Stanford ENGINEERING
Civil & Environmental Engineering

CIFE
CENTER FOR INTEGRATED FACILITY ENGINEERING

Predicting, Analyzing, and Educating on Wage Theft with Machine Learning Tools

By

Tessa Johnson, Forest Peterson, Michael Myers, Ruth Silver Taube, & Martin Fischer

CIFE Technical Report #TR229
October 2018

STANFORD UNIVERSITY

If you would like to contact the authors, please write to:

c/o CIFE, Civil and Environmental Engineering Dept.,
Stanford University
The Jerry Yang & Akiko Yamazaki Environment & Energy Building
473 Via Ortega, Room 292, Mail Code: 4020
Stanford, CA 94305-4020

Predicting, Analyzing, and Educating on Wage Theft with Machine Learning Tools

Tessa Johnson,[1] Forest Peterson,[2] Michael Myers,[1] Ruth Silver Taube,[3] Martin Fischer[2]

[1]Insight Data Science

[2]Stanford University Center for Integrated Facility Engineering

[3]Santa Clara University School of Law

Abstract

Could novel data science approaches to wage theft assist both public policymakers and enforcement investigators? Wage theft reduces the resources available to the workforce, that results in reduced health of both the wage earner and their dependents. There are wages that society has agreed on as living wages. Wage theft undercuts living wages. Detecting wage theft has proven difficult. To-date, wage theft organizing has focused on public policy by advocating for ordinances. Wage theft is a value proposition, like any business decision, decisions follow a test if income exceeds expenses. Therefore, those least able to resist wage theft become victims. Pragmatically that has been low-income workers. Vulnerable populations feel a disproportionate impact of wage theft due to their already low wages. As a measure of success, this research uses a pragmatic review by a wage theft investigator. The data science solutions presented here are novel in that they are first data science solutions. This pilot found a value in data science for policymakers and enforcement investigators. The authors recommend continuing research on data science to support public policy and investigation. They recommend a focus on data features of the context to support machine learning. With this, pilot implementations with partner organizations can proceed.

Introduction

Every year, tens of billions of dollars are stolen from US workers through wage theft,[1] which can take the form of shorted hours, unpaid overtime, and pay below minimum wage, among other things. The problem is widespread and under-investigated, and even in the case of court-ordered payment of back wages to workers, the payout can go unenforced. Local advocacy can

[1] "What is wage theft," UCLA Labor Center, last accessed 10/17/2018 www.labor.ucla.edu/wage-theft/

Note: Portions of this paper were published by Dr. Tessa Johnson in Medium.com as "Fighting Wage Theft in California with Data," June 28, 2018.

significantly help vulnerable workers to receive their fair pay. The Wage Theft Coalition of Santa Clara County is one such advocacy organization of volunteers that help local workers in need. In collaboration with the Wage Theft Coalition of Santa Clara County the first author researched a data science approach to public policy and enforcement investigation.

Literature review

Institutional protocols underlie the degree of wage theft prevalence. For example, legal and immigration.

In legal protocol, a provision called Rule 68 is an offer by a defendant just before trial to settle through a judgment. If that offer is not accepted by a plaintiff, then the plaintiff could be responsible for the defendant's legal costs. As a result, if a plaintiff has low resources, they must carefully consider the risk they take. For some demographics, such as low-income workers, there is a strong incentive to accept an offer. There is evidence that this incentive is sufficient to accept unjust offers (Ruan, 2010). These accepted offers are despite that it is clear at that point that they will prevail with a decision for a just compensation. For this reason, Rule 68 offers are routinely declined by courts in race and sex discrimination actions. Rule 68 offers are not routinely declined by courts for wage actions.

Before the protocol of Rule 68 deflates a strong case, there are initial challenges to bring an action. One challenge is a threat of deportation for those with an unsettled immigration status. That is a strong deterrent to immigrant workers. There is evidence that immigrant workers are less likely to bring an action (Fussell,2016). This threat is known by employers, who then target immigrant workers for underpayment.

If victims do bring an action against their employer, the legal incentive is for those victims to accept less than they should have received. There is a combination of low risk of action, a discounted liability for wages, and no penalty. There are known workforce demographics that are not likely to bring a complaint. This is known by employers. Therefore, there is an increased prevalence of wage theft in those demographics. Given these examples, there is not much incentive for an employer to follow the law. Employers have no incentive to provide immigrant workers a fair wage.

In real terms, what if bank robbery were given a similar protocol. Say that there was a demographic that was not thought of as bank robbers. Therefore, that demographic is not investigated for bank robbery. Combine that with if a bank robber is caught, then they are only required to return a portion of what they stole. This means they keep a portion of what they stole. There is no penalty. Would you rob banks if you were in that demographic? You could rob a bank every day, if caught, you could simply pull your mask and say, “well played, well played,” grab a handful of cash out of the bag and hand the rest back; going on about the rest of your day.

One proposed solution to wage theft relies on a network of employers. The current situation is one where firms that subcontract find lower costs of business. An unstated premise is that enforcement of wage laws across many subcontractors is difficult, therefore, some of the lower cost of business is due to an increase in wage theft prevalence. Rogers (2010) argues for a third-party liability for wage theft based on a test of reasonable care to monitor wage theft. There is a longstanding legal precedent of responsibility as a third party. If a firm does not exercise reasonable care in observing a potential subcontracted firm and then declines to do business with those firms that use wage theft then that firm is jointly liable for the wage theft. The goal is to disincentivize wage theft.

However, a third party liability relies on A) an ability to define, observe, and measure reasonable care, and B) an ability to investigate and penalize wage theft. As shown above, B is a problem. There are two limits with B, investigation efforts are limited given an immigrant demographics and penalties are limited.

Taking a third party liability as a solution, is there a way to enforce that liability. Are there published methods to assist investigators to monitor third-party liabilities and methods to assist employers to avoid doing business with wage theft firms?

- Public policy: Monitor firms avoidance of wage theft
- Private policy: Avoid wage theft by prioritizing firms to investigate.

The closest was statistical approaches to show the topography of wage theft. An example is Galvin (2016) which found a statistically significant positive correlation between stronger penalties and the prevalence of wage theft.

There has been a focus on technology approaches to timecard apps, exemplified by the "Location History Analyzer"[2] developed by Hack4Impact in collaboration with Community Legal Services of Philadelphia.

Early efforts by the authors to use a data science approach focused on a timecard app and a machine learning model to weight features of the Federal investigation dataset, however, the results provided motivation to continue researching (Starkman, 2016).

This literature review for a data science approach to wage theft found no academic publications.

Methodology

For my data project as a fellow in the Silicon Valley Insight data science program, I worked with members of the Wage Theft Coalition of Santa Clara County to utilize past investigative data in two ways:

1. Deploying a web application[3] which provides a local landscape of wage theft for cities in California through a presentation of statistics for similar cities

[2] Hack4Impact "Location History Analyzer" hack4impact.org/projects/cls-lha
[3] "Find Wage Theft Statistics for Cities like Yours!", last accessed 10/17/2018 wagetheftincitieslikemine.site

2. Building a predictive model to target repeat offenders of wage theft, helping local governments to prioritize investigations (and/or businesses given a scenario of third-party liability).

I used the Wage and Hour Compliance action data from the Department of Labor. Each datum is a wage theft case. These data include the name and location of the companies which were investigated, how many case violations were found of a certain type, and the amount of back wages found owed to employees. The case finding dates range between 2005–2017. There are 250,000 cases in the Wage and Hour Compliance action data and can be found here.[4]

Notable cases from the investigation data:

- New United Motors Manufacturing in Fremont owed over 800,000 dollars to over 3,500 employees in just a one year period.
- Nestle in Glendale owed over 2.7 million dollars in overtime wages to almost 3,000 workers between 2006 and 2008.
- Taco Bell on Glen Bell Way in Irvine owed over 250,000 dollars in overtime back wages to over 1,000 workers over a two-year period starting in 2014.

Tools for local advocacy

The first project was to create a web application meant to help with local advocacy in California cities. A piece of information that I received from the Wage Theft Coalition of Santa Clara County was that local governments were most responsive to information about what was going on in their local community. For many cities in California, the landscape of wage theft is not available due to a lack of past investigations and lack of community advocacy. The map of California below shows the cities in blue which have less than twenty investigations since 2005 (for cities with a population greater than 5,000).

[4] U.S. Department of Labor Data Catalog, Wage and Hour Compliance Action Data, last accessed 10/17/2018 enforcedata.dol.gov/views/data_catalogs.php

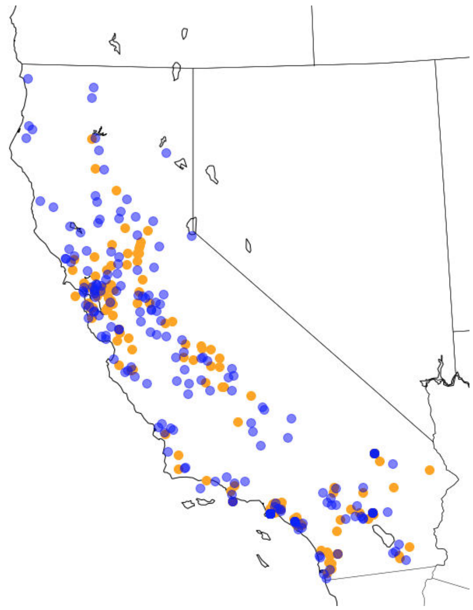

Figure 1: Blue points represent cities that have had less than twenty investigations into wage theft since 2005 (with a population above 5,000). Orange points represent cities that have had greater than or equal to twenty investigations.

To supplement statistics for a single community while providing a local story, I grouped like cities together. Grouping was with a K-means clustering based on eleven economic and demographic variables. I then built a web application for advocacy groups and press which delivers simple statistics about wage theft for a California city with information collected from all cities within its cluster. The web application can be found here: wagetheftincitieslikemine.site.

Prioritizing investigations

Based on past advocacy, I expect that the volume of company investigations into wage theft will increase given advocacy using localized information from the above "wagetheftincitieslikemine.site". If investigative resources are limited, a predictive model will benefit sorting pending investigations. A sort will place near the top of the list those investigations likely to find a high volume of violations. Investigations could be by public agencies or private firms. With a sorted list, the cost of investigations are reduced.

The investigative data does not include context information on companies themselves. The information available is company trade, legal name, city, number of case violations, and back wages owed per violation type. This was a limitation as that is not much to work with in terms of features for a machine-learning model. This finding parallels results of a 2016 machine learning pilot by the Santa Clara County Wage Theft Coalition with DataKind SF (Starkman, 2016). The ideal features would be the size of a company, its yearly revenue, and the number and demographic makeup of its employees. Obtaining that data would require engaging in an extensive search which would span far outside of the Insight project timeframe. Since that data was not available, I supplemented the investigative data from three sources:

- Economic and demographic information that was used for the K-means clustering of California cities
- Business type, taken from the suffix of the legal name of the company (corporation, LLC, FLC, etc.)
- Business franchise, inferred from whether or not the same business legal name from a different location was present in investigative data.

The industry type and business type are categorical variables and were converted to binary vectors before their inclusion in the model.

Since the size of a company is unavailable, an actual number of case violations has little meaning. I decided to separate investigations into two classes and predict between them with a logistic regression model. The two classes that the model predicts between are “non-systematic”, meaning that an investigation will find zero or one violation in the company, and “systematic”, meaning that the company will find more than one violation. My underlying assumption is that a single violation may have occurred through an oversight, where multiple violations indicate a systematic problem within the company. The range in the number of case violations can be seen in the figure below, with a maximum value of over 3,000 in one company.

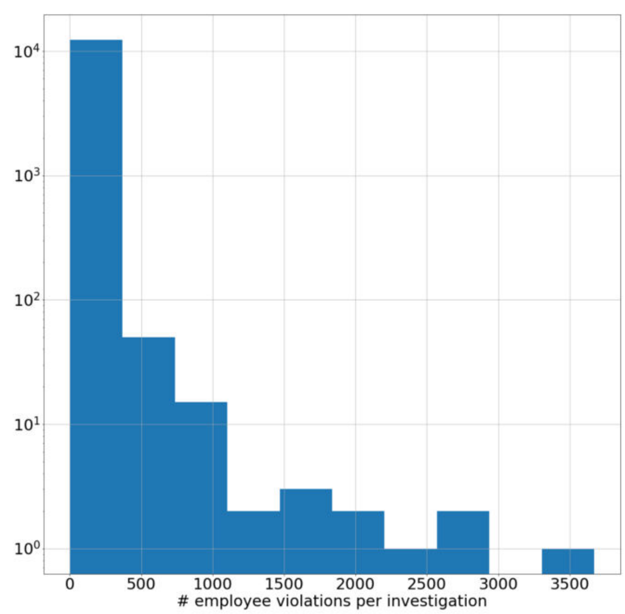

Figure 2: The number of employee violations per investigation is shown above, for investigations into wage theft in California between 2005 and 2017.

For logistic regression modeling, I used the US Department of Labor Wage and Hour Compliance Action data. The compliance action data has 22,000 investigative data points in California. The data were shuffled, with 60% used for training, 20% used for cross-validation, and 20% set aside for final testing and quoting on the performance of the model.

My choice of a metric to describe the model's performance is based on a desire to use investigative resources most efficiently — a goal is high precision in predicting companies with a systematic wage theft problem so that those companies can go to the top of an investigative list. However, I don't want such high precision that no companies are flagged as systematic. I chose to set a minimum of 10% for the fraction of companies flagged as "systematic," which was upheld while using cross-validation data to tune for the model's decision boundary.

I attempted various combinations of variables. The highest precision was achieved with all three sources of supplemental data (listed above) plus the company industry type. A 0.67 decision boundary was found when optimizing for precision with the cross-validation data, which leads to a precision of 70% while flagging 10% of companies as "systematic," measured with the data set aside for final testing. However, recall at this decision boundary is poor, just 14%, giving an F1 score of 0.23. The poor recall leads to a caveat that, if resources allow, all companies on an investigative list should, in fact, be investigated, This is a point in the caveats section below.

Bias in the model

The model is built on investigation data from companies that were not chosen by a random sampling (reiterated in the caveats section). The investigations were prompted due to situations like a worker tip-off, notice of a bad report, or local advocacy. The bias in the model due to this selection of investigations can be inferred by comparing the importance of different features to the model.

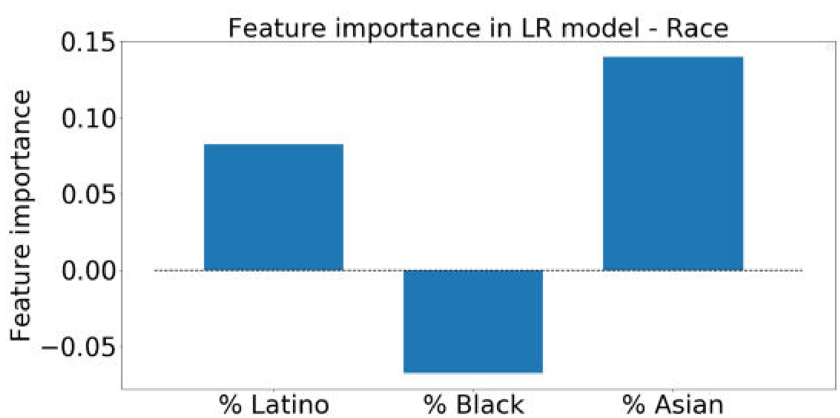

Figure 3: Feature importance as a function of race, possibly showing indications of data selection bias in the model itself.

The bar chart above shows a comparison of the importance of a communities race to systematic cases. It shows that as the percentage of Black citizens in a town increases, the chance of systematic wage theft decreases — the opposite is seen with the percentage of Asian citizens. Rather than taking this relationship at face value, we need to remember that local advocacy plays a big role in whether or not companies are investigated. It could be the case that Black communities are less likely to have local advocacy for wage fairness, and the opposite with

Asian communities. In fact, there is a large worker advocacy group in Los Angeles called the Koreatown Immigrant Worker's Alliance[5] which has successfully advocated for wage theft in the past, possibly contributing to this feature importance.

Caveats of the model

This model can be used to prioritize a list of investigations, but the caveats of the model must be listed:

- The model is built on an inherently biased sample of companies — the companies are often investigated due to local advocacy, a suspicious report, or a tip-off from an employee. This investigative list is not a random sampling, and therefore cannot be used to represent a random sampling of companies.

- At the metric chosen for precision, the recall is very poor. If investigative resources are limited, and only a subset of companies can be investigated, the model can be used to increase the likelihood of finding companies with a systematic wage theft problem. Because the model's recall is so poor, companies flagged as non-systematic should also be investigated if resources allow.

Summary

Wage theft is a prevalent problem for low-wage workers in the United States, and advocacy is usually achieved through reporting on the local problem. This paper presented two novel data science implementations: One is a web application that presents local statistics of wage theft for cities in California, supplemented by data from cities of similar demographic and economic types (clustered with K-means.) The second is a logistic regression model to help prioritize pending investigations of wage theft, however, the model is built on data with a selection bias and should not be used on a random sampling of companies.

The authors recommend that before attempting further data science implementations, that research is completed to define case features (fields) that support machine learning.

The authors thank the many community groups that advocate for wage theft ordinances and diligently reach out to immigrant workforces to raise awareness about wage theft laws and direct victims to community legal clinics.

Bibliography

Fussell, E." The Deportation Threat Dynamic and Victimization of Latino Migrants: Wage Theft and Robbery," The Sociological Quarterly, 52:4 (2011): pp. 593-615

[5] Koreatown Immigrant Worker's Alliance, last accessed 10/17/2018 kiwa.org

Galvin, D. “Deterring Wage Theft: Alt-Labor, State Politics, and the Policy Determinants of Minimum Wage Compliance,” American Political Science Association, Vol 14, no. 2 (June 2016): pp. 324-350

Rogers, B. “Toward Third-Party Liability for Wage Theft,” Berkeley Journal of Employment and Labor Law, Vol. 31, no. 1 (March 2010): pp. 1-64

Ruan, N. "Facilitating Wage Theft: How Courts Use Procedural Rules to Undermine Substantive Rights of Low-Wage Workers," Vanderbilt Law Review 63, no. 3 (April 2010): 725-760

Starkman, R. “Tackling Wage Theft in Silicon Valley,” Huffington Post, November 29, 2016