Employee Injury Claim Project
- 95720 Final Report

Lansing (Yifan) Chen,
lansingchen@cmu.edu

Lanxuan Zhou,
lanxuanz@andrew.cmu.edu

Jiayuan Zhang,
jiayuanz@andrew.cmu.edu

Shanyue Wan,
shanyuew@andrew.cmu.edu

Hongqin Wang,
hongqinw@andrew.cmu.edu

Nathan Wang,
xinyangw@andrew.cmu.edu

May 2022

Contents

1 Executive Summary

Pennsylvania Department of Labor deployed an analytic system developed by CMU students in the previous semesters where the data integration, basic time-series prediction and visualization were completed.In this semester, we continued to work with the PA department to improve the prediction web application.

The overall project objective is to provide client insights on injury rate and build tools on how to predict the future rate. Through this project, we hope to assist the staff in the PA department to do more proactive work that could reduce injury rate.

We started by working on data cleaning and preparation. Then we moved to the exploratory data analysis phase, where trend analysis and various granularity data analysis were conducted. Based on the previous finds, we developed further clustering analysis. Next, we moved to the model stage, where we do time series models and non time series models in parallel. We studied the existing predictive models and added the prediction evaluation function. We also implemented a neural network model to predict injury nature. Eventually, we reached the final stage to redesign the web application to incorporate new features. We conducted tests to make sure the application is operating as expected.

At the end of the project, we were able to deliver an updated web application, with adding the neural network prediction model and time-series evaluation. We also prepared a comprehensive presentation and report, which includes other data analytics.

2 Project Objectives

The core mission of our project is to help the Pennsylvania Department of Labor to mitigate employee injury claims by discovering key factors and developing effective predictive models, so that the client can make data driven decisions. In the last phase of the project, the previous team delivered a web platform that generated an injury rate prediction by integrating two time series prediction models Prophet and SARIMAX. The prediction is generated based on the input dataset along with the user selected county and time span. The web application also incorporated a PowerBI Dashboard to present prediction results. However, time series models limit the underlying factor to solely seasonality, which is not significant for certain industries and injury types. Additionally, the prediction of those two models lack evaluation metrics in terms of the result validity. Therefore, our project targets three main objectives. First, we provided insights on identifying prevention priorities, which is achieved by trend analysis and clustering analysis. Second, we aim to improve the user experience in using prediction models, which is achieved by providing an evaluation report per prediction run as well as an additional neural network model to predict the top N injury nature types for each employee given his or her basic information. Last, we enhanced the web application by integrating new functionalities.

3 Values and Business Impact

This project helps the client to reduce the injury claim count, save injury reimbursement costs and improve the employee satisfaction. In the year of 2020, 147,980 claims were filed, leading to 33,902 days off in total. With an estimate of 8-hour workday and $20 hourly wage, the financial loss from the employee's absence is worth $5.4 million, not to mention the medical treatment costs. Under this calculation, reducing the claims count by 1% may save the client over $1 million annually.

4 Project Methodology

4.1 Phase 1: Understand client needs and examine available resources

Given this is the ninth phase of the PA project, we initiated our project by setting kick-up meetings to learn the requirements, key contacts from the client side as well as available resources from the client. In addition, we requested the project deliverables of the last semester's team from our advisor to generate a better understanding of the final output expectation. With the previous Git Repository supplied by the client, we read through the ReadMe file as well as the tutorial video together during the internal meeting to understand the main functionalities of the web application. Then, we learned the code structure and set up the local environment accordingly to get familiar with the technical resources.

Based on a basic understanding of the client needs and available resources, we decided to manage our project in an agile style. We listed product backlogs while splitting our project cycle into three sprints. Any progress will be reflected on a team-wise Trello board to ensure the transparency of the task status and member contribution. At the end of our planning phase, we confirmed the project scope, resource availability and weekly meeting schedule with the client.

4.2 Phase 2: Explore data and evaluate existing models

We explored the data collectively as a team by extracting six main aspects from the injury claim data set. Each member is assigned with one aspect and present findings during the internal discussion to help the whole team form a holistic view of the data set efficiently. We recorded questions and collected clarifications from the client during the 2nd and 3rd client meeting. In the meantime, two of the members who are more familiar in the machine learning field researched on the existing two time-series models to inspect their principles and effectiveness.

Given key findings from the analytics, we discovered that seasonalities are only significant in certain industries or injury causes. This finding motivates us to implement a non-series model which relies on features other than the injury date. Additionally, our analytics identified key features for injury occurrences which are optimal candidates for the feature engineering.

4.3 Phase 3

4.3.1 Phase 3.1: Implement new prediction model and add model evaluation

In order to offer a brand new prediction perspective other than a high level injury rate per county or industry, we implemented a neural network with word embedding to predict the probability of each injury based on the basic information of an employee.

Besides, considering the underutilization of the prediction web app by the client, we planned to add evaluations to the time-series prediction, in order to enhance the interoperability of the prediction validity. Through the evolution, we also identified key factors to generate high quality prediction in terms of input data size.

4.3.2 Phase 3.2: Integrate functionalities into Web App

Once the backend codes and data were fully prepared, two members started to integrate new functionalities into the web application. We enhanced the existing web application with a neural network prediction moduled, an evaluation report option for time-series predictions and a static dashboard of key analytical findings.

4.3.3 Phase 3.3: Review code and conduct local tests

Before merging code to the group repository, we conducted local tests individually and assigned peers to review code.

4.4 Phase 4: Finalize deliverables and pack materials

Before we started to finalize our deliverables, we drafted the report or presentation structure and collected feedback from the client to make sure no key elements were missing. At this last phase, we prepared video scripts, recorded videos and prepared final presentation slides. With the feedback from the final presentation, we finished our report and packaged all deliverables for submission.

5 Data Analysis

Our data analysis mainly focused on different injury types and industries.

5.1 Injury Cause Type

Below shows us the monthly claim number by different injury cause types from 2017 to 2020. Each curve represents one work-related injury cause. As we can see all these curves go up and down simultaneously. Injury causes, such as Fall, Slip or Trip Injury which is represented by the yellow curve has a very significant seasonality. This is reasonable, considering during wintertime, injuries like falling on ice or snow have a higher probability to occur. It also in some way corroborates the usability of previous models like SARIMAX. Also, from the figure below, we can easily visualize the impact of the pandemic especially in April 2020, when all other cause types went down except the type of Struck or Injured By which is represented by the dark gray curve. This cause curve presents the impact from COVID-19.

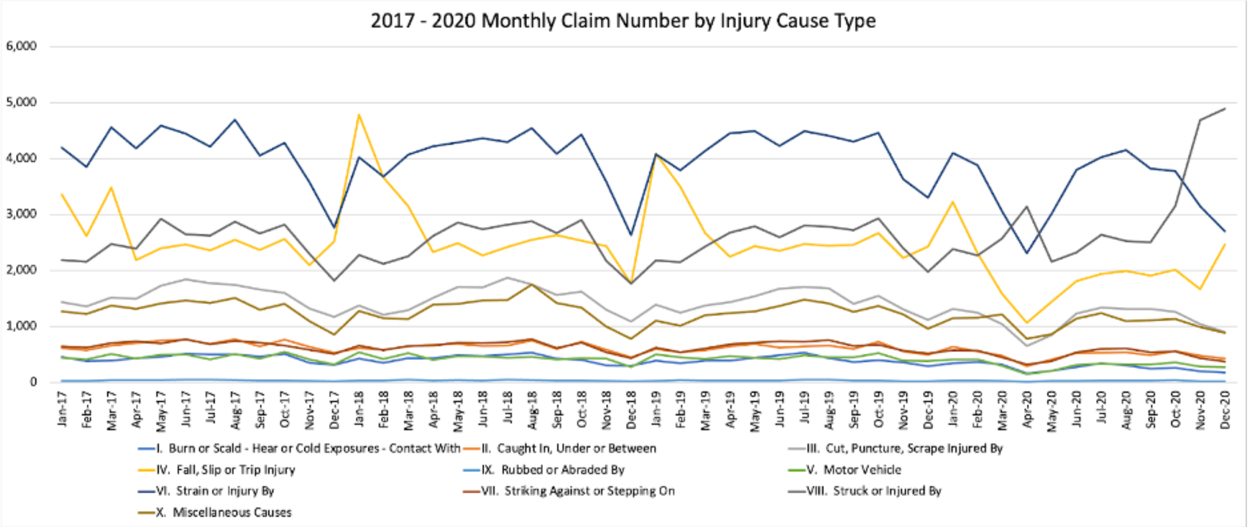


Figure 1: 2017 - 2020 Monthly Claim Number by Injury Cause Type

5.2 Death Cases

Below shows us the trend of the monthly death count and death rate from 2017 to 2020. The death rate is calculated by dividing the death count by the employee population. As we can see, there is no significant seasonality for the death count and death. Since the variation of the death count is way more than that of the number of total employees, the fluctuation of the death rate is fairly consistent with that of the death count.

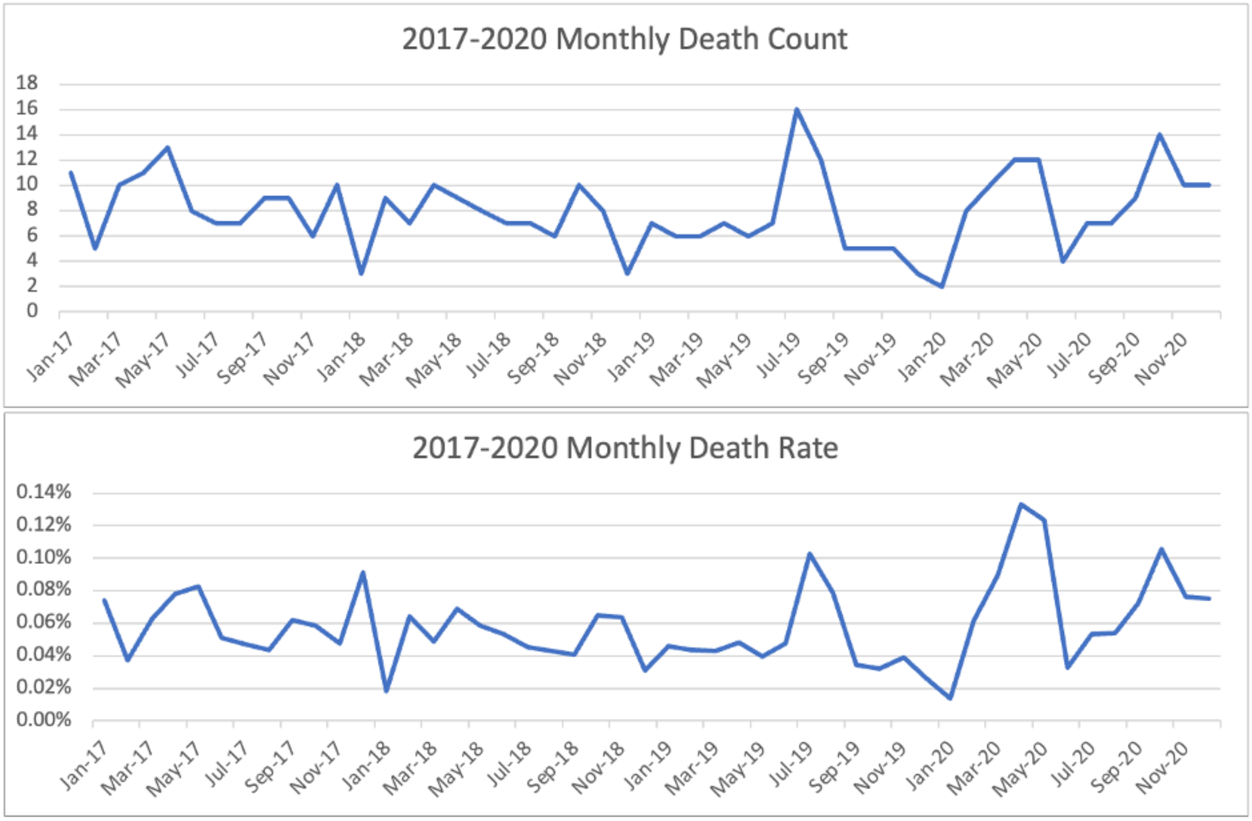


Figure 2: 2017 - 2020 Monthly Death Count and Death Rate

We also investigated top industries that have a high death count each year. Below shows us the top 10 industries ordered by the death count from 2017 to 2020. Those industries are also top industries that have a very high injury count in these four years. As to the death count, we filtered out death cases caused by heart attack and gun shots which cannot be predicted. From this figure, we can see that there are several industries that appear repeatedly in these four years, including Solid Waste Collection; Power and Communication Line and Related Structures Construction; Highway, Street and Bridge Construction; and General Freight Trucking, Long-Distance Truckload and so on. The most common causes of deaths are motor vehicle, collision, or sideswipe, and electrical current. The most common natures of deaths are electric shock, crushing and COVID-19 typically in 2020.

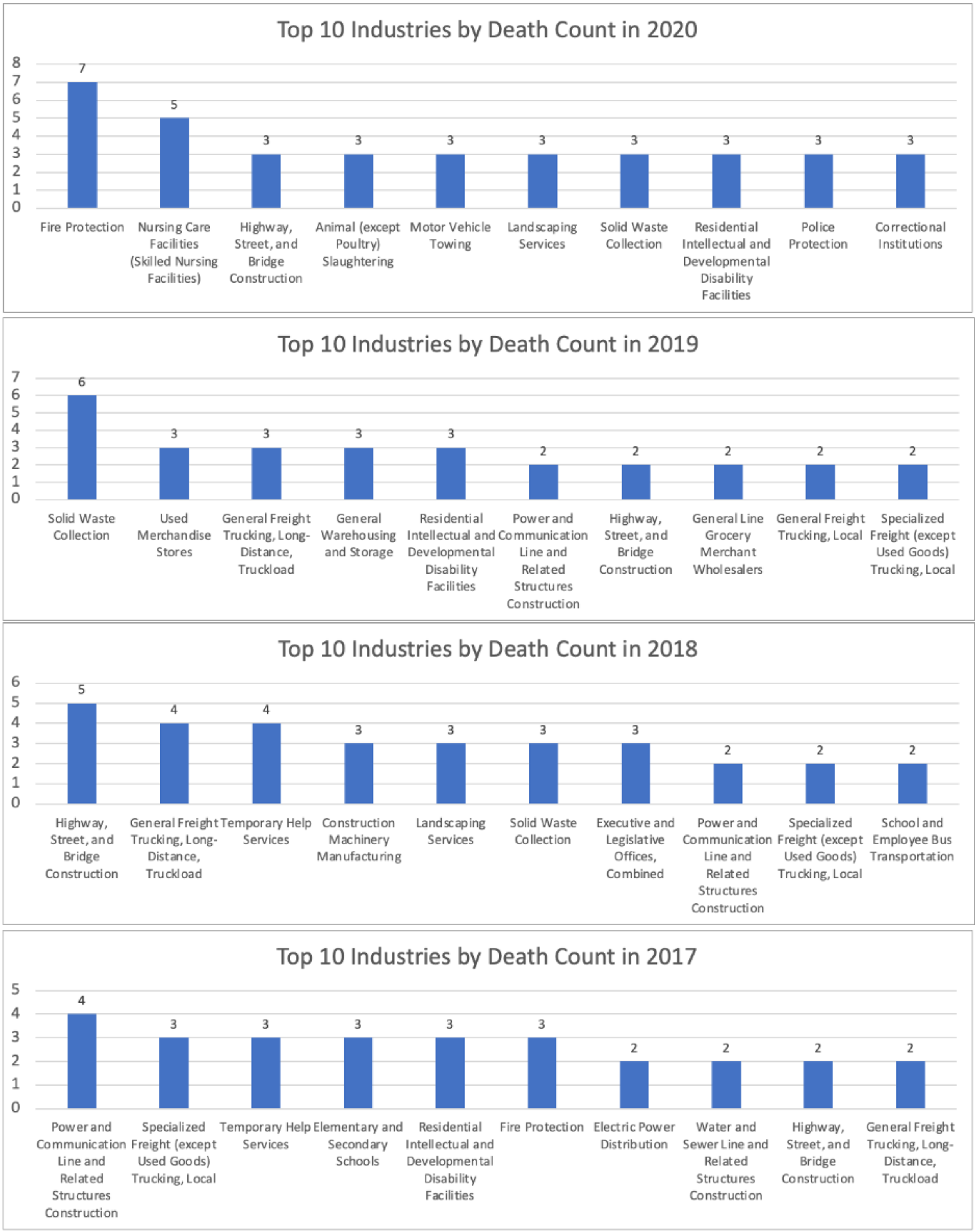


Figure 3: Top 10 Industries by Death Count from 2017 to 2020

5.3 Different Injury Cause Types in Different Industries

We also investigated in detail about different injury causes of the top 5 industries that have the most claims in 2020. We found out that some industries such as the General Medical and Surgical Hospitals Industry and Nursing Care Facilities (Skilled Nursing Facilities) Industry are significantly affected by COVID-19. Some industries such as Home Centers Industry have a quite obvious seasonal fluctuation due to these industries' market demand fluctuation. Some industries have dominant injury cause types. For example, in the General Warehousing and Storage Industry, more than 50% of injuries are caused by Strain or Injured By within which more than 45% are directly caused by Lifting. However, not all industries have features to conclude. For instance, the Supermarket and Other Grocery Stores Industry is neither affected by the market demand's seasonal variation nor by the pandemic.

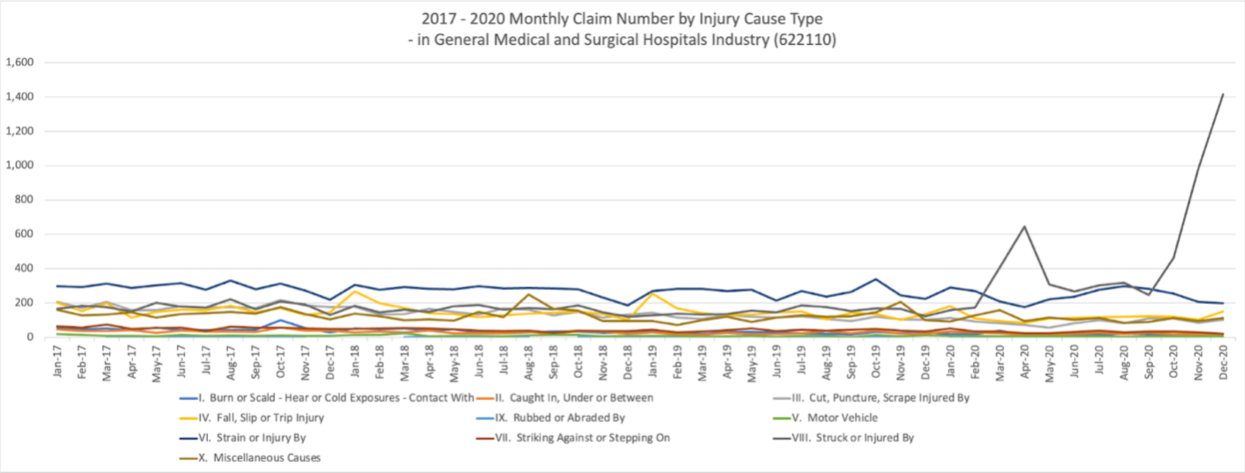


Figure 4: 2017 - 2020 Monthly Claim Number by Injury Cause Type in General Medical and Surgical Hospitals Industry

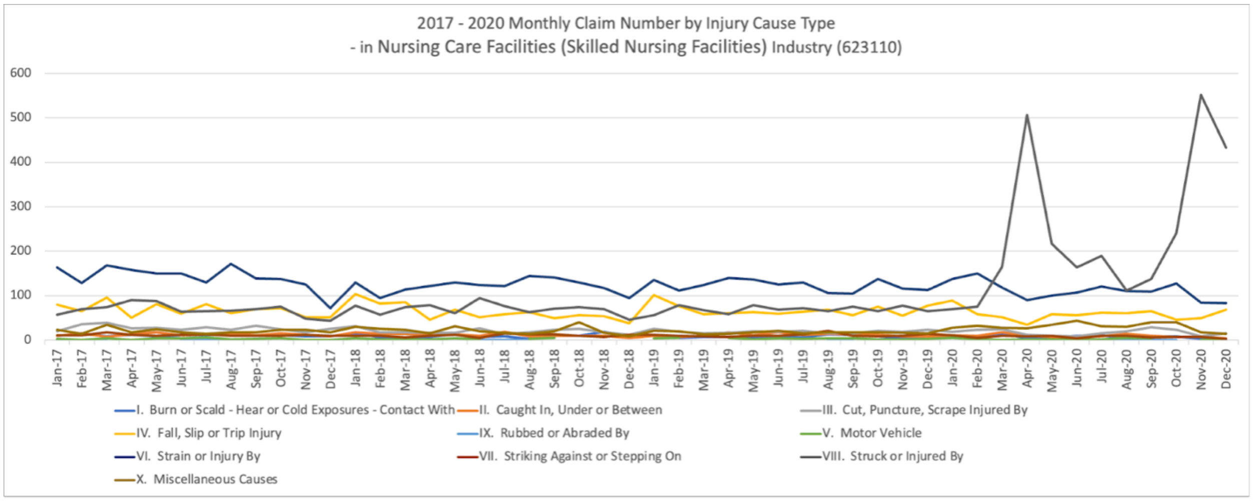


Figure 5: 2017 - 2020 Monthly Claim Number by Injury Cause Type in Nursing Care Facilities (Skilled Nursing Facilities) Industry

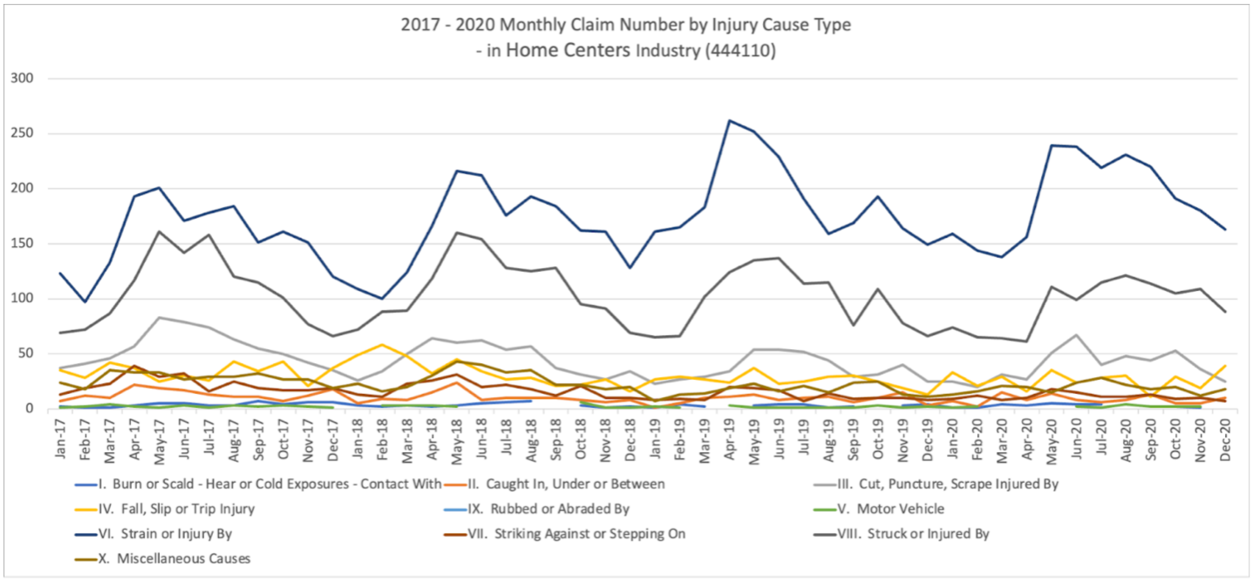


Figure 6: 2017 - 2020 Monthly Claim Number by Injury Cause Type in Home Centers Industry

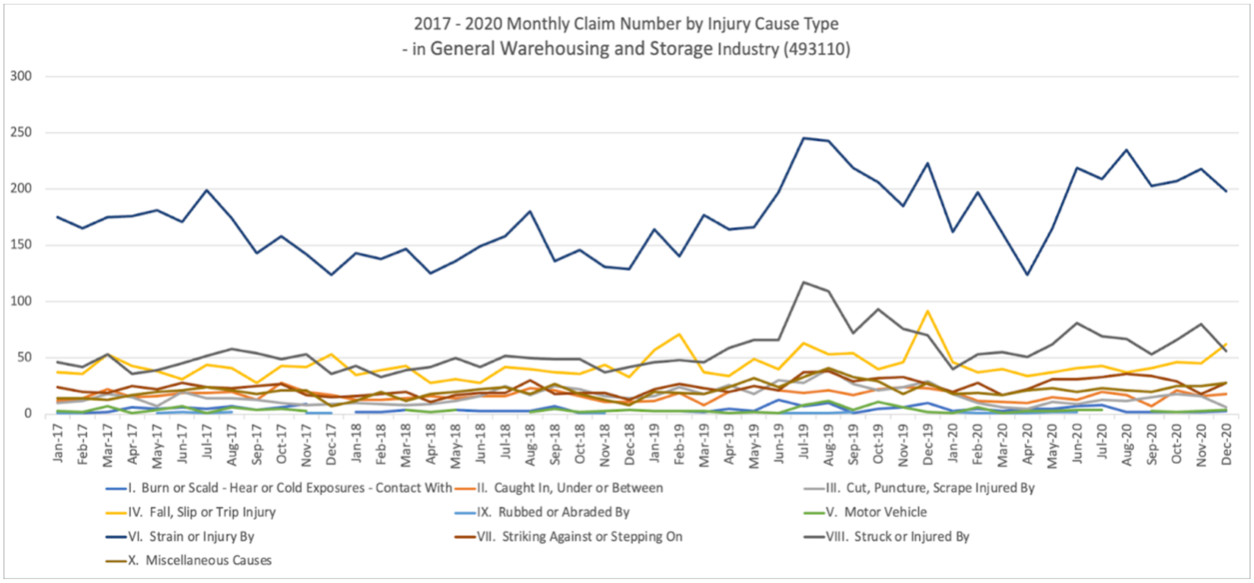


Figure 7: 2017 - 2020 Monthly Claim Number by Injury Cause Type in General Warehousing and Storage Industry

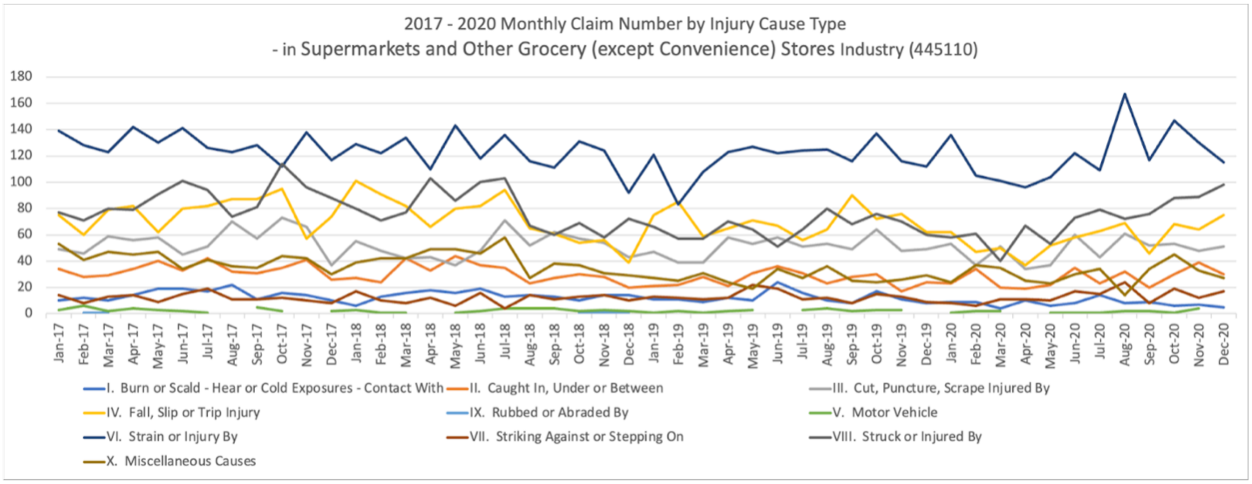


Figure 8: 2017 - 2020 Monthly Claim Number by Injury Cause Type in Supermarkets and Other Grocery (except Convenience) Store Industry

6 Neural Network Model

6.1 Features

We want to benefit this project by predicting something else other than time series predictions, so we build a neural network model to predict most probable injury natures based on some basic personal information.

This neural network takes NAICS code, employer state code, employee static code, gender, and age as the inputs, and it will output the probabilities that each injury will happen in the future. However, it is not as easy as it looks. We have 4 categorical features and one continuous feature. When we combine them together, we get heterogeneous inputs, which is a little bit difficult for just fully-connected neural networks to learn. So we have to take advantage of other kinds of structure of the neural network. But we first tried the fully-connected neural network, which is the easiest one, although we know it will not yield an ideal result, we can find where it goes wrong and try to improve it.

6.2 Fully-connected Neural Network

Here is the structure of our first try, the fully-connected neural network. Of course we can not just put the raw categorical features into the network. We have to encode it into another format for the network to learn.

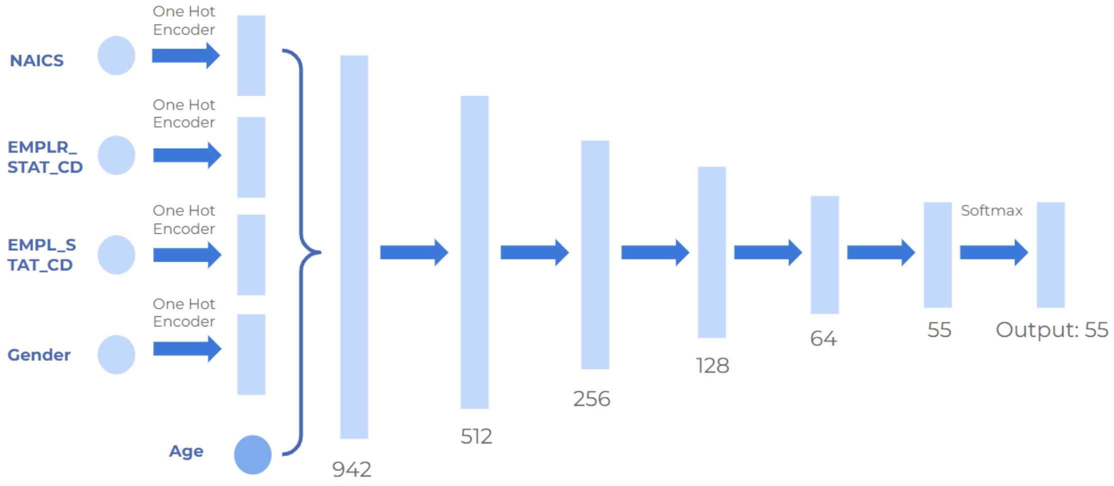


Figure 9: Fully-connected Neural Network

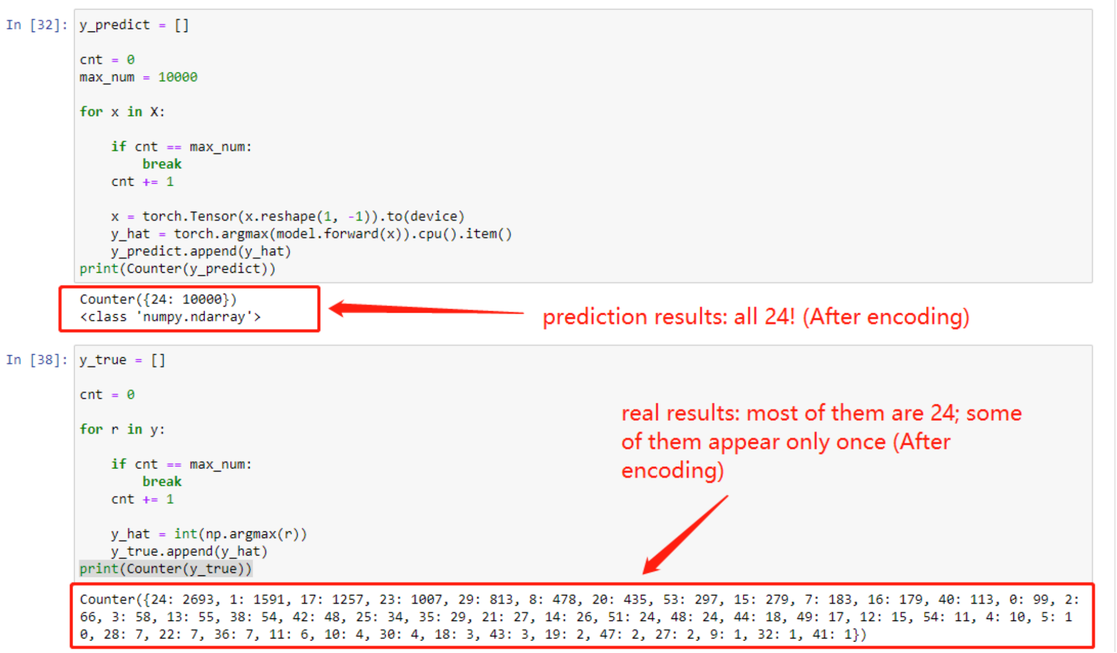
Usually there are two options to encode them: One-Hot and Ordinal. For an ordinal encoder, it will just encode them as some ordinal number like 1, 2, 3, 4. It is a little weird to do that because they will become comparable after encoding. For example, there will be PA and CA for employer state code. They are just location information and not comparable. However, they can become such as 2 and 5 after ordinal encoding. There will be a relationship that 5 is greater than 2, but it doesn't mean CA is greater than PA. So ordinal encoding will add some extra information which is unnecessary or even noise to the network.

On the other hand, one-hot encoding does not have such drawbacks. After one-hot encoding, categorical features will become column vectors with consistent shape. The length of the vectors is the number of values of these features. All the elements in this vector will be 0 except for the element with corresponding index to the categorical feature, and the value of this element will be one. For example, in ordinal encoding, PA and CA might become 2 and 5, and in one-hot encoding, they will become 010000 and 000010. It can be learnt by the network with as little noise involved as possible.

Here we used one-hot for the encoding. We concatenate the encoded categorical features and continuous features altogether as the input, and it will go through lost of linear layers, or we say, fully-connected layers. Finally, it will go through a softmax layer that makes it probabilities. For this network, the output will have 55 neurons, each of them are just the probabilities of each kind of injury natures because there are 55 kinds of injury natures. The sum of them will be one. This network surely learnt something but the result is not ideal.

6.3 Problems

This figure shows the difference of our output and the real injury natures. We can see that in the real result, there will be 55 injury natures and number 24 is the majority. For our output, we only have number 24. We think there are 2 reasons that can yield this problem.



```
In [32]: y_predict = []

         cnt = 0
         max_num = 10000

         for x in X:

             if cnt == max_num:
                 break
             cnt += 1

             x = torch.Tensor(x.reshape(1, -1)).to(device)
             y_hat = torch.argmax(model.forward(x)).cpu().item()
             y_predict.append(y_hat)
         print(Counter(y_predict))

         Counter({24: 10000})
         <class 'numpy.ndarray'>
```

prediction results: all 24! (After encoding)

```
In [38]: y_true = []

         cnt = 0

         for r in y:

             if cnt == max_num:
                 break
             cnt += 1

             y_hat = int(np.argmax(r))
             y_true.append(y_hat)
         print(Counter(y_true))

         Counter({24: 2693, 1: 1591, 17: 1257, 23: 1007, 29: 813, 8: 478, 20: 435, 53: 297, 15: 279, 7: 183, 16: 179, 40: 113, 0: 99, 2:
         66, 3: 58, 13: 55, 38: 54, 42: 48, 25: 34, 35: 29, 21: 27, 14: 26, 51: 24, 48: 24, 44: 18, 49: 17, 12: 15, 54: 11, 4: 10, 5: 1
         0, 28: 7, 22: 7, 36: 7, 11: 6, 10: 4, 30: 4, 18: 3, 43: 3, 19: 2, 47: 2, 27: 2, 9: 1, 32: 1, 41: 1})
```

real results: most of them are 24; some of them appear only once (After encoding)

Figure 10: Prediction result (Top) and Real result (Bottom)

First is that the range of the value of the encoded categorical features are just 0 and 1, but we still have a continuous feature, the age, which has the range from 10 to100. We are not using normalization so that the value of the continuous feature will be much greater than the value of one-hot vectors. Which will increase the bias of the model.

Second is that the dataset is skewed. Some categories are the majority and most categories are minority, which could also increase the bias. Therefore, we need to find some other ways to improve the very first model.

6.4 Embedding Layers

After searching for some references, we started working on embedding layers. What is the embedding layer? It is initially designed for natural language processing. It will convert words into vectors. Moreover, in the new vector space, some relationships will also be maintained. For example, after encoded by the embedding layer, king - queen will be similar to man - women. This is a very good attribute and some papers talk about using embedding layers to encode categorical features is plausible. The method is simply adding embedding layers after the one-hot encoder.

Another advantage of the embedding layer is that it is part of the neural networks so we don't need to train it before we train the network. It will be trained during the training process.

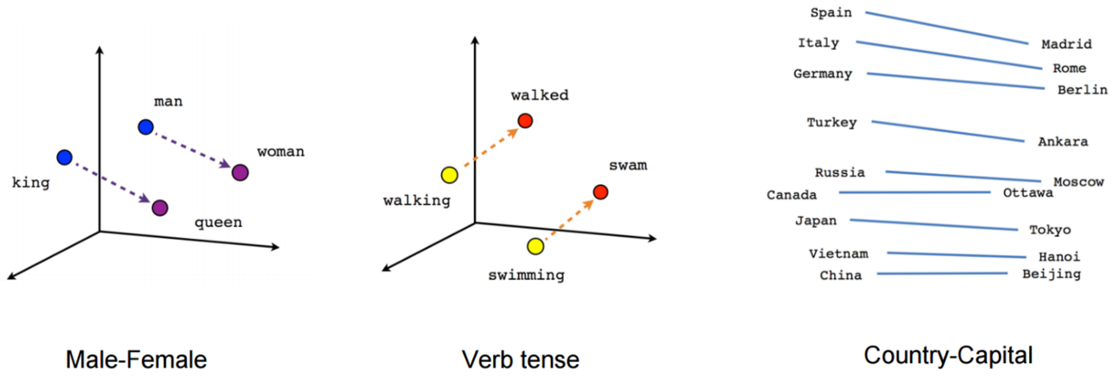


Figure 11: Embedding Layer

6.5 Neural Network with Embedding Layers

Here is the improved network. We added embedding layers to represent categorical features better. Also we add some batch normalization layers to restrict the range of the values and dropout layers to prevent overfitting. The result is much better than before.

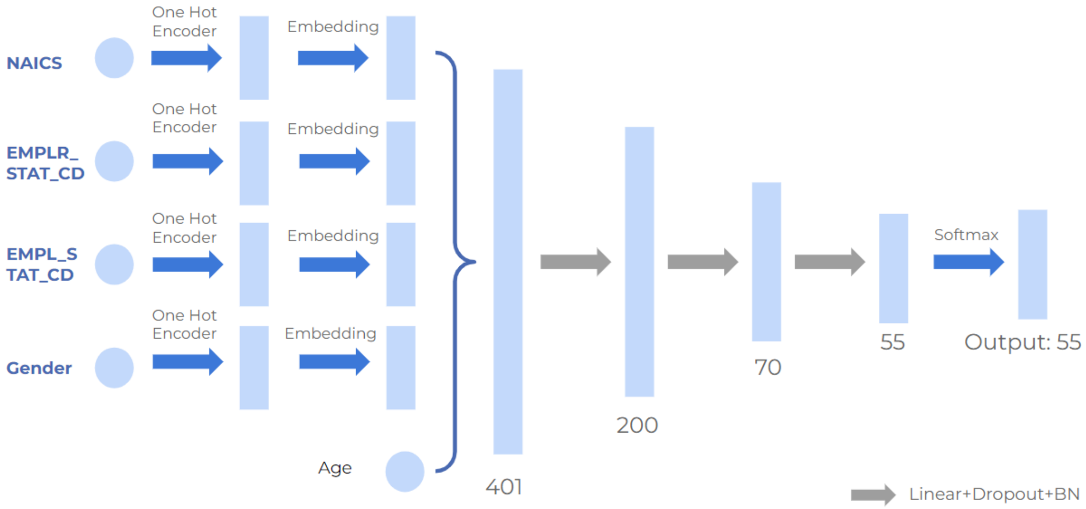


Figure 12: Neural Network with Embedding Layer

6.6 Result

As we can see, the model can predict all kinds of injury natures and the distribution is quite similar to the real distribution of injury natures. If we only look at whether the top 1 nature fits the real value, the accuracy will be just about 27%. If we look at whether the top 3 natures include the real value, the accuracy will be 56% and for top 5 natures, the accuracy will be 72%. So we recommend using this model to predict the top 5 injury natures that the employee might face in the future.

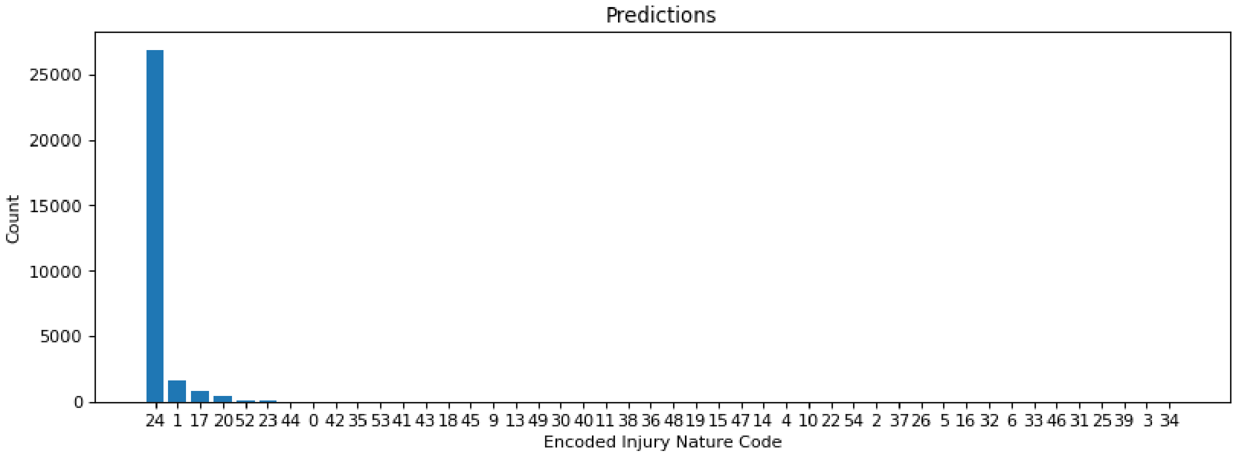


Figure 13: Prediction Result

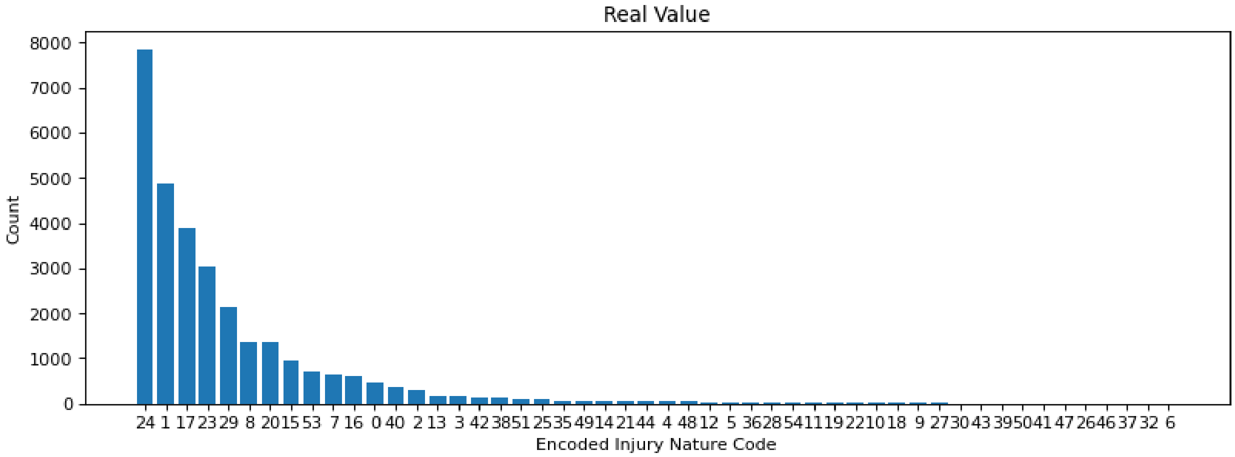


Figure 14: Real Distribution

7 Model Evaluation

7.1 Introduction

Previous project provided the solution to predict injury rate with two time-series models: Prophet and SARIMAX. The previous workflow is: 1) users select their prediction criteria, including NAICS code, start date, end date, and the prediction model; 2) then, users need to submit the prediction task; 3) finally, the model would predict the average injury rate of the next year. However, this workflow does not consider prediction reliability and future trend analysis. In this project, we would continue to work on the time-series model and address these problems.

7.2 Problem Statement

In our prediction workflow, prediction results depend on the user input, including NAICS code, start date, end date, and the prediction model. We can conclude several major factors that would potentially cause us to fail our prediction: 1) The selected NAICS or the selected time range includes too few data points, so the model will predict the result using a lot of zero rate instead of a valid value rate. 2) The injury rate by itself might not follow any time series cycle, so it is not reasonable enough to use a time-series model. Because the prediction model has the potential to fail, result evaluation and model diagnostics are important for both models in our case; however, they are missing in the previous project. In this section, we will build an evaluation analysis to answer how our prediction result is reliable, and how much confidence we are sure about our conclusion.

7.3 Solution

To evaluate the prediction result, we conduct the following analysis:

1. Confidence interval analysis
   - Description: Blue points are actual rate; red line is the prediction; Gray area is the confidence interval when $\alpha = 0.01$.
   - Diagnostics: A good estimate requires the actual rate is located as much close to the red line as possible; zero points is bad for the prediction.

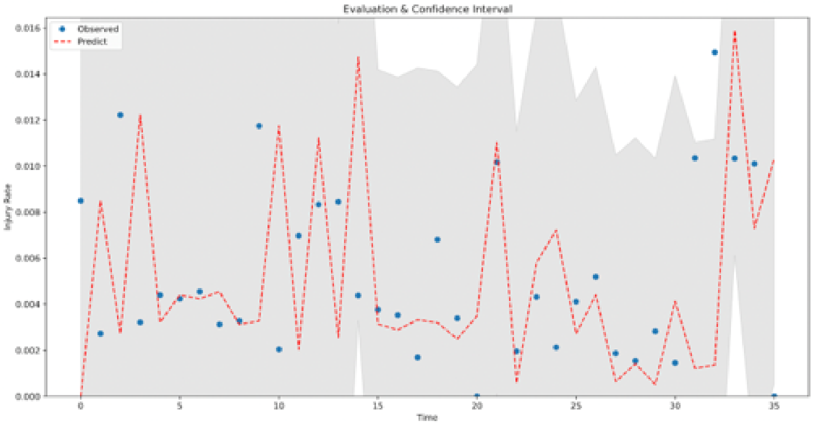


Figure 15: Confidence Interval Analysis Example

2. 5-year prediction trend

    - Description: Blue line is the prediction over the selected time range, while red line is the next 5-year prediction; Gray area is the confidence interval when $\alpha = 0.01$.
    - Diagnostics: A good estimate requires the confidence interval (gray area) to be as narrow as possible when the prediction is more reliable.

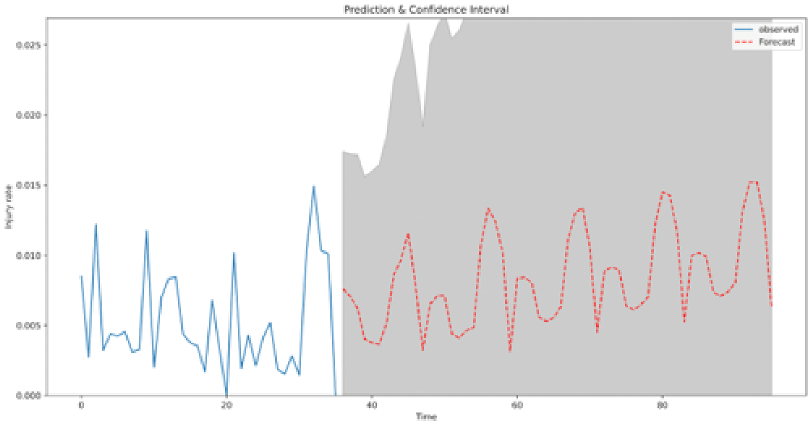


Figure 16: 5-year Prediction Example

3. Residual analysis (SARIMAX only)

    - Description: Standard plot of residual; histogram and estimated density; Normal Q-Q plot; Correlogram
    - Diagnostics: A good estimate requires the residual follows the normal distribution with a mean of zero. So, we would like to see the residual plot close the zero, and the density plot fit with the normal distribution. In the normal Q-Q graph, we would like to see blue points fit close to the red line.

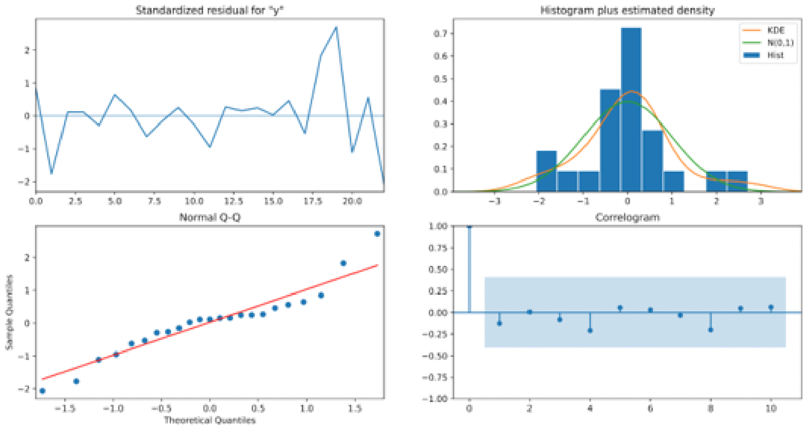


Figure 17: Residual Analysis Example

4. RMSE (or MAPE) analysis (Prophet only)

- Description: A plot of RMSE (or MAPE) and the moving average trend line.
- Diagnostics: A good estimate requires the RMSE to be ideally close to zero and stable.

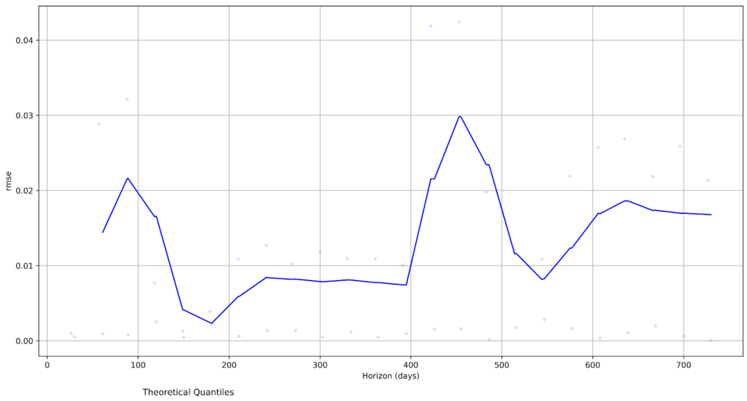


Figure 18: RMSE Plot Example

7.4 Recommendation

Result reliability is very important to consider when we draw any conclusion based on our prediction. Improving prediction reliability is very desired, and we would like to provide some recommendations that would help improve prediction result reliability.

1. Use high level NAICS code. In our case, higher level NACS code includes more aggregated data, so they provide more evidence to the prediction.

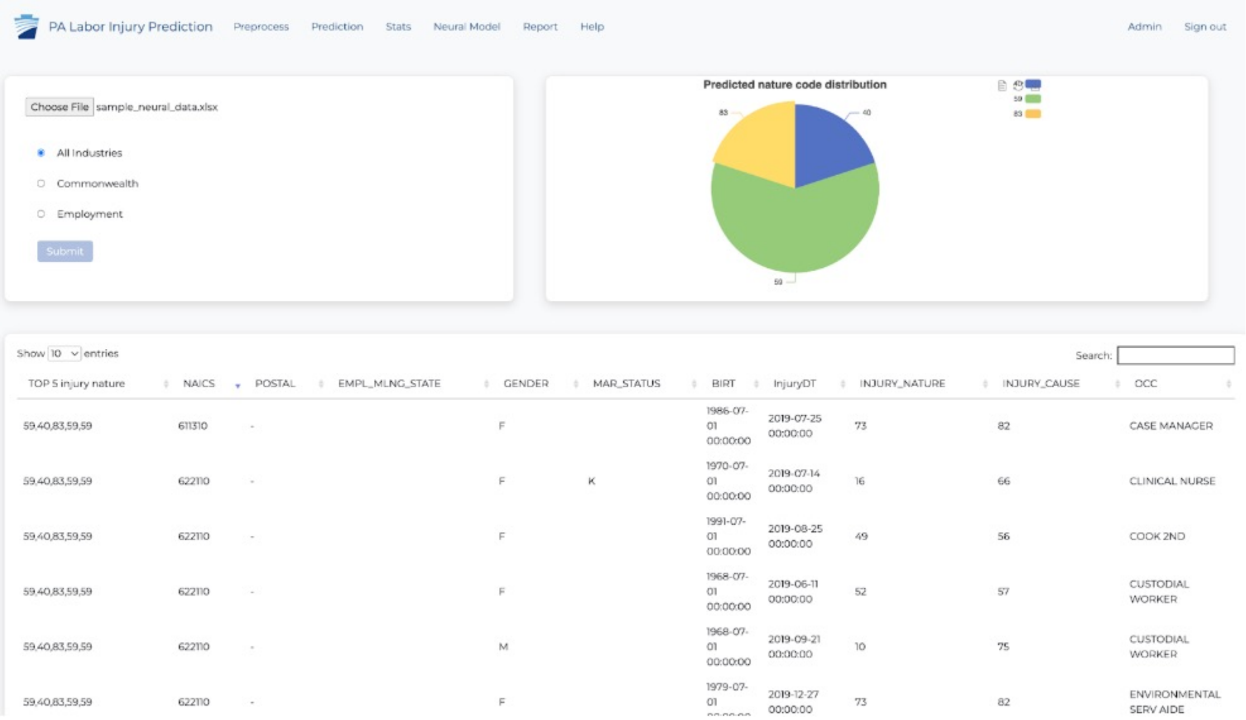
2. Select to include at least 3-year data. In our case, the time series cycle of the injury rate is year to year. So, 3 years of data is at least recommended to select.

3. Read the evaluation report together with the Power BI prediction. In our case, prediction quality depends on each selection of NAICS code, start date, end date, and the model. Result quality varies case by case, and no single model could work effectively on all cases. So, it is recommended to read the evaluation whenever the selected input changes. It would guide you to understand the result reliability before drawing any conclusion.

8 Web Application Enhancement

8.1 Neural Model Prediction

We integrated the neural network model into the current software. As we could see in the following figure, there is a extra tab called neural model. Users are allowed to upload the data in the system. Once they click on the submit botton, a detailed predictive results will be displayed in the following table. The first column is the predicted injury nature. A predicted nature code distribution is shown in the up right pie chart.



Choose File sample_neural_data.xlsx

● All Industries

○ Commonwealth

○ Employment

Show 10 entries

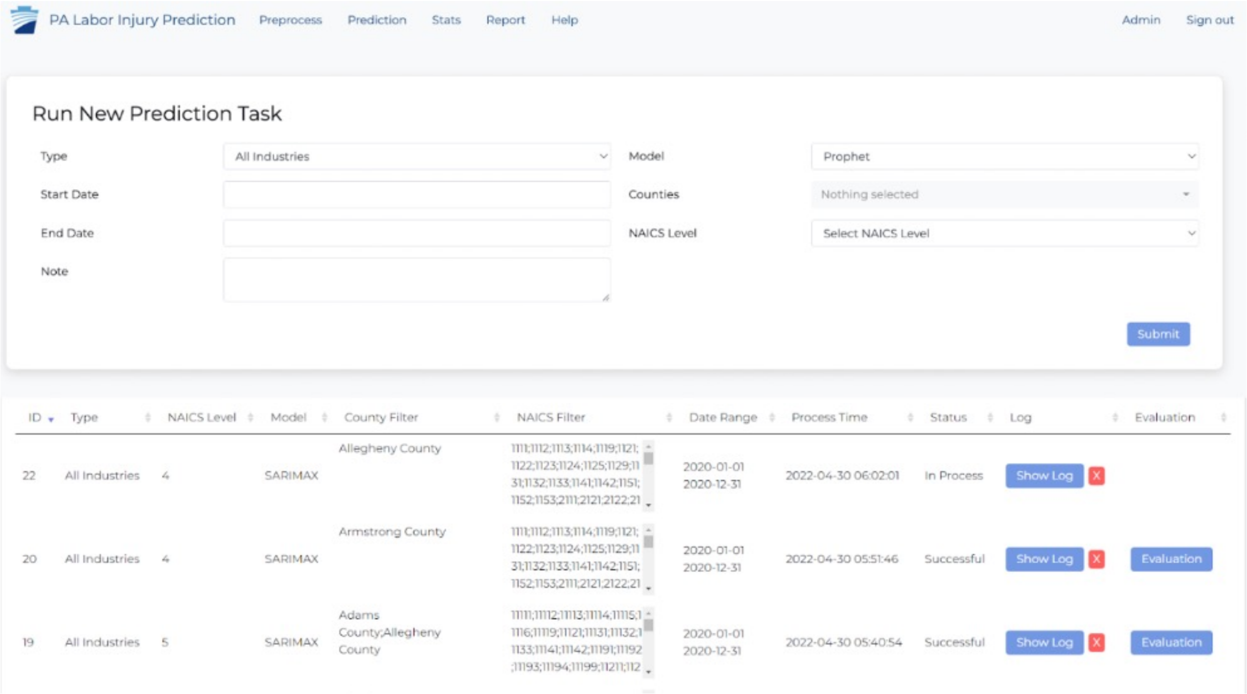
Search:

TOP 5 injury nature	NAICS	POSTAL	EMPL_MLNG_STATE	GENDER	MAR_STATUS	BIRT	InjuryDT	INJURY_NATURE	INJURY_CAUSE	OCC
59,40,83,59,59	611310	-		F		1986-07-01 00:00:00	2019-07-25 00:00:00	73	82	CASE MANAGER
59,40,83,59,59	622110	-		F	K	1970-07-01 00:00:00	2019-07-14 00:00:00	16	66	CLINICAL NURSE
59,40,83,59,59	622110	-		F		1991-07-01 00:00:00	2019-08-25 00:00:00	49	56	COOK 2ND
59,40,83,59,59	622110	-		F		1968-07-01 00:00:00	2019-06-11 00:00:00	52	57	CUSTODIAL WORKER
59,40,83,59,59	622110	-		M		1968-07-01 00:00:00	2019-09-21 00:00:00	10	75	CUSTODIAL WORKER
59,40,83,59,59	622110	-		F		1979-07-01	2019-12-27 00:00:00	73	82	ENVIRONMENTAL SERV AIDE

Figure 19: Neural Network Webpage

8.2 Evaluation Method

Previously, we used SARIMAX and Prophet to predict 1-year average injury rate of a selected NAICS code. To strengthen the reliability of the prediction result, we included a residual analysis and a confidence interval analysis. On the prediction page of the web application, we created a new view button on the right side of each prediction task, so that users can access the new prediction reports. We will provide a guidance document to help users to understand the new reports.



Run New Prediction Task

Type	All Industries	Model	Prophet
Start Date		Counties	Nothing selected
End Date		NAICS Level	Select NAICS Level
Note			

ID	Type	NAICS Level	Model	County Filter	NAICS Filter	Date Range	Process Time	Status	Log	Evaluation
22	All Industries	4	SARIMAX	Allegheny County	1111;1112;1113;1114;1119;1121;1122;1123;1124;1125;1129;1131;1132;1133;1141;1142;1151;1152;1153;2111;2121;2122;21	2020-01-01 2020-12-31	2022-04-30 06:02:01	In Process	Show Log X	
20	All Industries	4	SARIMAX	Armstrong County	1111;1112;1113;1114;1119;1121;1122;1123;1124;1125;1129;1131;1132;1133;1141;1142;1151;1152;1153;2111;2121;2122;21	2020-01-01 2020-12-31	2022-04-30 05:51:46	Successful	Show Log X	Evaluation
19	All Industries	5	SARIMAX	Adams County;Allegheny County	11111;11112;11113;11114;11115;11116;11119;11121;11131;11132;11133;11141;11142;11191;11192;11193;11194;11199;11211;112	2020-01-01 2020-12-31	2022-04-30 05:40:54	Successful	Show Log X	Evaluation

Figure 20: Evaluation Webpage Enhancement

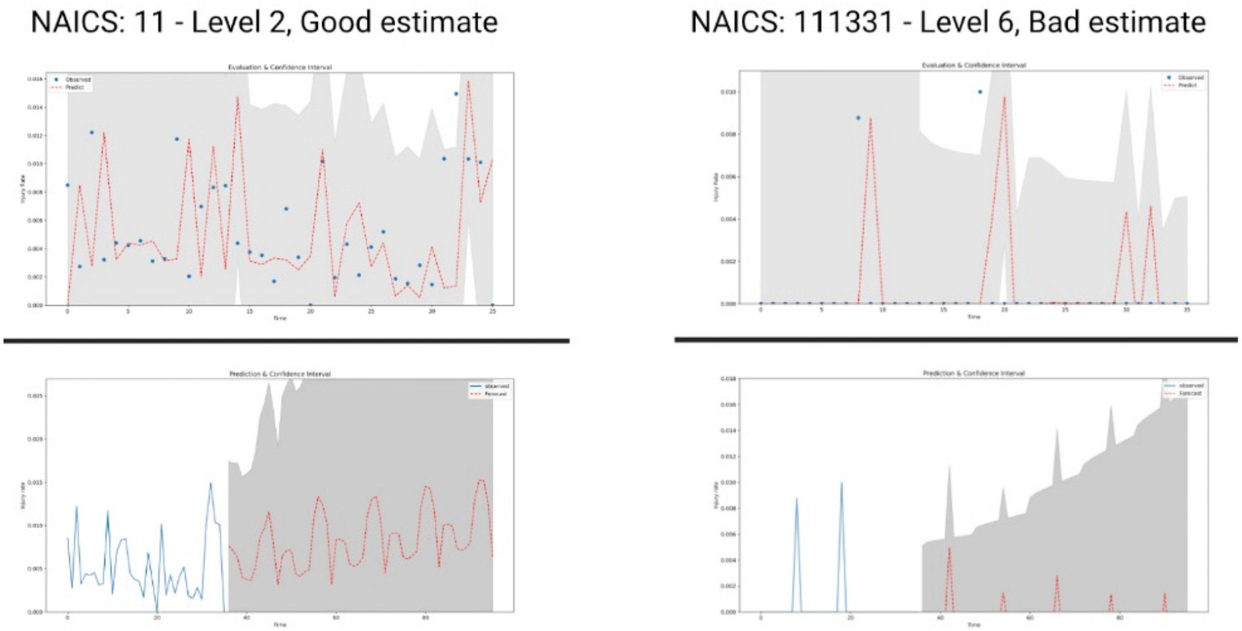
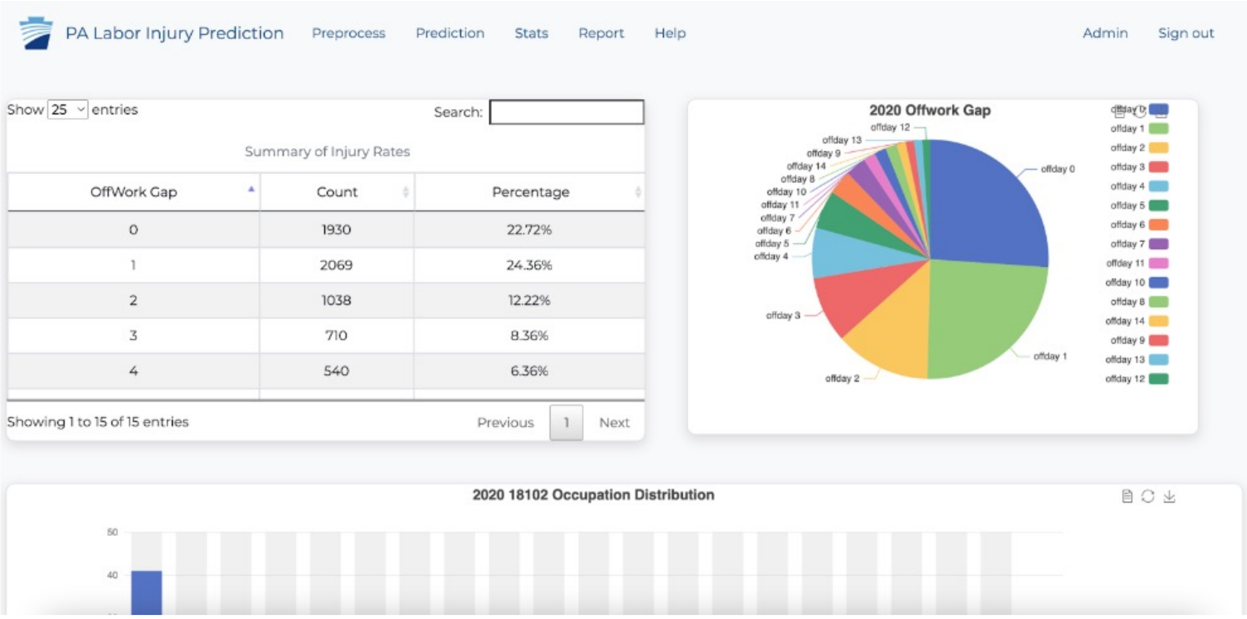


Figure 21: Evaluation Report

8.3 Statistics Dashboard

Also, we provided a dashboard to illustrate the statistics. For example, the summary of the injury rate will be displayed in the table. Also, the offwork gap pie chart is provided in the up right chart.



Summary of Injury Rates

OffWork Gap	Count	Percentage
0	1930	22.72%
1	2069	24.36%
2	1038	12.22%
3	710	8.36%
4	540	6.36%

Showing 1 to 15 of 15 entries

Figure 22: Statistic Dashboard Webpage

9 Conclusion and Recommendation

In conclusion, we work can be segmented into three aspects, including extracting analytical findings to advise the injury mitigation, complementing existing prediction with a new neural network model as well as evaluation report and enhancing the web application with new functionalities mentioned above.

9.1 Analytics

Based on the descriptive analytics, seasonality is an essential factor of injury occurrences for several injury causes, especially the Fall, Slip or Tip injury. Besides, more than 50

Therefore, we would recommend the PA department to treat injury cases relevant to strain and lifting with the highest priority, in order to mitigate the overall injury claim count within the state efficiently. Given those two types of injury cases are preventable, the government could take preventive actions proactively. For example, the government could examine the working environment safety regularly, especially from January to March. Additionally, employers could host safety seminars to develop the employee's safety awareness to avoid injuries at work as well their knowledge to react to injuries appropriately.

As to medical care trainers, the government could allocate trainers based on the popularity of injury causes and reserve more medical supplies for frequent injuries, such as bandages, alcohol prep pad and so on.

9.2 Prediction

The new neural network produces the top five likely injury natures for each employee based on their age, gender, location and industry. Employers or the government are encouraged to feed the basic information of labor workforces into our model and envision the most likely injury types. Therefore, they could prevent those types of injuries accordingly, such as placing security alerts at work places, arranging safety training or adjusting budgets relevant to those injuries in advance.

9.3 Web Application

The web application now is equipped with two types of prediction models to cater to various user needs. The time-series models, targeting macro-level prediction, could facilitate users to picture the injury rate trend in the foreseeable future within specific counties. The new neural network focuses on a micro level for each employee's individual cases.

Besides, the generated evaluation report provides insights on the validity of for each time-series prediction run. In this way, the users are able to interpret the prediction results and thus use them with a certain level of confidence.

10 Lesson Learned

10.1 Business

Though weekly meeting with clients, we understood how the business in the real company is working. Information technologies are playing a more and more important role in the traditional industries. Through this project, we learnt in practice about the business working pipeline and information technologies power.

10.2 Technological Enhancements

Through this project, we gained lessons on how data could be utilized to guide the development of business.For example, we applied neural networks into the real data application, and gained good results in terms of the prediction accuracy. Also, we relied on python analytics techniques to clean, visualize and analyze the data and gained valuable insight underlining the business data. Finally, web application is becoming a natural demonstration method for us when we try to make the system easy to use.

10.3 Team Collaboration

Through whole semester's collaboration, we, as a team, learnt to how to contribute what we are good at into the team's efforts. For example, Lanxuan is familiar with machine learning algorithms and machine learning applications, he is working on the neural network model training and inference part. Shanyue is good at system development and he applied this skills when the team needs to integrate the neural network model into the web application.

11 Suggestions for Future Work

11.1 Augment the PowerBI dashboard

The current PowerBI dashboard only presents injury rate prediction results under three dimensions, county, industry and time. To enable users to visualize the injury claim data under other granualities, such as death cases, injury causes and injury natures, the future team could incorporate graphs into the powerBI dashboard.

11.2 Consolidate Data Cleaning Logics

We would assume that each semester, teams will preprocess the data to a certain extent during the data exploration and analytics process. Therefore, we would suggest the client work with future teams to consolidate the data cleaning logics, such as how to define invalid values for each future, how to deal with empty values and so on. This will improve the consistency of analytics and input data for prediction across different project phases.

11.3 Document data source and FAQs

Similarly, if future teams could keep collaborating with the client to improve the documentation for the data source, such as the data dictionary and record the FAQs into a document, this will save time for clients and support future teams to get familiar with the resources efficiently.

11.4 Acquire medical cost rankings

Due to the unavailability of medical costs data from a third party organization, we utilized the average hourly rate to predict the potential savings from injury cases. If data privacy is still a concern, the client could consider requesting aggregated and sanitized data or just the rankings for key dimensions. For instance, the rankings of average medical cost per case of each injury cause or nature would be insightful enough for future teams to quantify the severity and priority of injury types financially.

11.5 Optimize neural network model

Last but not least, future teams could consider optimizing the neural network model by modifying the dropout layer to improve the prediction stability from the neural network model.