Predicting US Federal Court Case Outcomes

BrainStation Capstone Project | Jason Taylor | 11APR2023

Business Problem Definition (Intro)

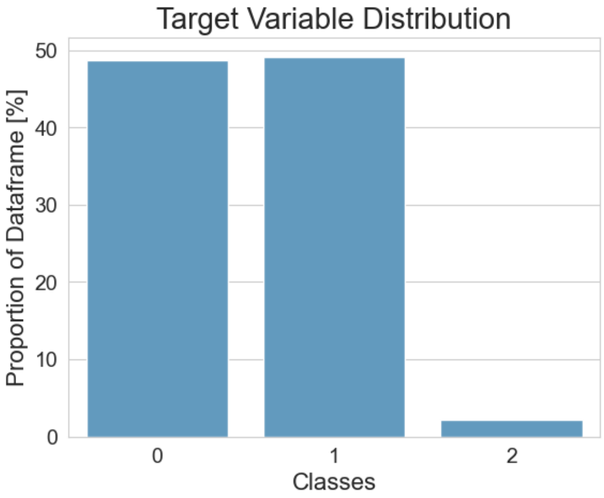
It is widely accepted in psychology and cognitive science that humans are susceptible to biases which can affect their decision-making processes[1]. These biases can be influenced by a variety of factors, such as personal experiences, beliefs, and social pressures, among others[2]. Ideally, the justice system is above these factors, providing a fair and unbiased platform upon which to arrive at a defendants' sentence. This concept is key as the justice system is a cornerstone of our society. The wide range of factors that could potentially influence sentencing unfairly make this an ideal project for machine learning models as we may uncover complex and otherwise hidden patterns in judges' behavior. This project applies machine learning models to data from US federal court cases to examine the ability of judges to separate themselves from bias when passing sentences.

Data Acquisition, Quality and Completeness

Court case data in the U.S. is, by law, publicly available information. However this information is distributed across multiple sources and can require specific tools, knowledge and considerable effort to effectively pull together the full picture. In October 2020, a group of researchers created 'JUSTFAIR': a free database of sentencing decisions that collates information from multiple public sources[3]. This collated database was used as the dataset for this project.

Preprocessing, Exploration and Analysis

At the time of sentencing, judges are provided with a minimum and maximum sentence guideline. These guidelines are complex variables taking into account a large number of factors related to the case. The court may depart from this range if they wish, giving a sentence either above or below the guideline range[4]. This is the basis for our target variable. The sentence of each case was classified as either below (lenient), within (standard), or above (harsh) the guidelines, thus forming our class 0, 1 and 2 target variables respectively. It was found that this target variable was highly imbalanced with only 2.2% of all cases being sentenced above the maximum guideline (class 2) while class 0 and 1 were very similar at 48.6% and 49.1% respectively.



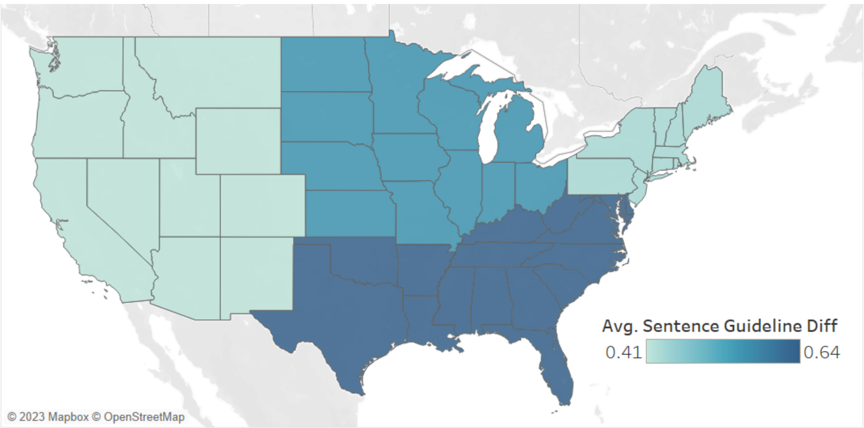
[1] https://www.sciencedirect.com/topics/neuroscience/cognitive-bias

[2] https://www.techtarget.com/searchenterpriseai/definition/cognitive-bias

[3] https://qsideinstitute.org/research/criminal-justice/justfair/

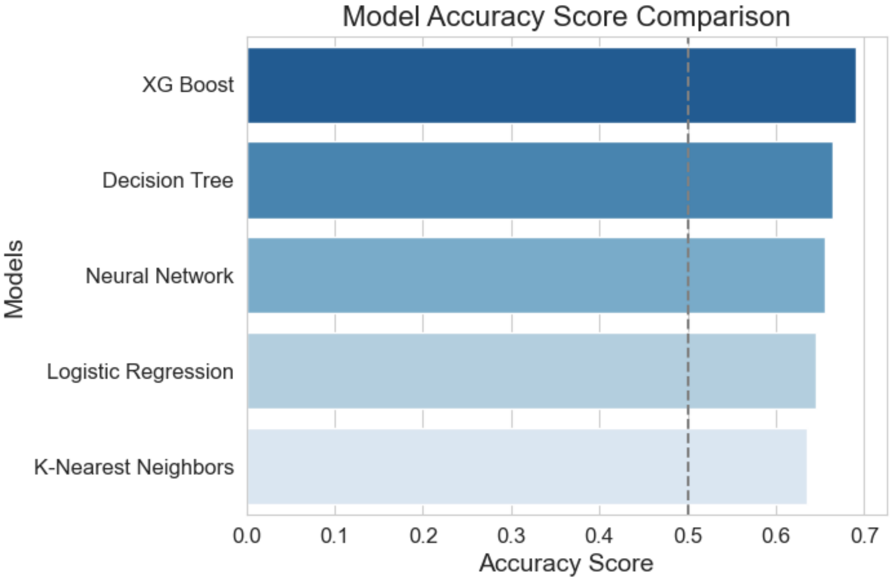
[4] https://www.ussc.gov/sites/default/files/pdf/about/overview/Overview_Federal_Sentencing_Guidelines.pdf

During Exploratory Data Analysis (EDA) it was found that the relative occurrences of cases that were sentenced outside of the guidelines increased relatively linearly from 2003 till 2014, suggesting year may be a strong predictor. It was also found that the region of the case may be a strong predictor with southern states having an average target variable value of 0.64 compared to western states with an average value of only 0.41 as can be seen in the map below.

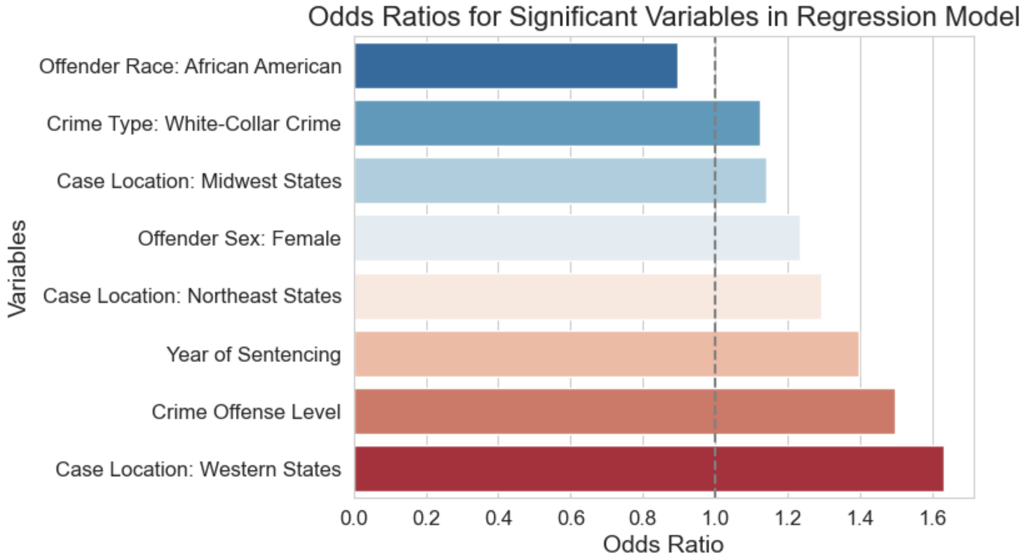


Modeling and Model Evaluation

As previously discussed, our engineered target variable is highly imbalanced with very few class 2 instances. Numerous balancing techniques were applied to try and resolve any patterns existing in this minority class however none were successful. Following this, our target class 2 was dropped leaving us with a binary problem. Thus our problem space shifted slightly to try and predict between lenient (class 0, below the minimum guideline) and standard cases (class 1, within the guidelines). The relative success of each applied model is shown below ranging from XG Boost with 69.1% accuracy down to K-Nearest Neighbors with 63.5% accuracy. These findings, although not impressive scores at face value, are significant. Recall that if there was no bias in the system, we would not be capable of predicting class 0 vs class 1 any better than random chance, resulting in accuracy scores of 50%. Thus we have found some level of bias in the system.



Following the success of these results, an attempt was made to isolate significant predictors to see if we are still able to predict our target classes. Note that for this study the target class is our lenient decisions. One of the benefits of a logistic regression model is we can calculate the odds ratio for each input variable. An odds ratio is a measure of the strength of a predictor where an odds ratio of 1 indicates no association with the target variable. A selection of 'significant' variables are shown below where the absolute values of the odds ratio minus 1 is greater than 0.1 (at least 0.1 away from 1 in either direction).



Technical note: Our categorical variables have been one hot encoded, thus the odds ratios for these variables is relative to our baseline (the variable(s) dropped). There is some loss of interpretability here as the variables were dropped based on solving multicollinearity within our dataframe. The odds ratio for numeric variables can be interpreted as the increase in likelihood of being a lenient decision per unit increase of the variable.

Recall that our target class are our lenient cases, thus all variables with a positive ratio are strong predictors of a lenient decision, whilst negative odds ratio variables are strong predictors of a standard decision. Passing only these variables back into a logistic regression model produces a model with 63.4% accuracy, a loss of only 1.1% compared to utilizing all of our variables. Thus these 8 variables are highly influential and mostly responsible for the bias seen in our data. Of note, we can see that having a case in the western states is highly predictive of a lenient decision whilst being an African American offender is highly predictive of a standard decision. Other categorical variables indicative of a lenient case can be seen in the figure. We also have 2 numeric variables, namely: The crime offense level and the year. This shows that as these variables increase, we are more likely to see a lenient decision. The association with year supports our finding from our EDA where decisions tended to diverge from the guidelines with time. We can also see that as our crime offense becomes more severe, judges are more likely to give a sentence under the minimum guideline. This may suggest that the judges' perception of cases begins to level off as cases become more severe whereas the guideline calculations may continue to increase.

Conclusion

This project aimed to investigate bias in the US federal justice system using machine learning models. It was found that sentence harshness can be predicted with 69.0% accuracy showing there is some bias in the system. Furthermore, we can isolate this bias down to 8 key variables with the highest influence on a judges' decision to predict sentence harshness with 63.4% accuracy.