Predicting Asylum Granting Decisions in Immigration Courts

Dagshayani Kamalaharan and Jacqueline Gutman
New York University

Data Description and Feature Generation

Types of features we included or generated in our model:

- Demographic information about each immigration judge: *e.g. gender, years of experience, academic and professional background, President appointed by*
- Demographic information about each country of origin *(merged from WorldBank and Pew data): e.g. level of economic development, percentage of population Muslim, Christian, Hindu, geographic region and subregion etc.*
- Time series data about the previous n decisions for that judge *(or for that judge and nationality combination, etc.)*
  - We used Last Value Carried Forward imputation to fill in the previous decisions where data was missing, going back in time a maximum of 30 days
- The average grant rate over all previous decisions for that judge *(or grant rate for that judge in that year, for that judge and that nationality, etc.)*
- Information about the asylum seeker*: e.g. nationality of the asylum seeker, number of family members, defensive/affirmative asylum seeker, whether the asylum seeker has an attorney*
- Court-level information: *e.g.* average grant rate for previous decisions for all other judges in that court, daily case load of the court, daily case load of the judge

We generated all features related to the country demographics and re-computed the time series and average grant features to exclude data from the future from our features

Splitting the Data for Model Validation

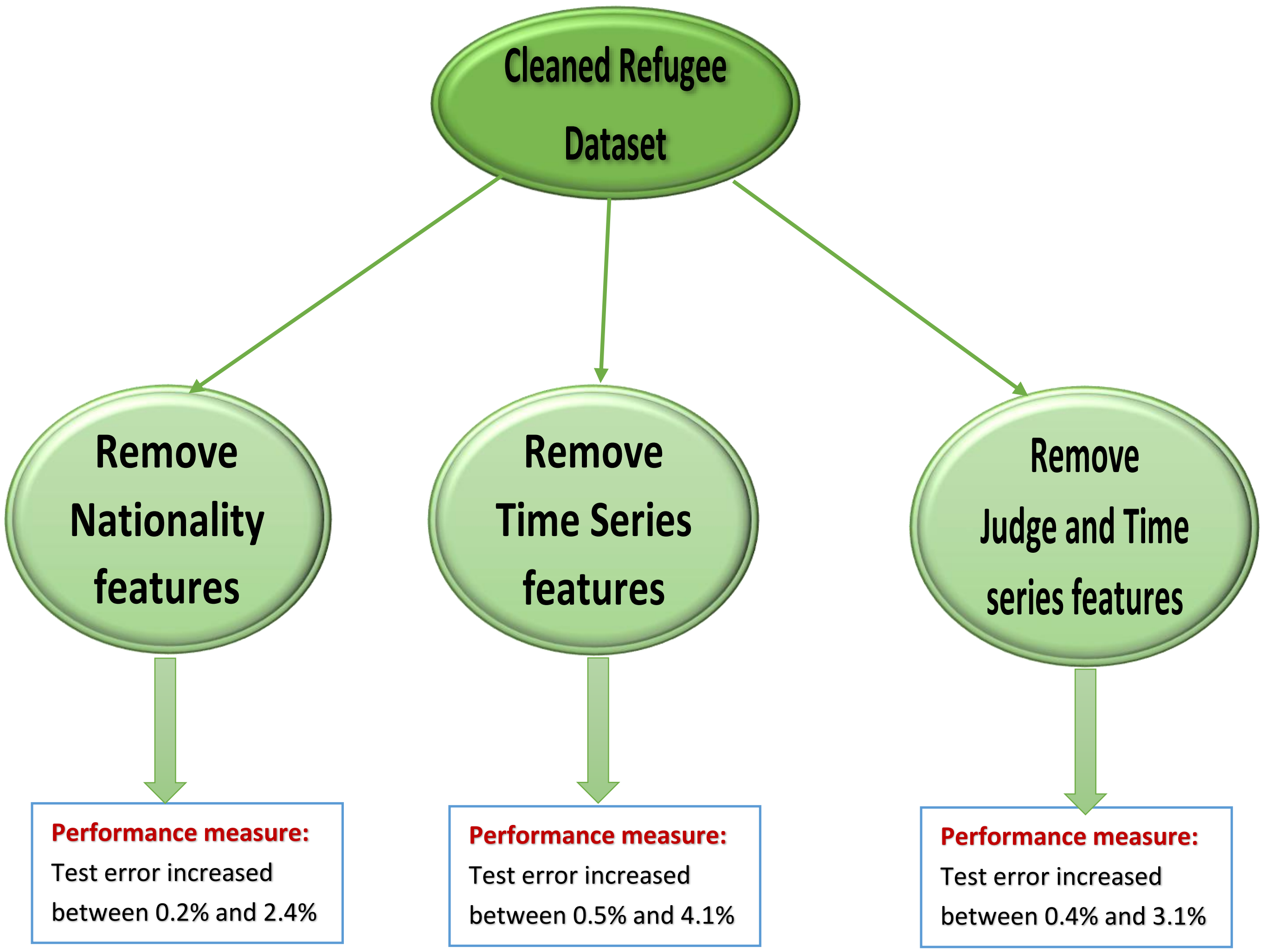
We are interested in the performance of our model on new judges – does the model generalize to judges not seen in the data, given that we have data on their past court decisions?

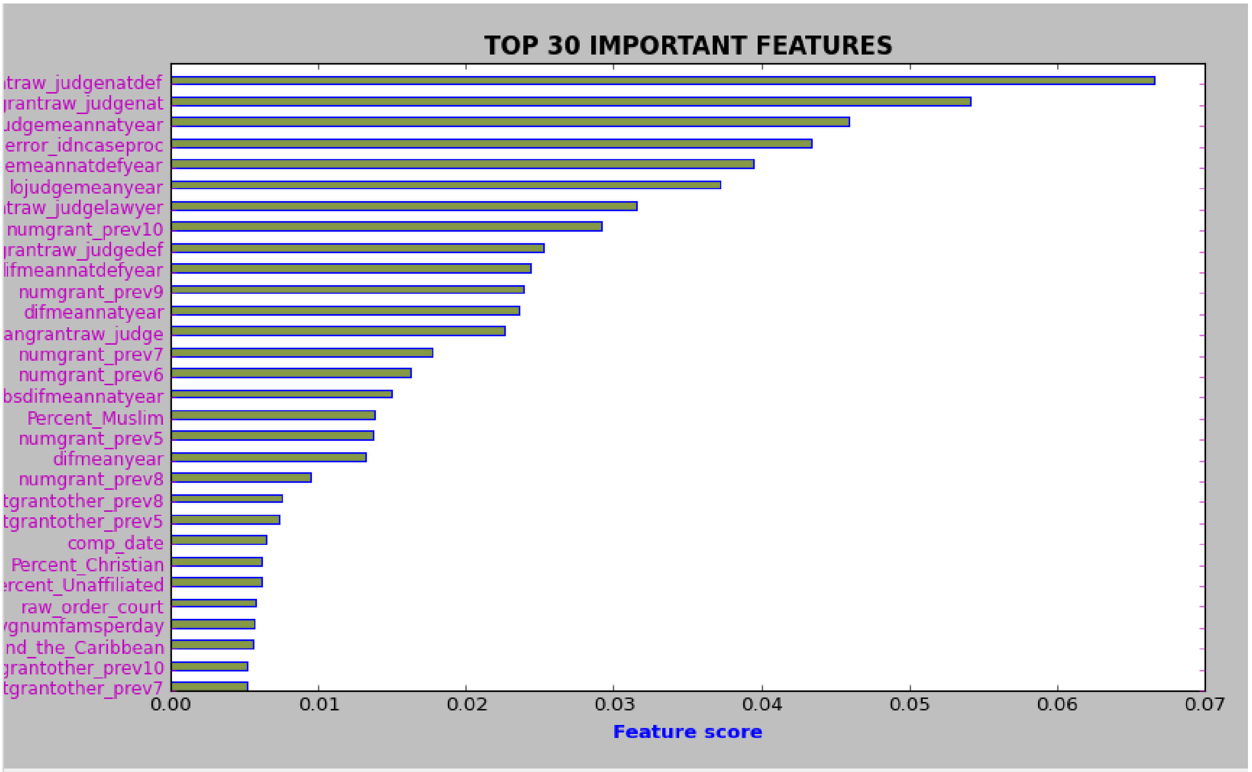
We split the data into training and test sets such that for any given judge, all of their cases were either included in the training set or in the test set.

We generated features summarizing the previous decisions made by that judge – how often did they grant asylum to individuals of nationality x in their previous cases? How often did they grant asylum overall?

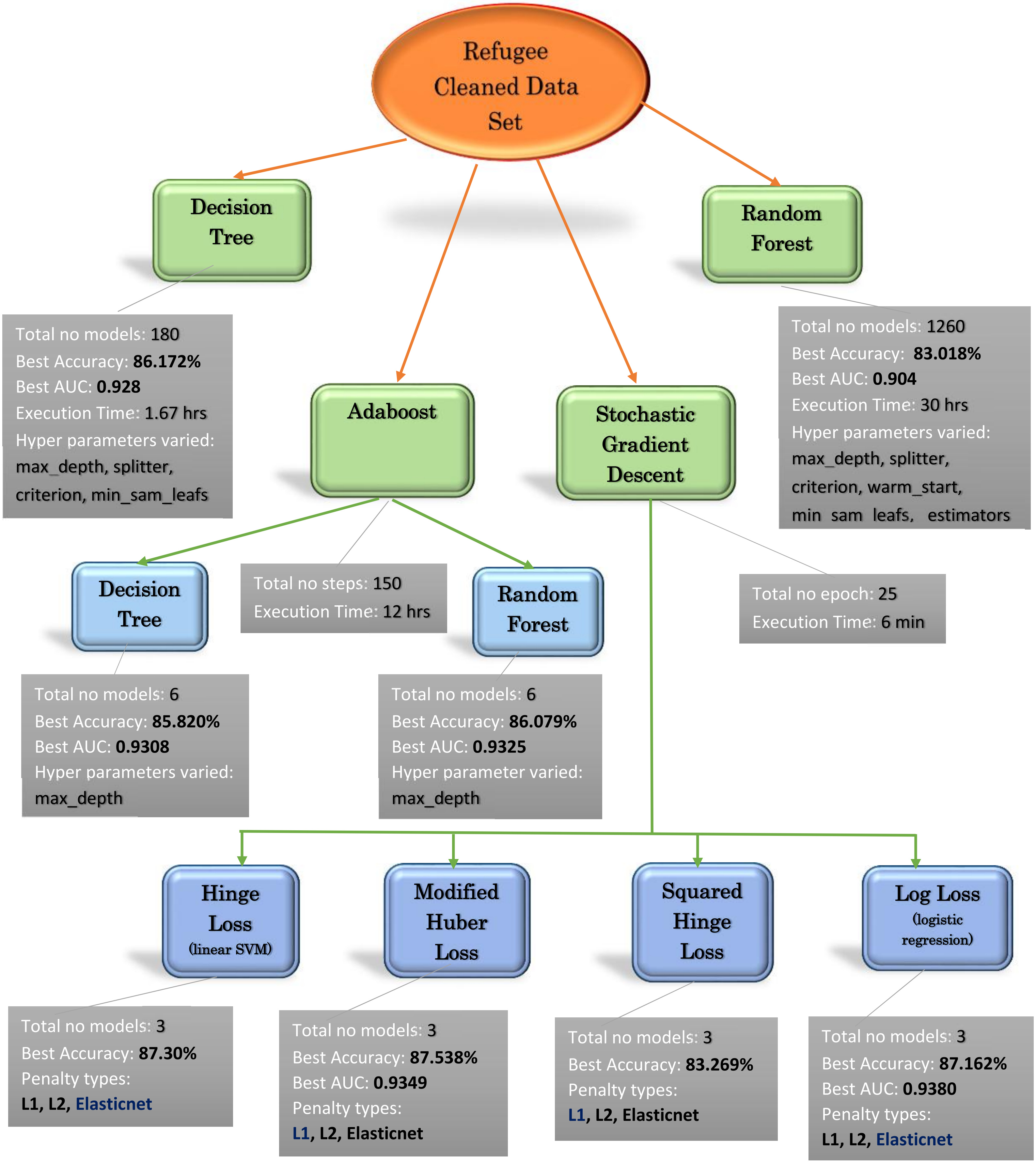
The time series aspect of the data is encoded in the averages over previous decisions for that court, judge, etc.

FEATURE IMPORTANCE



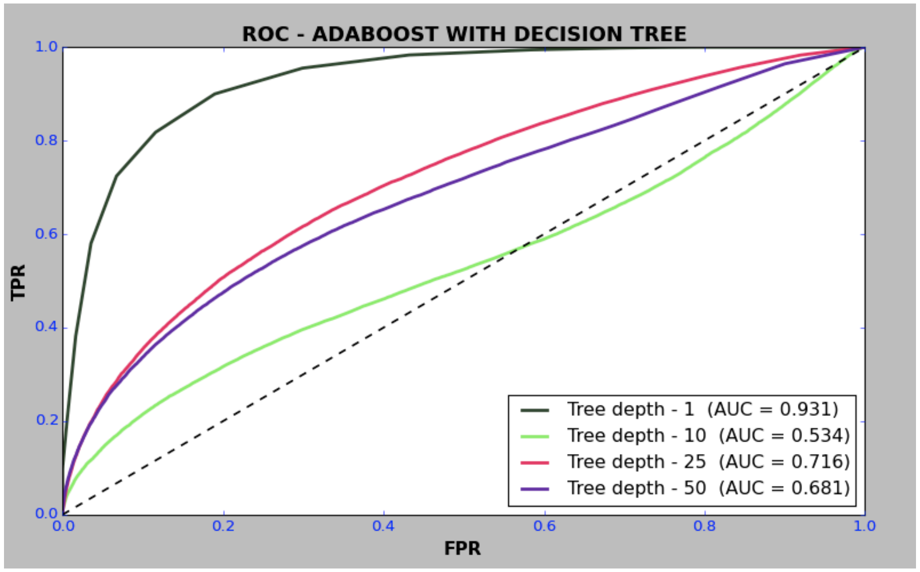


- Top 3 important features were related to the average grant rate for the judge for a specific nationality.
- Other important features were related to the number of grants for that judge for previous set of ‘x’ cases.
- Among the features we add, 3 of them came up within top 30 features: Percent_Muslim in asylum seekers country of origin, Percent_Christians in asylum seekers country of origin and whether the asylum seeker’s country of origin belongs to Latin_America_and_the_Caribbean region.

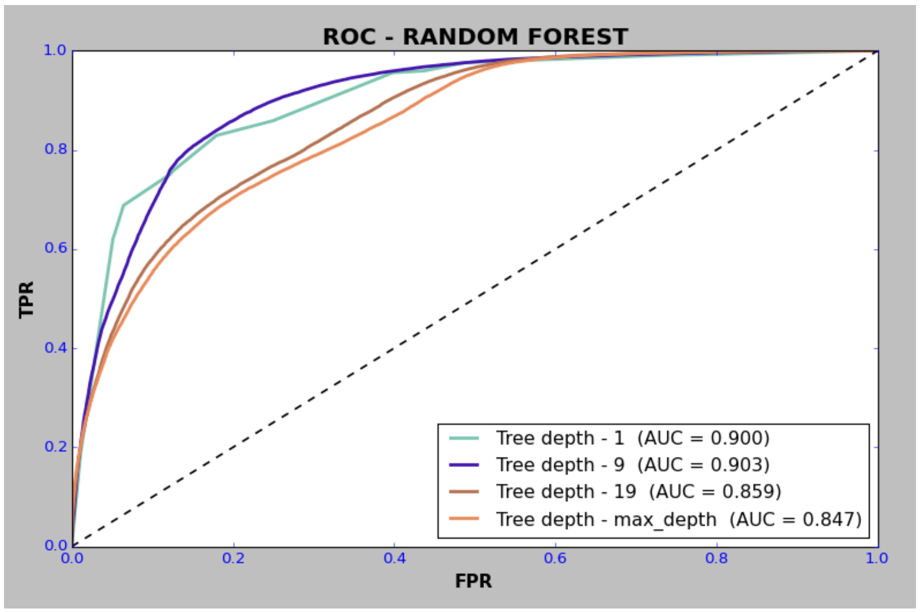


Refugee Cleaned Data Set
Decision Tree
Random Forest
Total no models: 180
Best Accuracy: 86.172%
Best AUC: 0.928
Execution Time: 1.67 hrs
Hyper parameters varied:
max_depth, splitter,
criterion, min_sam_leafs
Total no models: 1260
Best Accuracy: 83.018%
Best AUC: 0.904
Execution Time: 30 hrs
Hyper parameters varied:
max_depth, splitter,
criterion, warm_start,
min_sam_leafs, estimators
Adaboost
Stochastic Gradient Descent
Decision Tree
Total no steps: 150
Execution Time: 12 hrs
Random Forest
Total no epoch: 25
Execution Time: 6 min
Total no models: 6
Best Accuracy: 85.820%
Best AUC: 0.9308
Hyper parameters varied:
max_depth
Total no models: 6
Best Accuracy: 86.079%
Best AUC: 0.9325
Hyper parameter varied:
max_depth
Hinge Loss (linear SVM)
Modified Huber Loss
Squared Hinge Loss
Log Loss (logistic regression)
Total no models: 3
Best Accuracy: 87.30%
Penalty types:
L1, L2, Elasticnet
Total no models: 3
Best Accuracy: 87.538%
Best AUC: 0.9349
Penalty types:
L1, L2, Elasticnet
Total no models: 3
Best Accuracy: 83.269%
Penalty types:
L1, L2, Elasticnet
Total no models: 3
Best Accuracy: 87.162%
Best AUC: 0.9380
Penalty types:
L1, L2, Elasticnet

Prediction with Adaboost and Random Forest

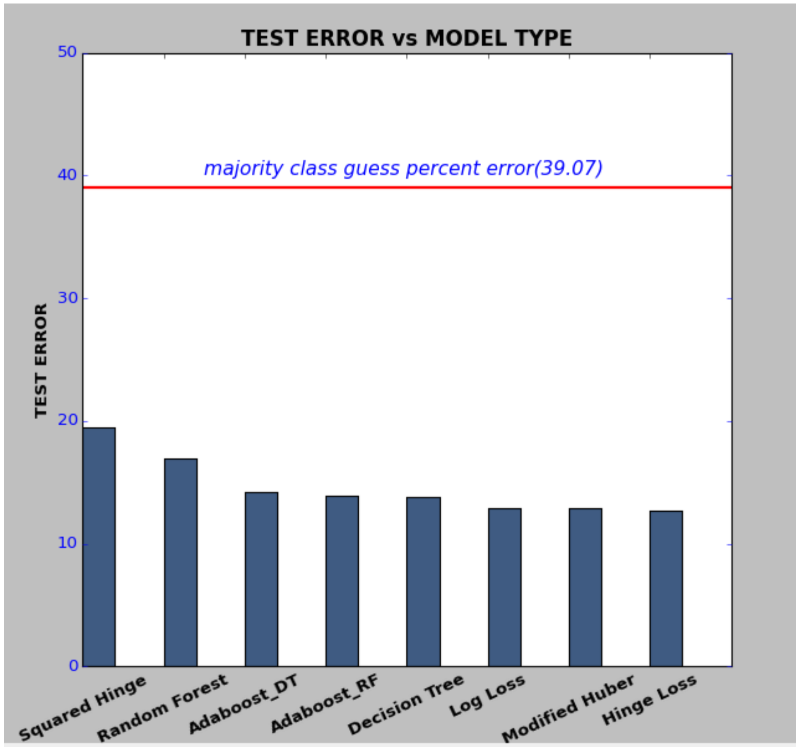


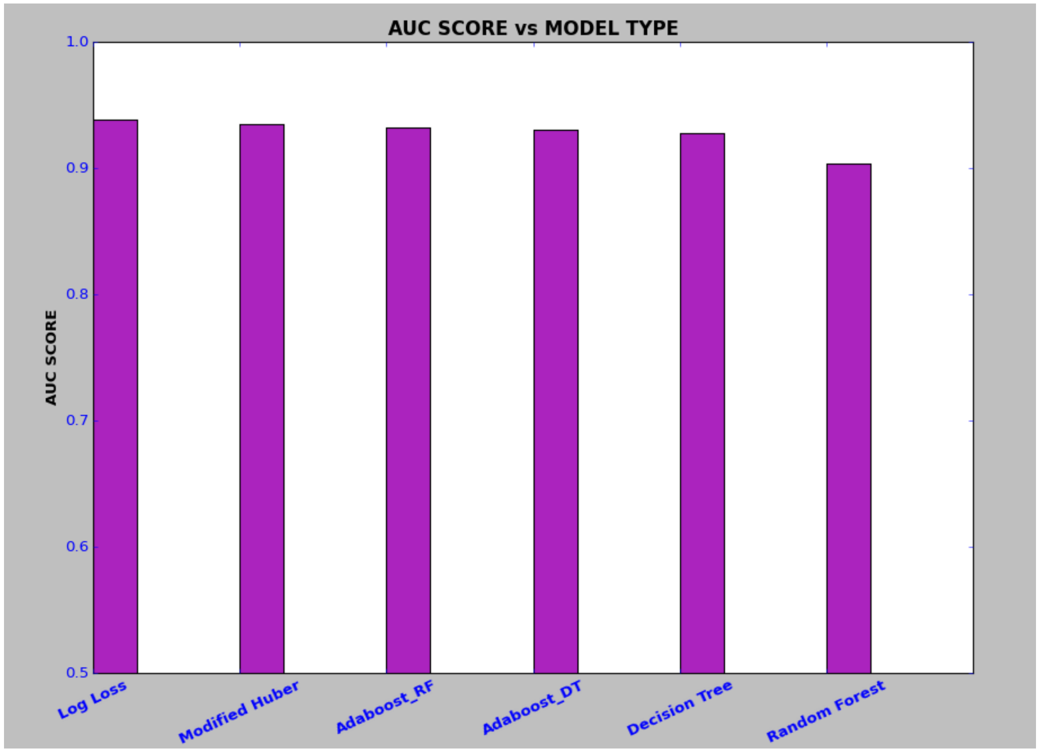
Adaboost with Decision Stumps performed very well – achieving a cross-validated AUC = .931 for predictions on new judges



Random Forest performed very similarly to Adaboost – predictions were strongly correlated and with optimal hyperparameters AUC = .903

Model Comparison





We see pretty consistent performance between models using hyper-parameters chosen in the validation process. The stochastic gradient descent algorithms, using logistic loss (with elastic net or L1 regularization) achieved the best trade-off between performance and computation time.

Leave One Judge Out Model Evaluation

For every judge in the data:

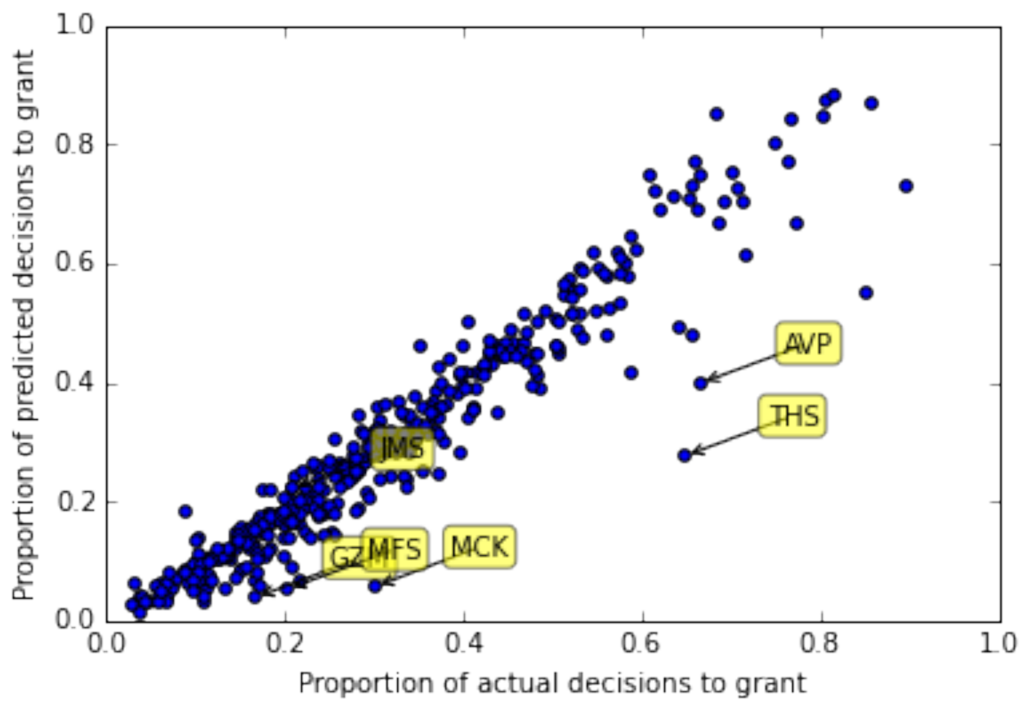
1. Fit a model to the data from all judges except judge X
2. Use the model to predict asylum decisions for judge X
3. Evaluate a performance metric (we used the phi coefficient) on the confusion matrix for judge X

Overall, model performance was good
Mean percent error across all judges: **11.8%**
Median pct error across all judges: **10.6%**
(for both Adaboost and random forest)
Mean phi square coefficient: **.465** (with random forest, strong association)
Mean phi square coefficient: **.502** (with Adaboost, strong association)

We want to identify the judges for whom our model did not generalize well in order to study their decision-making further:

We can see that for the judges who were not predicted well by the other judges' data, we typically underestimated their decision rate.

These judges were more liberal in granting asylum than expected given the features of the case and the behavior of other judges.



Limitations in our Analysis

No Ethnic conflicts Information

No Reliogious background Information

- Gender of the refugee unknown
- Age of the Refugee Unknown
- Marital Status of the refugee unknown
- Educational background of the refugee unknown

Conclusions

- We are able to achieve good performance. The strength of our performance raises potential concerns about whether there is any leakage, so we have carefully analyzed the most important features in our models and split the data into validation sets stratified by judge to allay these concerns.
- The most predictive features are related to the judge's past decision-making in the judge's most recent cases as well as in the previous cases featuring an asylum seeker of the same nationality. This indicates that country of origin is extremely important in asylum decisions, but certain judges may have implicit biases—country of origin is not interpreted the same by every judge.
- There is not much difference in performance from one model family to another, as long as we have cross-validated the hyper-parameters. However, computation time is a real limiting factor here, and we can do best by choosing the fastest-performing algorithms (e.g. linear SVM or regularized logistic loss with stochastic gradient descent algorithm)
- Overall our generalization performance to new judges is very good. We can predict a judge's decision without explicitly training on that judge if we are told the results of the judge's previous asylum cases.
- Some judges do break the trend, however, and we can use the phi squared/ contingency coefficient to identify those judges. Our model seems to drastically underestimate the grant rate for just a few judges, but further investigation is needed to understand why.