CS791 Final Report

Project Title: Identifying Labor Trafficking

Team Members: Andrew Polevoda and Ziang Leng

Date: 4/30/2020

Motivation

The motivation of the project is to combat Fair Labor violations committed by employers. The Massachusetts Attorney General's Office receives a large number of complaints of such violations, but since resources are limited, not all of them can be addressed. Also, it is suspected that certain issues are underreported. The Massachusetts Attorney General's Office publishes data of Fair Labor Division complaints and enforcement actions. The Attorney General's Office wants to try and use this data to predict whether an employer is likely to be committing a fair labor violation. In a pragmatic sense, the aim of the project is to use state-of-the-art machine learning techniques to inform more efficient utilization of resources, which will ultimately help to combat these violations.

Data Exploration and Visualization

Our first task, as identified by our clients, was to work with and identify the overlap between the Fair Labor Division Complaints and Fair Labor Division Enforcement data published by the Massachusetts Attorney General's Office. Some visualizations of the data are given on the following pages.

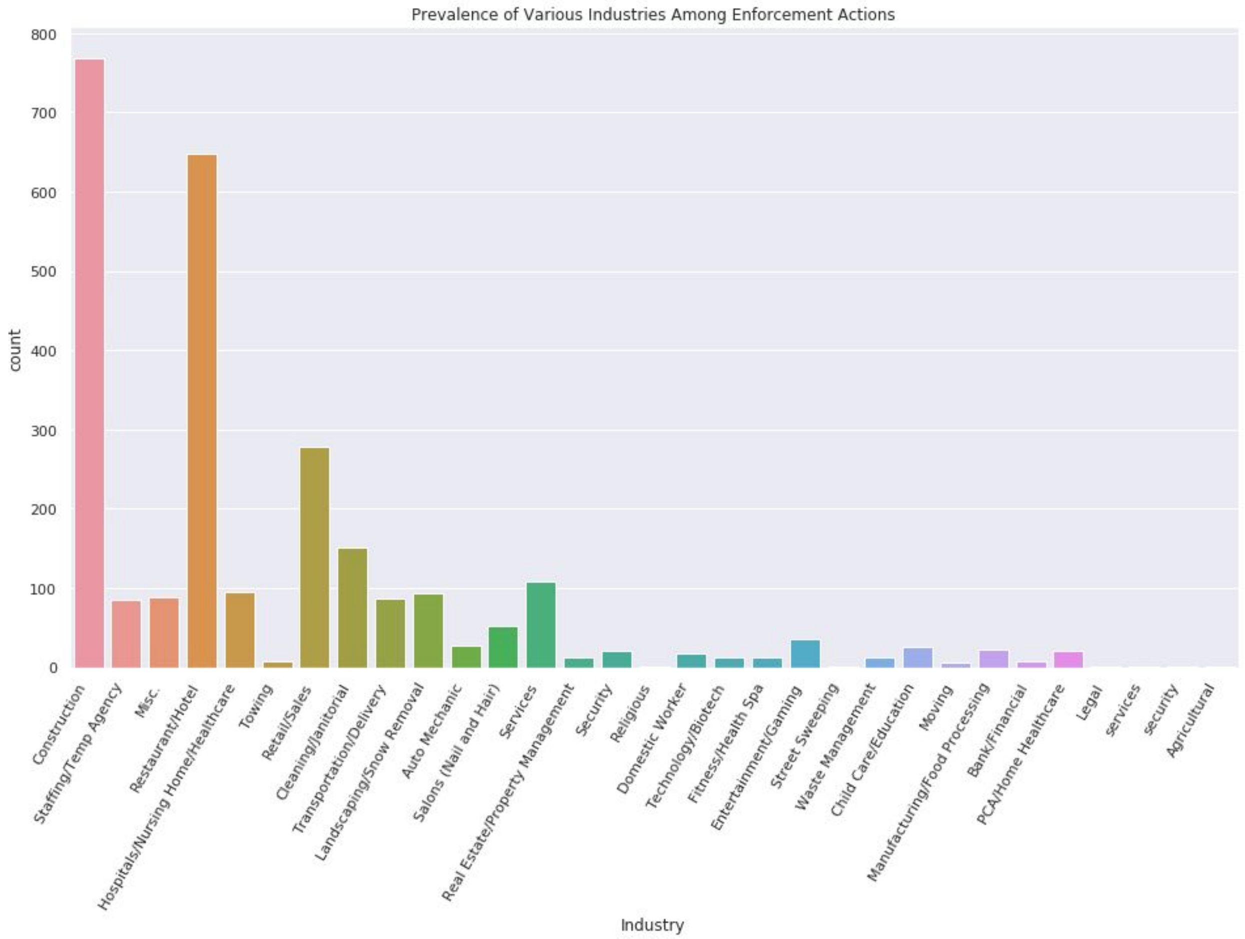

Figure 1. (code available at https://github.com/apolevoda/CS791-ML-PublicHealth-2020/blob/master/cs791-data-visualization.ipynb)

We found that most Fair Labor Division enforcement actions are taken against employers in the industries of construction and restaurants/hotels, as shown in Figure 1. The industry with the next highest amount of enforcement actions has less than half as many as either of the aforementioned two. This lines up with the observations of our clients, as well as what they told us during our kick-off meeting.

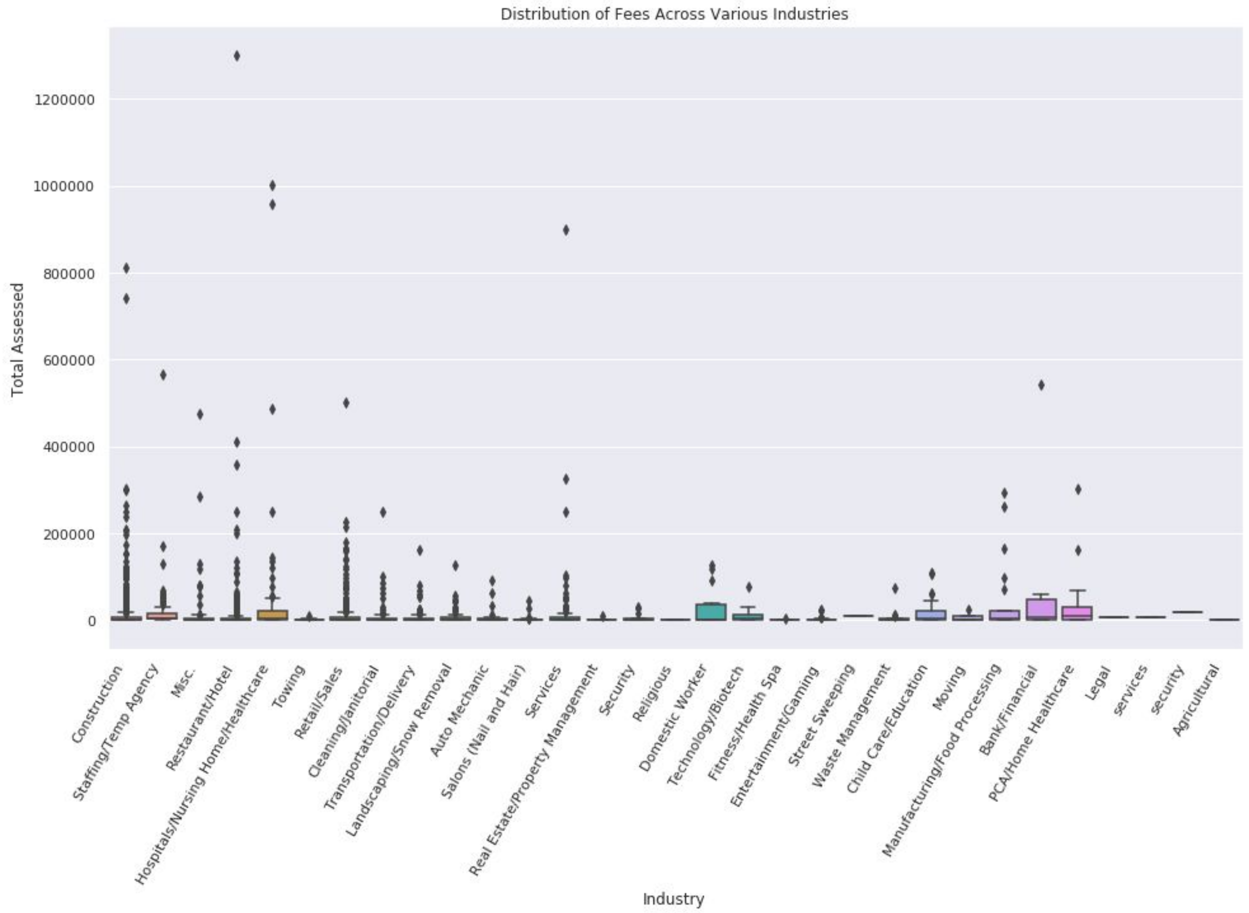

Figure 2. (code available at https://github.com/apolevoda/CS791-ML-PublicHealth-2020/blob/master/cs791-data-visualization.ipynb)

When an enforcement action is taken against an employer, that employer has to pay a restitution and/or a penalty. For each enforcement action taken, the Fair Labor Division Enforcement data contains the restitution, penalty, and the total of the two. We plotted the distribution of these totals across the different industries represented in the data. We see from the graph that across all industries, these totals are generally low, however there is a small number of cases where the totals are exceptionally high. The industries which have the highest individual total of restitution and penalty are Restaurant/Hotel, Hospitals/Nursing Home/Healthcare, Services, and Construction. Despite this, the industries with the highest 75th percentiles are Banking/Financial and Domestic Work.

We find that although the Massachusetts Attorney General’s Office is the entity that the complaints are filed with and takes the enforcement action, a few locations outside Massachusetts are represented in the complaint and enforcement action data. This is somewhat unexpected to us, and our clients seemed to be under the impression that all of the complaints and enforcement were in Massachusetts. Nonetheless, the locations represented in the complaints and enforcement action data are predominantly in Massachusetts.

The most prevalent category of violation in the enforcement action data is non-payment of wages. This is more than twice as common as the next most prevalent category, failure to furnish records for inspection.

Recall that our first task was to identify overlap between Fair Labor Division complaints and enforcements. Originally, we were thinking of training a classifier that, given a complaint, would predict whether that complaint would lead to an enforcement action. A major issue with this approach is that not only is the number of enforcement actions in the data set quite small (around 2700), but as it turns out, for only about 300 of the enforcement actions was there a complaint filed against that employer in that city and state at any point in time. The enforcement action data set is already small as it is, and the fact that it is so unbalanced only exacerbates the issue and makes it that much more difficult to train a good classifier when the problem is formulated this way.

We reported our findings to our clients. Given the aforementioned complications with the approach we previously considered (training a classifier that, given a complaint, would predict whether that complaint would lead to an enforcement action), and the fact that we ideally would predict Fair Labor violations even without complaints being filed, our original approach is not conducive towards our new objective. Instead, we will train a classifier that, given information about an employer, predicts whether that employer is likely to be committing a Fair Labor violation. Since we have the enforcement action data set, we know all of the employers that have been found guilty of Fair Labor violations from early January 2015 to early February 2020.

Methods

We decided to try multiple approaches and compare the results. Andrew mainly focused on SVM but also tried logistic regression and linear discriminant analysis for comparison. For the SVM, radial basis function kernel was used. Ziang focused on Random Forest. We were originally considering focusing primarily on linear discriminant analysis, but after much discussion, we believe that a discriminative classifier would be better than a generative classifier. For our purposes, we have no real need to generate new features. Additionally, discriminative models typically perform better than generative ones at classification. Thus, we focused on other models and then used linear discriminant analysis for comparison.

Results - SVM, Logistic Regression, LDA

The following scores are all averages over 50 runs with different test/train splits of the data, rounded to four decimal places.

SVM:

Metric	Score
Test Accuracy	0.9016
ROC AUC	0.9016
F1 Score	0.9044
Recall	0.9296
Precision	0.8809

Table 1. Result for SVM with radial basis function kernel

Logistic Regression:

Metric	Score
Test Accuracy	0.8708
ROC AUC	0.8714
F1 Score	0.8740
Recall	0.9083
Precision	0.8426

Table 2. Results for logistic regression

Linear Discriminant Analysis:

Metric	Score
Test Accuracy	0.8501
ROC AUC	0.8502
F1 Score	0.8603
Recall	0.9244
Precision	0.8048

Table 3. Results for linear discriminant analysis

Code available at:

https://github.com/apolevoda/CS791-ML-PublicHealth-2020/blob/master/cs791_ago_andrew.ipynb

As initially suspected, SVM outperformed both logistic regression and linear discriminant analysis. Recall was consistently higher than precision. It is important to note what recall and precision actually mean in the real-world context of the project. Recall tells us what fraction of employers that have committed fair labor violations my model has identified as having committed fair labor violations. The precision tells us what fraction of employers that my model has identified as having committed fair labor violations have committed fair labor violations. That we are getting high recall tells us that our model is correctly identifying a very large fraction of employers that have committed fair labor violations. The precision score tells us that the model is predicting some employers that have not committed fair labor violations as having committed fair labor violations. Since we are trying to inform efficient utilization of resources, it is important to have high precision. There is room for improvement here, and in the recall as well. However, for the SVM results especially, the metrics are overall still decent nonetheless.

Challenges - SVM, Logistic Regression, LDA

For additional features, I supplemented the AG’s Office data with data from the Massachusetts Department of Licensure and federal data from the US Department of Labor. From the former, I worked with DPL disciplinary action data; from the latter, I worked with Occupational Safety and Health Administration (OSHA) data and compliance action data from the Wage and Hour Division. The biggest challenge was probably getting all this data in one place so that I could train a model with it. The DPL data was organized in different spreadsheets for different fiscal years and within each spreadsheet, the data was scattered across multiple sheets based on the industry of the licensees. The US Department of Labor data was also difficult to work with for several reasons. The OSHA data in particular was split across many different spreadsheets, with each spreadsheet being limited to 1 million records. Additionally, for all of the US Department of Labor data, the column labels are not always intelligible and there are separate data dictionaries which list what each of the column names actually means. Thus, it was rather involved and time-consuming to even load the data into pandas and determine which columns were relevant.

There were a couple of features, such as employer industry and location, which were high cardinality categorical features. This made it not so practical to one-hot encode them. Thus, I employed feature hashing for these high cardinality categorical features.

There is a significant imbalance between the classes. There are approximately 1400 distinct employers represented in the Fair Labor enforcement actions data set, and there are approximately 22,000 distinct employers in the Fair Labor complaints data set, excluding those that are also present in the enforcement actions data set. When I first attempted the naive approach of training the SVM on all of the data, I got accuracy close to 1, but the F1 score was close to 0 and the ROC AUC was close to 0.5. This indicates that the model was most likely just guessing the majority class most of the time, and that it is not very useful for the task of classification. I worked around this issue by randomly undersampling the majority class so that the two classes were about equally represented. Doing this resulted in a small drop in test accuracy; however, the F1 score and ROC AUC both improved dramatically. This tradeoff is most definitely worth it as the model becomes far more useful for the purpose of classification.

Challenges - Random Forest

In the random forest method, we used the bagging method. We have selected 15 different features out of 26 features. In the random forest method, we have used 10 trees inside the random forest. We have discussed the dataset above which is an imbalance dataset. The Random Forest method is good at dealing with high dimensional data like this. However, the imbalance dataset definitely leads to an overfitting result which we can see in the chart below. We have a good accuracy which is about 98.6%, but the F1 score is moderately low which is only 0.1. This implies the classifier overfitting with the imbalance data. So in our data, there are more than 26,000 data which are labelled as “false” and 400 are labelled as “true.” After I play some tricks on the dataset, where I downsampled some data from those 26000 “false” data to 4000. However, this wasn’t fixing the problem essentially. One reason for this is the essence of random forest is to split at each node. So, those 400 data points didn’t quite share the same features. So when we train the model, the machine will try to achieve the best score which will lead to an undesired result. However, if the data is balanced and if there are somehow more clearer features which are shared by “true” data. I believe random forest will perform well.

Random Result

Random Forest

Metric	Score
Test Accuracy	0.986

F1 Score	0.1
Recall	0.98
Precision	0.505

Conclusion

Using the AG Office’s data together with federal data, we trained classifiers to predict whether an employer is committing a Fair Labor violation. With fairly good accuracy, F1 score, and ROC AUC, our classifiers identified which employers in our test data had committed Fair Labor violations. In the process, we learned a lot about data cleaning and preprocessing, as well as working with imbalanced datasets.