Final Report – Crime Finder - DavisA

1.0 Executive Summary

This study sheds light on current events by analyzing crime statistics. Historical data was analyzed and a model built to predict situations and locations in which specific offenses are likely to occur, such as justifiable homicide. The business objective of this study is to lower the number of justifiable homicides. Predictions will show the government, community, and media where changes need to be made the most. Funding can be raised in at-risk areas, increasing the sense of community and providing an added sense of protection to all. The source and application of funding is outside the scope of this project. The model objective was to create a model in which the predicted output is the likelihood that a specific criminal offense will occur. This document describes the technical methods used in building the predictive model and model evaluation results. The final model predicts the type of offense that will occur with an 84% accuracy rate, given geographic and victim demographic information. This model can be used for predictions for any location across the United States. A threshold can be set to redirect focus for funding and training. For example, if a city is predicted to have over two (2) justifiable homicides per 10,000 people, you can take action for change.

2.0 Intro/Background of the Problem

Protests throughout the country have shed light on racial profiling, excessive use of force, and unjust biases. The publicly broadcast death of George Floyd has brought questions around the world about police tactics and whether they are justified. Throughout the years, time and again, unjust incidents have brought up the question of systemic racism. There has been an increase in reported incidents of excessive use of force by police. This increase has heightened visibility,

and communities across the country are calling for change. The police have a reputation for targeting minorities, making some afraid to call the police when needed. Change is needed to deter crime while providing a sense of protection to the entire community.

3.0 Methods

3.1 Data Preparation

A lot of time was spent on data preparation, reviewing data codebooks, linking data files, and identifying appropriate category codes. Data types were converted from categorical to numeric. Derived features were created for month and day of week for date fields. The victim dataset was merged with precinct data to obtain city location for reporting precincts. This dataset included 6,034,725 victim records from 2016. Some fields had nothing but N/A values and were removed. There were 217 incidents labeled as "justifiable homicide" reported to the Uniform Crime Reporting Program (UCR) for 2016. To increase the amount of data analyzed, the study was expanded to include the 235,811 incidents labeled as "aggravated assault."

3.1.1 Outliers

Outliers were found in date fields. There were records leftover from the previous year that were not reported. These records were removed.

3.1.2 Missing Data

Missing data in numeric fields were imputed. The age field was imputed with the median since the distribution was slightly skewed. For city population, however, missing values were imputed with the mean.

There was a large disproportion of incidents reported by states. Since some states have switched over to the mandated National Incident-Based Reporting System (NIBRS), the UCR data only

includes those that remain. For this reason, only the following states were selected for the study: Tennessee, Michigan, South Carolina, Massachusetts, Ohio, and Washington.

3.2 Exploratory Data Analysis (EDA)

The following research questions were addressed using visualizations, such as histograms, boxplots, and scatterplots:

- What area in the United States has the highest crime rate?
- Are crimes reported only against certain demographics?
- What is the most common crime reported?
- In what areas are justified homicides committed most frequently? What about reports of suspicious activity?

3.4 Modeling

The modeling objective was to create a model in which the predicted output is the likelihood that a specific criminal offense will occur. For this reason, offense code was chosen as the target variable, creating a multi-class classification problem. Several different models were used and results compared to determine the best model fit for the data. The following models were assessed: Decision Tree, k Nearest Neighbor, Keras Neural Network, Random Forest Classifier, Bagging Classifier, and a Booster Regressor model.

4.0 Results

4.1 Data Analysis

1) What area in the United States has the highest crime rate?

- The top area reporting incidents in the UCR system was from the East North Central Region of the US.

2) Are crimes reported only against certain demographics?

- More crimes reported were committed against whites, almost 30% more than other races reported. This is half of what is represented in the US population.
- Slightly more crimes were committed against females, although the gender distribution is close.

3) What is the most common crime reported?

- The top six (6) offenses reported in 2016 to the UCR system were: Destruction/Damage/Vandalism of Property, Simple Assault, All Other Larceny, Drug/Narcotic Violations, Burglary/Breaking and Entering, and Theft From Motor Vehicle.

4) In what areas are justified homicides committed most frequently? What about reports of suspicious activity?

A picture is worth a thousand words. See Figure 1 for justifiable homicide locations and Figure 2 for locations of suspicious activity reporting. Note that no justifiable homicides were reported in Ohio or Massachusetts in 2016.

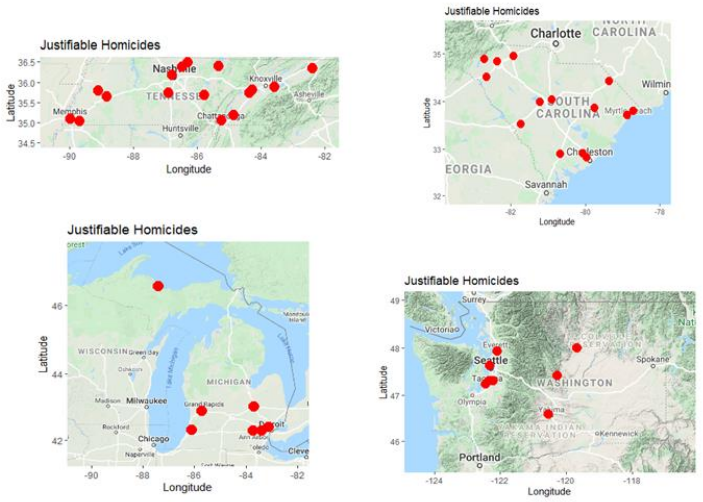

Figure 1: Justifiable Homicide Locations

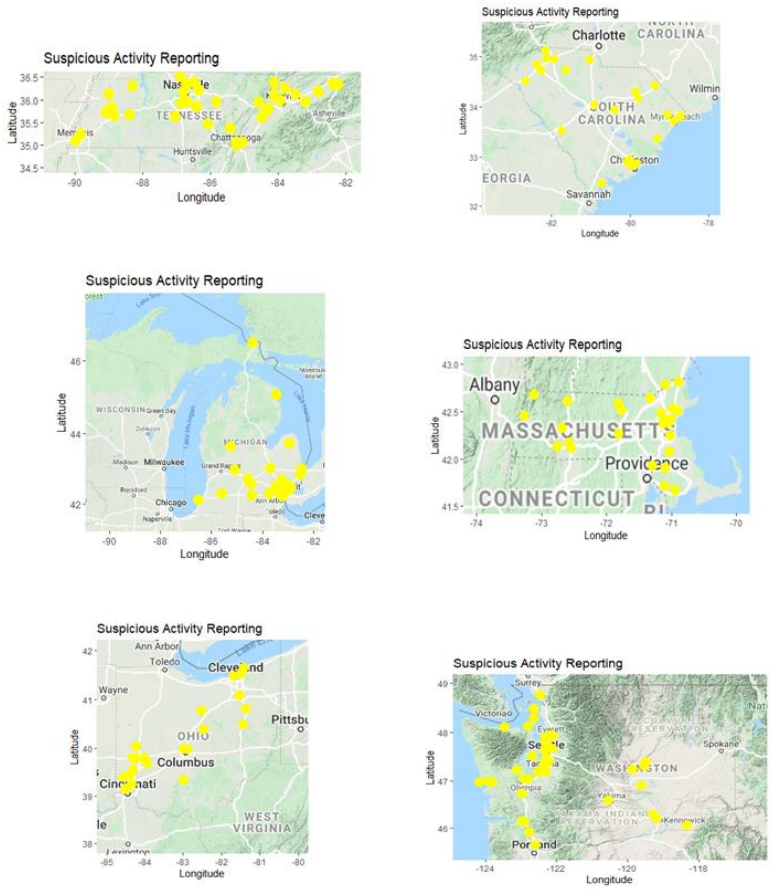

Figure 2: Locations of Suspicious Activity Reporting

4.2 Feature selection

After analysis of the codebook, some fields were not needed and were removed. Identification numbers were dropped since they were constants. The city population distribution showed similar distribution to the last census population recorded, so it was decided to use only the current city population data. The relationship the victim has to the offender is out of scope of this project, so the related fields were removed. Each reporting district can associate up to five (5) metropolitan areas. However, for the scope of this project, the area was limited to the first one reported.

4.2 Modeling

4.2.1 Decision Tree

This was originally my top pick for modeling due to the vast number of categorical features presented. Unfortunately, the scikit-learn decision tree classifier used required that categorical data to be converted to numeric before processing, which diminished the justification for choosing this model. Originally the decision tree model produced only 27% accuracy with default hyperparameters. Although that looks pretty bad, it was the most accurate during my initial assessments. When running through the full dataset, the accuracy decreased to 26%, so I kept the restriction to the six (6) states. There was no improvement whether Gini Index or Entropy was selected as the measurement of impurity. After adding victim demographic features, I was able to improve model accuracy to 38%

4.2.2 Keras Neural Network

I built a neural network model using Keras, which is an Application Programming Interface (API) for Tensorflow. I used a softmax activation function due to the categorical distribution, categorical_crossentropy as the loss function, and added four (4) hidden layers. This initial

model only produced 13% accuracy, which was quite surprising for a neural network. I adjusted the number of epochs and batch size. Increasing the number of nodes in each hidden layer from 50 to 400 made quite the difference and increased the accuracy to 23%. Once all features were included, this went up to 31%.

4.2.3 Random Forest Classifier

When using location features only, I initially created a random forest model with 26% accuracy, surprisingly lower than the original decision tree model. I used the random grid to search for the best hyperparameters, which saved time. After optimizing hyperparameters for the number of features and maximum depth, I was able to reach to 28%, which was still only slightly higher than the original decision tree model. Although the accuracy of random forest was initially lower, I was able to improve accuracy 2% by optimizing hyperparameters. The biggest benefit from running the random forest ensemble model was the outlook to the features that most influenced the model. The top three features that had the most influence on the determination of the type of offense were the victim's age (25%), the month of the incident (17%), and the day of the week of the incident (13%).

4.2.4 k Nearest Neighbor (k-NN)

Initially, I used the K Neighbors Classifier from Python's scikit-learn package to build a k-NN model. This model was built using a value of 5 for k. All other values were left as default. This model took about 18 hours to run and only resulted in 11% accuracy. Due to the length of processing time and low accuracy, I did not pursue optimizing this model in Python. Instead, I used R's knn function. I ran the model for values of k of 3,5,10, and 15. The best accuracy was found when k was set to 3 at 84%.

4.2.5 Bagging Classifier

I tried out scikit-learn's Bagging Classifier ensemble, using default settings, which resulted in a model with 40% accuracy. Since I was less familiar with this model, and it did not come close to k-NN's accuracy, I did not pursue optimizing it further.

4.2.6 Booster Regressor

Scikit-learn's Gradient Boosting Regressor was not intuitive but allowed the number of estimators and the maximum depth to be adjusted, which was similar to the random forest model. For the regressor, I used Mean Absolute Error (MAE) as a measurement at 50%.

4.2.7 Single-Classification Decision Trees

I created single decision tree models to predict the likelihood of a justifiable homicide being committed, as well as an aggravated assault. Using confusion matrices, I assessed accuracy, as well as sensitivity and specificity. Although accuracy was high at 99.8% and 72.4% respectively, sensitivity was quite low. This was caused by the large number of records with very few positive results. The low sensitivity led me to discount these single-class models for deployment.

4.2.8 Random Forests By State

I also built random forest prediction models by state to determine if models would fit better for each one. For these models, the input dataset was restricted to the specified state. Accuracy for the state models actually decreased from the joint model, with the exception of Washington state. So, it was decided to stick with the combined dataset.

4.3 Model Evaluation

I attempted to use a confusion matrix to assess the multi-class classification models, but with 49 different codes, it was just not feasible. I opted, instead, to use a classification report, which included precision, recall, f1-score, and accuracy. k-NN far out-performed the remaining

models. Due to its fast performance and good accuracy (in R), this was the model I will use for deployment. The vast difference in the k-NN model tells me that the data was able to be grouped together with Euclidean distances but did not necessarily follow rules that would have been uncovered with decision trees. I also saw that boosting was better than bagging, which was better than the single decision tree model, as expected.

Discussion/ Conclusion

The deployment method chosen will be in the predictive modeling software, in this case R. No additional resources will be required. Data preparation is already complete. No model translation will be required. I recommend reassessing the model annually. Data from precincts reporting can be applied to other precincts for predictions. Data from previous years can be applied to the current year for predictions. I recommend setting a threshold at two (2) justifiable homicides per 10,000 people. If a city is predicted to have over two (2) justifiable homicides per 10,000 people, take action for change. As changes are made, we expect the number of justifiable homicides to go down in those areas. We can decrease the threshold to identify new "hot spots" requiring focus. Although the accuracy of most of my models were low, value was obtained by determining the importance of features towards the types of offenses committed. There were also a lot of lessons learned in feature selection and hyperparameter tuning. Per Siegel (2014), "A little prediction goes a long way."

Acknowledgements

I want to thank my friends who shared their stories while discussing the systematic racism that exists in our country and ideas of where funding can help improve communities that are most affected. I want to also thank my peers who provided constructive criticisms during each

milestone. I could not improve this project without your insights. Special thanks to my daughter for cooking dinner, while I worked on my project.

References

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst.* Indianapolis, IN: Wiley.

Albon, C. (2018). *Machine learning with Python cookbook: practical solutions from preprocessing to deep learning.* O'Reilly.

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language Aware Data Products with Machine Learning.* Sebastopol, CA: O'Reilly Media, Incorporated.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. Retrieved November 6, 2019, from http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Resource Guide: National Incident-Based Reporting System. (n.d.). Retrieved June 7, 2020, from https://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS/

Siegel, E. (2016). *Predictive analytics: The power to predict who will click, buy, lie, or die.* Hoboken, NJ: Wiley

Data Sources

Uniform Crime Reporting Program Data: National Incident-Based Reporting System, [United States], 2016; United States Federal Bureau of Investigation; Inter-university Consortium for

Political and Social Research (ICPSR), University of Michigan;

https://www.icpsr.umich.edu/icpsrweb/NACJD/NIBRS/

US Cities Database: https://simplemaps.com/data/us-cities

State Geodetic Centers: https://www.latlong.net