Optimizing Refugee Integration via Learned Assignment

Max Darling & Jason Ginsberg

Overview

» Our project improves the process by which refugees are assigned to resettlement locations within the United States.
» Current policy randomly assigns refugees based on capacity constraints and proximity.
» Our model, in contrast, learns to place refugees in regions of optimal employment opportunity.
» The results of our supervised learning approach demonstrates that two refugees of similar backgrounds face different outcomes based solely on where they are placed.

Dataset

The Annual Survey of Refugees
» 4,683 refugees who entered the US from 2011 to 2015
» 100 Questions
» 4 regions: Northeast, South, Midwest, and West
» Reduce to 32 features, separate by region, label by employment status, one-hot encode (categorical)

Table 1: ASR2016 Feature Categories

Marital Status	Ethnicity	Country of Origin	Country of Birth
Age	Gender	Household Relation	Arrival Date
Education	Profession	English Skills	Health Status

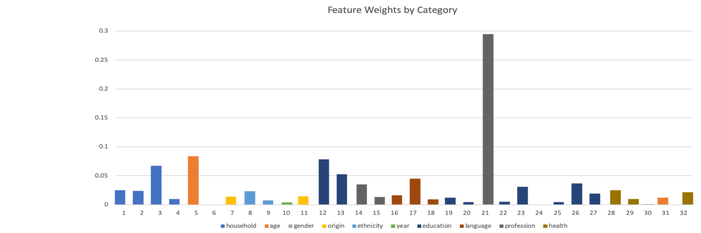
Extreme Gradient Boosted Trees

» Goal: predict a refugee's employability given a region
» Decision tree: classify by splitting the input space of a node
Gradient Boosting: ensemble of multiple decision trees, correct residual error via iterative augmentation of tree.
» XGBoost: 2nd order derivative of loss with L1 and L2 norm

Our implementation
» Train 1 employability classifier per each region via XGBoost over 200 iterations
» Logistic loss for classification at each split
» Output logits at inference (via Softmax function)

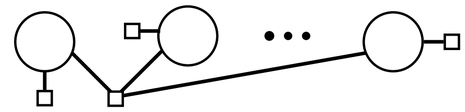
Parameter Search and Feature Selection

» 80-20 train-test split, 5-fold cross-validation
» 40 uniform random search over 5 parameters
» Select features by pruning lowest magnitude weights



Refugee Assignment

» Allocate refugees to 1 of 4 US regions such that employment probability is maximized and regional constraints are obeyed
» Factor graph: N-ary regional population constraint, unary employment probability factors.
» Optimize variable assignments (refugee allocations)



Results

Cross-Validation

Region 0: learning rate=0.068, max_depth=2, avg accuracy=0.910

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.882	0.907	0.924	0.915	0.924

Region 1: learning rate=0.033, max_depth=2, avg accuracy=0.900

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.897	0.901	0.906	0.919	0.879

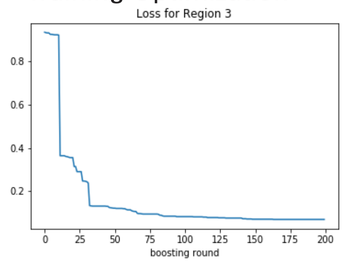
Region 2: learning rate=0.040, max_depth=2, avg accuracy=0.928

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.936	0.932	0.932	0.923	0.918

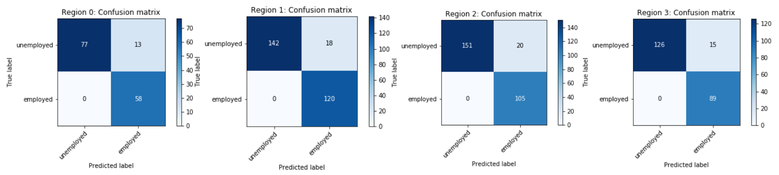
Region 3: learning rate=0.064, max_depth=3, avg accuracy=0.923

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.935	0.935	0.918	0.902	0.923

Training Optimization



Precision-Recall



Weighted F1 Scores
» Northeast: 91.316%
» South: 93.604%
» Midwest: 92.843%
» West: 93.548%

Percent of refugees whose employability improved after our re-assignment
» 71.948%

Analysis

» 66% improvement over baseline accuracy
» Model generalizes well based on cross-validation
» Model simplifies to 20 features
» Re-assignments greatly improve likelihood of employment
» Complex regional factors greatly affect employment
» Professional information, language skills, age, and education matter most
» Gender, ethnicity, origin, and year of arrival matter least
» Issues: accuracy of data, 2nd order assignment effects, small size of dataset, employment-only, family separation
» Greater nunber of false positives (worse misprediction-type)

Future Work

» Evaluation of feature pruning on accuracy across models
» More complex constraint-based model accounting for families, regional laws, and policies
» Application of model on synthesized inputs in order to make policy recommendations about future refugee influxes
» Larger dataset with state-level rather than regional data
  » Inter-dependent factors even more complex
  » Less bias in outcomes based on region of origin
  » Greater confidence in generalizability