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 thirty-two 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 [2], we predict twenty-seven 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.

Introduction

We like to believe that the legal system defends human and civil rights while promoting equality and fairness. Our judicial system is inundated with processes, precedent, and procedures to enforce this very ideal. 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. As a starting point, we draw attention to the correlation between the grant-deny ratio and our feature matrix. Interesting patterns emerge related to whether judges become harsher before lunchtime or the end of the day [7], 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 on the authors that focused narrowly on specific questions of casual inference [3] [5], and by another that considers a behavioral question about judges' choice to acquire information[4].

By 2013, using data only available up to the date of the trial, our model accurately predicts 82% of asylum hearing outcomes. Additionally, we detail the consistent outperformance of the random forest method versus extremely randomized and decision tree classifiers. 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.

1 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 [10]. 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.

Datasets and Preprocessing

Taken together, the final fully merged set contained approximately five hundred thousand cases and 137 features. We classified each feature into one of six buckets: case information, court information, judge information, news, trend, or weather.

Case information

A number of case-centric variables are included in our feature space. Generally speaking, we have some intuition about the relevance of these 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.

Court and judge information and trend

As a secondary source, 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, 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.

Weather and news

We integrated a time series of weather statistics, from NOAA [9], 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', which are shown in Table 1. Bloomberg [1] Trends provides daily reports on the volume of specific words across a host of multinational news sources. Through the Bloomberg API, we scraped thousands of news outlets and amassed a time-series of the frequency of our keywords. Over the past three decades media coverage has grown exponentially with the Internet. In order to account for the changing times, we regularized each feature on a rolling basis using historical z-scores before mapping them into the final feature space.

Table 1: News Trend Keywords

Refugee	Genocide	Displaced
Crisis	Ethnic	Fled
War	Ethnic Cleansing	Asylum seeker
Asylum	Migrant	Migrant Crisis

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. With this at top of mind, we took a page from our econometrics book [6] and introduced 'dummy' variables and 'dummy' indicators to the space. To 'dummy' the feature matrix, we replaced missing values with a known constant and simultaneously created a binary flag feature, which indicates whether a variable has been dummied.

2 Data Characteristics

With the full data set in hand, we assessed correlations of our feature space to the average grant ratios. 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 as shown in the bottom-right plot of figure 1.

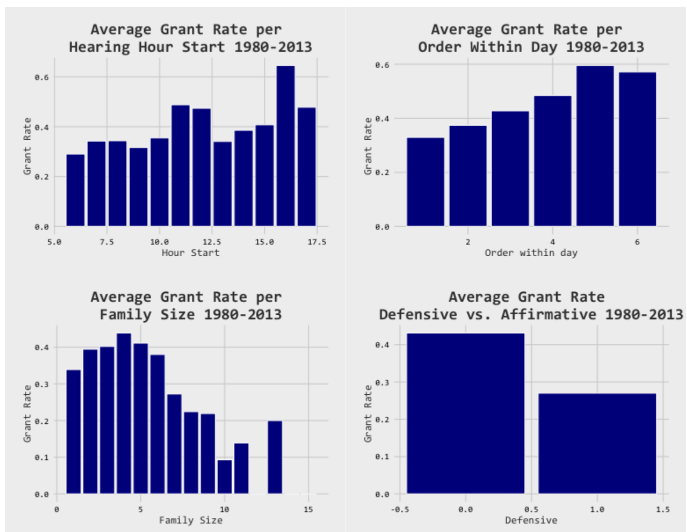

Figure 1: Case Information Charts

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. Also, the number of years of experience for each judge appears slightly positively correlated to the average grant rate.

At stark contrast with our intuition, there does appear to be some correlation between the weather and the average grant rate. The top left chart 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.

Our 'genocide' news trend indicator is less correlated to average grant rate. The trending variables, are significantly correlated to the outcome of the hearing. The bottom-left chart in figure 3 illustrates the increased likelihood of a refugee being granted asylum conditional on the previous five decisions.

The bottom right 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.

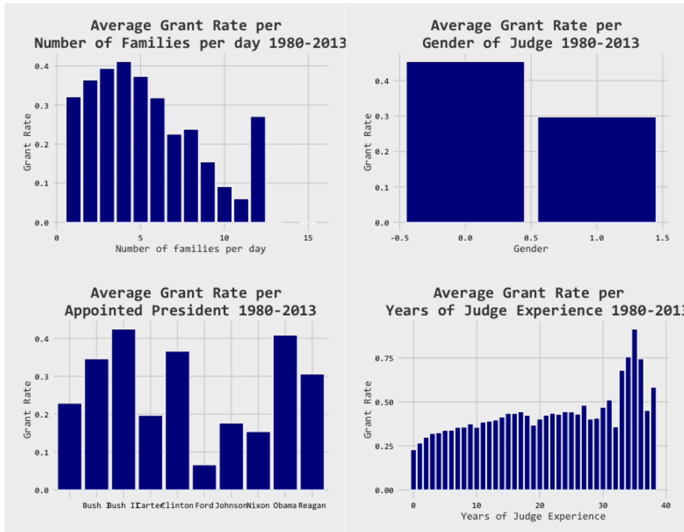

Figure 2: Court Information Charts

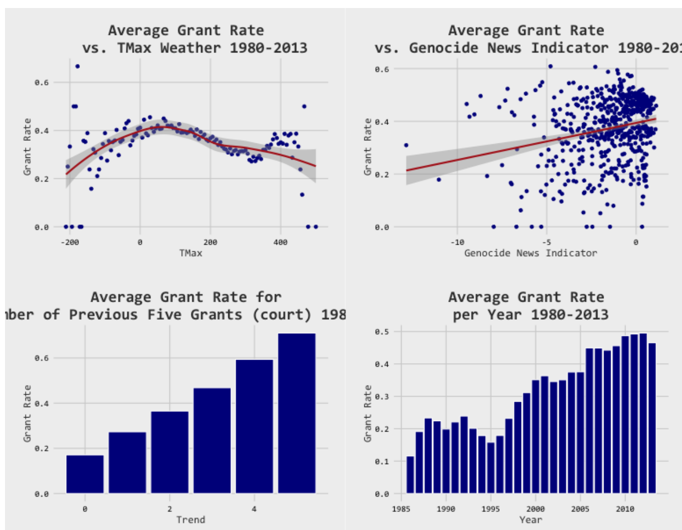

Figure 3: Trend, News, Weather Information Charts

3 Predictive Models

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. Below we have written our pseudo-code from implementing time series analysis with Python's sklearn Random Forest classifier.

```
for i in range( len( stata_dates ) ):

end_train_date = stata_dates[i]
end_test_date = end_train_date + 365

df_subset_train = subset_df_by_dates( df, 1, end_train_date )
df_subset_test  = subset_df_by_dates( df, end_train_date + 1,
        end_test_date )

X_train = df_subset_train.values[:, :-1]
y_train = df_subset_train.values[:,-1]
y_train = np.array([0 if y_train == -1 else
                    y_train for y_train in y_train])

X_test = df_subset_test.values[:, :-1]
y_test = df_subset_test.values[:,-1]
y_test = np.array([0 if y_test == -1 else y_test for y_test in y_test])

# Random Forests
forest = RandomForestClassifier(n_estimators=1000,
        max_features = 9, max_depth=None, min_samples_split=1,
    criterion = 'entropy', random_state=0, oob_score = True,
    min_samples_leaf = 25)

forest.fit(X_train, y_train)
my_forests.append(forest)

y_hat = forest.predict(X_test)
error_rates[i] = error_rate
error_y_mean[i] = error_y.mean()
oob_scores[i] = forest.oob_score_
running_importances[ : , i] = forest.feature_importances_
error_summaries[ : , i ] = error_df.describe().transpose()['mean'].values
my_error_dfs.append(error_df)
```

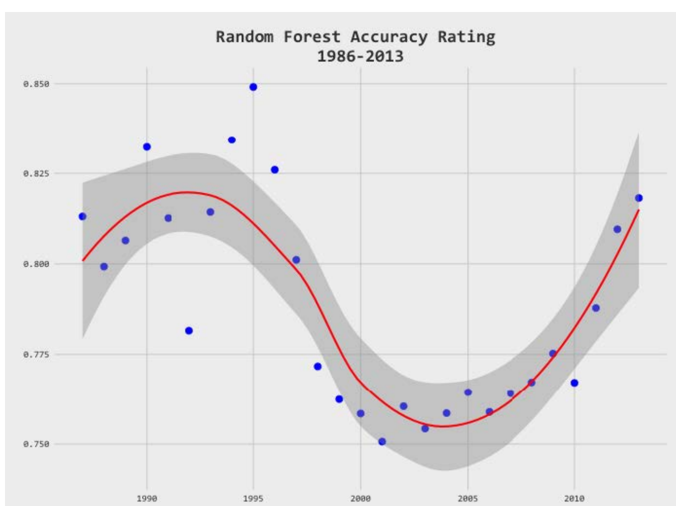

Figure 4: Random Forest Performance 1986-2013

The overall accuracy of the Random Forest reached 82% by 2013, shown in figure 4. Interestingly, in the mid-2000's there is a meaningful dip in the performance on the test set. In our error analysis we contend that this is mainly a function of two feature variables that might have some historical context.

Category	Feature	Weight
Case Information	Attorney ID	0.01
	Court ID	0.01
	Defensive	0.01
	Early Start	0.00
	Hour Start	0.004
	Lawyer	0.02
	Lunchtime	0.001
	Morning	0.001
	Nationality	0.024
	# in family	0.002
	Order in day	0.002
	Start time	0.004
	Other	0.11
	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
	Graduation Year	0.006
	Military Years	0.001
	# of Cases	0.014
	President Appointed	0.002
	Year Appointed	0.005
	Other	0.051
	Total Judge	**0.10**
News Trends	Asylum	0.006
	Cleansing	0.005
	Crisis	0.006
	Genocide	0.006
	Refugee	0.006
	Aggregate	0.006
	Total News	**0.07**
Trend Features	Judge Avg. grant	0.179
	Avg. grant for nationality	0.14
	Previous five	0.058
	Other	0.115
	Total Trend	**0.49**
Weather	Cloud Coverage	0.004
	Precipitation	0.002
	Snow	0.001
	Other	0.017
	Total Weather	**0.02**
	Total	**1.00**

In table 3, 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 echos 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 top left plot of figure 2.

Despite showing a promising correlation in our initial assessment of the data, as alluded to in top left chart of figure 3, 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.

Extremely Randomized Trees

Extremely randomized trees are similar to Random Forests, but do not rely on a bagging procedure. Instead, the same training set is used to train all trees and each tree is split randomly across an index and value. We compared the extremely randomized trees method and decision tree classifiers with our random forest approach. The random forest algorithm outperformed its counterparts consistently over the time horizon. We speculate that this is due to the bootstrapping procedure we detailed earlier. Figure 5 details those results.

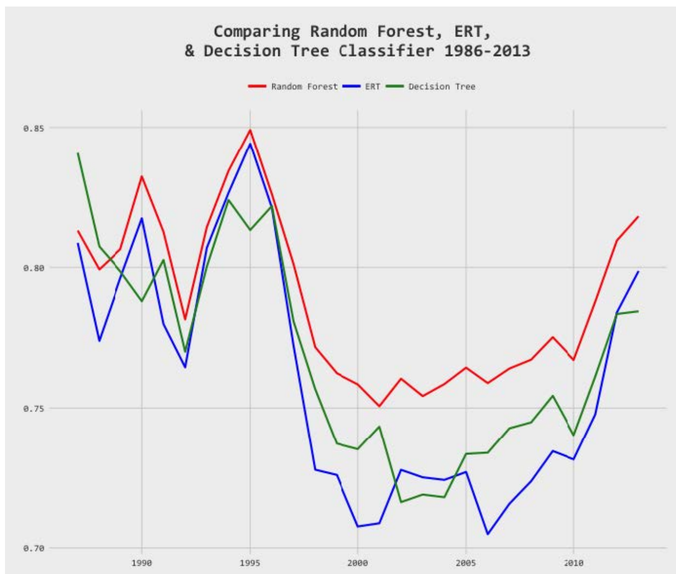

Figure 5: Comparing RF ERT and DT Classifiers

4 Error Analysis

Tantamount to any modeling exercise is understanding the characteristics of misclassified predictions. After each iteration of the random forest, we logged a data frame of the incorrect classifications. In table 4 we detail our confusion matrix and the breakdown the 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.

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**

Of particular interest to us, is how our error series evolves over time. Below we have plotted 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. One negative takeaway from figure 6 is that we consistently regress in our ability to forecast denied applications.

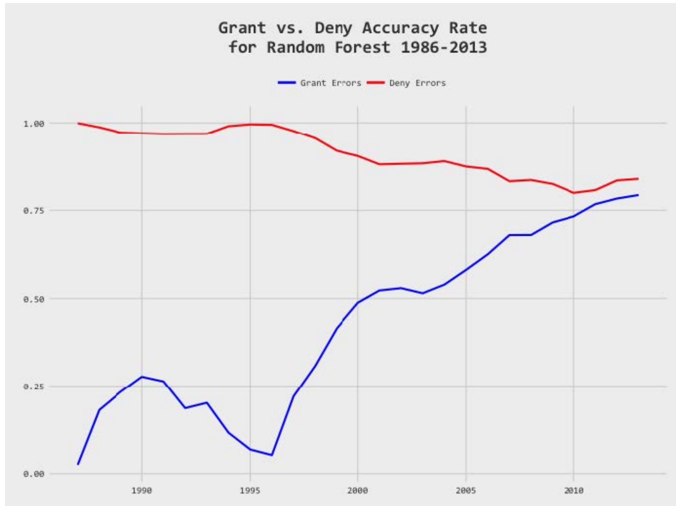

Figure 6: Grant vs. Deny Errors 1986-2013

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.

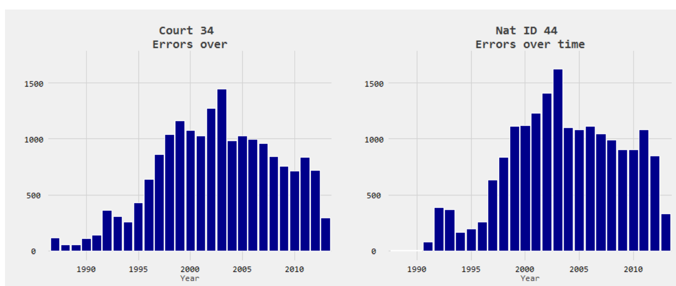

Figure 7: Errors for court 34 and nationality 44

The Second CongoWar 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.

5 A Fully Predictive Model

In the feature set we outlined, there were a few features that gathered significant weight in all three ensemble methods. 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 trending. This model produced a 79% accuracy rating on average over the time series. Table 2 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 ten times more than its original weight. Despite removing the forward looking trending features, other time sensitive variables still amass significant weighting. Number of cases granted asylum out of the previous five decisions by the judge and number of cases granted asylum out of the previous five decisions at the court account for a 9% and 3% weighting, respectively.

Table 2: 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

6 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. Additionally, we provided a comparison of our preferred random forest approach versus two other non-linear learning algorithms. 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.

Lastly, at one point we pondered the idea of penalizing false positives *(i.e. predict deny versus actual accept)* more than false negatives *(predict accept versus actual deny)*, if our tool were to advise asylum decisions. Key to our thinking, was the notion that denying anyone who was truly at risk in their home country was worse than letting a few applicants who might be less deserving of refugee status through the doors. This idea echoes, in part, the ‘beyond a reasonable doubt’ burden of proof standard. More simply, it is better to have a few guilty people in the streets than it is to have anyone innocent behind bars. On the other hand, if our tool were to advise asylum seekers, we might wish to penalize false negatives more giving an applicant false hope (you are likely to be accepted) and then have that hope taken away (application rejected).

References

[1] Bloomberg L.P. www.bloomberg.com

[2] Breiman, L. Random Forests. Machine Learning, 455(1):55-32, 2001.

[3] Chen, D., T. J. Moskowitz, and K. Shue (2014). Decision-making under the gambler’s fallacy: Evidence from asylum judges, loan ocers, and baseball umpires.

[4] Chen, D., Dunn, M., Hale, S.. Early Predictability of Asylum Court Decisions.

[5] Chen, Daniel L. "This Morning’s Breakfast, Last Night’s Game: Detecting Extraneous Influences on Judging." (n.d.): n. pag. Web.

[6] Cohen, Jacob, et al. Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, 2013.

[7] Danziger, Shai. "Extraneous Factors in Judicial Decision." (n.d.): n. pag. 11 Feb. 2011. Web.

[8] Katz, D. M. and Bommarito, M. J. and Blackman, Josh, Predicting the Behavior of the Supreme Court of the United States: A General Approach (July 21, 2014). Available at SSRN: http://ssrn.com/abstract=2463244

[9] National Oceanic and Atmospheric Administrator. www.weather.gov

[10] Ramji-Nogales, J. and Schoenholtz, A. and Schrag, P. G., Refugee Roulette: Disparities in Asylum Adjudication. Stanford Law Review, 60(2):295-412, 2008

Table 3: Feature Definitions

Feature Name	Definition	Category
comp_date	Date of ruling	Case Information
lawyer	Binary- lawyer	Case Information
defensive	Binary - affirmative/defensive	Case Information
natid	Nationality ID	Case Information
written	Binary- written/oral decision	Case Information
adj_time_start	Time of day for hearing	Case Information
eoirattyid	Attorney ID	Case Information
famcode	Family code of applicant	Case Information
numinfamily	Number of family members	Case Information
orderwithinday	Order in day	Case Information
order_raw	Order of the case for judge	Case Information
comp_dow	Day of the week of hearing	Case Information
raw_order_court	The order of the case in the courthouse	Case Information
natdefcode	Nationality of defensive applicants	Case Information
samenat	Binary- whether nationality is same as previous case	Case Information
hour_start	Hour of day for start	Case Information
morning	Binary - morning hearing	Case Information
lunchtime	Binary - hearing at lunchtime	Case Information
flag_unknowntime	Flag for unknown start time	Case Information
flag_mismatch_base_city	Flag for mismatch base city	Case Information
flag_mismatch_hearing_code	Flag for mismatch hearing code	Case Information
flag_earlystarttime	Flag to indicate timing error	Case Information
ij_code_index	Judge code	Judge Information
Male_judge	Binary - male / female	Judge Information
Year_Appointed_SLR.y	Year appointed	Judge Information
YearofFirstUndergradGraduatio	Year of undergraduate graduation	Judge Information
Year_College_SLR	Year finished college	Judge Information
Year_Law_school_SLR	Year graduated law school	Judge Information
Government_Years_SLR	# years in govt.	Judge Information
Govt_nonINS_SLR	# years in govt. outside immigration/naturalization	Judge Information
INS_Years_SLR	# years in govt. in immigration/naturalization	Judge Information
INS_Every5Years_SLR	# years in last 5 govt. in immigration/naturalization	Judge Information
Military_Years_SLR	# of military years	Judge Information
NGO_Years_SLR	# years worked in NGO	Judge Information
Privateprac_Years_SLR	# years private practice	Judge Information
Academia_Years_SLR	# years in academia	Judge Information
FirstUndergrad_Index	Identifies first undergraduate college	Judge Information
JudgeUndergradLocation_Index	Identifies location of undergraduate college	Judge Information
LawSchool_Index	Identifies Law school	Judge Information
Bar_Index	Identifies Bar passed	Judge Information
President_SLR_Index	Identifies President when appointed	Judge Information
numcases_judgeday	# cases granted asylum in this courthouse bv judge that day	Judge Information
numcases_judge	# cases granted asylum in this courthouse bv judge	Judge Information
experience	# years experience	Judge Information
experience8	Binary - judge has experience >8 years	Judge Information
courtid	Identifies the city of the courthouse	Court_Information
ij_court_code	identify judge courthouse	Court_Information
hearing_loc_code_id	Identifies the hearing location within a base city	Court_Information
zip_code	Zipcode of the hearing location	Court_Information
numfamsperslot	# families with hearing in the court in same time slot	Court_Information
numfamsperday	# families with hearing in court at that day	Court_Information
numcase_court_hearing	# Cases granted asylum in that court	Court_Information

Table 4: Feature Definitions Continued

Feature Name	Definition	Category
Refugee	Z-score of word count of 'refugee' in Bloomberg News - Refugee	News Trend
Crisis	Z-score of word count of 'crisis' in Bloomberg News	News Trend
War	Z-score of word count of 'war' in Bloomberg News	News Trend
Asylum	Z-score of word count of 'asylum' in Bloomberg News	News Trend
Displaced	Z-score of word count of 'displaced' in Bloomberg News	News Trend
Fled	Z-score of word count of 'fled' in Bloomberg News	News Trend
Genocide	Z-score of word count of 'genocide' in Bloomberg News	News Trend
Ethnic	Z-score of word count of 'ethnic' in Bloomberg News	News Trend
Ethnic_Cleansing	Z-score of word count of 'ethnic cleansing' in Bloomberg News	News Trend
Migrant	Z-score of word count of 'migrant' in Bloomberg News	News Trend
Asylum_Seeker	Z-score of word count of 'asylum seeker' in Bloomberg News	News Trend
Regularized	News Trend - Regularized	News Trend
acmh	average cloud coverage in hours	Weather
prcp	precipation	Weather
snwd	wind	Weather
snow	binary - snow	Weather
acsh	hours of sun	Weather
tsun	time of sun	Weather
tmax	Maximum temperature at the day of the hearing	Weather
tmin	Minimum temperature at the day of the hearing	Weather
numgrant_prev5	# of asylums granted in previous five decisions by judge	Trend
prev5_dayslapse	# of days lapsed between current case and 5th last case of judge	Trend
numcourtgrant_prev5	# of asylums granted in prev. five decisions (court)	Trend
numcourtdecideself_prev5	# of cases in prev. 5 in court decided by current judge	Trend
numcourtgrantother_prev5	# of asylums granted in prev. 5 in court ex-judge	Trend
courtprevother5_dayslapse	# of days laped curr. Case& 5th last case in court ex-judge	Trend
year	Year of hearing	Trend
numdecisionsraw_judgenatdef	# of asylums granted per judge x nationality x defensive	Trend
lomeangrantraw_judgenatdef	Mean grat rate per judge x nationality x defensive, ex- current	Trend
judgenumdecnatdefyear	# of asylums per judge x court x nat. x def x year	Trend
lojudgemeannatdefyear	mean grant rate per judge x court x nat, x def, x year, ex-curr	Trend
moderategrantrawnatdef	binary - value of lojudgemeannatdef year btw 0.3-0.7	Trend
grantgrant	binary - for streak 2 grants	Trend
grantdeny	binary - grant followed by deny in prev 2	Trend
denygrant	binary - deny followed by grant in prev 2	Trend
denydeny	binary - for streak of 2 denies	Trend
flag_decisionerror_strdes	Flag for non-unique decions	Trend