https://doi.org/10.48009/2_iis_2020_150-157

A COMPUTATIONAL ANALYSIS OF THE IMPACT OF POPULATION AND NUMBER OF POLICE OFFICERS ON ASSAULT AGAINST LAW ENFORCEMENT OFFICERS

Rudra Panda, University of St. Thomas, rudrapanda@yahoo.com
Priya Sivaprakasam, University of St. Thomas, priya.sivaprakasam@stthomas.edu
Manjeet Rege, University of St. Thomas, rege@stthomas.edu

ABSTRACT

In the United States, law enforcement professionals are being killed feloniously, accidentally and assaulted in the line of duty. The algorithm proposed in this paper will help to predict the assault count based on the population size of a geographic area and number of police personnel. This paper focuses on analyzing the trends in the numbers of assaults using Jupyter Notebook and SASPy in SAS® University Edition and SAS® Enterprise MinerTM. A variety of methods will be used to determine the best model such as Gradient Boosting, Decision Tree, Linear Regression and Neural Network.

Keywords: Gradient Boosting, Decision Tree, Linear Regression, Neural Network, Jupyter Notebook, SASPy, SAS® University Edition, SAS® Enterprise MinerTM, LEOKA

INTRODUCTION

The Uniform Crime Reporting program's primary objective is to generate reliable information for use in law enforcement administration, operation and management; over the years, however the data have become one of the country's leading social indicators. The program has been the starting place for law enforcement executives, students of criminal justice, researchers, members of the media and the public at large seeking information on crime in the nation.

The Uniform Crime Reporting (2018) program defines law enforcement officers as individuals who ordinarily carry a firearm and a badge, have full arrest powers and are paid from governmental funds set aside specifically for sworn law enforcement representatives. Each year, the law enforcement agencies across the United States report to the UCR program the total number of sworn law enforcement officers and civilians in their agencies as of October 31. Civilian employees include full-time agency personnel such as clerks, radio dispatchers, meter attendants, stenographers, jailers, correctional officers and mechanics. Here we can't find the line-of-duty deaths of police officers.

The Law Enforcement Officers Killed and Assaulted (2020) Program offers information to law enforcement agencies about officers who were feloniously or accidentally killed or assaulted while performing their duties. The data collected is analyzed and the results are incorporated into the officer safety awareness training the FBI provides for partner agencies. The number of police officers being assaulted in line-of-duty is quite alarming. Reportedly, 106 police officers were killed in 2018 and 90 in 2019 in United States of America.

The study focuses on both the police employee data and LEOKA's assault data to identify if the increase of police personnel affects the number of assaults.

DATA PREPARATION AND CLEANING

The data is sourced from Crime Data Explorer. The Crime Data Explorer (n.d.) is part of the FBI's broader effort to modernize the reporting of national crime data at the national, state, and agency levels. It provides option to download bulk data and related documents. The assaults on Law Enforcement Officers and Police Employee dataset were chosen for this purpose. The assaults on Law Enforcement Officers contained assault counts while the Police Employee Data contained information about number of police officers and population.

The assault dataset had 288,401 observations and 31 variables. Figure 1 shows the correlation between variables of assault dataset. The police employee dataset had 1,439,467 observations and 21 variables. Figure 2 shows the correlation between variables of police employee dataset.

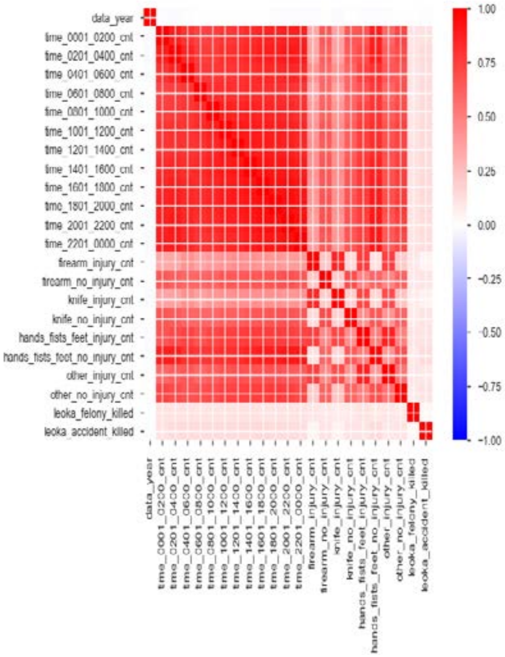


Figure 1. Correlation Matrix For Assault Dataset

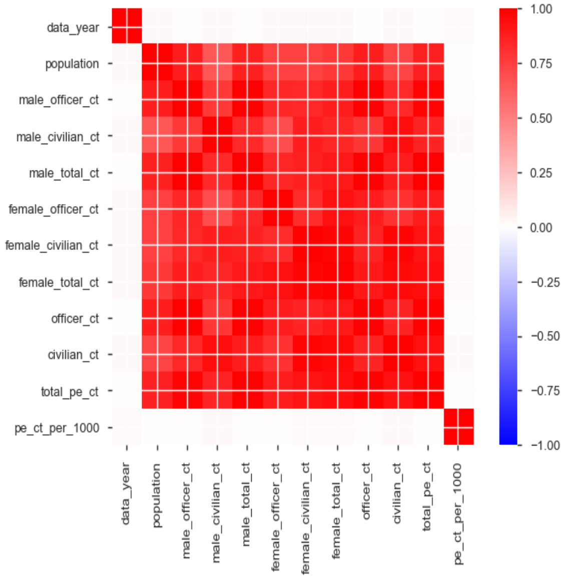


Figure 2. Correlation Matrix For Police Employee Dataset

Data Cleaning

The first step was to clean up the duplicate observations from police employee dataset. There were 19,085 (1.3%) duplicate observations. The assault dataset had no duplicate records. The assault dataset had records from the year

1995 to 2018 while police employee dataset had records from the year 1960 to 2018. The data in police employee dataset was cleaned up to retain data from the year 1995 to 2018 as to match with assault data.

The data analysis was performed to identify columns from both datasets and merged records that matched from assault dataset and employee dataset. During the analysis, the columns were identified and those that were not relevant for the problem were dropped. Here is the list of columns that were dropped from datasets.

- **Assault** 'pub_agency_unit','division_name','region_name','agency_type_name'

- **Employee** 'pub_agency_unit','pe_ct_per_1000','ori','division_name','region_name','agency_type_name'

In assault dataset, the number of assaults were given for every 2 hours interval over 24 hours period. The number of assaults for each interval wad summed to calculate total number of assaults for the observation. If the total number of assaults was zero, it was treated as missing value and those records were pruned from the dataset. This was done because assault count was the target variable and the objective were to identify factors that contribute toward assaults.

In employee dataset, if population and officer count (total_pe_ct) column values were zero, those were treated as missing values. This is because zero population or zero officer count data were treated as data accuracy issues and those records were pruned from dataset. Post data cleansing and enrichment, the assault dataset was left with 106,550 observations and 31 columns, and employee dataset was left with 344,967 and 15 columns.

The following columns identify records uniquely in both datasets:
'data_year','state_abbr,'population_group_desc', 'county_name', 'pub_agency_name'.

The combination of columns identified 105,786 over 106,550 (99.3%) records uniquely in assault dataset, and 344,957 over 344,967 (99.99%) records uniquely in employee dataset. The duplicate records (less than 1%) were replaced in both datasets by taking the average of the columns.

After cleansing both datasets they were merged using the combination of columns that identified records uniquely in both datasets. The final dataset had 93,846 number of observations and 41 columns. Though our columns of interest were assault count (target), population and number of officers (inputs), the other columns were retained to study their effect on target variable during model building process. Figure 3 shows that total_pe_ct and officer_ct is highly correlated.

To understand the relationship between the number of assaults and population size of a geographic area the pair plot was plotted. Figure 4 depicts the relationship between assault count (target), population (input), and total number of officers (input). It clearly shows that all three variables are heavily right skewed.

The assumption was that increasing number of police officers could reduce the number of assaults n. Figure 5 depicts the relationship between all three variables. Figure 5 shows some linear relationship between population and total number of police officers. However, there is no linear relationship between assault count and population or total number of police officers.

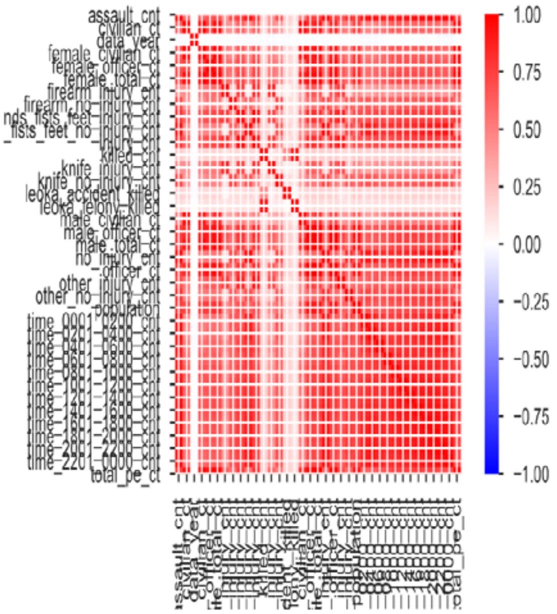


Figure 3. Correlation Matrix Post Data Cleaning

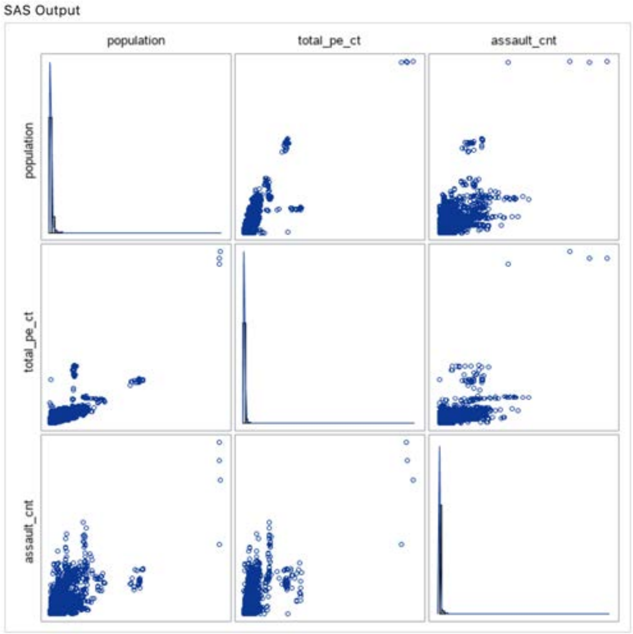


Figure 4. Relationship Between Assault Count, Population and Number Of Officers

To perform the cleansing and merging of both assault and police employee dataset, *SAS® University Edition* was used which supports SASPy and Python Jupyter Lab (2016). The merged dataset which shows some nonlinear relationship and skewed data was run through models where regression techniques were applied in SAS® Enterprise MinerTM to determine and choose the relevant model after comparing various models.

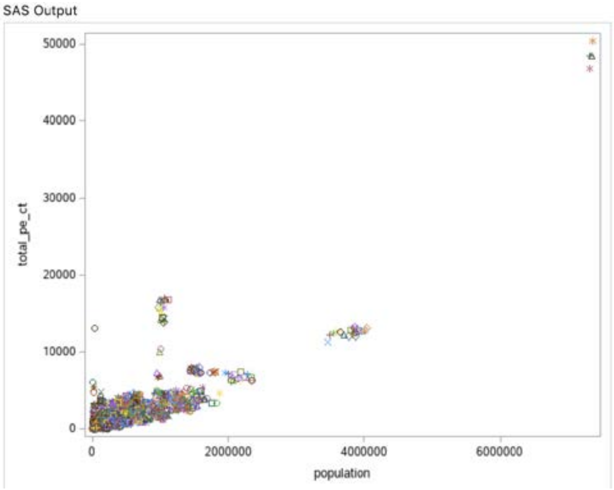


Figure 5. Relationship Between Population And Number Of Officers

MODEL BUILDING

The merged dataset was extracted to csv format in SASPy. The data file was then imported into SAS workspace and data source ASSAULT_TRAIN was created by using *SAS® University Edition.*

The data source ASSAULT_TRAIN was selected for the INPUT DATA node and the RMSE(2013) was chosen as selection criteria for the model. The RMSE is directly interpretable in terms of measurement units and so is a better selection criterion.

The model was built, and various regression techniques were applied to predict police assaults. In order to predict the response, four models were built using the Neural networks (2007), Linear regression (n.d.)., Decision tree (2016) and Gradient Boosting. Data was divided into 70% training and 30% validation. SAS® Enterprise MinerTM software has Model comparison node that selects the best model based on the value of a single statistic. For this analysis, Mean Squared Error was used as selection criteria.

Mean Squared Error is one of the common methods used to measure the performance of a linear model. For each point, it calculates square difference between prediction and target and then average those values. The higher this value, the worse the model is. Figure 6 shows the several models that were built using SAS® Enterprise Miner 15.1.

From the results of the Regression model, it was found that the variable that is significant in predicting is population. Figure 7 shows the results from Linear Regression.

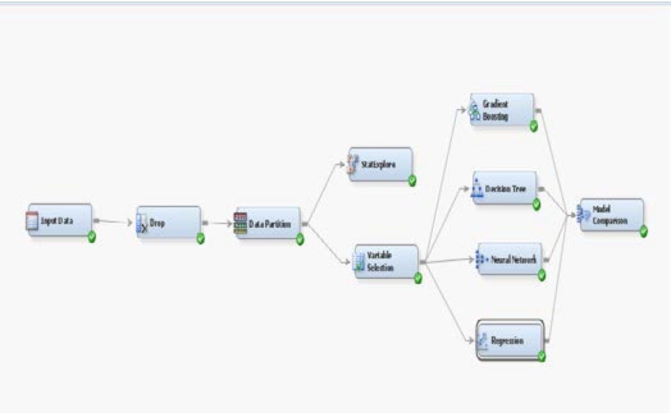
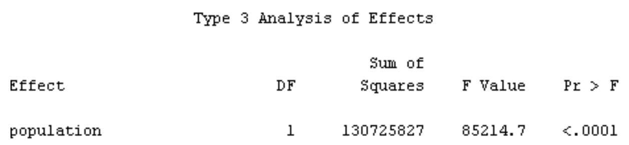


Figure 6. SAS EM Process Flow



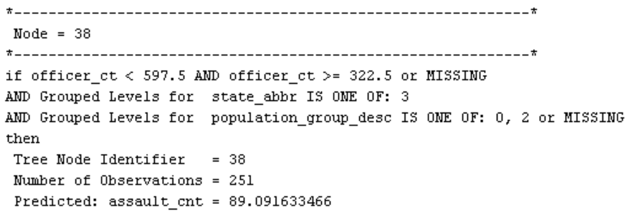
```
                     Type 3 Analysis of Effects

                                       Sum of
Effect                      DF        Squares    F Value    Pr > F

population                   1      130725827    85214.7    <.0001
```

Figure 7. Results From Linear Regression

From the Decision Tree model, the assault_cnt predicted can be identified based on certain conditions. Figure 8 shows the partial output from the decision tree node. The node rules help in explaining that assault_cnt predicted is 89% for the conditions based on officer_ct, state_abbr and population_group_desc.



```
*------------------------------------------------------------*
 Node = 38
*------------------------------------------------------------*
if officer_ct < 597.5 AND officer_ct >= 322.5 or MISSING
AND Grouped Levels for  state_abbr IS ONE OF: 3
AND Grouped Levels for  population_group_desc IS ONE OF: 0, 2 or MISSING
then
 Tree Node Identifier   = 38
 Number of Observations = 251
 Predicted: assault_cnt = 89.091633466
```

Figure 8. Decision Tree Node Rules

The Neural Network generated the best prediction based on the selection criterion RMSE. The model with the smallest RMSE value is chosen and Neural Network was chosen based on the criterion.

CONCLUSION

The predictive models built using the available data could estimate the number of assaults with a root mean squared error (MSE) as selection criterion. Though this larger error could work for areas with a large predicted assault count, it will not work for areas with small assault count. In general, the current model could not be used for estimating assault count with confidence. The large error is a result of skewness in the assault count and population data. Hence, with current study, no conclusions can be drawn between assault count and the population of an area or the number of police officers employed in that area. Figure 9, 10, 11 and 12 shows the model score distribution based on the Mean Predicted and Mean Target for assault_cnt. The red color in the figure indicates the Mean Target and the blue color indicates the Mean Predicted.

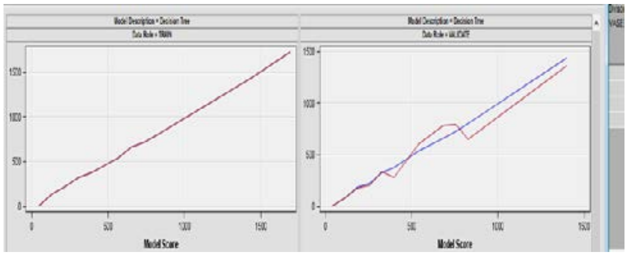


Figure 9. Decision Tree

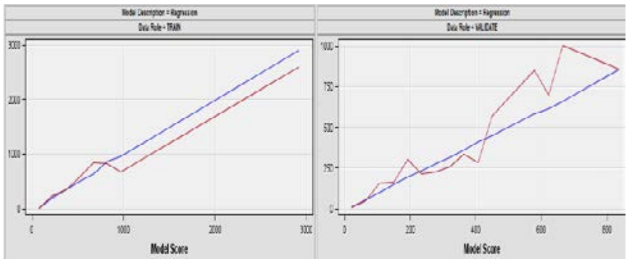


Figure 10. Linear Regression

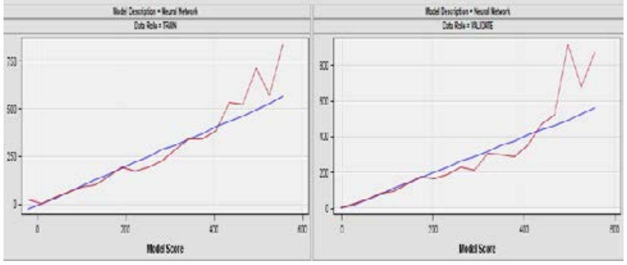


Figure 11. Neural Network

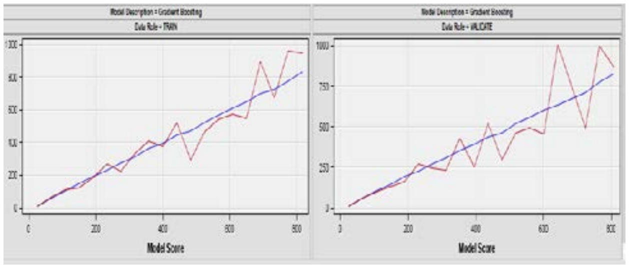


Figure 12. Gradient Boosting

Based on the RMSE, Neural network model was selected.

For future studies, we anticipate completing the research on the relationship between the number of assaults and the type of crimes in an area. The current analysis could be combined with crime (arrest data) and researched further to validate this assumption. Once the factors affecting the number of assaults incidents are identified, appropriate measures could be taken to reduce the number of assaults and ultimately reducing the number of deaths in line-of-duty. Once the factors affecting the number of assaults incidents are identified, appropriate measures could be taken to reduce the number of assaults and ultimately reducing the number of deaths in line-of-duty.

REFERENCES

Bykelly93. (2016, April 28). SAS Enterprise Miner – Decision Tree. Retrieved from https://bykelly93.wordpress.com/2016/04/28/sas-enterprise-miner-decision-tree/

Directed Data Mining: Simple Linear Regression. (n.d.). Retrieved from https://amadeus.co.uk/tips/directed-data-mining-simple-linear-regression/

Federal Bureau of Investigation. 2020. LEOKA | Federal Bureau Of Investigation. [online] Available at: <https://www.fbi.gov/services/cjis/ucr/leoka> [Accessed 25 April 2020].

Heuristicandrew, & Says:, J. (2013, July 12). Calculate RMSE and MAE in R and SAS. Retrieved from https://heuristically.wordpress.com/2013/07/12/calculate-rmse-and-mae-in-r-and-sas/

(n.d.). Retrieved from https://www.hsdl.org/?view&did=803672

Sarma, K.: Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications. SAS Press (2007)

SAS University Edition: Help Center. (2016, June 21). Retrieved from https://support.sas.com/software/products/university-edition/faq/jn_whatis.htm

“Uniform Crime Reporting (UCR) Program.” FBI, FBI, 10 Sept. 2018, www.fbi.gov/services/cjis/ucr