Early Predictability of Asylum Court Decisions

Matthew Dunn[1], Rafael Garcia Cano Da Costa[1],
Benjamin Jakubowski[1], Daniel Chen[2]
[1]NYU Center for Data Science,[2]Toulouse School of Economics

05/11/2016

Contents

1 Introduction

Problem

In the United States, foreign nationals who fear persecution in their home country can apply for asylum under the Refugee Act of 1980 (as illustrated in figure 1). Unfortunately, over the past decade, legal scholarship has uncovered significant disparities in asylum adjudication by region and by judge [1]. In 2009, Ramji-Nogales *et al.* demonstrated dramatic differences between some judges in their asylum grant rates for particular subgroups of asylum applications. They 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. This is problematic, since our justice system is premised on the consistent application of the law across cases. More immediately, the outcome of the affirmative asylum process can have life-or-death consequences for the asylum seeker.

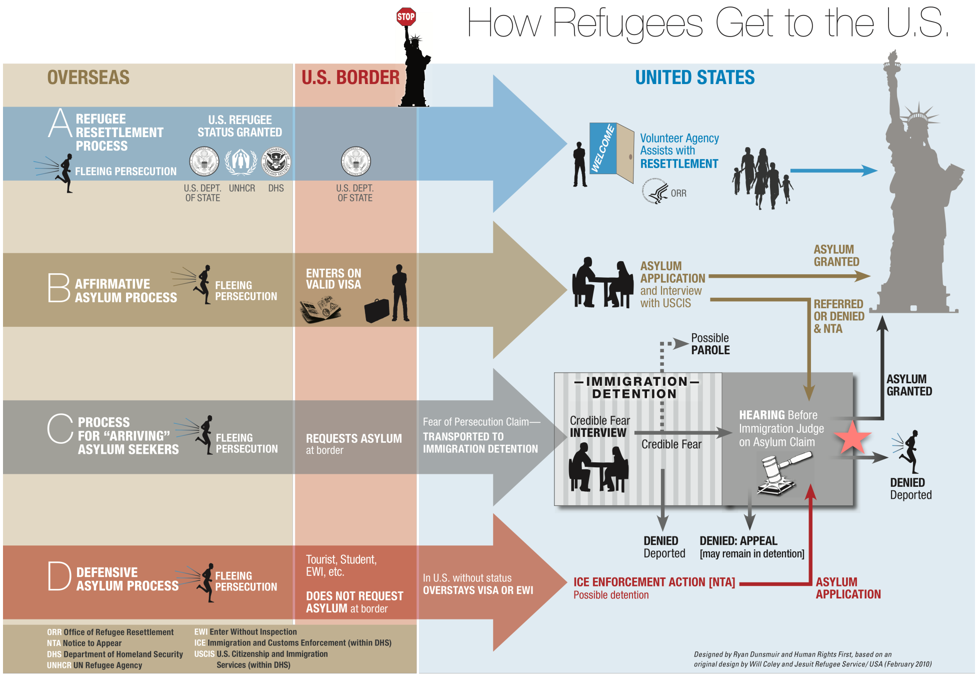

Figure 1: The asylum application process illustrated [2]. Note our model attempts to predict the the immigration judge's decision (the starred branch in the process diagram).

Objective

While Ramji-Nogales *et al.*'s groundbreaking work [1] 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 to assist legal professionals in advising asylum applicants. Specifically, we sought to develop a model to predict the probability of an applicant being 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 further explore potential predictors of asylum court decisions.

Use Case

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 inform their selection of cases, and potentially allow them to suggest interventions that improve the odds of the client receiving asylum. For example, if the predicted probability of an applicant being granted asylum was low, the clinic could advise him or her to move to another region with a more generous immigration court.

2 Methods

Data Source

Our model is built on Executive Office for Immigration Review (EOIR) data initially obtained by Dr. Sue Long. Dr. Long is the co-founder and co-director of the Transactional Records Access Clearinghouse (TRAC), which specializes in using the Freedom of Information Act (FOIA) to obtain federal government administrative records. We built our model on (i) EOIR courts data obtained by TRAC through a FOIA request, and (ii) initially processed by Dr. Chen for exploration of the Gambler's Fallacy in immigration courts [4]. The raw data provided by Dr. Long included:

- Data on the scheduled time and outcome of hearings:
  - 15,377,520 records
  - 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
  - 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.

Our project was based on two pre-processed views of these data:

- We used a merged view of the raw courts data, with each record corresponding to an asylum application decision.
- We used a manually-featurized version of the judge biographies provided by Dr. Chen

Data Engineering

Our objective was to build a predictive model that could predict whether an applicant would be granted asylum at the time they were notified of their 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 *Refugee Roulette* [1], as well as conversations with Dr. Chen and interviews with practicing immigration attorneys. Our final feature set admitted the following features:

- **Application decision (target)**: We viewed application decision as a binary variable, and aimed to predict whether or not an applicant was granted asylum.
- **Base city**: Each asylum seeker is assigned to one of several regional immigration courts.
- **Hearing location**: While each applicant is assigned to an immigration court (base city), hearings are often held elsewhere. For example, if an applicant is detained, his or her hearings may be scheduled at a detention center.
- **Nationality**: The applicant's nationality.
- **Case type**: Asylum applicants either apply affirmatively or defensively-see figure 1 for description.
- **Attorney**: Indicator of whether or not the asylum seeker was represented by an attorney.
- **Language**: The language spoken by the applicant.
- **Day of week**: The day of the week of the scheduled hearing.
- **Hearing time**: The time of the scheduled hearing, rounded to the nearest half-hour.
- **Judge features**: Since grant rates have been shown to vary dramatically between judges [1], our model includes features about the deciding judge. However we did not specify judges using unique identifiers. Instead, judges are represented using judge-level features, such as gender and work history. This model design decision reflects our objective of building a predictive model that can accommodate turnover in the judge pool. Specifically, if we had included judge ID as a feature, we wouldn't be able to make predictions for new judges; however, assuming we can determine relevant characteristics of the new judge (sex, work history, etc.), we can attempt to deploy the existing model (though this use would require additional validation).

After reducing the number of features to the subset available at the time of model deployment, new lag features were engineered to measure temporal trends in immigration grant patterns. Specifically, features were constructed to measure grant and deny rates over the past 1- and 5-years:

- At the assigned base court for other applicants of the same nationality as the applicant under consideration.
- At the assigned base court, for other applicants of the same nationality with the same application type (defensive or affirmative).
- At the assigned base court, for other applicants of the same nationality, speaking the same language, with the same application type (defensive or affirmative).

Note to avoid leakage cases were counted in these lag variables if they were decided during the 1- or 5-year window prior to applicant receiving notice to appear.

Next, we chose to exclude cases where the applicant received notice to appear prior to the year 2000. This threshold was chosen based on Ramji-Nogales *et al.*'s observation that immigration grant rates changed dramatically during George W. Bush's presidency (following the appointment of John Ashcroft as Attorney General) [1]. We chose to include only cases from 2000 onwards to account for this non-stationarity in the asylum process, while retaining a large dataset. Finally, we used an 80/20 training/testing set split. Thus, our final datasets included:

- **Training set**: 261820 records (asylum application decisions)
- **Testing set**: 65456 records (asylum application decisions)

Model Fitting

General modeling framework

We proceeded to build our predictive model as follows:

- **Model evaluation**: All models were evaluated using ROC AUC, since the dataset had imbalanced classes (asylum applications are denied more often than they are accepted).
- **Baseline** We first trained and evaluated a simple logistic regression model to establish a baseline for comparison.
- **Model Optimization** Next, we used 5-fold cross validation to train and evaluate a number of different model configurations.
- **Model Selection** Based on cross-validation performance over the training set, we chose our highest-performing model for evaluation against the test set.

Baseline Model

Prior to learning an optimized predictive model, we first established a baseline model. Based on Ramji-Nogales *et al.*, we identified application type[1]. and attorney representation as potentially strong predictors of an applicant receiving asylum. To build a simple baseline model on these features, we used unregularized logistic regression:

$$\log\left(\frac{\mathrm{P}(\mathit{Grant})}{1-\mathrm{P}(\mathit{Grant})}\right) = \beta_0 + \beta_1 \mathbb{1}_{\text{Type = A}} + \beta_2 \mathbb{1}_{\text{Type = M}} + \beta_3 \mathbb{1}_{\text{Attorney}}$$

After fitting this model on our training data, we determined baseline performance over the test set (measured using ROC AUC). The test set ROC curve for this baseline model is shown in figure 2. Recall the ROC AUC has a probabilistic interpretation; namely, it is the probability a random true positive (asylum is granted) instance is ranked above a random true negative (asylum not granted) instance by the model.

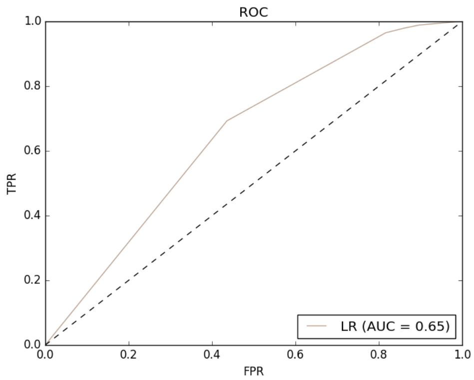

Figure 2: ROC for baseline model

Learning an optimized model

Next, we proceeded to specify and train a number of different candidate classifiers. Training was executed using sklearn's GridSearchCV [3] to optimize hyperparameters for AdaBoost, random forests, logistic regression, and support vector classifiers. Hyperparameters tested using GridSearchCV are described in Table 1. Note in addition to testing different hyperparameters, we also tested two different feature sets:

- **Feature set 1**: This was our original feature set, including the features described in the **Data Engineering** section.
- **Feature set 2**: This feature set included all original features, plus all interactions with the attorney indicator variable. We tested this expanded feature set because Ramji-Nogales *et al.*'s previous work indicated the significance of interaction terms with the attorney indicator variable [1].

Note the second feature set was only tested for classifiers which would otherwise not have included these interaction terms (logistic regression

[1]Note in this equation, case type is abbreviated as A = Affirmative and M = Missing

and SVM models); these feature sets were not tested for tree-based models that implicitly included interaction . Trained models were evaluated using mean 5-fold cross-validation (CV) ROC AUC. The results for all trained and optimized models are shown in table 1.

3 Results

First, selected modeling results are presented in table 1. For each model family, we present the maximal mean AUC ($\pm$ sd) from the 5-fold cross validation.

Model	**Hyperparameter Grid**	**Optimal Hyperparameters**	**AUC ($\bar{x} \pm$ sd)**
Random Forests	Decision Trees: [100, 300, 600, 900]	900	.8892 $\pm$ 0.0011
AdaBoost	Decision Trees: [100, 200, 400, 800]	100	.8089 $\pm$ 0.0089
SVC	C = $[10^{-5}, 10^{-4}, \cdots, 10^{2}]$ Feature sets = [1, 2]	C = 1 Feature set = 1	.8143 $\pm$ 0.0024
Logistic Regression	Regularization: ℓ_1, ℓ_2 C = $[10^{-5}, 10^{-4}, \cdots, 10^{2}, 200, 300, \cdots, 1000]$ Feature sets = [1,2]	Regularization: ℓ_2 C = 10^{-4} Feature sets = 1	0.7899 $\pm$ 0.0124

Table 1: Model performance for 5-fold cross validation on the training set

ROC AUC was maximized by a random forest, which achieved a cross-validation average ROC AUC of 0.8892. Thus, we chose this model for evaluation against the held-out test set. Note our choice of 900 estimators in the random forest is supported by a training set learning curve (CV performance vs. number of estimators) presented in figure 3. This curve illustrates the diminishing marginal returns in model performance achieved by inclusion of additional decision trees in the random forest. Given these diminishing returns, and the time-cost of learning additional trees, we are justified stopping growth of the forest after 900 trees.

In addition, figure 4 shows the test set ROC curve for our final classifier. For context, note this final test-set model performance represents a dramatic improvement over our baseline model, which achieved a test-set AUC of 0.65 (see the **baseline model** section).

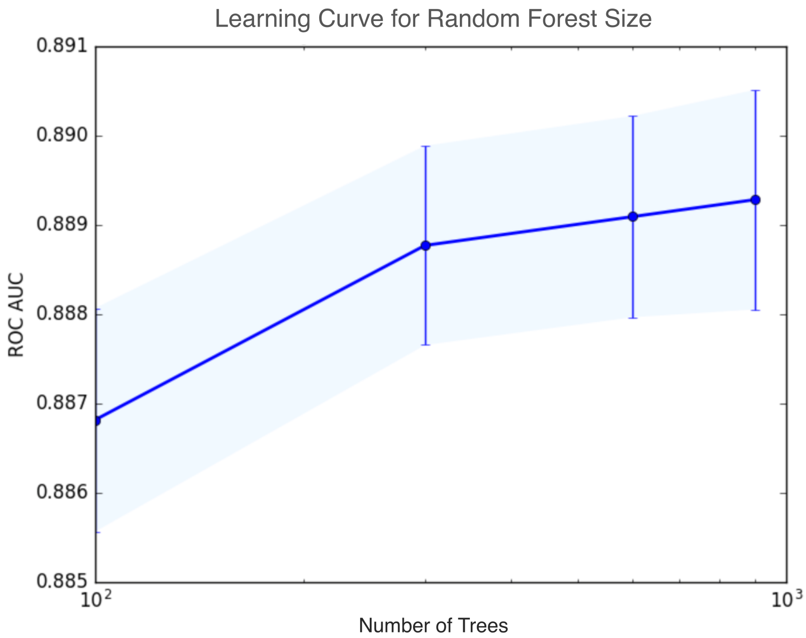

Figure 3: 5-fold cross validation AUC

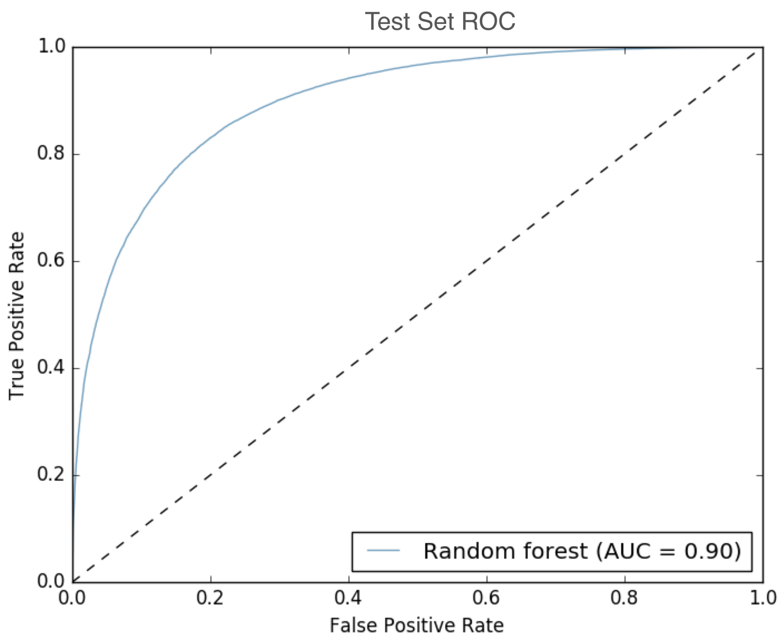

Figure 4: Test set ROC curve for the final model

4 Conclusion

Model implications

First, while asylum law may be applied inconsistently across courts and judges, its application is still highly predictable even at a very early stage in the asylum process. Ramji-Nogales *et al.*'s groundbreaking report was titled *Refugee Roulette* to reflect the reality that asylum seekers face differing odds- hence, applying for asylum involves "spinning a roulette wheel", where your[2] odds of being granted asylum are problematically determined by your location and random judge assignment, not just (i) your fear of persecution, and (ii) the credibility of that fear. Our results add to this picture (without fundamentally challenging it). In essence, our results suggest that given

- Your spin of the roulette wheel (i.e. your judge assignment, base city, hearing location, and assigned hearing day and time)
- Origin (language, nationality) and access to legal representation

the outcome of your asylum application is largely baked in (as indicated by an AUC of 0.9) significantly before your day in court. Moreover, the feature importance ranking (shown in 5) indicates much of this predictability arises from historic patterns *specific to your assigned court*.

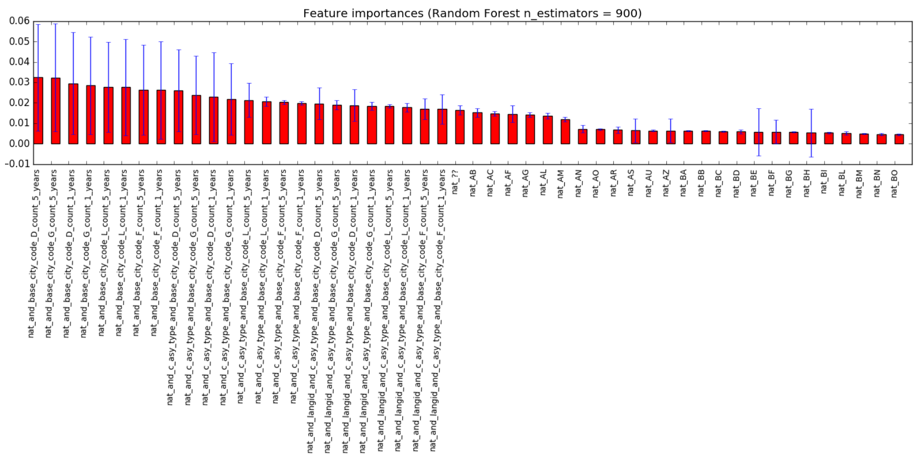

Figure 5: 50 most important features in our random forest

[2]Moving forward, we refer to an applicant as 'you' and 'your'. We do so to ease the exposition and to highlight that we are speaking about life-and-death decisions affecting individuals and families

Importantly, the reader should interpret figure 5 as a relative ranking of the importance of different features. As indicated by the error bars, the feature importance scores are highly variable across trees in the forest. This reflects the correlation between features within a group (i.e. between `nat_and_base_city_code_D_count_5_years`, the 5-year count of denied applications for applicants of your nationality in your specific immigration court, and `nat_and_base_city_code_G_count_5_years`, the 5-year count of granted applications for applicants of your nationality in your specific immigration court). Still, the implication of the relative feature importance rankings remains unchanged- the prevailing grant rate at your assigned immigration court is the most important predictor of your case outcome.

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 (since the available test set was a random sample of decisions from the year 2000 onward). Given the known non-stationarity of the asylum process (often attributed to political transitions), it would be necessary to assess performance under the current interpretation and enforcement of asylum law. 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 asylum is granted after an applicant receives notice to appear), it would be beneficial to develop supporting models that highlight potential margins the applicant can exploit to increase their 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), and a supporting model could indicate the anticipate affect of such a move.

Next Steps

Along with developing additional models to support the deployment of a predictive model, we would also consider evaluating a few additional models. Due to the high dimensionality of the final analytic data set, learning and storing the random forest was memory intensive. As such, we'd suggest trying alternative modeling libraries that are designed to handle very large data sets (specifically XGBoost). Another area of exploration we suggest is evaluating alternative loss functions (logistic loss, etc.) for forward additive stepwise modeling (boosting), since the AUC obtained using an exponential loss function (AdaBoost) was relatively low. Finally, we'd suggest constructing a probably calibration curve, to evaluate the performance of the model as a probability estimator (since probability estimation is the intended use of the final model).

References

Acknowledgements

The first three authors would like to thank Dr. Chen for his support in providing data and guidance in analysis. We would also like to thank Dr. Sue Long for initially sharing the EOIR data with Dr. Chen. Finally, we would like to thank Anna Boyle (attorney) for her insight into the lawyer and applicant's perspective on the asylum process.

Code base

The code base for this analysis can be found at `https://github.com/mattyd2/ds1003finalproject`.

Bibliography

[1] Ramji-Nogales, Jaya and Schoenholtz, Andrew and Schrag, Philip G., Refugee Roulette: Disparities in Asylum Adjudication. Stanford Law Review, Vol. 60, 2008; Temple University Legal Studies Research Paper No. 2007-12. Available at SSRN: `http://ssrn.com/abstract=983946`

[2] Refugee Council USA. Asylum and Detention. `http://www.rcusa.org/asylum-and-detention`. 2016.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011, 12:2825-2830.

[4] Chen, Daniel L. and Moskowitz, Tobias J. and Shue, Kelly, Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires (January 12, 2016). Fama-Miller Working Paper. Available at SSRN: `http://ssrn.com/abstract=2538147`