SDS Exam assignment M3 - 29/11-2019
Anders Kronborg, Aleksander Ommundsen, Mikkel Hartvigsen & Thomas Boye

# Introduction

Building upon the realization from our previous assignment regarding how big a factor race was in homicides, we wanted to use deep learning to calculate a profile on the perpetrator based on multiple variables of the victim and the perpetrator. Hereby the scope of this assignment is to illustrate the possibilities of neural networks and how this differs from machine learning algorithms and their limitations. This report and the neural network portrayed showcases a social data science issue as this illustrates that complex social issues such as murders and crime solving indeed can benefit from a neural network approach to gain unique insight in social phenomena.

# Selection of dataset

We chose to reuse our dataset from M1 which was obtained through Kaggle.com and contains over half a million different cases of manslaughters/homicides in the United States of America. The dataset covers every registered murder in the US from 1980 to 2014, spanning 34 years.

# Cleaning of data

As with the dataset, the cleaning was primarily done in the same manner as in the previous assignment.

This time we also fixed some of our earlier mistakes. In the previous assignment we assigned converted the 'Weapon' variable to integers, which could be interpreted as a score where a weapon with the value '1' was inferior to a weapon with the value of '7'. In this assignment it was was accounted for by by using a neural network as a solution to our problem. This is due to the fact that our machine is not required to only have integers as value but actually thrives on strings, for the approach that we have taken. However with the "age" variable we have indexed the different ages of the perpetrator to make it easier for the deep learning model to make a qualified guess. These indexes were set to:

- Age 6-12 = "kid"
- Age 13-17 = "teenager"
- Age 18-27 = "young adult"

- Age 28- 38 = “adult”
- Age 39-59 = “old adult
- Age 60-95 = “elderly”

We also realised that there were unknown values in the columns “weapon” and “relationship” that got removed compared to our earlier assignment.

Furthermore we realised that the model naturally would be better at finding white and black perpetrators compared to asian and native perpetrators, simply because the first two variables vastly outnumbered the latter. We fixed that problem by taking only 2500 perpetrators of each race (a little less than the amount of native perpetrators) to make sure that the new model wouldn’t get a racial bias. Another approach to this could have been to use the .get_dummies function in order to add dummy-entries to our dataframe and thereby ensuring that there were approximately the same amount of white, black, native American and Asain/Pacific perpetrators. However this was deemed unnecessary due to the size of the dataset.

Next, we removed the same unnecessary information as in the previous assignment. These were: Agency Code, Agency Name, Incident, Victim Ethnicity, Perpetrator Ethnicity and Record Source. Ethnicity would normally be fine to include, but since this only ranges from hispanic/non-hispanic it was deemed irrelevant. We also set the perpetrator age to be at least 6 years of age, based on the lowest convictable age in the USA, and a maximum age of 95 years, because we were afraid that perpetrators with an age set to 99 years, was just given a maximum value in the police system.

Furthermore we created a heatmap, as in the last assignment, to visualize the correlation between the different variables. This heatmap is nearly identical to the previous one, with the exception of missing clusters, as we did not perform a cluster analysis.

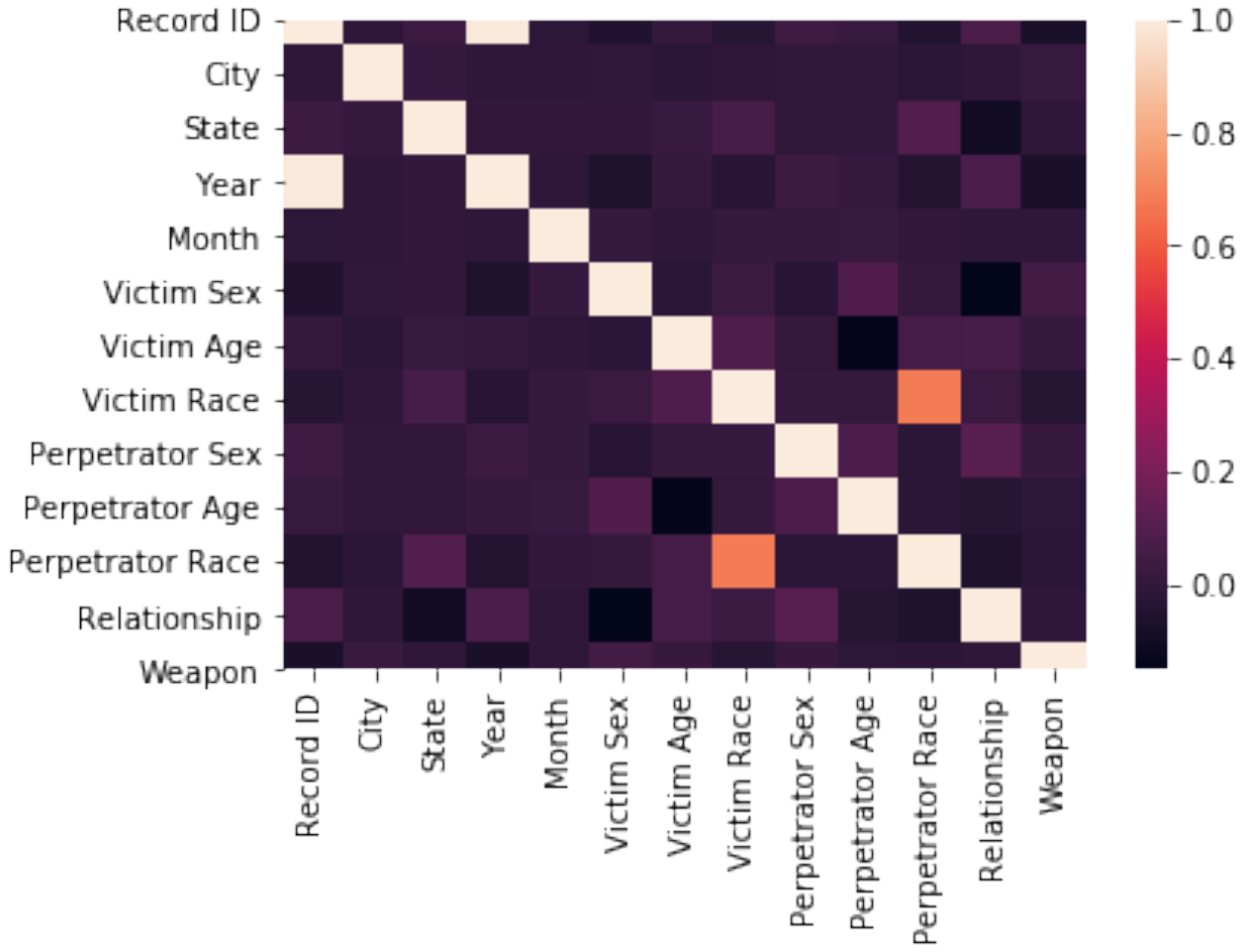


Figure 1. The figure illustrates the heatmap of the selected variables

In the heatmap we can see a clear correlation between Perpetrator Race and Victim Race. This is also the only correlation that we can see. We can also see that there is close to no correlation between Victim Sex and Relationship.

## Previous results and building upon that

In the previous assignment, we concluded that our hypothesis regarding race as a large factor was true. By using deep learning we were given more tools to explore and refine our machine to make more accurate predictions. The great aspect of deep learning compared to machine learning is, that it is now possible to make a full list of predictions on the perpetrator, instead of only one at the time.

From there, we wanted to create a deep learning algorithm that could output several characteristics about a perpetrator, from several variables regarding the crime.

Another interesting thing that we did was creating a generic police report out of the information giving in the dataset. The police report simply gives context to the information by always representing it in the same way and therefore increasing the overview. Stuff like this is pretty easy to do, but could actually be used in real-life scenarios in the police force, for example be letting a deep learning model write the police reports and giving the police more time to solve crimes.

## Trial and error

Everything wasn't flowers and sunshine during the creation of this report and the underlying Google Colab work. In fact there was a lot of work that didn't make it to the final report and in this part of the assignment we inspect why much of the coding didn't work, and what we did to fix it.

One of the things that we had trouble with was separating the different informations of the perpetrator in words instead of characters. This resulted in the predictions being very unreadable, as shown below:

| | prediction | real |
|---|---|---|
| 6743 | [ , ,, /, A, D, F, G, H, I, K, M, N, P, Y, a, c, d, e, f, g, i, k, l, m, n, o, p, r, s, t, u, v, w] | [ , ,, /, A, F, N, Y, a, c, d, e, g, i, k, l, m, n, o, q, r, s, t, u, v] |

Figure 2. The figure illustrates the results with character separation

Our model would simply predict on the upcoming characters instead of the words. We solved that by running a "x.split( , )", forcing the code to only separate by commas instead of individual letters.

Another troubling thing was that the model seemed to only make "safe guesses" and avoiding predicting on things that it were unsure about like the perpetrators age. Instead it would often guess "acquaintance" multiple times, in terms of what relationship the perpetrator had to the victim, which obviously wasn't ideal. This was due to the fact, that the machine did not differ between the different things found in the Perp_Information column. This meant that it prioritized labels, that occurred more often, and also the ones most likely to have a

match. Amongst other things, this led to the model being predisposed to predicting that the Relationship was equal to Acquaintance almost all the time, since Acquaintance was the safer bet. It also had a tendency to not predict on Perpetrator Age at any point due to the same reason. In other words, it prioritizes features with the least unique values. We sort of fixed the problem by indexing the ages as mentioned earlier. It was now possible for the model to make predictions on the perpetrators age, but the same problem still occured where it wouldn't predict on every parameter available. If the deep learning model would have predicted on both numbers and characters at the same time, it would have been necessary to have made a multi branch structure, but that was something out of our expertise.

# Results:

We can conclude that by using deep learning, we can get a very accurate result regarding the profile of the perpetrator. With machine learning, we had to rely on only one variable, whereas deep learning makes it possible to include multiple variables, to get a more accurate and consistent result. With a high accuracy, our deep learning algorithm was able to identify the profile of the perpetrator. In a lot of cases information regarding the victim is known so by using this information we were able to correctly predict the information regarding the perpetrator.

In the figure below (figure 3), we can see that our model was quite good at predicting all parameters on the perpetrator. But it excelled in some variables. It got very good results with the race but with the relationship it got mixed results. This corresponds with the heatmap we showed earlier where the correlation between victim race and perpetrator race was very strong.

| | prediction | real | victim_Information |
|---|---|---|---|
| 4456 | [ Acquaintance, Male, Young Adult, Black] | [ Acquaintance, Adult, Male, Black] | March, 1997 In the city of Arlington, Virginia we made a disturbing discovery as we found a dead Black, Male, who seemed to be no older than 44 years of age, we suspect, that the victim was killed with a Blunt Object |
| 1106 | [ Elderly, Male, Old Adult, Wife, White] | [ Acquaintance, Male, Young Adult, White] | June, 1983 In the city of Laurel, Kentucky we made a disturbing discovery as we found a dead White, Female, who seemed to be no older than 79 years of age, we suspect, that the victim was killed with a Handgun |
| 2666 | [ Acquaintance, Male, Stranger, Young Adult, Black] | [ Acquaintance, Male, Teenager, Black] | April, 1996 In the city of Nash, North Carolina we made a disturbing discovery as we found a dead Black, Male, who seemed to be no older than 23 years of age, we suspect, that the victim was killed with a Handgun |
| 9838 | [ Male, Young Adult, Asian/Pacific Islander] | [ Male, Stranger, Teenager, Asian/Pacific Islander] | November, 2005 In the city of Sacramento, California we made a disturbing discovery as we found a dead Asian/Pacific Islander, Male, who seemed to be no older than 13 years of age, we suspect, that the victim was killed with a Handgun |
| 105 | [ Acquaintance, Male, Stranger, Young Adult, White] | [ Friend, Male, Young Adult, White] | May, 2006 In the city of Bristol, Massachusetts we made a disturbing discovery as we found a dead White, Male, who seemed to be no older than 18 years of age, we suspect, that the victim was killed with a Knife |

Figure 3. The results from our model.

We tried to apply our trained model to predict on unsolved crimes but the problem was, that it wasn't possible to verify the results.

Additionally, there could be multiple other variables that could have been included to optimize our machine even more. So other future work could be to look at different other variables regarding crime statistics. This could for example be statistics on crime in each state or city.

If done in a elegant manner, this work could *maybe* be applicable for police investigation.