Using machine learning to predict H-2A investigations/violations from DOL applications

Jack Gourdeau
Dartmouth College

David Kantor
Dartmouth College

Cameron Wright
Dartmouth College

Abstract

The H-2A program allows United States based agricultural employers to temporarily hire foreign guest workers to fill jobs for which there has been no demonstrated interested from domestic workers. Since the guest worker's visas are tied to their employment oftentimes worker rights violations go unreported and unpunished. The purpose of our study was to determine if, using confirmed violations as recorded by the Wage and Hour Division of the Department of Labor, we can identify informative characteristics of violating employers to better focus investigative efforts. Our key aims and questions for this project include:

- Given a new employer can we accurately predict if they will be found to have committed a workers rights violation?
- Which employer characteristics are most informative for predicting investigation and or violation?
- Can we identify key characteristics of employers who are investigated but are not found to have violations to help focus DOL investigations?

1 Introduction

Every year, the United States issues a large number of temporary work visas, allowing foreign nationals to enter the country to pursue work for a fixed and non-permanent period of time. There are many different types of visas prospective workers can attempt to secure, with specific visas corresponding to specific industries or lines of work. Non-Americans seeking to work in the agriculture industry apply specifically for H-2A visas, issued to allow agricultural employers to hire foreign employees if a shortage of domestic workers is anticipated for a given season (Costa et al. 2). Due to the high number of migrant workers that United States agricultural workers typically employ, there is no cap on the number of H-2A visas issued in a given year, unlike other temporary work visas issued by the United States government. In theory, H-2A visas create a mutually beneficial scenario in which agricultural companies that anticipate a shortage of domestic workers can fulfill their employment needs, and prospective foreign employees are given a legal path into the United States to better either their own or their family's futures. In actuality, however, many workers who arrive in the country on H-2A visas suffer through workers' rights violations and other such abuses of powers by employers and recruiters throughout the process. These potential abuses include misleading job information as well as being charged illegal fees during the visa application process (Sherill et al. 28).

For a potential migrant worker to obtain an H-2A visa, they must be "sponsored" by the specific company for which they will work upon their arrival in the United States. Companies can either recruit these employees directly, have returning workers recruit others to join them, or hire third-party agencies to recruit workers (Sherrill et al. 25). Because of the structure of H-2A visas, migrant workers who enter the country to perform agricultural work are permitted only to work for the one specific employer who sponsored their visa. If the company terminates the employment of a worker who has entered the United States on an H-2A visa, the visa is likewise terminated, forcing the employee to return to their home country or remain in the United States unlawfully and search for new, likely worse, potential jobs. Because of this possible retaliation from employees, foreign workers on H-2A visas are often hesitant to report violations of their workers' rights (Sherrill et al. 37).

Both companies and potential employees are screened by federal agencies to ensure they qualify for H-2A visas. The Department of Labor ensures that companies seeking to hire workers on H-2A visas are not hiring for positions for which domestic workers would be available, before the Department of Homeland Security further screens the company's application and approves a total number of workers the company can hire on H2-A visas. It is at this step where the DHS can deny the petition if it determines the employer has violated a worker's rights by collecting fees from the worker that were not reimbursed. Unfortunately, the DHS does not make the data regarding these denials public. This can leave workers investigating potential companies to work for vulnerable, as they are unable to form a fully-informed opinion on a company without knowing its history and any possible violations it may have committed in the past (Sherrill et al.).

The goal of our project was to build a model that could effectively predict whether a given H-2A application would eventually lead to a Department of Labor investigation or violation. To do so, we ran two iteraations of our machine learning model, one aimed at predicting investigations and one aimed at predicting violations. If identifying characteristics for employers in these two separate groups could be pinpointed, the Department of Labor would, in theory, be able to streamline their investigative efforts. Given that each investigator for the Wage and Hour Division of the DOL (responsible for enforcing the labor laws designed to protect H-2A workers) is responsible for 175,000 workers (Costa et al. 3), being able to identify characteristics that are associated with H-2A violations would allow them to to more efficiently allocate their time and resources to identify employers who take advantage of their H-2A workers' tenuous legal standing would.

2 Related work

While the prevalence of abuses of the rights of workers on H-2A visas is extremely troubling, both government and independent groups have conducted investigations into these violations with the goal of analyzing and improving upon the systems that allow these violations to happen and go unreported at such a high frequency. A 2020 report from the Economic Policy Institute, a nonprofit think tank, found that a large share of the total number of violations come from a relatively small handful of employers (Costa et al. 57), and that farm labor contractors tended to be the worst violators, particularly in major farming states such as California and Florida (Costa et al. 56) - a troubling trend given that these contractors represent the fastest growing segment of agricultural employment (Costa et al. 1). The report also mentions that the WHD already practices a form of strategic enforcement, directing at least half of their limited resources towards proactive investigations of employers that are likely to violate the laws protecting workers' rights (Costa et al. 3). Our project builds upon this report by attempting to identify shared characteristics of violating employers, hopefully allowing us to identify traits that make an employer more likely to violate its workers' rights. If successful, the principles underlying our project could be used by the WHD to bolster their proactive investigations into likely violators. An additional report, given to United States Congressional Committees by the Government Accountability Office in March of 2015, detailed the results of their own investigation into violations of workers' rights laws committed by H-2A employers and recruiters, with a specific focus on third party recruitment of foreign laborers. The March 2015 report revealed that the Department of Homeland Security, despite collecting petition information from employers, does not capture detailed information on job available for H-2A workers or make the data they do collect publicly available. Because of these shortcomings, potential employees and people advocating for them cannot verify the job offers presented to them by recruiters - allowing recruiters to misrepresent positions, charge hidden fees, or otherwise violate the rights of foreign workers seeking H-2A visas (Sherill et al. 29). The Department of Labor similarly fails to consistently share its data with the public as well as with other departments within the U.S. government (Sherill et al. 41). Building upon this report, our project sought to easily match information on employers from one dataset to another, which could possibly aid in aggregating all relevant information from various United States government agencies to share with the public.

3 Data

For our project, we worked with the WHD Violations Dataset as well as the DOL H-2A applications data from FY 2018. We were able to access both of these data sets on the DOL's website and we have included the links to the data sets above. Please note, we chose to use H-2A applications from

2018 because we wanted to avoid any abnormalities caused by the COVID-19 pandemic. Of course, we understand that this limited time frame will hinder the predictive ability of our model; however, we figured it was a reasonable starting point to estimate our model with one year of data. As a next step, it would probably be helpful to expand this time window. This may help better-capture features that lead to investigations/violations.

3.1 WHD Violations Data

While the original WHD Violations data set includes data going as far back as the early 1900s, we were only interested in the data for investigations from the fiscal year of 2018 and onward. Accordingly, we subset this data to all of the cases for which the load date (the date of entry for the case) was past January 1st, 2017. We chose January 1st, 2017 because there were some applications listed in the H-2A data set that had job start dates in 2017. As we didn't want to miss any potential violations, we therefore subset past January 1st, 2017. The original unit of analysis in the data set was at the case level. As we will explain later on, we used this version of the data through matching; however, once we found our fuzzy matches we aggregated the results to the employer level. The data set includes every investigation opened by the WHD, regardless of whether or not the investigation discovered violations of H-2A workers' rights or not. In the data set there is a column called H-2A violation count which we used to subset our data for different iterations of our model. For the model testing whether the application led to a violation we subset this to H-2A violations greater than or equal to zero. for the model testing whether the application led to an investigation we removed this constraint (thus, the H-2A application could be any value greater than or equal to zero).

3.2 H-2A Application Data

The H-2A data set has applications where the start date ranges from 2014 to 2018. Since we were trying to match applications with potential violations we used this whole subset of time; however, as we noted earlier, we subset the earlier data set past January 1st, 2017. Therefore, for matching, we effectively looked at the H-2A applications past January 1st, 2017. When we later ran our ML model we used the entire date subset. An important subset to note here is that we only looked at applications where the application status was CERTIFICATION or PARTIAL CERTIFICATION. In theory, this subset should be our potential pool for investigations/violations since these were the only applications approved; however, if any applications were given an incorrect status, then there's a chance we could be missing those applications.

Ultimately, a big piece of data missing from both of these data sets that would have been extremely helpful is an application id code of some sort that is later matched for in the investigations/violations data set. Currently, we use fuzzy matching for name and date filtering to best estimate which violations are associated with which application; however, if there was an exact application number that we could retroactively match on, then we would be able to use exact matching on this variable to determine which cases led to investigations. Additionally, since we run our ML model on the data provided in the H-2A data, it would have been helpful to have more information on the employer that could be used as feature vectors. For example, perhaps it would have been helpful to have a column in the H-2A data set which identifies whether or not that employer had any previous violations. I imagine that would have been a very useful determinant in our model.

4 Methods

4.1 Data cleaning

In order to prepare both of our data sets for fuzzy matching, we had to clean our data sets.

First, we made sure that the relevant date columns in both of the data sets were date-time objects rather than just strings. This is particularly important for later sub-setting. In the WHD data set, the date columns of interest were the findings' start date, the findings' end date, and the date of entry for the report (denoted load date). In the H-2A data, the date columns of interest were the job's start date and the job's end date.

Additionally, we had to clean the names of the employers in each of the data sets. In the WHD data set, there were columns for both legal name and trade name. We chose to clean the legal name column by converting the names to upper-case and removing any punctuation after "LLC," "CO," and "INC." We applied a similar process for cleaning the names in the H-2A applications data set. Specifically, we applied our name-cleaning function to the employer name column in the H-2A data set.

Another way we cleaned our data prior to fuzzy matching was that we filtered out H-2A applications that had not received partial or full certification. We did this by pulling out the certification status from the case status column in the H-2A data set and then sub-setting to the applications where the case status was "PARTIAL CERTIFICATION" or "CERTIFICATION." We believed this would give us the most accurate representation of these applications because weŕe effectively ignoring applications that were not accepted, and, accordingly, should not result in investigations and/or violations.

After we did all of our major cleaning for these data sets, we did some final cleaning to aid our fuzzy matching. This included converting the city names in each data set to upper-case, removing rows in the WHD data set where the name was null (NAN), and sub-setting the WHD data to those entered prior to January 1st, 2017. We chose January 1st, 2017 specifically because although the H-2A applications were from fiscal year 2018, there were some applications which had job start dates in 2017. Therefore, we wanted to ensure that we didn't miss any potential matches.

4.2 Fuzzy matching

Once we had all of our data prepared for fuzzy matching, we used fuzzy matching on employer name to identify potential matches between the H-2A applications data set and the WHD violations/investigations data set. We decided to use fuzzy matching rather than exact matching because we noticed that the names used for the employers was not always consistent. For example, sometimes a given name may include LLC, whereas another time it wouldn't. Also, as we learned during our discussions with the TRLA, one potential difficulty in identifying repeat violators is that these companies may change their legal name. Fuzzy matching would help address both of these issues.

For our fuzzy matching model, we blocked on state code and matched on both employer name and city name. Blocking on state code and matching on city name both help to avoid false positives regarding name matches because it's relatively difficult to change the location of a business (note: it's certainly not impossible for an employer to change the city and/or state that their business is registered in; however, it seems relatively unlikely). To find the matches, we used the Jaro-Winkler distance function with a threshold of 0.85.

4.3 Data preparation for ML model

After we obtained our matches from our fuzzy matching, we had to further clean/prepare our data so that it could be used for our machine learning model(s).

First, we removed entries from our matches where the load date was earlier than the job start date listed on the application. We did this because we didn't want to consider matches that were not associated with old applications. For example, if a match has a job start date of 06/01/2018 and the load date of the investigation of 01/01/2018, then this investigation must have been from an earlier job posting. We chose to ignore these matches because they would confuse our machine learning models which are looking for key features in the H-2A applications that may lead to an investigation or violation.

Next, we generated an indicator column in our original H-2A application data set that would indicate if that application ended up leading to an investigation (or violation). Since we stored the H-2A application employer names in our matched data set, we were able to form the indicator successfully with the fuzzy matched names.

After we generated this investigated/violator indicator, we realized we had many matches for one given violation. This happened when there were numerous applications for a given employer and one (or possibly more) violations. In these scenarios, when we matched on employer name, it matched these sparse with violations with each of the applications. To account for this, we effectively built a "representative" matrix that would only contain one entry per employer. In other words, we effectively aggregated this data set from application-level to employer level. In doing this aggregation, we replaced the values of each column with a representative sample of that employer's values. More specifically, we took the mean value for all integer and float columns and the mode for any string columns. For date-time columns, we generally took the minimum date; however, if the date-time column contained "END" we took the maximum of this column. In doing this, we effectively got a range from the first start date to the last end date.

Once we had our representative matrix formed, we were able to move forward with more machine learning-specific data preparation. More specifically, this including converting all date objects in our representative data frame to date-time objects (rather than just the relevant columns we converted earlier); these included case received date, decision date, requested start date of need, requested end date of need, job start date, and job end date. Next, we dropped the second diploma major column because it only had null values so it would be useless in our model. Also, we pulled out our outcome variable (is_violator in both models) because the model would predict a 1:1 mapping if we left this variable in our feature matrix.

After this, we separated our feature matrix into numeric data and categorical data so that we can impute any missing values. We did these separately because we used different imputation methods for each of these types. For numeric imputation, we filled missing values with the most frequent value found (i.e., the mode) and for categorical imputation we just filled the missing values with the string "missing_value." This allows us to see if the missing data had a material impact on our results. After we imputed all of our data, we combined all of the columns back together and dropped a few columns that were simply unique identifiers that would be falsely predictive in our model. These columns included unique id, case number, employer name, and trade name.

4.4 Machine learning model specifications

With our data cleaned and prepared, we were able to run our machine learning model. For our model we used a train/test split of 80/20. Once we split our model into the training and testing sets, we applied one-hot encoding to our feature matrices (X_train and X_test) so that our model could best capture the impact of certain categorical variables. Further, we used a Random Forest Classifier to fit our model. We fit our model with the training data (X_train and y_train). Finally, once our model was fit, we used it to predict the output values for the test data. We compared this output with the true output (y_test) to assess the performance of our model.

4.5 Measurements of model performance

To assess our model's performance, we generated a few key metrics/outputs. First, we generated a confusion matrix which compares the number of violations or investigations correctly predicted, incorrectly predicted, correctly missed, and incorrectly missed. Additionally, we were able to generate the accuracy score, the precision score, the recall score, and the f1 score all from the predefined Scikit-Learn package in Python (which we also used to run our ML models). Please refer to the results section for the mathematical definitions of these performance metrics.

5 Results

5.1 Characterization and Analysis of input data

As a result of our fuzzy matching, using a jarowinkler threshold of 0.85, we were presented with $11,411$ potential H2A applications to WHD investigations based on company name after blocking on state. Upon manual inspection, it seems that our threshold allowed for the capture of some non-identical names which were highly likely to be from the same employer (table 1). However, it was identified that the fuzzy matching also resulted in the false identification of matches for names which, although sharing similar structure, were not likely to be true matches. Examples of such true and false positives can be seen in table 1. In combination with our downstream representative application formation which takes into account all applications matched to a name, these false positives potentially skewed the results of the aggregation process such as by shifting the mean number of employees away from the true mean of applications submitted by the employer. However, given that manual inspection seemed to show the majority of the fuzzy matches were likely matches and, in an attempt to reduce false negatives given the tendencies of employers to change their names following being cited for violations, we proceeded with a threshold of 0.85 to representative application creation as outlined in the above methods. Sample columns from an example result from one employer's applications in table 2 can be found in table 3.

After fuzzy matching, subsequent data cleaning and representative application formation, we were left with two primary data sets one of which (repMatrixforpredict_investigations.csv) contained a binary classifier with 1 denoting companies which had been investigated by the wage and hour division and 0 denoting companies which had not been investigated. The other, (repMatrixforpredict_violations.csv) contained the a similar classifier, however, in this case with 1 denoting companies which had been found to have violations by the wage and hour division and 0 denoting companies which did not have violations recorded. Companies without violations may have been investigated and not found guilty or had not been investigated at all.

Of the original $12,027$ applications in the 2018 fiscal year, from which the representative application matrix was formed, $9,931$ of the applications came from non-investigated companies while $2,096$ were submitted by companies who were ultimately investigated. These $2,096$ applications came from 858 unique companies and, therefore, within the investigations focused representative applications data set, of the $7,643$ unique companies, 858 were classified as investigated and $6,785$ were classified as non-investigated. Therefore, this suggests around 11% of the companies who applied in 2018 fiscal year and had one or more of their applications certified or partially certified were investigated at a later date.

Likewise, of the original $12,027$ applications in the 2018 fiscal year, from which the representative application matrix was formed, $10,131$ came from non-violating companies while $1,896$ were submitted by companies who were ultimately found to have committed violations of some type. The aforementioned $1,896$ applications came from 736 unique violating employers and therefore, within the violations focused representative applications data set, of the $7,643$ unique companies, 736 of the representative applications were classified as violators and $6,907$ were classified as non-violators. Again, this suggests that around 9.6% of the unique companies represented in the applications were found to be violators. Furthermore, this suggests that approximately 86% of the 858 employers who were investigated are ultimately found to have been in violation of the workers rights.

Each of these representative application data-sets retained all of the original data fields collected by the DOL within the 2018 Fiscal Year H2A Applications data set. The full documentation may be found within our GitHub repository, however, this included fields such as work site state, the name of the filing agent attorney, the filing law firm's company name and the primary crop produced among others (Table 4).

To gain a better understanding of the information contained in the variables which would be ultimately informing our machine learning models, we grouped representative applications by a selection of variables and examined the percentage of employers within that variable's levels which corresponded to a violation/investigation. When we examined the relationship between primary crop listed on the companies representative application and the percentage of investigations we fond that labels of 'oranges', 'citrus' and 'chickens' had the highest percentage of companies investigated (figure 1). Across the $7,643$ representative applications there were 146 unique primary crops with 'Agricultural Equipment Operators' (n=840), 'Tobacco' (n=738) and 'Flue-cured Tobacco' (n=509) representing the three most popular levels within the field. The presence of 'Tobacco', and 'Flue-cured Tobacco' suggests that, in the future, aggregation of similar crops may increase and improve signal. This reveals both a limitation in our methods and a potential future improvement. The mean percentage of investigations across all "crop" groups was 22.8% ($sd = 17.13\%$) When the relationship between primary crop and the percentage of violators was examined, unsurprisingly, the same three primary crops were had the highest percentage of violators (figure 2). The mean percentage of violations across all crop groups was 20.47% ($sd = 16.57\%$). The presence of the two separate but similar crop groups of oranges and citrus provides further evidence that the aggregation of similar and equivalent crops may increase signal and downstream predicting power. The same approach was taken to examine the relationship between the filing law firm and the percentage of companies investigated/found to be violators. There were 336 unique Law firms listed within the applications with "Agriculture Workforce Management Association" (n=620) and "SNAKE RIVER FARMERS ASSOCIATION INC" (n=543) being the two most popular firms across the companies. We found that across all listed legal firms the percentage of investigated companies and percentage of violators listing that legal firm was 25.13% ($sd = 17.81\%$) and 23.78% ($sd = 18.39\%$) respectively. There were, however, legal firms which showed higher percentages of violation and investigation as shown in figures 1 and 2, however, additional statistical analysis is needed to determine the significance of these differences. The lack of any robust statistical exploration is significant limitation to our study and suggests a future informative direction of study.

We also examined how the percentage of companies who were investigated or deemed violators differed across the work site state listed on their application and created a heat map to visually inspect both (figures 3, 4). Across all states the mean percentage of employers investigated was 18.41% ($sd = 13.69\%$) while the mean percentage of employers found in violation was 15.90% ($sd = 12.50\%$). Rhode Island led both categories with 66% of the companies applying eventually being investigated and the same percentage ultimately being found guilty of violations. However, upon closer inspection, both of these figures are most likely due to a small sample size (n=3) for Rhode Island work sites. The next highest states in terms of percent investigations, FL (n=91), PA (n=101) had 48.35% and 39.60% of employers investigated respectively. Likewise, FL (n=91) and WV (n=11) were the next highest states in terms of percent violations with 45.05% and 36.60% of employers investigated respectively. PA followed FL and WV with 34.65% of employers found to have committed a violation. The results of these analysis can be found in figures 3 and 4. Again, due to time limitations we were unable to perform informative statistical analysis, however, a potentially informative future direction would be to compare TRLA catchment states against non-TRLA catchment states and perform a simple two sample t-test or equivalent method to determine any statistically significant differences. Furthermore, a simple uni-variate analysis of investigation rates across states to determine the modality of distribution reveals a potentially bi-modal grouping of investigation rate and further investigation is required to determine the legitimacy (could simply be an artifact of bin size) and origin of this pattern (figure S1).

Table 1: **Fuzzy matching produced both True Positive matches and False positive matches.** Examples of true and false positives (TP and FP respectively) resulting from fuzzy matching. Shown are the original, uncleaned, names contained in each data set, with LLC and INC still retained. False positives are likely the result of exact matches such as "VINEYARDD MANAGEMENT LLC" which allows for poor first word matches to be accepted.

Result	**H2A Application Name**	**WHD Investigation Name**
TP	MCF4 SOLUTIONS, LLC	MCF4 SOLUTIONS, LLC
TP	BROTHERS BEST PRODUCE	BROTHERS BEST LABOR, INC
FP	BAZAN VINEYARD MANAGEMENT LLC	CORONA VINEYARD MANAGEMENT LLC
FP	ANOROC VINEYARD MANAGEMENT LLC	CORTINA VINEYARD MANAGEMENT LLC

Table 2: **Example of multiple H2A applications matching back to one company.** Example of multiple applications from one employer with a sample of columns. The resulting representative application can be found in table 3

CASE_STATUS	EMPLOYER_NAME	...CROP	NBR_WORKERS...	JOB_START_DATE	JOB_END_DATE
DETERMINATION ISSUED - CERTIFICATION	MARTIN AUZA SHEEP COMPANY	Sheep	1.0	2017-10-19	2018-07-08
DETERMINATION ISSUED - CERTIFICATION	MARTIN AUZA SHEEP COMPANY	Sheep	1.0	2017-11-20	2018-08-08
DETERMINATION ISSUED - CERTIFICATION	MARTIN AUZA SHEEP COMPANY	Sheep	2.0	2018-02-20	2018-09-17
DETERMINATION ISSUED - CERTIFICATION	MARTIN AUZA SHEEP COMPANY	Sheep	1.0	NaT	2019-02-05
DETERMINATION ISSUED - WITHDRAWN	MARTIN AUZA SHEEP COMPANY	Sheep	0.0	2018-05-24	2019-02-05

Table 3: **Example of representative application created from the multiple applications found in table 2.** Each column was summarized using a method suitable for the datatype.

CASE_STATUS	EMPLOYER_NAME	...CROP	NBR_WORKERS...	JOB_START_DATE	JOB_END_DATE
DETERMINATION ISSUED - CERTIFICATION	MARTIN AUZA SHEEP COMPANY	Sheep	1.0	2017-10-19	2019-02-05

Table 4: **Description of the two resulting data frames with binary classifiers corresponding to the prediction of violations and the prediction of investigation.** The unique number of analyzed levels for 3 of all the considered application input fields passed to the ML model are given alongside the percentage of the positive class.

Binary Classifier (Pos/Neg)	Total Rep. Applications	Unique Companies	Unique Primary Crops	Unique law firms	Percent Positive Class.
Violation/No-Violation	7643	7643	146	336	0.106558564
Investigation/No-investigation	7643	7643	146	336	0.126455416

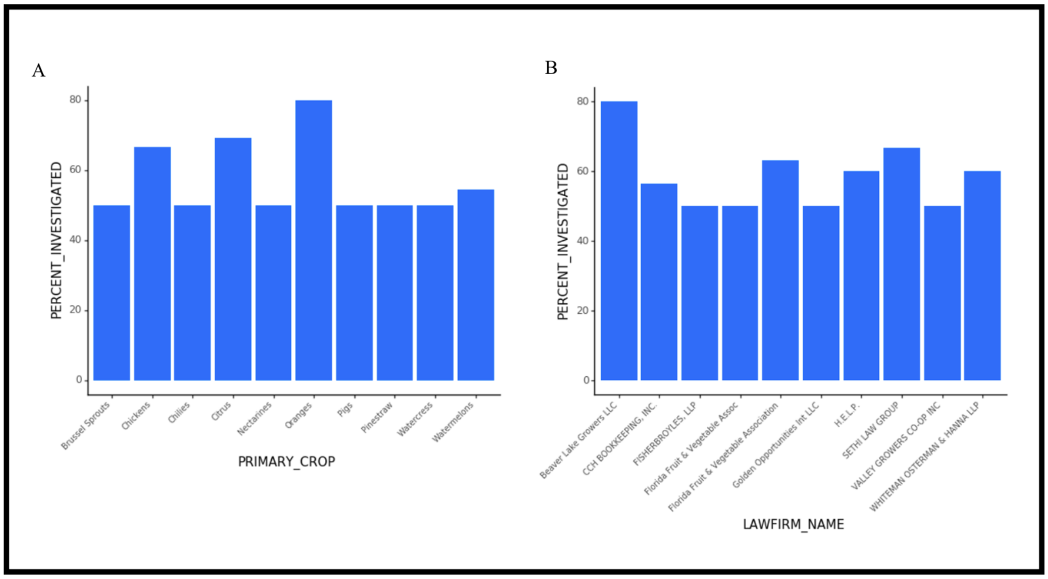

Figure 1: **The top category levels of Primary crop and law firm name by investigation rates demonstrate some levels have higher rates of investigations than others.** A) Top ten of 146 unique Primary Crop labels by percentage of employers listing that crop on their application who were investigated. B) Top ten of 336 unique law firm name labels by percentage of employers listing that law firm who were investigated.

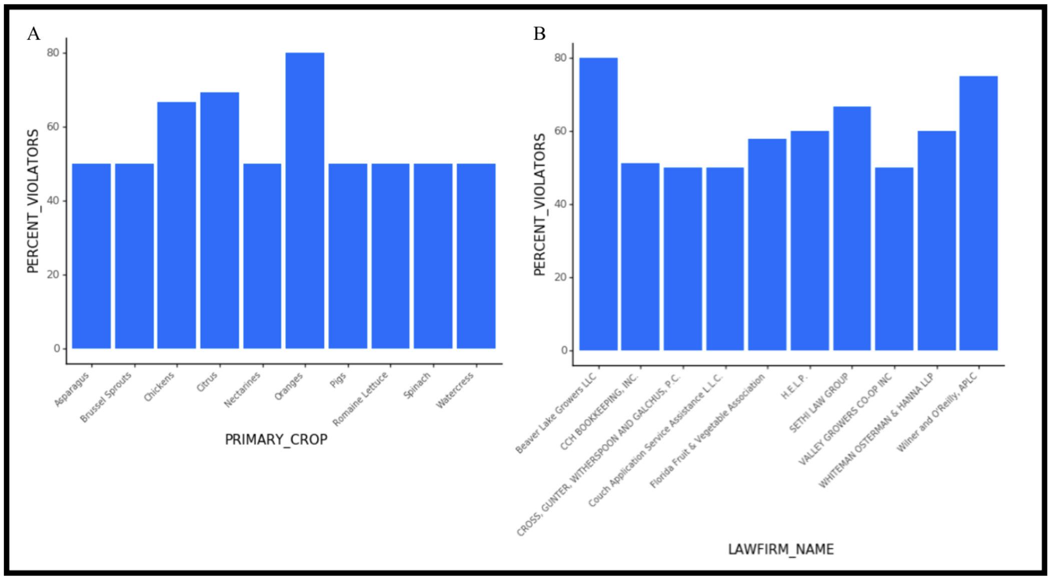

Figure 2: **The top category levels of Primary crop and law firm name by violation rates demonstrate some levels have higher rates of violations than others.** A) Top ten of 146 Primary Crop labels by percentage of employers listing that crop on their application who were found to be violating workers rights. B) Top ten of 336 unique law firm name labels by percentage of employers listing that law firm who were found to be violating workers rights.

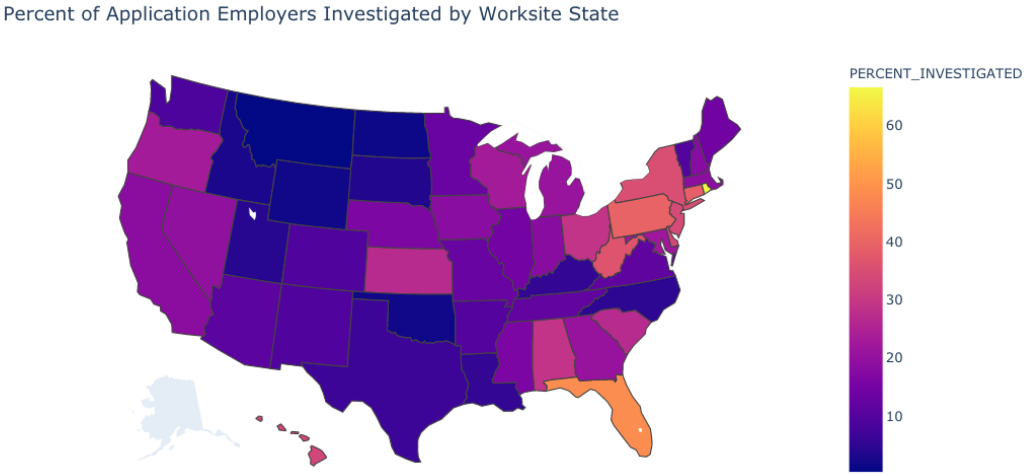
Percent of Application Employers Investigated by Worksite State
PERCENT_INVESTIGATED
60
50
40
30
20
10

Figure 3: **There is some variation in investigation rates across states.**.Heat map of the percent of all employers investigated by work site state with lighter colors denoting a higher percentage of employers. The mean percent of employers investigated across all states was 18.41%.

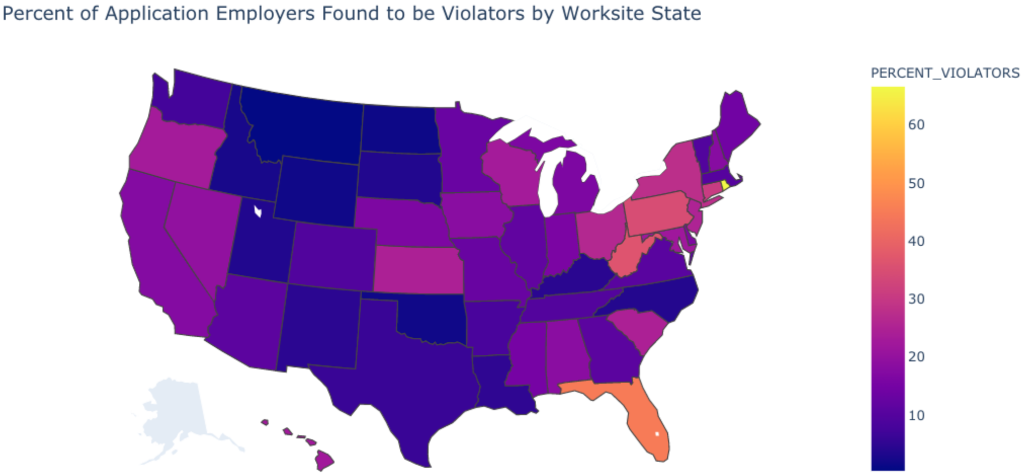
Percent of Application Employers Found to be Violators by Worksite State
PERCENT_VIOLATORS
60
50
40
30
20
10

Figure 4: **There is some variation in violation rates across states** Heat map of the percent of all employers found to be violators by work site state with lighter colors denoting a higher percentage of employers. The mean percent of employers found to be violators across all states was 15.90%.

5.2 Random Forest Machine Learning Results

The two representative applications data frames, one with a binary classifier for violations the other with a binary classifier for investigations were imputed and passed into the random forest model. The train test split on the original $7,643$ representative applications in both data sets resulted in a training set of $6,114$ representative applications and a testing set of $1,529$ representative applications.

We found that, when predicting violations, the model correctly predicted $1,389$ of the $1,520$ representative applications classified as non-violators therefore missing 131 applications which should have been classified as non-violators. The model correctly predicted 4 of 9 representative applications labeled as violators (figure 5A). Likewise, when predicting investigations from representative applications the model correctly predicted $1,364$ of the $1,514$ representative applications classified as non-investigated therefore missing 150 applications which should have been classified as non-violators. The model correctly predicted 6 of 15 representative applications labeled as violators (figure 5B).

A) Predicting Violations:

Predicted Actual	Non-Violator	Violator
Non-Violator	TN: 1389	FP: 4
Violator	FN: 131	TP: 5

B) Predicting Investigations:

Predicted Actual	Non-Investigated	Investigated
Non-investigated	TN: 1364	FP: 6
Investigated	FN: 150	TP: 9

Figure 5: **Confusion matrices for both random forest classifier models show difficulty in predicting violations and investigations given current data set.** True positive, True Negatives, False Positive and False Negative are denoted as TP, TN, FP, and FN respectively.

With True Positives, True Negatives, False Positive and False Negatives denoted as TP, TN, FP, FN, accuracy, precision, recall and F1 can be calculated as follows:

$$\mathit{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)}, \tag{1}$$

$$\mathit{Precision} = \frac{TP}{(TP + FP)}, \tag{2}$$

$$\mathit{Recall} = \frac{TP}{(TP + FN)}, \tag{3}$$

and

$$F1 = \frac{2(\mathit{Recall} \cdot \mathit{Precision})}{(\mathit{Recall} + \mathit{Precision})} \tag{4}$$

respectively. The results of these calculations for both models using the values in figure 5 are presented in table 5.

Table 5: **Model performance metrics for both random forest classifier models predicting the binary label of either violation/no violation or investigation/no investigation.** Metrics calculated using the values found in figure 5

Predicting	Accuracy	Precision	Recall	F1
Violations	0.912	0.556	0.037	0.069
Investigations	0.898	0.600	0.057	0.103

6 Discussion and Conclusion

Our data do not support the use of a Random Forest Classifier machine learning model to predict either violations or investigations based on H2A Applications data. For both models the precision, denoting the correctly predicted positive observations to the total predicted positive, and recall, denoting the correctly labeled positive observations to all the true positive observations, reveals poor precision and sensitivity in predicting either violations or investigations (table 5). The accuracy figure in both models is misleading given the rare occurrence of violations and investigations in the data sets. Machine learning models predicting a binomial classifier are known to work best on data sets containing an equal number of each class. It is therefore not surprising that we see such poor performance and our results suggest that further work could focus on producing an evenly split data set either through sampling from the rare class (violations/investigations) with replacement during formation of the training data or, alternatively, through the expansion of the data set beyond the fiscal year 2018.

Furthermore, by limiting our data to applications in FY 2018 there is inherently a great deal of variability in the statistics drawn from our data. This said, the methods employed in this study, by comparing investigations/violations to applications and developing a representative for those applications has the potential to allow for useful comparisons between different application groups (state, crop etc.) as we have shown on a small scale. As previously motioned future work may improve upon our fuzzy matching methods to improve performance and future studies would benefit greatly from powerful statistical analysis such as linear contrasts which could be applied with relative ease given our data structure. This would allow for the comparison of groups to better understand potential signals for violators and investigations which are not reflected in the machine learning outcomes. The machine learning methods themselves may also be refined through the use of alternative classifiers and the improvement of the encoding steps. In our current model a OneHotEncoder is applied to all of the input columns which may have resulted in informative numerical relationships being encoded as categorical data and the potential lost predictive power. Alternative approaches may include applying compatible encoders column wise and re-merging to create a feature matrix which preserves valuable relationships which were potentially lost. We believe that the expansion of our data set to additional years and the inclusion of outside categorical information as feature vectors (such as the text Analysis results from our classmates' project) may help to improve future machine learning performance.

Overall, while it is clear that the current methods in predicting violations and investigations from H2A applications requires refinement, we believe that this comparison approach will be useful and could potentially inform resource management for the Department of Labor and improve worker's protections. We suggest future work focuses on the application of robust statistical methods to understand potential relationships in the data resulting from our representative application methods and the refinement of machine learning models.

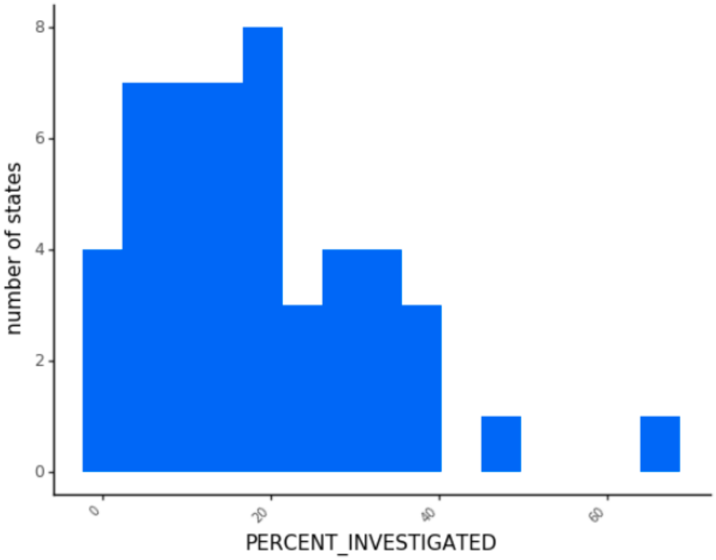

Figure S1: **Histogram demonstrating a potential bi-modal distribution of company investigation rate across states which may be informative for future analysis directions.** Future analysis is needed to determine the legitimacy of this distribution and its possible origins