The Contaminated U.S. Labour Market by COVID-19: Unemployment Spikes and Spreading Speed of the Novel Coronavirus

Yingying Zhou
1002916525
yingying.zhou@mail.utoronto.ca
MI Human Centered Data Science, First Year
Faculty of Information
St. George Campus

Ning Ye
1001806691
angela.ye@mail.utoronto.ca
MScAC, Second Year
Department of Computer Science
St. George Campus

Abstract

In this paper, the impact of the COVID-19 spreading speed on unemployment claims for each state in the U.S. across various sectors is investigated using a difference-in-differences (DID) approach with state-specific time trend techniques. Data forecasting is performed using machine learning methods including linear regression and neural networks. We use panel data concerning the weekly unemployment claims and cumulative cases in each state since the inception of the coronavirus outbreak. This paper strives to investigate the causal effect of the spreading speed of COVID-19 in the U.S. on weekly unemployment claims in different states across sectors as well as forecast of unemployment insurance (UI) claims patterns. We hope the thesis will provide guidance on the implementation of Central Bank's economy revival policy. Code and data are available at: https://github.com/angelaaye/covid19.

1 Introduction

Rationale The recent unemployment spikes in U.S. and Canada have taken a toll on the national economy; the self-quarantine scheme has cost many jobs in the previously healthy workforce. Furthermore, as one of the most severe economic consequences inflicted by the coronavirus, the negative spillover on unemployment has cast real effect on the national welfare, in terms of household aggregate demand, output production and international trade balances.

Empirical evidence so far shows that different spreading speed of COVID-19 across states is likely to cause different geographic layoff patterns in different sectors, depending on their heterogeneous economic resource allocations. If a significant effect is found in the estimation for sectors, then the state government will be able to identify the sources of layoff sectors. Consequently, a series of targeted employment stimulus policies can be made to revive productivity of those that are being hit the least and to lessen the economic damage for the most pandemic-vulnerable sectors.

Research Goals Therefore, the paper strives to investigate the causal effect of the spreading speed of COVID-19 in U.S. on weekly unemployment claims in different states across sectors and to identify and forecast industry-specific unemployment patterns.

Summary of Other Sections The subsequent sections will focus on (i) methodology on model building, theory and data; (ii) analytics and model evaluation; and (iii) conclusions with a brief summary of the research achievements, limitation and policy implications.

2 Methodology

Theory and Literature Review Difference-in-differences approach with fixed effects and state-specific time trend in linear regression is often used for policy impact detection in the file of econometrics. One example would be the paper by Duflo, in which she investigates the policy effect of school construction in Indonesia on education and wage [1]. Policies are often seen as an exogenous 'shock' to the whole system, being analogous to the pandemic outbreak. Thereby, this paper would like to use this method to estimate the impact of the pandemic 'policy' of COVID-19. The model potentially controls for time trend effects and individual characteristics, and takes into account the fluctuating state-specific time trend considering the various levels and focuses of economic development in each state.

We explore the following 23 sectors in our work:

Agriculture/Forestry/Fishing and Hunting	Wholesale Trade
Mining	Retail Trade
Utilities	Transportation and Warehousing
Construction	Information
Manufacturing	Finance and Insurance

Real Estate	Arts
Rental and Leasing	Entertainment and Recreation
Professional/Scientific/Technical Services	Accommodation and Food Services
Management of Companies and Enterprises	Other Services (except Public Administration)
Administration and Support/Waste Management and Remedial Services	Public Administration
Educational Services	Information Not Available
Healthcare and Social Assistance	

Model Design A linear regression model using difference-in-differences (DID) approach with state time fixed effects and state-specific time trend is used to capture the COVID-19 pandemic impact on layoff in each state and sector.

The difference-in-differences linear regression model expression is as follows:

$$Y_{s,t} = \alpha + \delta_{DD} COVID_{s,t} + \sum_{i=Alaska}^{Wyoming} \beta_i STATE_i + \sum_{j=Alaska}^{Wyoming} \gamma_j WEEK_j + \sum_{i=Alaska}^{Wyoming} \theta_i (STATE_i \times t) + \kappa_i sector_k + e_{s,t}$$

The description of the equation parameters is given below:

Target variable: $Y_{s,t}$: number of unemployment claims in state s, week t

Features:

$COVID_{s,t}$: dummy variable, equals 1 if the number of new reported cases in state s during week t is larger than the increased new cases in the previous week in the linear regression MLP model. normalized numeric variable in the panel data

$STATE_i$: dummy variable, equals 1 if the observation is in state i (from Alaska to Wyoming), reference group = Alabama

$WEEK_j$: dummy variable, equals 1 if the observation is in week j since the first case diagnosed in the U.S.

$STATE_i \times t$: state-specific time trend, t = week since the COVID-19 outbreak

$sector_k(\%)$: continuous treatment. The proportion of UI claims from each sector every week

Aside from the feature variables, we also try to predict a bias term, α, as well as an error term, or an unobserved random variable, $e_{(s,t)}$, that disturbs the model.

Although the DID approach captures the interaction between time and a group of treatment variables, it is only appropriate under the assumption that the features formulate a linear relationship with the target variable. However, such conjecture may not hold true due to real-world variances that are not unaccounted for. Therefore, a multi-layer perceptron (MLP) regression network was also explored to learn a more complex relationship between the feature sets and target variable. With the use of an intricate nonlinear activation function between layers, the model is able to capture more informative characteristics between the variables. As a result, the predictive accuracy on the target variable Y would be improved greatly.

To address the potential overfitting problem in the linear model due to the explicit inclusion of over 100 regressors in the regression computation, an alternative approach using the PanelOLS Python package was considered. The dataset was reconstructed in a panel data format, with the state and week being the multi-indices (outer and inner index respectively). PanelOLS package emulates the DID effect, meanwhile only requires four categorical variable groups in regression.

Data Collection Web-scraping was performed to collect weekly data from Jan 21, 2020, when the first case was confirmed in the U.S., to the newest available (week 28, end of July). The weekly number of initial unemployment insurance (UI) claims and monthly number of unemployment insurance claimants by industry are obtained from the U.S. Department of Labor [2, 3]. We have inquired the Department of Labor for weekly UI claimants by sector, but the information was not tracked. Daily reported COVID-19 cases in the U.S. by counties collected from the Center for Systems Science and Engineering at Johns Hopkins University [4] is used to calculate the spreading speed in each state.

Dataset Construction & Preprocessing Raw datasets from three sources were merged into one panel dataset, with observations being implicitly indexed by state and week. One-hot encoding was applied to transform the numeric values into dummy-variable formatting, which is required by the Difference-in-Difference regression approach. To account for the missing weekly initial unemployment insurance claimant data by sector, we substituted the values with the corresponding monthly data for UI claimants. We assumed that if a sector has reported claimants in a specific month, the proportion of claimants in each week of the month is the same. Therefore, we adopt the feature of UI claimants percentage as a proxy for the actual number of claimants.

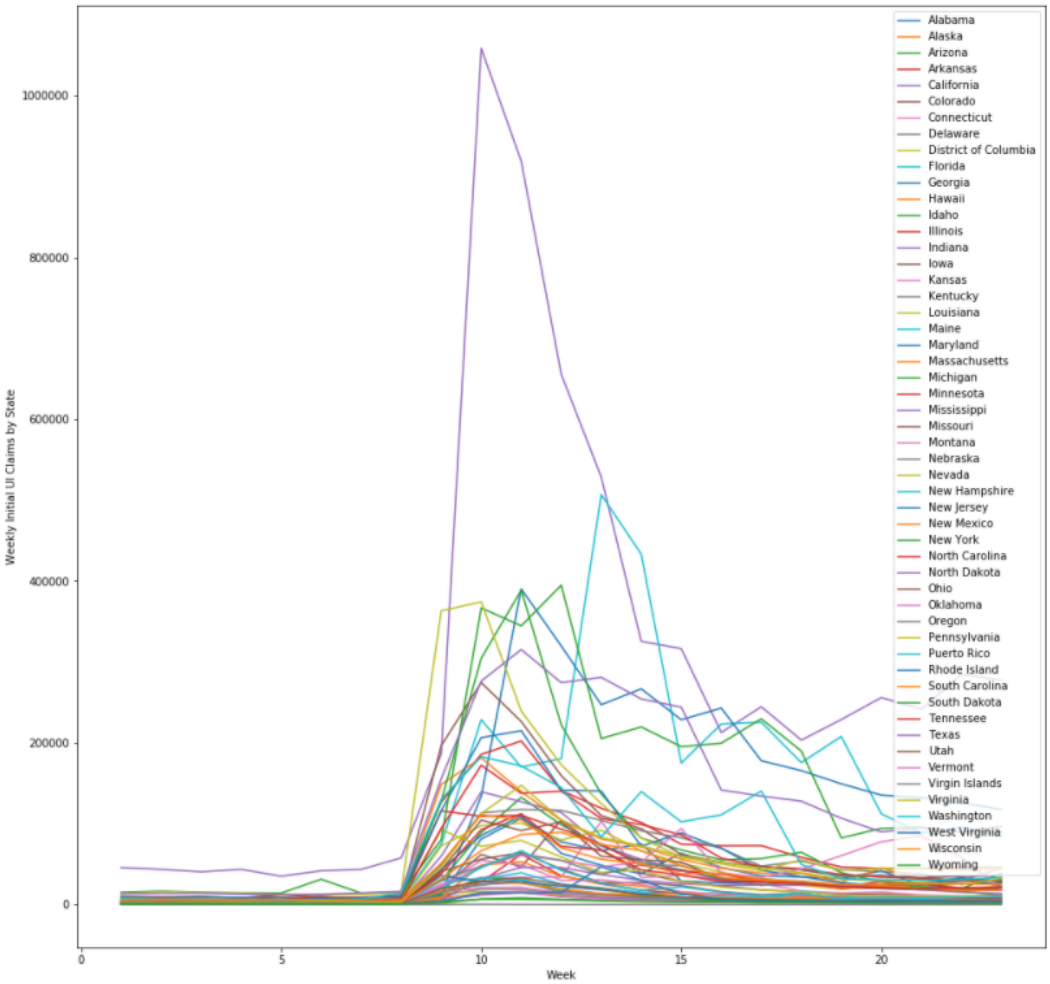

Figure 1: Weekly initial UI claims by state.

Feature Engineering Upon plotting the time series graph of the target variable Y as shown in Fig. 1, we have noticed that the values span from as little as a few hundred to as high as a million. To account for such a skewed distribution, we take the natural logarithm of the dependent variable before proceeding with regression. The motivation behind the transformation is to reduce the

dynamic ranges of the target variable into comparable values whilst maintaining the notion of distance. COVID-19 data in the panel dataset was normalized with mean-center and unit variance to re-collaborate its numeric influence on coefficients and Y.

Exploratory Data Analysis

Time Series Plot for $Y_{s,t}$: As illustrated by the time series plot from weeks 1 to 23 in Fig. 1 and the summary table of state rankings by Y in Fig. 2, the peak at which the number of UI claims is dominant occurs around week 10 or 11. The number of unemployment claims in California outnumbers all other states, reaching its first peak in week 10 at 1,058,325 claimants. It is also evident that California has began to undergo its second wave in week 22, totaling at 284,394. The state with the second highest UI claims is Florida, with 506,670 claims in week 13, only half of that of California. Unlike most states, there is a one-to-two-week delay in unemployment claims spike in Florida. The states with the next largest unemployment claims are New York, Georgia, Michigan and Pennsylvania, each following a similar pattern in trend development and peak.

State	Y_(s,t)	Week	Manufacturing %	Retail Trade %	Transporation and Warehousing %	Information %	Administration and Support/Waste Management and Remedial Services %	Educational Services %	Healthcare and Social Assistance %	Arts, Entertainment and Recreation %	Accomodation and Food Services %
California	1058325	10	7.6	6.4	1.9	5.6	6.7	2.4	5.3	1.1	3.1
California	918814	11	5.4	12.9	2.2	5.4	4.5	3.4	8.6	1.6	6.9
California	655472	12	5.4	12.9	2.2	5.4	4.5	3.4	8.6	1.6	6.9
California	528360	13	5.4	12.9	2.2	5.4	4.5	3.4	8.6	1.6	6.9
Florida	506670	13	2.3	11.5	2.5	1.0	10.7	1.8	9.7	4.5	25.2
Florida	433103	14	2.3	11.5	2.5	1.0	10.7	1.8	9.7	4.5	25.2
New York	394701	12	4.4	12.5	4.2	3.5	8.6	2.4	10.1	3.5	18.2
Georgia	390132	11	12.9	8.5	3.6	2.1	14.1	3.2	14.0	4.0	18.4
Michigan	388554	11	26.4	10.2	3.0	0.6	7.3	0.9	9.0	1.3	8.4
Pennsylvania	374056	10	10.5	7.3	5.0	1.0	12.8	1.4	8.5	3.7	5.7
New York	366595	10	4.9	8.0	3.4	3.5	15.6	1.9	7.7	4.9	9.8
Pennsylvania	363000	9	10.5	7.3	5.0	1.0	12.8	1.4	8.5	3.7	5.7
New York	344451	11	4.4	12.5	4.2	3.5	8.6	2.4	10.1	3.5	18.2
California	325343	14	5.4	12.9	2.2	5.4	4.5	3.4	8.6	1.6	6.9
Georgia	319581	12	12.9	8.5	3.6	2.1	14.1	3.2	14.0	4.0	18.4
California	316257	15	5.4	12.9	2.2	5.4	4.5	3.4	8.6	1.6	6.9
Texas	315167	11	6.7	13.2	2.8	1.3	8.7	2.1	12.2	2.2	17.3
Michigan	304335	10	14.0	5.8	3.5	0.9	14.3	0.7	4.2	3.6	5.1
California	284494	22	5.6	13.8	2.8	6.2	5.0	4.4	8.6	2.6	8.8
Texas	280761	13	6.7	13.2	2.8	1.3	8.7	2.1	12.2	2.2	17.3

Figure 2: Summary table of state rankings of Y. Top 5 states with sector proportions: California, Florida, New York, Georgia, Michigan

Correlation between Feature Variables and Target Variable: The Pearson correlation between the number of COVID cases and initial UI claims is around 47.92%, which is a moderate positive relationship. The correlation between the number of COVID-19 cases and sector claimants is also checked. Sectors related to Administration, Transportation Warehousing, Real State, Rental Leasing and Accommodation Food Services displayed a relatively strong relationship with the number of COVID-19 cases, at around 35%. Lastly, Construction, Wholesale and Retail Trade, Finance Insurance, Professional Services, Health Care, and Arts, Entertainment Recreation industries are moderately affected by the pandemic as well, with a correlation coefficient over 30%.

Top 5 State UI Claimants Characteristics on Sector Compositions: As for the sector compositions, the top 5 states on UI claims present some common patterns. Retail trade in California, Florida, New York and Michigan account for over 10% of the UI claimants. Accommodation Food services in Florida, New York, and Georgia were severely hit, amount to over 20% of the layoff total. Healthcare in California, New York, Georgia and Michigan were impacted as well, at around 10%. Administration-related jobs in Florida and Georgia also went through a heavy blow, taking up 10-15% of their state unemployment.

3 Analysis & Model Evaluation

Experimental Setup Limited by the data on UI claimants characteristics, we were only able to acquire 28 weeks worth of data starting from the week of January 20, when a COVID-19 case was first reported in the U.S. We set aside the data corresponding to the five weeks in July as test set for forecasting UI claims patterns, and used the remaining 23 weeks for training. We have evaluated the data on three models: linear regression, PanelOLS framework, and a multi-layer perceptron (MLP) network. While the former two models attempt to predict a linear relationship between the variables, the latter has more flexibility in model fitting. After hyperparameter tuning, it is determined that a network with five hidden layers, nonlinear ReLU activations, and a L2 regularization penalty of $\lambda = 0.1$ yields the best result. A high regularization parameter is selected to help reduce overfitting.

Model Evaluation Our research conducted linear regression with difference-in-differences approach to investigate the impact of COVID-19 on unemployment in each state by sector. The overall model is significant in 5% confidence interval with a P-value of 0.00.

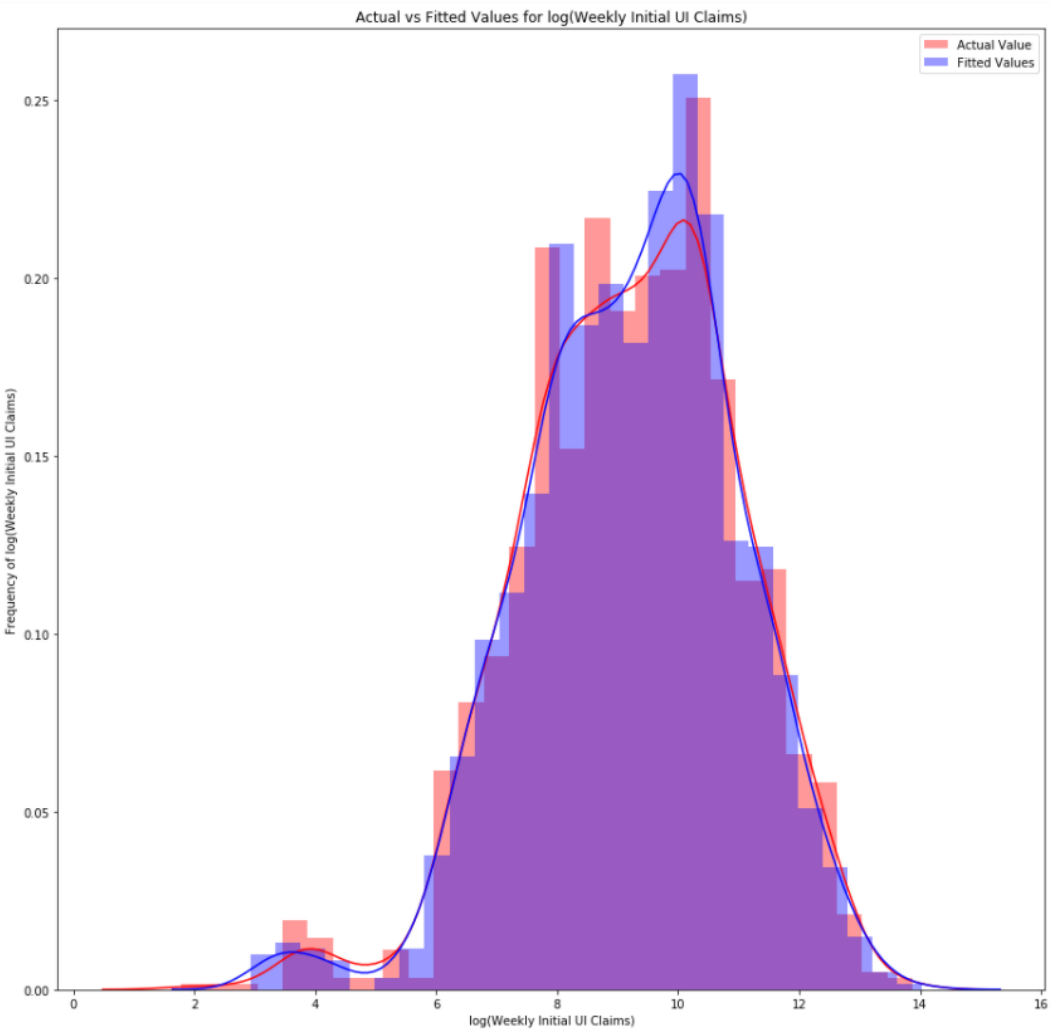

Figure 3: Actual vs. fitted values for log(weekly initial UI claims).

In terms of the model goodness of fit, the coefficient of determination, or R-squared, for both the linear regression and MLP model are surprisingly high, achieving 98% and 99% on the training set respectively. The actual and the predicted values for the target variable are illustrated in Fig. 3. However, as the linear regression model is unable to capture the nonlinear intricacies present in the variables, it performs poorly on the test set and has a negative coefficient of determination. Since

the model parameters are not trained on the unseen test data, a negative value is highly plausible, especially in the case when the test set has a different distribution than the training data. Under the MLP machine learning algorithm, the model prediction accuracy reaches an R-squared of 72% on the test set. Simply put, the model can forecast the target variable, the number of weekly initial claims, at a satisfactory level of accuracy. On the other hand, the PanelOLS method gives a R-squared of around 40%. Empirically speaking, it is a decent result for real-world analytics, indicating that the included feature characteristics can account for 40% of the variation in Y, the weekly initial UI claims in the U.S.

Model Interpretation Given the overall performance of the three models, we set linear regression as the baseline model, due to its inferior interpretability than the PanelOLS model and low prediction accuracy. With the test set score greatly outperforming the baseline, MLP will be mainly used for target variable forecast. Furthermore, due to its ease in interpretation, PanelOLS regression summary statistics table is referred to for the purpose of regression result explanation.

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	9.2532	0.1407	65.762	0.0000	8.9771	9.5293
COVID_(s,t)	0.0016	0.0103	0.1538	0.8778	-0.0185	0.0217
Agriculture/Forestry/Fishing and Hunting %	0.0175	0.0086	2.0458	0.0410	0.0007	0.0343
Mining %	-0.0061	0.0155	-0.3945	0.6933	-0.0366	0.0244
Utilities %	0.1744	0.1989	0.8770	0.3806	-0.2158	0.5646
Construction %	0.0107	0.0031	3.4843	0.0005	0.0047	0.0168
Manufacturing %	-0.0005	0.0024	-0.2163	0.8288	-0.0053	0.0042
Wholesale Trade %	-0.0773	0.0449	-1.7234	0.0851	-0.1654	0.0107
Retail Trade %	-0.0136	0.0189	-0.7228	0.4700	-0.0506	0.0234
Transporation and Warehousing %	-0.0795	0.0201	-3.9498	0.0001	-0.1189	-0.0400
Information %	-0.0614	0.0237	-2.5911	0.0097	-0.1079	-0.0149
Finance and Insurance %	-0.0832	0.0464	-1.7944	0.0730	-0.1742	0.0078
Real Estate, Rental and Leasing %	-0.0031	0.0807	-0.0380	0.9697	-0.1614	0.1553
Professional/Scientific/Technical Services %	-0.0537	0.0132	-4.0615	0.0001	-0.0797	-0.0278
Management of Companies and Enterprises %	0.3256	0.0498	6.5409	0.0000	0.2279	0.4233
Administration and Support/Waste Management and Remedial Services %	0.0386	0.0063	6.1198	0.0000	0.0262	0.0509
Educational Services %	0.0190	0.0097	1.9589	0.0504	-3.038e-05	0.0380
Healthcare and Social Assistance %	-0.0092	0.0096	-0.9649	0.3348	-0.0281	0.0096
Arts, Entertainment and Recreation %	0.0552	0.0130	4.2544	0.0000	0.0297	0.0806
Accomodation and Food Services %	0.0007	0.0043	0.1686	0.8661	-0.0077	0.0092
Other Services (except Public Administration) %	0.0441	0.0231	1.9091	0.0565	-0.0012	0.0893
Public Administration %	0.0033	0.0167	0.1957	0.8449	-0.0295	0.0360
Information Not Available %	0.0007	0.0021	0.3483	0.7277	-0.0034	0.0049

Figure 4: PanelOLS regression coefficient summary table.

As shown in Fig. 4, at 5% confidence interval, the sectors of Agriculture, Construction, Transportation and Warehousing, Information, Professional Services, Management of Companies, Administration Support/Waste Management, Educational Services, and Arts, Entertainment Recreation are significant factors that contribute to the layoff phenomenon. Surprisingly, what we have observed is an ongoing large-scale unemployment wave which followed the inception of COVID-19 outbreak. The model claims the virus itself has very limited impact on the layoff phenomenon, controlling for sectors, state, and time trend. Among the industries, layoff has increased significantly in the industry of Agriculture, Construction, Administration related jobs, Educational Services, as well as Arts, Entertainment Recreation. On the contrary, Transportation Warehousing, Information, and Professional/Scientific/Technical Services are witnessing a deduction in initial UI claimants, or potentially an increase in available positions.

Numerically speaking, administration related jobs will suffer the most from unemployment.

(For every 1% increase in management-level jobs, there will be an additional 3 to 32% increase in initial UI claims). Recreational industry also suffers from some noteworthy job losses. Meanwhile, Transportation Warehousing, Information, and Professional/Scientific/Technical Services benefit from a job market 'boom'. (For every 1% increase of jobs in those fields, initial UI claims drop by 6.3% on average.) These results reconcile with the ongoing job market turmoil that can be observed.

Interestingly, the sectors (Retail, Healthcare, Administration, Accommodation Food, and Manufacturing) identified as most vulnerable in layoff peak weeks for the top 5 states in the process of exploratory data analysis (EDA) were nowhere to be found in the regression result. Instead, Administration, Arts, Entertainment Recreation, Transportation Warehousing, Information, and Professional/Scientific/Technical Services were picked by the model as the determinant factors for Y, the number of unemployment claims.

To reconcile the two statements, note that the EDA only focused on the sector compositions for peak weeks in the top 5 state and several external elements were not included in our model such as the contingency plans enforced by the state authorities. It is also probable that some states rely on different pillar industries, depending on their internal social and geographical resource allocations and so on. The state-specific trait of industry concentration is absorbed by the state categorical feature. In addition, the industry concentration phenomenon might account for their high proportions in related-sector layoffs. We hope that the conjectures proposed above can clear up the puzzle.

4 Conclusion

Achievements Our research investigates the the impact of COVID-19 on industry unemployment in each state by applying linear regression with difference-in-differences approach, machine learning technique for unemployment patterns forecast, and PanelOLS regression for model improvement. This paper managed to identify the state-specific layoff trend in its most vulnerable sectors at unemployment peaks in the preliminary data exploration analysis, which was later cross-compared with the regression results.

Limitations Although the impact of COVID-19 outbreak is almost instantaneous, its damage on the economy might be profound and long-lasting. Therefore, it is hard to be teased out from the cyclical time trend effect. In terms of the feature engineering and model inclusiveness, by the time the research is being conducted, organizations and state authorities have taken actions to ease the economy hardships. Those external factors were not included in the model due to some difficulty in finding proper proxies and value measurement. Although our MLP model is able to achieve satisfactory performance in forecasting the amount of future UI claims, the explainability of the model is quite low, which is a common defect like many other machine learning techniques.

Policy Impact This paper is intended to serve as a guideline for the Canadian government to refer to in industry-related economy stimulus policy decisions. As U.S. and Canada display similar patterns in the socioeconomic structure and potentially the spreading pace of COVID-19 due to their geographical proximity in the early stage before the lockdown. As a result, the derived insights from the case study in the U.S. would have reliable reference value to the provincial authorities and the Central Bank in layoff pattern forecasting and policy implementation in Canada. Given the case study result, it is concluded that the central bank should direct funds and social resources (such as direct loans and funding, and career re-training programs in the communities) to the sectors which are identified as layoff determinant factors by the model. That way, the stimulus policy would be able to ease the pain for the wounded and to boost the job market for the 'benefited' sectors. On the other hand, those province-specific internally 'vulnerable' industries should also be financially aided to halt the panic and ease the economic hardships for the general public.

References

[1] E. Duflo, "Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment," *American Economic Review*, vol. 91, 2001.

[2] "Unemployment insurance weekly claims data." U.S. Department of Labor. https://oui.doleta.gov/unemploy/claims.asp. Accessed: 2020-05-16.

[3] "Characteristics of the unemployment insurance claimants." U.S. Department of Labor. https://oui.doleta.gov/unemploy/chariu.asp. Accessed: 2020-08-31.

[4] "Covid-19 data repository by the center for systems science and engineering (csse) at johns hopkins university." GitHub. https://github.com/CSSEGISandData/COVID-19. Accessed: 2020-08-31.