Daniel Deloughry

-

R00115920

-

SDH4

-

Machine Learning

-

Repeat Research Assignment - Report

-

Cork Institute of Technology

Contents

Abstract

In this project I shall be looking at the numbers of shootings occurring in each American state. After reading articles I found that some states have had higher levels of shootings than others. I found that some states have stricter gun control laws than others.

Chapter 1

Introduction

My motivation for writing software to predict the state in which an attack occurs, is because if it can be predicted from characteristics of an attack, then this knowledge could potentially prevent future shootings, or at least assist in protecting against them.

In the United States of America, guns are legal to own. As the country is a federal republic, each state does not have exactly the same laws on ownership. The country has, for a long time, had a problem with gun violence. Although some states, such as Illinois, have stricter gun laws, the number of shootings is still large. I am interested to see if the details in the dataset might help elaborate.

This document is divided into four chapters. As you have just read, *Chapter 1* is the motivation of the project. It details the motivation, summary and plan of the project. *Chapter 2* describes the algorithm/model and use of it for predicting the states in which attacks occur. *Chapter 3* talks about the results of the software's predictions. Finally, Chapter 4 is about the project conclusions, future work and what I would have done differently.

Chapter 2

Algorithm/Model Detail

Firstly, for the US shooting information, I downloaded a dataset. The dataset was downloaded in CSV format from www.kaggle.com. CSV files are very convenient in *Python*. In the program, using *Pandas*, I imported the data into a dataframe.

I then split the dataframe into two new dataframes. I did this for using with *Scikit-Learn*. I split it into the training dataframe and the testing dataframe. Before performing the algorithm, the datasets had to be processed. Firstly, I removed the columns that weren't relevant. These included the data sources. The dataframes' columns that were in string format then had to be converted to numerical. I did this as *Scikit-Learn* requires numerical values.

After the dataframes were encoded the final processing could be done. From the training dataframe, the incident IDs were dropped as they weren't needed for the algorithm. The state column was removed from the test dataframe as it was column that was to be predicted. Lastly the incident IDs were removed from the testing dataframe and saved as a *Pandas* series object.

Finally, the algorithm could then be run on the dataframes. The algorithm used was the nearest neighbour algorithm (known as *KNeighborsClassifier* in *Scikit-Learn*). The final dataframe was then created using the series containing the incident IDs and the results gotten from the algorithm. The results were then printed on a scatter graph as can be viewed in the next chapter. Lastly the results were printed to a CSV file.

Chapter 3

Empirical Evaluation

After the dataframes were processed and algorithms run, I used a scatter graph to show the results. The scatter graph's axes consisted of each incident's ID and the US state where the incident occurred.

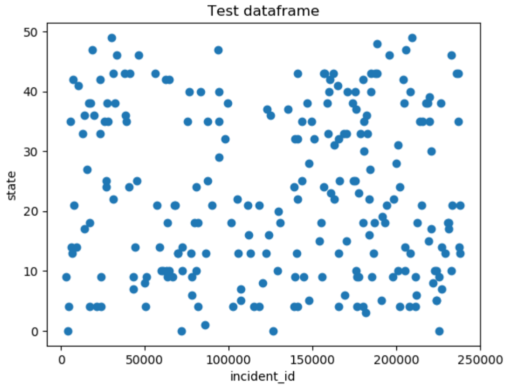

Figure 1: Scatter graph of test dataframe before being processed

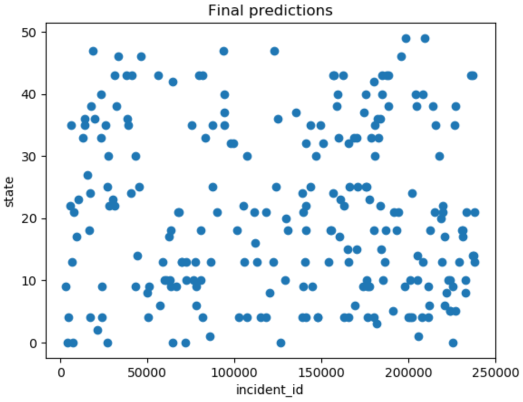

Figure 2: Scatter graph of final results

Before processing I also printed the test dataframe for comparing afterwards. After analysing the graph of the final predictions, the similarity of it and and the test dataframe's graph showed the accuracy of the algorithm. The algorithm I used was the nearest neighbour algorithm.

The most accurate of the predictions were for the states with the highest number of incidents. To find outliers in the data, I used boxplots.

Chapter 4

Conclusion and Future Work

In conclusion I feel that the use of this software could potentially help learn to protect better against possible shootings.

As with any project there is an endless amount of changes that could be made to the software. Firstly, I would try to make the software more accurate. I would do this by further processing the data. I would also like in future, for the algorithm to take into account, each state's bordering states. I would do this as researching, I found that bordering states having weaker gun control laws sometimes influenced the number of shootings in a state. Even if the state in question had stricter laws on guns.