Using Predictive Models to Determine Judge Bias in Asylum Refugee Court Cases

Ismail Mustafa
Courant Institute of
Mathematical Sciences
New York University
ismailmustafa@nyu.edu

Pavan Sudhir Nallam
Center for
Data Science
New York University
psn240@nyu.edu

Honor Pledge I pledge my honor that all the work described in this report is solely mine and that I have given credit to all third party resources that I have used.

Abstract

The purpose of this paper is to develop a predictive model for classifying whether or not a refugee is granted asylum in the United States, and to use that model to determine which features bias judges the most. The court case data used here is kindly provided to us by Daniel L. Chen, and examines asylum court decisions across 426 immigration judges over a period of 42 years from 1971 to 2013. Other data we use are NBA, MLB, NHL, and NFL sports data, weather data from all cities court decisions are made, and mood data analyzed from tweets across eight cities in the United states over the span of one year.

We focus on two key aspects we believe could cause bias in decisions, the outcome of sports games in cities associated with each specific judge, and weather at the time of the decision. Using this data, we develop a hidden Markov model for the Judge's mood, and pass this into a decision tree classifier to determine feature importance. Our results indicate that weather and sports have no observable effect on the decisions made in these cases, however we do find that the judge's past five decisions, where the judge is licensed, and the nationality of the asylum seeker, improve classification score indicating that there is bias in the way decisions are made. We attain a maximum train and test score of 78% using a gradient boosting technique.

1 Introduction

1.1 Background

Ideally, all court cases conducted in the United States should be free from bias. Unfortunately this is not the case for asylum court cases. According to Ramji-Nogales et al. in the Stanford Law Review (Refugee Roulette: Disparities in Asylum Adjudication,)

"... the most important moment in an asylum case is the instant in which a clerk randomly assigns an application to a particular asylum officer or immigration judge."

The very fact that different judges' decisions would vary wildly on the same case shows that there is inherent bias in the system. Additionally, Fatma E. Marouf in the New England Law review (Implicit Bias and Immigration Courts,) argues that the very nature of the job promotes implicit bias due to reasons such as

"... lack of independence, limited opportunity for deliberate thinking, low motivation, complex caseload, and the low risk of review..."

1.2 Problem Definition

The bias present in asylum court cases is a well known and well documented phenomenon. Our aim is to determine what features specifically tend to cause the most bias in these decisions. We look at features from as general as the nationality of the asylum seeker as well as their gender, to features as granular as the time of day the decision is made and whether or not the decision is written. We also take into account the weather conditions where the decision is being made that could have an effect on the judge's mood.

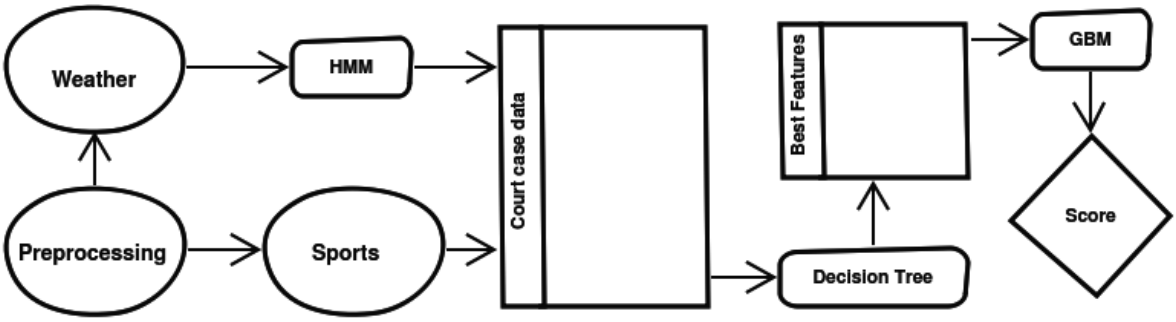

Figure 1: Approach

Additionally, we look at sports teams that could be related to the judge through their hometown, college town, or town of residence to see if the percentage of wins vs. loses could affect the way the judge is feeling that day.

Using these features, we hope to develop a model that tells us that that a judge's decision is indeed biased by the weather and outcome of sports games. We also intend to determine the most important features that affect a Judge's decision.

2 Related Work

There are many studies out there that look at bias in asylum court cases, however most of them do not have access to as much data as we do. For one, Ramji-Nogales et al. in the Stanford Law Review (Refugee Roulette: Disparities in Asylum Adjudication,) only uses data during the fiscal years 1999 - 2005, a time-span of 16 years, 26 years worth of data less than we are provided with.

Additionally studies such as Fatma E. Marouf's in the New England Law review (Implicit Bias and Immigration Courts,) studies the data in a much more qualitative manner as opposed to running algorithms to determine the most important features. In this paper, we aim to differentiate ourselves by using machine learning classifiers as a method of determining biasing features.

3 Data

There are four main types of data used:

- Court case data provided by Daniel L. Chen examining asylum court decisions across 426 immigration judges over a period of 42 years from 1971 to 2013. There is a lot of missing values associated with this data-set. Missing values are handled in the paper using surrogate splits and imputing. We combine this with biographical data provided about each judge to produce a single file containing 134 features.

- National Oceanic and Atmospheric Administration historical weather data over the same time period and across all cities in which asylum decisions were recorded. Features of interest include precipitation, snowfall, snow depth, max temperature, min temperature and minutes of sun for every day, five days before the decision.

- Twitter mood data calculated over the period of one year across eight different cities.

- Historical sports data across the four major sports leagues in the United States (NBA, MLB, NHL, NFL).

4 Approach

Our approach is highlighted in Figure 1. We begin by pre-processing all the sports and weather data to extract only the information that applies to the specific judge as well as the date on which the court case decision falls. We then use our data-set of twitter mood data as well as weather data to create a hidden Markov model in which the observable variables are the current state of the whether using a metric we define, and the hidden variables are the mood of the judge on the day the decision is made. We then use this mood to calculate the missing mood values for all other court case decisions.

Concurrently, we use the hometown, college town, and town of residence of the judge to parse through the sports data and extract teams that we believe are relevant to the judge. We then use the ratio of wins vs. total games played as a feature.

Next, we take the sports and mood features as well as all other features extracted from the raw data and pass them into a baseline decision tree classifier. We use this classifier to determine the most important features. We then use these features on a more robust classification algorithm, Gradient Boosting Machine (GBM) in order to improve the score on our important features.

4.1 Pre-processing

4.1.1 Sports Data

The four sports we consider are, MLB, the NBA, the NFL, and the NHL. For each court case decision, we look at the Judge in question as well as their hometown, their college town, and the town in which they currently reside. Using these variables, we then look up for each type of sport, any games that were played in the past five days of the date on which the judge has made their decision. For all of these games, each win is taken as a 1, and each loss as a 0. We then determine the sports score by diving the number of wins by the total number of games played. The equation for this metric is shown below:

$$sportsScore = \frac{games_{won}}{games_{played}}$$

4.1.2 Weather Data

The weather data was supplied to us by Daniel L. Chen and was taken from the National Oceanic and Atmospheric Administration (NOAA). In processing this data, we looked at the location in which the judge made the specific court case decision and the date of the decision.

We then looked at the weather in the past five days of the date of the decision and picked four features for each day. These four features are shown in Table 1. $tavg$ is simply the average of $tmax$ and $tmin$. The weather score was then calculated using the equation shown below. The maximum values for $tavg$ and $prcp$ are determined from the entire data-set.

This equation was developed iteratively in order to give the most intuitive results for emission probabilities when training our hidden Markov model.

$$weatherScore = \frac{tavg}{tavg_{max}} - 4 \times \frac{prcp}{prcp_{max}}$$

4.2 Hidden Markov Model

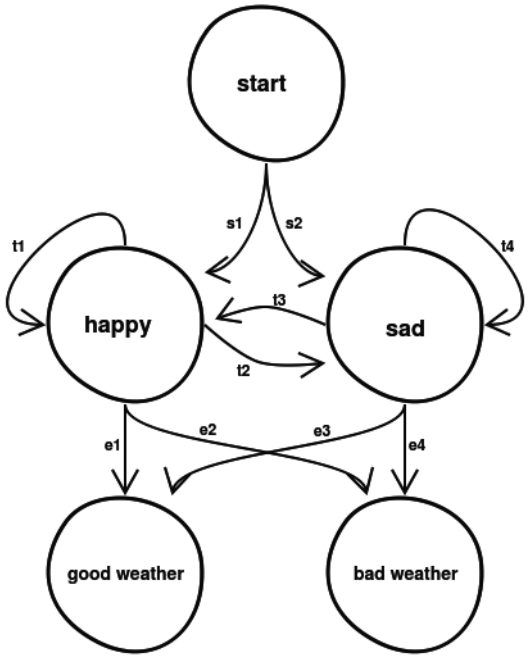

Figure 2: HMM used to model weather and mood.

The hidden Markov model (HMM) we developed uses the weather score metric we developed in the previous section as the observable variables, and the mood of the judge as the latent variables. The HMM was trained using twitter mood data calculated over the period of one year in eight different cities. By matching the city associated with each mood in the court data, we ended up with 10,000 usable court case decisions. The 10,000 usable court case decisions were split 50/50 into a training and testing set in order to train and evaluate the HMM.

Any mood data values that are above 6.01 are taken to be "happy", and the rest are taken to be "sad". The threshold of 6.01 was found iteratively as the optimal value.

The emission probabilities are denoted as e in Figure 2 and are calculated as follows:

prcp	precipitation in 10ths of mm
tmax	highest temperature in 10th of Celcius
tmin	lowest temperature in 10ths of Celcius
tavg	average of tmax and tmin

Table 1: Weather Values

e1	0.612
e2	0.388
e3	0.3
e4	0.7

Table 2: Emission Probabilities

$$e = \frac{P(latent, observed)}{P(latent)}$$

Our emission probabilities came out to the values shown in Table 2. These values match our intuition. $e1$ is the probability that the happy latent variable "emits" the good weather observed state. $e2$ is the probability that the happy latent variable "emits" the bad weather observed state. $e1$ is larger than $e2$ which is intuitively correct. The same goes for $e3$ and $e4$ where the probability values are inverted since the sad latent variable is emitting them.

The transition probabilities were optimized to produce the highest training score. The optimal transition probabilities are shown in Table 3. Using these transition probabilities, we get a test score of 96.8% when evaluating the HMM.

Now that both the emission and transition probability matrices are calculated, we use the HMM to calculate the mood state that the judges are in as the decision is made. Since we are using an HMM, we want to take the weather observations from previous days up until the day in which the decision is being made. We do this by considering the weather from the past five days. As an example, a sequence of weather scores from the past five days would look like this:

t1	0.8
t2	0.2
t3	0.2
t4	0.8

Table 3: Transition Probabilities

$$[0.33, 0.25, 0.68, 0.65, 0.23]$$

Where the first value is the weather score four days before the day of the decision and the last value in the list is the weather score on the day of the decision. Converting these into observations, we get the following:

$$[b, b, g, g, b]$$

Where b is bad weather, and g is good weather. To get these values, we took any weather score above 0.5 to be indicative of good weather, and anything below 0.5 to represent bad weather. Passing this particular sequence into the Viterbi algorithm using the transition and emission probabilities we calculated, we get the following sequence of latent variables:

$$[sad, sad, sad, happy, sad]$$

Where we find that the judge's mood on this particular day given the previous four days is sad. Figure 3 shows a visualization of the HMM sequence generated by the Viterbi algorithm.

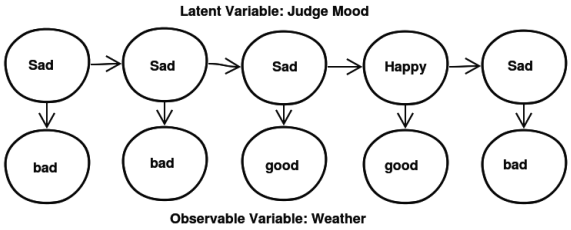

Figure 3: Generated HMM sequence.

4.3 Feature Selection and Baseline

Once we calculate the mood feature based on the weather as well as the sports features based on win/loss percentage, we pass those features in a baseline decision tree classifier model along with tens of other features available to us from the court data. The decision tree classifier we use is implemented using R. In order to handle missing values, we use surrogate splits. From this baseline implementation, we get a classification score of 60%. Additionally, we determine the most important features to be those shown in Table 4.

The feature with the highest weighting was found to be $numgrant\_prev5$ which is the number of times the judge decided in favor of the asylum seeker out of the judge's last five court cases. The next highest was bar which is the state

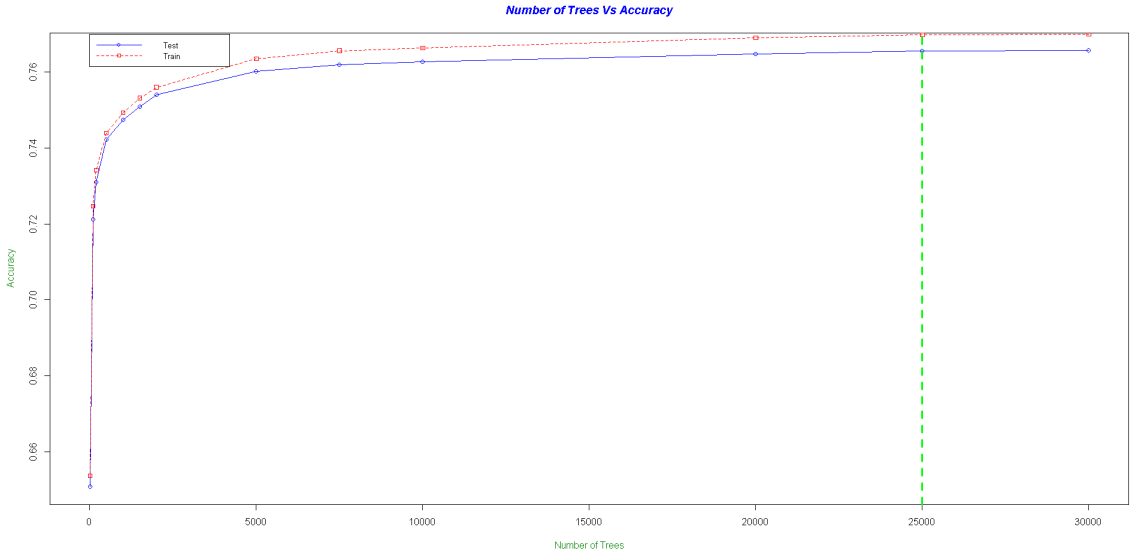

Figure 4: Gradient boosting algorithm run on important features.

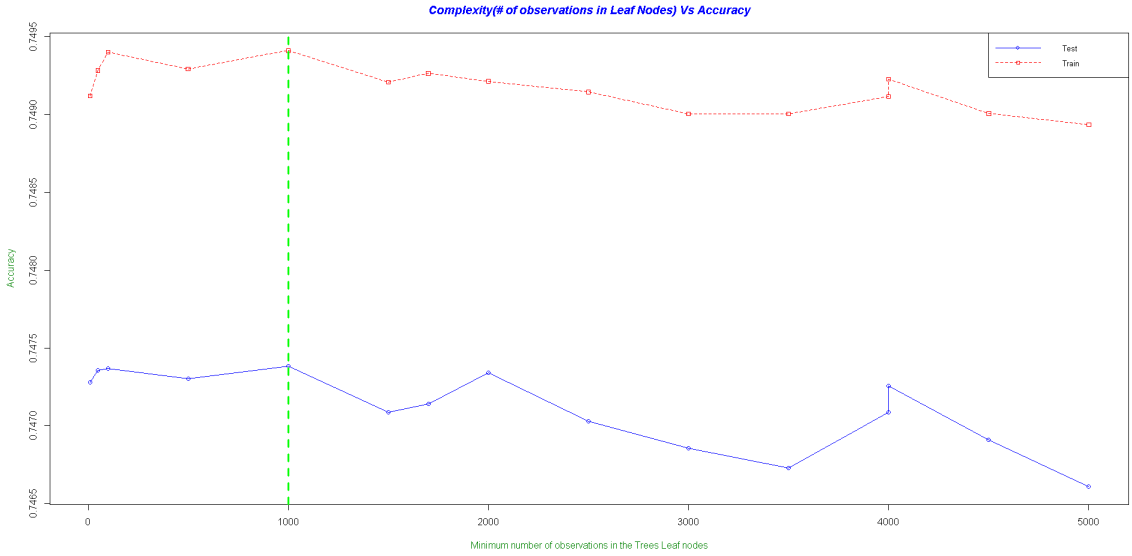

Figure 5: Number of observations in leaf nodes vs. accuracy for GBM.

Feature	Weight
numgrant_prev5	0.20142874
bar	0.13039
natid	0.110319
numcourtgrant_prev5	0.08797
natdefcode	0.0594
numcourtgrantother_prev5	0.05337
numcases_court_hearing	0.051319
lawyer	0.024951

Table 4: Most Important Features

in which the judge is licensed to practice. The third most important feature is $natid$ which is the nationality of the asylum seeker. Looking at these three features alone, we see that a judge's decision is strongly biased by the nationality of the asylum seeker, the state in which the judge lives, and the grant percentage of the judge.

The other features are all variations on the top three features. For example, $numcourtgrantother\_prev5$ is the number of times other judges in the same court as the current judge ruled in favor of the asylum seeker in the past five decisions. This shows that all judges in the same courthouse tend to bias in the same way. Additionally, $natdefcode$ is a higher order combination of the nationality of the asylum seeker, and whether or not the case was defensive.

The lowest weighted feature that is still significant is whether or not a lawyer is present. This makes sense intuitively as having a lawyer can boost your odds of winning a court case. Additionally, it appears as though from this analysis of feature importance, weather and sports have no observable influence on classifying the outcome of court cases as their weights came out to zero.

4.4 Gradient Boosting Machine

After determining the most important features, we use a gradient boosting algorithm to improve the score of our classification. By applying the GBM algorithm on only the most important features, we approach a score of 78%. We have implemented GBM in R, which handles missing data by creating surrogate splits.

4.4.1 Tuning GBM Parameters

We have used number of trees (steps), number of observations in leaf node and shrinkage (step size) as the GBM parameters to tune.

Figure 4 shows the score of GBM vs. the number of trees it was run on. The green line marks the point at which the score tends to plateau which occurs around 25,000 trees. Figure 5 shows complexity which is a way of "regularizing" GBM, by limiting the number of observations in leaf nodes. The green line shows that the score is maximized at a complexity of 1000. Figure 6 shows the shrinkage (step size) of 0.1 at which accuracy score reaches maximum.

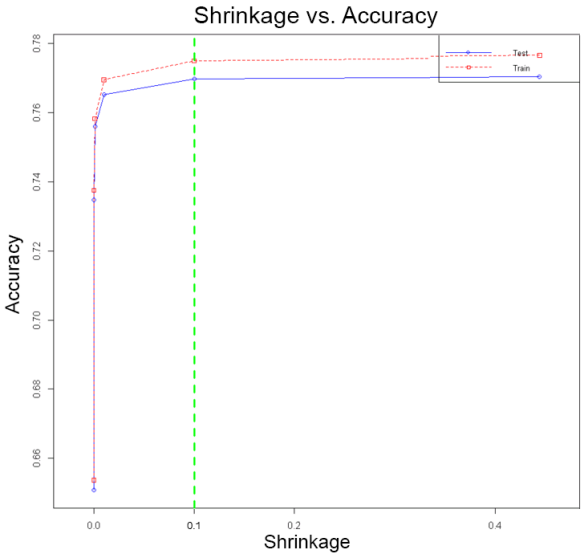

Figure 6: Shrinkage vs. accuracy for GBM.

4.5 Random Forest

In an attempt to improve the score further past 78% we used a random forest regressor. The data was imputed using the mean of all available values. Unfortunately, the score plateaus at around 35%. The algorithm was only run up to 140 trees.

Figure 7 shows the random forest score vs. the number of trees generated. The score plateaus at around 35% which is around 60 trees.

4.6 Unique Features

Other features not included in the data-set were used in an attempt to boost the score. Unfortunately many of them had no observable effect on

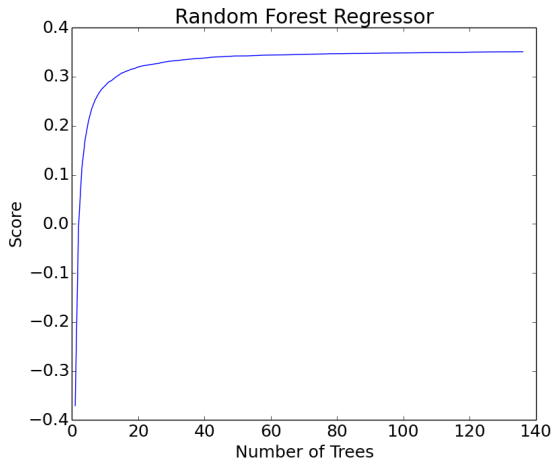

Figure 7: Random forest regressor trained on up to 140 trees.

the final score. many of these features are mentioned below.

4.6.1 State Mood Data

In a paper by Mitchell et. al. titled "The Geography of Happiness" (The Geography of Happiness,), they use a larger amount of data to classify the mood in all 50 states. At the end of their paper, they list the mood in all 50 states. One approach we tried was to use all 50 of these values, and take the mood of the city in which the court decision is being made to be equal to the mood of that state. While we were hopeful there would be some correlation since these results were more direct than being passed through an HMM, there was no statistically significant difference in classification score.

4.6.2 Events of September 11th

For this feature, we assume that the events of September 11th caused immigration judges to be more wary about who is let into the country. The assumption was that less grants would be given after September 11th than before. A binary feature was created in which any court cases after September 11th were set to 1, and the rest were set to 0. No statistically significant difference in score resulted from this.

4.6.3 Weather Score without HMM

The raw observed variable weather score was passed in as a feature. The idea here was that the HMM obfuscates the actual weather result. However, including the weather score alone does not seem to make a difference.

Classifier	Score
Decision Tree	60%
GBM	78%
Random Forest	35%

Table 5: Most Important Features

4.7 Results

There were two main goals of this paper. The first was to determine if the outcome of sports games, and the weather on the day of the decision had an appreciable effect on the judge's decision. The second goal to determine the most important features.

Regarding the first goal, including sports and weather data as features did not cause an increase in the score of our baseline classifier. In fact, when analyzing feature relevance, the sports and weather metrics we used had zero weighting.

The second goal was more successful in that we were able to determine three key features that cause the greatest biasing of judges. These features include the grant ratio accounting for the judge's previous five decisions, the state in which the judge is licensed, and the nationality of the asylum seeker.

Table 5 shows final scores of the different classifiers.

5 Conclusion, Future Work and Challenges

One direction we would like to take in the future is to gain access to more twitter mood data. We feel as though the limited amount of data points inflated our HMM score to an unrealistically high level. Additionally, we would like to look at other forms of mood data that are perhaps not related to twitter since we feel as though the mood of twitter could potentially not be indicative of the mood of immigration judges. This could be court case transcripts in which we could apply sentiment analysis in order to ascertain the mood of the judge.

Given more time, we would also like to increase the number of observable variables in the HMM to include sports data. This would be dependent on finding a larger mood data-set since the reason

it wasn't included this time was due to limited data.

We would also like to improve on the HMM we developed. We would like to reformulate the HMM so that the latent variables are weather, and the observable variables are the decision the judge makes. In doing so, we could have even more data to work with rather than being limited by the small data set of twitter data we had. Additionally, with this model, we could tune the HMM in order to tell us which weather features are most important in order to use in the classifier.

Acknowledgments

We are grateful to our adviser, Daniel L. Chen for his guidance and advice during the project. We would also like to thank our professor David Rosenberg for helpful advice along the way.

Source Code

The source code is available at https://github.com/ismailmustafa/predictingRefugeeAsylum-avengers.

References

"Refugee Roulette: Disparities in Asylum Adjudication" Stanford Law Review. Volume 60, Issue 2. Page 295

Marouf, Fatma, "Implicit Bias and Immigration Courts" (2011). Scholarly Works. Paper 787. http://scholars.law.unlv.edu/facpub/787

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place." PLoS ONE 8.5 (2013): n. pag. Web.