Machine Prediction of Appeal Success in U.S. Asylum Courts

Elliott Ash, Daniel Chen, Colin Andrus, Dustin Godevais, Gary Ng*

Abstract

This paper asks whether the decisions of the appeals boards of U.S. Asylum Courts can be predicted using machine learning tools applied to information on the lower-court decisions. We use a new data set of 830,000 asylum appeals for the years 1985 through 2013. We show that the decisions of asylum appeals can be predicted with 80% accuracy and 0.85 AUC. Comparable performance is obtained using only decisions in previous years as training data. Important predictors include the nationality of the asylee and the identity of the lower-court judge. Our model suggests that the individuals who do not appeal have a very low predicted success rate.

1 Introduction

Under the United States Refugee Act of 1980, the U.S. may provide asylum to individuals from other nations facing persecution due to "race, religion, nationality, membership in a particular social group, or political opinion."[1] Many people are already residing inside the U.S. at the time they file for asylum. If one's asylum application is denied by the U.S. Citizenship and Immigration Services Asylum Office or if removal proceedings begin against them, an asylum seeker can argue for their asylum to be granted in Immigration Court.[2] Decisions in U.S. Executive Office for Immigration Review (EOIR) Immigration Courts are made by a single judge who has the authority to grant or deny asylum.

In the case of a denial, the asylum seeker may appeal the decision the Board of Immigration Appeals (BIA).[2] Most appeals from EOIR Immigration Court to the BIA are also decided by a single member of the board, but in some cases panels of

*Elliott Ash: ETH Zurich, ashe@ethz.ch; Daniel L. Chen: Toulouse School of Economics, daniel.chen@iast.fr.

3 members may take place.[3] Due to the high stakes decision being made by a single judge at the Immigration Court, the possibility of appealing provides a valuable opportunity to the individual seeking asylum to seek a reversal of an asylum denial.

Previous studies have largely focused on EOIR Immigration Court decisions, exploring the decision making process of judges[4] and predicting decision outcomes.[5] In the current study, we seek to predict the outcome of the BIA appeals filed by respondents who were denied in the original proceedings. Our study is motivated by two purposes: to understand the predictive drivers of appeal success, specifically as it relates to judge biases, and to potentially use the predictive model to advise appellants on their likelihood of success.

We show that the information available on the lower-court case can predict appeals outcomes. In the tuned model, appeals decisions are predicted with 80% accuracy and 0.85 AUC score. In our follow up analysis, we indicate the importance of time features, nationality of the asylee, and the identity of the judge in the original asylum court hearings. We also find that the individuals who do not appeal a denial have a low predicted success rate, consistent with a rational response to private information about appeal success probability.

2 Data

2.1 Data Sources

Our prediction model utilizes 3 datasets: data from the original hearing EOIR Immigration Court, data from appeals to the BIA, and Immigration Court judge biographical data.

The EOIR Immigration Court data we use for modeling is a cleaned dataset assembled by Chen et al[4] and Dunn et al.[5] for use in their work modeling Immigration Court asylum decisions. Our BIA appeals dataset consists of the attributes and outcomes of Appealed Immigration Court decisions. Judge biographical data provides attributes for individual judges who make Immigration Court decisions.

Immigration Court asylum cases data contains attributes associated 602,500 asylum cases including the asylum seeker's nationality, language spoken and details about their case including whether they sought out asylum affirmatively or defensively. A binary label for the case gives outcome. Of cases included in the dataset 35.5% have been granted asylum and the remaining 64.5% were denied.

Our raw BIA appeals dataset consists of 870,388 total instances, 870,388 having an appeal ID, 868,758 having a case ID and 776,380 having a proceeding ID.

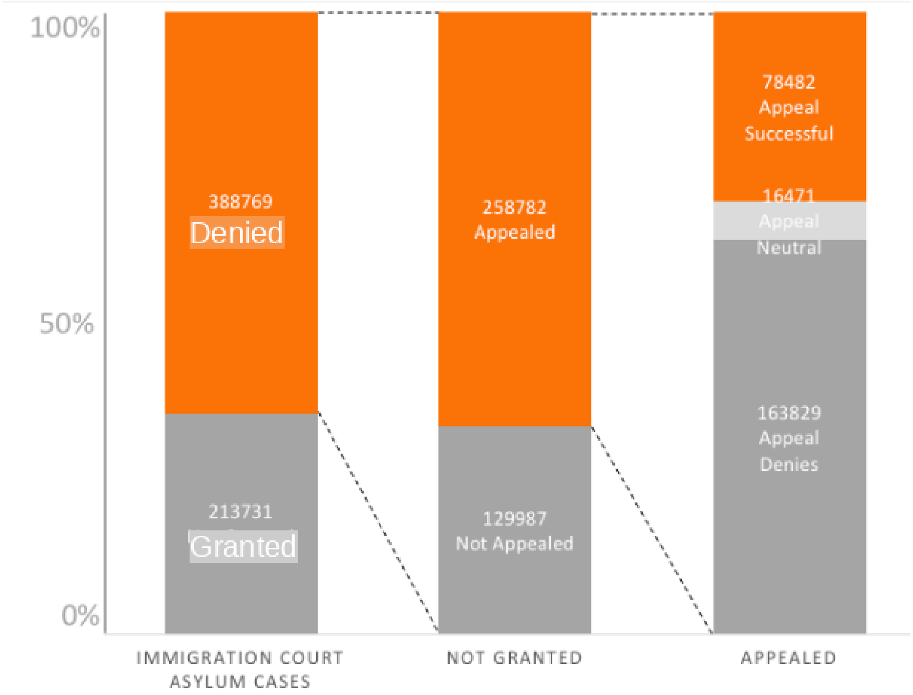
Instances lacking a value for appeal, case or proceeding. After dropping instances with NA values for any appeal, case or proceeding ID, we have 755,222 observations.

Judge biographical data contains information on 367 judges including when they began their role as a Immigration Court judge, schools they attended and the time they have spend working in different types of law (i.e. government, military, private practice).

2.2 Data Prep

In order to assemble a dataset of appealed asylum decisions, the Immigration Court asylum decision dataset is merged with BIA appeal data using values for case ID and proceeding ID. Of the 602,500 original Immigration Court asylum cases, there were 366,927 appeals. Of those appeals, 11,689 were appeals by the government when asylum was granted by the Immigration Court. Because they are relatively few, we drop these cases, but they are an interesting topic for future work. In the case of mulitple appeals, we remove all but the most recent appeal. This leaves 599,377 cases for analysis.

Figure 1: Asylum Case Outcomes



To use this data for building a binary classification model, it is necessary to classify the many possible outcomes of appeal as either positive or negative for the appellant. We split the 27 different decision strings into positive (granted), negative (dismissed), or neutral (other). Appeals resulting in neutral outcome are dropped from the dataset, reducing the data set to 242,466 appeals.

We create two additional judge experience features: these are the difference between the year of appeal and judge's law school graduation year, and EOIR Immigration Court appointment year. respectively. We also created two time horizon features; days elapsed between when charges were filed and when the proceeding began, and days elapsed between when the proceeding began and when the Immigration Court decision was made. To more effectively capture any trend effects, we also made two features representing the average grant rates of the last ten appeals for the same judge and same judge-nationality.

2.3 Summary Statistics

Our final dataset used for modeling consists of 242,466 appeals with 38 features (See Appendix 1 for complete list) describing the appellant, the case, and the lower-court judge. Of these appealed cases, 78,482 (32.4%) were successfully appealed and 163,829 (67.6%) were unsuccessful in the appeal. Appeal grant rate, however, varies widely across time and other factors. Figure 2 illustrates the number of appeals and average appeal grant aggregated over some of the key feature dimensions. First, we see that the number of appeals has increased over time. The grant rate decreased in the late 90s and early 2000s, but has increased since the mid 2000s. There is wide variation across judges in the number of cases, and especially in the grant rate. Most asylees are from China and Latin America; Haiti has relatively low grant rates. Finally, there is also variation across the cities from which asylum is requested.

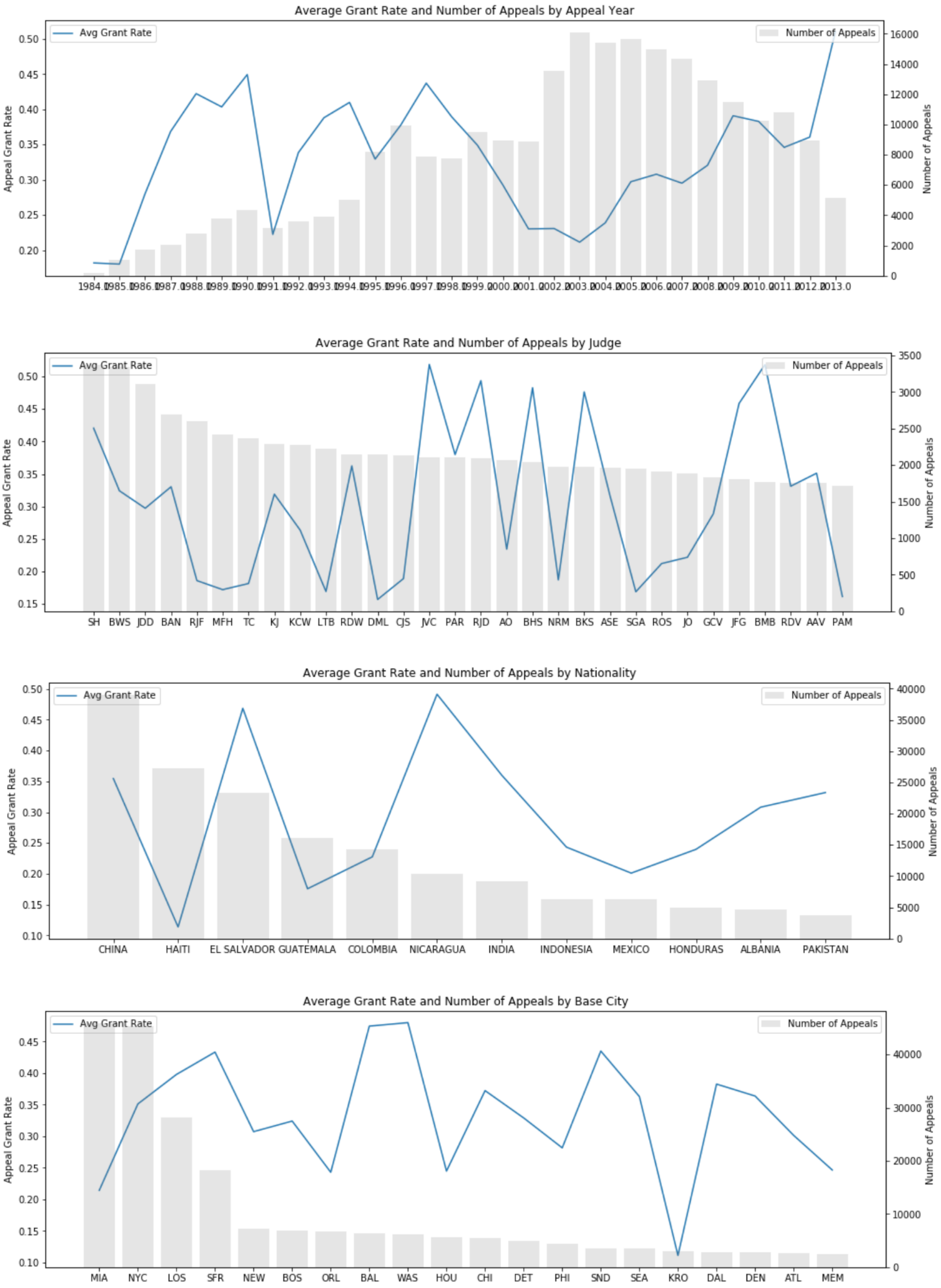
3 Pooled Model

The goal of this predict the appeal outcome of denied asylum cases. This section described the methods used.

3.1 Feature Selection

Using an untuned random forest model, we iteratively add features believed to be important in predicting the outcome of asylum appeals. Model performance greatly increases with the addition of nationality, judge identity, and year of appeal. As

Figure 2: Summary Statistics on Appeal Counts and Grant Rates



shown in Table 1, the recall of the model greatly increases with additional features. Performance as measured by AUC continues to increase with the addition of the remaining 38 features, but with diminishing returns as compared to the the addition of nationality, judge and appeal year.

Table 1: Random Forest Performance Using Different Groups of Features

Model	Accuracy	ROC AUC	Log Loss	Precision	Recall
Full Model	**0.792**	**0.840**	**0.612**	**0.750**	**0.538**
Nat + Judge + Year	0.741	0.765	1.095	0.628	0.501
Nat + Judge	0.704	0.701	0.769	0.578	0.337
Nationality Only	0.683	0.665	0.590	0.565	0.109
Judge Only	0.675	0.625	0.608	0.503	0.061

3.2 Model Selection

In selecting the algorithm to use for our classifier, we chose tree-based models over linear models given their ability to capture nonlinearities and interaction effects in the data. We tested four model algorithms out-of-the-box. We found that Random Forests significantly outperformed gradient boosting, XGBoost, and logistic regression in terms of AUC. The relative performance of the untuned models is illustrated in Figure 3.

We proceeded to tune the random forest model to yield the best AUC. To speed up the tuning process, we used 20 trees per parameter combination, varying values for max depth of trees (60, 80, 100, 120, 140, None), minimum samples to be split (2, 5, 10) and minimum samples per leaf (1,2,4), and max features (sqrt, log2). The best model returned has a max depth of 60, minimum sample split of 2, minimum of 1 sample per leaf, and max feature configuration of 'sqrt'. Finally we re-ran the model with best parameters and more trees (100) to get our best aggregate random forest model, which yielded an accuracy of 80.2% and AUC of 0.855 on the test set. These performance metrics are summarized in Table 2. In our context, it turns out that the default parameters in sklearn performed quite well in the prediciton task.

3.3 Validation

Testing our aggregate model on 48,494 unseen validation asylum appeals returns the confusion matrix reported in Table 3. Again, the data

Figure 3: Performance of Tree Based and Linear Models

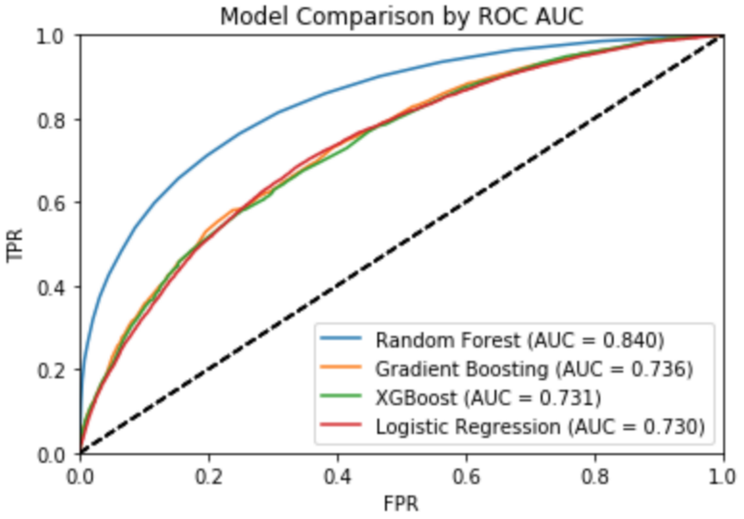


Table 2: Tuned and Untuned Random Forest Model Performance

Model	Accuracy	ROC AUC	Log Loss	Precision	Recall
Tuned Model	**0.802**	**0.855**	**0.443**	**0.760**	**0.570**
Untuned Model	0.792	0.84	0.612	0.750	0.538

4 Sequential Model

4.1 Motivation

The baseline model fitted above serves well in terms of understanding what factors drive appesequenal grant rates. However, it tends to exaggerate the predictive power of the model that we might use in practice to predict on future appeals, because in the real world we do not have the benefit of using data on concurrent or future appeals. In terms of using this model to advise current, possible real world appellants it is less useful. Due to changes in global events such as war and regime change, the attributes of who is granted asylum and who successfully appeals asylum denials vary greatly over time.

To see this better, let us look at one example of the benefits of a sequential model. In Figure 4 we plot the aggregate appeal success rates by month. There is an obvious dip in the aggregate grant and total number of appeals in 1991.

It turns out that this dip is due to the end of the Salvadoran Civil War, during

Table 3: Confusion Matrix for Tuned Random Forest

	Predicted Appeal Denial	Predicted Appeal Success
Actual Appeal Denial	29,912	2,829
Actual Appeal Success	6,770	8,983

Figure 4: Number of Appeals and Success Rate, by Month

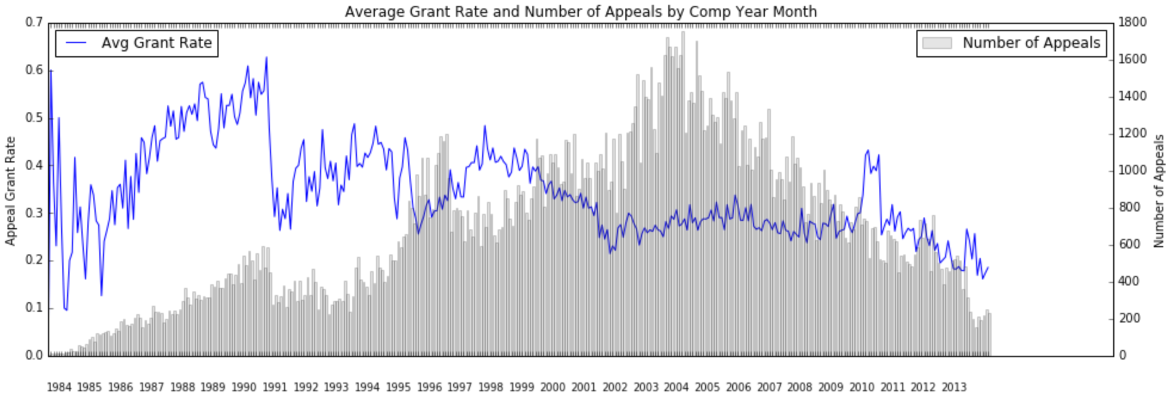
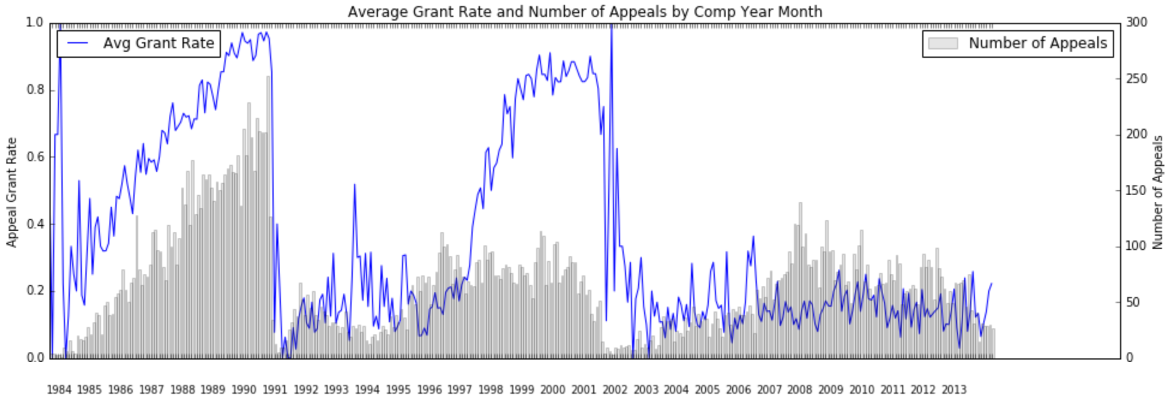


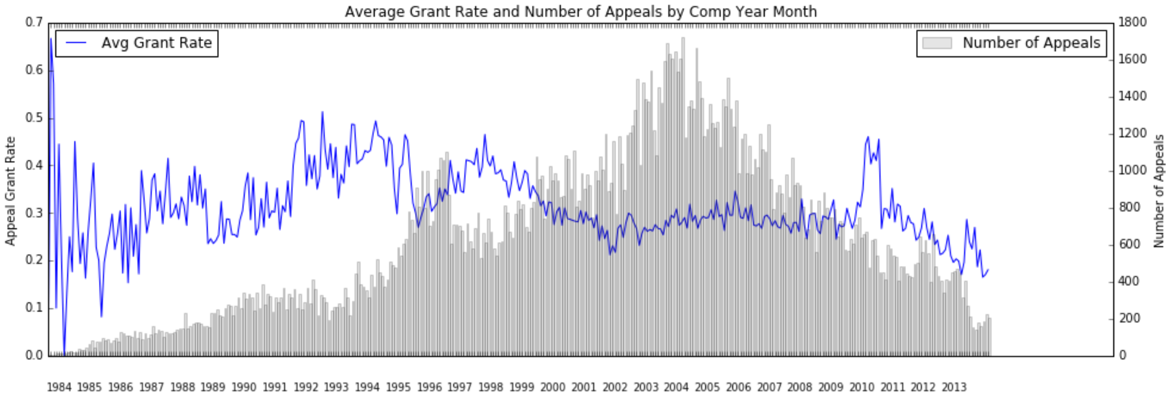
Figure 5: Number of Appeals and Success Rate, El Salvador Asylees Only, by Month



which a large number of people sought asylum from El Salvador. When appeal grant rates and total number of appeals are plotted only including data from applicants of Salvadoran nationality (Figure 5), the dip is even more pronounced. In addition, we can see another important event related to El Salvador in March 2001. This was when refugees from El Salvador were granted a temporary protected status to stay in the U.S. legally. This separate, simpler option to stay in the country led to fewer asylum applications, fewer appeals, and also a lower appeal success rate.

Plotting the aggregate data again without El Salvador (Figre 6) removes the dip at 1991, but presents additional points with large changes in grant rate. This example illustrates why for a real world, advisory application, using sequential models that more heavily weight recent data compared to data from appeals far in the past could be advantageous. Predicting the outcome of appeals for Salvadorans in 1992 would not be as successful as the results of our aggregate model suggest due to the model being trained on cases from before and after the end of the civil war.

Figure 6: Number of Appeals and Success Rate, Excluding El Salvador Asylees, by Month



4.2 Sequential Model Approach

To address this issue, we create an additional set of sequential random forest models, in which we build a separate model for each year between 1994 to 2013 by using only data from preceding years (e.g. to predict 1994 appeals, we train on data from 1993 and before).

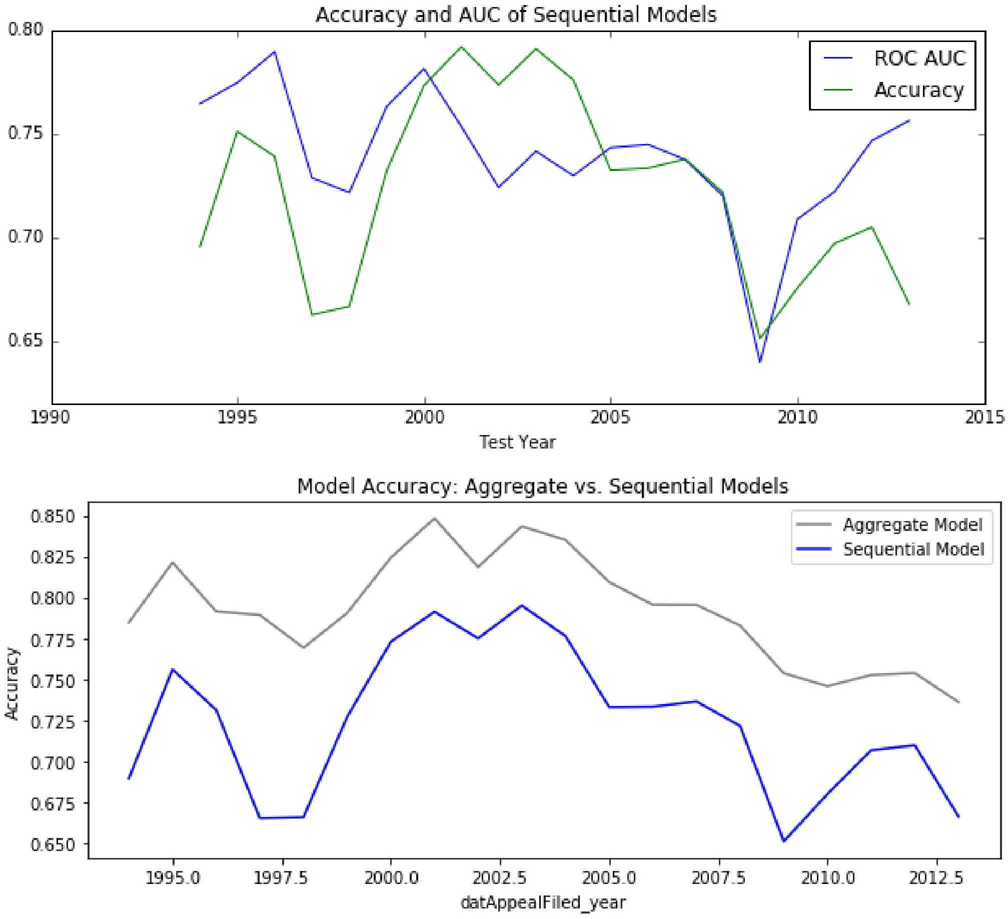
Second, given that exploratory data analysis suggests that appeal grant rates can vary drastically we time, we surmise that more recent data might be more informative to our prediction. Thus we employ a sample weighting technique wherein we weigh each sample exponentially less based on the "age" of the data. More formally, let τ be the number of years between the prediction year and the year of the previous appeal data. We assign each data point a weight of α given by

$$\alpha = 0.9^{\delta - 1}.$$

Cases in the current year and in the future are weighted to zero. This means that when we are predicting 1994 appeals, we would apply $\alpha = 0$ for 1994's data, $\alpha = 1$ on 1993's data, $\alpha = 0.9$ on 1992's data, $\alpha = 0.9^2 = 0.81$ on 1991's data and so on.

Applying the sequential method, we train models for the years 1994 through 2013. Figure 7 shows the performance metrics for these models for each year. Unsurprisingly, performance drops compared to our aggregate model with no single year having an AUC above 0.8 and some years falling below 0.7. Comparing the year by year performance of our sequential models to that of the aggregate model (bottom panel), we see that the aggregate routinely outperforms the sequential models. Still the sequential model performs quite well, especially given that it is only using historical data.

Figure 7: Performance of Sequential Models by Year



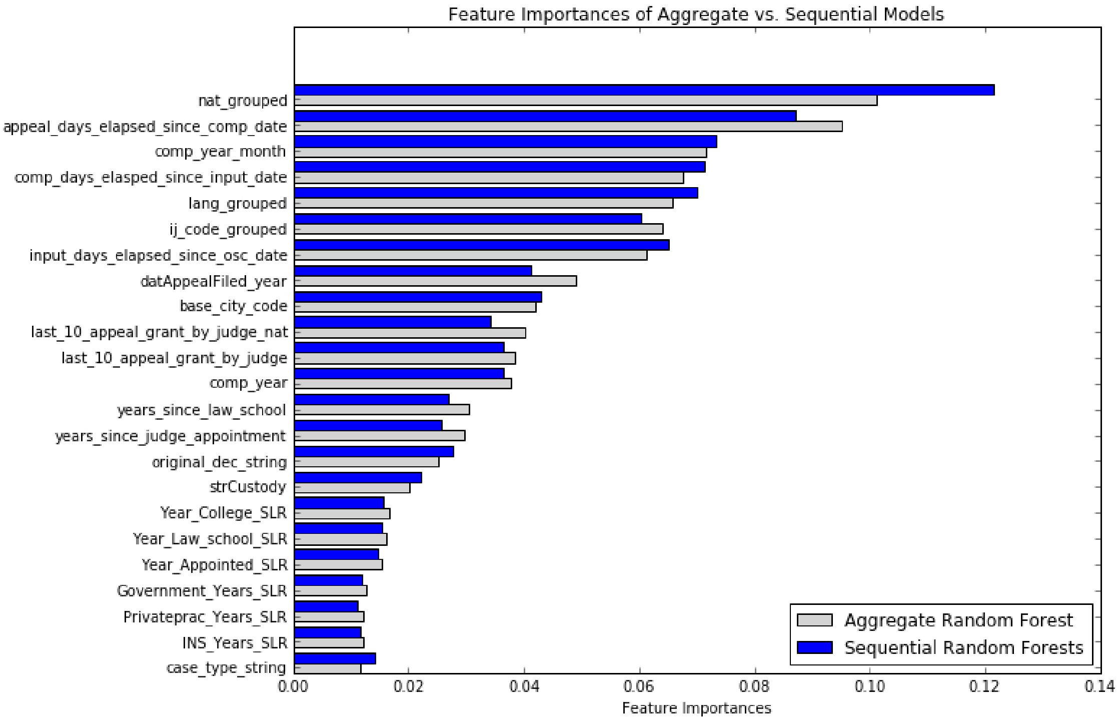
5 Analysis

Next we use the model to analyze the asylum process.

5.1 Feature Importance

This section analyzes the importance across features, as weighted by the random forest model. These are measured as the mean decrease of impurity, returned from both the aggregate model and the mean of the 20 sequential models' importances. The most important features are reported in Figure 8.

Figure 8: Feature Importances



The Nationality feature has the highest feature importance for both the aggregate and averaged sequential models, with the sequential models giving it more importance. This comes mostly at the expense of time horizon variables like year of the appeal being filed. This makes sense because in the aggregate model, the model is able to deduce current event and other trend effects by correlating appeals occurring during the same time period. However, the sequential model does not have the benefit of peeking into other concurrent or future events. This is a key reason why the sequential model has lower predictive power than the aggregate model.

Table 4: Feature Importance Grouped and Summed by Class

Feature Group	
Time Horizon Features	0.377804
Judge Features	0.277066
Respondent	0.177945
Trend Features	0.074494
Proceeding Features	0.060490
Location Features	0.042636

A successfully appealed denial of asylum effectively means the judge making the decision in Immigration Court made a mistake. To quantify the effect a judge has on appeal success, we split our feature importances into 6 groups. As shown in Table 4 the summed feature importance of judge related features is 0.277. This is greater than all but the time horizon features, which include important features such as the year of appeal. This shows that which judge made an asylum decision has a large but not overwhelming effect on the likelihood of appeal success.

Our asylum appeal model provides some useful insights to possible appellants. Judge biases, as evidenced by the large feature importances of judge related variables, show that the judge selected in Immigration Court can greatly affect their odds of appeal success.

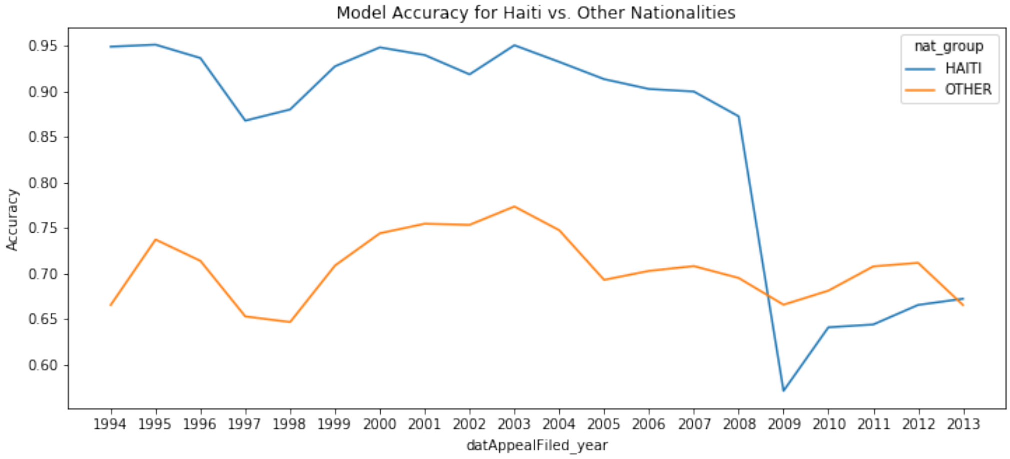
5.2 The Haiti Earthquake

As shown in Figure 9, both the aggregate and sequential models have a large drop off in prediction performance for appeals being filed in the year 2009. Much of this decrease in model performance can be attributed to the events following the catastrophic earthquake that struck Haiti in January 2010. Many Haitian individuals who had attempted to gain asylum in 2009 had their asylum denied. When they appealed, they were subsequently granted asylum through their appeal after conditions had significantly deteriorated. Changes regarding the stability of an asylum seeker's country can change the likelihood of appeal success.

5.3 Predicted Success in the Case of No Appeal

Since data on the outcome of appeals is only available for those who do choose to appeal, we are training the model on a selected sample. While appealing a denial of asylum would maximize an individuals chances of being granted asylum and

Figure 9: Aggregate Model Accuracy Split by Haitians and All Other Nationalities



being able to stay in the United States legally, other factors make not appealing a reasonable choice. If seeking asylum affirmatively, an individual may not bother with appealing their case to pursue a different avenue of remaining in the U.S. legally, such as a work visa. If an individual seeking asylum defensively is in detention for the length of their court proceeding, they may end a case by accepting the outcome of the Immigration Court decision to avoid more time in detention. The poor treatment of people in U.S. immigration detention centers[7] may discourage detainees from appealing their asylum decisions in order to improve their immediate conditions. For these reasons, the model likely would not predict appeal outcomes well for the individuals who do not choose to appeal.

That said, we undertake the analysis to see what can be learned. Of the 112,029 denied asylum cases that were not appealed, we predict only **7164 (6.4%)** would have been successful in their appeal and the remaining **104,865 (93.6%)** would be unsuccessful. This is much lower than the **32.4%** grant rate for the population that did appeal their case. Further, this is only based on information available to the model. The respondents who did not appeal likely have private information about their success odds that might have motivated them not to appeal. As such, their odds of appeal success might be even lower.

Put differently, of all cases that the model predicts to be successful in their appeals, an overwhelming majority of **84.3%** did appeal. This might imply a low ceiling on the number of additional respondents we might nudge into appealing successfully. Of our model's predicted successful appeals a large majority do appeal already. This suggests the usefulness of a predictive tool for advising denied applicants whether they should appeal.

References

1. “S. 643 — 96th Congress: Refugee Act of 1979.” www.GovTrack.us. 1979. March 22, 2018 https://www.govtrack.us/congress/bills/96/s643

2. “Immigration Court Practice Manual.” The United States Department of Justice, www.justice.gov/eoir/office-chief-immigration-judge-0.

3. “Board of Immigration Appeals Practice Manual.” The United States Department of Justice, www.justice.gov/eoir/board-immigration-appeals-2.

4. Chen, Daniel, et al. “Decision-Making under the Gamblers Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires.” The Quarterly Journal of Economics, vol. 131, no. 3, 1 Aug. 2016, pp. 1181–1242., doi:10.3386/w22026.

5. Dunn, Matt, et al. “Early Predictability of Asylum Court Decisions.” Proceedings of the 16th Edition of the International Conference on Artificial Intelligence and Law - ICAIL 17, 2017, doi:10.1145/3086512.3086537.

6. Chen, Daniel L., and Jess Eagel. “Can Machine Learning Help Predict the Outcome of Asylum Adjudications?” SSRN Electronic Journal, 2016, doi:10.2139/ssrn.2815876.

7. “Systemic Indifference | Dangerous & Substandard Medical Care in US Immigration Detention.” Human Rights Watch, 6 June 2017, www.hrw.org/report/2017/05/08/systemic-indifference/dangerous-substandard-medical-care-us-immigration-detention.

A Appendix

Table 5: Complete Feature List

Feature	Type
nat_grouped	object
lang_grouped	object
ij_code_grouped	object
Male_judge	float64
Year_Appointed_SLR	float64
Year_College_SLR	float64
Year_Law_school_SLR	float64
Government_Years_SLR	float64
Govt_nonINS_SLR	float64
INS_Years_SLR	float64
Military_Years_SLR	float64
NGO_Years_SLR	float64
Privateprac_Years_SLR	float64
Academia_Years_SLR	float64
judge_missing_bio	int64
years_since_judge_appointment	float64
years_since_law_school	float64
last_10_appeal_grant_by_judge	float64
last_10_appeal_grant_by_judge_nat	float64
lawyer	int64
defensive	float64
affirmative	float64
oral	float64
written	float64
case_type_string	object
original_dec_string	object
strCustody	object
strProbono	object
base_city_code	object
hearing_loc_match_base	object
datAppealFiled_year	float64
datAppealFiled_month	float64
datBIADecision_year	float64
datBIADecision_year_month	float64
comp_year	int64
comp_date	int64
comp_days_elasped_since_input_date	float64
input_days_elapsed_since_osc_date	float64