University of Arkansas - CSCE Department
Capstone II - Final Report - Spring 2019
Unsupervised Call Classifier
Liam Dillingham, Spencer Erdman, Damian Wells, Jonah Keller

1 Abstract

Spam calls are an issue that affects nearly everyone with a cellphone. Partly due to security flaws from other companies, our personal information is freely available on the web. This makes us vulnerable to regular phone calls from bots, telemarketers, and scammers. We want to explore the structure of phone call data, and use data science techniques in order to mitigate the occurrence of spam calls.

2 Problem

Since the ground truth of the data we are given is not known, it is up to proper preprocessing techniques to wrangle the data into a format where unsupervised clustering techniques can be properly used. Then it is up to experimentation to find a technique that properly separates the data into a binary classification using only the natural struture of the data. These classes can be investigated further to derive the truth of the classes.

3 Objective

The goal of this project is to assist the company First Orion in improving upon their already existing models which detect and stop phone calls. Whether through their carrier technology, which stops these detected calls at the network level, or on user's phones through an app which performs an automatic hang up, we want to take given data, derive useful results using unsupervised techniques, and use those results to improve their data on spam callers, and thus improve the technologies and products they offer their customers.

4 Background

4.1 Key Concepts

- **<u>Unsupervised Learning:</u>** When we have data which is unlabeled, and we need to derive a result from (in this case classification), we cannot use traditional methods of learning. Since the **ground truth** is not known, we have to use techniques which explore the natural structure of the data, and do an investigation of the results

- **<u>Ground Truth:</u>** The value, or label of an observation or row of data, given that the truth is known. i.e. not the predicted truth.

- **<u>Data Wrangling:</u>** The act and process of cleaning data, removing outliers, and processing it such that it can work with known and proven techniques and algorithms.

4.2 Related Work

The First Orion data science team is regularly working on models which detect and classify phone numbers as spam or legitimate, and our goal is to improve upon the techniques and give insight so that the team can provide a better product to carriers and users.

5 Design

5.1 Requirements and/or Use Cases and/or Design Goals

The most important part of these insights is to reduce false positives. That is, to not block people who may be high-volume, but not actually spam. This is important because then people who have legitimate needs to make many phone calls in a short amount of time will be blocked from calling, and this could have serious implications for business or otherwise.

5.2 Detailed Architecture

Our data consisted of two sets. The first set was a 6-minute frame of streaming phone call data. It consisted of 3 features. The encrypted phone number, an EPOCH time for when the call was initiated, and a flag to discern between a landline and a mobile call. The task was to first split the data into separate sets between landline and mobile. Since the dataset was rather larger ($> 70,000$ rows) this first step reduced the size of the dataset substantially. The we ran over the data and counted the number of phone calls made by each number. After an interquartile range analysis, we decided than anyone who makes more than 12 phone calls in 6 minutes on a landline and 11 on a cellphone is considered an anomaly. This doesn't necessarily mean that all those callers are spam, but they are certainly an anomaly.

The next step was to take our anomalies detected in the streaming phone data, and use them to filter from our other dataset. This dataset, contains a list of historical phone call data about many phone numbers. It was about 80,000 rows with 31 features. To reduce the size of this dataset we then used both the reduced cellphone and landline data to pull out matching phone numbers. This was then the data we were to work with.

We wanted to use principal component analysis (PCA) to figure how many dimensions were required to explain at least 85% of the variance in the set as a whole. However, there were some extreme outliers in the data. A few of the feature columns had a single outlier which dominated the variance of the entire set and broke the PCA algorithm.

(a) Correlation matrix of data with outliers (cellphone)
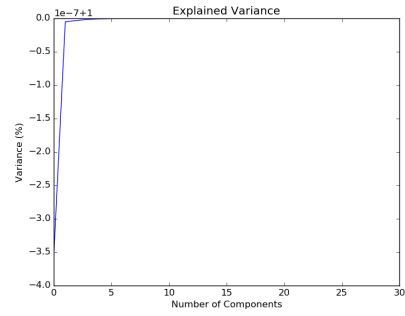
(b) Variance elbow curve (cellphone)

We can see from the variance plot, and the correlation matrix above that a single feature column had variance so high that the relationship between all other features is obfuscated by such a large variance. In addition, many of these columns have a variance of 0 (that is, all fields have the exact same value for the whole column) so they are useless when deriving some sort of insight for the data. In most cases, they can ruin the results of a model. So we needed to find a way to remove these useless columns and model-destroying outliers.

So we had to experiment with several different methods of outlier removal in order to pre-process the data in such a way that the algorithms would work. We used a modified z-score approach to remove rows with outliers that dominated the variance of their columns. The standard z-score approach is to divide the difference of the column mean and the data value by the standard deviation of the column. The modified z-score uses the difference between the median and the value, and divides by the mean absolute deviation instead. This method is more robust to outliers, and helps us calculate them. After doing these two steps, we end up with a much better looking correlation matrix and elbow curve
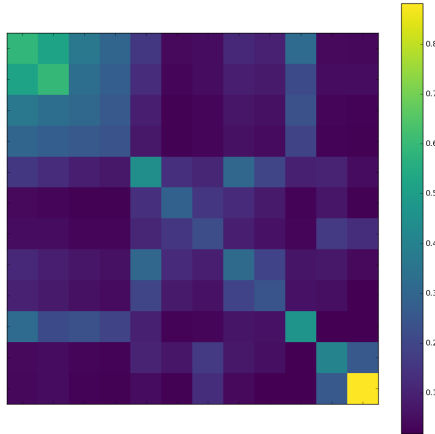

(a) preprocessed correlation matrix (cellphone)

(b) Variance elbow curve (cellphone)

3
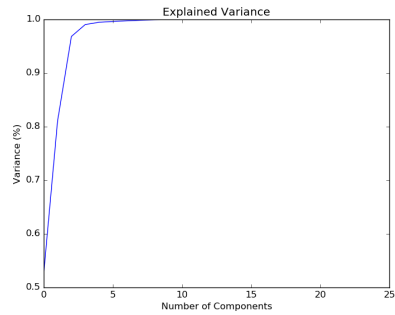
Keep in mind that the data we are looking at here is specifically the cellphone data. The landline data, while suffering from the same problems, ended up being a much simpler result after applying the same methodologies: Below are the before and after results of preprocessing the landline data:

(a) unprocessed correlation matrix (landline)

(b) Variance elbow curve before preprocessing (landline)
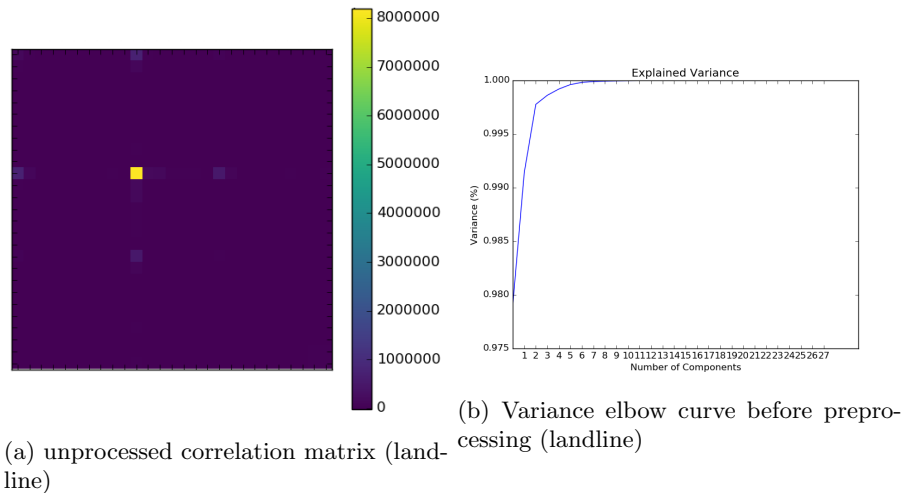
And after feature selection, and outlier detection and removal:

(a) preprocessed correlation matrix (landline)

(b) Variance elbow curve (landline)
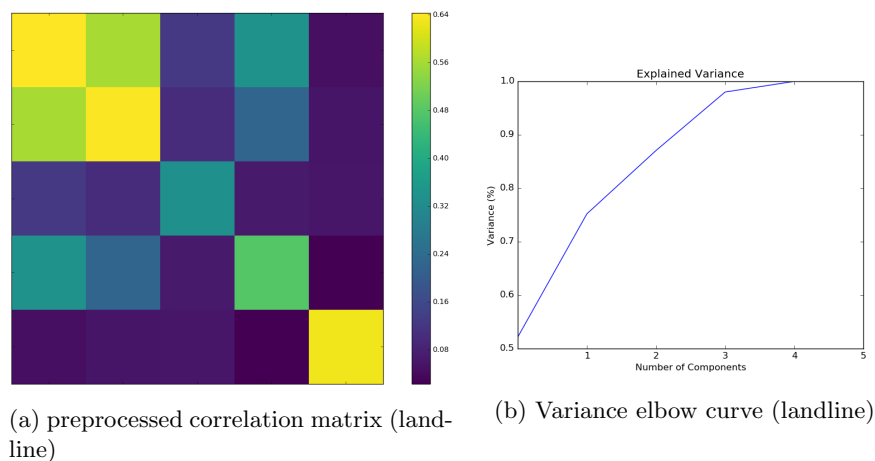
Once this was completed, that is, the data is successfully cleansed/wrangled, we now needed to explore several clustering techniques to try and find a natural separation in the data.

5.3 Risks

- **False Positives:** We reduce this risk by first setting a threshold below which "normal" callers are placed. By doing so, we reduce the amount of variance in the data and can focus entirely upon the anomalies. Since we are only dealing with the anomalies, the average person won't be accidently detected, and as a consequence the different between a legitimate anomaly (a person with real needs to make many calls) and a spam caller

will become more apparent, as the structure of the data overall won't be obscured by the average person.

5.4 Tasks

5.5 Schedule

5.6 Deliverables

- **R code:** This is the code use to wrangle some of the data and split it into subsets, as well as detect and remove outliers

- **Python code:** This is the code used to run clustering and plotting algorithms on, produce plots, etc.

- **Insights:** Our project won't neccesarily have a deliverable, the deliverable is, the insights we can provide with our techniques, that is differentiating between high-volume legitimate callers and spam callers. The spam callers will have to be investigated on their own once the results are complete.

6 Key Personnel

- **Liam Dillingham** - A senior Computer Science and Mathematics (Statistics) double major in the CSCE and FullBright college at University of Arkansas. Complete relevant courses: STAT 4003 (Statistical Methods), Machine Learning (CSCE 5063), Big Data Analytics and Management (CSCE 5273), and Computational Statistics (STAT 5443). Relevant experience comes from previous internships as a dual-role software engineer/data scientist, as well as course-related projects. Tasks included working with industry champions to produce satisfactory results and develop code and insights from data

- **Industry Champion: Nysia George** - Nysia is a senior level data scientist at First Orion, and was tasked to be our mentor during the course of the project. She is who we brought our results to, asked questions, and took guidance in order to provide her with results that aided in her work and improving the company's products.

7 Facilities and Equipment

Initially we believed we needed a larger computer in order to run some of the more complex algorithms on the data, however, after sufficient preprocessing, we were able to reduce the size and complexity of the data to a much more manageable scale.