William Orgertrice

Vilexia Jackson

Sanghwan Park

**Data Mining Final Report**

**Abstract:**

The goal is to figure out how often fraud occurs and what is the estimated amount that people tend to consume. We hope to find fraudulent activity and make it easier to detect. Our results from the dataset we used illustrated that the age range that is most likely to experience fraud is between the ages of twenty-six and thirty two.

### **Introduction:**

We decided to utilize a bank’s database that provides information regarding customers personal information and specific transactions. The information provided will allow us to compare and contrast the similarities between customers that have experienced fraud and those that have not. There will be a percentage provided at the end of the database that informs the user on the likelihood that they will and will not experience fraudulent activity on their account.

The main purpose of the bank database is to improve communication between the customer and the bank. As well as inform the customer as to whether or not they should consider gaining fraud protection or choosing another bank to bank with. We will use an algorithm that will mine through and analyze the bank’s data regarding the different customers. If this algorithm identifies any attributes similar to those who have experienced fraudulent activity, the user will be notified.

### **Problem Statement:**

Fraud is a trend that has continuously increased and affects people throughout the world. There are several thousand cases composed of people having their identity stolen or have experienced fraudulent activity on their bank account. Because having a bank account is a necessity for most, the issue of fraud affects millions of people. It is important to be aware of the possibility that fraud can happen, especially when one cannot afford to pay for the overdraft fees as well as the amount of money stolen.

Therefore, our project is centered around detecting the likelihood of one experiencing fraud based on a few attributes. Our project is outlined from a bank’s data. Our biggest issue was cleaning the large dataset. We originally had 10 attributes but we had to shorten it to 3 attributes: age, amount, and fraud.

**Data Description and Pre-processing:**

We used a synthetic dataset for financial transactions that has nominal, binary and numeric attributes. The data is in the format of a matrix. The data that we utilized for our project included ten different attributes. The data consists of age, gender, zip code, category, gender, zip merchant, amount of the transactions and if fraud was detected or not. The age attribute is from a range of 1 to 6, with 1 being less than or equal to 18 and with 6 being greater than the age of 65. The zipcode of the dataset comes from a bank in Madrid, Spain. The category attribute is the item that the customer purchased. The fraud attribute is a binary attribute that marks the occurrence of fraud within our dataset.

In order to make our results from the algorithms more accurate, we did not utilize all of the attributes. The data we utilized included over five-hundred thousand rows, as a result we had to clean the data. Therefore, we only utilized three of the ten attributes, which gave us more accurate results. The three attributes chosen was age, the transaction amount and if fraud was detected.

We did not include the zipcode seeing that everyone that used the bank lived in the same area. The gender of the customer was not vital information because fraudulent activity was equal amongst all genders. However, the age and fraud attributes contributed the most information. After analyzing our results, the group realized that the age group that was affected the most by fraud was between the ages of twenty-six and thirty-two. This data provided enough important detail to help us find similarities amongst those who experienced fraud.

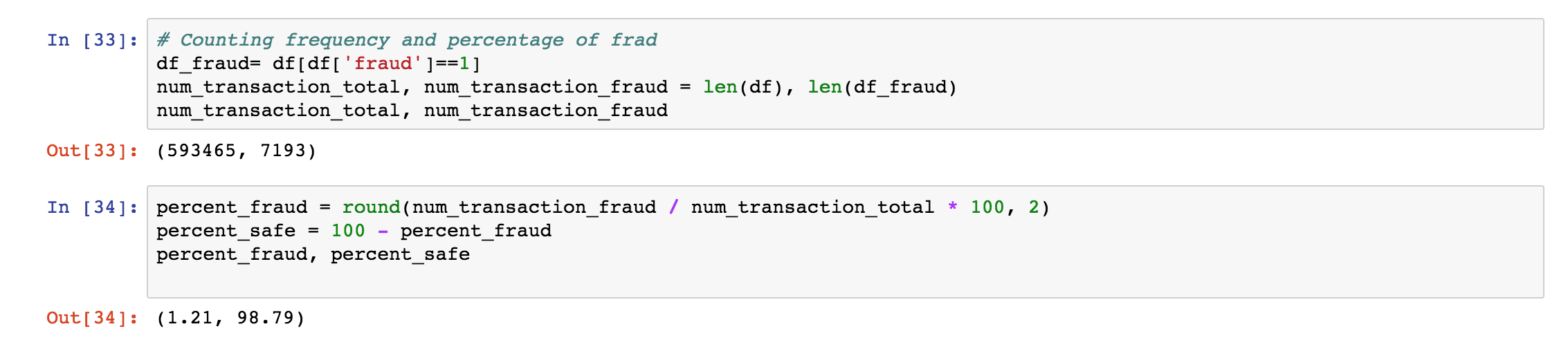
### **Description of Algorithms:**

We utilized the K-Means algorithm. We chose the K-means algorithm because it is easy to apply to a large data set. K-means is a method of vector quantization, which originated from signal processing. The K-means algorithms uses cluster observation into a group of related observation without any prior knowledge of those relationships. The goal of the algorithm is to partition a particular amount of observations into a specific amount of clusters. Depending on the amount of clusters, will determine which observations goes where. Each observation will cluster with the nearest mean, which in turn serves as a prototype of the cluster. The K-means algorithm is one of the simplest clustering techniques and it is commonly used often in computer science fields.

For our project we also utilized a decision tree regression. A decision tree is a flowchart-like tree structure, where each node has two or more branches that is tested on an attribute value. A decision tree can test both categorical and numerical attributes. Each branch represents an outcome of the test and the tree leaves represent classes or class distributions.

**Experiments:**

Our original data set contains exactly 59,4643 rows of data. That data was distributed amongst ten different attributes. According to our data, fraudulent activity occurred 7,193 times out of the entire data set. If you take the average of both the number of rows and the actual times that fraud occurred, the percentage that fraud happened is approximately 1.21%. Therefore, at this specific bank, fraudulent activity occurs less than ten percent of the time.

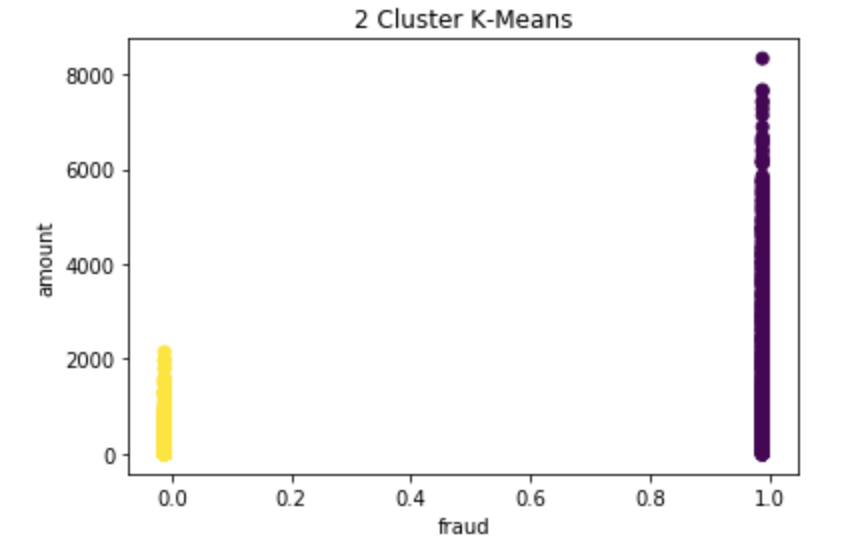


The chart above highlights the percentages of fraudulent activity that occurred within our dataset.

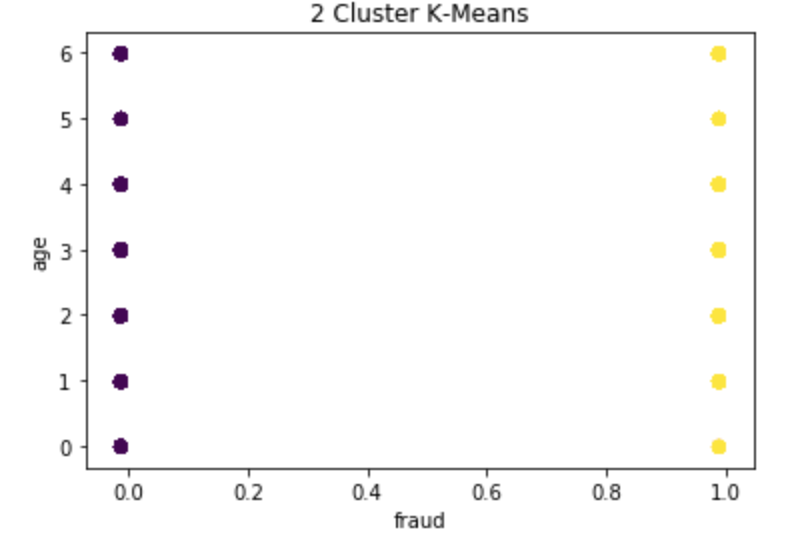
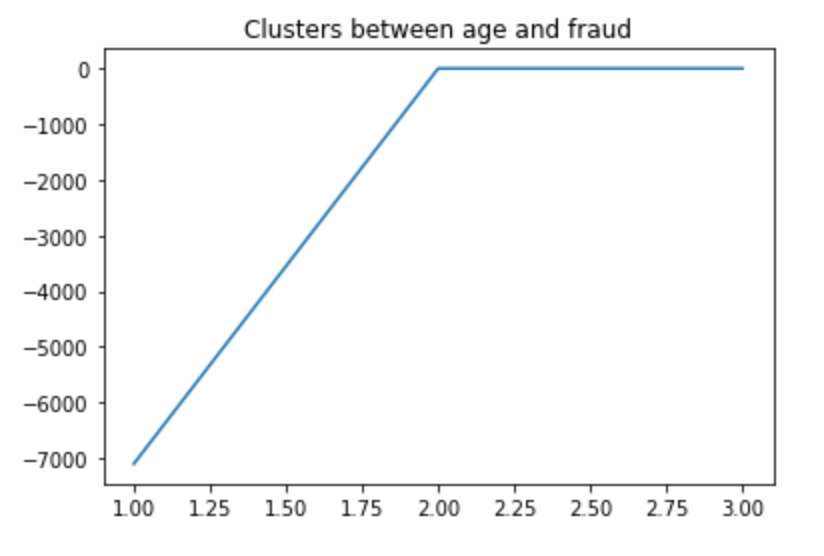
**Algorithm 1: K-Means**

The first step we needed to take was to analyze our data set and then implement the K- means algorithm. Therefore, we used the elbow method. The elbow method is a method that highlights interpretation and the validation of consistency, within cluster analysis. The elbow method is designed to help find the appropriate amount of clusters within a data set. We analyzed the dataset and its attributes. When implementing the age, amount, and fraud attributes, our results gave us two clusters. However, when we implemented the attributes of the customer’s age and the transaction amount, the algorithm gave us over ten different clusters. Due to developing our project with a large dataset, running the program required an abundant amount of time.

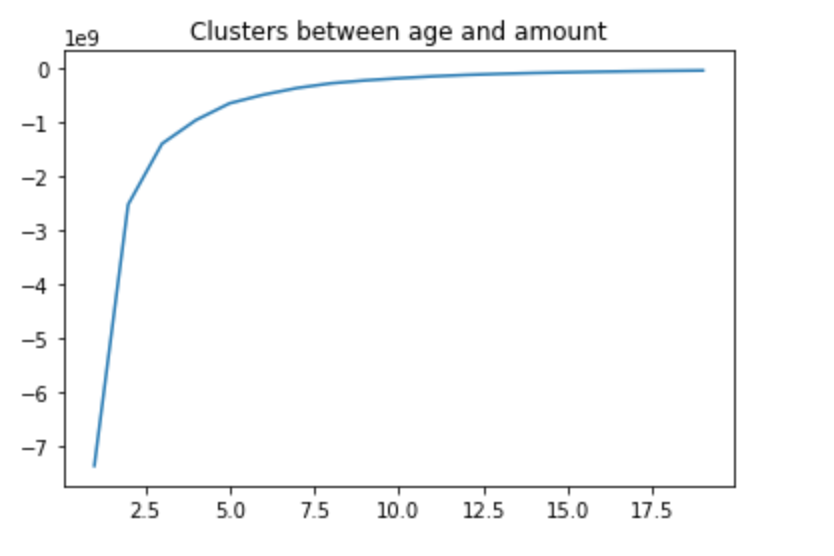
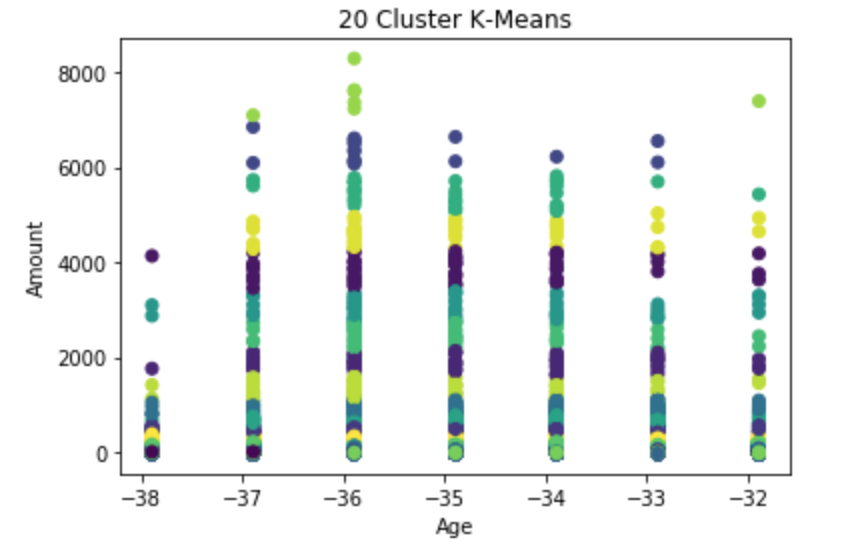
The figures below are the results from fraud and amount using the K-means algorithm.



The figures below are the results of using age and fraud with the K-means algorithm.



The underlying fact regarding our data is that fraudulent crimes can happen to anyone, regardless of any given attribute provided by our data.

The figures below are the results of using age and amount: 

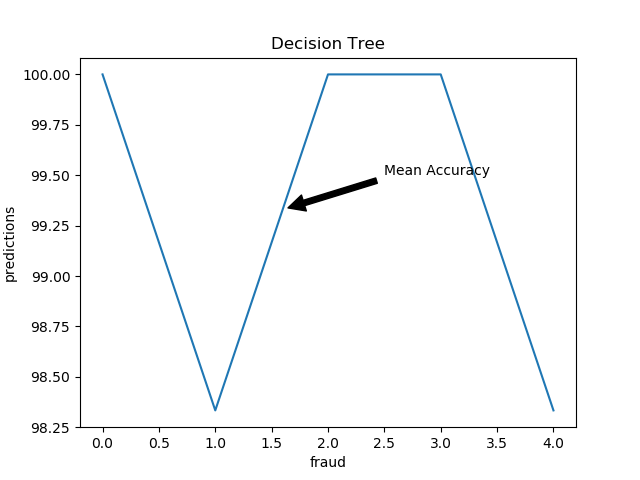
We implemented the k-means algorithm between different attributes. As stated previously we ran a program with all three attributes and one with age and the transaction amount. We decided to run different programs so that we could know how the transaction amount corresponds to the customer’s age. Compare to the original authors report, our amount of consumption graph looks same with theirs report. Because of the large amount of data from our dataset, we did not get several clusters. Fortunately, we could estimate the clusters from the elbow method graph. As the x value is approaches twenty, the y values share the same slope. Therefore, there are at least twenty clusters between them.

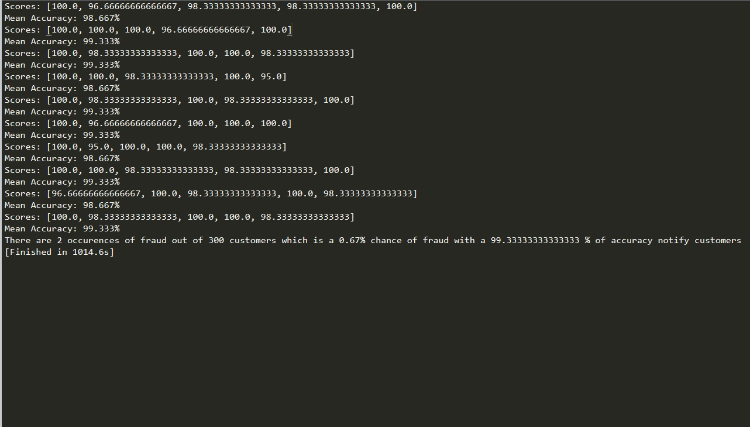
**Algorithm 2: Decision Tree Regression**

The decision tree algorithm loads a csv file and converts the cells of the dataset into floating point numbers. It then splits the dataset into k folds and we set k to be 5. It splits the dataset by the amount of attributes and attribute values. By doing this it starts to build the tree by splitting into smaller subsets starting from the root node causing the tree to get incrementally larger.

The program outputs the accuracy prediction starting from the root node then grows down into the leaf nodes going from left to right node. The algorithm uses standard deviation in order to form its prediction. If there are no more leaf nodes then the program stops and gives its final result.

Below is the predictions vs. fraud graph created by the decision tree algorithm



Below is the output of the decision tree algorithm:****

The decision tree algorithm outputs the accuracy prediction of the dataset with only 300 rows in order for it to run quickly for testing. We get two occurences of fraud with the smaller dataset which results is 0.67%

### **Conclusions:**

In conclusion, we were able to implement both the K-means and Decision tree regression algorithms. Our data set has five-hundred and four-thousand rows and dispersed amongst ten attributes. However, we only utilized 300 data for the demo live to shorten the run-time. However, we have included all the data from the original data set. Fraudulent activity occurred over seven-thousand times, which was only 1.21 percentage occurrence from the data.

According to our data results, we have come to the conclusion that fraudulent activity can happen to anyone, regardless of their financial stability or age. Anyone can become a victim of fraud. Though, it is likely for anyone to be affected by fraud, we were able to predict the probability of one experiencing fraud on their account. This prediction allows us to notify customers with the probability of fraud in their area.

### **References**

*[Fraud Detect] Visualization to Classification | Kaggle*, www.kaggle.com/funkyong13/fraud-detect-visualization-to-classification/notebook.

“KDnuggets.” *KDnuggets Analytics Big Data Data Mining and Data Science*, www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html. accessed September 16, 2017

Alam, Sarfaraj. “21 Best Data Mining Project Ideas For Computer Science Student.”*Programming Assignment Help, Java Assignment Help*, www.lovelycoding.org/2015/08/21-best-data-mining-project-ideas-students.html. accessed September 16, 2017

Website: Www.ijetae.com. “Data Mining Techniques and Its Applications in Banking Sector.”*Data Mining Techniques and Its Applications in Banking Sector*. accessed September 17, 2017

*Datasets*, www3.dsi.uminho.pt/pcortez/data/. Accessed September 18, 2017

*Synthetic Data from a Financial Payment System | Kaggle*, www.kaggle.com/ntnu-testimon/banksim1/data.

“How To Implement The Decision Tree Algorithm From Scratch In Python.” *Machine Learning Mastery*, 11 Aug. 2017, machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/.

“Pyplot Tutorial.” *Pyplot Tutorial — Matplotlib 2.0.2 Documentation*, matplotlib.org/users/pyplot\_tutorial.html.

### **Division of Work and Contributions**

|  |  |
| --- | --- |
| Task | People |
| 1. Collecting and preprocessing data | Sanghwan Park, William Orgertrice |
| 2. Implementing Algorithm 1 | Sanghwan Park |
| 3. Implementing Algorithm 2 | William Orgertrice |
| 4. Evaluating and comparing algorithms | Vilexia Jackson |
| 5. Slides, demo and Presentation | William Orgertrice, Sanghwan Park, Vilexia Jackson |
| 6. Writing report | William Orgertrice, Sanghwan Park, Vilexia Jackson |