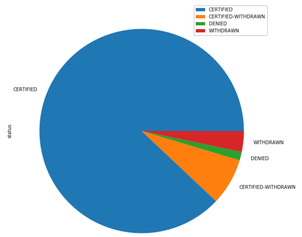
Classifier Performance Analysis through H-1B Visa Certification Status Classification

Yunhao Yang | CS 370F Undergraduate Research and Writing | University of Texas at Austin, Department of Computer Science

Background

The H-1B is a visa in the United States under the Immigration and Nationality Act, section 101(a)(15)(H) that allows U.S. employers to temporarily employ foreign workers in specialty occupations. The application specifies the applicant's professional knowledge by requiring a bachelor's degree or equivalent work experience.

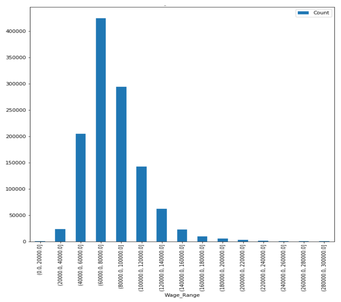


Based on the dataset of 2017-2018 H-1B Application Statistics, around 88 percent of applications were certified and moved to the next stage. The second-largest portion was Certified-Withdrawn, which comprises around 7.4 percent. Around 3.3 percent of applicants withdrew their application before it was reviewed. The last 1.3 percent of applications were denied.

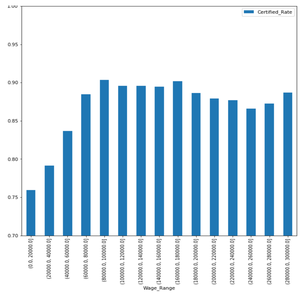
Top 10 Jobs with Highest Certified Rate

Job	Application Number	Certified Rate
COMPUTER SYSTEMS ANALYST	31164	98.86
COMPUTER SYSTEMS ENGINEERS/ARCHITECTS	2645	98.34
SOFTWARE QUALITY ASSURANCE ENGINEERS AND TESTERS	4169	96.86
WEB DEVELOPERS	9492	96.5
INFORMATION SECURITY ANALYSTS	5472	91.7
MANAGEMENT ANALYSTS	25955	90.72
COMPUTER NETWORK ARCHITECTS	3668	90.65
FINANCIAL SPECIALISTS, ALL OTHER	5626	90.44
COMPUTER OCCUPATIONS, ALL OTHER	114407	90.2
PHYSICAL THERAPISTS	5792	90.02

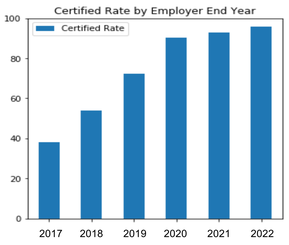
Certified Applicants Wages



Certified Rate vs Wage Range



The year that applicants finish their employment is related to the certified rate. As demonstrated by the organization of the data by employment end year, the majority of the applicants will end their employment in 2017 ~ 2022 (the dataset is from 2017). There was a positive correlation between Employment End Year and certified rate. The earlier applicants end their employment, the lower the chance their application will be certified.



Objective

The objective of this research is to be able to predict whether an H1B Visa application will be certified or not, based on the details in the application. This could enable those applying for H1B Visas to improve their applications if it is predicted that they will not be certified.

Methodology

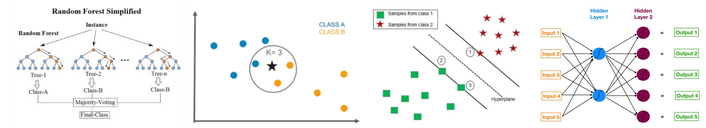
Feature Engineering:

- Remove features that are not related to the certification status, or highly correlated to another chosen feature.
- Identify anomalies using statistical anomalies detection and remove records that contain features out of three standard deviations.

Train each baseline model first and then integrate their results in the final model

Training data: In order to avoid the effect of class imbalance, each model randomly chooses 50,000 records on both certified group and denied/withdrawn group to form a balanced data with 100,000 records.
Testing data: In a random selection of 10,000 records, around 88% of the records are certified.

Baseline Models



Decision Trees are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Random Forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset, and uses averaging to improve the predictive accuracy and control over-fitting.

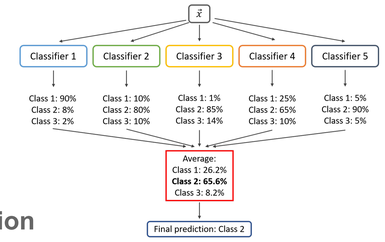
AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm. It can be used in conjunction with many other types of learning algorithms to improve performance.

K-Neighbors Classification is computed from a simple majority vote of the nearest neighbors of each point.

Support-Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function. It is a class of feedforward artificial neural network

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.



Integration

Voting Classifier is one of the simplest ways of combining predictions from multiple machine learning algorithms. It is a wrapper for a set of different algorithms that are trained and evaluated in parallel in order to exploit the different strengths of each algorithm.

Stacking Classifier can be described as an ensemble learning technique where the predictions of multiple classifiers are used as new features to train a meta-classifier.

Results

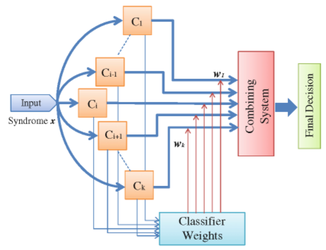
Accuracy indicates the overall percentage of the test records that were accurately predicted.
Precision is the fraction of the number of correct predictions over the number of the applications that are predicted denied.
Recall is the fraction of the number of correct predictions over the number of the applications that are actually denied.

Algorithm	Precision	Recall	Accuracy
decision tree	74.34	27.52	89.99
random forest	97.73	24.73	90.74
ada boost	97.73	24.73	90.74
nearest neighbor	68.32	18.02	88.97
svc	92.22	19.41	89.96
neural network	95.76	12.94	89.3
logistic regression	93.67	12.12	89.17
voting	99.23	21.49	90.59
stacking	48.66	28.93	87.85

The majority of the baseline classifiers get precision above 90%, except for the decision tree and nearest neighbor (around 70%). The recalls are below 30%. In terms of accuracy, two ensemble classification algorithms- Random Forest and AdaBoost- achieve the best accuracy and precision. Decision Tree works best in terms of recall.
The voting classifier achieved the highest precision at 99.23%. Though its recall is slightly lower, the voting classifier looks best overall in terms of accuracy, precision, and recall.

Conclusion

Final Model



The research develops a way of classifying H-1B visa certification status, which achieves 91% accuracy, and 99% precision.

The final model is constructed by seven distinct classifiers listed in baseline models. Each classifier is tuned independently. Then, use Voting Classifier to integrate the seven trained classifiers and tune it to fit the training data.

Ensemble classifiers have a significant improvement in performance compared to other classifiers. In this research, Random Forest and AdaBoost achieve the best performance among the baseline models.