**Denver Crime**

Radjamin Hukom, rhukom@bellarmine.edu

**ABSTRACT**

This paper seeks to understand the factors surrounding crime in Denver that is available to be analyzed in the Denver Crime data. Through this investigation, the paper seeks to be able to understand whether certain factors correlate with crime or not.

1. **INTRODUCTION**

Using the database of Denver crime acquired at kaggle, I want to investigate what are the different factors that affect the type of crime that is being done in Denver through logistic regression machine learning. Some of the factors that I would be looking into would be the location of the crime, the crime\_id, offense\_id, time, among others.

The dependent variable would mainly be the type of crime that was perpetrated, meaning, it would be a logistic regression as it is a discrete variable. Other possible outputs would be the district or precinct id showing whether certain types of crimes or other factors affect where the crime would be perpetrated.

Some interesting parts of this project would be (hopefully) a map that shows where the crime was perpetrated, a timetable showing where crime often occurred, and when it is often reported among other things. This is a big dataset.

1. **BACKGROUND**

The data set documents criminal offenses in the city and county of Denver. It is a data set that is continually renewed. However, the data set is not whole as certain information is omitted due to legal requirements. The data is not without errors as it relies on data that cannot always be verified. The data set omitted sexual assault, child abuse, juvenile perpetrators, victims, and witnesses of certain crimes due to legal restrictions.

It is not stated explicitly in the data card why this information is collected other than for documentation purposes. It is also not clear what potentially useful information in the data set that is being sought by the city of Denver. However, by looking at the data set, a clear directive to investigate which factors correlating with (or even causing) crime seems obvious.

1. **EXPLORATORY ANALYSIS**

This data set contains 386865 rows and 20 columns with various data types.

**Table 1: Data Types**

|  |  |  |
| --- | --- | --- |
| *Variable Name* | *Data Type* | *Additional Notes* |
| Incident\_id | Float64 | Not usable in logistic regression due to large numbers |
| Offense\_id | Float64 | Not usable in logistic regression due to large numbers |
| Offense\_code | Int64 | Not usable in logistic regression due to large numbers |
| Offense\_code\_extension | Int64 |  |
| Offense\_type\_id | Object (string) | A further elaboration on offense\_category\_id |
| Offense\_category\_id | Object (string) |  |
| First\_occurrence\_date | Object (string 🡪 datetime64[ns]) |  |
| Last\_occurrence\_date | Object (string 🡪 datetime64[ns]) | 175556 null values |
| Reported\_date | Object (string 🡪 datetime64[ns]) |  |
| Incident\_address | Object (string) |  |
| Geo\_x | Float64 | 15503 null values |
| Geo\_y | Float64 | 15503 null values |
| Geo\_lon | Float64 | 15769 null values |
| Geo\_lat | Float64 | 15769 null values |
| District\_id | Object (string) | 57 null values |
| Precinct\_id | Int64 |  |
| Neighborhood\_id | Object (string) | 689 null values |
| Is\_crime | Int64 (binary) |  |
| Is\_traffic | Int64 (binary) |  |
| Victim\_count | Int64 (binary) |  |

1. **METHODS**
   1. *Data Preparation*

First, due to programming language issues, I needed additional code to enable the program to be able to read the dataset. Then, to enable to make the visualizations that I later needed, I converted the date columns into datetime64[ns] from string to make the AM and PM also show up.

Then I dropped ‘incident\_address’, ‘is\_crime’, ‘is\_traffic’, ‘victim\_count’. I wanted to drop as few columns aas possible in the beginning. I dropped the first one because I saw that there were too many variations in that column. The last three I dropped due to is\_crime being all 1s, is\_traffic being all 0s, and victim\_count being all 1s.

To eliminate the null values, I started with district\_id. I noticed that almost all was using integer values except U which I presumed to denote unknown which I converted to 0 to be able to convert district\_id to int from string. I decided to drop neighborhood\_id due to there being 78 different neighborhoods. Then, I separate first\_occurrence\_date into fid, fid\_time, and fid\_part, which denotes the date, the time, and the AM or PM of the crime. Then, I did the same for last\_occurrence\_date (lod), and reported\_date (rd). After rearranging the columns, I decided to drop last\_occurrence\_date as I don't want to remove that big of a chunk of my data and I also don't want the hassle to clean it, there's no good value to put in for empty values as there's not enough knowledge available for me to know what the empty values meant. Tried researching it, can't come up with a good answer. Wanted to put current date time, but then need a function to continually update to current time, which doesn’t really makes sense.

I decided to drop precinct\_id to there being too many unique inputs in that column.

Now, I’m left with geo locator columns. I tried to make a program to estimate the geo locators using the district\_id or precinct\_id. after the upteenth time of attempting to fix this 'bug' where it clearly says that the boolean function outputs True so it should do what I asked it to do, but it clearly didn't when I checked so I decided to dropna.

Then I made fid\_day and rd\_day, new columns that tell the day of the week the crime took place. For logistic regression, I converted offense\_category\_id into offense\_category\_num which is the integer version of the same information.

* 1. *Experimental Design*

x = ds[['incident\_id', 'offense\_id', 'offense\_code\_extension', 'geo\_x', 'geo\_y',

'geo\_lon', 'geo\_lat', 'district\_id']]

y = ds[[‘offense\_category\_num’]]

Table 2: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All 8 raw features with 66/16.5/16.5 split for train, validate, and test |
| 2 | All 8 raw features with 80/10/10 split for train, validate, and test |
| 3 | All 8 raw features with 50/25/25 split for train, validate, and test |

x= ds[['incident\_id', 'offense\_id', 'offense\_code\_extension','offense\_category\_num', 'geo\_x', 'geo\_y',

'geo\_lon', 'geo\_lat']]

y= ds[['district\_id']]

Table 3: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 4 | All 8 raw features with 66/16.5/16.5 split for train, validate, and test |
| 5 | All 8 raw features with 80/10/10 split for train, validate, and test |
| 6 | All 8 raw features with 50/25/25 split for train, validate, and test |

1. *Tools Used*

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Asus computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1, datetime, and warnings. Provide a brief explanation of why you chose these tools.

All the ones listed above were tools that was used in class and were ones that I was familiar with in usage due to extensive tutelage under Dr. Sarkar. The datetime and warnings were used for convenience and to speed up the process as converting certain columns (like the date ones) into datetime format would make it easier for data manipulation and analysis while the warnings allow for me to tell the computer to forgo printing the warnings, meaning, faster and easier to read outputs.

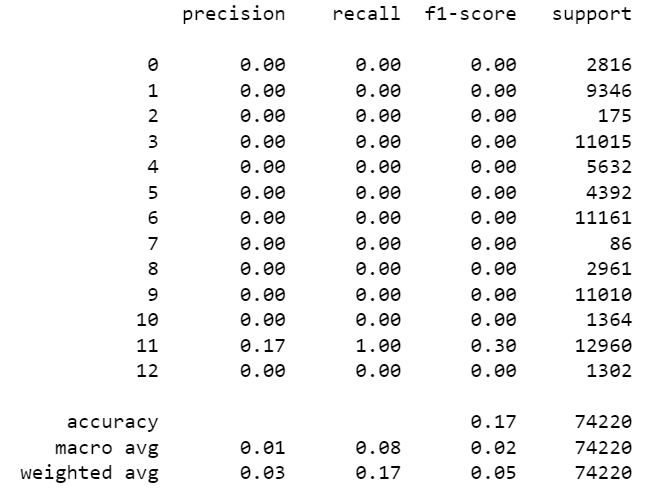
1. **RESULTS**
   1. *Classification Measures/ Accuracy measure*

**Figure 1: Experiment 1**

A screenshot of a table

Description automatically generated

**Figure 2: Experiment 2**



**Figure 3: Experiment 3**

A table of numbers with text

Description automatically generated

**Figure 4: Experiment 4**

A screenshot of a table

Description automatically generated

**Figure 5: Experiment 5**

A table of numbers and letters

Description automatically generated

**Figure 6: Experiment 6**

A screenshot of a computer screen

Description automatically generated

* 1. *Discussion of Results*

My models that were trained to output the district\_id seems to be more accurate with accuracy score of .22 compared to the .17 for the ones meant to output the offense\_category\_id. It might be due to the number of variables that was available to support each possible output being so lopsided, more lopsided for the offense\_category\_id than for the district\_id.

* 1. *Problems Encountered*

The project went surprisingly well even though I spent a disproportionate amount of time on the preprocessing. However, I found that I learned a lot about data preprocessing through the process, so I am not too sad about that. However, what bugged me the most is the bug that I encountered that ultimately made me decide to dropna. Another problem that I encountered is my models being so low in accuracy. After further discussion with my instructor, Dr. Sarkar, in the future I should modify my preprocessing process in light of my goal of logistic regression ML. In light of it, I should have focused more on a smaller amount of columns, a smaller amount of possible outputs (say top 10 if they have more than 15) for each column, and then using dummyvariable, made all the columns into binary columns as instructed in class instead of what I did, which is more for visualization.

* 1. *Limitations of Implementation*

My models SHOULD NOT be used to predict the target data due to its low accuracy (below .5). However, the model that was discussed in the previous section might possibly be better. However, due to the lopsided amount of data supporting each variable (even in the top 10), I have reason to believe that it might just increase the accuracy to around .5, which is usable but not advised, certainly more usable than the ones I produced.

* 1. *Improvements/Future Work*

As previously discussed, in the future I should modify my preprocessing process in light of my goal of logistic regression ML. In light of it, I should have focused more on a smaller amount of columns, a smaller amount of possible outputs (say top 10 if they have more than 15) for each column, and then using dummyvariable, made all the columns into binary columns as instructed in class instead of what I did, which is more for visualization.

Also, I would have liked more time devoted to the creation of an interactive map to be able to show if crime is more frequent in certain areas to make use of the geo locators. Sadly, I had to abandon that idea due to time constraints. I have found the answers, but not the time to implement them.

1. **CONCLUSION**

In conclusion, my models are bad and needs more work. My visualizations were good as I was able to glean useful information from them. However, ultimately my goals were not visualization but logistic regression, which means I failed this time. Additionally, in the visualization area, I failed to section off more time for map creation. There were a lot of good and new things that I learned through this project, but I ultimately failed to meet my personal objectives.