Carnegie Mellon University Africa

Course: 04637-A

Course Name: Mobile Big Data Analytics and Management

Instructor: Emily Aiken

Assignment number: 2

Submitted by:

Name: Hakizimana Uwizeye Placide
Andrew ID: hup
Email: hup@andrew.cmu.edu

Introduction

In this report for assignment 2, documents the results and insight gained while working on it. And steps performed to achieve the desired output.

Question 1

While solving this question, First I have to analyse the data and load them

1.1 Create a line plot showing the weekly counts of new cases.

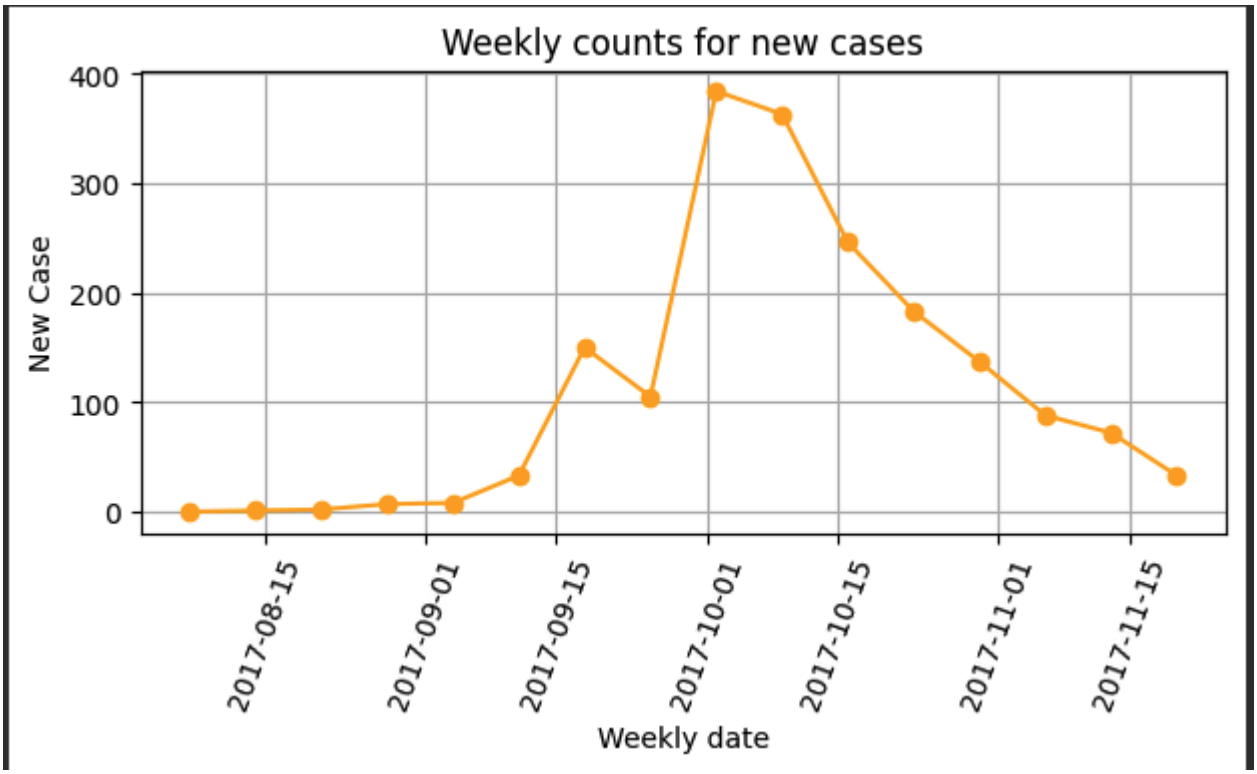

Figure 1: epidemologival_weekly_data

The above figure clearly shows how the cases evolved on a weekly basis where on October 02,2017 we observed a high number of cases reaching to **396** approximately.

1.2.1 When was the first plague case?

The first recorded case in the dataset occurred on **August 14, 2017**

1.2.2 When was the last plague case?

The last recorded case was on November 20, 2017

1.2.3 When was the first plague case?

The outbreak peaked during the week of **October 2, 2017**, which had the highest number of new cases reported.

Question 2

2.1

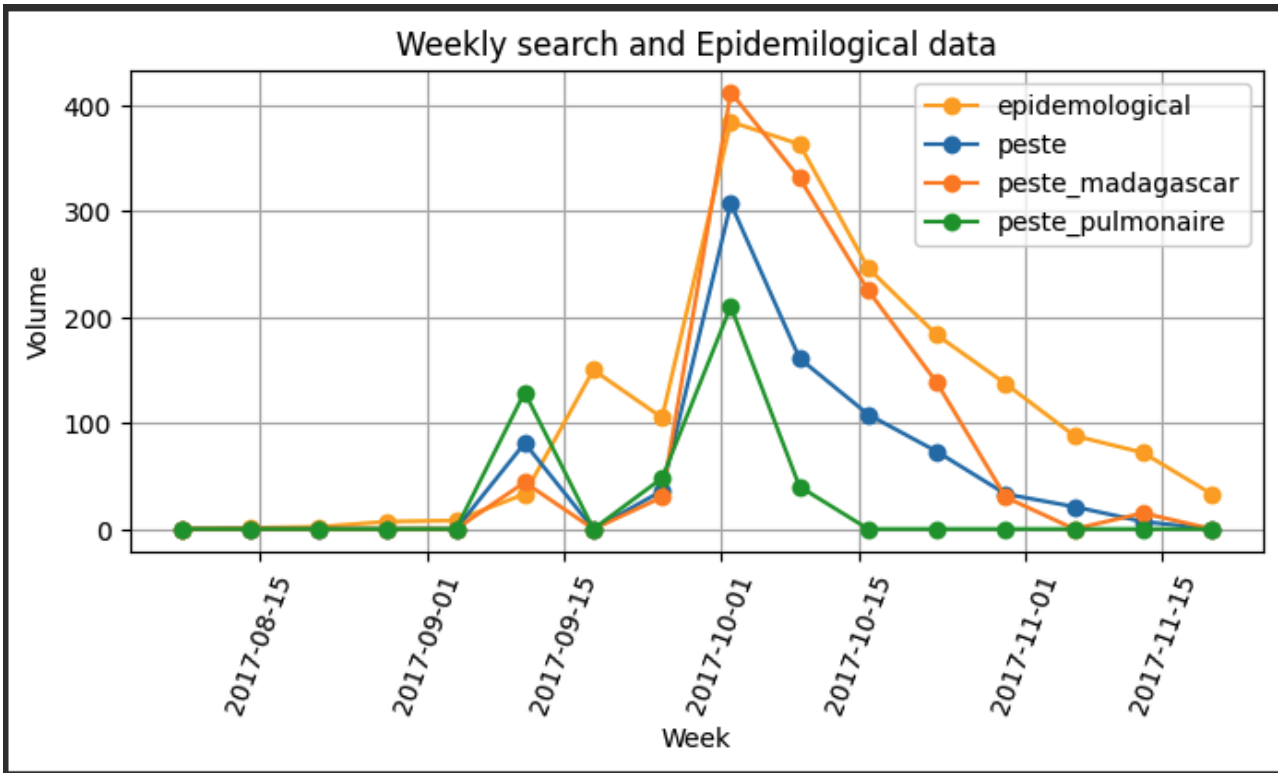

Figure 2: Weekly peste search and epidemiological data

From Figure 2, that shows weekly search trends during the 2017 plague outbreak, with a focus on “epidemiological”, “peste”,”peste_madagascar”, and “peste_pulmonaire”. The searches rose highly in mid-September, and reached their peak in early October, as there was an increase of the outbreak severity in the community. The terms “epidemiological” and “peste” had the highest search volumes, while “peste_madagascar” stood out with a notable spike, surpassing 400 cases in the October week.

2.2. What was the first day that each search term appeared,What was the last day that each search term appeared, When did the use of each search term “peak” (which week had the largest
number of searches for each term)?

Search term	**first day**	**last_day**	**peak_week**
Peste	2017-09-11	2017-11-13	2017-10-02
Peste Pulmonaire	2017-09-11	2017-10-09	2017-10-02
Peste Madagascar	2017-09-11	2017-10-09	2017-10-02

Question 3

3.1 Correlation between searches for “peste” in Madagascar and the ground truth case counts and other search terms

Peste **Madagascar** correlation searches: **0.9310496462758787**

- This high correlation coefficient, clearly shows a very strong positive correlation between for “peste Madagascar” and the ground truth data “epidemiological data”
- We can then infer that as the number of cases increased, public awareness specifically in Madagascar also increased significantly, thus having more searches.

Peste **correlation** searches: **0.8690834587124728**

- Likewise, this also represents a strong positive correlation, but somehow lower to the “Peste Madagascar”.
- This indicates that searches globally or with the keyword of “peste”, reflect the trend of the outbreak though it may include searches which are not directly related to the Madagascar case.

Peste **Pulmonaire** correlation searches: **0.5009197203586835**

- This correlation is indeed positive, though suggesting a weaker linear relationship between the searches for “Peste Pulmonaire” and the actual case counts.
-

3.2 Based on the above correlation coefficients the **Peste Madagascar ,**proves to have the highest correlation with the Ground truth case counts. The **Peste Pulmonaire** is the lowest correlation with the ground truth case as it has **0.5009 .**

3.3

Search volumes may align well with epidemiological data for several reasons. First, during an outbreak, people search for symptoms, ways to prevent the disease, or updates. This can show how the disease is spreading and how worried people are. Second, news reports and health campaigns can make more people search for information, which often happens when cases increase. This makes search trends a good way to track public awareness of the outbreak.

On the other hand, search volumes do not always match epidemiological data. News coverage can make more people search for information, even in places with few cases, which can create a gap between searches and real cases. Also, internet access and education levels affect how people search. Some regions may have fewer searches simply because fewer people have the internet or know how to use it, leading to limited data in certain areas.

Question 4

4.1 after splitting the training and test set below are the length of each dataset and corresponding rows and columns

```
Number of weeks in the Training data: 7 and the rows: (7, 4)
Number of weeks in the Testing data: 9 and the rows : (9, 4)
```

4.2

Univariate Regression: Correlation **= 0.931,** Mean Absolute Error **(MAE) = 145.883**

4.3

Multivariate Regression: Correlation **= 0.894, MAE = 147.667**

4.4

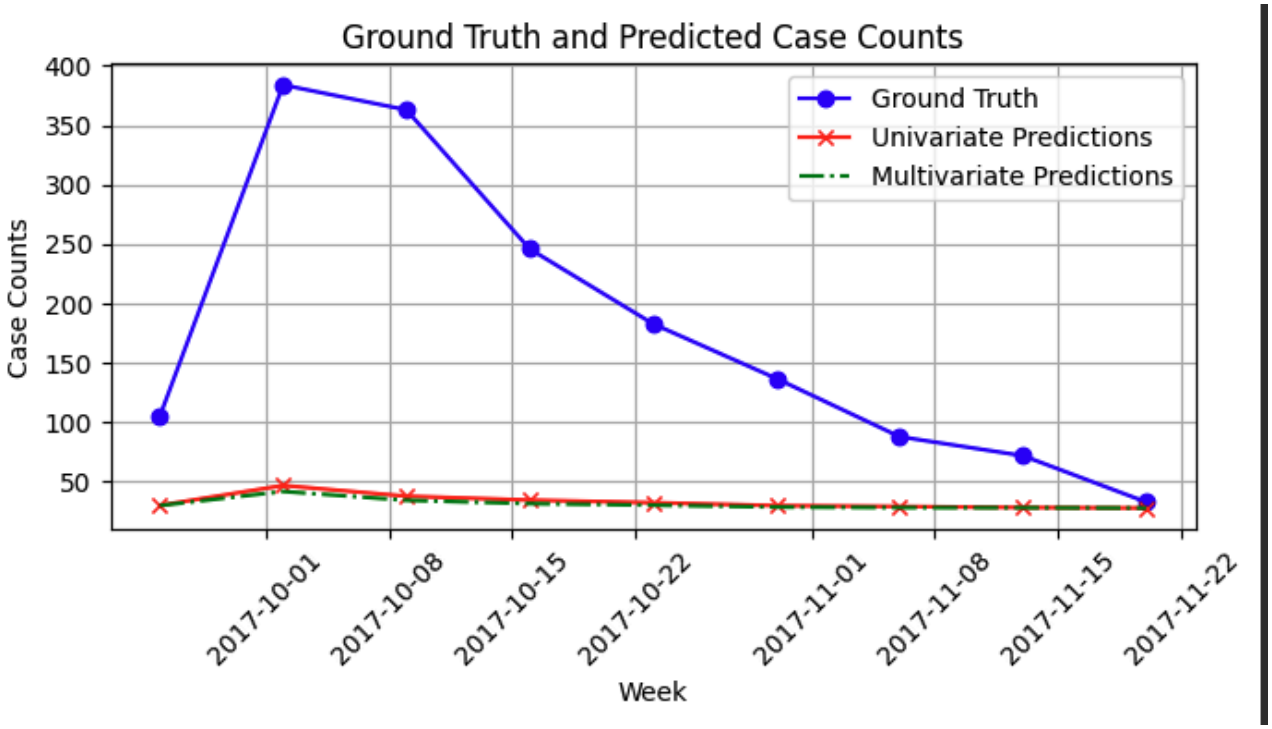

Figure 3: ground_truth and predicted case counts

4.5

From the metrics shown in sub-answer(4.2 and 4.3) on question 4, it appears that the univariate models show slightly better performance in both metrics as opposed to multivariate model.

The reason why univariate seems to be more effective is due to its simplicity and direct relevance of the independent input variable used, which could be more significantly correlated with the outcome than the additional variables included in the multivariate model. Moreover the independent variables used on the multivariate model may not provide relevant information and could introduce noise which could potentially lead to

overfitting.

4.6 I am not surprised with the accuracy of the prediction derived from two models, this is due to the fact that the model's correlation coefficients are relatively low, from which there might be a weak linear relationship between the independent input variables and dependent variables.

Question 5

5.1

Correlation between predictions and actual case counts on the test set:
0.5659406539941269

Mean Absolute Error (**MAE**) on the test set: **703.202**

5.2

Correlation between lags predictions and actual case counts on the test set:
0.33410935118331003

Mean Absolute Error (**MAE**) on the test set: **1586.798**

5.3

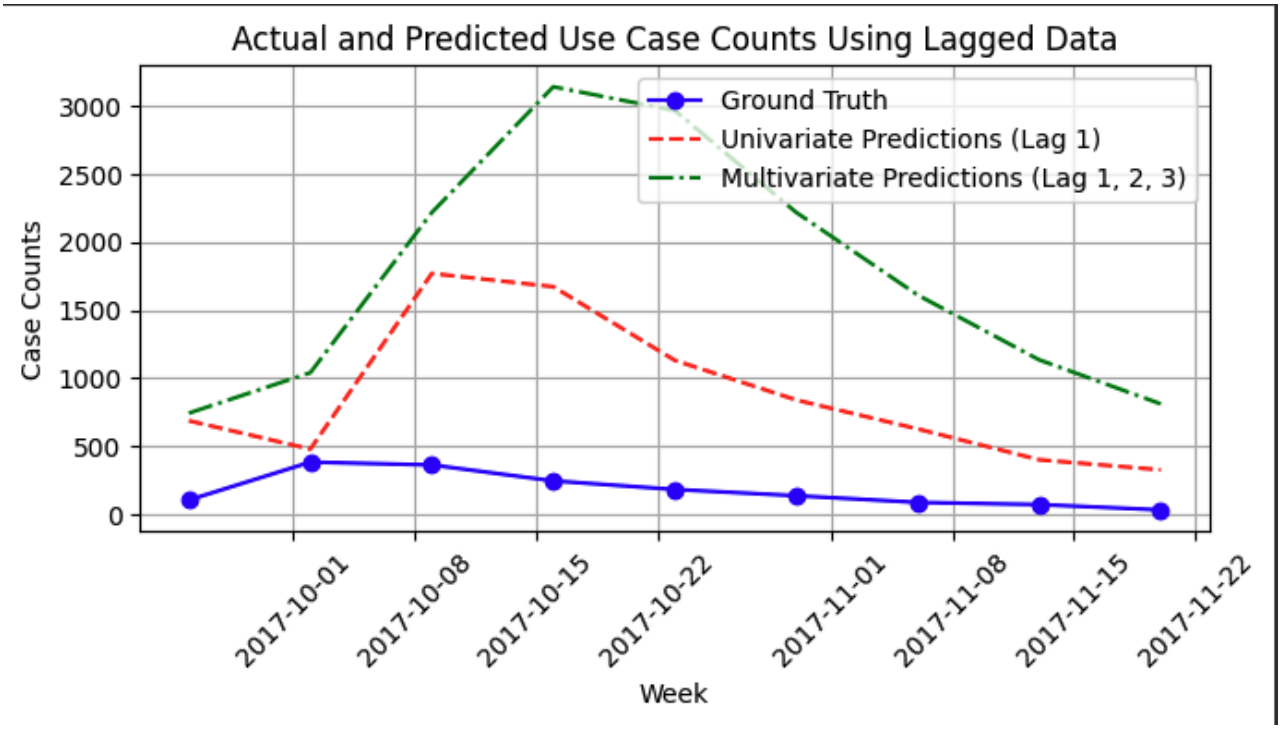

Figure 4: actual and predicted use ase counts using lagged data

5.4

Based on the graph on figure 4 it clearly shows that Univariate predictions are close to the Ground Truth compared to the Multivariate predictions, mostly during the earlier weeks.

Based on the correlation and mean absolute error for the Multivariate regression, as it has 3 lagged variables (lag 1,2,3) , this can lead to overfitting.

However, the simplicity of the univariate regression, which only uses lag 1, is way close to generalizing better to unseen data and avoid unnecessary complexity. And this makes it more accurate as it aligns better with the actual data.

5.5

With Observing the figure 3 which represents the predictions using search term data and figure 4 which represents predictions using lagged epidemiological data, the more accurate, as seen it is lagged epidemiological data as it shows better alignment with overall trends, especially with the univariate model.

The reason why lagged epidemiological data might be more accurate is also due to it directly reflects on the past case trends whereas search trends data might not correlate strongly with the actual case counts.

Secondly, the model complexity could be one of the reasons, as the univariate model for the figure 4 performs better because it avoids overfitting or introducing noise from additional lagged variables.

Question 6

6.1

correlation between lags and search terms predictions and actual case counts on the test set: **0.5715670618844626**

Mean Absolute Error (MAE) on the test set: **724.730**

6.2

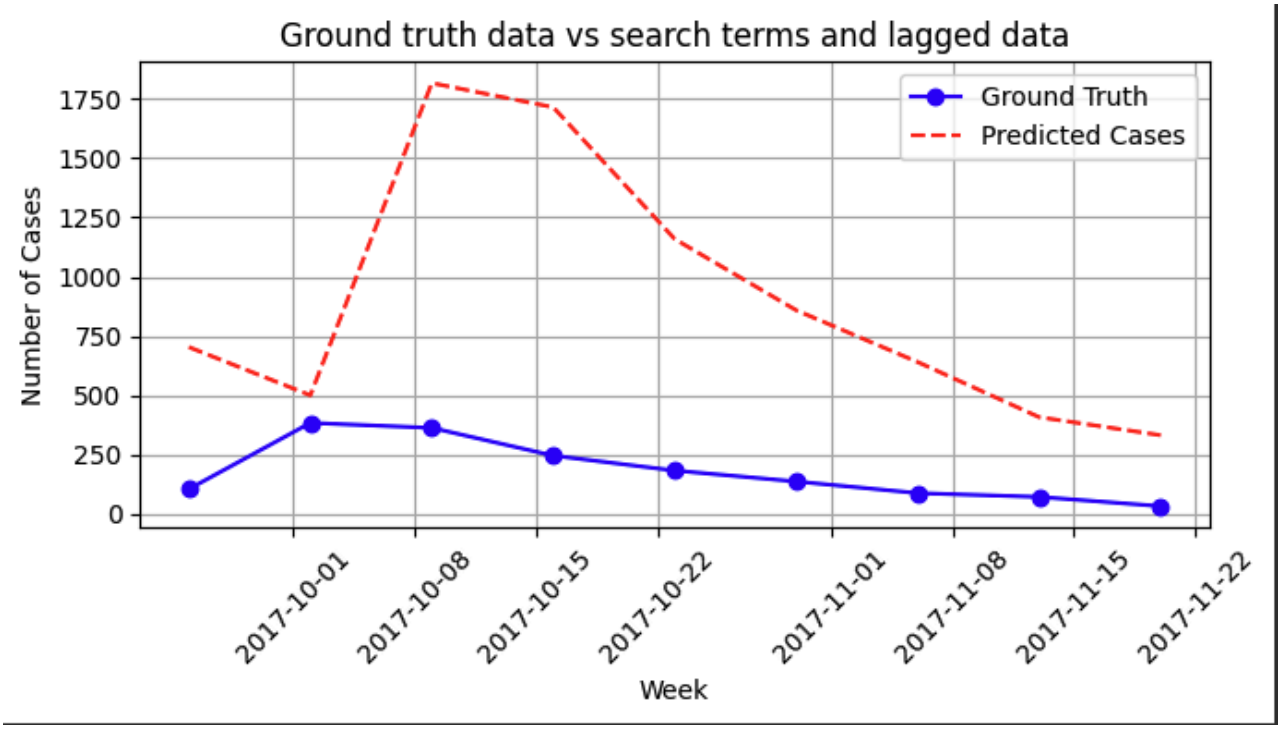

Figure 5: Ground truth vs search terms and lagged data

Question 7

7.1

Based on the three graphs, plotted using different methods, the methods that seems to perform best out the three is from figure 4 with model that combines both search terms and epidemiological, because it contains more comprehensive information, allowing it to approximate the trends in the ground truth more effectively than models relying on either search terms or epidemiological lags only.

7.2

The multivariate model combining **search terms and epidemiological lags** achieves the lowest **MAE,** because it has complementary data sources, enabling better prediction of trends and reducing deviations from ground truth. The other methods either fail to capture key dynamics (Figure 3) or show large overestimations (Figure 5).

7.3

Both 7.1 and 7.2 are the same, I think that having a multivariate model, which uses both **search terms and epidemiological lags**, consistently achieves the best performance in correlation and mean absolute error (MAE). As it combines different data sources, which leads to providing a more accurate representation of trends and reduces errors, making it more effective than models than others defined in our scenario.

7.4

The models show some big errors, sometimes predicting too few or too many cases. Relying only on these predictions for important public health decisions, such as mask mandates, curfews, or lockdowns, would be risky. having improved model accuracy would require at least to have more data sources and better modeling techniques are needed.

7.5

In a global outbreak like COVID-19, search terms can be useful for tracking the disease widespread by the use of the internet, as it can leverage public awareness, and as well strong links between searches and infections trends. However, their effectiveness still depends on regional variations in internet access and different ways of searching.

7.6

Since many rural areas in Rwanda lack internet access and most healthcare workers are concentrated in the capital, Kigali, public awareness of the Marburg virus is limited. This suggests that search terms would be less effective for tracking Marburg in Rwanda compared to pneumonic plague in Madagascar**.**