Predicting H1N1 and seasonal influenza vaccines by using data modeling and analysis methods

Yi Yuan
School of Computer Science
The University of Nottingham
Nottingham, United Kingdom
psxyy19@nottingham.ac.uk

Xinsheng Yao
School of Computer Science
The University of Nottingham
Nottingham, United Kingdom
psxxy10@nottingham.ac.uk

Abstract - In this study, we have employed multiple machine learning models to predict the likelihood of individuals receiving their H1N1 and seasonal flu vaccines. The 2009 flu pandemic in the United States caused by the novel strain of influenza A/H1N1 virus was devastating. The Center for Disease Control and Prevention estimated that between 151,700 and 575,400 people died worldwide during the first year the virus circulated. We have proposed ten machine learning models, including Linear Regression, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Artificial Neural Network, XGBoost, LightGBM, Gradient Boosting, and CatBoost. Among these models, CatBoost, Gradient Boosting, and LightGBM achieved the top three positions in terms of their performance. In summary, CatBoost demonstrated the best performance in this predicting flu vaccines problems.

Keywords—machine learning, influenza vaccine, CatBoost, Gradient Boosting

I. INTRODUCTION

On June 11, 2009, the World Health Organization decided to raise the alert level for influenza [7] A (H1N1) to the highest level. H1N1[8] is an RNA virus belonging to the family Orthomyxoviridae, with birds and some mammals as its main hosts. H1N1 is an abbreviation of the virus name, where "H" refers to hemagglutinin and "N" refers to neuraminidase, both of which are antigens on the virus. Therefore, the meaning of H1N1 is hemagglutinin type 1 and neuraminidase type 1. The pandemic was devastating, causing more than 150,000 deaths worldwide within twenty months, especially causing tremendous damage in the United States. Compared to H1N1, seasonal flu is more common and mild. It is an acute respiratory disease caused by influenza A, B, and C viruses, and belongs to the category of infectious diseases. It is mainly transmitted through droplets in the air, contact between people, or contact transmission with contaminated goods.

This dataset collected data on the intention of U.S. adults to receive vaccination at a specific point in time. The purpose of this research is to predict whether individuals would choose to receive the H1N1 and seasonal influenza vaccines under various circumstances. Researchers conducted interviews with individuals who had lived in the United States for more than six months to gather the data. In addition to questions related to H1N1 and personal information of respondents, the survey also included questions about whether they had received a seasonal flu vaccine, given the high correlation between the rejection of these two labels (H1N1 vaccine and seasonal vaccine).

The ultimate goal of this article is to improve the accuracy of machine learning algorithms used to predict which individuals are more likely to receive vaccination for H1N1 and seasonal flu. We also address other research questions that are of lesser significance. Our focus is on identifying the factors that influence people's decisions to receive or decline H1N1 and seasonal flu vaccines, as well as the barriers to vaccination uptake among different populations, such as age, race, and socio-economic status. We explore the correlations between these factors and vaccination behaviors to gain a better understanding of public health strategies.

II. LITERATURE REVIEW

Maurer, Jürgen, et al. [7] used a student’s t-test to compare the means of two groups and assess the statistical significance of uptake intentions for the novel H1N1 vaccine among individuals who had and had not received the seasonal influenza vaccine. They found that the average likelihood of being vaccinated against novel H1N1 was 49.6%. Moreover, their research indicated a strong positive relationship between the stated probability of injecting the H1N1 vaccine and having received the seasonal influenza vaccine. Specifically, individuals who had received the seasonal influenza vaccine were twice as likely to receive the H1N1 vaccine as those who had not.

Ayachit, Sai Sanjay, et al. [8], introduced multiple machine learning algorithms, such as linear regression, MlBox, XgBoost, and CatBoost, to predict the probability of whether citizens have received the H1N1 vaccine or seasonal flu vaccine. Their research used one-hot encoding to solve the problem of weight affecting categorical data and employed univariate selection to find the best features. The output of this article showed that the CatBoost algorithm achieved the highest accuracy among the algorithms they used.

Zhang and Nawata[9] utilized the LSTM classifier to predict influenza. Their study focused on four prediction algorithms: multi-stage prediction, adjusted multi-stage prediction, multiple single-output prediction, and multiple-output prediction. The results showed that the multiple single-output prediction algorithm achieved the best performance among the four algorithms.

In their study, Joshi, Aditya, et al. [10] proposed multiple deep learning-based approaches to detect vaccination behaviors. They found that using task-specific features with statistical classifiers or deep learning models with LSTM classifiers had better performance in the detection field.

III. METHODOLOGY

1. Data Pre-processing

We utilized the python programming language and the pandas package to wrangle and pre-process our dataset. The pandas package is both advanced and user-friendly. We divided the entire training feature dataset into two groups: psychological features and physiological features. The former group represents people's thoughts and behaviors regarding the flu, while the latter group represents people's personal information and status. This dataset does not contain any unordered numbers, so we did not need to deal with duplicates. We removed irrelevant features that contained mixed and disorderly words. We used various methods from the pandas library to handle missing data. We dropped the rows with more than two NaN values and the column named "health_insurance," which contained too many NaN values. For psychological features, we filled the missing values by calculating the mean of the same age group, education, race, and sex. For physiological features, we used dummy encoding[1] to handle unordered data. This method transforms categorical variables into an N*N matrix, where N is the number of different values of the data, and the matrix contains only binary variables. Using the dummy encoding method ensures that variables are not affected by weights. We assigned values to ordinal variables, such as education. These ordered variables are transformed into discrete numbers to help us analyze them easily. After data pre-processing, the dataset's size changed from 26,707 rows * 35 columns to 20,466 rows * 45 columns.

2. Data Pre-processing

For the pre-processing of the dataset, we used the Python programming language and its rich library of data analysis classes for data processing, such as NumPy and Pandas. The variables in the dataset were divided into two parts: behavioral and non-behavioral. Most of the variables in the non-behavioral part are strings, which we need to convert to multivariate numeric variables, which are numerically intuitive and better suited to most model training and prediction. By analyzing the dataset, we found that the hhs_geo_region,employment_industry,and employment_occupation variables had too many values and had little impact on the prediction of the results, so we removed these variables to reduce the impact of irrelevant variables on the prediction results. To get as much available data as possible to improve the accuracy of the machine-learning model, we used forward padding to deal with missing values. Specifically, starting with the first observation of the data, the missing values are filled with the value of that observation. The data processing resulted in a dataset of 24190 rows and 35 columns by the above method.

3. Linear Regression

Linear regression[2] is a crucial statistical modeling technique that focuses on the relationship between independent and dependent variables in a linear field. The following graph shows the multivariable linear regression model[3] which used in this study to predict two labels: h1n1_vaccine and seasonal_vaccine.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$

Figure 1. Multivariable linear regression model

In Figure 1, y represents the dependent variable, and each x represents an independent variable. Each β is the parameter corresponding to an independent variable, and ε is the residual of the function, which is the distance between the real and predicted value.

In this article, we need to predict two labels which are h1n1_vaccine and seasonal_vaccine. Therefore, we built two linear regression models to predict these two values respectively. The dataset contains forty-three columns of numerical variables. However, most of variables are not related to these two labels. We employed correlations between each feature and two labels, and we only remained the value of correlations which are larger than 0.15. For selected variables, we regularized them through sklearn library. The aims of regularization[4] are balancing complexity and performance as well as to avoid overfitting as much as possible. Rider regression[5], an advanced multivariable linear regression method, is used to build the model. This estimation method can address the collinearity and overfitting problems which happened in traditional linear regression frequently. Here is the function of each parameter:

$$\widehat{\beta}(k) = (\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}\mathbf{x}'\mathbf{y},\ k \geq 0,$$

Figure 2. Parameter in rider regression model

In Figure 2, we can calculate the value of each parameter, and I is the identity matrix. By using all the parameters gained from the function to the test features, we can predict the values of the test labels.

4. Artificial Neural Networks

Artificial neural networks[6] are computer technologies based on biology, designed to simulate the working process of the human brain. Rather than using programming to train the model, artificial neural networks detect patterns and relationships of each unit to learn epoch by epoch. Compared with traditional algorithms, artificial neural networks algorithms have higher model complexity. What's more, it also has more robust data processing ability and model generalization ability to process high-dimensional and data with large scale as well as adaptability of new data. It is an operational model composed of a vast number of nodes that relate to each other. Each node represents a specific output function called the activation function. The memory of the artificial neural network is a weighted value between any two nodes.

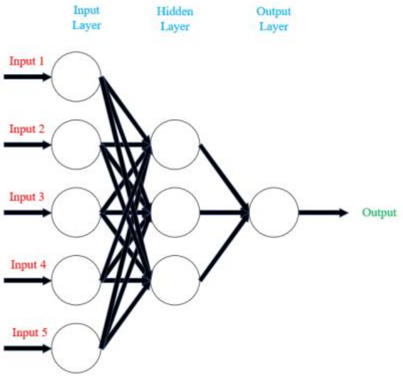


Figure 3. The example structure of artificial neural network

Figure 3 depicts the neural network architecture used in this study. The left layer is known as the input layer, and the neurons within this layer are called input neurons. The right layer is called the output layer, containing output neurons, with only one neuron in this example. The middle layer is referred to as the hidden layer, with its neurons fully connected to the input layer. Each line connecting neurons represents the functions with different weights.

In this article, we divided the dataset into training, validation, and testing sets, and created a sequential model based on the Keras library. The architecture of this model consists of an input layer with forty-one effective features as neurons, two dense layers, and an output layer. The first dense layer includes sixty-four neurons and uses the Rectified Linear Unit (ReLU) activation function, while the last dense layer has neurons with half of the first dense layer, and it also employs the ReLU activation function. The output layer has two neurons since this task is a binary classification problem, predicting the probability status (vaccinated or not vaccinated) of two categories. This layer uses the sigmoid activation function. The ReLU function and the sigmoid function are defined as Figure 4 and Figure 5.

$$g(x) = \max\{0, x\} = \begin{cases} 0 & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases}$$

Figure 4. Rectified Linear Unit function

$$g(x) = \left(1 + \operatorname{erf}\left[\frac{x}{\sqrt{2}}\right]\right) = 2\int_{-\infty}^{x} dz \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

Figure 5. Sigmoid function

5. Decision Tree

Decision tree is a traditional supervised learning method used for classification. Each sample has a set of attributes and a known classification result. By learning from these samples, a decision tree model can be obtained and used to provide classification for new data. Decision tree has a tree structure, and each decision is made at a leaf node for classification, with each leaf node representing a classification model that may be binary or multi-class. In this article, a decision tree classifier was used to train the features and two labels.

6. Random Forest

Random forest is an ensemble learning algorithm based on decision tree. Ensemble learning achieves the goal of learning by integrating multiple learners. Random forest uses bootstrap sampling to randomly select n samples from the original dataset to create a new training data subset. This new subset of data is used to train a new decision tree model. To address classification problems such as predicting vaccine injection, random forest selects the majority classification result as the final output.

7. Support Vector Machine (SVM)

Support vector machine[11] is a classic linear classifier that maps feature vectors of instances to points in space. The objective of SVM is to draw a line that can best distinguish the two types of points so that if new points are added, this line can also perform well in classification. Its decision boundary is the maximum hyperplane of learning sample solution. It is suitable for classification problems with small or medium-sized data samples, non-linear, and high-dimensional data.

8. Gradient Boosting Machine (GBM)

Gradient Boosting Machine algorithm is a type of Boosting algorithm that combines multiple weak learners with high bias. The main idea is to use different weights to reduce the overall bias and form a strong learner.

9. Logistic Regression

Logistic regression is an efficient and powerful way to analyze the effect of a group of independent variables on a binary outcome by quantifying each independent variable’s unique contribution [12]. It works by converting a combination of input features and a linear function to the predicted values of the H1N1 vaccine and seasonal vaccine for this problem and mapping the results to probabilities between [0,1] using a sigmoid activation function. The following equation for the sigmoid function is given:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

In the Equation, Z is the predicted value of the output of a combination of linear functions. When Z tends to positive infinity, $\sigma(z)$ tends to 1, indicating a high probability of belonging to a category. Conversely, when Z tends to negative infinity, $\sigma(z)$ tends to 0, indicating a low probability of belonging to a category.

The analysis of the dataset and the prediction problem revealed the following features. (1) The output of the problem is the probability of belonging to a binary classification problem, such as h1n1_vaccine and seasonal_vaccine. (2) Most of the feature values in the dataset are numeric and most of the non-numeric data can be converted to numeric. (3) Both the number of features and the sample size of the dataset are small, with the number of features less than 100 and the sample size of the dataset less than 50,000. (4) The features are independent of each other or only weakly correlated. The above characteristics are extremely significantly in line with the requirements of logistic regression models, thus we try to use logistic regression models to solve the problem.

In this problem, we demand to train models for the two labels h1n1_vaccine and seasonal_vaccine for the prediction of the outcome values respectively. We divide the pre-processed dataset into a training set and a test set in the ratio of 8:2 for training and testing of the model. The model was also trained using the machine-learning functions provided in the Scikit-Learn library. The choice of hyperparameters for the logistic regression model is decisive for the performance of the model, such as regularization parameters, learning rate, number of iterations, etc. To obtain the extreme prediction results of the model, we must find the optimal hyperparameters for model training. Therefore, We use a combination of stochastic search and cross-validation to obtain the optimal combination of parameters. Random search is a hyperparameter optimization technique where a random sample is taken within a given range of hyperparameters and the best-performing combination of hyperparameters is selected. Cross-validation is a method to evaluate the performance of the model for each hyperparameter combination in the model and to ensure that the model is not overfitted on the training set. By using the above method, we can acquire the best logistic regression model for predicting the results based on this dataset.

10. Gradient Boosting Decision Tree

Gradient boosting decision trees[15] (GBDT) is a combination of gradient boosting and decision tree working principles. The basic principle of a decision tree is to partition the data set into smaller subsets from the top down, and then repeat the operation for each subset until some termination condition is met. The fundamental concept of gradient boosting is to fit the model iteratively, constantly updating the parameters in the model so that the error of the model on the training set is gradually reduced. The gradient boosting decision tree is therefore an integrated learning algorithm, based on the decision tree boosting method. It works by combining multiple decision trees and training a series of tree models through iterations, with each new tree model attempting to compensate for the error in the previous model, resulting in a smaller and more accurate model after a certain number of iterations:

$$F_t(x) = F_{t-1}(x) + \gamma_t h_t(x) \tag{2}$$

$$r_t = y - F_{t-1}(x) \tag{3}$$

In the Equation, $F_{t-1}(x)$ denotes the prediction of the first t-1 trees, $h_t(x)$ denotes the prediction of the t tree, and γ_t denotes the weight of the t tree. In each iteration, we calculate the residual γ_t and use it as the target variable for the t tree. y denotes the true value of the sample.

The gradient decision tree is an excellent algorithmic idea. However, traditional gradient decision tree models still have some drawbacks, such as high computational complexity, easy overfitting, sensitivity to outliers and sample imbalance. To eliminate these defects, we have adopted improved models based on gradient decision trees, such as CatBoost, XGBoost and LightGBM, which have been optimized in several aspects to make the models more robust and improve the computational capabilities significantly. It tackles the shortcomings of traditional models well and provides more accurate prediction models. Therefore, we have selected these three models and performed predictions for each of the datasets and compared the results between the three models to extract the best-performing model. They are all based on the principle of gradient decision tree algorithm. Figure 6 shows the relationship between these four algorithm models.

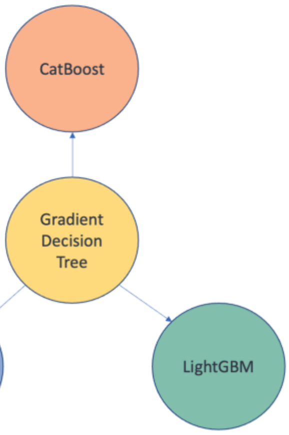


Figure 6. Schematic diagram of the relationship between CatBoost, XGBoost, LightGBM and Gradient Boosting

10.1 CatBoost

The main advantages of the CatBoost[14] are the usage of Ordered Boosting techniques to speed up training, the ability to automatically handle category-based features, missing values and outliers, and the use of techniques such as symmetric trees and weighted quartiles to improve the generalization performance of the model. To avoid overfitting problems, an L2 regularization-based approach is introduced to control the complexity of the model and an Early Stopping technique, which stops training when the model performance starts to degrade. In summary, CatBoost is a highly adaptive machine learning algorithm that can assist us in reducing errors during data pre-processing and improving training efficiency. Hence, we omit the data pre-processing part when training with the model and let the model help us process and populate the data automatically. To further optimize the model, we also applied a random search and cross-validation method of tuning hyperparameters on its foundation to obtain the best parametric model. The model ended up with good results in the prediction of the dataset. The following figures are based on the prediction model to draw the receiver operating characteristic curve (ROC). TPR stands for True Positive Rate and FPR stands for False Positive Rate.

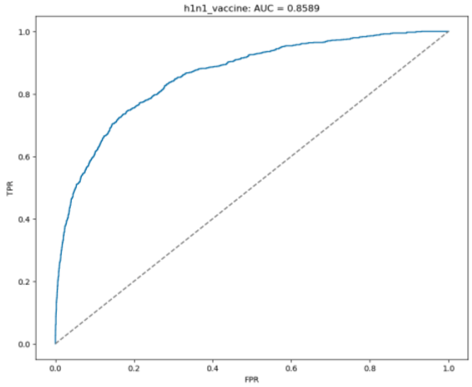


Figure 7. Prediction model of h1n1 vaccine

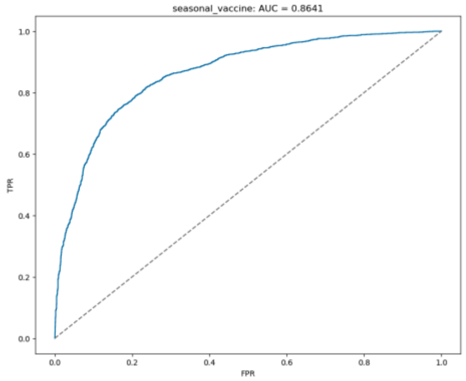
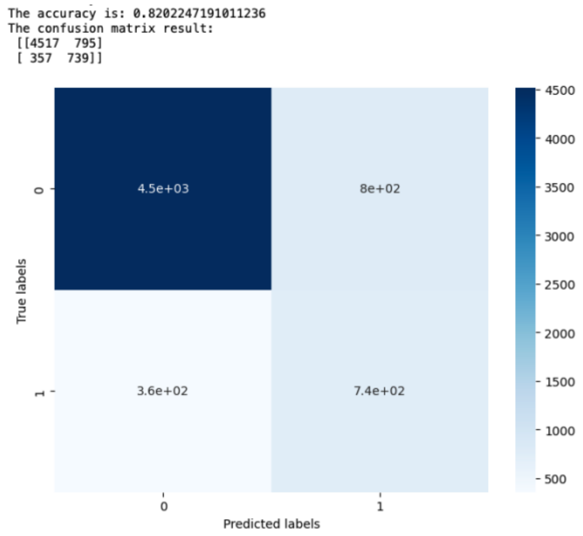


Figure 8. Prediction model of seasonal vaccine

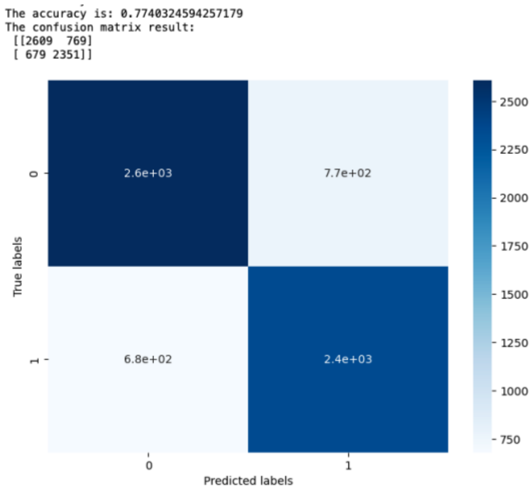
10.2 XGBoost

The core features of the XGBoost[13] are high-level accuracy and scalability, powerful parallel processing, and high forecasting speed. The model achieves efficiency, flexibility, and lightness by making the loss function more accurate through second-order derivatives, which can be computed in parallel through Block storage. However, the use of XGBoost requires data pre-processing of the dataset in advance, such as missing value padding, data type conversion and feature value extraction. For this reason, we decided to train and predict the model based on our data pre-processing method, to observe the prediction results and make a comparison with the CatBoost and LightGBM models, in which we also can verify the suitability of our pre-processing method. The figures below are the result of visualizing the confusion matrix for the model's heat map.



```
The accuracy is: 0.8202247191011236
The confusion matrix result:
 [[4517  795]
 [ 357  739]]
```

Figure 9. Heat map of h1n1 vaccine prediction model



```
The accuracy is: 0.7740324594257179
The confusion matrix result:
 [[2609  769]
 [ 679 2351]]
```

Figure 10. Heat map of seasonal vaccine prediction model

10.3 LightGBM

The LightGBM has the benefit of overcoming the problem of GBDT's low efficiency when dealing with large-scale data. It utilizes two unique mechanisms to enhance the efficiency of the algorithm. The first one is a histogram-based decision tree splitting approach, which divides the feature values into different intervals and calculates their gradient information, reducing the number of sorting and computations. The second is a leaf-wise growth strategy with a depth constraint, which prioritizes the growth of trees with less depth in the leaf nodes to reduce the depth of the tree and reduce the risk of over-fitting. We have chosen to use this model with the same data pre-processing method as XGBoost for training predictions and compare the results with the CatBoost model which does not require data processing. We can explore the prediction efficiency of the three principal machine learning models based on the gradient decision tree in this dataset. The figures below are the receiver operating characteristic (ROC) curve drawn according to the model results, where TPR stands for True Positive Rate and FPR stands for False Positive Rate.

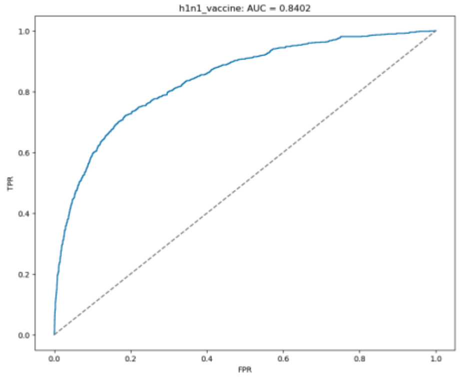


Figure 11. Prediction model of h1n1 vaccine

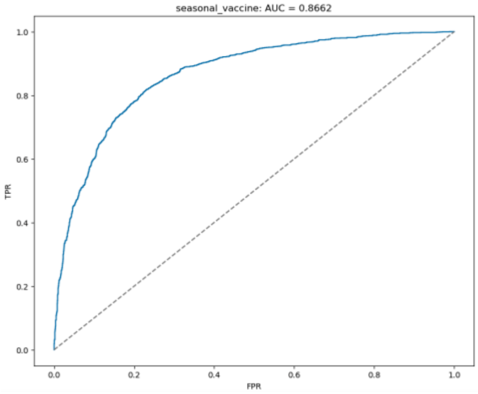


Figure 12. Prediction model of seasonal vaccine

IV. RESULT

1. Model results

Based on practice with the algorithms mentioned in the Methodology section, it was concluded that the best-performing algorithm was CatBoost and the worst-performing algorithm was Linear Regression. The Ensemble Model algorithms were more efficient than the Single Model algorithms. The following figures below show the accuracy(Table 1) and visualization results(Figure 13) of all used algorithms.

Algorithm	Accuracy	Model Type
CatBoost	0.8591	Ensemble
Gradient Boosting	0.8354	Ensemble
LightGBM	0.8344	Ensemble
XGBoost	0.8311	Ensemble
Artificial Neural Network	0.8290	Single
Support Vector Machine	0.8267	Single
Random Forest	0.8208	Ensemble
Decision Tree	0.8107	Single
Logistic Regression	0.7958	Single

Table1. Algorithm Result

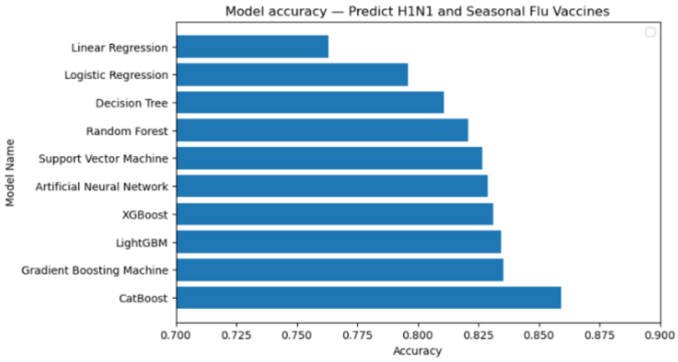


Figure 13. Model Accuracy in Predict H1N1 and Seasonal Flu Vaccines

2. Correlation Result

The following conclusions were drawn from the results of the correlation between features and vaccines(Figure 14). Among all the features, the ones with a greater impact on the vaccine are opirion_seas_vacc_effective,opinion_seas_risk,doctor_recc_seasonal, doctor_recc_h1n1 and opinion_h1n1_risk.

Those with a positive impact are doctor_recc_h1n1 and opinion_h1n1_risk, with doctor_recc_h1n1 having the greatest positive impact. Opirion_seas_vacc_effective,opinion_seas_risk and doctor_recc_seasonal had a negative impact on the vaccine, with opinion_seas_risk having the greatest negative impact. The following figure shows the correlation between the features and the vaccine.

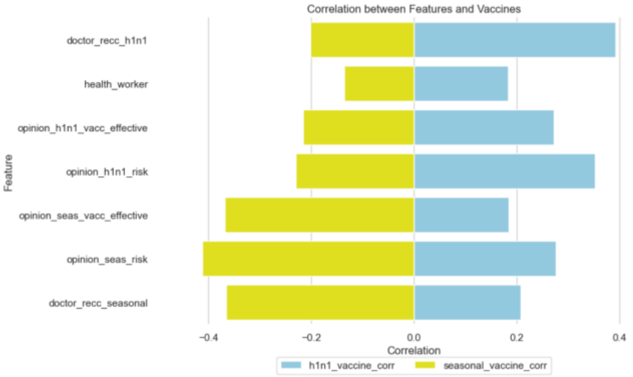


Figure 14. Correlations between Features and Vaccines

3. Evaluation metric

Accuracy is an excellent metric for evaluating the prediction capability of a model and measures the ratio of the number of correct classifications of a model to the total number of samples. It assists us in comparing the predictive accuracy of different models in the same dataset directly. Accuracy can be defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

In the formula, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. Higher Accuracy values represent the better prediction accuracy of the model. Based on Accuracy, we concluded that the best algorithmic model to solve this problem is CatBoost.

V. DISCUSSION

Xinsheng Yao has pointed out that the main weakness of Yi Yuan's work is the data preprocessing stage. Specifically, Yi's approach to categorical data such as sex or race involves converting them directly into multivariate numeric variables. This approach can lead to unreasonable weights, which can compromise the accuracy of the resulting model. In addition, Yi's use of forward padding to handle missing values can result in increased computational costs and a high degree of inaccuracy. However, Yi Yuan's research proposes a model that achieves the highest accuracy using CatBoost, an advanced boosting algorithm that automatically handles categorical features, missing values, and outliers. This method employs L2 regularization to prevent overfitting, which can improve the generalizability of the model. Furthermore, Yi Yuan's model incorporates random search and cross-validation methods to further enhance accuracy, ensuring that the model is both robust and accurate. Overall, Yi Yuan's work achieved higher accuracy and used more advanced machine learning algorithms, but with a potential limitation due to the method of data preprocessing used. Xinsheng Yao proposed a better data preprocessing approach that may lead to more accurate and robust models.

The aim of this research is to find the best algorithm to solve this problem. After experimenting with different algorithms, we finally found the best algorithm is CatBoost. In this study, Xin Sheng Yao used a traditional machine learning approach, while Yi Yuan used an ensemble learning approach. Xin Sheng Yao's method has a variety of data pre-processing, not only for missing values but also for exploring the relevance of feature values and selecting those with strong relevance for training on different machine learning models. This greatly enhances the efficiency of model training and allows for fast prediction results. However, it also screens out the impact of subtle feature values on the overall results, resulting in a model that is not as accurate as expected. The ensemble learning approach used by Yi Yuan automatically processes and transforms the eigenvalues for the dataset, reducing the impact of human filtering on the prediction results while constantly searching for the best combination of hyperparameters to achieve the best performance. This method is time-consuming and requires large amounts of memory, CPU, and other computing resources. The parameter combinations obtained by the random search method and the cross-validation method are only relatively close to the optimal parameters, hence they are difficult to reproduce. The methods mentioned in the literature review for finding hyperparameter combinations are more efficient. Therefore, this field could be further extended and optimised, for example by trying the grid search method to find the optimal parameters. In addition, a combination of LSTM and other deep learning algorithms will also give better prediction results, while these methods will take longer to train the model and require more resources in terms of memory, CPU, GPU, etc. In conclusion, Xinsheng Yao's method is diversified in terms of data pre-processing, shorter time to obtain model results and limited accuracy. Yi Yuan's method does not require much processing of features to achieve more accurate results. But it consumes more computer resources and takes longer to train the model.

VI. CONCLUSION

In this study, we utilized ten machine learning algorithms to build models for predicting the likelihood of individuals receiving the H1N1 and seasonal flu vaccines in the United States in 2009. The CatBoost algorithm achieved the best performance, with an accuracy of 0.8591. Therefore, we conclude that the model using the CatBoost algorithm is the most effective, and it can be used to predict future data in this category. Furthermore, we can apply these algorithms which we used in this article in future research with similar fields.

REFERENCES

1. Dahouda, Mwamba Kasongo, and Inwhee Joe. "A deep-learned embedding technique for categorical features encoding." IEEE Access 9 (2021): 114381-114391.

2. Su, Xiaogang, Xin Yan, and Chih-Ling Tsai. "Linear regression." Wiley Interdisciplinary Reviews: Computational Statistics 4.3 (2012): 275-294.

3. Uyanık, Gülden Kaya, and Neşe Güler. "A study on multiple linear regression analysis." Procedia-Social and Behavioral Sciences 106 (2013): 234-240.

4. Bickel, Peter J., et al. "Regularization in statistics." Test 15 (2006): 271-344.

5. McDonald, Gary C. "Ridge regression." Wiley Interdisciplinary Reviews: Computational Statistics 1.1 (2009): 93-100.

6. Agatonovic-Kustrin, S., and Rosemary Beresford. "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research." Journal of pharmaceutical and biomedical analysis 22.5 (2000): 717-727.

7. Maurer, Jürgen, et al. "Does receipt of seasonal influenza vaccine predict intention to receive novel H1N1 vaccine: evidence from a nationally representative survey of US adults." Vaccine 27.42 (2009): 5732-5734.

8. Ayachit, Sai Sanjay, et al. "Predicting h1n1 and seasonal flu: Vaccine cases using ensemble learning approach." 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE, 2020.

9. Zhang, J., and K. Nawata. "Multi-step prediction for influenza outbreak by an adjusted long short-term memory." Epidemiology & Infection 146.7 (2018): 809-816.

10. Joshi, Aditya, et al. "Shot or not: Comparison of NLP approaches for vaccination behaviour detection." Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. 2018.

11. Chauhan, Vinod Kumar, Kalpana Dahiya, and Anuj Sharma. "Problem formulations and solvers in linear SVM: a review." Artificial Intelligence Review 52.2 (2019): 803-855.

12. Stoltzfus, Jill C. "Logistic regression: a brief primer." Academic emergency medicine 18.10 (2011): 1099-1104.

13. Al Daoud, Essam. "Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset." International Journal of Computer and Information Engineering 13.1 (2019): 6-10.

14. Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support." arXiv preprint arXiv:1810.11363 (2018).

15. Wang, Jidong, et al. "A short-term photovoltaic power prediction model based on the gradient boost decision tree." Applied Sciences 8.5 (2018): 689.