Prediction on Homicide Reports, 1980-2014

Shen Liu
UC San Diego
A53094622
shl425@ucsd.edu

Linlin Wang
UC San Diego
A53097909
liw087@eng.ucsd.edu

Xiaonan Zhi
UC San Diego
A53087211
xizhi@ucsd.edu

ABSTRACT

In this project, we use the homicide dataset from 1980 to 2014 to predict the perpetrator's sex, age and race. First we analyze the dataset in terms of State, Year, characteristics of victims and perpetrators, Relationship and Weapon. And then we calculate the correlation between each pair of useful features to help us choose features. After describing these three prediction task, we dig into each task by analyzing the pre-processing method, the features chosen, the evaluation method chosen, the model chosen and draw a conclusion from our results by comparing with baseline results. The goal of our project is to help FBI to narrow down the search range for the perpetrators to solve the homicides more efficiently.

1. EXPLORATORY ANALYSIS OF DATASET

The dataset used for this project is the homicide reports from 1980 to 2014 which is getting from Murder Accountability Project. The Murder Accountability Project is a nonprofit group which dedicates to emphasize the importance of investigation of homicides within the United States. It seeks to obtain information from federal, state and local governments and consists of retired law enforcement investigators, investigative journalists, criminologists and other experts on various aspects of homicide. The dataset includes murders from the FBI's Supplementary Homicide Report and Freedom of Information Act data on more than 22,000 homicides that were not reported to the Justice Department. It consists of around 620K valid homicides data of which 440K are solved and 180K are unsolved. Each homicide data includes the following important features:

1. State: The state where the homicides happened
2. Year: The year when the homicides happened
3. Crime Type: It includes two type of crime , i.e. Murder or Manslaughter and Manslaughter by Negligence
4. Crime Solved:It indicates whether the homicide has been solved or not.
5. Victim Sex
6. Victim Age
7. Victim Ethnicity
8. Victim Race
9. Perpetrator Sex
10. Perpetrator Age
11. Perpetrator Ethnicity
12. Perpetrator Race
13. Relationship: The relationship between the victim and perpetrator. It has 27 different kinds of relationship.
14. Weapon: The weapon used for the homicide. It has 15 different kinds of weapons.

Homicide here includes Murder and Manslaughter by negligence. In our project, we assume that future homicides can be predicted on basis of various characteristics of the past cases, i.e. they may be similar in terms of the victim's characteristics such as age, race and sex, weapon of the homicide, location of the homicide and relationship between victim and perpetrator in the homicide. So we should investigate both the perpetrators and the victims. In addition, We believe that it is instructive to investigate the overall homicide trends to select the important determinants of homicide. These features may have a great impact on predicting the information of perpetrator. So we investigate the trend in the dataset first, and then we investigate the correlation between each pair of these useful features.

1.1 Homicides distribution in terms of State

We calculate the number of total homicides in each state and plot it on the map showing in Figure 1. From Figure 1, we can get that top 5 states with high total number of homicides are: California, Texas, New York, Florida, Michigan.

1.2 Homicides distribution in terms of Year

We calculate the number of total homicides happened per year and plot it in Figure 2. From Figure 2, we can get that the year 1993 has the largest number of total homicides and from 1993 to 1999, the number of total homicides has decreased noticeably. After 1999, the number of total homicides tends to be stable each year and remains around 15000 homicides each year which is a good tendency and represents the low homicide rate in recent years.

1.3 Characteristic of victims and perpetrators

We analyze the characteristic of victims and perpetrators in terms of their sex, age, race and ethnicity.
In terms of sex, we can see from Figure 3 and Figure 4 that most victims and perpetrators are male.
In terms of age, we can get from Figure 5 and Figure 6 that older teens and younger adults have the highest homicide victimization and offending rates. What's more, the homicide victimization and offending rates increase as the age increasing when the age is below around 20, and then the homicide victimization and offending rates decrease as the age increasing. The relatively high homicide victimization

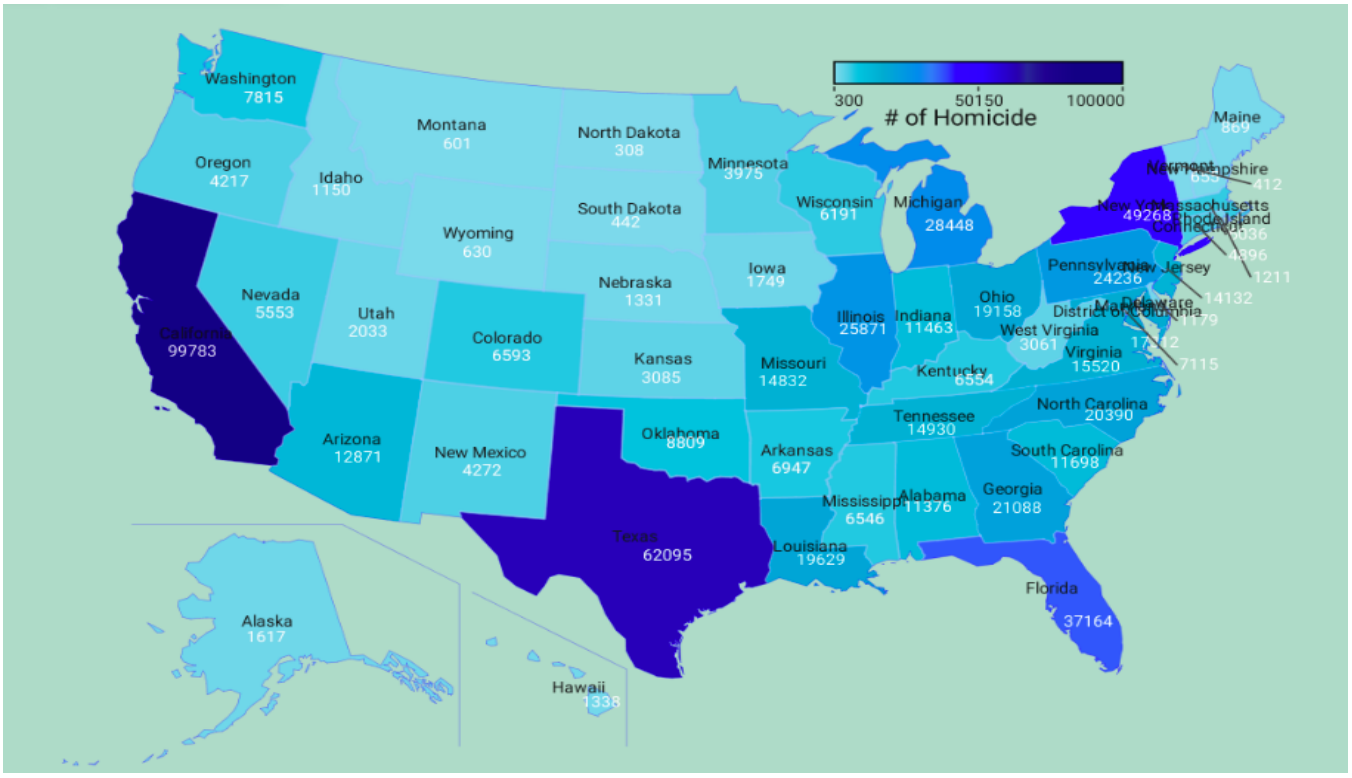

Figure 1: Homicides distribution in terms of State

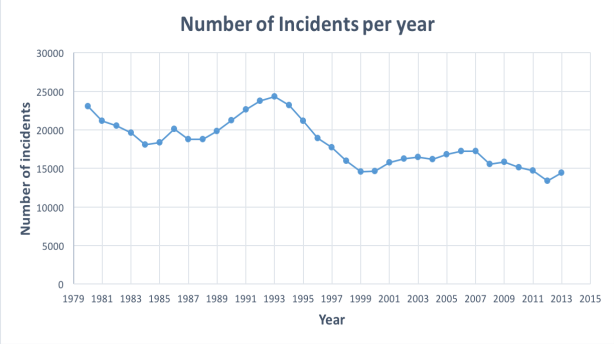

Figure 2: Homicides distribution in terms of Year

rate when victim age equals to zero may come from some error data and we can drop these data for future use in our prediction task.

In terms of ethnicity, we can get from Figure 7 that for the victim ethnicity, a person who is not Hispanic is more likely to be the victim than person who is Hispanic. From Figure 8, we can get that for the perpetrator ethnicity, a person who is not Hispanic is more likely to be the perpetrator than person who is not Hispanic. But the reason leads to this condition may be that in real life, there are more people who are not Hispanic than people who are Hispanic.

In terms of race, we can get from Figure 9 and Figure 10 that white and black people have the highest homicide victimization and offending rates. Same as the condition for ethnicity, the reason may be that there are more white and black people in America than other races. So we need to move forward to investigate the relationship between the race of victim and the race of perpetrator which is more valuable for the future prediction task. From Figure 11, we can get that most victims are killed by the perpetrator with the same race.

1.4 Relationship

From Figure 12, there are 27 different kinds of relationship between victim and perpetrator. Acquaintance and stranger are two relationship with the top 2 highest homicide rate. Wife killed by husband, friend killed by friend and girlfriend killed boyfriend also have relatively high homicide rate. Other relationship such as son, family, husband, daughter, boyfriend, neighbor have relatively smaller homicide rate than the ones mentioned above. We may categorize the relationship into three categories, i.e. Stranger or acquaintance, families or in a relationship, stepparents or employment relationship. The first category has the highest homicide rate, the second one has relatively lower homicide rate and the third one has the lowest homicide rate. So it is most likely that the victim is killed by someone he or she doesn't know or the person he or she has no intimate relationship.

1.5 Weapon trend for homicide

We can get from Figure 13 that there are 15 different kinds of weapons. Homicides are mostly committed using handgun. Knife, blunt object, firearm, shotgun and rifle are also involved in many homicides.

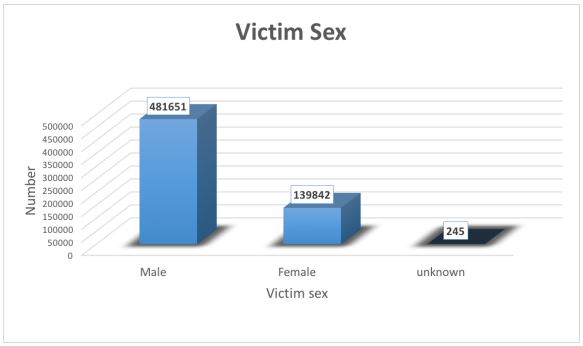

Figure 3: Victim Sex distribution

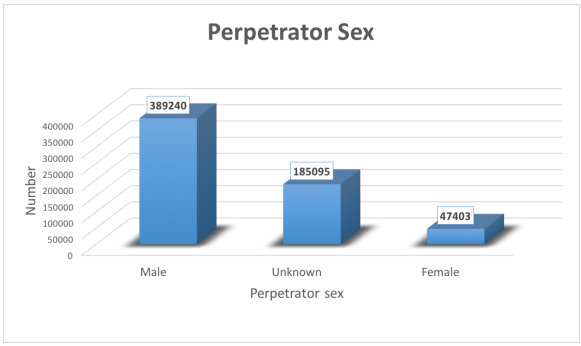

Figure 4: Perpetrator Sex distribution

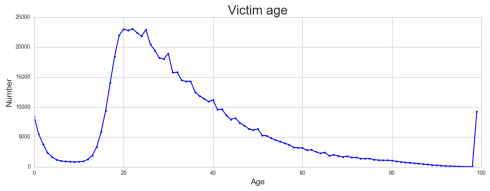

Figure 5: Victim Age distribution

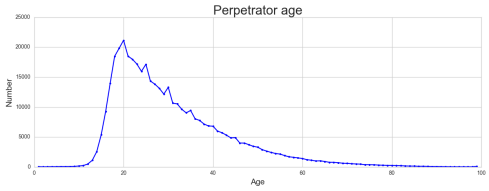

Figure 6: Perpetrator Age distribution

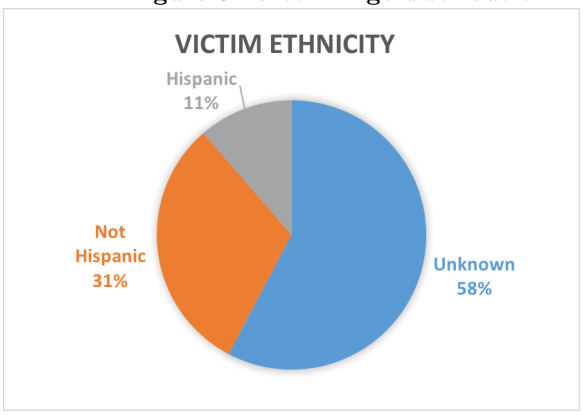

Figure 7: Victim Ethnicity distribution

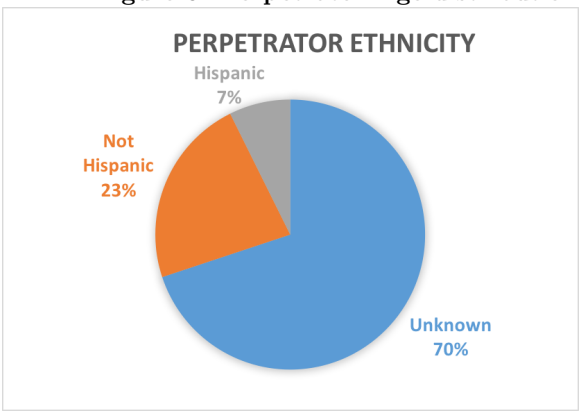

Figure 8: Perpetrator Ethnicity distribution

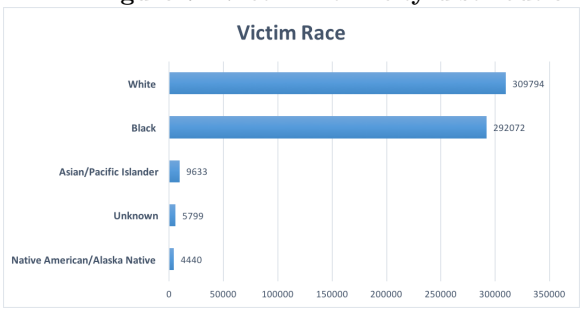

Figure 9: Victim Race distribution

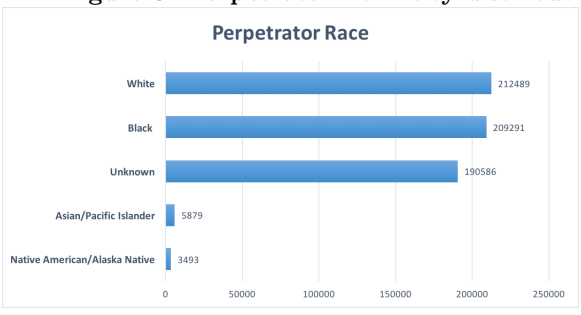

Figure 10: Perpetrator Race distribution

Victim Race	Perpetrator Race	Number of Victims
White	White	136945
	Black	21152
	Unknown	1623
	Asian/Pacific Islander	965
	Native American/Alaska Native	937
Black	White	10847
	Black	125121
	Unknown	1100
	Asian/Pacific Islander	340
	Native American/Alaska Native	123
Asian/Pacific Islander	White	1042
	Black	748
	Unknown	90
	Asian/Pacific Islander	2570
	Native American/Alaska Native	19
Native American/Alaska Native	White	806
	Black	241
	Unknown	25
	Asian/Pacific Islander	20
	Native American/Alaska Native	1366
Unknown	White	503
	Black	455
	Unknown	964
	Asian/Pacific Islander	21
	Native American/Alaska Native	11

Figure 11: Race relationship between victim and perpetrator

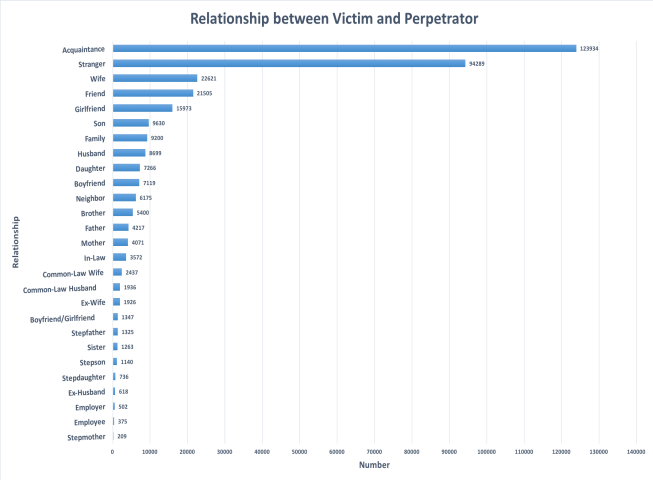

Figure 12: Relationship between victim and perpetrator

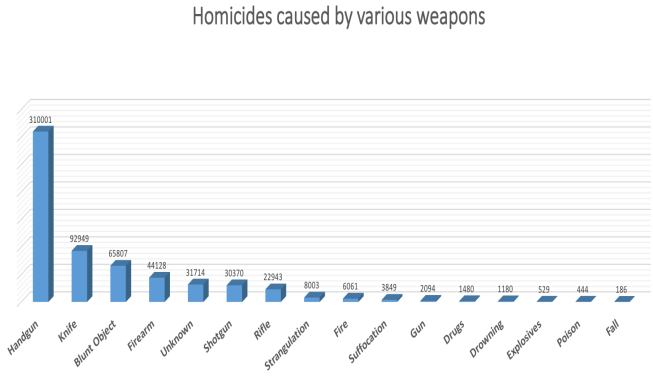

Figure 13: Weapon trend for homicide

1.6 Correlation between useful features

After analyzing all the features of victim and perpetrator as well as the relationship between them, we can investigate all these features to see how they relate to each other which is useful for choosing features in the future prediction task. In correlation analysis, we estimate a training set correlation coefficient, more specifically the Pearson Product Moment correlation coefficient, which ranges between -1 and +1 and quantifies the direction and strength of the line association between each pair of two variables. If the correlation coefficient between two variables is positive, the two variables move into the same direction and are proportionate to each other. Otherwise, they move into the opposite direction and are inversely proportionate to each other. The magnitude of the correlation coefficient indicates the strength of the association. Thus from the heat map shown in Figure 14, we can see a strong correlation between Victim Ethnicity and Perpetrator Ethnicity which has the correlation coefficient as 0.7. Moreover, if we take a look at the relationship between the perpetrator race and victim race, we could see that the correlation coefficient between these two features is 0.56 which is the highest correlation coefficient for perpetrator race feature. Based on our life experience and the analyzing in section 1.3, the race of victim and perpetrator are more likely to be related to each other. By analyzing the magnitude of correlation coefficient between victim sex and other features, we could see that the most correlated features with victim sex are perpetrator age, relationship and weapon, which makes sense by our common sense. Perpetrator Sex is highly correlated with Perpetrator Age, Perpetrator Ethnicity and relationship between victim and perpetrator. And Perpetrator Sex is correlated with Victim Race and Victim Sex. Perpetrator Age is correlated with Perpetrator Sex, Perpetrator Ethnicity, relationship, Victim Sex, Victim Age and Victim Race. Relationship between victim and perpetrator is correlated with Perpetrator Age, Victim Sex, Perpetrator Ethnicity, weapon and Perpetrator Sex.
The correlation coefficient matrix gives us an idea about how to design features when we make linear regression. In another words, given some features such as the information about the victim, we can find useful features using the correlation coefficients and build a prediction model to predict some unknown features which can be helpful for solving the homicide.

2. PREDICTION TASK

2.1 Prediction Task description and importance

In the project, we are interested in building the perpetrator characteristic prediction model for the Homicide incidents dataset coming from Murder Accountability Project from 1980 to 2014. In each incident, the police wants to get more information or makes more reasonable guess from victim information and the evidence that the perpetrator left. The most efficient way to downsize the scale of possible perpetrators is to filter all the possible perpetrators by their sex, age and race. As a result, in our project, we use more than 620k raw dataset to train our model to predict the sex, age range and race based on the information we have already had. In real life, our predictions could help the police automatically filter out a large number of possible perpetrator suspect and help the police focus their investigation on a

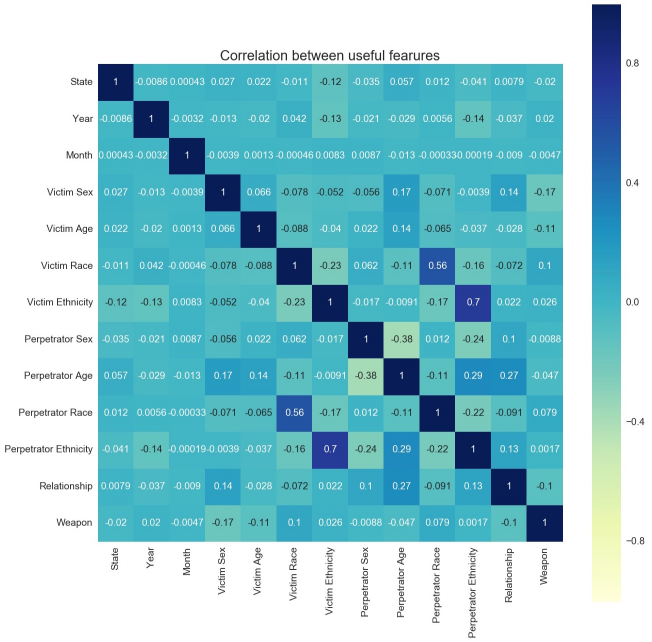

Figure 14: Correlation between useful features

certain group of people who have higher possibility to murder the victim. In addition, we notice that there are a great number of unsolved incidents from 1980 to 2014, our model may give police a new guide to solve the old cases, which is a good news for both police officers and the victim's family.

2.2 Data Preprocessing

Before the beginning of training our model, we must clean the data. The raw dataset we get includes the following columns of features: Record ID, Agency Code, Agency Name, Agency Type, City, State, Year, Month, Incident, Crime Type, Crime Solved, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, Perpetrator Count, Record Source. It is easy to see that there are lots of features that we will not use in our prediction like Agency Code, Agency Name and Agency Type. So we need to delete those redundant columns first of all. Then we have the reduced columns of features: City, State, Year,Month, Crime Type, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity, Relationship, Weapon. After that, there are lots of unsolved cases records in the dataset, which could not be used to train or even test our model. Then we delete the records whose Crime solved feature is unsolved. Thirdly, for the solved incidents there are unreasonable information. For example, the age of perpetrator and victim could not be zero or more than one hundred. As a result, we cut off the data who has the age information more that 100 or less than 0. In this case, we need to take out the nonsense incidents and use the rest of information to train and test our model. Following the three steps above, we finally get a clean dataset to train and test our model. In a nutshell, we have around 400k clean dataset to train and test our model. Moreover, we need to assign each type of label to a digital representation. For instance, we use binary representation zero and one to represent male and female. Then we could have several digital dictionaries which yield several maps for the digital label and original description in words. Finally, we have a clean dataset with each feature setting as a digital number to represent it. And we have a bunch of dictionaries to map the digital label back to the word's description of that characteristic.

2.3 Feature chosen and Model chosen vs. exploratory analysis results

In section 3, we will analyze the process of choosing features for each prediction task. But the process should all be related with our exploratory analysis results. Taking the prediction of perpetrator's sex as an example, after analyzing the exploratory analysis of perpetrator's sex we find that the dataset is really imbalanced and try to solve this problem by Soft-Margin SVM and oversampling method. What's more, we choose different evaluation method to fit into each task by analyzing the exploratory analysis results. And for the feature chosen for each predicting task, we look into the correlation matrix shown in Figure 14 and find the highly correlated features to include in our feature array. By analyzing the prediction results on the dataset, we can choose the features from these highly correlated features by including one at a time and see the results.

3. METHODOLOGY

In this section, we will introduce the model we choose and the method we use. After that, we will analyze each prediction task by preprocessing data, evaluation method, baseline chosen, model chosen and feature chosen.

3.1 Model Introduction

3.1.1 Linear Regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and an array of independent variables denoted X. Our linear regression model is fitted using the least square approach based on the Euclidean distance .

$$\| x \|_2 = (\sum_i x_i^2)^{\frac{1}{2}} \tag{1}$$

In prediction task, the linear regression assumes a predictor of the form as following.

$$X\theta = y \tag{2}$$

X denotes the matrix of features, theta represents the unknown parameters we want to train, and y is the vector of outputs. Solving the parameter theta by deviation, we get the following equation.

$$\theta = (X^T X)^{-1} X^T y \tag{3}$$

3.1.2 Logistic Regression

Logistic Regression model aims to model $P(label|data)$ by training a classifier of the form

$$y_i = \begin{cases} 1, & if\ X_i\theta > 0 \\ 0, & otherwise \end{cases} \tag{4}$$

Using sigmoid function to convert a real-valued expression $X_i\theta$ into a probability $P_\theta(y_i|X_i) \in [0, 1]$. Then we need to

optimize the probability showing in equation 5 by taking logarithm, subtract regularizer which is showing in equation 6. And then we compute the gradient and use gradient ascent to solve the problem. The gradient is showing in equation 7. After getting theta from equation 7, we do the classification by equation 8. In terms of regularization, it is of the process of penalizing model complexity during training, we use the validation set to tune the regularizer parameter.

$$argmax_\theta \prod_{y_i=1} P_\theta(y_i|X_i) \prod_{y_i=0} (1 - P_\theta(y_i|X_i)) \parallel \theta \parallel_2^2 \quad (5)$$

$$argmax_\theta \sum_i -log(1 + e^{-X_i\theta}) + \sum_{y_i=0} -X_i\theta - \lambda \parallel \theta \parallel_2^2 \quad (6)$$

$$\frac{\partial l}{\partial \theta_k} = \sum_i X_{i,k}(1 - \sigma(X_i\theta)) + \sum_{y_i=0} -X_{i,k} - 2\lambda\theta_k \quad (7)$$

$$y_i = \begin{cases} 1, & if\ \sigma(\ X_i\theta) \geq 0.5 \\ 0, & otherwise \end{cases} \quad (8)$$

3.1.3 Support Vector Machine(SVM)

SVM training algorithm builds a model which is a non-probabilistic linear classifier. The goal of SVM is to represent the examples as points in space and build a hyperplane to separate the examples into separate categories and makes the gap between two categories as far as possible. We implement two kinds of linear SVM in our project, which are hard-margin SVM and soft-margin SVM. We have a classifier of the form, and we want to minimize the number of mis classifications.

$$y_i = \begin{cases} 1, & if\ X_i\theta - \alpha > 0 \\ -1, & otherwise \end{cases} \quad (9)$$

$$argmin_\theta \sum_i \delta(y_i(X_i\theta - alpha) \leq 0) \quad (10)$$

Solving the problem by applying Numerical Optimization method, we have the expression of hard-margin SVM as following.

$$argmin_{\theta,\alpha} \frac{1}{2} \parallel \theta \parallel_2^2 \quad \forall_i y_i(\theta X_i - \alpha) \geq 1 \quad (11)$$

However, in some situation, misclassifying a certain type of problem into another one will cause larger damage. For example, in our dataset, the number of male perpetrators are more than female. When we train the SVM to predict the sex of perpetrator, we need to take this unbalanced fact into consideration. In this case, we want to make the margin to be as wide as possible while penalizing points on the wrong side of it. Then we introduce the soft-margin SVM formation as following.

$$argmin_{\theta,\alpha,\xi_i>0} \frac{1}{2} \parallel \theta \parallel_2^2 + C \sum_i \xi_i \quad (12)$$

$$\forall_i y_i(\theta X_i - \alpha) \geq 1 - \xi_i$$

3.1.4 Oversampling for imbalanced data

Imbalanced data means the data with prediction task where the classes for the classification are not represented equally. For example, in our sex prediction task, we have a binary classification problem for the perpetrator, i.e. male or female with an imbalanced dataset which has the ratio of male to female at around 8:1. After we creating a classification model, we evaluate the classification accuracy which is the first measure we commonly use for classification problem and get around 90% accuracy immediately. But when we analyze the result, we find that all the data belongs to one class. The reason is that our model look into the data and decide the best thing to do is to always predict Male to achieve high classification accuracy which is the outcome of using an imbalanced dataset in terms of sex. To handle with this problem, we may change the evaluation method from precision to true positive rate, true negative rate and balanced error rate. In addition to the soft-margin SVM method we mentioned above, we can also change to dataset to make it more balanced, i.e. oversampling. Oversampling in data analysis is the technique we use to adjust the class distribution of a dataset. We can add copies of instances or generate synthetic samples of the under-represented class to make the dataset more balanced. There are many oversampling techniques for classification problems, we choose SMOTE for our sex prediction task.

3.2 Perpetrator Sex Prediction Task

For the perpetrator sex prediction task, we first reshuffle the data since we may use the Year and State feature in the prediction task. And then we use the first 200K raw data to train, 100K raw data to validate and 200K data to test. To use the data for our prediction task, we need to pre-process the dataset by filtering out all the unsolved cases and all the cases with victim's age or perpetrator's age that are less equal than zero or bigger than 100, then we get around 140k data to train, 67k data to validate and 138k data to test. To evaluate our model for predicting perpetrator's sex, we need to consider not only the accuracy but also TPR, TNR and BER which are listed in detail as below. In our prediction task, we label Male as 1, i.e. positive and Female as -1, i.e. negative.

1. ACC(Accuracy):
   Correct predictions / number of predictions
2. TPR(True Positive Rate):
   True Positives / number of labeled positive, where True positives are number of predictions which are positive and with positive labels.
3. TNR(True Negative Rate):
   True Negatives / number of labeled negative, where True negatives are number of predictions which are negative and with negative labels.
4. BER(Balanced Error Rate):
   1 - (TPR + TNR) / 2, is 0.5 for a random classifier and 0 for a perfect classifier.

The reason for choosing these evaluation methods is that the dataset is really imbalanced in terms of sex for perpetrator. By just evaluating the model using the classification accuracy is really misleading since the model will predict all to Male to get high accuracy. Thus for evaluating the model for imbalanced data, we need to consider the TPR, TNR and BER.

For the perpetrator sex prediction task, we use "all predict to Male" as the baseline due to the imbalanced feature in the dataset. We build three different kind of model:

Evaluation / Feature	TPR	TNR	BER	ACC
Victim features	0.71008	0.60504	0.34243	0.69867
Add State	0.72252	0.60026	0.33860	0.70947
Add Year	0.72581	0.58131	0.34643	0.71055
Add State Year	0.73141	0.58292	0.34282	0.71572

Table 1: Feature chosen

1. Soft-margin SVM
2. Oversampling data and SVM
3. Oversampling data and logistic regression

Considering about the realty for this kind of prediction, we first include all victim's features as the feature array. For the victim's features, we create the feature array by using zero one arrays. Taking the Victim_Sex as an example, it is $[1, 0]$ if Victim Sex is Male and $[0, 1]$ if Victim Sex is Female. Given all the characteristics of the victim, we need to predict the sex of the perpetrator. We use the Soft-Margin SVM model to help use evaluate the model for choosing appropriate features. After analyzing the correlation between useful features in Figure 14, we find that State and Year are highly correlated with perpetrator's sex. So we create the Year feature by count all the possible year in our dataset and create the zero one array to append on the feature array we just create. And we create the State feature using the top 5 high homicides rate states in section 1.1, i.e. State = [1, 0, 0, 0, 0] if the state in data is California, etc. Then we can use the features adding State, the features adding Year, the features adding State and Year to train the model and determine what features to include. The feature array are showing as below:

- Using all victim features:
  feature_array $= [1, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race]$
- Using all victim features and State feature:
  feature_array_State $= [1, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, State]$
- Using all victim features and Year feature:
  feature_array_Year $= [1, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, Year]$
- Using all victim features, State feature and Year feature:
  feature_array_State_Year $= [1, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, State, Year]$

By comparing the outcomes showing in table 1, we finally choose all victim features and State feature for our perpetrator sex prediction task.

3.3 Perpetrator Age Prediction Task

Perpetrator age prediction model utilizes some given features of victims and perpetrators to predict the possible age of the perpetrator for each crime. The first 300K datasets will be trained to generate the model. Since FBI provided the dataset in the order of year (1980 - 2014), we need to randomize the data before training if year is considered as a feature. The perpetrator age of the trained datasets is almost well-balanced with a range of 0 to 100, and only 18 to 60 year old perpetrators will be used to train since most perpetrators' age fall in-between this range.
The baseline of predicting age can be achieved by simply calculating the average age of perpetrators among all dataset. In contrast, using linear regression model is more acceptable since we are predicting an integer in this model:

$$X\theta = y$$

In another word, we need to select certain features of victims and perpetrators to optimize the prediction. If we are only given information of victims, we chose the following features which produce the best result:
feature_array $= [1, Year, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race]$
Here, we use 0-1 represents some of the features. For example, if the victim is male, $Victim\_Sex = [1, 0]$, and $[0, 1]$ otherwise.
If we assume some of the information about the perpetrators is exposed after the investigation such as perpetrator's race and sex, we can then append two more features to the array:
feature_array $= [1, Year, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, Perpetrator\_Sex, Perpetrator\_Race]$

3.4 Perpetrator Race Prediction Task

In the perpetrator race precision task, we use the first 300K raw data to train and 100K raw data to test. To use the data to train, we firstly need to filter out all the unsolved cases. Then we also need to delete the cases in which the label is unknown(perpetrator's race). After cleaning the data, we get the size of training dataset is around 211k . The size of testing dataset is around 72k. Then for the evaluation purpose, we evaluate our prediction results by computing the accuracy of test dataset. In addition, we also evaluate the prediction accuracy for each race of perpetrator. The reason why we choose this evaluation method is that we have four different types of race which are White, Black , Native American/Alaska Native and Asian/Pacific Islander. The White and Black are two kinds of dominant race and the other two have much smaller size. We implement two models for perpetrator race prediction task.

1. Logistic Regression
   In this model, we train four different models for the four different kinds of race respectively. Each prediction model will give us a score of the probability of perpetrator in that particular race. Then we find the highest score and yield our prediction result as that specific race.
2. SVM
   Similarly, we train four models for four kinds of race respectively. The difference between SVM and Logistic Regression is that here we do not predict the score of each race. We just have a binary result that is whether the perpetrator belongs to that race or not. If more than one kinds of race give us the positive result, we choose to predict the race of perpetrator using the result with smallest distance to the hyperplane.

Based on our data analysis in the first part, using the cor-

relation matrix, we could see that the features which have highest correlation coefficients with perpetrator race are victim race, perpetrator ethnicity, and victim ethnicity. Then we design the feature vector based on two assumptions, because sometimes we have no idea about the perpetrator in real life but in other case we may get some evidence from the alibi.

1. Knowledge of perpetrator is not available:
   feature_array $= [1, Victim\_Race, Victim\_Ethnicity]$
2. Knowledge of perpetrator is available:
   feature_array $= [1, Victim\_Race, Victim\_Ethnicity, Perpetrator\_Ethnicity]$

According to the analysis from the first part of our report, we could have a good baseline in the prediction of race, because the correlation coefficient between perpetrator race and victim race is more that 0.5. As a result, a reasonable way to simply predict the perpetrator race is to predict the perpetrator and victim are from the same race. This baseline gives a good prediction because of the large correlation, and we will discuss the improvement of our model in the result part.

4. RELATED LITERATURE

Unlike the common crime, homicide includes murder and non-negligent manslaughter which is the willful killing of one human being by another [1]. Compared to other kinds of crimes, homicide has worse influence and is needed to be solved as soon as possible. Our dataset comes from the Murder Accountability Project which is a nonprofit group dedicated to help solve homicides within the United States. Government and institutions use this data set to analyze the trend of the homicides and predict homicides in some specific regions. Lots of works have been done to find the relationship between perpetrators and victims and to build the database for each incident by institutions of government. [2] Decades of research have consistently upheld that age, poverty and race are the key issues that cause homicides[3]. Adolescents especially who have opportunities to get access to guns have high risk to perform serious crime. In addition, the trends of homicides happened each year differ from each other which may be due to the political strategies and global environment. Another article gives us a Logistic Regression Model Presenting the Predictors of Intimate Partner Homicide in Portuguese[4]. They stated that men who killed their intimate partners were more likely to have used violence against a previous partner and to have used a weapon or instrument. Despite the findings they gave, the predictor has some limitations. First, the homicide reports can be affected by desirability of victims and perpetrators. And this is also a problem we may face in our prediction model. As a result, we need to delete some incomplete data from the database to train our model. Second, they have used only static linear predictive relationships and did not include interaction effects in the prediction. Moreover, those investigations usually focus on the relationship between perpetrators and victims and statistic results. According to all the investigation and statistic results from the above articles, we could find that several factors such as sex, age, race, relationship.etc contribute to the characteristics of perpetrators. Given a situation that we could only have access to the victim information and part of the perpetrator's information as we do not know who is the perpetrator, an efficient prediction based on the information we have is quite useful for police to find the perpetrator and solve the incident quickly. Then why do not we use those the huge amount of solved incidents to predict the possible characteristics of perpetrators given the information of victims and the evidence we may get from an alibi. In this case, our model could help the agency to solve homicide cases or give polices a good guess of direction to find the perpetrators.

Inspired by the linear regression, logistic regression and SVM models we have implemented in Amazon review rating predictions[5], our team consider to use the three models to predict sex, age and race given the information of victims and some incomplete information of perpetrators obtained by police. Then we could build a feature vector using all the information we have, and implement supervised learning given the label of the characteristic of solved homicide incidents. This analogy may give us the most possible range of detection for the perpetrator and save more time for police to solve new homicide incidents.

5. RESULTS

In this section, we will evaluate the results for the feature array and model we mention in section 3 by comparing it with baseline results. Then we can select the best model and interpret the parameters and features we choose and draw conclusion from the results we get for each prediction task.

5.1 Perpetrator Sex Prediction Result

For the perpetrator Sex Prediction Task, we end up choosing the following feature array:
feature_array $= [1, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, State]$
We use the victim features, i.e. sex, age, race, ethnicity and the state feature to compose the feature array. And then we use three different models to do the classification task. From Table 2 and 3, we can get that our model beats the baseline in terms of BER and TNR on train dataset and test dataset. In our baseline, the female perpetrator can never be found, however, using the three different model we build, the TNR has improved a lot and the balanced error rate has improved a lot. From Table 2, we can get that on the train set, oversampling the data will perform better than the Soft-Margin SVM method in terms of TNR and BER. And the SVM model for oversampled balanced data performs a little better than the logistic regression model for oversampled balanced data. However, from Table 3 we can get that on the test dataset, the Soft-Margin SVM model performs better in terms of TNR and BER and ACC. The oversampling method cannot perform really good on test dataset since the test dataset is still imbalanced. Thus we finally choose the Soft-Margin SVM model. In Soft-Margin SVM model, we choose the class weight vector as one for Male and six for Female. And then the penalty parameter C will be set as class weight * C(C = 1 in this model).
In oversampling method, we use the SMOTE(Synthetic Minority Over-Sampling Technique) to oversample the dataset.It works by creating synthetic samples from the minor class.
In the SVM model for oversampled dataset, we choose the penalty parameter C of the error term as one.
In the logistic model for oversampled dataset, we choose the regularizer λ as seven.
Thus from all the discussion above, we finally choose the Soft-Margin SVM model using all victim features as well as

Model \ Evaluation	TPR	TNR	BER	ACC
Baseline	1.0	0.0	0.5	0.89323
Soft-Margin SVM	0.72252	0.60026	0.33860	0.70947
Oversampling + SVM	0.750400	0.743392	0.253104	0.746896
Oversampling + logistic regression	0.71859	0.70039	0.29050	0.70949

Table 2: Perpetrator Sex Prediction Evaluation for different models on train set

Model \ Evaluation	TPR	TNR	BER	ACC
Baseline	1.0	0.0	0.5	0.87563
Soft-Margin SVM	0.67679	0.57640	0.35339	0.68959
Oversampling + SVM	0.70854	0.50495	0.39327	0.67942
Oversampling + logistic regression	0.67342	0.49960	0.41348	0.64859

Table 3: Perpetrator Sex Prediction Evaluation for different models on test set

the state feature. Our model beat the baseline in terms of TNR and BER. The value of our model is to increase the accuracy for predicting the female perpetrator while keep the accuracy for predicting the male perpetrator as an acceptable rate to decrease the balanced error rate.

5.2 Perpetrator Age Prediction Result

For the model of predicting perpetrator's age, we end up choosing the following two feature array:

- **feature_array** = $[1, Year, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race]$
- **feature_array** = $[1, Year, Victim\_Ethnicity, Victim\_Sex, Victim\_Age, Victim\_Race, Perpetrator\_Sex, Perpetrator\_Race]$

The first feature array assumes we are only given information of victims, and the second one adds two more characteristics of perpetrator indicating FBI has found some clues about the perpetrator. We test 50K dataset after training 300K crimes using linear regression model with a regularizer. We choose $\lambda = 0.01$ since our feature model is not too complex. The resulted MAE for the first feature array is 8.70920315682 and the second MAE given some information of perpetrators is 8.5399694501. The second one has a lower value since we know more related information.
The results indicate that the FBI will be able to largely narrow down their search targets with assistance of the model. For example, if the predicted age of a perpetrator is 30, the FBI can put more focus on the suspects within the range of 21 to 39 year old. The age predictions with linear regression are much better than the trivial solution (always predict the average age, MAE = 9.98066221103).
In addition, we can optimize our solution even further by adding a bias term to every single age prediction. This bias term is calculated based on the difference between the average age of the perpetrator for the state where the crime takes place and the average age of the perpetrator among all dataset:

$$\textbf{bias} = (State\_AgeAvg - Global\_AgeAvg) * factor$$

We add this extra term to each predicted result which will even lower the MAE. When factor term is approximately 0.5, we will obtain the minimum MAE result. The MAE becomes 8.69205066191 and 8.52921334012 with or without some information of the perpetrator.
To conclude our age prediction, whenever predict an integer result, linear regression should always among our top choices. The age of perpetrator is related to victim information such as age and sex. For example, a perpetrator tend to murder a victim with close age. And the more information about the perpetrator we have while training, the more accurate we can predict the result. In addition, Year might also be a useful feature since it somehow related to the age of people involved in the crime.

5.3 Perpetrator Race Prediction Result

For the perpetrator race prediction task, we end up choosing two models for two different situations in real life. We choose:
feature_array = $[1, Victim\_Race, Victim\_Ethnicity]$
when we have no knowledge about the perpetrator.
Otherwise, we choose:
feature_array = $[1, Victim\_Race, Victim\_Ethnicity, perpetrator\_Ethnicity]$
which indicates that we know the ethnicity of perpetrator.
First of all, the ACC of baseline is 0.865, which indicates that our baseline is a wise choice. After implementing the feature vector without knowledge of perpetrator information, the Logistic Regression model gives us the ACC is 0.868 which is the same as the result that SVM model gives us. Moreover, for the feature vector with knowledge about the perpetrator ethnicity, the ACC of Logistic Regression model is 0.874, which is also almost the same as the SVM model. Based on the results we get above, it is easy to calculate that the ACC of feature vector without knowledge of per-

Model \ Evaluation	ACC_all	ACC_white	ACC_black	ACC_native	ACC_asian
Baseline	0.865	0.871	0.829	0.992	0.988
logistic regression	0.868	0.878	0.831	0.991	0.988
SVM	0.868	0.877	0.830	0.993	0.987

Table 4: Perpetrator Race Prediction Evaluation for different models on train set with no knowledge of perpetrator

Model \ Evaluation	ACC_all	ACC_white	ACC_black	ACC_native	ACC_asian
Baseline	0.865	0.871	0.829	0.992	0.988
logistic regression	0.874	0.881	0.832	0.994	0.987
SVM	0.875	0.882	0.831	0.993	0.988

Table 5: Perpetrator Race Prediction Evaluation for different models on train set with knowledge of perpetrator

petrator information is about 0.003 higher than the ACC of baseline and the ACC of feature vector with knowledge of perpetrator information is about 0.009 higher than the ACC of baseline and 0.006 higher than the ACC of feature vector without knowledge of perpetrator information. This conclusion is reasonable, because the more information we have about the perpetrator, the much easier for us to predict the other unknown information about the perpetrator. And this is the reason why it is important for policeman to collect as much as evidence when they deal each incident. Beyond that, it is shown that the two models we choose give us the similar predictions. Moreover, from Table 4 we could see that our prediction models for Native American/Alaska Native and Asian/Pacific Islander who are the minority part of United States compared with White and Black is quite accurate. The ACC for Native and Asian/Pacific Islander in Logistic Regression model is about 0.991 and the ACC for Asian/Pacific Islander in Logistic Regression model is about 0.988. This could be explained by the fact that those minorities live in a smaller life circle compared with majorities. So the victim and perpetrator have closer connection with each other means higher correlation which could give us a more accurate prediction.

5.4 Future Scope

The datasets provided by FBI mostly contain basic information of victims and perpetrators. More features such as witnesses and suspects' testimony can be utilized to improve the accuracy of our model in the future. We could even further combine text prediction and feature prediction together to build a even better model. In linear model, we assume the features we use are independent. However, in real life, these features may depend on each other. We may consider to use a more complicated model like Neural Network to predict characteristics of perpetrator.

6. REFERENCES

[1] Cella, Matthew, and Alan Neuhauser. "Race and Homicide in America, by the Numbers." U.S. News & World Report. U.S. News & World Report, 29 Sept. 2016. Web. 09 Mar. 2017.

[2] Loeber, Rolf, Dustin Pardini, D. Lynn Homish, Evelyn H. Wei, Anne M. Crawford, David P. Farrington, Magda Stouthamer-Loeber, Judith Creemers, Steven A. Koehler, and Richard Rosenfeld. "The Prediction of Violence and Homicide in Young Men." Journal of Consulting and Clinical Psychology 73.6 (2005): 1074-088. Web.

[3] "Homicide Trends in the United States, 1980-2008." Bureau of Justice Statistics (BJS). N.p., n.d. Web. 09 Mar. 2017.

[4] Cunha, Olga Soares, and Rui Abrunhosa GonÃğalves. "Predictors of Intimate Partner Homicide in a Sample of Portuguese Male Domestic Offenders." Journal of Interpersonal Violence (2016): 088626051666230. Web.

[5] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013