Using NHFS Data To Forecast Vaccination Rates For The H1N1 And Seasonal Flu

Ibrahim Atomanson
20265875
psyia3@nottinham.ac.uk

Xiaonan Wang
20419606
alyxw54@nottingham.ac.uk

ABSTRACT

Are there relations that lie in the features of people who are vaccinated? After the H1N1 Flu was declared a global pandemic in 2009, a national survey was conducted to track vaccinations with H1N1 and seasonal influenza.We aimed to use this dataset to answer our research questions which explore possible relations or trends leading to someone being vaccinated.To achieve this purpose various data wrangling and pre processing methods are applied to produce insightful visualisations produce different classification models which we look to compare and evaluate and possibly receive answers from.

INTRODUCTION

The National 2009 H1N1 Flu Survey is a US dataset used to track the trivalent seasonal influenza vaccine and pH1N1 vaccination coverage rates in 2009 and 2010. It contains 36 parameters that were gathered through a telephone survey of homes, including knowledge, opinions, and demographic elements that could affect vaccination rates.Vaccines provide immunisation and can reduce the spread of diseases through herd immunity, a full description of features can be found using the link in the reference section[1].

For our research questions we wanted to target different factor categories as to why/ why not the vaccines are taken, for example, demographic factors, behavioural factors , the education level of the respondent and their opinion in regards to the vaccines. The following questions are:

How do levels of concern and knowledge about the H1N1 flu influence the likelihood of receiving the H1N1 vaccine?

How do different behavioural factors, such as avoidance of close contact or reduced time at large gatherings, relate to the likelihood of receiving the H1N1 and seasonal flu vaccines?

Does the perceived effectiveness and risk of getting the H1N1 and seasonal flu vaccines impact an individual's likelihood of getting vaccinated?

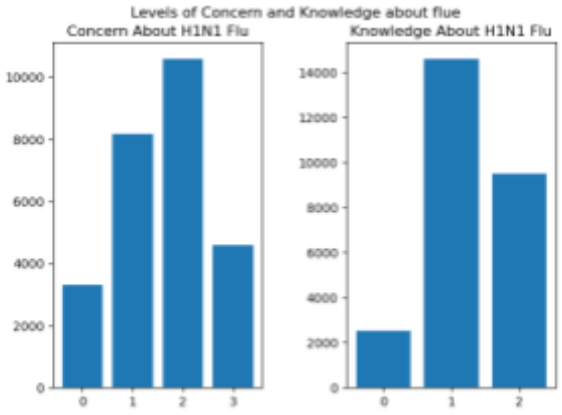

Figure 1 - Plot exploring concerns of each flu

These figures provide a statistical overview of the dataset related to the questions. Figure 1 displays the distribution of concern and knowledge about the H1N1 flu in relation to Q1. The median values for knowledge and concern are 1 and 2 respectively, indicating that most people have low knowledge and moderate concern.

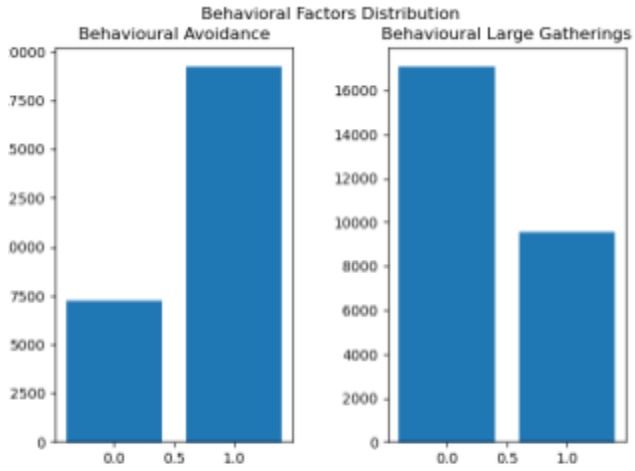

Figure 2 - Plot exploring the distribution of behavioural categories

Figure 2 shows the distribution of results based on behaviour, required for Q2, it tells us the majority of people avoid close contact with others with flu-like symptoms, however the majority have not reduced time at large gatherings.

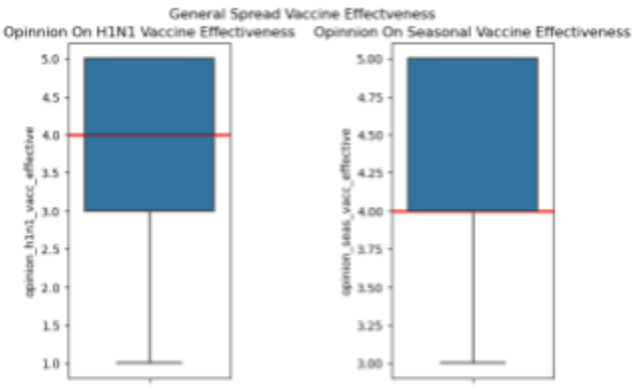

Figure 3 - Box plot showing the spread of flu vaccine effectiveness (opinions)

Figure 3 depicts the variance in opinions on vaccine effectiveness, as well as showing the median value in

regards to Q3.Both boxplots tell us the majority of records have the opinion both vaccines are somewhat effective to very effective which would suggest they would receive the vaccine.

LITERATURE REVIEW

When analysing influenza vaccination among Italian Public Health medical residents [3] multivariable logistic regression was used to analyse factors involved in the decision to get vaccinations therefore logistic regression will be used in the same way in this data analysis journey. The only difference is that this data set consists of categorical and binary variables therefore from a report regarding H1N1 and Seasonal Flu : Vaccine Cases using Ensemble Learning approach[4] the same approach of using one hot encoding to be able to parse categorical values will be used.

METHODOLOGY

For our methodology we aimed to show the idea that data analysis is not a linear, step by step process but rather a highly iterative and non linear process as described in the Art of Data Science [2]. As shown in figure 4 our plan was to outline a clear question, perform exploratory data analysis (EDA) consisting of various cleaning and pre processing techniques before analysing through visualisations. We would then create our models using separate classification techniques.

Figure 4 - Diagram representing our approach

To begin the training features and training labels were loaded into separate dataframes.To get an overall understanding methods such as describe() from the pandas library were used to inspect the dataframes values.The pandas method columns() was used to inspect differences in features between the training dataframe and the labels, they were then merged using merge() in order to be look at the records with their results and also be able to remove labels if their feature needs to be removed.Duplicates were checked for and removed using duplicated(keep=false).Next we used replace() to rename the values in the education column to make it clearer, we merged college graduates and people who attended college to just 'higher education' as they all attended college regardless of how long.the '12 years' value was also changed to '>=12 years' for more clarity.

For data imputation separate techniques were used simply to explore different methods.Xiaonan attempted to use median imputation, all the rows with missing values were identified then imputed with the median value for each column using a for loop combined with the median() method applied to each column then using the fillna() method to impute the values.

Ibrahim used Random hot deck imputation in order to try to maintain possible relations or trends.First all the columns and sum of null values were printed.Records with more than 50% of null columns were dropped using thresh parameter as too many imputed values can skew data.Hot deck imputation was then performed using columns regarding age, race and sex because they all has no null values and with aim to retain patterns.Nested for loops were used in order to iterate through columns then rows.As mentioned EDA is not a step by step process , Ibrahim noticed he would need multiple imputation methods as there were still null variables. First he investigated large null values in columns. Columns to do with employment were replayed with unemployed variables as it was decided that's what they represent. As for the rest of columns containing null values, mode imputation was used to fill the null values.The data frame was then narrowed to only include features relevant to the research question.

Analysis was carried on the same data but using different visualisation techniques each.

Ibrahim decided to first look at the dataframe as a whole through a simple pie chart to explore what percentage of the population are vaccinated. Sums of each target variable were created using sum(). Matplotlib was used to create pie charts of these values and place the pie chart for each vaccine into a subplot to be able to view in the same plot.for Q1 a barplot was created using the seaborn library, the data frame was reshaped to be grouped by the h1n1_vaccine and h1n1_concern column using groupby(). The size() function was used to calculate the sum for each group.When using sns.barplot() the hue parameter was set to h1n1_vaccine in order to show the different bars for vaccination status.The same was done for knowledge.For question 2 a pie chart was used to compare how different behaviour affected vaccination rates using pie() method from matplotlib and grouping the data by columns which was often done as dealing with categorical values.For question 3 the same

method of grouping the data by relevant columns was used to then plot using seaborn barplot() where the vaccination status was shown using the hue parameter, the 1's and 0's wer explained in the legend using the parameter title.

In order to compare and visualise the relationship between the variables, Xiaonan utilised a stacked bar plot. He examined how awareness and anxiety about the H1N1 virus affected people's likelihood of getting the vaccine. A well-known Python drawing package called "Matplotlib.pyplot" was imported in order to build charts and visualisations. 'h1n1_concern' and 'h1n1_knowledge' were grouped using the 'groupby()' method, and the sum of 'H1N1_vaccine' was computed. The 'plot()' method was used to create the stacked bar plot, which represents many data sets. The 'xlabel()', 'ylabel()', and 'title()' functions were used to add axis labels and titles for clarity. 'legend()' was used to include legends to explain 'H1N1 Knowledge'. With the help of the show()' method, the graph was presented.

Question 2 uses a grouped bar plot to analyse the different behavioural factors (avoiding close contact and large gatherings).The 'groupby()' function is used to group 'behavioral_avoidance' (indicating the degree of close contact avoidance) and 'behavioral_large_gatherings' (indicating the degree of reduction in large gathering time) and to calculate the ' H1N1_vaccine' (the number of people who received H1N1 vaccine) and 'seasonal_vaccine' (the number of people who received seasonal flu vaccine) in total, and later plot the grouped histogram using the 'plot()' function. Set kind='bar' to indicate plotting a bar chart and set the rest of the chart properties in the same way as in question 1.

Question 3 uses violin plots to visually compare the opinion of the perceived effectiveness of taking the h1n1 vaccine and its vaccination rate.The seaborn library was used for the function violinplo() to generate the plot.The horizontal and vertical axis variables by specifying the x and y parameters, where 'x='h1n1_vaccine" indicates vaccination with H1N1 vaccine and 'y='opinion_h1n1_vacc_effective", and use 'split=True' to split the data in the Violin chart, followed by setting the chart properties in the method is the same as in question 1.

For classification we decided to settle upon these 2 different classification methods, Decision Tree Classification and Logistic Regression.

Xiaonan used decision tree classification because it provides a clear understanding of the importance and impact of features and also displays how the model makes predictions through branches and nodes on the decision tree.This makes the decision tree model very useful in cases where the results need to be interpreted and understood. Since the method used for all three problems is the same but the variables in the data are different, the general method will be explained .

First 'DecisionTreeClassifier' class was imported from the 'sklearn' library, which is the class used to build the decision tree classification model. It also implements the decision tree algorithm. After that, the 'train_test_split' function is imported from the 'sklearn' library.' train_test_split' function is used to divide the dataset into a training set and test set for model training and evaluation. It is also necessary to import functions for evaluation metrics such as 'accuracy_score', 'precision_score' and 'recall_score', which are used to calculate the classification model's accuracy, precision and recall. Finally, it is also necessary to import the tree module, which provides functions for drawing the decision tree model, including drawing a graphical representation of the decision tree .The features and target variables related to each question are first selected from the dataset. For example in Q1, we select 'h1n1_concern' and 'h1n1_knowledge' as features and 'h1n1_vaccine' as the target variable. The data set is read and the classification variable is encoded using 'LabelEncoder' to convert it into a numerical representation for the classification task. The data sets are then divided into training sets and test sets for use when training models and evaluating performance. Here, we split the data set into a 70% training set and a 30% test set, via the 'train_test_split' function. We then adopted our 'DecisionTreeClassifier' as our classifier. Decision tree is a kind of supervised learning algorithm based on tree structure. It constructs classification models by learning to split features. Then the training set is used to train the decision tree classifier, and the model is fitted by fit method. The model will learn how to predict the classification outcome of the 'h1n1_vaccine' based on the 'h1n1_concern' and 'h1n1_knowledge' characteristics by visualising the trained decision tree model using the 'plot_tree' function. This presents a graphical decision tree, showing the splitting of features and the classification rules for nodes. Finally, the trained model is used to predict the test set, and the predicted results are obtained by the predict method. accuracy, precision and recall of the predicted results were calculated, which could evaluate the performance of the model in the classified tasks.

Ibrahim decided to use logistic regression considering the fact that the research questions explore relations between different features in the dataset, furthermore it is simple to encode categorical features and place them into a model to obtain reliable results and easy to interpret results to help answer our questions.Inspiration came from study into the influenza vaccination among Italian public health residents [3] where logistic regression was used to analyse factors in vaccination. First important libraries were imported such as pandas for data manipulation and analysis, LogisticRegression from sklearn.Linear_model used to create a regression model and classification report to do as stated.Next the data frames were narrowed down to the variables specific to each question, e.g q1 included

h1n1_concern and h1n1_knowledge. One hot encoding was then performed to convert the categorical variables to numerical dummies.For the model training an instance of the logistic regression model was created using LogisticRegression ().The regression model was then fitted with the training data and corresponding target variable using the fit() method.The trained model was then used to predict the target variable using the predict() method.Ibrahim would then evaluate the accuracy of the prediction using 'accuracy_score' from sklearn.metrics, this can be used for comparison with other classification models and for hyper parameter turning.Finally a classification report was created to show other evaluation metrics such as recall and precision. Across all three questions the same methodology is used

Following the prediction to help answer our research questions the logistic regression coefficients were printed using the .coef_[0] method to obtain the list.Next the coefficients were interpreted by defining the categorical variable labels for each question and assigning that coefficient value to its label, this was performed through a for loop.These values would be placed into a dataframe using pandas Dataframe() method.Next I would plot barplots clearly showing the influence of each variable through its coefficient.

RESULTS

Starting with pre-processing, we found the results for hot deck imputation were best to stick with because having compared the distribution of data before and after imputation , the shape remained, revealing that the imputation did not skew the data. You can see an example of this through comparison of fig 1 and fig 5.

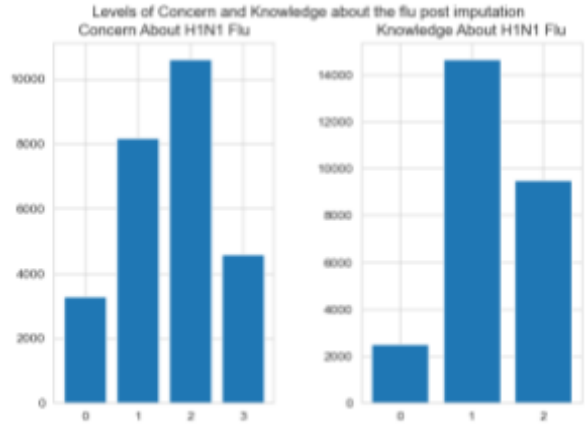

Figure 5 - Used to compare shape of data after Imputation

Looking into the results of Question 1 the bar plot results were expected, the proportion of vaccinated records increased as concern over the h1n1 flu increased this was the same for the proportion of vaccinated records increasing as knowledge of the h1n1 flu increased.This can be seen in figure 6, the same trend is also for h1n1 knowledge.

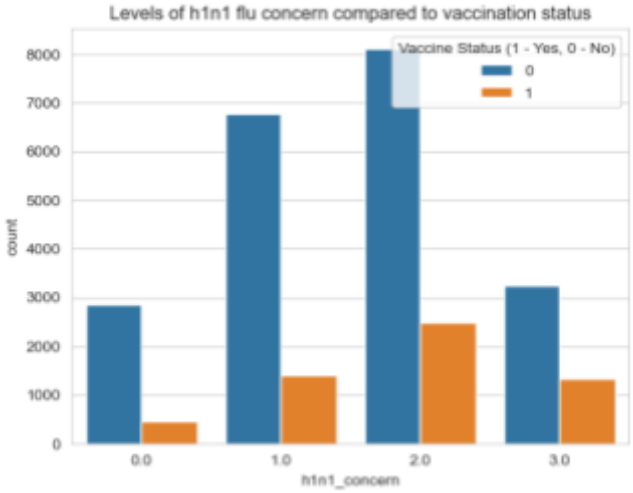

Figure 6 - Showing how the proportion of vaccinated records increases with flu concern

Despite a pie chart being used to answer question 2 the results were telling.The first pie chart inspected vaccination statuses based on how they avoided close contact with possible flu victims or not.Out of the 72% vaccinated only 16% of the total were vaccinated compared to 5% of people who are vaccinated nut do not avoid contact with people who show flu symptoms.When comparing results of time spent at large gatherings compared to vaccination out of the 36% who are vaccinated, only 8% have reduce their gathering time. These numbers suggest behaviour factors may not have as strong an influence on vaccination numbers as expected.

With question 3 we were able to obtain interesting results.Firstly our violin plot that visualised opinions on the h1n1 vaccine effectiveness against the actual vaccination rate shows us that for unvaccinated records despite having a large spread of records that believe it is at least somewhat effective, they still did not take the vaccine.It also shows the records who have a better opinion on the effectiveness of the vaccine is where majority of vaccinated records are which helps in answering the question, this can be seen in figure 7.

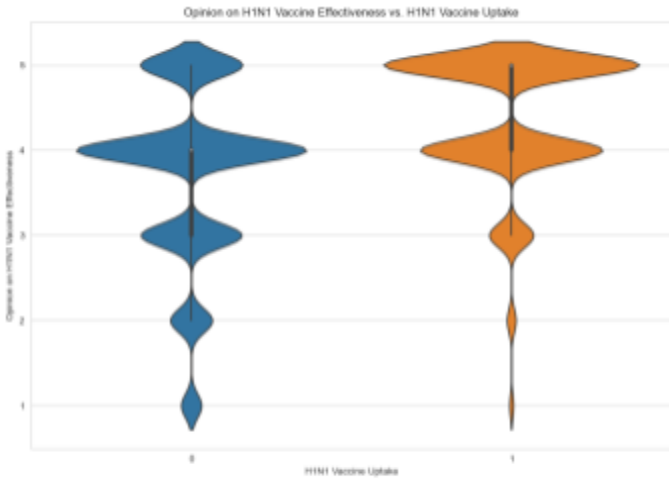

Figure 7 - Violin plot showing how vaccinated people tend to have stronger opinion on vaccine effectiveness

The same trend follows for the seasonal flu vaccine and its effectiveness where the majority of vaccinated records believe the vaccine is somewhat or very effective. This mirrors onto both grouped bar charts regarding both flu's and the risk of getting the flu, where people who are vaccinated tend to believe the flu have more risk however in figure 8 the slight anomaly is when records gave number 3 as a response of risk .

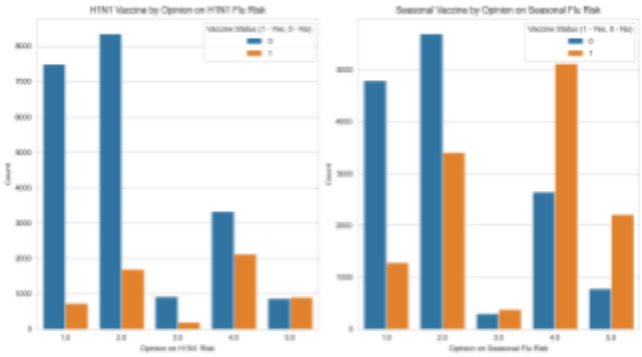

Figure 8 - Showing how opinions on each flu's risk compare to vaccination rates

However this can be explained because 3 means that participants do not have an opinion/ do not know about the matter.

For the first classification method, Decision Tree Classification, the method used is for all the questions.

The image for the first question is shown in Figure 9. The decision tree diagram shows a decision tree for predicting the "h1n1_vaccine" based on two characteristics: "H1N1 concern" and "h1n1 _ knowledge" h1n1_knowledge ". Each node represents a feature and the corresponding split criteria used to classify the sample. The branches of the decision tree represent different decision paths until the leaf nodes represent the final classification result, which is whether or not to vaccinate against H1N1. Accuracy is 0.7834239639977498, which means that the classification accuracy of the model is very good. A Precision of 1.0 and a Recall of 0.0 usually means that the model is completely correct in predicting positive class samples, but has not successfully found any negative class samples.

Figure 9 - Showing the Decision Tree for question 1

The image of the second question is shown in Figure 10. The decision tree classifier was used to study the effect of different behavioural factors (behavioral_avoidance and behavioral_large_gatherings) on the likelihood of receiving the H1N1 vaccine and the seasonal influenza vaccine, and the figure shows that samples with the "behavioral_avoidance" feature less than or equal to the samples with "behavioral_avoidance" less than or equal to 0.5 are classified to this node. In this problem, the accuracy is 0.4575, which means that the model predicts 45.75% of the H1N1 and seasonal influenza vaccines to match the actual situation. An accuracy rate of 0.5212 means that the model predicted vaccination with 52.12% of the predicted results being true vaccination, and a recall rate of 0.2383 means that the model was able to correctly predict vaccination for 23.83% of all actual vaccinated samples.

Figure 10 - Decision Tree for question 2

The third problem, shown in Figure 11, was divided according to the values taken for the feature variable 'opinion_seas_risk', with a divided Gini index of 0.416 and a total number of samples of 21331. There were 16820

samples in the 'Not Vaccinated' category and 4511 samples in the 'Vaccinated' category. This information provides a description of the statistics and division of the current nodes. The accuracy rate of 0.7819 means that the model's predictions for the H1N1 influenza vaccination and seasonal influenza vaccination match the actual situation by 78.19%. The accuracy rate was 0.8137, meaning that the model predicted vaccination with 81.37% of the predicted results being the actual vaccination. The recall rate was 0.9277, meaning that the model was able to correctly predict vaccination for 92.77% of all actual vaccinated samples.

Figure 11 - Decision Tree for Question 3

Ibrahim's results from logistic regression were informative and helped to provide an answer with our research questions.

Starting with question 1 after training and testing the dataset the accuracy, precision and recall of the logistic regression model could be calculated. An accuracy of 0.79 indicates just under 80% were predicted correctly which is fairly accurate, there was also a recall rate of 1 means that the model was able to correctly predict vaccination for all vaccinated samples.Looking into how we can answer the question of how do levels of concern and knowledge about the H1N1 flu influence the likelihood of receiving the H1N1 vaccine.Figure 12 displays boxplot showing the coefficient of each categorical variable value and how that leads to a positive target value (vaccinated).We can see the coefficient increase and goes from negative to positive as concern about the flu increases and as knowledge of the flu increases.

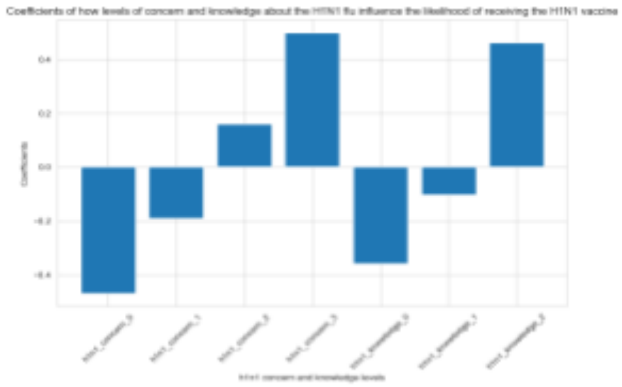

Figure 12 - Bar plot of how concern and knowledge values effect vaccination rate

After training and testing the dataset for question 2 the accuracy, precision and recall of the logistic regression model could be calculated. Starting with the h1n1 vaccines' behavioural categories (possible flu victims avoidance and avoiding large gatherings) An accuracy of 0.79 indicates just under 80% were predicted correctly which is fairly accurate, there was also a recall rate of 1 means that the model was able to correctly predict vaccination for all vaccinated samples.For the seasonal flu vaccines' behavioural categories (possible flu victims avoidance and avoiding large gatherings) An accuracy of 0.54 indicates just above 50% were predicted correctly which does not provide the most confidence answering this question, there was also a recall rate of 0.74 which means that the model was able to correctly predict vaccination for 74% of vaccinated samples. Again the coefficients of each category variable is used to try to answer our research question of how do different behavioural factors, such as avoidance of close contact or reduced time at large gatherings, relate to the likelihood of receiving the H1N1 and seasonal flu vaccines? In figure 13 you can see when records avoided people showing h1n1 flu symptoms there was a weak positive correlation just under 0.25 which was stronger than records who would avoid large gatherings, this was also a weak positive correlation but was under 0.10.

Figure 14 shows when records avoided people showing seasonal flu symptoms there was a just weak/moderate positive correlation just under 0.30 which was stronger than records who would avoid large gatherings, this was also a weak positive correlation but was under 0.25.

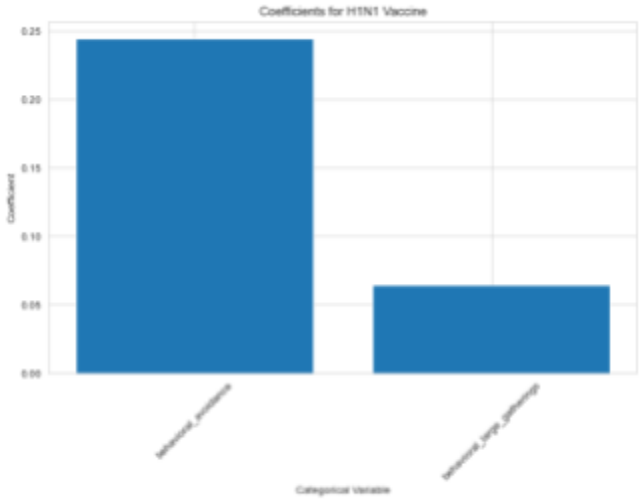

Figure 13 - Box plot showing the coefficients of h1n1 behavioural categories

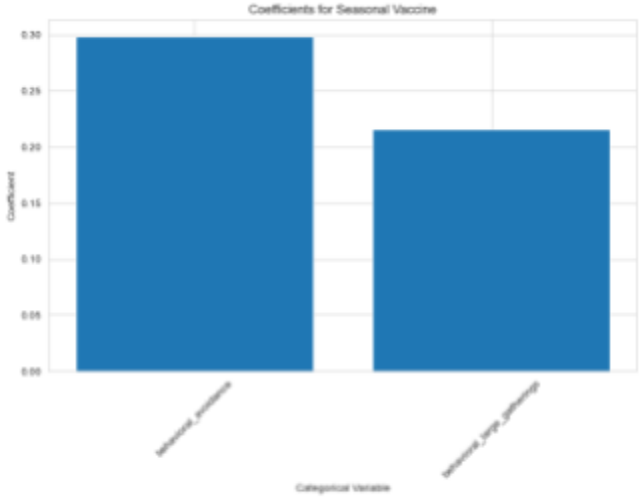

Figure 14 - Showing the coefficient of behavioural categories for the seasonal flu

After training and testing the dataset for question 3 the accuracy, precision and recall of the logistic regression model could be calculated. Starting with the h1n1 vaccines' categories regarding opinions (on flu vaccine effectiveness and the risk from not getting the vaccine) An accuracy of 0.80 indicates just 80% were predicted correctly which is the most accurate out of all the questions, there was also a recall rate of 0.88 means that the model was able to correctly predict vaccination for all vaccinated samples.For the seasonal vaccine categories regarding opinions (on flu vaccine effectiveness and the risk from not getting the vaccine) An accuracy of 0.71 indicates just above 70% were predicted correctly which provides confidence in trusting this model, there was also a recall rate of 0.70 which means that the model was able to correctly predict vaccination for 70% of vaccinated samples. The coefficients of each category variable is used to try to answer our research question of Does the perceived effectiveness and risk of getting the H1N1 and seasonal flu vaccines impact an individual's likelihood of getting vaccinated? In figure 15 we can see h1n1 flu risk and vaccine effectiveness coefficients plotted for each possible value entered by a record. Both variables are similar in that as the opinion on effectiveness and risk becomes higher (believe the vaccine is more effective or there is higher risk of catching h1n1) the coefficient rate increases.

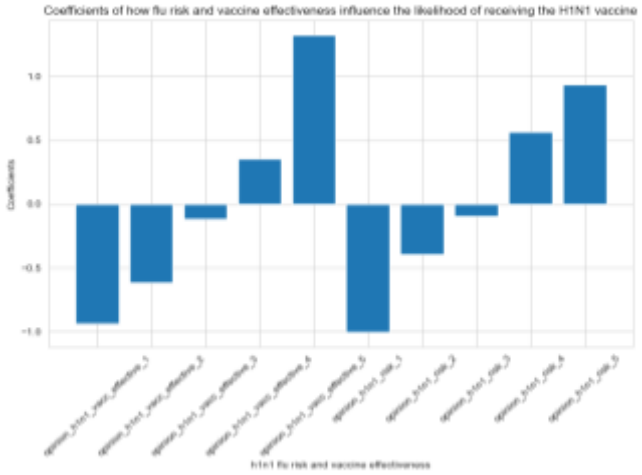

Figure 15 - Showing opinion features and their value's coefficient for H1N1

In figure 16 the results are very similar, how the stronger the opinion leads to a stronger correlation to being vaccinated.Only difference is the correlation coefficient is slightly weaker. A slight improvement Ibrahim noted was how the x axis variable labels could be vertical in order to avoid confusion between what bar belongs where.

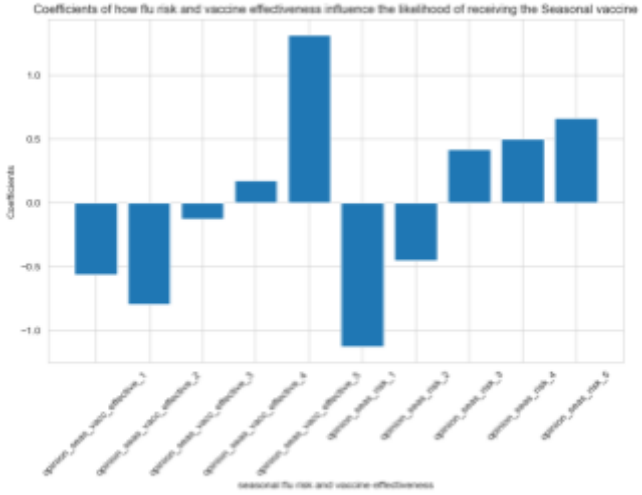

Figure 16 - Showing the seasonal flu risk and vaccine effectiveness value coefficients

DISCUSSION

Looking at the results as a whole we believe the data was cleaned and pre processed well, in terms of imputation we believe hot deck imputation turning to multiple imputation was the correct method to go for as It kept patterns, trends

and shape of the data, the only improvement were if we were to use hot deck imputation for one column, then add that column to the criteria next to try give more accuracy.

For our data analysis techniques one improvement Ibraim would have liked to make is with his legends, he intended to use specific labels detailing the vaccination status rather than 1 and 0 , however there was a struggle trying to use the same handle colour that was used for the diagram's hue. Ibrahim Also would have liked to use different visualisations rather than rotating between bar plots, pie charts and box/violin plots however there may not be a way around this as the dataset consisted of categorical and binary data. One improvement that could be made regarding Xiaonan's visualisations is labelling of x axis variables, for example initially figure 7 attempted to show both h1n1 and seasonal data at once however that provides too much of an overload and an audience of non data scientists would struggle to interpret.

When comparing classification models in general the logistic regression models proved more accurate, despite having the same level of accuracy for question 1, for question 2 and 3 the logistic regression model had significantly more accuracy.In terms of increasing accuracy for the decision tree classifier Ibrahim suggested possibly tuning hyperparameters in order to find the best parameters that yields the best accuracy.

Looking at both models we wanted to evaluate the sample size, we felt a greater sample size may have benefited the models as the precision for both variables were never high, only results siding for one outcome would be fully explored.We also wanted to evaluate the survey, considering it is carried out over the phone it explains the simplicity and lack of details in the possible answers, however if there were greater possible answers within the categories, they could provide more answers, this was shown by the length of boxplot sizes, showing the results were all very close together.

After seeing the poor accuracy from the decision tree looking back it would have been more beneficial to use a random forest classifier, this is due to the fact there is improved accuracy as it combines multiple decision trees rather than just one, there is also reduced variance , meaning the random forest classifier would not be victim to small changes in the training data unlike a decision tree, their tree structure and results can be skewed.

It would have been nice to discuss feature importance from the decision tree classifier to compare with the logistic regression classifier coefficients however this was not possible as it was not done for the decision tree, however based of the similar results from data analysis it can be assumed they would give the same pattern.

CONCLUSIONS

Overall we feel that this data analysis process provided insight into answering our research questions. This was due to the decision of using hot deck imputation which did not skew the data, the various plots for each question foreshadowing the future trends our coefficients would confirm and also presenting possible features to explore. To reiterate some improvements can be made in terms of visualisation so the data is interpretable for all audiences, especially medical staff who can use these results for research purposes.The use of Logistic Regression classification is deemed very beneficial in terms of accuracy and answering our 3 questions.For Question 1 we discovered that as levels of concern and knowledge increase it increases the chances of a record being h1n1 vaccinated hence proving there is an influence . For question 2 we can answer the question of how do different behavioural factors, such as avoidance of close contact or reduced time at large gatherings, relate to the likelihood of receiving the H1N1 and seasonal flu vaccines? We learnt that for both h1n1 flu and seasonal flu there was a greater influence from records who avoided people who showed flu symptoms, this was proved through a positive weak correlation for h1n1 and slightly stronger for the seasonal flu. There was a very weak influence for h1n1 records who avoided large gatherings, compared to the seasonal flu where there was a slightly stronger coefficient, this builds on the idea Ibrahim saw in his data analysis that the seasonal vaccine has a stronger vaccination rate. Finally the question, does the perceived effectiveness and risk of getting the H1N1 and seasonal flu vaccines impact an individual's likelihood of getting vaccinated? There is a clear showing of how the greater the perceived effectiveness or risk of getting a flu, the greater there is a correlation to being vaccinated hence answering the question as yes it does. For future work Xiaonan could include feature importance in order to fully be able to compare and evaluate our findings and be sure of our research question's answers.

For future work after assessing some of our feature relations through our EDA, for question 3 we could further explore by looking at how age and education level may affect people's opinions on the different flus.Maybe people with a higher education level may have greater insight on these flus and may want to be vaccinated? This is something that can be explored.

CONTRIBUTION

Any thing not stated was joint effort. Notebook is labelled.

Ibrahim - Interim Report, Abstract, Literature Review, conclusion, Data pre processing,

Xiaonan - Additional Research

REFERENCES

[1] DRIVENDATA, Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines, https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/

[2] Roger D. Peng & Elizabeth Matsui, The Art of Data Science, 2017-04-26

[3] J Prev Med Hyg, Logistic regression of attitudes and coverage for influenza vaccination among Italian Public Health medical residents, 12/2014
[4] Sai Sanjay A, Predicting H1N1 and Seasonal Flu : Vaccine Cases using Ensemble Learning approach