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

Soorya Paturi (sopaturi@gmail.com)

***Abstract*—The H-1B visa application is a highly competitive application process that rewards 85,000 applicants per year with an H-1B visa out of approximately 650,000 applications annually. A preliminary step needed to apply for the H-1B visa 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 that the employer is providing fair work conditions for the foreign born worker who is applying for the H-1B visa compared to a citizen who would be applying for the same job. The H-1B LCA application contains information that pertain solely to the employer, not the employee. An application varies in many categories such as job title, employer name, prevailing wage, and work site city. The goal is to predict the outcome and the likelihood of acceptance of the H-1B LCA application in order to inform an employer what they can change in their application to increase acceptance rate.**

I. Introduction

THE H-1B visa is is an immigrant visa that allows foreign born workers to enter the United States and work temporarily for up to three years with a possibility of extension for up to six years. In order for a worker to receive an H-1B visa, an employer must offer them a position and then submit an H-1B visa application with the Department of Immigration. H-1B visas are commonly applied to by international students are who are looking to work after completing their education in the United States.

A preliminary step before an H-1B visa application can be filed is the submission of the H-1B LCA (Labor Condition Application) to the Department of Labor. The H-1B LCA contains information about the job title being offered, duration of the job, whether job is full time, rate of pay, location of the job, and the prevailing wage in the area. The purpose of the H-1B LCA is to bind the employer into paying a fair wage and providing benefits to a foreign born worker that greater than or equal to the prevailing wage and benefits in the applicant's location.

Currently, there is not a published deterministic cause for H-1B LCA application denial. However, the objective is to find a relationship between responses to fields of the application to application outcome, also known as case status, so that the likelihood of acceptance can be predicted.

The client, or an entity that is interested in the findings of this report, is any company that wishes to hire a foreign born worker. The client should care about H-1B LCA outcomes because they do not want to spend time filing and looking for candidates that will eventually not be certified with a successful H-1B LCA application. Based on my analysis, 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. The client will have a better chance at having their H-1B LCA certified by using the findings to change the application responses to responses that have a high H-1B LCA certification rate.

The outline of this report is summarized here with Section II introducing the data set. In section III, the data wrangling of the data set is summarized by explaining how the data set was obtained, cleaned, and wrangled. In section IV, exploratory data analysis of the data is summarized by a field by field analysis that determines which fields contain values that may be predictors for application outcome. In section V, the prediction task is described. In the last section, section VI, conclusions and next steps are discussed.

II. Data Set

The data set was downloaded from, www.foreignlaborcert.doleta.gov/performancedata.cfm [1]. The data set was collected by the United States Department of Labor's Office of Foreign Labor Certification. The data set contains one year's worth of data from fiscal year 2017 and approximately 625,000 records of H-1B LCA application results.

In order to prevent employers from having their H-1B LCA application denied, employers should be able to identify and communicate what the employment characteristics are for a high acceptance rate and low acceptance rate. If the reason an application is denied is mainly due to the fields defined by the employer such as wage offered or work site location, the employer will know what to change in their application in order to be certified.

The field defined as case status or application outcome is divided into four classes: (1) Certified (2) Denied (3) Withdrawn (4) Certified Withdrawn.

Based on the application counts listed in Table 1, it becomes clear that the number of certified applications outnumber the amount of denied applications. There are several steps that can be taken in the data wrangling step that can simplify the imbalanced data set.

TABLE I
CASE STATUS LABELS

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

A selected number of fields, six to be exact that may cause confusion, are described below so that interpretation of the data set above is easier. The full description of all fields is shown in the link below.

-VISA_CLASS: Indicates the type of temporary application submitted for processing. R H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as Program in prior years.

-SOC_CODE: Occupational code associated with the job being requested for temporary labor condition, as classified by the Standard Occupational Classification (SOC) System.

-NAICS_CODE: Industry code associated with the employer requesting permanent labor condition, as classified by the North American Industrial Classification System (NAICS).

-PREVAILING_WAGE: The average wage for the occupation in the same MSA (Metropolitan Statistical Area Code).

-WAGE_RATE_OF_PAY_FROM: The wage offered by the employer.

-LABOR_CON_AGREE: Y = Employer agrees to the responses to the Labor Condition Statements as in the subsection; N = Employer does not agree to the responses to the Labor Conditions Statements in the subsection.

Description of all fields in data set can be found at http://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2017/H-1B_FY17_Record_Layout.pdf [2].

III. DATA WRANGLING

It is evident from Table 1 that the data set is highly imbalanced towards certified applications. A smaller sample of the certified applications may need to be taken when comparing to the denied applications so that a machine learning model has a more balanced data set. The applications that are labeled with a certified-withdrawn status are still certified but later withdrawn by the employer when the employer decides not to hire the applicant. The certified-withdrawn applicants can be treated as certified and were relabeled as certified. The acceptance rate for applications once this relabeling occurred was 98.6%. The withdrawn applications were removed because this action was taken by the employer and does not result in an application outcome determined by the Department of Labor.

There are 624,650 rows in the data set. A lot of these rows contain missing values. When all rows with any missing values are deleted, only 98,155 rows remain. Therefore, one should be cautious before deciding to remove rows based on missing values.

A missing value can be due to two reasons, the applicant filling out the H-1B LCA application chose not to input a value or the mechanism by which the data set was created contained scraping errors that resulted in only some values being picked up. After reviewing the Office of Foreign Labor Certifications iCERT Visa Portal System, where the application is filled out, it is clear based on the application instructions, that all the fields in the application are discretionary, meaning the applicant did not have to enter a value if they did not want to. Because missing values are due to the optional nature of the application, having a missing value may affect application outcome.

Removing rows with missing values may skew the data set towards certified applications because

denied applications would be filtered out by removing rows with missing values in certain columns. Therefore no columns in the H-1B LCA application data set had the entries for their corresponding missing values removed. In many cases, missing values were imputed or labeled to signify there exists a missing value at that position.

If a missing value is a categorical variable, the missing value was replaced with the string 'None'. If the missing value is a numerical variable, the missing value was replaced with an imputed average value. Many missing values for the attorney name field were represented with a comma. These commas were replaced with the string 'None'. Currently, the industry from which the job is in is represented by a NAICS_CODE. The NAICS codes were converted to an industry name. NAICS code values were first replaced in the dataframe with industry names and then the codes without an entry in the dictionary were replaced with the string 'None'. Replacing the numerical codes with industry names makes the field more readable and easy to understand.

There is one more field left that is numerical and had a missing value. It is the prevailing wage field and it only had one entry with a missing value. The value cannot be replaced with 'None' because the field is a numerical variable. Therefore, the missing value is replaced with 0.00 because this application was denied and all the other applications that had a prevailing wage of 0.00 were denied. By replacing the missing value with 0.00 the application is grouped with other denied applications and maintains its own denied identity.

A possible reason for a declined H-1B LCA application is the difference between the prevailing wage and the wage offered by the employer. If the wage difference is negative, it may be a reason for the application receiving a rejection because a lower wage offered than the prevailing wage is considered unfair. The units for the prevailing wage and the offered wage may be in different units such as monthly, hourly, and yearly. The units for wage were all converted to yearly so that prevailing wage and wage offered units matched.

Another reason the application could be rejected is that prevailing wage listed is not correctly listed. By converting all of the salaries to annual salaries, the salaries can be compared to a database with prevailing wage values.

Because wage offered is hypothesized to be a large determining factor in case status as part of the data wrangling step, outliers must be identified and determined if they should be removed because they skew the data.

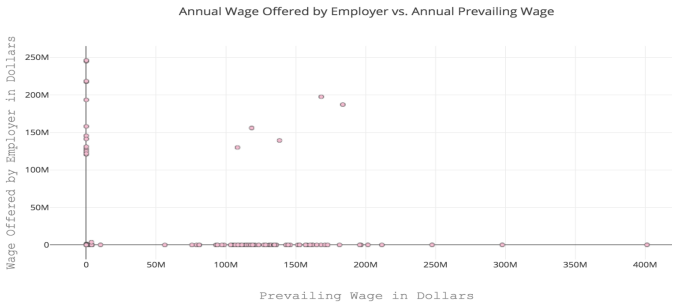

Fig. 1. Scatter plot of annual wage offered vs. annual prevailing wage for each Application

It is clear from Figure 1 that there are outlier values in the prevailing wage axis. There are some entries with a prevailing wage greater than 50 million. These are not correct prevailing wages and should be removed from the data set. However, when looking at the case status of these applications with a prevailing wage greater than 50 million, it is seen that all the applications are denied. It can be hypothesized that an unrealistic prevailing wage listed, one that is not commensurate with wage offered, can lead to an application denial. Removing these rows would remove rows that received a denial status and make the data set even more unbalanced towards the certified entries. Therefore, the outliers are kept because prevailing wage may be a large factor in the application case status.

To summarize the data wrangling section, the dependent variable case status was relabeled such that all values were labeled certified or denied. Missing values were dealt with and outliers were explored for two numerical variables, wage offered and prevailing, that are hypothesized to affect application outcome. The next section will explore the relationship between the independent variables, also known as the application responses, and how they might affect application outcome.

IV. Exploratory Data Analysis

The distribution of values in a field that may affect application outcome are explored in this section. A single field's entries can correspond to

many different acceptance rate values. This section is structured as a field by field analysis that identifies which fields contain values that may be predictors for application outcome.

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. As seen in Figure 2, this maximum application count is claimed by Infosys Limited.

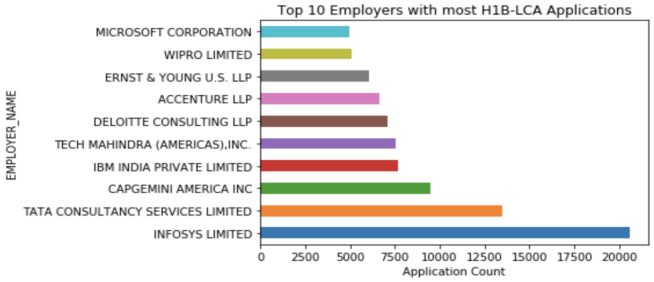

Fig. 2. Top 10 employers by application count

The word cloud in Figure 3 depicts the top employers by application count [3]. It is interesting to see that technology and consultancy companies are the majority of the top employers per application count. The mean application count per employer is 8.56 applications while the standard deviation is 128.79 applications. This means that the company with the 10th most applications per employer, Microsoft, is an astounding 39 standard deviations away from the mean. The large companies in this data set take up a disproportionate amount of applications submitted.

Fig. 3. Word Cloud of Top Employers

There are 70,537 unique employer names in the H-1B LCA data set. The different employers were sorted according to their acceptance rates. The acceptance rate is defined as the number of accepted applications per employer over the total number of applications per employer. From Figure 4, we see that the graph of every 300th serialized employer vs acceptance rate is not a uniform distribution. And hence, employer name can be a good predictor of case status in the prediction task.

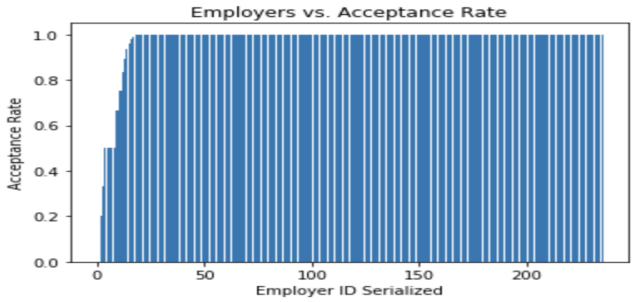

Fig. 4. Employers sorted by acceptance rate in increasing order

Occupation Code and Job Title: The occupation code field gives more information regarding the type of work that the employer will be providing as opposed to agent representation which does not give information about the type of work offered. The main goal of the H-1B LCA application is to ensure fair work conditions. Some job occupations have more unfair work conditions than others and the acceptance rate of each occupation should be non uniform. The graph in Figure 5 summarizes the top occupation codes. It is apparent that the top occupation codes are inhabited by technology related roles such as a software developer and computer systems analyst.

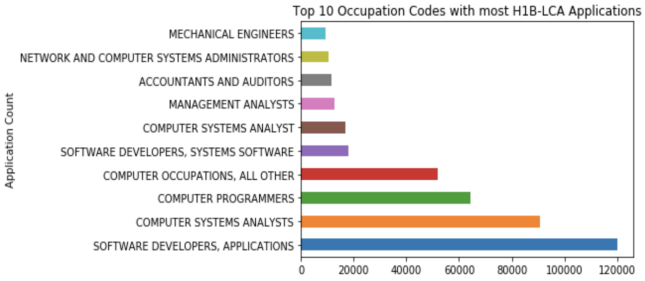

Fig. 5. Top Occupation Codes in submitted applications

The word cloud in Figure 6 depicts the top employers by application count. It is interesting to see that technology related occupations are the majority of the top occupations per application count.

There are 862 unique occupation codes in the H-1B LCA data set. There are more unique job ti-

Fig. 6. Word cloud of top occupation codes

tles than occupation codes because each occupation code can have many job titles and this is seen in the fact that the unique job title count is at 91,646, approximately 100x greater than occupation code count. The different occupation codes and job titles were sorted according to their acceptance rates. The acceptance rate is defined as the number of accepted applications per occupation/job titles over the total number of applications per occupation/job title. In Figure 7, the graph of every 20th serialized occupation vs acceptance rate and every job title vs acceptance rate are show to not have uniform distributions. And hence, occupation and job title can be good predictors of case status in the prediction task.

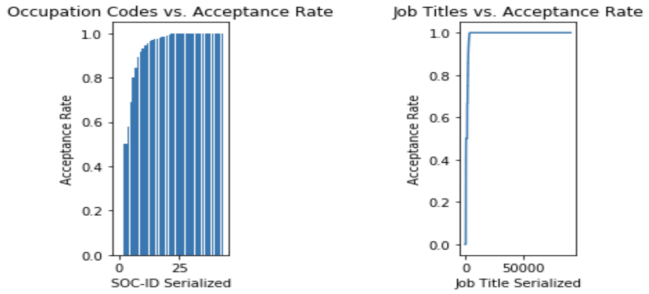

Fig. 7. Occupation code and job title sorted by acceptance rate in increasing order

Prevailing Wage: There are many outliers for the prevailing wage. In Figure 8, all prevailing wages vs. acceptance rates were graphed. It is seen that any prevailing past the red line at 0.5e8 has an acceptance rate of 0.0. Prevailing wages on the order of 1e8 were removed in order to see a how the points less than 1e8 were distributed. We can see from the second scatter plot in Figure 8 that the acceptance rates are not uniform and thus can be a good predictor for the case status prediction task. A Pearson correlation value could not be established for data in either graph below as there is no correlation due to outliers with values much higher than the majority of the data.

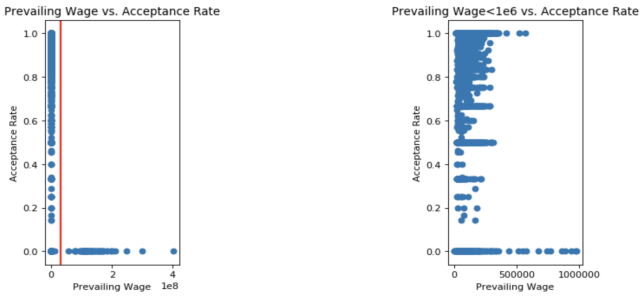

Fig. 8. Prevailing wage vs. acceptance rate for all values and when prevailing wage is less than one million

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 [4]. The varying prevailing differences were found by joining the data set with with wage tables that group wages based on occupation and metropolitan statistical area code. Prevailing wage differences were sorted according to their acceptance rates. The acceptance rate is defined as the number of accepted applications per prevailing wage difference over the total number of applications per prevailing wage difference. In Figure 9, prevailing wage difference vs. acceptance rate is not a uniform distribution. And hence, it can be a good predictor of case status in the prediction task.

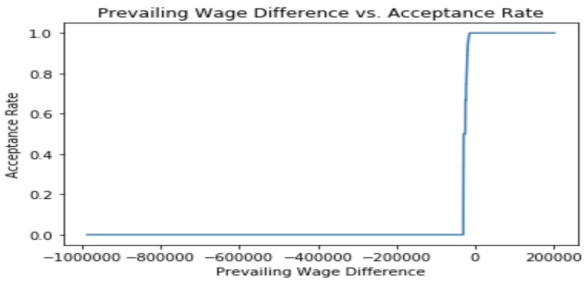

Fig. 9. Difference between prevailing wage in OFLC database and listed prevailing wage on application vs. acceptance rate

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. In Figure 10, negative wage differences appear to have a very low acceptance rate, close to 0.0% percent across all negative values. Two out of 2012 negative wage difference applications are accepted while the other 2010 applications applications were rejected. Applications with a negative wage difference have a 99.9 percent rejection rate. Wage difference seems to be a very strong indicator for application outcome.

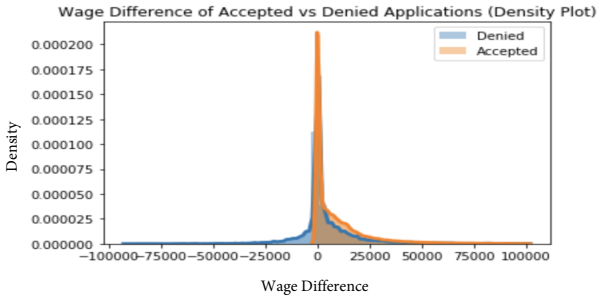

Fig. 10. Wage difference density plot with certified and denied applications denoted by color

A hypothesis test with random sampling cannot be performed to check if the mean wage difference for accepted applications was significantly different than denied applications. In the case of this H-1B LCA data set, since all data points for accepted and denied applications are in 2017, the data represents the a sample of the population that is not random as it is one year of the total and therefore bootstrapping methods as well as a t-tests cannot be used. If the test was done, it could show that the means are different. In addition, a difference in mean wages will confirm that wage difference is a differentiating factor between accepted and denied applications.

Wage offered vs Prevailing Wage: The wage offered vs. prevailing wage graph is shown in Figure 11 with accepted applications colored in blue and denied applications colored in red. The best fit line is depicted and appears to split the data into two groups with the accepted applications concentrated above the best fit line while the denied applications are concentrated below the best fit line.

Missing Value Count: A missing value in the H-1B LCA application means that the applicant chose not to fill out a field on the application. Choosing not to fill out a field may have unwanted consequences to the result of the application. We can see from

Fig. 11. Wage Offered vs. Prevailing Wage with case status labeled with color

Figure 12 that as missing value count increases, the acceptance rate decreases. The missing value count is strongly negatively correlated with acceptance rate. Missing value count might be a strong predictor for application outcome when the prediction task is looked at.

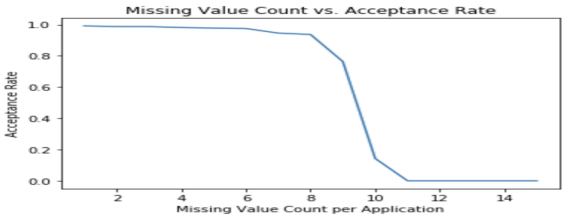

Fig. 12. Missing value count in an application vs. missing value count's acceptance rate

State: Utilizing a heat map, the the distribution of acceptance rate across states was visualized in Figure 13. Color intensity depicts 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 and this is confirmed by the geographical representation.

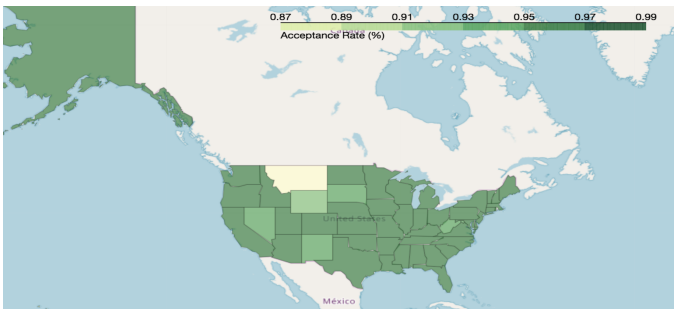

Fig. 13. Acceptance rate by State

In addition, the median wage across states was visualized. The heat map in Figure 14 shows the distribution of the median wage where states with darker shade attribute to more wages compared to others [5]. The states on the west coast appear to have a higher median wage than the center of the country and the east coast. Because acceptance rate and wage varies across states, state can be used as a predictive variable in the prediction task.

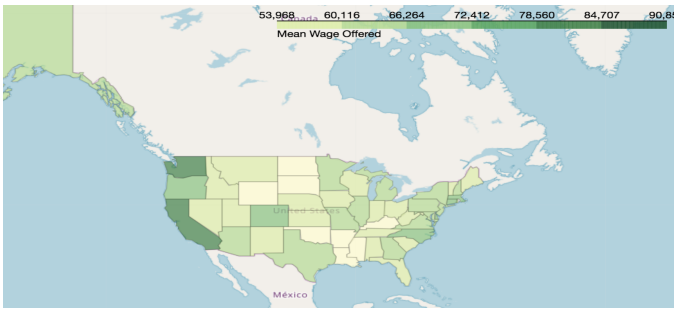

Fig. 14. Mean Wage Offered by State

V. Prediction Task

The dataset for this section was obtained from the exploratory data analysis section. In the previous section, feature engineering was done to create new fields such as the number of missing values in an H-1B LCA application, the wage difference (difference between wage offered and prevailing wage), as well as wage ratio and prevailing wage difference. Wage ratio refers to the ratio of prevailing wage to wage offered while prevailing wage difference is the difference between prevailing wage listed in the application and the prevailing wage for the position on the Foreign Labor Certification Data Center Online Wage Library.

The goal of the prediction task, the final step in the report, is to identify a classifier that best classifies applications as denied or certified. In addition, H-1B LCA application fields that highly impact the application outcome need to be identified so that employers can improve their application success rate in the future. A classification task will be performed with all of the relevant fields initially so that no important field is mistakenly left out. The prediction will determine the probability of an application being certified and whether it is denied or certified.

A. Classification Metrics

The following metrics were used to compare the performance of the baseline classifier and other classifiers against one another.

Classification Accuracy: Fraction of applications correctly classified

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

where
tp is the number of true positives,
tn is the number of true negatives,
fp is the number of false positives,
and fn is the number of false negatives.

False Positive Rate: Fraction of denied applications labeled as certified.

$$FPR = \frac{fp}{fp + tn}$$

False Negative Rate: Fraction of certified applications classified as denied.

$$FNR = \frac{fn}{fn + tp}$$

Balanced Error Rate: Average of the false negative rate and false positive rate.

$$BER = \frac{FPR + FNR}{2}$$

F1 Score: Another measure of accuracy that uses precision and recall.

- Precision is the ratio of true positives to applications that were predicted to be positive.

- Recall is the ratio of true positives to applications that were actually positive.

$$F1\ Score = \frac{2 * precision * recall}{precision + recall}$$

AUC: Area under the curve of the ROC-curve (True positive Rate vs. False Positive Rate). AUC represents the probability of a randomly drawn certified application having a higher probability of being certified compared to a randomly drawn denied application.

B. Feature Selection

All of the relevant features as well as the engineered features were used as inputs to the classifiers.

The missing values in the wage ratio and prevailing wage difference fields were replaced with the average value while the attorney name and prevailing wage source year fields were replaced with None and the year 2017.0 respectively. The year 2017 was used as it is the most recent and common year for prevailing wages.

Several columns were dropped because they contain information that is similar to other columns that were kept or the fields contained irrelevant information to the target variable. Dropping unnecessary columns prevents overfitting of the classifier as these columns add more information that a classifier has to conform to. Employer location information was deemed not as important as work site location for describing the workplace of the applicant, so these columns were dropped. `SOC_CODE` and `SOC_NAME` also known as occupation code are redundant fields as there is already a job title and industry field. Employment start date, end date, case number, and case submitted do not have any relationship with application outcome. Finally, work site county and work site zip-code was removed as there is already a work site state and city field. Finally, case status indicator was dropped because it is a duplicate of the case status column which is the target variable.

The categorical fields that had relatively low cardinality or number of unique values were encoded with one-hot k vectors. This encoding method was not used for all the fields in order to prevent adding too many dimensions to the dataset. For all categorical fields with less than ten unique values, the fields were one-hot encoded.

The rest of the categorical variables with more than 10 unique values were target encoded. This means each unique value was replaced with the corresponding adjusted acceptance rate of the unique value in a field. Replacing the unique values in a categorical field captured information about historical acceptance rate and prevented too many new dimensions from being added to the data.

C. Models

A benchmark classifier is used to compare all the other classification methods against.

-ZeroR (Baseline): The ZeroR classifier used as a baseline simply classifies all applications as the majority class which in this case is certified.

A.) Logistic Regression: A logistic regression classifier produces a probability of the target variable having a certain value based on a combination of the predictors. Logistic regression chooses coefficients that maximize the likelihood of observing the target variable. The benefit of logistic regression is that the probability of a receiving a certified application is outputted. However, a downside is that logistic regression assumes the target variable is linearly separable.

B.) Gaussian Naive Bayes: The Gaussian Naive Bayes classifier determines the conditional probability of the target variable according to Bayes Theorem. The probabilities of the predictors given the occurrence of the target variable are approximated using a Gaussian distribution. The advantage of Naive Bayes is that it is computationally fast and does not require hyperparamter tuning. The disadvantage is that the classifier assumes independence of all predictors even though this is rarely the case.

C.) Decision Tree: A decision tree classifier is the simplest of tree methods as it only constructs one tree to perform a classification. The decision tree finds the best split in a predictor's distribution that separate the classes of the target variable until a specified tree depth is reached. An advantage of using a decision tree is that they are simple to understand and interpret as well as visualize. The disadvantage of a decision tree is that they can easily overfit if the max tree depth is too large and only one tree is constructed.

D.) Random Forest: The random forest classifier is an ensemble method that uses many decision trees that have had bootstrap aggregating or bagging done on the samples. Bagging means that samples are chosen at random with replacement to reduce variance in any single tree. Random forests, in addition, take a subset of the predictors when constructing a tree so that each tree is more different from one another. The combination of the trees made from bagged samples and subsetted predictors is the random forest result. Compared to a decision tree, a random forest improves accuracy by determining class from the majority of votes for each sample in many trees.

E.) Gradient Boosted Classifier: A gradient boosted classifier involves combining a large num-

ber of trees. Each subsequent tree that is constructed is done so based on the least pure splits created by the previous tree. The trees are added back to the previous tree so that the tree can learn slowly and improve classification accuracy. An advantage to gradient boosted trees is that compared to a decision tree that may overfit to one set of the data, gradient boosted trees learn by focusing on the worst split in a tree and improve over time. The disadvantages of gradient boosted trees is that they may overfit compared to random forests.

VI. Evaluation and Results

The original dataset was split into a training and test set with 80% of the data going to the training set while the other 20% went to the test set. The same split ratio was used for all of the models but a random seed was used in each split so that the same samples were not picked for the training and test sets.

Before any of the machine learning models were run on the training data, stratified sampling was used so that the original case status ratio was maintained after the sampling. The number of samples in the minority class was 500 which means the total number of samples in the training data that was fed into all the models was 49798. Cross-validation was used with five folds so that the ML models would fit the training data in such a way that the fit was representative the whole dataset and not just the data from the train-test split.

Two methods of converting categorical fields to numerical fields were used before ML methods were applied because the methods require that all fields are in numerical format. The encoding was done while making sure the number of dimensions in the dataset did not increase too much as too many dimensions may lead to overfitting. For fields with less than 10 unique values, the fields were encoded with one hot k-vectors. While categorical fields with more than 10 unique values were target encoded. Target encoding means that each unique categorical value in a field was replaced by its average acceptance rate. A smoothing factor with a value of 1 was introduced to the average acceptance rate formula so that acceptance rates far from the global mean were closer to the global mean. The average acceptance rate formula requires there to be one sample in the average, if there is only one unique value, the global mean of acceptance rate is used. Target encoding does not add extra fields because the categorical values are replaced with numerical ones.

Logistic regression was used with an L1 penalty. Because all of the fields were inputted into the model, the L1 penalty was used in order to perform feature selection by adding the absolute value of the magnitude of the coefficient to the loss function. This penalty causes small coefficients to go to zero and thereby removes negligible features. A stochastic average gradient descent solver was used in order to decrease computation time as only a subset of the data would be used at a time to fit the model. The optimal regularization parameter was found to be 0.01 and the tolerance for the stopping criteria was changed to 0.01 from 0.0001 because the classification metrics were similar after the change and computation time decreased.

The logistic regression model identified the features that most influence rejection based on the fact that these features had the lowest coefficient values. The top ten features were `WAGE_RATIO`, `PREVAILING_WAGE`, `PW_SOURCE_None`, `PW_UNIT_OF_PAY_None`, `PW_SOURCE_DBA`, `VISA_CLASS_E-3 Australian`, `AGENT_REPRESENTING_EMPLOYER_N`, `LABOR_CON_AGREE_Y`, `PW_WAGE_LEVEL_None`, and `WILLFUL_VIOLATOR_None`.

The Gaussian Naive Bayes classifier did not have any hyperparameters that needed to be tuned. In addition to this more basic classifier, a Support Vector Classifier (SVC) was run on the training data but the fit was not converging in time so this model was not looked at.

A single decision tree was constructed so that it could be visualized and the splits in the important fields could be identified. The decision tree had a max tree depth of three after using GridSearch over the range of three to ten. The top features in the decision tree that determined whether an application was certified or not were `WAGE_RATIO`, `EMPLOYER_NAME`, `JOB_TITLE`, and `WORKSITE_CITY`. According to the decision tree visualization, a wage ratio below -0.011 standard deviations from the mean caused all 129 samples to be denied. Low wage ratio appears to be a sure predictor for application denial. The other three fields were converted to acceptance rates

by target encoding and then scaled so that the decision tree only shows that an employer name with an average acceptance rate less than or equal to -9.165 standard deviations from the mean acceptance rate is predicted to be denied. A similar reasoning can be used for the other two fields but this does not tell us which employer names are predicted to be denied.

The random forest ensemble method used 250 estimators and the function to measure quality of a split was gini impurity. More trees could have been used but according to *How Many Trees in a Random Forest?* [6], "the analysis of 29 datasets shows that from 128 trees there is no more signicant dierence between the forests using 256, 512,1024, 2048 and 4096 trees". The square root of the number of predictors being fitted were used in each tree. The minimum sample leaf was not tuned in order to decrease computation time. The top features according to the random forest that contribute to application outcome were `EMPLOYER_NAME`, `AGENT_ATTORNEY_NAME`, `JOB_TITLE`, `WORKSITE_CITY`, `NAICS_CODE`, `WAGE_RATIO`, `WAGE_DIFFERENCE`, `PW_SOURCE_OTHER`, `WAGE_RATE_OF_PAY_FROM`, and `WAGE_RATE_OF_PAY_TO`.

The gradient boosted classifier also used 250 estimators. The minimum sample leaf as well as learning rate were not tuned in order to decrease computation time. The top features in the gradient boosted classifier that contribute to application denial were `EMPLOYER_NAME`, `WAGE_DIFFERENCE`, `JOB_TITLE`, `WORKSITE_CITY`, `AGENT_ATTORNEY_NAME`, `PW_SOURCE_OTHER`, `NAICS_CODE`, `PW_SOURCE_OES`, `PREVAILING_WAGE`, and `NEW_EMPLOYMENT`.

Below are the classification metrics for all the ML models that were run.

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

The baseline classifier represents a benchmark that should be surpassed by a good classifier. All of the classifiers have a lower false positive rate and an AUC higher than the benchmark, which means all of the classifiers performed better. The macro averaged F1 score was used as a classification metric because as opposed to a micro averaged F1 score, the macro averaged metric is insensitive to the imbalance of the classes. The gradient boosted classifier had the highest F1 score and the lowest false positive rate out of all of the classifiers.

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

It is more important for the false positive rate to be low compared to the false negative rate in this problem. A false positive means that a denied application is predicted as certified. The employer will have to resubmit the application and have to deal with the cost of another application as well as the loss of time. A false negative means that the certified application is predicted as denied. This situation is not as bad because the employer can attempt to adjust the application according to a ML model's top features in the application before it is submitted so that another application does not have to be submitted.

The ROC (Reciever Operating Chararacteristic) curve was constructed for each model so that the area under the curve (AUC) could be computed. The AUC is a better measure than accuracy in this problem because of the large class imbalance. A baseline classifier that classifies all applications as accepted will receive a 98.596% accuracy, which sounds good, but would not correctly identify a single rejected application. AUC takes into account the false positive rate at each probability threshold and this metric, false positive rate, needs to be minimized by the classifier. The model with the highest AUC was the random forest with a value of 0.80291. This means there is a 80.291% chance a randomly chosen certified application will have a higher predicted probability for certification than a randomly chosen denied application.

VII. CONCLUSIONS AND FUTURE WORK

The best performing machine learning classifier was the random forest classifier because it had an AUC of 0.80291, the highest AUC. Even though

the gradient boosted classifier had the lowest false positive rate of 0.72041, 0.1481 lower than the random forest classifier, this false positive rate is only applicable at a probability threshold of 0.5. The most important features for the random forest classifier were `EMPLOYER_NAME`, `WAGE_DIFFERENCE`, `JOB_TITLE`, `WORKSITE_CITY`, `AGENT_ATTORNEY_NAME`, `PW_SOURCE_OTHER`, `NAICS_CODE`, `WAGE_RATIO`, `WAGE_RATE_OF_PAY_TO`, and `WAGE_RATE_OF_PAY_FROM`. The gradient boosted classifier's top features that were different from the random forest classifier's top features were `NEW_EMPLOYMENT` and `PW_SOURCE_OES`. These features are also important to consider because the AUC for the gradient boosted classifier was close to the random forest's with a value of 0.79205 and the F1-score was the highest with a value 0.70.

Based on the top features for both ensemble tree methods, it is apparent that employer names, job titles, worksite cities, agent/attorney names, prevailing wage sources, and industry codes with low acceptance rates tend to lead to application denial. When the difference between prevailing wage and wage offered and the ratio of prevailing wage to wage offered is low, the application tends to be denied. Applications where the maximum wage offered, the prevailing wage, and the number of first time applicants were higher had a higher acceptance rate. When a prevailing wage source is listed anything other than OES (Occupation Employment Statistics) on the application, the application had a lower acceptance rate.

With the top features that are changeable in mind, it is recommended that an employer who wants a certified H-1B LCA application pays more than the prevailing wage so that wage difference and wage ratio listed are higher. The maximum wage offered and the number of first time applicants should also be as high as possible as there is a positive correlation between these field and acceptance rate. In addition, the employer should hire an attorney with a history of a high H-1B LCA acceptance rates and use a prevailing wage from the OES.

A random forest classifier can be used by employers to check their H-1B LCA applications before they are submitted in order to save the cost of having to submit another application if the first submission fails.

In the future, in order to improve the AUC for the random forest classifier, more samples need to be used to train the classifier as only 500 samples of the minority class were used to decrease computation time. In addition, the number of certified applications in the training data can be downsampled so that the signal from a denied application is stronger. Finally, more hyperparameters can be searched over for the gradient boosted classifier such as the learning rate and minimum sample leaf but were not tried in order to decrease fitting time. Not all of the fields in the application were listed in the original dataset and can possibly be retrieved and included so that performance metrics are improved in a future iteration of this project.

REFERENCES

[1] https://www.foreignlaborcert.doleta.gov/performancedata.cfm
[2] https://www.foreignlaborcert.doleta.gov/pdf/PerformanceData/2017/H-1B_FY17_Record_Layout.pdf
[3] https://www.datacamp.com/community/tutorials/wordcloud-python
[4] http://www.flcdatacenter.com/
[5] https://github.com/python-visualization/folium
[6] https://www.researchgate.net/publication/230766603_How_Many_Trees_in_a_Random_Forest