the mid 1990s. Given the regulatory environment of the United States and the massive increase in the amount of transactions, corporations have a large incentive to prevent fraud in real time. These circumstances present a ripe environment for automation through the use of machine learning, which has been relatively common. Banks, payment processors, and tech companies such as Apple, Amazon, and Microsoft devote copious resources to the development of automated fraud detection systems.

Despite those efforts, billions of dollars are lost each year due to fraud in a diverse range of fields. Bad actors and corporations are in a constant arms race when it comes to fraudulent activity. Whether it's Facebook attempting to detect fraudulent ad buys on their site or a payment processor such as Vesta preventing fraud from occurring at the transaction level, being able to effectively detect and prevent fraud is in the interest of businesses and consumers. While potentially relatively simple, automated fraud detection through machine learning present an easy avenue into the use of algorithms for a typical business.

Methods

Data

The *data sets* used in this analysis was found through a Google search of "fraud detection data sets".

First and foremost, a few circumstances forced me to select a random sample of the provided training and test data sets. The main issue was the fact that I did not have enough memory on my machine to effectively analyze it locally. The second issue was the fact that Github limits the file sizes to 100MB. My preference is to upload my data sets to Github and reference the raw format of those so it isn't necessary to download

So, I begin with a simple exploration of the data and identification of some existing relationships that are easily apparent. Then I construct principal components that are used for a logit model to classify the variable of interest, isFraud.

#Methods #Results #Conclusion