Abstract

In the United States, foreign nationals who fear persecution in their home country can apply for asylum under the Refugee Act of 1980. Unfortunately, over the past decade, legal scholarship has uncovered significant disparities in asylum adjudication by judge, by region of the United States in which the application is filed, and by the applicant’s nationality. These disparities raise concerns about whether applicants are receiving equal treatment before the law. Using machine learning to predict judges’ decisions, we document another concern that may violate our notions of justice: significant variation among the degree of predictability of judges at the time the case is assigned to a judge. Highly predictable judges are those who almost always grant or deny asylum. Our predictive model corroborates prior work as the final outcome of the case is overwhelmingly driven by the adjudicating judge and the applicant’s nationality. We are able to predict the final outcome of a case with 80% accuracy at the time the case opens. Additionally, this study shows that highly predictable judges tend to make use of fewer hearing sessions before making their decision. The contribution of this study is twofold. First, early prediction of a case with 80% accuracy could assist asylum seeker in their process of application. Secondly, by demonstrating the variation of predictability among the judges, based solely on a minimal subset of case information, this study raises questions about whether the specifics of each case are being given their due weight in asylum adjudications.

1 Introduction

In the United States, foreign nationals who "demonstrate that they were persecuted or fear persecution due to race, religion, nationality, political opinion, or membership in a particular social group" ref ([2016]4) can apply for asylum under the Refugee Act of 1980 (as illustrated in Figure 1), in compliance with international law, specifically the United Nations Protocol relating to the Status of Refugees of 1968 (Rottman et al., 2009). Altogether, asylum officers, immigration judges, members of the Board of Immigration Appeals, and judges of U.S. courts of appeals render approximately 79,000 asylum decisions yearly (Rottman et al., 2009).

Over the past decade, legal scholarship has uncovered significant disparities in asylum adjudication by judge, by the region in the United States in which the application was filed, and by the nationality of the applicant Ramji-Nogales et al. (2007). This is fundamentally at odds with the principle that all cases should receive equal treatment before the law. In other words, consistency (similar cases

*Equal contribution. Order arranged alphabetically according to the last names.
having similar outcomes) is desired as it indicates fairness. In addition to fairness, and perhaps even more importantly, consistency is desirable because it demonstrates that the adjudications determining the asylum seeker’s future do not depend on the personal opinions and prejudices of the individual judges to which the applicant happens to be assigned. To investigate the degree of consistency in the decisions made about similar asylum cases [Ramji-Nogales et al., 2007] demonstrated the existence of dramatic variation in decision-making among different offices, regions and officials, stating that “the variation is particularly striking when one controls for both the nationality and current area of residence of applicants, and examines the asylum grant rates of the different asylum officers who work in the same regional building, or immigration judges who sit in adjacent courtrooms of the same immigration court.” (Rottman et al., 2009) (p.302).

There could be a number of factors causing the dramatic differences in grant rates between judges. These include judge ‘burnout’, a theory that "the overwhelming caseloads and long hours worked without overtime...can potentially affect the outcome for applicants whose fates rest in judges’ hands” [Lustig et al., 2008]. The work environment that immigration judges face has also been suggested as a factor producing implicit bias that can drive their decision-making [Marouf, 2010]. Some anecdotal evidence has also been used to suggest that there are simply a few ‘bad apples,’ or a subset of immigration judges whose decisions are deemed to be unfair toward the applicant [Legomsky, 2010].

We also want to make a conceptual distinction between inter-judge disparities in predictions and inter-judge disparities in prediction accuracy. Inter-judge disparities is inconsistent with equal predictions of outcomes before an applicant comes to court based on her case facts. If, after a judge is assigned, case outcomes could be completely predicted prior to judicial inquiry into the case, this would be inconsistent with judges taking into account differences between cases. There may be cases for which the country and date of application should completely determine outcomes, for example in the case of a large-scale violent conflict in a country. However, significant inter-judge variation in early predictability suggests that this understanding does not apply to cases in general.

This current study focuses on predicting whether asylum is granted or denied based on the common features of any given asylum case: application decision (target), nationality, language, notice to appear (NTA), base city, hearing location, case type, attorney, and judge. The goal is to provide better information to asylum seekers regarding the strength of their application at the moment when they are scheduled to appear before an immigration judge. The asylum seeker may believe that her case rests primarily on the specifics of her story and may not be aware of the importance of external factors, such as the judge to whom her case is assigned. Therefore, we sought to build a predictive model that would allow applicants to understand how these external factors might affect their application.

Our model allows an asylum applicant to predict the final outcome of her application with 80% accuracy. The model uses all features about the application available at the time that the applicant receives a Notice to Appear (NTA). In addition, our model allows us to evaluate the relative impact of specific features. Echoing the findings reviewed above, the features that have the strongest impact on an application’s final outcome are the adjudicating judge and the nationality of the applicant. In addition, although the cases are randomly assigned to the judges (Ramji-Nogales et al., 2007), this study shows that a certain percentage of the judges are highly predictable, almost always granting asylum or almost always rejecting asylum regardless of the specifics of the case at hand. Although this does not indicate an inconsistent treatment of cases, it suggests that personal predilections might be a factor in the decision making.

2 Data and Methodology

2.1 Summary statistics

In the course of sorting the data to build our model, we found several statistics worthy of note. Table 1 shows the number of cases per state. We see that the vast majority of asylum applications are filed in three states, New York, California, and Florida. Table 3 shows the mean "grant rate" for each country as their citizens apply for refugee status in the United States. We see substantial variation in grant rates across countries.

The model is built using data from the Executive Office for Immigration Review (EOIR). A pre-processed version of the same date has been used in a previous work in Chen et al., 2016, which explores the gambler’s fallacy in immigration decisions. In our work, however, we focus on predicting
How Refugees Get to the U.S.

Figure 1: The asylum application process illustrated [rcu] (2016). Note that our model attempts to predict the immigration judge’s decision (the starred branch in the process diagram).

outcomes using machine learning techniques. The raw data include multiple records for a given case. Below is further details on the EOIR data:

- Data on the scheduled time and outcome of hearings: 15,377,520 records and includes approximately 70 additional features about the hearing, such as location and the presence/absence of attorney.
- Data on the outcome of asylum cases: 6,084,435 records and features include nationality of applicant, case type, asylum seeker type, base city, hearing location, decision type, attorney present/absent, unique judge identifier.
- Data on the biography of the judges: 455 text files with paragraph biographies of the judges.

The project is based on two pre-processed views of these data: (1) a merged view of the raw courts data, with each record corresponding to an asylum application decision. (2) a manually-featurized version of the judge biographies. Once the data is merged we have:

- **Total Preprocessed Data Set**: 602,500 cases
- **Total number of grants**: 213,731 cases (35.5%)
- **Training/Test split**: 482,000/120,500 cases

### 2.2 Data Engineering

Our objective was to build a model that could predict whether an applicant would be granted asylum at the moment he was notified about his initial hearing time, location, and judge assignment. In order to avoid data leakage, we first constructed a data dictionary defining each feature and indicating whether it would be available at the time of model utilization. We based our data dictionary on information from [Ramji-Nogales et al.] (2007), as well as conversations with practicing immigration attorneys. Our final feature set admitted the following features: Application decision (target as a binary variable), Language (spoken by the applicant), Nationality (of the applicant), Base city
Table 1: Ten states with highest number of cases.

<table>
<thead>
<tr>
<th>State</th>
<th>Count</th>
<th>Percentage</th>
<th>Grant Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY</td>
<td>155656</td>
<td>0.26</td>
<td>0.51</td>
</tr>
<tr>
<td>CA</td>
<td>134697</td>
<td>0.22</td>
<td>0.32</td>
</tr>
<tr>
<td>FL</td>
<td>100368</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td>TX</td>
<td>29674</td>
<td>0.049</td>
<td>0.26</td>
</tr>
<tr>
<td>NJ</td>
<td>22893</td>
<td>0.038</td>
<td>0.35</td>
</tr>
<tr>
<td>MD</td>
<td>19291</td>
<td>0.032</td>
<td>0.4</td>
</tr>
<tr>
<td>VA</td>
<td>17680</td>
<td>0.029</td>
<td>0.39</td>
</tr>
<tr>
<td>MA</td>
<td>14598</td>
<td>0.024</td>
<td>0.32</td>
</tr>
<tr>
<td>IL</td>
<td>13179</td>
<td>0.022</td>
<td>0.4</td>
</tr>
<tr>
<td>PA</td>
<td>12621</td>
<td>0.021</td>
<td>0.32</td>
</tr>
</tbody>
</table>

(asylum seeker is assigned to one of several regional immigration courts), Attorney (indicator of whether or not the asylum seeker was represented by an attorney), Notice to Appear (NTA) (the date when the applicant received notice from the EOIR), Hearing location (hearings can be held elsewhere, for example, if an applicant is detained, their hearing may be scheduled at a detention center), Case type (affirmative and defensive: Affirmative applications are made by the asylum seeker voluntarily, within one year of arriving in the United States, and are not triggered by a removal order from the U.S. government [usc, 2015]). On the other hand, a defensive application is one in which the asylum seeker has requested asylum as a defense against removal from the U.S. The case for a defensive application is presented in front of an immigration judge in adversarial proceedings and is subject to cross-examination by a government attorney (Rottman et al., 2009). A defense attorney is not provided by the U.S. government. At her own expense or with the support of a nonprofit advocacy group, the applicant can request that her case be presented by an attorney (Rottman et al., 2009). See Figure [ ] for description.), Judge features (unique identifiers for individual judges in addition to general features about the judge, such as gender and work history).

2.3 Model selection

There are two main characteristics of the data that lead us to use random forest algorithms: (1) The dominance of categorical features, and (2) collinearity when restricted to our baseline models. Most of the features are categorical variables with categories at the order of hundreds, one-hot encoding expands the dimension of the feature space in a way that leads to memory issues. On the other hand, assigning numerical values to categories keep the feature space small, but the lack of meaningful scale in the features rule out nearest neighbor classifiers. The data used for early prediction is linearly separable only when one includes the date and time specific variables that are not inherently binding to the case. Without those, there are plenty of rows that have the same input but different results. Even in the full model with 32 features that include year and month information, there are 72853 cases. This rules out linear classifiers, and even some mildly nonlinear SVM’s. Due to the combination of the above two reasons, neural networks are also not suitable for the task at hand. Indeed, our preliminary results indicate no significant improvement over random guessing when the data is trained with these models, thus we moved forward analyzing the data using random forests algorithms.

Random forests appear to be a powerful tool in the face of such limitations in data. Tree algorithms are powerful in handling categorical variables due to their inherent branching structure. Moreover, ensemble methods combined with randomness can overcome the ambiguity that is created by the same-input-different-output cases. Random forests combine these two strengths.

Ramji-Nogales et al. (2007) also uncovered a number of significant predictors of grant rates, including the gender and employment history of the immigration judge, as well as significant regional differences in grant patterns between asylum courts.
3 Results

3.1 Incremental Evaluation of Feature Space

From our discussions with practicing lawyers, we suspected that the presiding judge would be a significant factor in determining whether an applicant’s case would be granted or denied. To test this belief and to understand how an applicant’s characteristics (nationality and language) and their case information (case type and application type) impacted the Random Forest model, we incrementally added features. We did this by training the model on each set of features and then testing its performance on a hold-out test set. We chose to train the model on the smallest reasonable feature space in the first instance. From there, we added features and trained the model on each set of data as the feature space became more complex, illuminating how attributes of the applicant and judge impacted the accuracy of the model.

Baseline Model Part 0 - To test whether the presiding judge was a significant factor in predicting the outcome of an asylum application, our baseline model used a single feature, the judge ID. With this single feature, the trained Random Forest model was able to predict whether someone would be granted or denied asylum with a mean accuracy score of 71%.

Baseline Model Part 1 - When we added a single feature, the Notification to Appear (NTA) date, and trained the model, we see a minor improvement in the prediction accuracy of 2%. This suggests that variation over time in the conditions of the country of origin appears to have little additional impact on the outcome of asylum decisions.

Baseline Model Part 2 - Next, to understand how fundamental information about the applicant impacted the accuracy of the Random Forest model, we added two basic features about the applicant, Language and Nationality identifiers. These two features should, in theory, have very little bearing on the outcome if the adjudication of the case is strictly based upon whether the applicant faces a credible fear in the country she is fleeing.

When the model is trained on these two additional features, we see nearly a 5% increase in the mean accuracy score of the model, from 73% to 78%. This is rather striking since we have limited ourselves to only the most basic information about the asylum applicant. One interpretation of this finding is that a large component of the final decision on the asylum application is already set prior to the judge’s review of the application. Additional review of the case by the judge may have less impact than may be desired on the outcome of the application.

Full Model - Our final model included all features noted in the Data Engineering section, which yields a mean accuracy score of over 80% on the hold-out test data set. Our final dataset includes 602,500 records, of which 35% of the cases have been granted asylum. With 80/20 splitting we have 482,000 rows in the training set and 120,500 in the test set, and the percentage of cases granted asylum is 35% for each. For all the models shown in the table below we used a grid search of the random forest algorithm with \{128, 256, 512, 1024\} trees and 7-fold cross validation over the training set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model Part 0</td>
<td>0.71453</td>
<td>0.74101</td>
</tr>
<tr>
<td>Baseline Model Part 1</td>
<td>0.73219</td>
<td>0.77484</td>
</tr>
<tr>
<td>Baseline Model Part 2</td>
<td>0.77972</td>
<td>0.83952</td>
</tr>
<tr>
<td>Full Model</td>
<td>0.81589</td>
<td>0.88137</td>
</tr>
</tbody>
</table>

4 Discussion of Results

With only the unique identifier of the presiding judge we are able to predict with 71% accuracy whether an applicant will be granted or denied asylum in the United States. We can see from Figure 2a that when we train the Random Forest model on the complete feature space, Nationality is still a primary driver of whether or not an individual applicant will be granted asylum. We can see this in the fact that the average grant rate varies substantially even though there is not much variation in
(a) Grouped by the applicant’s nationality. For more information on large dots please refer to Table 3.

(b) Grouped by individual judges (cases are randomly assigned).

Figure 2: x-axis shows the variance in grant rates. y-axis shows the accuracy of our model on the cases when they are grouped according to the given class.

the predictive accuracy. This raises the question of how much of the individual application’s facts or litigation strategy is actually materially relevant to its final outcome.

Including all the features available on the date the case opens, the predictive accuracy is 80%. Additional variation in predictive accuracy appears across judges even holding fixed the judge’s grant rate per nationality, as seen in Figure 2b. This figure shows that some judges are fairly conservative, granting asylum to less than half of their applicants, and as their grant rate falls, prediction accuracy increases. Some of this may be mechanical, as in the extreme, a judge who never grants or always grants asylum will be very predictable. Some of the variation in the y-axis dimension can shed light on whether some judges vary in their attention to applicants from certain nationalities. Alternatively, applicants from certain nationalities may be very homogeneous, but homogeneity in unobserved applicant characteristics is unlikely to explain the variation in predictability across judges since cases are randomly assigned.

Table 3: Top ten countries by the number of applications and their grant rate.

<table>
<thead>
<tr>
<th>Country</th>
<th>Count</th>
<th>Percentage</th>
<th>Grant Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHINA</td>
<td>107964</td>
<td>0.19</td>
<td>0.53</td>
</tr>
<tr>
<td>HAITI</td>
<td>42013</td>
<td>0.074</td>
<td>0.16</td>
</tr>
<tr>
<td>EL SALVADOR</td>
<td>41626</td>
<td>0.074</td>
<td>0.087</td>
</tr>
<tr>
<td>GUATEMALA</td>
<td>34705</td>
<td>0.061</td>
<td>0.11</td>
</tr>
<tr>
<td>COLOMBIA</td>
<td>27713</td>
<td>0.049</td>
<td>0.35</td>
</tr>
<tr>
<td>INDIA</td>
<td>19161</td>
<td>0.034</td>
<td>0.37</td>
</tr>
<tr>
<td>MEXICO</td>
<td>19031</td>
<td>0.034</td>
<td>0.073</td>
</tr>
<tr>
<td>NICARAGUA</td>
<td>15987</td>
<td>0.028</td>
<td>0.2</td>
</tr>
<tr>
<td>ALBANIA</td>
<td>12036</td>
<td>0.021</td>
<td>0.52</td>
</tr>
<tr>
<td>INDONESIA</td>
<td>11399</td>
<td>0.02</td>
<td>0.32</td>
</tr>
</tbody>
</table>

5 Conclusion

While Ramji-Nogales et al. (2007) noted regional differences in asylum grant patterns and identified several judge-level variables correlated with asylum grant rates, they did not explicitly aim to build predictive models on these features. Thus, we sought to extend their work by building a predictive
model. Our goal was to develop a model that could predict whether an applicant will be granted asylum, using only information available at the time an applicant receives notice to appear (NTA) before an immigration judge. In the process, we also aimed to interpret what early predictability might mean.

University and pro-bono asylum law clinics have limited resources and a large number of prospective clients. A predictive model could potentially assist these organizations by allowing them to estimate the probability of an applicant being awarded asylum prior to any case assistance. This could also potentially allow them to suggest interventions that improve the odds of the client receiving asylum.

An additional application could be to use the model as a feedback tool for individual judges, who through it could gain information about their past granting patterns. A judge could thus enter basic information about the case and their own judge ID, and be presented with a probabilistic prediction of her decision on the application, based on past behavior.

5.1 Model deployment considerations

Finally, while our model achieved a high test set AUC, prior to deployment we would suggest validating the model against a recent test set. Given the known non-stationarity of the asylum process (often attributed to domestic political transitions), it would be necessary to assess performance under the current interpretation and enforcement of asylum law.

Moreover, incorporating international or domestic policy influences could be a further area of investigation for the improvement of this predictive model. In [Salehyan and Rosenblum (2008)] the authors look at the factors that play a role in asylum decisions at the level of domestic and foreign policy. They show that while humanitarian conditions are a consistently important predictor of asylum decisions, they also show that applicants were significantly less likely to be granted asylum if they fled countries that were U.S. military allies in the 1980s, U.S. trade partners in the 1990s, or important sources of undocumented migration in the 1980s and 1990s [Salehyan and Rosenblum, 2008].

In addition, to increase the utility of the predictive model, an additional area of work would be to develop another (simplified) model to determine which asylum court offers an asylum seeker the highest estimated probability of being granted asylum. Given the anticipated use case for our model (estimating the probability that asylum is granted after an applicant receives notice to appear),
it would be beneficial to develop supporting models that highlight the potential margins that the applicant can exploit to increase the odds of success in court. Moving to the catchment area for a different asylum court represents one of few such margins (beyond seeking legal representation). A supporting model could indicate the anticipated effect of such a move.

5.2 Early predictability

This paper raises the concern that judges might not be incorporating sufficient case information into their decisions, which raises a separate question from the significant inter-judge disparities documented by prior work. We have found that judges who are highly predictable tend to use fewer hearing sessions on average before rendering their decision as shown in Figure 5. We also found that judges who are almost always granting asylum are using more hearing sessions per applicant than judges who are almost always denying asylum. This distinction raises interesting questions as to which sets of judges are more fair in taking into account the individual differences between cases. To be sure, it is possible that judges who use fewer hearing sessions are trying to be more sparing of court time and resources and, as a consequence, become more predictable (and we see that judges who almost always deny asylum are using hearing sessions far less than other judges).

References


