Beyond the Picket Line: A Data-Driven Exploration of Strike Predictors and Socio-Economic Factors

Leon Zong, Tarun Paravasthu, Madeleine Johnson, and Anuraag Kumar

Abstract— **In this predictive final project, a variety of datasets related to strikes, wages, and employee satisfaction were analyzed in order to discern the main contributors to unions deciding to strike. Our analysis incorporated advanced data science techniques such as feature engineering, clustering our strike dataset by industry, utilizing classification models to predict strike occurrences, and employing PCA dimension reduction for visualization. Contrary to conventional beliefs, our findings highlight that employee job satisfaction stands out as the key predictor of strikes, emphasizing the crucial role of the human element in workplace dynamics.**

I. Introduction

Navigating the intricate dynamics of labor relations requires an understanding of the factors that influence unions' decisions to begin a strike. In the political landscape, where union actions have once again captured public attention, our research details an exploration into the predictors and socioeconomic factors that drive unrest in unionized industries. This research utilizes a diverse collection of socioeconomic datasets.

As unions' actions have once again entered the public consciousness, our project seeks to uncover the variables that contribute to unions deciding to strike. Despite the rising visibility of union activities on the national stage, there has been a stark decline in union membership, from 20.1% in 1983 to the current 10.1% of Americans [1]. This further complicates the changing landscape of organized labor. Our aim is to bridge the gap between the perceived rise in union actions and the historical and modern patterns of labor participation.

The primary dataset for our analysis is sourced from the Bureau of Labor Statistics (BLS) Work Stoppages Program (WSP), supplemented by additional data from the Federal Reserve Economic Data (FRED), Kaggle, and Gallup. These datasets collectively provide a rich source of information, allowing us to paint a picture of how strikes, wages, and job satisfaction relate to each other.

Our analysis incorporated a large range of data analysis techniques –feature engineering, clustering our strike dataset by industry, utilizing classification models to predict strike occurrences, and employing PCA dimension reduction for visualization.

A key finding that emerged from our analysis is that employee job satisfaction from Gallup Polls stands out as the most significant predictor of strikes. This finding challenges conventional assumptions that strikes are mainly motivated by underpay and highlights the importance of the human element of the workplace that we often neglect.

II. Literature Review

To lay the foundation for our analysis, we did a literature review, drawing insights from previous research to inform our analysis of the history of unions, strikes, and employers.

Strikes have a long history in the United States. We first conducted a historical analysis of strikes to learn about their history. Strikes range in size, length and date of occurrence. From strikes like the first recorded – the Lowell Mill Girls strike in 1834 – to the 2018 interstate teachers strike, strikes play a major economic and social role, especially in the pursuit of reducing wage inequality [2], [3]. We identified laws like the National Labor Relations Act of 1935 and the Labor Management Relations Act (1947) as foundational legislature in respectively enabling and restricting union activity [4].

According to Encyclopedia Brittanica, "Strikes arise for a number of reasons, though principally in response to economic conditions or labour practices (intended to improve work conditions). Other strikes can stem from sympathy with other striking unions or from jurisdictional disputes between two unions…" [5].

Interesting and related research has been done regarding both the economic effectiveness, political effectiveness, and other impacts of strikes. Strikes in the education sector have been shown to result in an increase in support for teachers and labor action, suggesting that an increased number of strikes has potential large-scale impact on support for worker's rights [3]. Decreasing union effects have been shown to be linked to rising inequality, comparable to the stratification of wages due to education [6].

Recent models for labor conflict have also been developed. One such model "predicts that the higher the union bargaining weight, the higher are the workers' average rents and rates of work stoppages in equilibrium" [7]. This provides an incentive into our work, showing that predictive modeling for work stoppages is possible.

III. Data

The primary dataset we are using for our analysis is the Bureau of Labor Statistics (BLS) Work Stoppages Program (WSP) [8]. This dataset contains listings for every major work stoppage "involving 1,000 or more workers lasting one full shift or longer". The dataset includes a total of 655 observations ranging from 1988 to 2023, as shown in Figure 1. Each observation consists of state, involved union, start and end dates, and number of workers participating. They also include the NAICS industry code, a 6 digit code that refers to the specific industry the strike occurred in.

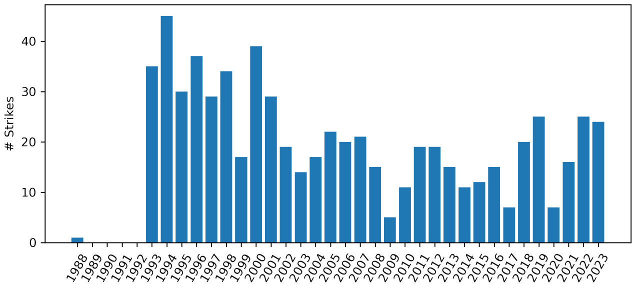

Fig. 1: Strikes per Year

Category	Median (2006)	Median (2023)
Transportation and Warehousing	19.04	29.58
Administrative and Support...	16.16	27.13
Retail Trade	15.24	24.05
Health Care and Education	20.02	33.27
Mining and Quarrying	22.78	38.42
Manufacturing	16.69	26.68
Arts, Entertainment, and Recreation	11.49	21.36
Construction	21.85	37.0

TABLE I: Median hourly nominal wages for selected categories

We match the largest NAICS categories (two-digit codes) on additional data from the Federal Reserve Economic Data (FRED) database [9]. For the top eight categories found in the WSP data, we find monthly time-series median hourly wage data. The series IDs used are listed in the Appendix. A brief description of the wage data is shown in Table I.

We additionally found Consumer Price Index (CPI) data to match our timeframe of analysis. CPI is an index on changing prices of various consumer goods, providing a measure of inflation. This data was downloaded from FRED series MEDCPIM158SFRBCLE, and consists of monthly median CPI from 1984 to the present. Nominal wage can be converted to real wage.

Our final piece of economic data is minimum wage per state. This data was collected from a Kaggle dataset, who originally sourced the data from the U.S. Department of Labor [10]. The Kaggle dataset has already generated an adjusted minimum wage for each state and the nationwide, which we used in our analysis (in 2020-equivalent dollars). Figure 2 shows the adjusted minimum wage for each state.

Our sole social variable dataset is a job satisfaction poll conducted by Gallup [11]. It is a yearly interview-based poll with a large set of variables. We are concerned solely with the job satisfaction data. Data was collected in four categories – satisfied, somewhat satisfied, somewhat dissatisfied, completely dissatisfied – and a fifth category for no response. The distribution over time is shown in Figure 3. Note that several years in the 1990s are missing data.

IV. Hypotheses and Goals

With this data, we aim to analyze strike frequency with regards to the socio-economic factors discussed above: real wage, minimum wage, buying power (CPI), and job satisfaction.

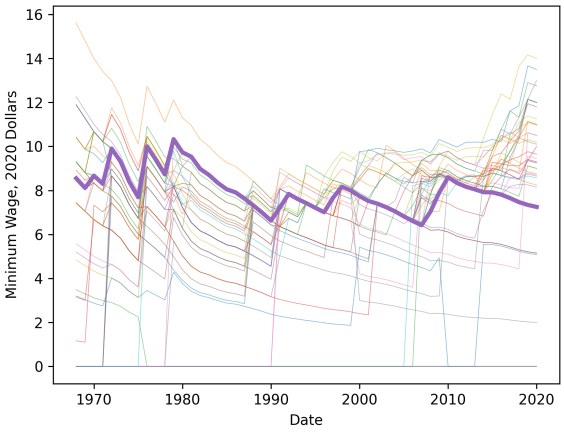

Fig. 2: Standardized minimum wage by state, purple line is federal minimum wage

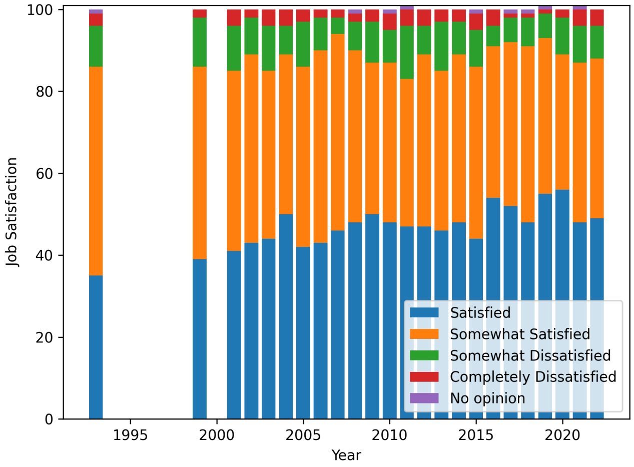

Fig. 3: Gallup Job Satisfaction Overtime

Our primary goal is generate a model which can predict (1) if a strike will occur in a given timeframe and (2) to quantify this by predicting whether an above average number of strikes will occur. We expect that real wage and job satisfaction will be the key features in predicting this. We additionally aim to identify patterns in historical strikes based on the same features.

V. Methods

In this section, we detail the methodology for our data analysis.

A. Feature Engineering

We augmented the strike dataset with the additional socioeconomic data mentioned above. For each strike, we added columns for nominal wage by matching on industry and calculated the real wage from this using the formula $\frac{\text{wage}}{1+\text{cpi}} = \text{real wage}$. We then added the job satisfaction data and minimum adjusted wage, matching by date and state.

On top of using our dataset of strikes, we found it useful to create a dataset of month by month data where we counted how many strikes happened in each month in each industry. We moved through our strike dataset and detected if a strike in each industry happened during a certain month and then created a total. We performed the original dataset merge for every month, generating points with no recorded strike(s).

Additionally, we used Binarization and Fixed Width Binning to convert some of our numeric data into categorical variables to improve classifcation. Specifically, we created binary variables “strike flag” and “above avg” for if a given time period had a strike, or an above average number of strikes. We also used fixed width binning to turn strike length into a categorical variable with values 1-5 for different lengths.

B. *Clustering*

We used a variety of different clustering algorithms and feature cleaning techniques to obtain results. We first aimed to cluster our strike dataset on the wage of the workers, job satisfaction, number of workers and Consumer Price Index at the time of the strike (to quantify purchasing power). We used two different methods to cluster this dataset. We first clustered the dataset by industry (coloring dots depending on the industry of the strike). We also did KMeans clustering on the dataset. In order to visualize our clusters (which you will see in Results), we used PCA dimension reduction.

We then tackled our manufactured dataset of the month by month data. Our month by month data contained data and reported if there was a strike during that time. We clustered our data on average wage and CPI using KMeans Clustering, Spectral Clustering, and by the number of strikes per month.

We used the Calisnki-Harabsz score to optimize the number of clusters we had to use to get an optimal clustering with our KMeans and Spectral Clustering in our above mentioned situations.

C. *Classification*

We employed several classification techniques in order to best predict strike information. Our output variables were number of strikes in a time period, whether a quarter had an above average number of strikes, and whether a month had any strike at all. Our input variables were CPI, Real Wage, Real Minimum Wage, and Job Satisfaction Data from Gallup polls. We iterated and determined that these would are the most useful variables in predicting strike frequency.

With this data, we ran Ordinary Least Squares (Linear) Regression and Logistic Regression, using the Sci-kit Learn (sklearn) package. Depending on the ouput variable, we also partitioned the data into both months and strikes, as outlined in section A. This ensured that categories were relatively balanced. We also used a train-test split of 80-20 on most of the analysis, in order to avoid overfitting.

To verify our results we used the coefficient of determination R^2 for linear regression and accuracy for logistic regression. Accuracy was determined to be useful because due to the different partitions, our data was fairly balanced with about half in each category.

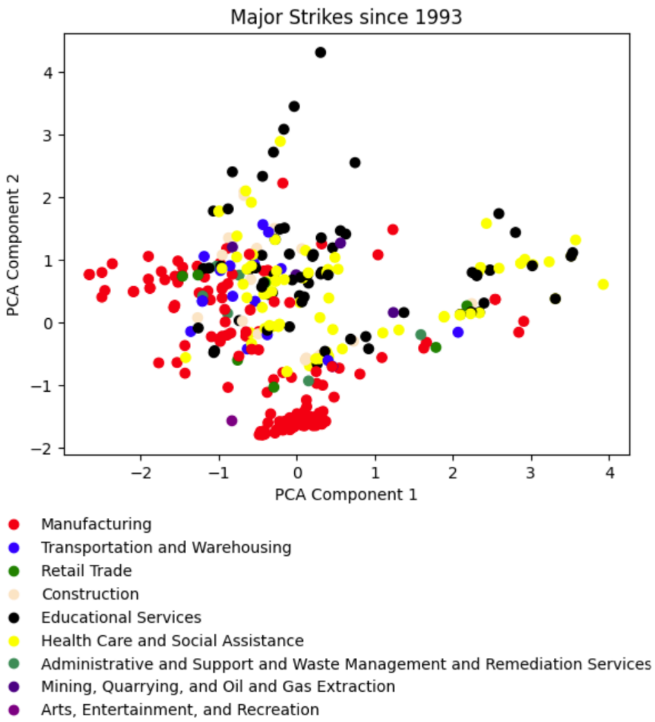

Fig. 4: PCA-reduced visualization of the data, colored by industry label

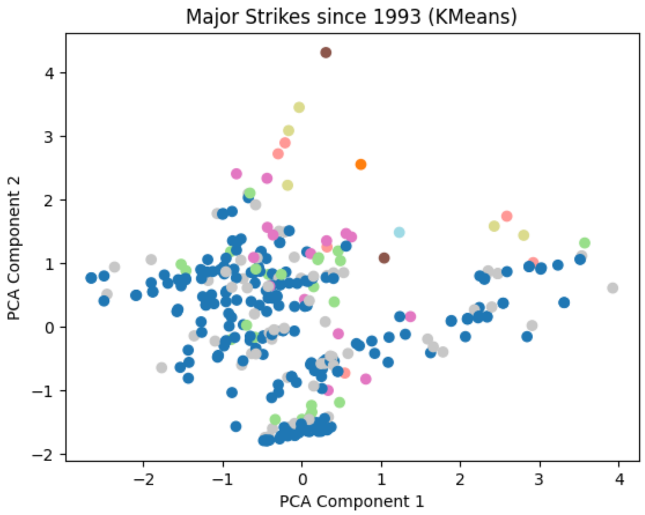

Fig. 5: PCA-reduced visualization, colored by KMeans clustering result

VI. Results

A. *Clustering*

As we mentioned before, we are clustering on two datasets. We first clustered our strikes dataset by a known cluster – industry – shown in Figure 4. We then used KMeans Clustering to create a clustering on 9 clusters (the expected number if strikes would be clustered by industry) to get a Calinski Harabsz (CH) score of 2906.68. The clustering in Figure 5 had a CH score of $\approx$ 1.6.

Despite the large difference between these two scores, we find that the clustering in general looks quite similar. This might be the result of PCA reduction skewing our view or CH score not being an good measure of clustering for this instance. For instance, it seems that manufacturing cluster mostly matches up with the blue cluster in the KMeans clusters.

We followed the analysis of the actual strikes with an analysis of our feature-engineered month dataset using wage and cpi as our variables. We did not need to do PCA reduction as we only clustered on two variables. The clustering in Figure 6 was by the categorical variable – total number of strikes each month. One should notice that the clustering is neither neat nor close to optimal.

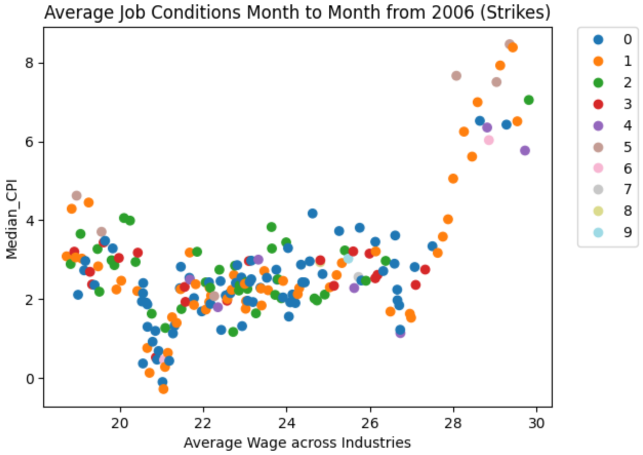

Fig. 6: Clustering using total strikes

The clustering in Figure 6 had a Calinski-Harabsz score of ≈ 1.7 which implies poor performance. We clustered using KMeans in Figure 7 and Spectral in Figure 8 and got a CH score of 522 and 400 respectively which implies a much stronger performance.

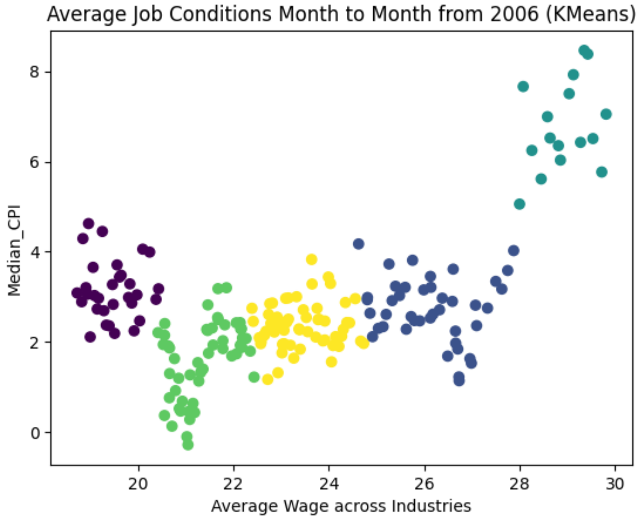

Fig. 7: Clustering using k-means

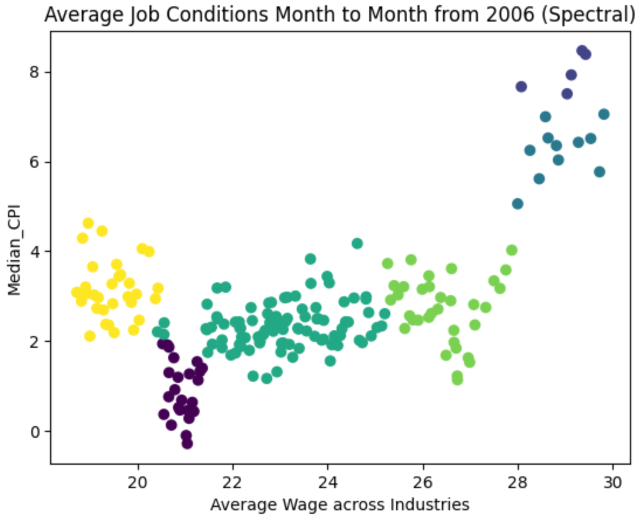

Fig. 8: Clustering using spectral clustering

B. *Classification*

First, we built a linear regression model to predict the number of strikes in a given period. This model was unsucessful, with an R^2 of 0.18 to 0.28 which decreased further when attempting cross validation. Interestingly, the model seemed to capture the trends of strikes, but not the extremity, which gave us hope that it would do well at predicting categorical variables with less possibilities. A graph of these results is shown in Figure 9. Similar results were obtained with monthly data.

Given this result, we then turned to a binary logistic regression model instead, predicting which quarters saw an above average number of strikes. This was much more promising; the model had an accuracy of 85 percent on the test set and 76 percent overall. A confusion matrix is shown in Figure 10.

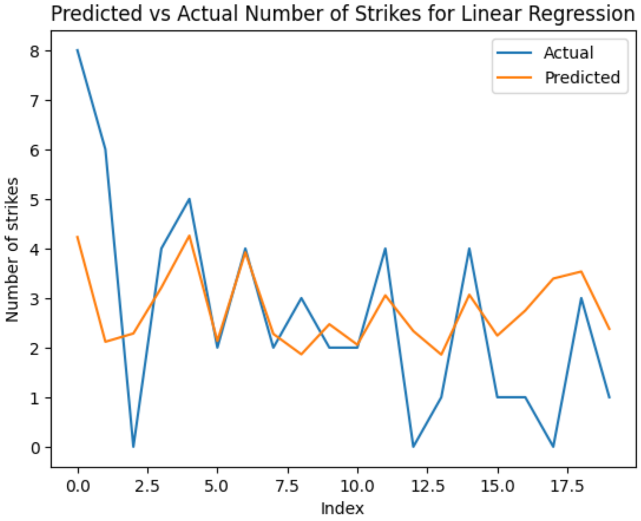

Fig. 9: Linear Regression on test set for Quarterly data

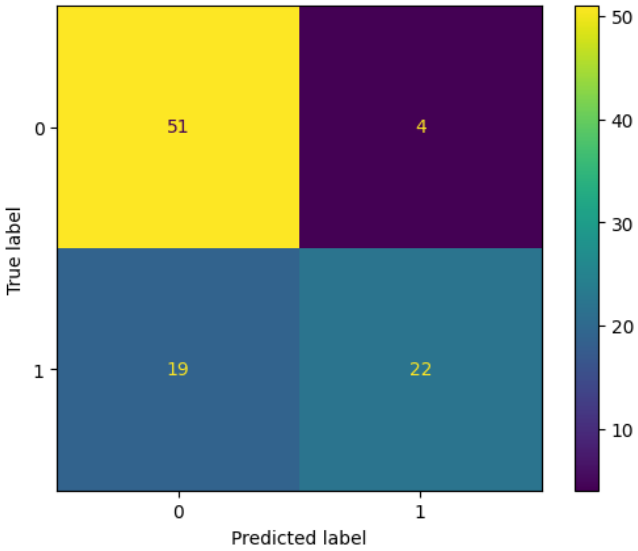

Fig. 10: Confusion Matrix for Quarterly Above Average Prediction

Notably, this is significantly better than just predicting below average for every week, which would only yield an accuracy of 55 percent.

We then built a similar logistic regression model to classify whether or not each month had a strike, using similar variables. This model had an accuracy of 60 percent on the training set, and 68 percent on all the data. While not as good as the previous model, this was still a promising result. A confusion matrix is shown in Figure 11.

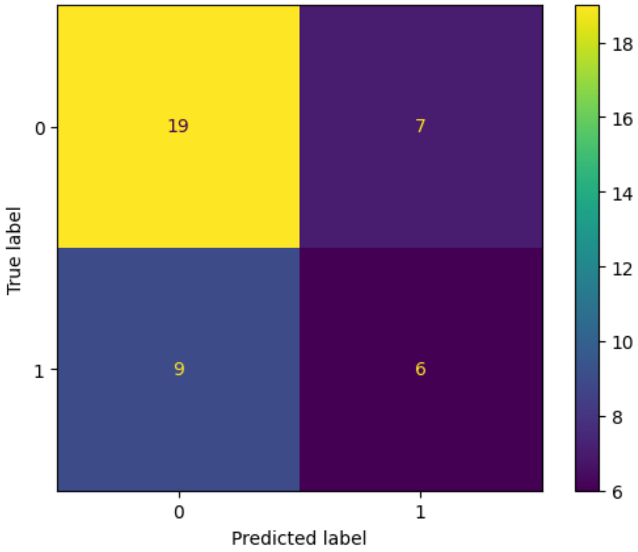

Fig. 11: Confusion Matrix for strike occurrence by month on test set

Finally, we wanted to explore which variables were actually making an impact. Using the a permutation importance algorithm, we generated a graph that shows permutation feature importance for the input variables. Essentially, this is how much worse the model score would be if the feature was random. This is shown in Figure 12.

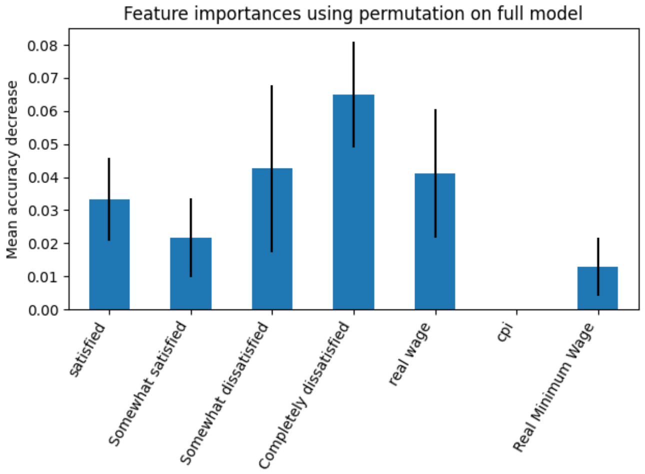

Fig. 12: Confusion Matrix for strike occurrence by month on test set

The results of the feature importance reveal that the Gallup feature – job satisfaction – made the most impact, with real wage close behind. Specifically, the amount of people who answered that they were "completely dissatisfied" made the largest impact, which makes sense. Real Minimum Wage also made some impact. Interestingly, the inflation value itself made no impact.

Model	Accuracy Split	Accuracy Full
Above Average (Quarterly)	0.85	0.76
Above Average (Monthly)	0.71	0.70
Strike Happened (Monthly)	0.60	0.68

TABLE II: Accuracy For All Classification Models

VII. Conclusion

We believe that these findings are interesting and could be useful to policymakers; strikes are generally an indicator that workers are struggling or that the economic situation is dire. Knowing which factors specifically contribute to strikes is also useful to find which metrics are accurate in measuring the labor market.

They are also useful to companies, both those involved and not involved in causing the strike, so they can minimize these conditions from happening, or be prepared for labor strikes as they can create a major negative impact on the economy.

We found when clustering strikes using industries as our cluster labels that our features (Real wage, CPI, Number of workers, job satisfaction) were not able to reveal a pattern in the industries of the strikes. However, on a look at our scatter plot in Figure 4 we find that there are some differences. For example, it seems that manufacturing strikes are in some way different than the rest of them. Also, we found that there seems to be similarities in the conditions leading up to strikes in Educational and Health Care sectors. If given more time, we would've explored this similarity.

Our second set of clustering (on the month by month data) seems to suggest that there is a only a small degree separation between the number of strikes occurring at various levels of Wage and Purchasing Power (quantified as CPI). This implied that strikes are mostly not predicted by wage alone and instead by other variables.

In terms of predicting future strikes, we found the Gallup Poll data on job satisfaction was excellent at predicting periods with strikes or an above average number of strikes. In terms of economic measures, Real Wage and Minimum Wage made a positive impact on the model, which could impact discussions about increasing the minimum wage. The fact that inflation made almost no impact on the model is also interesting, considering what a hot topic it is on the news.

Overall, strikes are difficult to predict! There are many different causes and unique situations with specific companies that may prompt a strike. Furthermore, some workers may strike in solidarity with others. Some things that could be improved in this analysis is the sample size and adding more features to potentially reduce omitted variable bias- we were somewhat constrained by the limitations of our dataset. In any case, strikes are back on the rise in the last couple years (Fig. 1) and this study may help to better understand and inform the actions of workers, companies and policy makers.

References

[1] "Union members summary." https://www.bls.gov/news.release/union2.nr0.htm. Accessed: 2023-12-14.

[2] P. Dray, *There Is Power in a Union: The Epic Story of Labor in America*. Knopf Doubleday Publishing Group, 2011.

[3] A. Hertel-Fernandez, S. Naidu, and A. Reich, "Schooled by strikes? the effects of large-scale labor unrest on mass attitudes toward the labor movement," *Perspectives on Politics*, vol. 19, no. 1, p. 73–91, 2021.

[4] B. Callaway and W. J. Collins, "Unions, workers, and wages at the peak of the american labor movement," Working Paper 23516, National Bureau of Economic Research, June 2017.

[5] https://www.britannica.com/money/topic/strike-industrial-relations. Accessed: 2023-12-14.

[6] B. Western and J. Rosenfeld, "Unions, norms, and the rise in u.s. wage inequality," *American Sociological Review*, vol. 76, no. 4, pp. 513–537, 2011.

[7] S. D. Alder, D. Lagakos, and L. Ohanian, "Labor market conflict and the decline of the rust belt," *Journal of Political Economy*, vol. 131, no. 10, pp. 2780–2824, 2023.

[8] "Work stoppages program." https://www.bls.gov/wsp/.

[9] F. R. B. of St. Louis, "Federal reserve economic data." https://fred.stlouisfed.org/.

[10] "Kaggle minimum wage dataset." https://www.kaggle.com/datasets/lislejoem/us-minimum-wage-by-state-from-1968-to-2017.

[11] "Gallup 2022 work and workplace poll." https://news.gallup.com/poll/1720/work-work-place.aspx.

Appendix

The following series IDs are from FRED and were used to generate the wage feature for each strike: CES4300000003, CES6056000001, CES4200000003, CES6500000003, CES6500000003, CES1000000003, CES3000000008, CES7000000003, and CES2000000003. Data and intermediate files are available upon request from the group. Code is attached in separate files, one for each of data processing and visualization and classification and two for clustering. Thank you!