Predicting the Likelihood of Acceptance of an H-1B LCA

Soorya Paturi

What is an H-1B LCA?

-The H-1B visa application is a highly competitive application process

-A preliminary step is to apply for an H-1B LCA (Labor Condition Application), which is done by the employer

-The H-1B LCA application's main goal is to ensure is to ensure fair work conditions

Why is it important to predict application outcome?

-The client is any employer who is looking to employ foreign born workers

-The client should care about H-1B LCA outcomes because they do not want to spend time looking for candidates that will eventually not be certified

-With help of a classifier, the employer will be able to spend more time finding competitive candidates for a position instead of having to resubmit an H-1B LCA application.

Background on Data Set

-The data set was collected by the United States Department of Labor's Office of Foreign Labor Certification.

-The data set contains on year's worth of data from fiscal year 2017 and approximately 625,000 records of H-1B LCA application results.

-Data set: www.foreignlaborcert.doleta.gov/performancedata.cfm

Data Wrangling

-Imbalanced data
-Imbalance reduced with relabeling
-Missing values accounted for
-Outliers explored

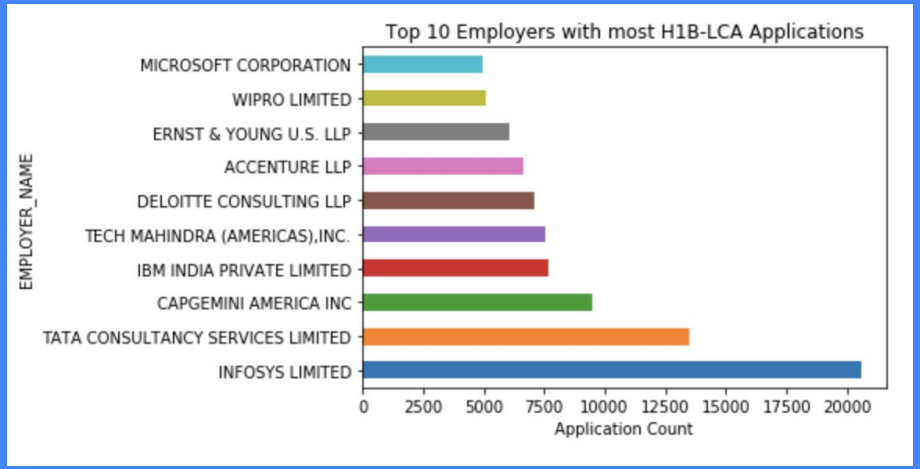
TABLE I
CASE STATUS LABELS

Case Status	Application Count
Certified	545694
Certified-Withdrawn	49704
Withdrawn	20772
Denied	8480

Employer Name

-The top employers that submit H-1B LCA applications each submit at least 5,000 applications with a maximum of just over 20,000 submitted

-The maximum application count is claimed by Infosys Limited

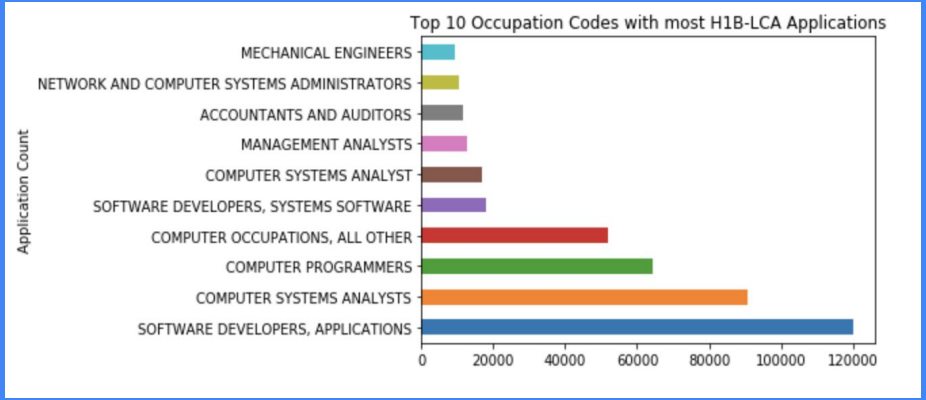


Occupation Code

-The occupation code field gives more information

-The main goal of the H-1B LCA application is to ensure fair work conditions

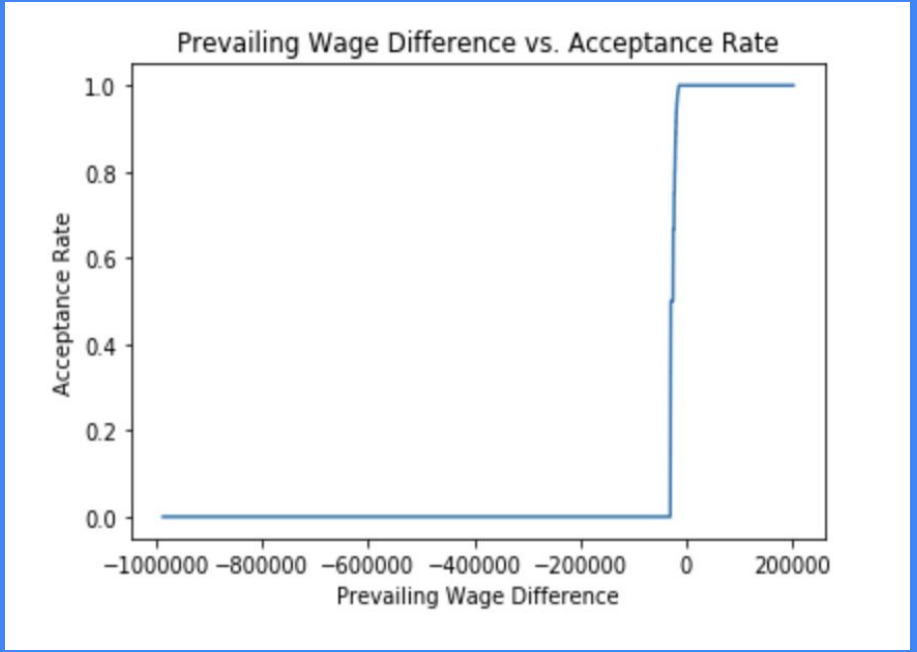
- Top occupation codes are inhabited by technology related roles: software developer and computer systems analyst



Prevailing Wage Difference

-The prevailing wage difference is defined as the difference between the prevailing wage in the OFLC wage database and the prevailing wage listed in application

-Acceptance Rate = the number of accepted applications per prevailing wage difference over the total number of applications per prevailing wage difference

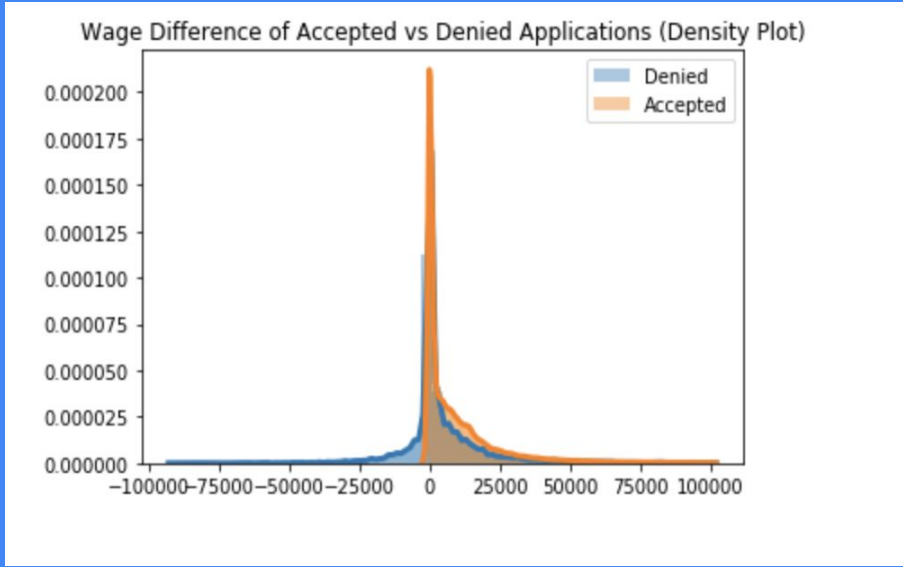


Wage Difference

-Wage difference is a calculated field that was thought to possibly be a predictor for application outcome

-A sign of an unfair work condition is if the wage offered to a foreign worker is less than the prevailing wage

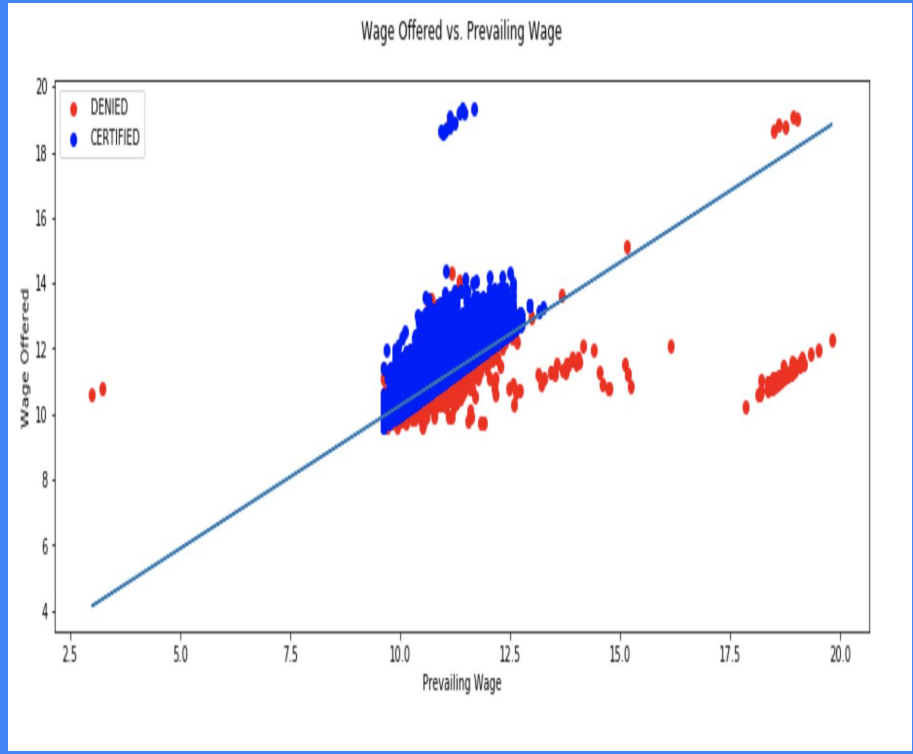
-The wage difference field is the difference between the wage offered and the prevailing wage



Wage Offered vs. Prevailing Wage

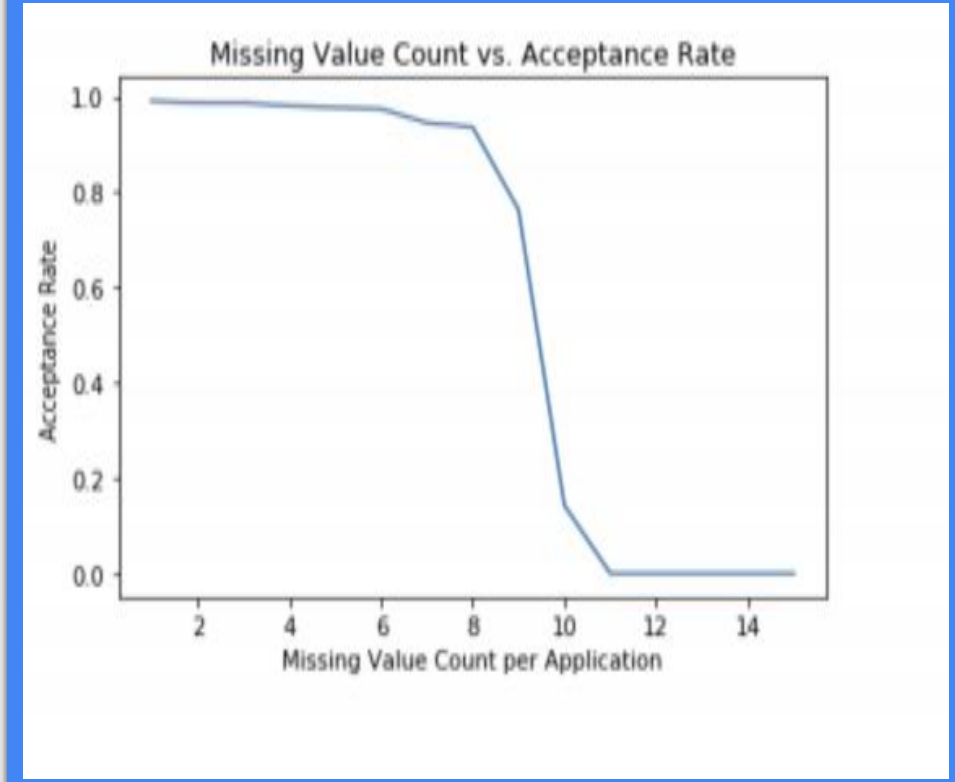
-Combination of wage offered and prevailing wage affects outcome

--Outcome may be linearly separable



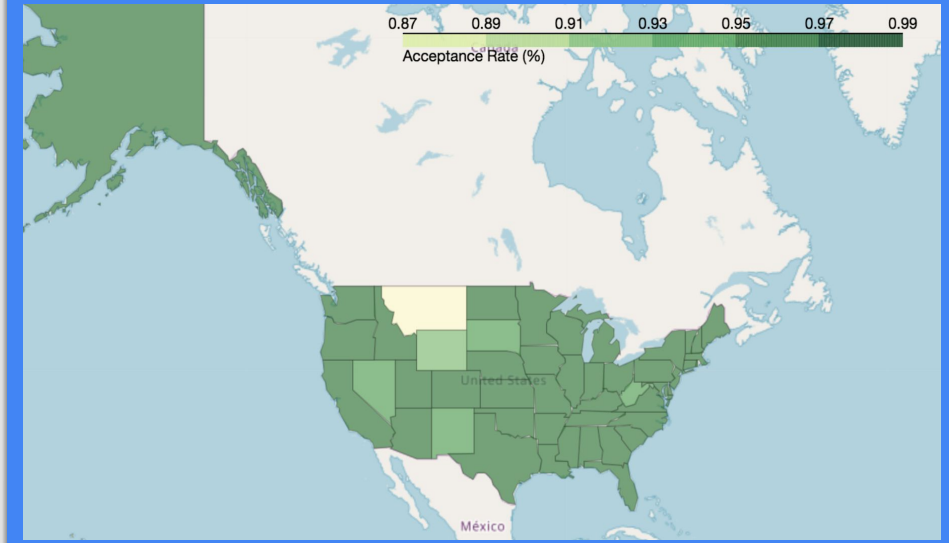
Missing Value Count

-Missing value count is a blank in application
-Strong negative correlation
-Might be a strong predictor



State

-Utilizing a heat map, the the distribution of acceptance rate across states was visualized
-Color intensity shows the variation.
-Higher acceptance rate is indicated by a darker shade.
Montana, Nevada, New Mexico, South Dakota, Wyoming, West, and Virginia have the lowest acceptance rates.



Feature Selection

-All relevant features were used

-Redundant features were removed to prevent overfitting

-Missing values replaced with None or mean value

-Categorical fields were one hot encoded & target encoded

Models

Benchmark-Label all applications as the majority class

Logistic Regression-L1 penalty (feature selection), stochastic average gradient descent solver

Gaussian Naive Bayes-No hyperparameters

Decision Tree-Single tree with an optimal depth of 3.

Random Forest-250 estimators, minimum sample leaf=1

Gradient Boosted Tree-250 estimators, minimum sample leaf=1, learning rate =1

Results

-AUC is looked at instead of accuracy because of class imbalance

-Classifier with highest AUC: Random Forest. Gradient Boosted is close in 2nd place but has highest F1-score

TABLE II
BENCHMARK CLASSIFICATION METRICS

Model	Accuracy	FPR	FNR	BER	F1-Score	AUC
ZeroR Baseline	0.98596	1.0	0.0	0.5	0.5	0.5

TABLE III
MACHINE LEARNING CLASSIFICATION METRICS

Model	Accuracy	FPR	FNR	BER	F1-Score	AUC
Logistic Regression	0.98448	0.98231	0.00176	0.49203	0.51	0.70458
Naive Bayes	0.95695	0.82405	0.03198	0.42802	0.54	0.58880
Decision Tree	0.98821	0.75463	0.00134	0.37799	0.68	0.62475
Random Forest	0.98726	0.86851	0.00055	0.43453	0.61	0.80291
Gradient Boosted Classifier	0.98770	0.72041	0.00190	0.36116	0.70	0.79205

Recommendations & Conclusion

Top unique features in random forest and gradient boosted classifier:
EMPLOYER_NAME, WAGE_DIFFERENCE,
JOB_TITLE, WORKSITE_CITY,
AGENT_ATTORNEY_NAME, PW_SOURCE_OTHER,
NAICS_CODE, and PREVAILING_WAGE, WAGE_RATIO,
WAGE_RATE_OF_PAY_TO, WAGE_RATE_OF_PAY_FROM,
NEW_EMPLOYMENT, and PW_SOURCE_OES.

.

Recommendations: **1.)** Offer more than the prevailing wage **2.)** Highest maximum wage offered and number of first time applicants should be high, **3.)** Use a prevailing wage from OES 4.) Hire an attorney with a history of high H-1B LCA acceptance rates.

Future Work

-More samples in minority class

-Majority class can be downsampled

-More hyperparameters can be tuned

-Additional fields in the application can be retrieved