Predicting HPV and MMR vaccination usage with linear regression modeling.

Ken Luy

April 14th, 2016

1. INTRODUCTION

In this study, it was desired to use Google Trends data to predict vaccine usage [1]. It was intuitive that an increase in searches of words relating to a vaccine would result in high usage of that vaccine. The model used to predict vaccine usage was an ordinary least squares linear regression model.

The HPV and MMR vaccines were chosen to do the study on. No distinction was made between HPV-1, HPV-2, and HPV-3; and similarly, no distinction was made between MMR-1, MMR-2, and MMR-3. HPV was assumed to encompass HPV-1, HPV-2, and HPV-3; and MMR was assumed to encompass MMR-1, MMR-2, and MMR-3.

2. METHODOLOGY

The descriptions from two webpages for each of the vaccines in this study were looked at. For each vaccine's two webpages, words that appeared in both webpages were chosen as the query words. Stop words (words that are common in the Danish language) did not count as query words. Also, punctuations were replaced with spaces, and all the words were converted to be lowercase.

One Google Trends query was performed for each query word. The Google searches that originated at Denmark were used for this study. Data from January 2011 to September 2015 were queried. The PyTrends library was used to programmatically perform the queries and save the results [2].

One of the challenges was that some of the data had monthly data, and some had weekly data. But it was desired that the data be bucketed into months. So of the data that were weekly, there was a conversion that was done. This conversion process involved making the average of a month's weekly frequencies to be the month's overall frequency. To handle what to do with weeks that start in one month and end in another, it was decided that the month that a week starts at would be used. So a month's weeks were decided to be all weeks that start that month.

After the frequency data files were cleansed and processed, they were used for linear regression modeling. The model chosen was Ordinary Least Squares. The data from Google Trends were used as training data, and clinical data from Hansen et al. were used as ground truth [3]. The sklearn Python module was the machine-learning library used [4].

Five-fold cross-validation was used to train the model. The cross-validation was manually done, whereby the training data were split into 5 groups, and only 4 out of 5 of them at a time were trained for each fold. Five models were generated, and predictions from each of them were generated. The average of the five models' predictions were used as the prediction.

3. FINDINGS

From January 2011 to September 2015, there were 57 months. The first 50 months were used as the training data, and the remaining 7 months were used as the test data. The first 50 months of the clinical data were used as the ground truth for the training data, and the 7 months after that were used as the ground truth for the test data.

There were 38 query words used for HPV, and 32 query words used for MMR. Two of HPV's query words and five of MMR's query words had zero frequency throughout January 2011 to September 2015, so the data of these query words were not use as features in the linear regression modeling.

Table 1: Monthly HPV Vaccinations in 2015

	Mar	Apr	May	Jun	Jul	Aug	Sep
Predicted	12.6	61.3	51.0	56.8	26.4	39.1	45.5
Actual	40.8	31.0	21.3	18.7	9.68	16.7	14.2
RMSE	28.8						
Min. of 57 months	9.69						
Max. of 57 months	73.9						

Table 2: Monthly MMR Vaccinations in 2015

	Mar	Apr	May	Jun	Jul	Aug	Sep
Predicted	153	128	131	97.8	84.9	98.2	94.7
Actual	161	109	104	102	91.9	106	82.2
RMSE	14.2						
Min. of 57 months	64.9						
Max. of 57 months	161						

It is shown in Tables 1 and 2 the predicted vaccinations of HPV and MMR from March to September in 2015. The RMSE of the predicted values were calculated. Also, the minimum and maximum monthly frequencies of the 57 months for each vaccine are shown on the tables.

4. CONCLUSIONS

The linear model for HPV was not so good, whereas the linear model for MMR was quite successful. The RMSE of the linear model for HPV was 28.8. This is quite high considering that it is about half of the range of the minimum and maximum ground truth (9.69 and 73.9). On the other hand, the linear model for MMR was quite low. It was 14.2, which is less than a sixth of the range of the minimum and maximum ground truth (64.9 and 161). Overall, using the linear model for MMR would be a good predictor, whereas using the linear model for HPV would not.

5. REFERENCES

[1] "Google Trends." Google Inc. N.p., n.d. Web. 11 Apr. 2016.

[2] "Pytrends." General Mills. n.d. Web. 11 Apr. 2016.

[3] N. D. Hansen, C. Lioma, and K. Molbak. Predicting vaccination uptake using web search queries. in press, 2016.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Web Science 2016 DIKU, Denmark

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$10.00.