“Big Brother” is Watching You！

—Mining on Crime Data and Suggestions to the Police

CS699 A1 Project Report

Yuandi Tang & Tianjun Ma

I. The Data Mining Goal

The crime issue, especially violent crimes such as robbery, rape and murder, is one of the biggest problems in the United States. According to the statement of related research, “Even though the violent crime rate has been decreasing since 1990, the United States tops the ranking of countries with the most prisoners”[1]. These kinds of issues not only cause negative impacts on people’s safety and happiness in their lives, but create severe challenges to the government and police department as well. To deal with this issue is never an easy task, as the violent crime rate can be influenced by various factors, from education to income, from family to society, from population density to the police force, etc. Therefore, we want to pick several major factors that may relate to the violent crime rate, and do some analysis and prediction based on their statistical data, and figure out the most influential, or, most crucial in other word, factor(s) of the violent crime rate in the United States. This goal is meaningful, as it can help the government and the police department to figure out and work on the most important direction to continually reduce the violent crime rate, thus creating a safer environment for the people no matter whether they suffer the threat of violent crimes or not.

II. Detailed Description of the Dataset

This dataset is named as “Crimes in US Communities” with a csv file named “crimedata.csv”, found in Kaggle (https://www.kaggle.com/datasets/michaelbryantds/crimedata). The description of this dataset on the website demonstrates, “This is a dataset of 2018 US communities, demographics of each community, and their crime rates. The dataset has 146 variables where the first four columns are community/location, the middle features are demographic information about each community such as population, age, race, income, and the final columns are types of crimes and overall crime rates.” Besides these, the columns in the dataset also include information about education, employment, marriage, immigration, police, etc. In general, this dataset takes all main factors that may influence crime rate into account, and list them in the csv file along with the crime rate itself.

For our research, we only need part of it, so after data preprocessing (which will be demonstrated later), we created a new dataset which only contained eight attributes: PctUnemployed (percentage of unemployment), PctImmigRct5 (percentage of immigrants in recent 5 years), TotalPctDiv (total percentage of divorce cases), PopDens (population density), PctLess9thGrade (percentage of people whose education level is lower than 9th grade), medIncome (medium income), ViolentCrimePerPop (violent crimes per population), and levelofcrime (a new classifier created by us to classify the violent crime rates with 2 levels: low and high).

III. Brief Description of Data Mining Tool(s) We Need

To begin with, we need to complete data preprocessing. The first step will be to eliminate all NA elements in the csv file, so we read the csv file and delete all NA elements. After this step, we also create a new dataset which only contains attributes we need and generate training and testing dataset from it. We complete this part by Rstudio, as its coding grammar is simple, more specifically, in our research, the coding for splitting the dataset into training data and test data is more simple than other programming languages such as Python. Moreover, the time complexity of the execution of R codes is shorter than other programming languages.

Next, we use Weka to make further calculations and apply multiple algorithms on our dataset, and come up with our final results and conclusions. We choose Weka because it can demonstrate the data more directly: it visualizes the results but no coding or programming needed. It also has a low error rate, as it already implements all algorithms we need into its system, what we need to do is only to execute the existing algorithm in it. However, if we choose other tools such as R or Python, we have to write and execute the code by ourselves, which may contain some bugs or errors in the codes and lower the accuracy of final results.

IV. Brief Description of Classification Algorithms We Use

In selecting the classification algorithm, we have chosen five categories with simple structure but large differences in methods, they are also known as the five common classification algorithms for machine semester. They are Naive Bayes, Logistic, Sequential Minimal Optimization, Decision Tree and KNN.

As the most basic and commonly used machine learning algorithm, the input variables of Naive Bayes algorithm are independent and are not influenced by or for other variables. And it has good characteristics of high efficiency and high accuracy for large data values. Especially for our crime data where some text is missing, its low sensitivity to missing data is a great advantage, and its results are easier to interpret. However, its classification decision process has a certain error rate and is sensitive to the form of the data, which can be verified in the results.

Logistic is also one of the reasons for machine learning algorithms, and the main reason we choose it is because of its very good mathematical properties, simple and easy to understand formulas, low complexity of operation time, and also less memory consumption resources. Moreover, Logistic does not need to assume the data distribution in advance for the case that the data distribution is no single, which avoids the problems caused by the inaccurate assumption of distribution. The problem is that crime data may have some multicollinearity in the indicators of population distribution, which is a difficult problem for Logistic to solve. Moreover, since its data form is relatively simple, the effect of fitting complex data will be discounted.

Sequential minimal optimization (SMO) was proposed by John Platt in 1998 to address the complexity of the previous generation of SVM training methods and to avoid iterations. It minimizes the optimization problem by solving a large optimization problem into several small optimization problems. Our crime data is tested by F-test and linear regression of R, and it is proven to have better linear features, so it can get better classification results.

As a common algorithm of machine learning, Decision Tree algorithm simulates signal control or human decision process, which has better understandability and subsequent operability. The algorithm is efficient and can be used repeatedly if it is constructed once. The mining results of crime data are given to people with non-statistical backgrounds to apply in real life, which is a more obvious advantage. The correlation of single attributes of the crime dataset is not strong, and at the same time, the preparation of data does not need to do a lot of cleaning and deletion operations and is not sensitive to data containing missing values similar to the crime dataset. However, due to the large number of categories of the crime dataset, its error rate may also increase accordingly.

KNN is in the family of lazy algorithms, which has a long time of theoretical development, while the core operation is number + majority decision, without complex mathematical operations. For large data operations, the time and cost are more cost-effective compared to SMO. And compared with plain Bayes, it does not make assumptions and has high accuracy. The crime data with a large sample size is more suitable. However, due to the large number of features in the crime dataset, it is computationally intensive in the adjacency calculation. We selected KNN=1 to make the bias degree low.

V. Brief Description of Attribute Selection Methods We Use

During the process of data preprocessing, we make some modifications towards this dataset, but do not actually change any data inside. We firstly remove all NA in the csv file to prevent any possible inaccuracies. This step also excludes some factors as they do not have enough data for us to do further analysis. After that, based on the completion of each attribute and the common sense in our lives and society, we select six attributes that may show the most obvious and important impact on the crime rate. These six attributes are: PctUnemployed (percentage of unemployment), PctImmigRec5 (percentage of immigrants in recent five years), TotalPctDiv (total percentage of divorce), PopDens (population density), PctLess9thGrade (percentage of people whose education level is less than 9th grade), and MedIncome (medium income). We confirm that these attributes have no NA values, and they are also the factors that may show the most significant impact on violent crime rates. Therefore, we select these six attributes to do our further analysis. We use R to accomplish this part.

VI. The Set of Attributes Selected by Each Attribute Selection Method

In the selection algorithm we used “ClassifierSubsetEval” in Weka, which is evaluating a subset of attributes on the training data or a separate holding test set. The classification algorithm is used to estimate the "merit" of a set of attributes. The classification algorithm is more suitable for the choice of crime data, and for consistency, we used the same algorithms used for data classification, namely Naive Bayes, Logistic, Sequential Minimal Optimization, Decision Tree and KNN.

After classifying the training data, we conclude that unemployment, immigration and divorce rate indicators are selected after evaluation using Naive Bayes; unemployment, immigration, divorce rate, population density, education and income are selected after evaluation using KNN; immigration, divorce rate, education and income are selected after evaluation using Logistic; unemployment, immigration and divorce rate and education are selected after evaluation using SMO; and immigration, divorce rate and education are selected after evaluation using Decision Tree.

After setting and saving the reduced training data, we selected the same attributes for the test datasets and generated five pairs of test-train data. These data will be classified according to the previously selected Naive Bayes, Logistic, Sequential Minimal Optimization, Decision Tree and KNN, and produce 25 classification results for the final model selection.

VII. Detailed Description of Data Mining Procedure We Actually Follow Including All Data Preprocessing We Perform

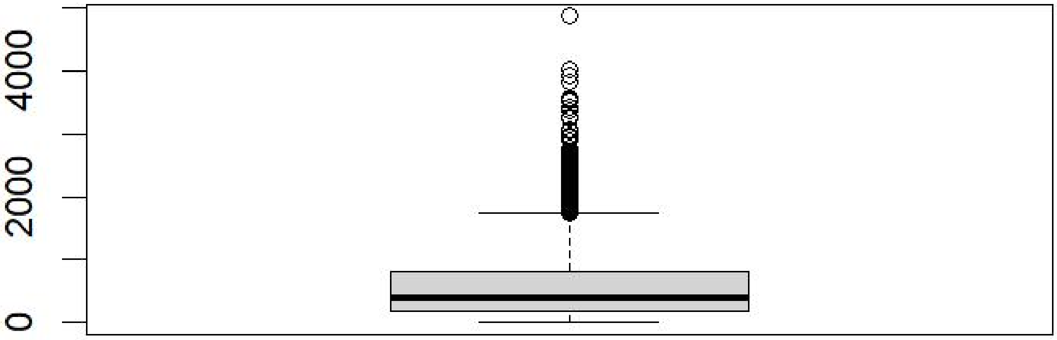
1. Dataset Cleaning and Preprocessing

Detect and remove NA values via R codes. This step excludes all NA values along with some attributes without enough data to analyze, which is also an essential part of data preprocessing. We finally select six attributes and confirm that they have no NA values. Then, we make a classifier named levelofcrime with two levels: low and high. We make the standard according to the relevant statistical result, that the violent crime rate in the United States was 395.7 cases per 100,000 of the population in 2021[2]; so, any data lower than 395.7 will be classified as low, while those higher than 395.7 will be high.

2. Dataset Validation

Skewed Data Test:

We run the screwed data test first to see if our dataset is skewed or not, by plotting a boxplot function using the *boxplot()* function in R, we can see clearly that the output of this dataset is highly skewed with many outliers. The boxplot is shown below.



Correlation Test:

We then run the correlation test to see if the attributes we pre-selected from the original dataset our dataset is correlated with the dependent variable *levelofcrime*, by using the *cor()* function in R, we can see clearly in the output that these variables are all correlated, but not strong correlated with each other. Among them, the median income is a reversed attribute with a negative correlation.

```
> cor(data$levelofcrime,data$PctUnemployed)
[1] 0.4432989
> cor(data$levelofcrime,data$PctImmigRec5)
[1] 0.1994695
```

```
> cor(data$levelofcrime,data$TotalPctDiv)
[1] 0.533617
> cor(data$levelofcrime,data$PopDens)
[1] 0.2013907
> cor(data$levelofcrime,data$PctLess9thGrade)
[1] 0.3821782
> cor(data$levelofcrime,data$medIncome)
[1] -0.412019
```

ANOVA Test:

We run the anova test to see if the attributes is significant in the predicted linear regression to the dependent variable *levelofcrime*, by using the *anova()* function in R, we can see clearly in the output that these variables are all significant enough to the dependent variable in this model.

```
Analysis of Variance Table

Response: data$levelofcrime
                       Df  Sum Sq Mean Sq F value
data$PctUnemployed      1  97.827  97.827 633.604
data$PctImmigRec5       1  11.652  11.652  75.468
data$TotalPctDiv        1  71.505  71.505 463.120
data$PopDens            1   4.066   4.066  26.337
data$PctLess9thGrade    1   5.522   5.522  35.762
data$medIncome          1   0.452   0.452   2.927
Residuals            1987 306.789   0.154
                       Pr(>F)
data$PctUnemployed    < 2.2e-16 ***
data$PctImmigRec5     < 2.2e-16 ***
data$TotalPctDiv      < 2.2e-16 ***
data$PopDens          3.147e-07 ***
data$PctLess9thGrade 2.637e-09 ***
data$medIncome          0.08727 .
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Backforward and Backward Selection:

We then run the selection method to gain the best fit model with lowest AIC, by using the *step()* function in R, we can see both backward and forward selection select the model with full attributes.

```
Forward Selections
Step:   AIC=-3718.24
```

```
data$levelofcrime ~ data$TotalPctDiv + data$PctLess9thGrade +
data$PctUnemployed + data$PctImmigRec5 + data$PopDens + data$medIncome
```

Backward selection:

```
Start:   AIC=-3718.24
data$levelofcrime ~ data$PctUnemployed + data$PctImmigRec5
+ data$TotalPctDiv + data$PopDens + data$PctLess9thGrade
+ data$medIncome

                     Df Sum of Sq    RSS     AIC
<none>                            306.79 -3718.2
- data$medIncome      1     0.452 307.24 -3717.3
- data$PopDens        1     2.137 308.93 -3706.4
- data$PctImmigRec5   1     4.291 311.08 -3692.6
- data$PctLess9thGrade 1    5.973 312.76 -3681.8
- data$PctUnemployed  1     6.479 313.27 -3678.6
- data$TotalPctDiv    1    61.007 367.80 -3358.6
```

3. Dataset Split

We split the data into the test and train dataset with the portion of 66% for test and 34% for train and save the reduced datasets into csv files by using sample() and logical indexing in R.

4. Original Dataset Classification

We then run our original dataset in Weka, Choose Classify with the Selector of Naive Bayes *(NaiveBayes*), Logistic (*function->Logistic*), Decision Tree (*RepTree*), SMO (SMO), and KNN (K=1) (*lazy->ibk*), and got buffer output and confusion matrix.

5. Attributes Selection for Training and Testing Dataset

We open the train dataset and use Attribute Selection in Weka, choose ClassifiersubsetEval and set decision algorithm as Naive Bayes, Logistic, Decision Tree, SMO and KNN (K=1), then we got the output with the attributes selected, we mark down these attributes save the reduced dataset, then use the same attributes to select in testing dataset and save it. Finally we have 5 pairs of training and testing datasets. They are:

ClassifiersubsetEval+Naive Bayes: testing1 and training1
ClassifiersubsetEval+Logistic: testing2 and training2
ClassifiersubsetEval+Decision Tree: testing3 and training3
ClassifiersubsetEval+SMO: testing4 and training4
ClassifiersubsetEval+KNN(K=1): testing4 and training4

6. Reduced Dataset Classification

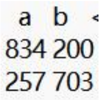
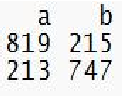
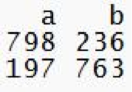
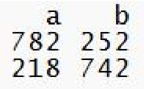
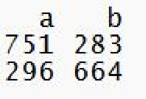
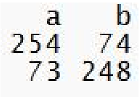
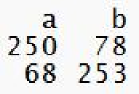
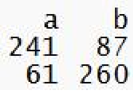
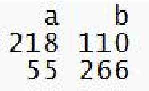
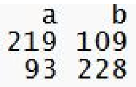
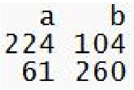
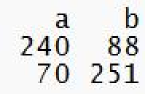
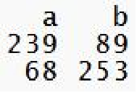
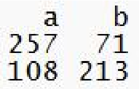
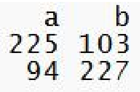
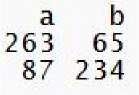
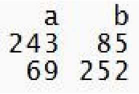
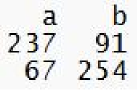
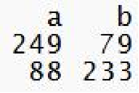
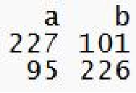
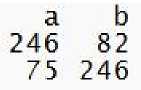
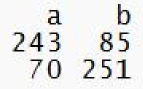
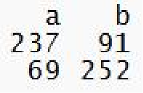
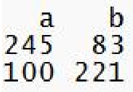
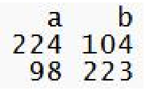
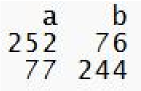
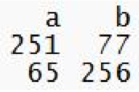
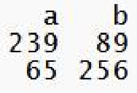
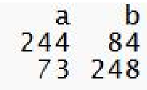
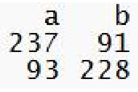
We open each train dataset in Weka, Choose Classify with the Selector of Naive Bayes *(NaiveBayes*), Logistic (*function->Logistic*), Decision Tree (*RepTree*), SMO (SMO), and KNN (K=1) (*lazy->ibk*), and test it with the testing dataset with the same attributes accordingly. We got buffer output and a confusion matrix.

Naive Bayes
Logistic
Decision Tree
SMO
KNN(K=1)

VIII. Data Mining Result and Evaluation

1. Performance Measures of All 25 Models

The confusion matrices of 5+25 models are shown below.

Naïve Bayes  a b < 834 200 257 703	Logistic  a b 819 215 213 747	SMO  a b 798 236 197 763	Decision Tree  a b 782 252 218 742	KNN  a b 751 283 296 664
Bayes-Bayes  a b 254 74 73 248	Bayes-Logistic  a b 250 78 68 253	Bayes-SMO  a b 241 87 61 260	Bayes-Tree  a b 218 110 55 266	Bayes-KNN  a b 219 109 93 228
Logistic-Bayes  a b 224 104 61 260	Logistic-Logistic  a b 240 88 70 251	Logistic-SMO  a b 239 89 68 253	Logistic-Tree  a b 257 71 108 213	Logistic-KNN  a b 225 103 94 227
SMO-Bayes  a b 263 65 87 234	SMO-Logistic  a b 243 85 69 252	SMO-SMO  a b 237 91 67 254	SMO-Tree  a b 249 79 88 233	SMO-KNN  a b 227 101 95 226
Tree-Bayes  a b 246 82 75 246	Tree-Logistic  a b 243 85 70 251	Tree-SMO  a b 237 91 69 252	Tree-Tree  a b 245 83 100 221	Tree-KNN  a b 224 104 98 223
KNN-Bayes  a b 252 76 77 244	KNN-Logistic  a b 251 77 65 256	KNN-SMO  a b 239 89 65 256	KNN-Tree  a b 244 84 73 248	KNN-KNN  a b 237 91 93 228

Note: a=low, b=high; Bayes=Naïve Bayes, Tree=Decision Tree.

In each matrix, first row = a; second row = b

The performance of 5+25 models is shown as below. The models with best performances are bolded.

Original Dataset

Classify Algorithm	Accuracy (%)	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Naïve Bayes	77.0812	0.771	0.232	0.771	0.771	0.770	0.541	0.849	0.835
KNN	70.9629	0.710	0.292	0.710	0.710	0.710	0.418	0.709	0.659
Decision Tree	76.4293	0.764	0.235	0.765	0.764	0.764	0.529	0.820	0.789
Logistic	**78.5356**	**0.785**	**0.215**	**0.785**	**0.785**	**0.785**	**0.570**	**0.864**	**0.858**
SMO	78.2849	0.783	0.216	0.784	0.783	0.783	0.566	0.782	0.722

Reduced Training-Testing Datasets

Selection-Classify	Accuracy (%)	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Naïve Bayes-Naïve Bayes	77.3498	0.773	0.227	0.774	0.773	0.740	0.547	0.839	0.844
Naïve Bayes-KNN	68.8752	0.689	0.311	0.689	0.689	0.689	0.378	0.689	0.630
Naïve Bayes-Decision Tree	74.5763	0.746	0.252	0.753	0.746	0.744	0.500	0.795	0.766
Naïve Bayes-Logistic	77.5039	0.775	0.225	0.775	0.775	0.775	0.550	0.840	0.828
Naïve Bayes-SMO	77.1957	0.772	0.227	0.774	0.772	0.772	0.546	0.772	0.711
KNN-Naïve Bayes	76.4253	0.764	0.236	0.764	0.764	0.764	0.528	0.844	0.831
KNN-KNN	71.6487	0.716	0.284	0.716	0.716	0.716	0.433	0.716	0.655
KNN-Decision Tree	75.8089	0.758	0.242	0.758	0.758	0.758	0.517	0.789	0.740
KNN-Logistic	**78.1202**	**0.781**	**0.218**	**0.782**	**0.781**	**0.781**	**0.563**	**0.852**	**0.843**
KNN-SMO	76.2712	0.763	0.237	0.764	0.763	0.762	0.527	0.763	0.701

Decision Tree-Naïve Bayes	75.8089	0.758	0.242	0.758	0.758	0.758	0.516	0.833	0.819
Decision Tree-KNN	68.8752	0.689	0.311	0.689	0.689	0.689	0.378	0.689	0.630
Decision Tree-Decision Tree	71.8028	0.718	0.283	0.718	0.718	0.718	0.436	0.787	0.758
Decision Tree-Logistic	76.1171	0.761	0.238	0.762	0.761	0.761	0.523	0.835	0.823
Decision Tree-SMO	75.3467	0.753	0.246	0.755	0.753	0.753	0.508	0.754	0.692
Logistic-Naïve Bayes	74.5763	0.746	0.253	0.751	0.746	0.745	0.497	0.825	0.811
Logistic-KNN	69.6456	0.696	0.303	0.697	0.696	0.696	0.393	0.697	0.637
Logistic-Decision Tree	72.4191	0.724	0.277	0.727	0.724	0.723	0.451	0.789	0.757
Logistic-Logistic	75.6549	0.757	0.243	0.757	0.757	0.756	0.514	0.834	0.822
Logistic-SMO	75.8089	0.758	0.241	0.759	0.758	0.758	0.518	0.758	0.696
SMO-Naïve Bayes	76.5794	0.766	0.235	0.767	0.766	0.765	0.532	0.842	0.828
SMO-KNN	69.7997	0.698	0.302	0.698	0.698	0.698	0.396	0.698	0.638
SMO-Decision Tree	74.2681	0.743	0.258	0.743	0.743	0.743	0.485	0.796	0.756
SMO-Logistic	76.2712	0.763	0.237	0.763	0.763	0.763	0.526	0.844	0.830
SMO-SMO	75.6549	0.757	0.243	0.758	0.757	0.756	0.515	0.757	0.695

2. Results

For the initial data, we used the Weka site and selected the 10-fold test and derived the output of the above five classification methods, where the best is Logistic algorithm with 78% accuracy, Naive Bayes, Decision tree and SMO between 76% - 78%, the worst is KNN at only 70%.

Next, we want to check if the best models will be different when we want to focus on specific data, like only TPR or precision. By observation of the results and tables we make above, we find that the models with the highest accuracies also have the highest TPR, precision and MCC. Therefore, we have sufficient evidence to conclude that the best models can make better predictions than others, no matter in terms of the overall accuracy or any single measurement, like TPR or precision.

3. Justification for My Selection of the Best Model

According to our decision rules-best accuracy among all 30 models. We can say the best model is the not reduced dataset classified by Logistic with an accuracy of 78.53%. Also, this classification has the best result in all assessment indexes.

IX. Discussion and Conclusion, Including What We Learn from This Project

In this project, we can see immigration, divorce rate, education and income are very influential to violent crime rate and they are also the best to predict the value of the crime rate. The governments, police departments, and other relevant departments may need to be more mindful of the impact of immigration, divorce rate, education and income to be precautious of rising crime trends, and thus reduce crime. More specifically, the relevant departments should work harder on detecting and solving problems including but not limited to illegal immigrants, domestic violence, dropout rate and unemployment rate.

Besides these factors, the other two factors also show some impact on the violent crime rate, as they are not completely uncorrelated to the crime rate. But of course, if we have to make a plan according to the priority of these factors, the relevant departments can take these factors into consideration later, but we strongly suggest that to keep an eye on these factors also.

According to this project, we also find the most accurate models to predict the violent crime in the future, which are Logistic and KNN-Logistic. If experts in the relevant departments can use these models to analyze and predict the violent crime rate and tendency in the future, they can be a helpful and useful tool to assist them and apply some precautions to reduce and stop the violent crime before they happen, and safeguard the innocent people.

Even though we find the best models after all the analysis we make, there are some potential improvements for our accuracy. Although the six attributes we select are representative factors in our society, there may be other factors which are related to the violent crime rate. For instance, in the original dataset, we have attributes about police, such as PolicPerPop (police per population), PolicCars (the amount of police cars), and PolicOperBudg (the budget of police operations). The police force is a vital factor to influence the security of an area by our knowledge and common senses, but we do not take these attributes into account as they have too many NA values – more precisely, most of their data are NA. In other words, they do not have enough data to analyze. Plus, we also exclude the attributes about race and ethnicity, such as racePctBlack (percentage of black people), racePctWhite (percentage of white people), etc., to prevent any possible sensitive information or unintentional offense. What is more, we notice that the attribute ViolentCrimesPerPop (violent crimes per population) contains some NA values, so our accuracy can be increased if these data can be completed in the future. Last but not least, this research focuses on overall violent crime rate in the United States, but we speculate that the specific situations in different states or communities may vary, which requires deeper and more detailed studies.

For better classification, the thing we can improve out of this project could probably be using data from each year to sketch a clearer trend of the violent crime rate. If possible, we may also investigate some special factors in each state or community, like if the violent crime rate will increase or decrease in certain periods of a year. According to our life experience, in China, the crime rate usually has an obvious increase just before the Spring Festival, as most people are going back to their hometown and thus increase the population mobility and chaos in the whole society. Therefore, we speculate that similar things may also take place in the United States, during some festivals or special events. Unfortunately, this idea cannot be reached under current sources, and requires far more research to respond. However, once figured out, the accuracy of the analysis and prediction will be much higher and will make more contribution to the government, the police, and other relevant departments.

X. References

[1][2]: Statista Research Department. “Reported violent crime rate in the U.S. 1990-2021”, https://www.statista.com/statistics/191219/reported-violent-crime-rate-in-the-usa-since-1990/, Oct 10, 2022.

XI. Acknowledgements

We would like to thank our professor, Jae Lee, for all helpful instructions and guidance.

For this project, each member of our team contributes equally. Yuandi Tang completes part IV, V, VI, VII and VIII, and Tianjun Ma completes part I, II, III, VIII and IX.