Kenneth Tran and Minji Yoon
Prof McCauley
CSE190 – Data Mining
2 June 2015

CSE190 Assignment 2 Report

Our dataset is a comprehensive list of convicted criminals from the Texas Department of Justice. It consists of the names, gender, ethnicity, crime, and sentence for over 130,000 people. It is well-known that the US justice system is inconsistent with sentencing severity for the same crime; sentencing length tends to be dependent on the judge in most cases. While the situation has been improved since our current criminal justice system moved away from the indeterminate sentencing structure that gave judges full discretion in length of sentences, we were not sure that our criminal justice system has reached a completely equitable sentencing system yet. Especially when incorporating differences such as race and age, we believe that the inconsistencies will become even more exacerbated.

Data for crime can be interesting to search for. Due to the personal nature of the data, it is often times unavailable to the public. While information on certain types of criminals, such as sex offenders, are readily available for the good of the public, many states prefer to keep personal information under wraps as long as they do not present an immediate danger to the general public. Criminal data related to narcotic cases or ones that involved juveniles are also highly restricted. And there is a big need to protect the identity of the victim. Due to these restrictions, it was pretty difficult locating a large enough dataset that did not infringe on these boundaries. Texas was one of the few states that offered a comprehensive list of convicted criminals.

Our predictive task for this dataset is to calculate the approximate prison term of a person based on their racial background and crime. And to compare the sentencing length with other people's sentences regarding similar crimes. We also wanted to see if we could identify trends where certain demographic groups were less or more likely to be incarcerated from a conviction.

The use of data mining and predictive analytics in relation with criminal datasets is not new. In one study, a researcher at Oracle used data mining to identify and model crime patterns. Shyam Varan Nath used clustering to identify groups of similar kinds of crime based on geographic location, which can indicate a crime spree or be used to identify a crime pattern.

In a study similar to the premise of our predictive task – conducted by Professor Jeffery Walker, Richard Hartley, Sean Maddan, Amy VanHouten, and Gwen Ervin-McLarty – linear regression was used to study the sentencing disparities under the US sentencing guidelines. Using data regarding sentencing practices from Arkansas, the disparities were examined under the context of whether a prison sentence was given and the length of that sentence. While the study found that there was negligible evidence to suggest that "extra legal variables" such as race and age

had any influence on whether the offender was sentenced to jail or prison, they did find that offenders who were non-white were more likely to receive 2.5 months more in prison/jail. Their results also suggested that males tended to receive lengthier sentences than females. On the other hand, age seemed to have little significant effect.

To obtain our results, we first converted the files we obtained from Texas's Department of Justice into CSV format, cleaned out any columns with missing data fields, and then split the dataset. We used one half of the data as our scoring set and the other set as our training set. To analyze the data, we utilized linear regression in order to model the relationship between race, crime, and sentence time. We took the data set that was previously split and trained one half of the data while we took the other half and scored it based on the three listed features stated earlier. The trained data set ended up with results that showed how much a person's racial background played a role in the length of their assigned sentenced, which is shown in the figure below.

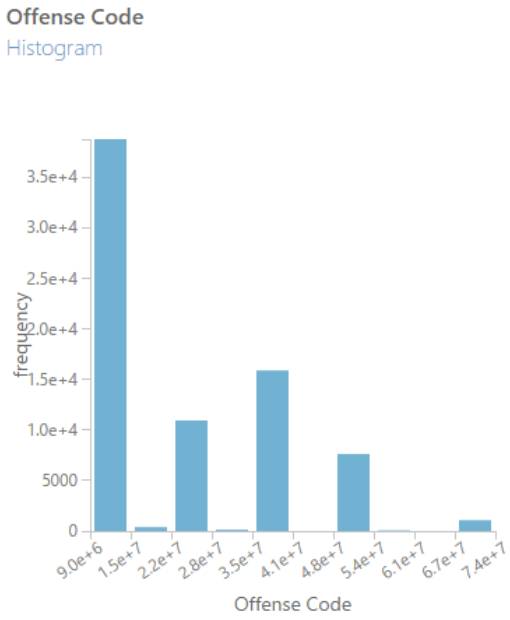
Feature	Weight
Bias	102.895
Race_O_4	41.167
Race_I_3	17.6771
Race_W_5	14.4103
Race_B_1	14.3048
Race_H_2	13.0895
Race_A_0	2.24618

The data in the chart on this page quantifies how much influence race can play in determining the duration of a person's punishment for their crime. The codes each represent a given race: W- White, which consists of anyone that have ancestral origins from Europe, North Africa, or the Middle East, B- Black, which represents the people who have origins or ties to the black racial group of Africa, H- Hispanic, which represents people who have Mexican or Latin American origins, I- for people of American Indian or Alaskan Native descent, A- for Asian or Pacific Islander peoples, and O, which is used to specify any of the numerous ethnicities not previously mentioned. According to our results, a person's racial background did have an impact on the terms of their sentencing. And having a Caucasian background had one of the strongest impacts (after O, which represents a mix of different ethnicities). Immediately after in level of impact are Blacks.

It is important to note that this analysis was the result of only one half of the available dataset – our training set. The result of our analysis on the other half of the dataset, which was used to be scored along with the trained data, are shown later in this report.

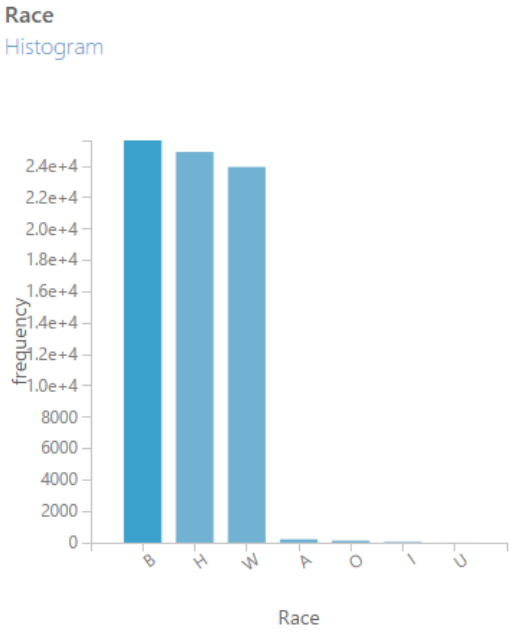
The chart on the next page, title "Offense Code", is a representation of the different offenses committed based on the legal crime code and the frequency of their occurrences. This analysis was done to see how often certain crimes appeared in our chosen dataset. This gives us a better idea of how much of the dataset we can use in our comparisons of sentence terms, since comparing the sentence length of petty theft in comparison to some violent crime would

not be very logical. As you can see from the results, crimes tend to be skewed into four major categories, with one (which represents theft) making up the majority of convicted crimes.



The next chart, labeled as “race”, shows the statistics for incarceration rates of different racial profiles. If you compare the values from the trained data vs. this chart, you can see that without the other factors such as the crime or the number of years sentenced, race code B, which represents the people of African descent had the highest frequency despite their race having a smaller weighted influence when it comes to their sentencing years for their crime.

The race code W, which represents white persons are the lowest of the top three race categories that make up the majority of the incarcerated population. This is interesting to note, since we saw that this same group of people had the largest impact on their sentencing.



Finally, we have our last chart below, labeled as “Sentence Years”. It is a representation of our predictions of the likely length a person would be imprisoned based on the two parts of data given above. As you can see from the chart, the predictions for the most likely amount of years you’re sentenced lies between 5-10 years, which makes sense considering the most frequent crime category is theft.