Can Machine Learning Help Predict the Outcome of Asylum Adjudications?

Daniel L. Chen
Toulouse School of Economics
daniel.chen@iast.fr

Jess Eagel
New York University CIMS
jae383@nyu.edu

ABSTRACT

In this study, we analyzed 492,903 asylum hearings from 336 different hearing locations, rendered by 441 unique judges over a 32 year period from 1981-2013. We define the problem of asylum adjudication prediction as a binary classification task, and using the random forest method developed by Breiman [1], we predict 27 years of refugee decisions. Using only data available up to the decision date, our model correctly classifies 82 percent of all refugee cases by 2013. Our empirical analysis suggests that decision makers exhibit a fair degree of autocorrelation in their rulings, and extraneous factors such as, news and the local weather may be impacting the fate of an asylum seeker. Surprisingly, granting asylum is predominantly driven by trend features and judicial characteristics- features that may seem unfair- and roughly one third-driven by case information, news events, and court information.

CCS CONCEPTS

• **Computing methodologies** → *Classification and regression trees*;
• **Applied computing** → *Law*;

KEYWORDS

Legal Prediction, Refugee, Machine Learning, Data Science

ACM Reference format:
Daniel L. Chen and Jess Eagel. 2017. Can Machine Learning Help Predict the Outcome of Asylum Adjudications?. In *Proceedings of ICAIL '17, London, United Kingdom, June 12-16, 2017,* 4 pages.
https://doi.org/10.1145/3086512.3086538

1 INTRODUCTION

We like to believe that the legal system defends human and civil rights while promoting equality and fairness. In this paper we detail one such area, the asylum adjudication process, where such impartiality may be less than what one might hope for or expect. Specifically, our goal was to show that the outcome of asylum proceedings is predictable from a set of known variables. Strikingly, historical trends of the judge's decisions contribute a great degree to prediction, and this autocorrelation could proxy for learning, habit formation, or tastes.

We begin by outlining the asylum adjudication process and the raw data files used in our study. Interesting patterns emerge related to whether judges become harsher before lunchtime or the end of the day [8], how family size is associated with grant rates, and how the day's caseload is associated with grant rates. These correlations are novel since this data is new and has not been examined other than by some prior papers by one of the authors that focused narrowly on specific questions of casual inference [4] [2], and by another that considers a behavioral question about judges' choice to acquire information[3].

By 2013, using data only available up to the date of the trial, our model accurately predicts 82% of asylum hearing outcomes. We show that approximately 40% of the misclassified hearings can be attributed to one nationality in a single court during the early 2000s, which reveals the presence of a major historical event not accounted for in the feature set. We conclude by offering additional areas for further research.

2 THE ASYLUM PROCESS AND DATASETS

An individual may apply for refugee status in the United States either affirmatively or defensively. Affirmative asylum applicants voluntarily identify themselves to the Department of Homeland Security. Defensive applicants are those who have been placed in removal proceedings by the DHS [9]. The details of the full asylum process are beyond the scope of this paper, as we are focusing on only those applicants who make it into the refugee court system. These applicants are randomly assigned to judges across the country to have their case heard, and ultimately this justice determines whether or not the individual or family shall remain in the country.

2.1 Datasets and Preprocessing

Our main dataset originates from the Transactional Records Access Clearinghouse (TRAC). We combined the TRAC dataset with data from NOAA [6] and Bloomberg [7]. Taken together, the final fully merged set contained approximately 500,000 cases and 137 features. We classified each feature into 1 of 6 buckets: case information, court information, judge information, news, trend, or weather.

2.1.1 Case information. Generally speaking, we have some intuition about the relevance of the case-centric factors. Among the twenty-two case information variables, were nationality, number of family members, date of hearing, and whether the application was affirmative or defensive. The affirmative/defensive speaks directly to the refugee's reason for immigration.

2.1.2 Court and judge information and trend. We also integrated 19 features, such as law school graduation year and gender, for 441 judges. The judge feature space included the President whom they were appointed by, whether or not they served in the military,

ICAIL '17, June 12-16, 2017, London, United Kingdom

ACM ISBN 978-1-4503-4891-1/17/06…$15.00
https://doi.org/10.1145/3086512.3086538

and experience years. The court information had seven features including the court ID and the number of hearings per day. Included in the court and judge feature space are 17 historical factors, which are meant to capture any time varying component in the ideology of a specific hearing location or justice.

2.1.3 Weather and news. We integrated a time series of weather statistics, from NOAA [6], for each hearing location. Six weather features are embedded in the feature matrix. Additionally, we hypothesized that current events and media coverage may weigh on a justice's consciousness when ruling. To this end, we computed the most frequently used words from the Wikipedia page for 'refugee'. Our goal was to proxy for the general security situation of asylum seekers at the time of the trial. Bloomberg [7] Trends provides daily reports on the volume of specific words across a host of multinational news sources. Through their API, we scraped thousands of news outlets and amassed a time-series of the frequency of our keywords. We regularized each feature on a rolling basis using historical z-scores before mapping them into the final feature space.

2.1.4 Missing data and dummy variables. The fully merged data set was rife with missing and placeholder values. For context, 80% of the cases in the original asylum data file were missing at least one feature. We and introduced 'dummy' variables and 'dummy indicators to the space [5] by replacing missing values with a known constant and simultaneously created a binary flag feature, which indicates whether a variable had been dummied.

3 DATA CHARACTERISTICS

Figure 1 illustrates some observable patterns in our case-centric feature matrix. The top-left plot depicts the average grant rate versus the start time of the hearing. Curiously, two periods, just prior to lunch and just before the end of the day reveal noticeable spikes in the mean grant rate. The top right bar graph supports the claim that a refugee case heard earlier in the day is less likely to be granted asylum than one heard later in the day.

Family size also exhibits a non-random pattern. For instance, the chance a family of four being granted asylum is 30% higher than for an individual and 100% greater than a family of eight. Perhaps less surprisingly, defensive applicants are 50% more likely to be granted asylum than affirmative applicants.

An analysis of the judge feature space reveals similar non-random patterns, shown in figure 2. The number of hearings per day for a given judge versus the average grant rate appears to exhibit a Poisson-like distribution. Female judges had an average grant rate of 45% compared to males, which had just a 30% grant rate.

At stark contrast with our intuition, there does appear to be some correlation between the weather and the average grant rate. The left-most plot in figure 3 shows the average grant rate versus the maximum temperature reading (in tenths of degrees Celsius) on the date of the hearing. Extreme weather, in either direction, may be impacting the decision to deny or grant an applicant.

The middle plot in figure 3 illustrates the increased likelihood of a refugee being granted asylum conditional on the previous five decisions. The right-most chart in figure 3 speaks to the heart of the model we propose. It is clear that the grant-deny ratio is not independent of time. In the following section, we propose a fully predictive model that takes into account only the data available up to the date of the trial to calibrate the parameter set.

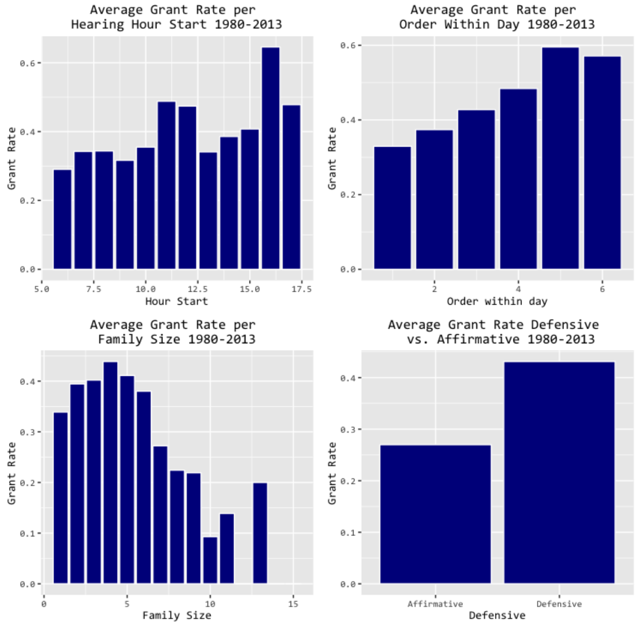

Figure 1: Case Information Charts

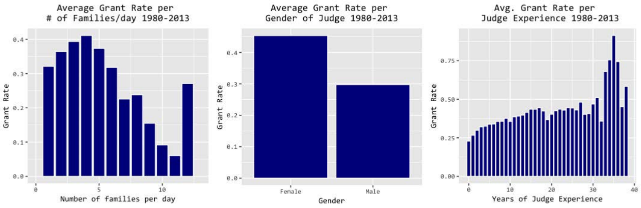

Figure 2: Court Information Charts

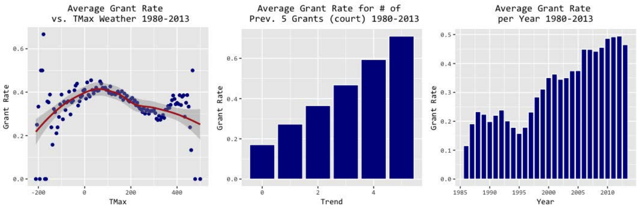

Figure 3: Trend, News, Weather Information Charts

4 PREDICTIVE MODELS

Our dataset contains approximately a 3-1 ratio of deny-to-grant applicants, which will serve as a baseline classifier for our statistical methods. To calibrate the time series models, we trained our parameter set on all asylum cases up to December 31^{st} of the prior calendar year. We used this parameter set to make predictions on all the incoming cases for the following twelve months.

Random Forests

Random forests is an ensemble method of a set of decision trees that grows in randomly selected sub-spaces. The trees are grown from a bootstrapped training set of size N. For a classification problem with p features, $\sqrt{p}$ features are used in each split in order to reduce the variance of the estimator. Typically, trees are grown to the largest extent possible with no pruning. However, due to computational hurdles we stop growing our trees when there were twenty-five samples in a leaf-node. We also stipulated that 1000 estimators were grown at each calibration stage.

The overall accuracy of the Random Forest reached 82% by 2013 (figure 4). In the error analysis section we contend that the meaningful performance dip in the mid 2000's is a function of two feature variables that might have some historical context.

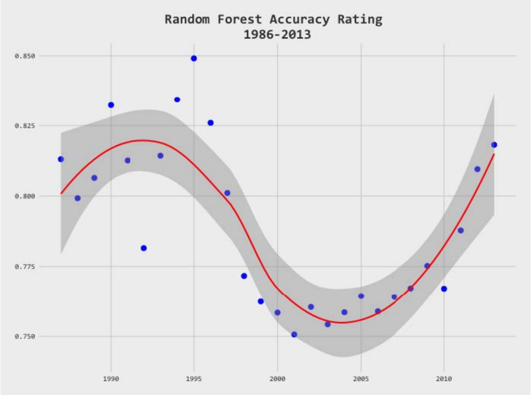

Figure 4: Random Forest Performance 1986-2013

In table 1[1], we show the relative weightings in our feature space at the end of 2012. It is easy to see that the trend factors gather significant weight in our test set, amassing 49% of the total importance. The second largest contributor was the case-centric information followed by judge information. The significant weight on trending features echoes our analysis in the previous section in figure 3. Moreover, the number of cases heard by a judge on any given day amassed 1.4% weighting in the random forest, which corroborates our finding in the left-most plot of figure 2.

Despite showing a promising correlation in our initial assessment of the data, the weather features were unable to garner meaningful weight in our random forest. We suspect that this is due to co-linear relationships with other features. The weather data was expressed in absolute degrees, not deviation from the mean. Therefore, the temperature was already embedded in other feature variables such as 'hearing location' or 'zip code'. Had the temperatures been expressed in z-scores, we may have been able to conclude whether or not a judge's verdict was influenced by extreme weather. Nevertheless, with an 82% accuracy, the random forest approach demonstrated significant improvement over the stated baseline.

Table 1: Random Forest Final Importances

Category	Feature	Weight
Case Information	Attorney ID	0.01
	Court ID	0.01
	Defensive	0.01
	Lawyer	0.02
	Nationality	0.024
	Order in day	0.002
	Other	0.12
	Total Case	**0.20**
Court Information	Hearing Location	0.01
	Other	0.06
	Total Court	**0.07**
Judge Information	College	0.007
	Judge ID	0.007
	Experience	0.006
	Male/Female	0.004
	Law School	0.007
	# of Cases	0.014
	Other	0.070
	Total Judge	**0.11**
News Trends	Crisis	0.006
	Genocide	0.006
	Other	0.026
	Total News	**0.07**
Trend Features	Judge Avg. grant	0.179
	Previous five	0.058
	Other	0.255
	Total Trend	**0.49**
Weather	Cloud Coverage	0.004
	Snow	0.001
	Other	0.017
	Total Weather	**0.02**

5 ERROR ANALYSIS

After each iteration of the random forest, we logged the incorrect classifications. In table 2 we detail our confusion matrix and the breakdown of errors. On an absolute basis, we mis-classified denied applicants one and half times more than granted applicants. Normalizing for the amount of actual grants versus denies, we performed better on granted applicants than denied.

Table 2: Confusion Matrix

Confusion Matrix	Actual Grant	Actual Deny	Total
Predict Grant	94,465	78,067	**172,532**
Predict Deny	30,009	290,362	**320,371**
Total	**124,474**	**368,429**	**492,903**

Figure 5 shows the misclassified grants and denies over the time series. Our model performs very poorly on actual granted applicants early on, however, the accuracy rate for each error converges gradually overtime. We consider this evidence that our model is 'learning' more about the feature spaces as time progresses. However, the model consistently regresses in its ability to forecast denied applications.

[1] A full table of the variable names and definitions is available in the appendix at https://www.tse-fr.eu/fr/publications/can-machine-learning-help-predict-outcome-asylum-adjudications

Another take away from our error analysis was the concentration of misclassified refugees during the early-2000s. Approximately 40% of our errors were unique to one nationality, *natid 44*, in one court ID, *courtid = 34*, at one hearing location, *hearingloc = 173*. Nationality ID *44* is Zaire, which is now known as the Democratic Republic of the Congo.

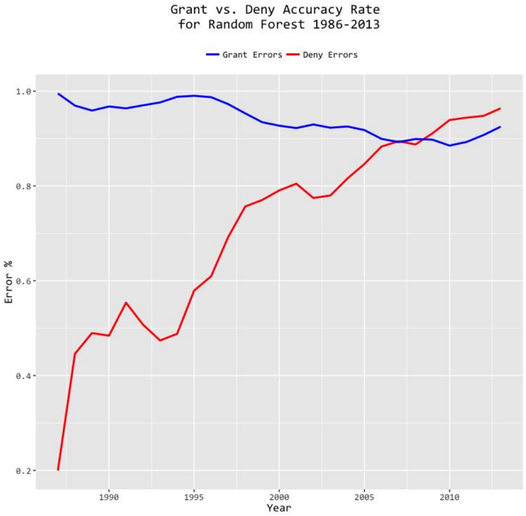

Figure 5: Grant vs. Deny Errors 1986-2013

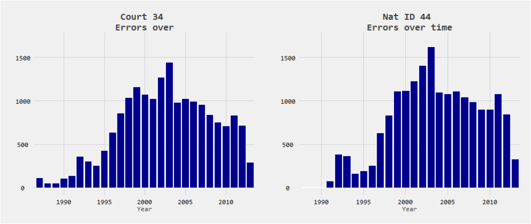

Figure 6: Errors for court 34 and nationality 44

The Second Congo War began in 1998 and ended in July 2003, perhaps putting some historical context to our errors. While we do not have a concrete name for court 34, these errors correlate highly with location 173, which is New York City.

6 A FULLY PREDICTIVE MODEL

In the feature set we outlined, there were a few features that gathered significant weight. The trend components carried 49% of the weighting in the final feature set. A few of these features were forward looking, such as the judge average grant variable (the average grant was always calculated excluding the current decision, but included future decisions). In one final iteration, we re-ran our random forest algorithm on a dataset devoid of forward looking trend features. This model produced a 79% accuracy rating on average over the time series. Table 3 highlights the change in the weightings for each category.

After removing all the forward looking trend components the case-centric features become more pronounced. Nationality accounts for 10% of the final feature weightings, which is five times more than its original weight. Despite removing the forward looking trending features, other time sensitive variables still amass significant weighting. The number of cases granted asylum out of the previous five decisions by the judge and the number of cases granted asylum out of the previous five decisions at the court account for a 9% and 3% weighting, respectively.

Table 3: Delta Random Forest Weights

Feature Space	Weight-Original	Weight-No Means
Case-centric	0.20	0.28
Trend	0.49	0.27
Judge	0.11	0.20
News	0.07	0.09
Court	0.07	0.09
Weather	0.02	0.03

7 CONCLUSION AND FURTHER RESEARCH

We have shown that through a complex non-linear learning system that we can predict with a high degree of accuracy whether an asylum applicant would be granted refugee status. Furthermore, we argued that our ability to forecast has improved over time, and by 2013 we were 82% accurate in our predictions. Finally, we provided some insight into the misclassified hearings.

Surely, there are plenty of additional avenues to explore with this dataset and machine learning approach. Random forests, and hard classification in general, are not without their drawbacks. Currently our model predicts 0 or 1, for deny versus grant. However, we could have predicted a probability distribution, so that we could forecast with what likelihood a person would be granted asylum status given a feature vector.

While we tackled the problem of time series analysis, we could have focused on what, if any, type of advice we could offer future refugee applicants to increase their chances of asylum. While small decision trees are easy to interpret, complex systems are rather difficult. With 137 features, we cannot explicitly advise a refugee applicant on what, if anything, they can do to skew the odds in their favor.

REFERENCES

[1] L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (January 2001), 5–32.
[2] Chen. "This Morning's Breakfast, Last Night's Game: Detecting Extraneous Influences on Judging.". *TSE Working Paper* (????).
[3] Hale Chen, Dunn. 2017. Early Predictability of Asylum Court Decisions. *TSE Working Paper* 17-781 (March 2017).
[4] Shue Chen, Moskowitz. 2016. Decision-making under the gambler's fallacy: Evidence from asylum judges, loan ocers, and baseball umpires. *The Quarterly Journal of Economics* 131, 3 (August 2016), 1181–1241.
[5] West Aiken Cohen, Cohen. 2003. *LaTeX User's Guide and Document Reference Manual*. LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS, Mahwah, New Jersey.
[6] NOAA-ESRL Physical Sciences Division. NOAA.
[7] Retrieved in March 2017. Bloomberg L.P.
[8] Avnaim-Pesso Levav, Danziger. 2011. Extraneous factors in judicial decisions. *PNAS* 108, 17 (April 2011), 6889–6892.
[9] Schrag Philip Ramji-Nogales, Schoenholtz. 2008. Refugee Roulette: Disparities in Asylum Adjudication. *Stanford Law Review* 60 (January 2008).