CIS 3200-01 Flu Analytics Term Paper 12/16/2019

Mario Arguello Jr.
Michael Brumwell
Brianne Phillips
Christian Rodriguez
California State University, Los Angeles

Abstract

This document goes into detail about how we used Clinical Vaccination Data to identify vulnerable personnel and problems that can occur as a result of not being vaccinated. We used both ElasticSearch with Kibana (ELK) Kit and Microsoft Azure Machine Learning Studio to import our data set retrieved from the California Department of Public Health. We found important insights of our research through data analysis in visualizations and linear regression. There is a possibility based off of our results and predictions that many more employees will receive the flu shot in 2020 due to the exposure that these employees face in their workplace by patients that are admitted with the flu. With this being a domino effect, it is crucial that the personnel receive the vaccine.

I. Introduction

We based our project on vaccination data with hospital employees to emphasize the risks associated with not getting vaccinated. Specifically, unvaccinated hospital workers due to their high risk of getting infected, but also because the infected personnel pose a risk to spreading it to vulnerable patients, such as the elderly and children. It is important to make sure all personnel get vaccinated to reduce the potential risk of contracting the flu as much as possible. Reducing the risk with employee vaccination is on-track to hit 90%, which is represented in our data and which will be discussed further on.

II. Body

For our experiment, we decided to focus on the State of California because it will affect the authors of the paper located in CA. We also chose to specifically look at the largest counties since there is more population, hence a higher possibility of the flu spreading amongst people. With viewing the larger counties with the most amount of hospitals, we wanted to view how many of these hospitals had personnel vaccinated with the flu shot. Again, it is crucial for these employees to have the vaccine as it decreases the chance for everyone they come in contact with to accidentally attract the virus. Once we saw the first bar chart, coincidentally our own county of Los Angeles was the largest with a high amount of clinics. We were also able to identify six other larger surrounding counties such as Riverside, Orange, etc., that also had a high amount of reported vaccinated personnel. With these results, we then chose to create a pie chart that would visualize the amount of counties with clinics that had the lowest percentage of vaccinated personnel.

These results can truly be an eye opener for some while visualizing the amount of clinics in these huge counties. The potential for millions of transfers of the flu virus are shown in areas with unvaccinated personnel. The flu is easily transferred from person to person or from object to person. A simple cough or sneeze in the area you're in or even touching the same items (then touching your nose or mouth) as a person whom is infected by the virus. With this being said, if a doctor were to see one patient who is contaminated, and then see more patients, the likelihood of these patients receiving the flu is very high. These patients are now exposed to the bed, materials, room, air, and doctor that the flu virus was around.

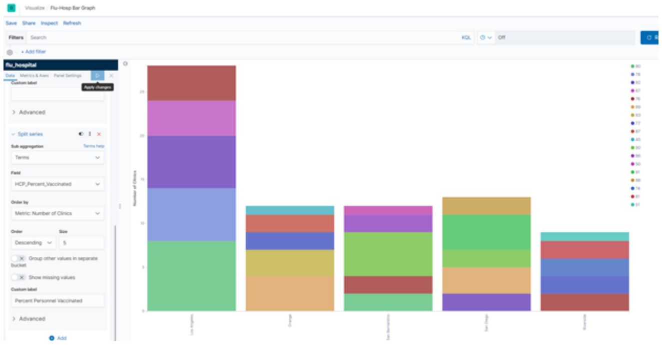
On our third chart, we wanted to start the analysis of hospitals that are on track to have at 90% of their personnel vaccinated by the year 2020. The reason for this is that we want to view the dedication that the hospitals are making to give patients peace of mind when visiting the clinics. Just as some hospitals require people with the cough or cold to put on a mask, it is imperative to have their staff be vaccinated to reduce the spread of the flu. Not only to reduce the chance of spreading the flu, but it does benefit the employees as they are avoiding the additional risks of giving a potentially deadly virus to others, including their own families. On our chart, we separated three bar charts using DNR (does not report), YES, or NO. There were some results with DNR, a great amount in the NO category, with Los Angeles being the worst county, and then we received a nice amount of clinics report being on track.

With all these charts in mind, we lastly decided it would be beneficial for the public to view these results on a map in order to identify the best regions of California that are safer or on-track to having less exposure to the flu in the hospitals. In excel, we produced a color coded map showing the best locations in a shaded color, with the more damaging areas in the light colored areas. However, do not be alarmed with the white sections, keep in mind some clinics did not report to the State.

2.1 Kibana and ElasticSearch

Kibana is one of two visualization tools we used to create graphs for our data. By uploading our data to ElasticSearch, which is primarily responsible for indexing imported data, we can create different types of charts to suit our needs. To discover different insights, we made three graphs and a dashboard to neatly organize them.

Our first graph was a bar graph we created for the purpose of identifying the counties with the highest number of clinics. From those clinics, we then separated the data in order to look at the percentage of vaccinated personnel in those clinics. For cleanliness, we only did the top five clinics with the highest percentages, but altering the size in Kibana will give you more clinics.



Visualize / Flu-Hosp Bar Graph

flu_hospital

Split series

Sub aggregation: Terms

Field: HCP_Percent_Vaccinated

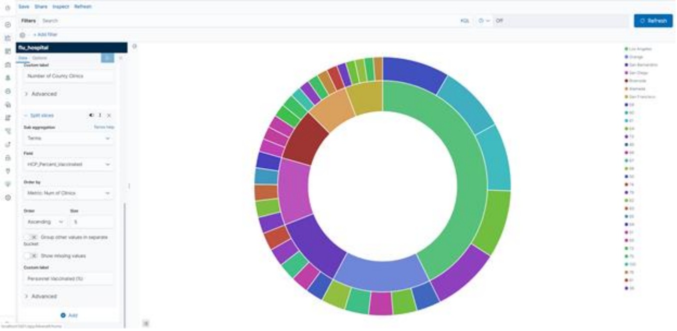
Order by: Metric: Number of Clinics

Order: Descending Size: 5

Custom label: Percent Personnel Vaccinated

Figure 1. Kibana bar graph of clinics.

The second graph is a pie graph we created as a way to easily check the clinics with the lowest percentages of vaccinated personnel. These clinics pose the most risk to vulnerable persons both inside and outside of a clinic, so it's these clinics that will need the most improvements. As with the previous graph, the data size is limited for cleanliness but can be adjusted to have a wider array of data available.



flu_hospital

Custom label: Number of County Clinics

Split slices

Sub aggregation: Terms

Field: HCP_Percent_Vaccinated

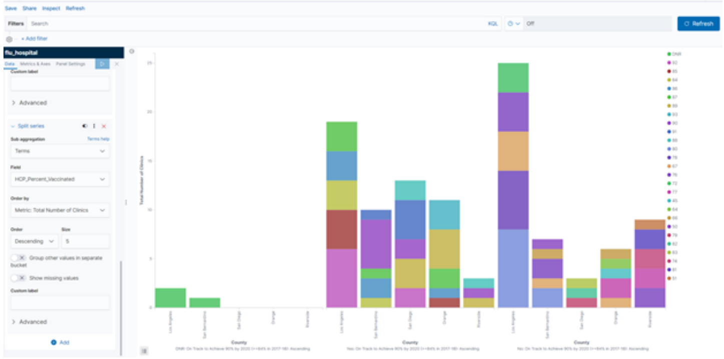
Order by: Metric: Num of Clinics

Order: Ascending Size: 5

Custom label: Personnel Vaccinated (%)

Figure 2. Kibana Pie graph of largest clinics with least personnel vaccinated.

Finally, the third graph represents the goal of each clinic to reach 90% vaccinated personnel by 2020. In order to be on track for this goal for 2020, a clinic must have at least 85% of their personnel vaccinated. The following chart is split into three smaller charts to represent which clinics were on-track and which were not. The third graph represents that there is currently no data available for those clinics which can easily be seen in the excel portion of this paper.



flu_hospital

Split series

Sub aggregation: Terms

Field: HCP_Percent_Vaccinated

Order by: Metric: Total Number of Clinics

Order: Descending Size: 5

Figure 3. Kibana Bar chart of clinics to be on track of 90% personnel vaccinated by 2020.

2.2 Excel

While Kibana is able to create geographic maps using geographical coordinates, the data set we are using does not have such data. As a result, we used Excel to create a geographic representation for our data. Excel is able to map our data based on location fields such as state, county, and country. Because of this, we are able to create a geographical map where Kibana would otherwise be unable to.

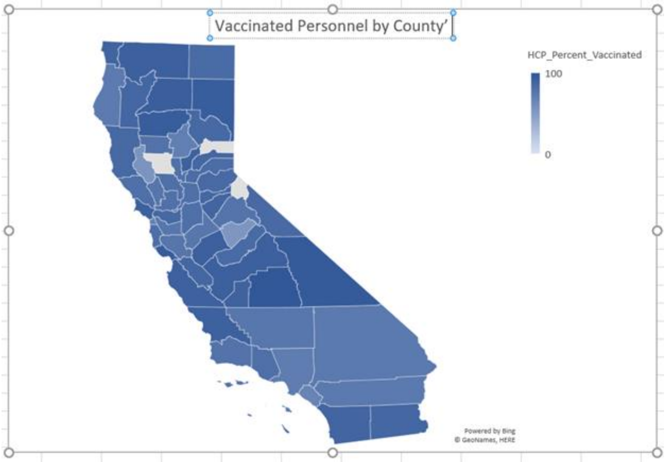


Figure 4. Excel map of vaccinated personnel in California.

2.3 Azure ML studio

During our research we also used Microsoft Azure Machine Learning Studio. We found many insights and strange predictions by using a data science approach called the Bayesian Linear Regression. This data set was retrieved from the flu season of 2017-2018. Hospital in California either reported the percentages of personnel who were vaccinated during this period

We used the train, score, and split data modules in order to receive the final evaluation result from the Bayesian Linear Regression module. The following image is the final result of our research conducted on Microsoft Azure ML Studio.

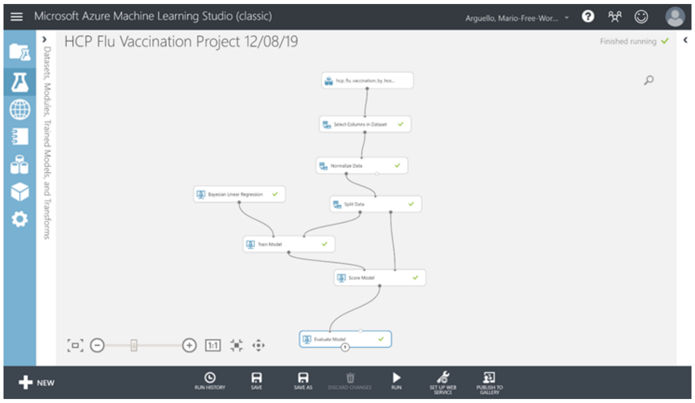

Microsoft Azure Machine Learning Studio (classic)
Arguello, Mario-Free-Wor...
HCP Flu Vaccination Project 12/08/19
Finished running
Datasets, Modules, Trained Models, and Transforms
hcp_flu_vaccination_by_hos...
Select Columns in Dataset
Normalize Data
Bayesian Linear Regression
Split Data
Train Model
Score Model
Evaluate Model
1:1
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

Figure 5. Azure train model of personnel vaccinated using Bayesian Linear Regression.

During our research we saw some negative predictions that personnel may not receive flu vaccinations. According to the visualizations, we discovered there were clinics whose personnel were not getting vaccinated in the Southern California region, specifically in these three counties: Los Angeles, Orange, and San Bernardino.

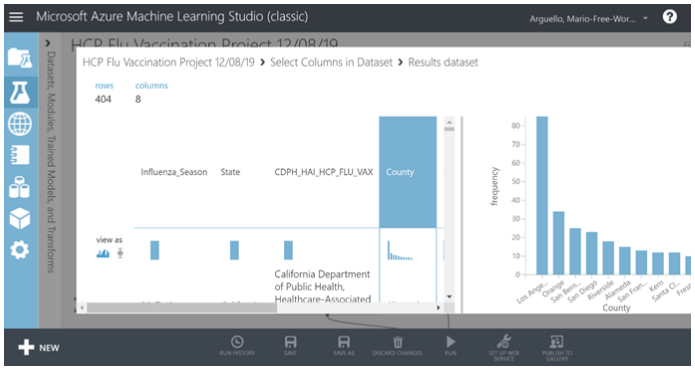
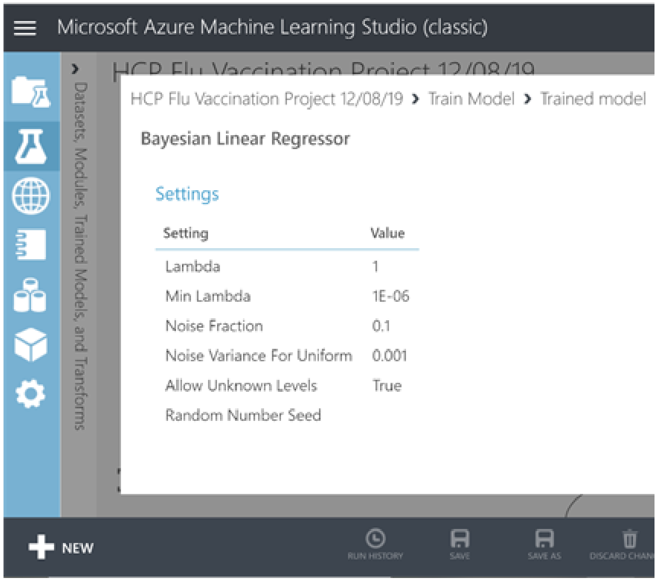

Microsoft Azure Machine Learning Studio (classic)
Arguello, Mario-Free-Wor...
HCP Flu Vaccination Project 12/08/19
Datasets, Modules, Trained Models, and Transforms
HCP Flu Vaccination Project 12/08/19 › Select Columns in Dataset › Results dataset
rows
404
columns
8
Influenza_Season
State
CDPH_HAI_HCP_FLU_VAX
County
view as
California Department of Public Health, Healthcare-Associated
frequency
80
70
60
50
40
30
20
10
0
Los Ange...
Orange
San Bern...
San Diego
Riverside
Alameda
San Fran...
Kern
Santa Cl...
Fresn
County
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

Figure 6. Azure model of 3 main counties not getting vaccinated.

We also observed that the least percentage of personnel getting vaccinated were in the rural counties of Central California. Oddly, Santa Clara county, reported the least personnel getting vaccinated than in any other Bay Area hospital. It appears there may be a lack of hospitals that are located in this county. Perhaps personnel preferred getting vaccinated in other areas and not in Santa Clara county hospitals. We also visualized the predictions of this data set using the Bayesian Linear Regression modules. Due to the underreporting of data for some hospitals using the option of not reporting, we observed that there were incorrect and inaccurate predictions of the personnel who will get vaccinated in California. These hospitals were separated in another category to reflect this DNR status.


Microsoft Azure Machine Learning Studio (classic)
HCP Flu Vaccination Project 12/08/19
Datasets, Modules, Trained Models, and Transforms
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHAN

HCP Flu Vaccination Project 12/08/19 › Train Model › Trained model

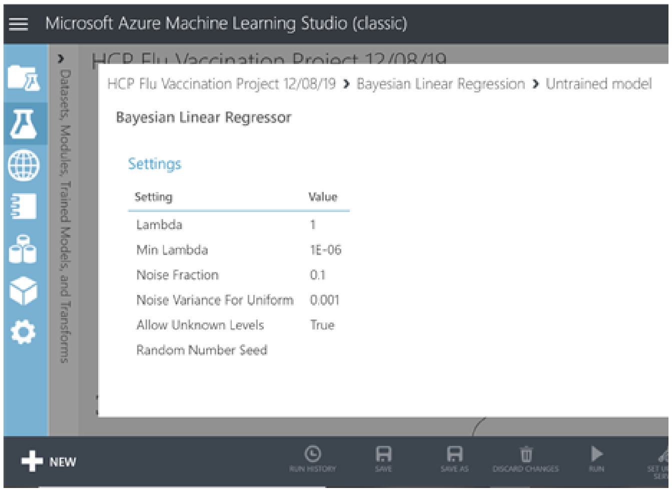
Bayesian Linear Regressor

Settings

Setting	Value
Lambda	1
Min Lambda	1E-06
Noise Fraction	0.1
Noise Variance For Uniform	0.001
Allow Unknown Levels	True
Random Number Seed	

Figure 7. Azure trained model statistics of “Do not Report Status”.

We used a trained model and we compared our prediction using our untrained model, below you will see our results.


Microsoft Azure Machine Learning Studio (classic)
Datasets, Modules, Trained Models, and Transforms
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN

HCP Flu Vaccination Project 12/08/19 › Bayesian Linear Regression › Untrained model

Bayesian Linear Regressor

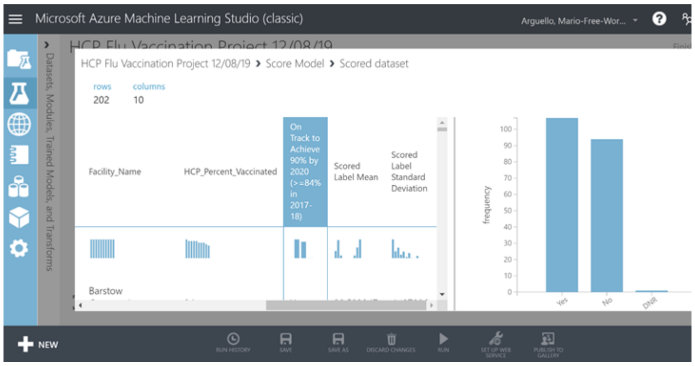
Settings

Setting	Value
Lambda	1
Min Lambda	1E-06
Noise Fraction	0.1
Noise Variance For Uniform	0.001
Allow Unknown Levels	True
Random Number Seed	

Figure 8. Azure model of trained versus untrained model.

We agreed that our prediction to be true with the “allow unknown levels” setting. Even through this module came untrained, we were in the right direction.

Finally we opened our visualization in the evaluation module. When we opened it we were able to get a better picture of our final conclusion from our prediction image below.


Microsoft Azure Machine Learning Studio (classic)
Arguello, Mario-Free-Wor...
Datasets, Modules, Trained Models, and Transforms
frequency
100
90
80
70
60
50
40
30
20
10
0
Yes
No
DNR
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

HCP Flu Vaccination Project 12/08/19 › Score Model › Scored dataset

rows	columns
202	10

Facility_Name	HCP_Percent_Vaccinated	On Track to Achieve 90% by 2020 (>=84% in 2017-18)	Scored Label Mean	Scored Label Standard Deviation
Barstow				

Figure 9. Azure model of visualization full scored sheet on clinics that were on track to have 90% vaccinated by 2020.

Upon clicking the question “on-track to achieve 90% by 2020” column, our histogram reads that there will be personnel getting vaccinated by 2020. We found that the high majority of personnel who will get vaccinated are located in the counties in Southern California and the San Francisco Bay Area. We included the names of the clinics, although it was originally thought of as unnecessary data. We discovered that hospital names told us that many of these hospitals are located in affluent communities. Additionally, we observed that 40 percent of hospitals are in underserved communities. Overall we concluded that personnel did receive vaccinations as needed for the flu season of 2017-2018. Another reason could be that personnel could not afford these shots and were not covered in their benefits package as medical staff.

III. Conclusion

We discovered in Kibana and Microsoft Azure that Health Care clinics in California have a goal to reach a certain percentage of vaccinated personnel. We utilized the visualization tools of Kibana and Excel to identify clinics that need the most improvements and the Bayesian Linear Regression tool to find insights and predictions with this data set. We also agreed that while some of these modules were incorrect, our prediction results were closely aligned with the results of the highest percentage of patients getting vaccinated, Los Angeles county, and the least amount of patients getting vaccinated, Fresno county.

References

“Datasets.” *California Open Data*, https://data.ca.gov/dataset.

“Health Care Personnel Influenza Vaccination.” *California Open Data*, 12 Dec. 2019, https://data.ca.gov/dataset/health-care-personnel-influenza-vaccination.

“Health Care Personnel Influenza Vaccination.” *California Open Data*, 12 Dec. 2019, https://data.ca.gov/dataset/health-care-personnel-influenza-vaccination.

Xiaoharper. “Bayesian Linear Regression - ML Studio (Classic) - Azure.” *ML Studio (Classic) - Azure | Microsoft Docs*, https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/bayesian-linear-regression.

“Microsoft Azure Studio link” https://gallery.azure.ai/Experiment/HCP-Flu-Vaccination-Project-12-08-19-2