School of Computing
FACULTY OF ENGINEERING AND
PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Finnal Report

<FluVaxML: A machine learning-based model for predicting H1N1 and seasonal flu vaccine uptake>

Xiuping Ouyang

Submitted in accordance with the requirements for the degree of

BSc Computer Science with Mathematics）

2022/23

COMP3931 Individual Project

The candidate confirms that the following have been submitted.

Item	**Format**	**Recipient(s) and Date**
Final Report	*PDF*	Uploaded to Minerva (09/05/2023)
code repository	shared access	Sent to supervisor and asses sor (09/05/2023)

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

Xiuping Ouyang

Summary

Influenza vaccines are the best way to respond to the severe global pandemic of seasonal and H1N1 influenza that began in 2009. However, there is a serious lack of research explaining the decisions of predictive models in predicting vaccination behaviour. In this sense, we designed a project to predict the likelihood of vaccination by analysing the performance of residents and their initial behaviour. Subsequently, we constructed multiple machine learning models and compared them to select a prediction model with cutting-edge results, where the area under the curve (AUC) of the xgboost-based H1N1 influenza vaccine prediction model was 0. 89.

To further facilitate human-artificial intelligence collaboration, we propose a variety of methods to explain our prediction model decisions, such as using a game-theoretic based framework (SHAP) that attempts to explain our non-linear prediction model in a linear fashion, where we can analyse the resulting results and judge the importance of features based on visualisations. With further refinement of the predictive model, this model can be further used for medical engineering follow-up and government policy adjustment to ensure a transparent process for key stakeholders (government, healthcare providers).

Acknowledgements

I would like to thank my supervisor Netta for her guidance and feedback during the session. I would also like to thank my parents for their support.

Catalogue

Chapter 1

Background Research

1.1Introduction

Influenza vaccination is critical globally for the prevention and control of seasonal influenza and H1N1 influenza epidemics. To ensure universal access to influenza vaccine, several recent studies have proposed methods for early prediction of vaccination, however, currently available models on predicting influenza vaccination are static analyses on existing data or similar analyses using historical data, such as using past injections in the same region, age and gender(Brown et al., 2018). However, this non-dynamic approach does not provide sufficient sustainability for the development of the models and the results obtained are not sufficiently theoretical to allow for further upgrading over time. It also has significant shortcomings in terms of predictive power and interpretation. This has led to significant challenges for public health departments and relevant policy makers in developing targeted vaccine roll-out strategies. To better address this issue, new prediction models still need to be developed to improve prediction accuracy and provide reasonable explanations for model decisions.

This study will look at the limitations of existing prediction models and explore how prediction methods can be improved with a view to achieving better results in predicting H1N1 influenza vaccination and seasonal influenza vaccination. We will compare different prediction models and delve into the differences in their performance in predicting influenza vaccination status. In addition, we will also investigate how modern machine learning techniques, such as the game theory framework (SHAP), can be used to improve the interpretability of the models, thereby helping policy makers and healthcare professionals to better understand the basis for the prediction results.

Through this study, we hope to provide new ideas for improving existing influenza vaccination prediction models and provide a scientific basis for influenza vaccination policy development, ultimately improving influenza vaccination coverage and reducing the impact of influenza epidemics on global population health.

1.2Influenza

1.2.1 Seasonal influenza

Seasonal influenza is a respiratory infection caused by influenza viruses that usually circulate in the fall and winter. There are many subtypes of seasonal influenza viruses, the most common of which are influenza A and B viruses(Galanis,2023). The virus is spread by airborne droplets and causes fever, cough, sore throat, headache, muscle pain and other symptoms when it infects people. In some populations, such as the elderly, children, pregnant women and people with chronic illnesses, infection with the flu virus can lead to more serious complications and even life-threatening illnesses(Liu,2923).

1.2.2 H1N1Influenza

H1N1Influenza, also known as swine flu, is a disease caused by the influenza A (H1N1) virus. H1N1 is an RNA virus belonging to the family Orthomyxoviridae. H1N1 is similar to seasonal influenza, but is more contagious and the virus mutates more rapidly. Among the developed swine influenza vaccines, the most technically mature and used in production are mainly monovalent or bivalent inactivated swine influenza whole virus vaccines of the H1NI and H3N2 subtypes(Li,2009) . Although inactivated swine influenza whole virus vaccines have been reported to be safe, efficient, low cost of production and long duration of effective antibodies, they also have disadvantages such as slow antibody production, weak ability to induce cellular immunity and strong stress response. In particular, most vaccine strains are derived from circulating strains. Once the virus is leaked during vaccine preparation, it can easily cause environmental contamination and induce a new outbreak[Jain,2009] .

Chapter 2

Methodology

2.1Integrated learning algorithms

2.1.1 Principle of the integrated learning algorithm

In supervised learning, ensemble methods combine weak classifiers to achieve a better, more comprehensive model (Zhang, Li, and Yang, 2012). Divided into sequential (e.g., AdaBoost) and parallel (e.g., Random Forests) approaches, ensemble learning leverages base learner dependencies or independence to improve accuracy (Breiman, 2001). Key features include combining multiple methods, superior performance compared to individual classifiers, and decision-making similar to multiple decision-makers.

2.1.2Bagging Random Forest

Random forests use several decision trees to improve classification and regression outcomes (Breiman, 2001). Randomly selecting training data and features at each node improves decision tree performance. Randomness reduces overfitting and increases generalisation.

Bagging uses bootstrap aggregation to randomly pick many subsamples with replacement from the original dataset and train a base learner on each training set. Averaging (regression) or voting (classification) all base learners yields the prediction. Bagging and random forests simply reduce variance by merging several base learners, not bias (Breiman, 1996).

2.1.3Boosting integrated learning models

Boosting integrated learning models are often used because they improve classification using shallow learning approaches and are thus among the most popular ensemble classifiers. AdaBoost is one of the best booster classifiers available today. The goal is to combine poor and strong learners in order to increase the efficiency of the learning paradigm. AdaBoost is a sample-based training technique in which weights are incrementally added to failing classifiers. By giving more weight to the misclassified samples and less to the properly labeled ones, a new weak learning classifier may be derived from the reweighted data. This process is performed as many times as necessary until all samples have been properly labeled. Boosting is an integrated learning approach with the primary goal of improving the performance of weak classifiers. It is feasible to strengthen weak

classifiers inside a PAC (probably roughly correct) learning paradigm(Uddin, 2022).

In the next part, we will discuss many widely used Boosting integrated learning models:

（1）Gradient Boosting Decision Tree (gbdt algorithm)

GBDT stands for Gradient Boosted Decision Tree and GBDT is arguably one of the best machine learning algorithms.

A Gradient-Boosted Decision Tree is a decision tree model that prevents overfitting and demonstrates improved prediction accuracy. In a gradient-augmented decision tree, it is presumed that F(x) is an approximation of the output y given a set of input variables x. The squared error function is used as the loss function L in Equation 2.5 to estimate the approximation function. (Uddin, 2022).

$$L(y,F(x))=[y-F(x)]^2 \tag{2.5}$$

Assume that each regression tree has a number of partitions J. Each regression tree divides the input space into J disjoint regions R1m, ..., Rjm and predicts a constant value bjm for region Rjm. In this case, each decision tree behaves in an additive form as shown in Equation 2.6.

$$h_m(x)=\sum_{j=1}^{j} b_{jm} I(x\in R_{jm}),\ \mathrm{I}=\begin{cases}1 & if\ (\mathrm{x}\in R_{jm})\\ 0 & \text{其他}\end{cases} \tag{2.6}$$

Using the training data, the gradient augmentation model iteratively constructs M decision trees $\mathrm{h}_1(x)$ ..... $\mathrm{h}_M(x)$ 。 Updating the approximation function $F_m(x)$ and gradient descent step ρ_m can be defined by equation 2.7 and equation 2.8.

$$F_m(x)=F_{m-1}(x)+\sum_{j=1}^{J}\gamma_{jm} I(x\in R_{jm}) \tag{2.7}$$

$$\rho_{\mathrm{m}}=\arg\min\sum_{i=1}^{N} L(y_i,F_{m-1}(x_i)+\rho\sum_{j=1}^{J} b_{jm} I(x\in R_{jm})) \tag{2.8}$$

$$F_m\left(x\right)=F_{m-1}\left(x\right)+\xi\cdot\sum_{j=1}^{J}\gamma_{jm} I\left(x\in R_{jm}\right),0<\xi\le 1 \tag{2.9}$$

In equation 2.9， The smaller the valueξ, the higher the learning rate. With the learning rate strategy, the overfitting problem can be avoided by reducing the influence of additional trees.

Key advantages of GBDT:

1) Flexibility in handling various types of data

2) High prediction accuracy

3) Some robust loss functions, such as huber, are used to handle outliers well

Disadvantages of GBDT:

Difficult to parallelise due to dependencies between base learners, but partial parallelism can be achieved by subsampling SGBT (Uddin, 2022).

（2）XGBoost

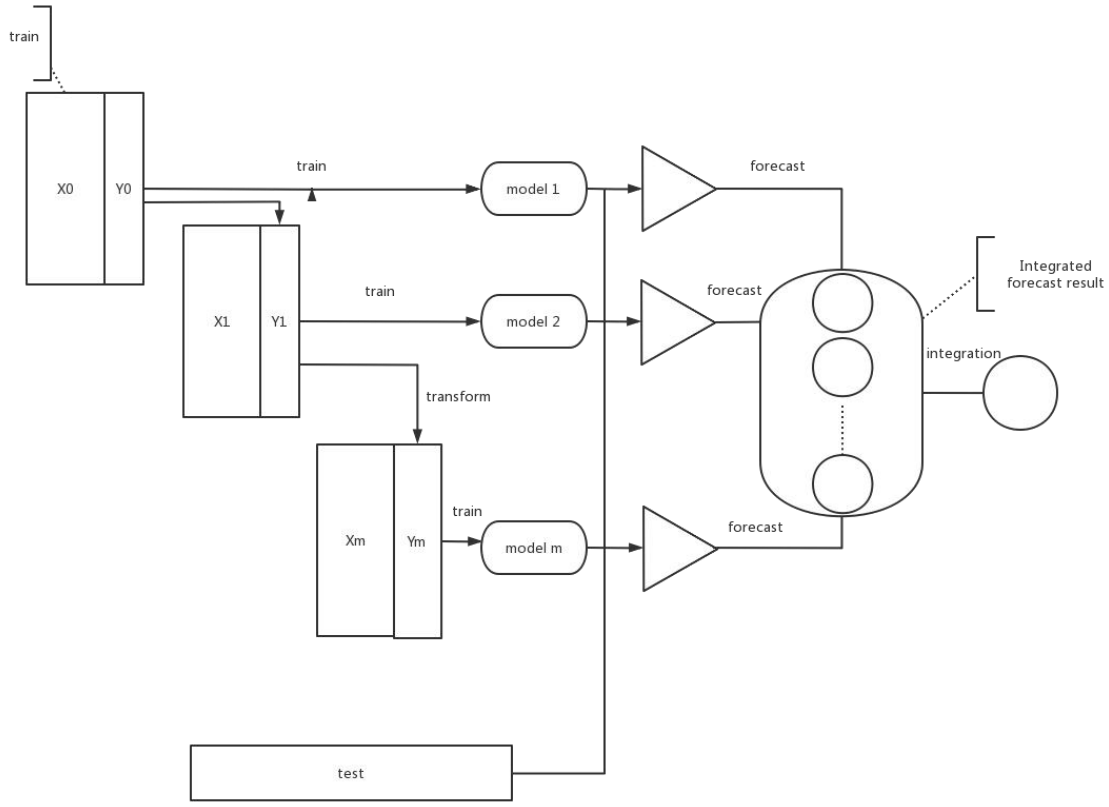
XGBoost was proposed by Dr. Tianqi Chen of the University of

Washington(Chen & Guestrin, 2016). It was used in Kaggle's Higgs subsignal recognition competition and has received a lot of attention for its excellent efficiency and high prediction accuracy.

The boost algorithm is called gradient boost if the weak prediction model in each step of the boost algorithm is generated according to the direction of the gradient of the loss function. the XGBoost algorithm uses a stepwise additive model but no longer needs to compute the coefficients after the weak learner is generated in each iteration. The model form is as follows:

$$F_T(X) = \sum_{m=0}^{T} f_m(X) \qquad 2.12$$

The XGBoost algorithm achieves weak learner generation by optimising the structured loss function (adding regular terms to the loss function reduces the risk of overfitting). instead of using search methods, XGBoost uses the first and second order derivatives of the loss function directly, and greatly improves the performance of the algorithm by pre-sorting, weighted quantile and other techniques. generation of XGboost trees The algorithm flowchart is shown in Figure 2.2:



2.1.2.1.1Flow Chart of XGboost Tree Generation Algorithm

（3）LightGBM

LightGBM is an algorithm based on GBDT. Because traditional GBDT requires traversing all training data multiple times during each iteration, GBDT training will be memory constrained. Particularly, GBDT cannot meet its requirements when dealing with enormous data of industrial quality. lightGBM facilitates efficient parallel training with faster iterations, reduced memory utilisation, and improved

precision, as well as rapid distributed processing of massive amounts of data (Ke et al., 2017).

2.2Hyperparameter optimisation methods

All machine learning models must go through the process of hyperparametric optimisation. The classification accuracy of 90% in previous training models, or what they perceive to be sufficient accuracy, is insufficient. In order to succeed in the Kaggle competition, it is essential to have both excellent features and the finest hyperparameters. Especially when utilising integrated learning techniques such as Random Forest and XGBoost, the hyperparameters have a significant impact on the model's efficacy.

2.2.1 Principle of Bayesian optimization algorithm

Utilising probabilistic models with multivariate interactions is an efficient method for resolving boundedly challenging problems. The Bayesian Optimisation Algorithm (BOA) is a distribution estimation algorithm that uses a Bayesian network to model the data. The Bayesian Optimisation Algorithm (BOA) constructs and samples Bayesian networks in order to generalise candidate solutions. BOA can be applied to black-box optimisation problems where candidate solutions are represented by fixed-length strings over a finite alphabet; however, for the sake of simplicity, only binary strings are considered in this section.

BOA generates an initial padding of strings with a random distribution over all conceivable strings. Each of the multiple iterations (generations) of the overall consists of four phases. Initially, plausible solutions are chosen from the current population utilising genetic algorithm selection techniques, such as race or truncated selection. Then, a Bayesian network is constructed to accommodate the population of promising solutions. Then, by sampling the constructed Bayesian network, new candidate solutions are generated. Finally, the new candidate solutions are added to the original population, potentially replacing some or all of the old solutions.

The four preceding stages are repeated until specific termination criteria are met. For instance, the run can be terminated when the overall population converges to a single case, when the overall population contains a solution of sufficient quality, or when a maximum number of iterations is reached. Figure 4 depicts the pseudo-code for the Bayesian Optimisation Algorithm (BOA）(Pelikan et al., 1999).

```
Bayesian optimization algorithm (BOA)
   t := 0;
   generate initial population P(0);
   while (not done) {
      select population of promising solutions S(t);
      build Bayesian network B(t) for S(t);
      sample B(t) to generate O(t);
      incorporate O(t) into P(t);
      t := t+1;
};
```

Figure 1.1Pseudocode of Bayesian Optimization Algorithm (BOA)

2.2.2 Principle of grid search and cross-validation

(1) Principle of grid search

The fundamental principle of grid search is to divide each interval of parameter values into a series of cells, calculate the target values (typically errors) determined by the combination of the corresponding parameter values in order, and select the best value one by one, so as to obtain the minimum target value and the optimal parameter value in the interval. This method assures that the search solution is optimal or near-optimal on a global scale and prevents significant errors. (Hutter, Hoos, & Leyton-Brown, 2011).

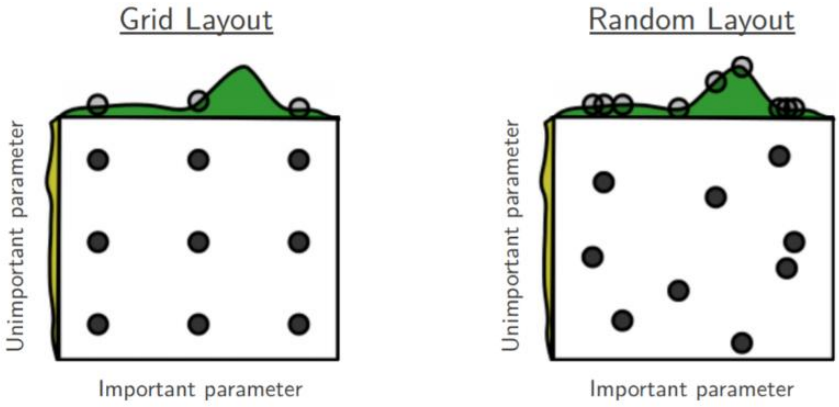


Figure 1.2Principles of Grid Search

(2) Principle of cross-validation

Cross-validation (CV) methods are used in machine learning algorithms to utilise data when the quantity of data is insufficient. Cross-validation, as its name implies, is the reuse of data (Kohavi, 1995). The data is divided into a training set, a validation set and a test set. The training set is used to train n models with n randomly selected sets of data. The test set evaluates the n models and chooses the best one.

2.3Machine learning model interpretability methods

2.3.1 Theoretical approaches to interpretability

Understanding a model's fundamentals is as crucial as its accuracy in many

applicable investigations. The best accuracy of current huge datasets is typically attained using complicated models that are difficult to comprehend even for specialists, such as ensemble or deep learning models, creating a conflict between accuracy and interpretability (Gilpin et al., 2018). Thus, professionals in adjacent domains have developed many approaches to assist users evaluate complicated model predictions, but it is frequently unclear how these strategies are connected and when one is better than another.

Last decade saw major developments in artificial intelligence. Controlled and researched, it may be beneficial in many applications. The AI community is struggling with interpretability (XAI) to satisfy expectations as fast as feasible. XAI is necessary for real-world AI models. Given that XAI looks to be one of the AI community's important future research topics, it is appropriate to examine what has been accomplished with XAI and concentrate on its future potential. Thus, AI researchers should concentrate on trustworthy AI ideas by studying XAI and adopting AI technologies at scale in organisations that value explanatory power, fairness, and accountability). Efficient deep learning algorithms in a huge parameter space with hundreds of layers and millions of parameters have helped machine learning models succeed recently. Deep learning models are complicated black boxes (Gilpin et al., 2018). Stakeholders in the AI business are increasingly demanding openness in crucial situations where black-box machine learning algorithms make vital predictions. Thus, AI conclusions may be unreasonable, illegal, and unjustified. Model output interpretation is crucial (Arrieta et al., 2020).

2.3.2Interpretable methods

（1）LIME

Most model simplification techniques are founded on rule extraction techniques. Local Interpretable Model Imperceptible Explanation (LIME) is the most widely used technique for local ex post explanation(Ribeiro, Singh, & Guestrin, 2016). In LIME, a locally-generated linear model is used to explain the predictions of intransigent machine learning models. LIME is classified as rule-based local simplified interpretation. By simplifying the explanation of the training model to be conveyed, a new system is constructed. Typically, the simplified model is designed to maintain performance while reducing the complexity of its previous model functions.

（2）SHAP

Shapley Additive Explanations (SHAP） (Lundberg & Lee, 2017). Shapley values are used to evaluate feature significance and explain a trained complicated machine learning model SHAP. Shapley values may be determined for each sample feature component and averaged. The sample feature components' shapley values indicate how much the sample values impact the predicted output. The feature significance may be calculated by averaging the absolute value of the shapley values of all samples on each feature. The total relevance helps us comprehend how each variable affects complicated model prediction outcomes.

2.3.3 Evaluation of interpretable methods

(1) Trust

Ethical artificial intelligence is needed since humans don't accept unexplained or untrustworthy approaches (Mittelstadt et al., 2019). Focusing on AI model performance is making these systems undesirable. AI model performance and transparency are trade-offs, yet interpretability and explainability can help fix model defects (Arrieta et al., 2020).

(2) Biases

Interpretability is a social notion that ensures machine learning model fairness (Barocas et al., 2018). XAI models should be able to see how intermediate factors impact output as part of a fairness or ethical examination (Weller, 2017). One of XAI's goals is to identify biassed parameters in data (Bellamy et al., 2018). XAI can assist avoid unethical AI model use as their practical applications increase fast to encompass human lives.

(3) Privacy evaluation

XAI models may need to analyse user privacy (Patel et al., 2019). Complex machine learning models make it hard for users to grasp the model structure and data contained in its internal representations, putting privacy at risk (Cramer et al., 2018). Unauthorised third parties being able to comprehend training model internal linkages may violate data source differential privacy (Abadi et al., 2016). Given their relevance in important industries, XAI model implementation should incorporate privacy and secrecy (Rahman et al., 2020). Some academics have worked on creating interpretable machine learning models that can interact with consumers (Kim et al., 2018). This aim applies to areas where users make all significant decisions and user model interaction is crucial to success.

Chapter 3

Results

3 .1Dataset acquisition

The data of this study is from **Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines**

3.2 Preparing the Dataset

3.2.1 Data pre-processing and feature engineering

The large volume of data in the influenza vaccine dataset collected in this paper has led to problems such as collection errors, large format differences, large null values and large differences in data ranges. Therefore, it is not possible to train a model without processing the raw data, so a series of processes are first applied to the raw data. The purpose of data pre-processing is to remove worthless information and to format the data according to standards so as to obtain a valid dataset. The following are the steps to pre-process the raw data.

(1) Data coding to view the shape of the data

(2) Missing value processing：In the dataset we collected, there were no missing values, so all variables could be considered in the modelling.

(3) Visualising the data

(4) Viewing the distribution of labels and, for strings, the coding of the data set

(5) Outlier handling

The presence of unreasonable values in the original data can affect the prediction accuracy of the model. A box-line plot-based outlier detection method is used, utilising the upper and lower quartiles (Q3, Q1) and the interquartile spacing (IQR) in the data. The interquartile spacing is defined as the difference between the upper quartile and the lower quartile, i.e. IQR = Q3-Q1. Variable values greater than Q3+1.5IQR and less than q1-1.5iQR are outliers.

(6) Separating features and labels and performing data partitioning where a random seed is set so that the experiment can be repeated multiple times.

3.2.2 Feature importance and correlation analysis methods

(1) Characteristic importance

The importance of a feature is the distribution of the relationship between the magnitude of the influence of each decision variable on the target variable, and the more important the feature selected, the greater the contribution of the feature to the training model. XGBoost is a type of Boosting tree model, which can be used to select features through its own generation process, and the number of feature splits, the average gain of features and the average coverage of features are used to measure the influence of different features on the target variables. The size of the relationship.

The number of feature splits is the number of times a feature acts as a split node during the splitting process of the XGBoost tree model, and is denoted by FS. The average gain of a feature is the average gain of a feature as a split node, denoted by AvgG. The average feature coverage is the average size of the number of samples covered by a feature acting as a split node, denoted by AvgC. AvgC is suitable for datasets with a large number of features, while this paper has a relatively small number of features and is not suitable for feature filtering by this method. The method of feature screening by AvgG reduces the shortcomings of FS and AvgC, so this paper uses the method of XGBoost feature average gain for feature screening, and the three calculation formulae are expressed as follows:

$$FS = |X| \tag{4.1}$$

$$AvgG = \frac{\sum Gain_x}{FS} \tag{4.2}$$

$$AvgC = \frac{\sum Cover_x}{FS} \tag{4.3}$$

(2) Feature correlation

In this paper, feature relevance screening is carried out by means of mutual information after feature importance screening, with the aim of eliminating redundant features, and the formula for calculating feature relevance is shown in 4-4.

$$I(A,B) = H(A) - H(A|\ \mathrm{B}) = \sum_{y\in B}\sum_{i\in A} p(\mathrm{a},\mathrm{b}) \log\left(\frac{p(\mathrm{a},\mathrm{b})}{p(\mathrm{a})\,p(\mathrm{b})}\right) \tag{4.4}$$

Where $I(A,B)$ takes values between [0,1], the closer $I(A,B)$ takes to 1, the more correlated the two features are, and the closer $I(A,B)$ takes to 0, the less correlated the two features are. $p(\mathrm{a})$ and $p(\mathrm{b})$ denote the marginal probability densities of a and b, respectively, while $p(\mathrm{a},\mathrm{b})$ denotes the joint distribution function of a and b.

3.2.3 FCG_XGBoost algorithm flow

In this paper, the FCG_XGBoost feature screening algorithm is constructed based on a comprehensive consideration of the impact of both feature importance and relevance on the accuracy of the model, and the algorithm is calculated as follows:

Input: Raw data D

Output: a collection of variables and feature importance scores at $\{(T_1,\mathrm{s}_1),(T_2,\mathrm{s}_2),....,(T_i,\mathrm{s}_i)\}$, where T_i is the variable name and s_i is the feature importance score.

1: Calculate the importance score of each feature by using the XGB tree model $T = \{(T_1,\mathrm{s}_1),(T_2,\mathrm{s}_2),...,(T_n,\mathrm{s}_n)\}$;;

2: Feature relevance screening by mutual information, filtered feature combinations $TF = \{F_1, F_2 ... \mathrm{F}_c\}$

3: For each F_i in the TF, the average importance within the group is calculated by $A_i = \frac{SUM(\mathrm{s}_i)}{COUNT(\mathrm{s}_i)}$ and the different groups are arranged in descending order of feature importance, and the groups are also arranged in descending order according to s_i to form a new set of features T ;

4: The least influential features are removed in reverse order in the feature combination T T_x , and the constructed new features are retrained in the XGB model.

5 Finding the combination of features with optimal results $\{(T_1,\mathrm{s}_1),(T_2,\mathrm{s}_2),....,(T_i,\mathrm{s}_i)\}$

In this case, the feature relevance screening by means of mutual information is calculated as follows:

Input: raw data D, set of variables and feature importance scores T= $\{(T_1,\mathrm{s}_1),(T_2,\mathrm{s}_2),....,(T_i,\mathrm{s}_i)\}$, where T_i is the variable name and s_i is the feature importance score.

Output: Combination of filtered features $TF = \{F_1, F_2 ... \mathrm{F}_c\}$

1: // feature combination $TF = \{F_1, F_2 ... \mathrm{F}_c\}$, initially empty; initial value of maximum correlation is 0, and max_R

// Index of the max_R group max_F_idex , initially -1

2. $for\ \mathrm{T}_i\ in\ \mathrm{T}$:

3: $max\_R = 0$

4: $max\_F\_idex = 1$

5: $for\ \mathrm{F}\_idex, F_i\ in\ \mathrm{TF}$

6: $num = 0$ // indicates the number of feature correlations $R > \theta$ in T_i and F_i , initially 0

7: $CR = 0$ // denotes the sum of the feature correlations of T_i and the previous λ in F_i , initially 0

8: $for\ \mathrm{f}_i, \mathrm{fcr}_i\ in\ \mathrm{F}_i$

9: $if\ num > \lambda: break$ // former λ feature to calculate correlation

10: $R = metrics.normalized\_mutual\_minfo\_score(D[T_i], D[\mathrm{f}_i])$ // Relevance R

11: $CR += R$ //Summing up correlations for different groups

12: $if\ R \geq \theta: num += 1$

13: $if\ (len(F_i) < \lambda\ \&\ num >= 1)\ ||\ num\lambda\ \&\ CR > \max_R$:

// Are the correlations of the former λ features in T_i and F_i all satisfied $R > \theta$

// If more than one combination meets the condition, the combination needs to be put into the group of the sum of the maximum correlations

14: $\max\_F\_idex \leftarrow \mathrm{F}\_idex$, $\max\_R \leftarrow CR$ //Update maximum relevance index

and maximum relevance sum

15: *if* $\max\_F\_idx \geq 0$

16: $TF[max\_\text{F\_idex}] \cup T_i$ //put T_i into the group indexed by $\max\_F\_idx$ in the combination TF

17: $else: TF \cup T_i$ // otherwise group T_i separately

18: *end for* //Get a combination of filtered features $TF = \{F_1, F_2 ... \text{F}_c\}$

The following is a detailed explanation of the FCG_XGBoost feature importance and relevance selection algorithm.

Step1:Firstly, we calculate the importance score of each feature based on the XGB tree model and remove the features with relatively low impact.

Step2:Grouping by mutual information, the initial feature $T1$ is automatically grouped as F1, if a new feature T2 is input, first calculate the correlation R between T1 and T2, if it meets $R > \theta$, then T2 and T1 are grouped into the same group, otherwise they are each divided into separate groups. When a new feature T_i is added again, compare the correlation between T_i and the first λ features of each group , if F_i $R > \theta$ is satisfied, the new feature is added to the group, otherwise continue to judge whether the next group meets the condition; if more than one group meets it at the same time, choose the group with the largest sum of feature correlation to join. If none of them meet the condition, the new feature will be added to the group independently. $TF = \{F_1, F_2 ... \text{F}_c\}$ If there are more than one grouping, select the group with the highest correlation.

F_i Step3:After step 2 is completed, the average feature importance score is calculated for different groups A_i ; the features are arranged from largest to smallest between groups A_i ; the features are arranged from largest to smallest within each group s_i ; the new feature combinations are formed by cyclically traversing through the features of different groups $(T_1, T_2, T\text{j})$; the cyclic traversal can be expressed as follows: the feature combinations are {{T1}{T2,T3} ... {Tn,Tn-1}}, and the new combination of features T={T1,T2.... .Tn-1,T3...Tn}.

Step4:Train the XGB tree model by removing one feature Tx at a time in the set T, forming a new combination of features to calculate the model's goodness-of-fit R^2 until the number of features drops to 1, completing the model training.

Step5:Select the combination of features with the best results. Through the previous 4 steps, the selection of feature relevance and importance is completed. With the principle of strong intra-group correlation and weak inter-group correlation, the features with high correlation are put together, and at the same time, the importance of the features is calculated to ensure that each feature contributes as much as possible to the model. The process of feature correlation analysis is shown in the figure below.

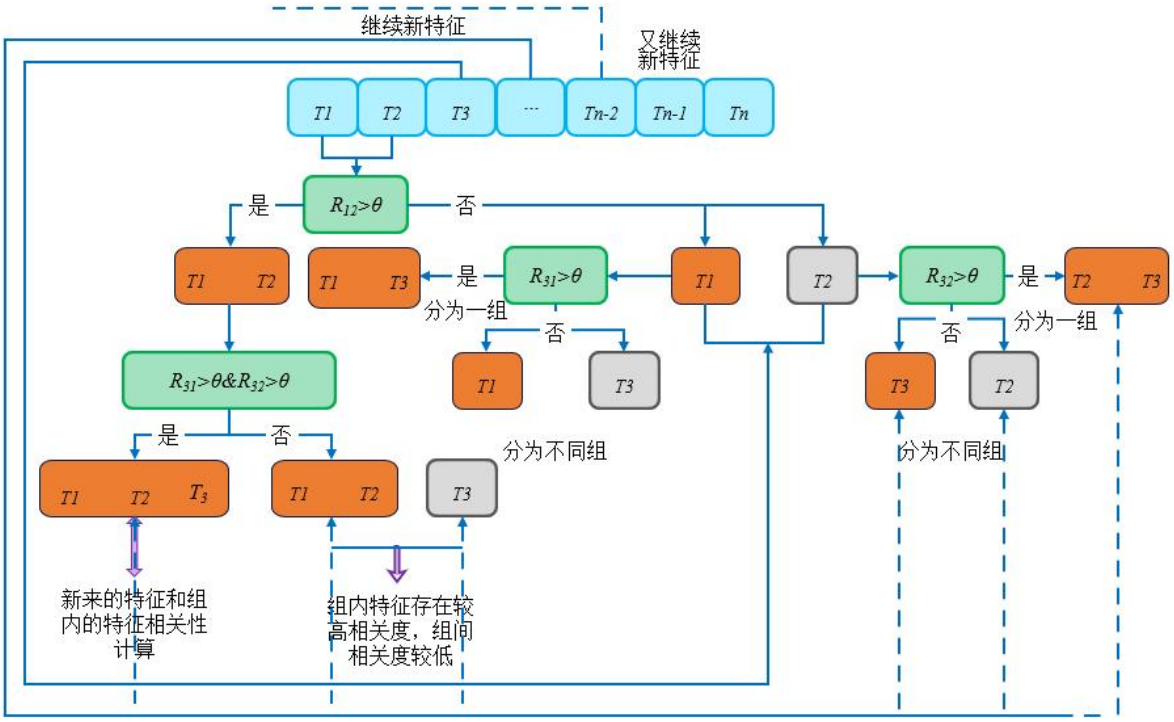
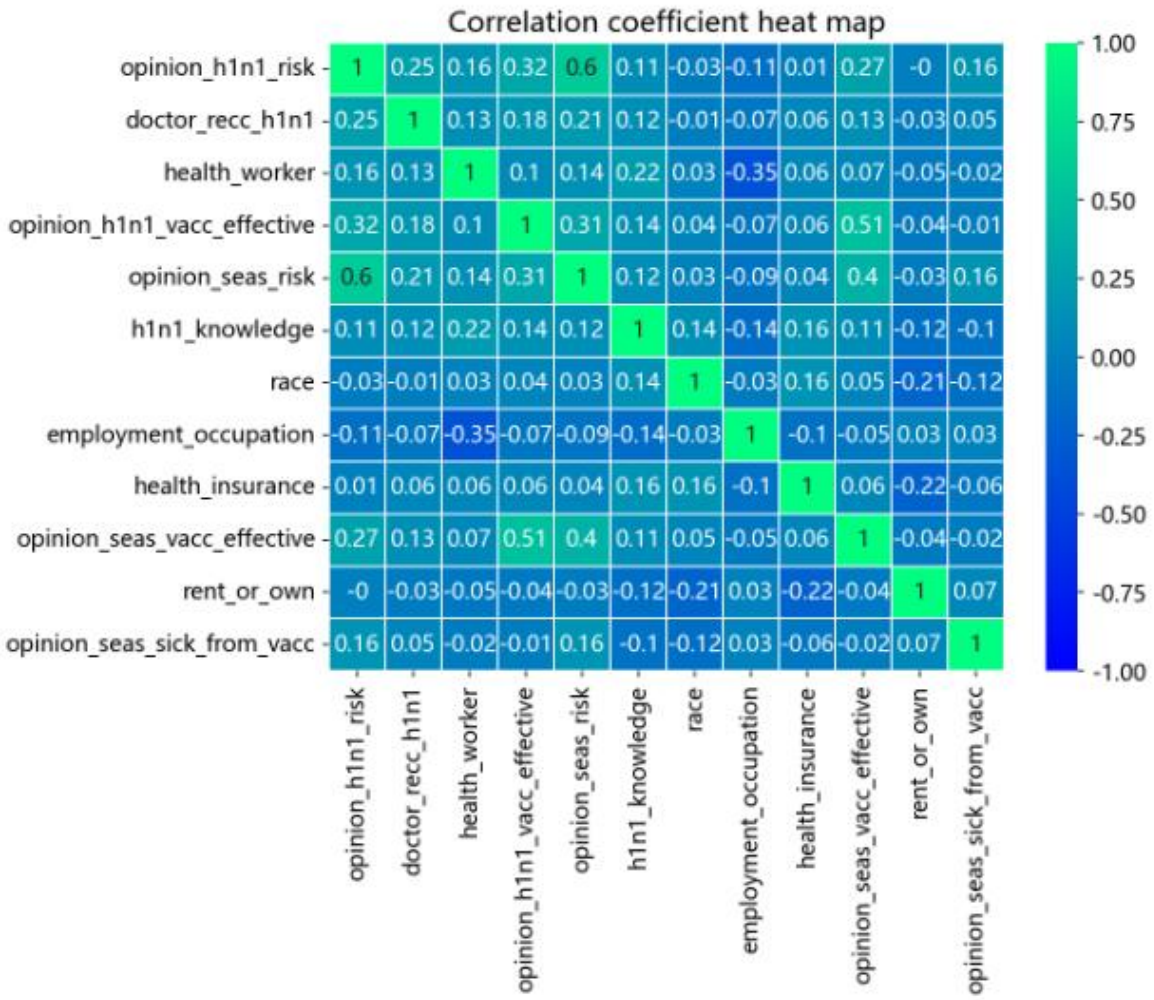


Figure 3.2.3.1 Feature correlation flowchart

The feature screening based on FCG_XGBoost algorithm can find the effective features with high feature importance and low correlation among features, which can ensure high feature importance and effectively eliminate the problem of large redundancy among features. After eliminating redundancy by this algorithm, 12 optimal feature combinations affecting H1N1 influenza vaccination were finally screened, and the correlation coefficient heat map matrix is shown in Figure 4 below, in which the correlation between opinion_h1n1_risk and opinion_seas_risk is the largest at 0.6.

Fig. 4


Correlation coefficient heat map
opinion_h1n1_risk 1 0.25 0.16 0.32 0.6 0.11 -0.03 -0.11 0.01 0.27 -0 0.16
doctor_recc_h1n1 0.25 1 0.13 0.18 0.21 0.12 -0.01 -0.07 0.06 0.13 -0.03 0.05
health_worker 0.16 0.13 1 0.1 0.14 0.22 0.03 -0.35 0.06 0.07 -0.05 -0.02
opinion_h1n1_vacc_effective 0.32 0.18 0.1 1 0.31 0.14 0.04 -0.07 0.06 0.51 -0.04 -0.01
opinion_seas_risk 0.6 0.21 0.14 0.31 1 0.12 0.03 -0.09 0.04 0.4 -0.03 0.16
h1n1_knowledge 0.11 0.12 0.22 0.14 0.12 1 0.14 -0.14 0.16 0.11 -0.12 -0.1
race -0.03 -0.01 0.03 0.04 0.03 0.14 1 -0.03 0.16 0.05 -0.21 -0.12
employment_occupation -0.11 -0.07 -0.35 -0.07 -0.09 -0.14 -0.03 1 -0.1 -0.05 0.03 0.03
health_insurance 0.01 0.06 0.06 0.06 0.04 0.16 0.16 -0.1 1 0.06 -0.22 -0.06
opinion_seas_vacc_effective 0.27 0.13 0.07 0.51 0.4 0.11 0.05 -0.05 0.06 1 -0.04 -0.02
rent_or_own -0 -0.03 -0.05 -0.04 -0.03 -0.12 -0.21 0.03 -0.22 -0.04 1 0.07
opinion_seas_sick_from_vacc 0.16 0.05 -0.02 -0.01 0.16 -0.1 -0.12 0.03 -0.06 -0.02 0.07 1
opinion_h1n1_risk
doctor_recc_h1n1
health_worker
opinion_h1n1_vacc_effective
opinion_seas_risk
h1n1_knowledge
race
employment_occupation
health_insurance
opinion_seas_vacc_effective
rent_or_own
opinion_seas_sick_from_vacc
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

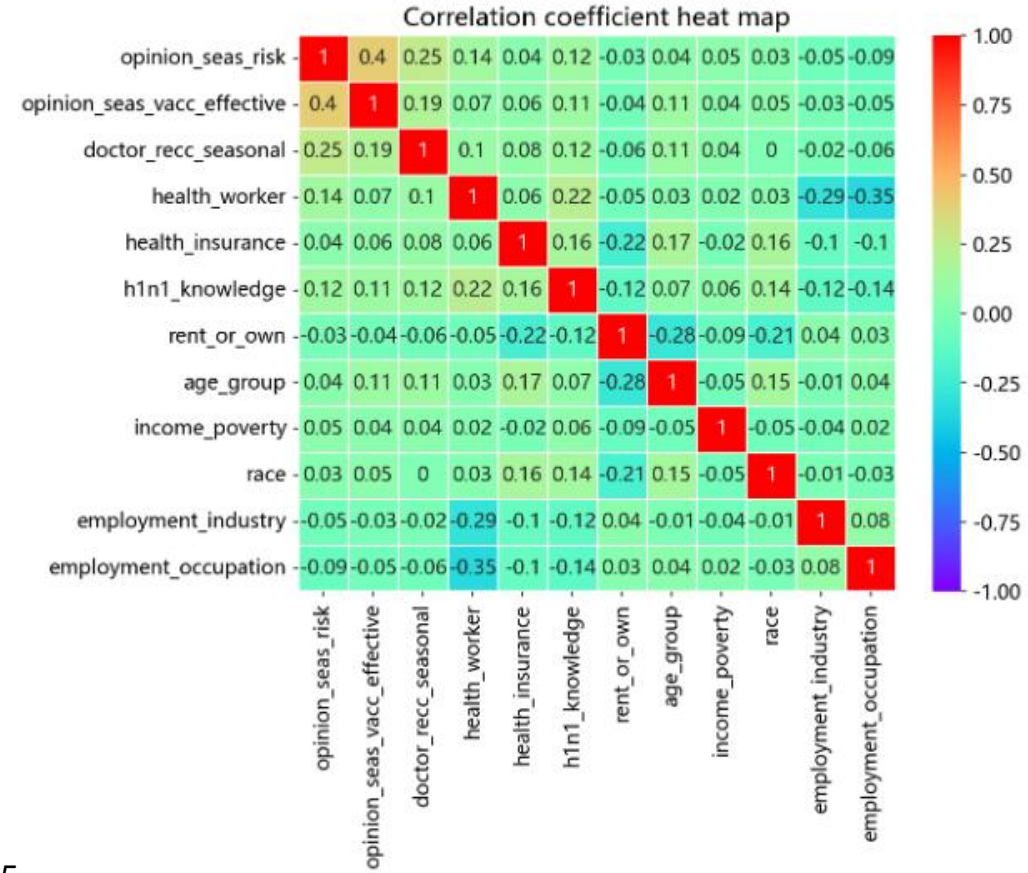

Correlation coefficient heat map
opinion_seas_risk 1 0.4 0.25 0.14 0.04 0.12 -0.03 0.04 0.05 0.03 -0.05 -0.09
opinion_seas_vacc_effective 0.4 1 0.19 0.07 0.06 0.11 -0.04 0.11 0.04 0.05 -0.03 -0.05
doctor_recc_seasonal 0.25 0.19 1 0.1 0.08 0.12 -0.06 0.11 0.04 0 -0.02 -0.06
health_worker 0.14 0.07 0.1 1 0.06 0.22 -0.05 0.03 0.02 0.03 -0.29 -0.35
health_insurance 0.04 0.06 0.08 0.06 1 0.16 -0.22 0.17 -0.02 0.16 -0.1 -0.1
h1n1_knowledge 0.12 0.11 0.12 0.22 0.16 1 -0.12 0.07 0.06 0.14 -0.12 -0.14
rent_or_own -0.03 -0.04 -0.06 -0.05 -0.22 -0.12 1 -0.28 -0.09 -0.21 0.04 0.03
age_group 0.04 0.11 0.11 0.03 0.17 0.07 -0.28 1 -0.05 0.15 -0.01 0.04
income_poverty 0.05 0.04 0.04 0.02 -0.02 0.06 -0.09 -0.05 1 -0.05 -0.04 0.02
race 0.03 0.05 0 0.03 0.16 0.14 -0.21 0.15 -0.05 1 -0.01 -0.03
employment_industry -0.05 -0.03 -0.02 -0.29 -0.1 -0.12 0.04 -0.01 -0.04 -0.01 1 0.08
employment_occupation -0.09 -0.05 -0.06 -0.35 -0.1 -0.14 0.03 0.04 0.02 -0.03 0.08 1
opinion_seas_risk
opinion_seas_vacc_effective
doctor_recc_seasonal
health_worker
health_insurance
h1n1_knowledge
rent_or_own
age_group
income_poverty
race
employment_industry
employment_occupation
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

Figure 5

3.3 Construction of a vaccination prediction model

3.3.1Model parameter setting

In order to compare the performance of different algorithms for prediction models affecting H1N1 influenza vaccine and seasonal influenza vaccination, this paper uses integrated learning algorithms such as XGB, GBDT and LGBM, and also compares the performance of random forest and decision tree models on this basis, and evaluates the experimental results of these five models through the classification model evaluation index to find the prediction model with the best performance for each. The main parameters of each model were set as follows:

Tab. 1.1Main parameter values of each model

Models	Parameter settings	
DT	Maximum tree depth: max_depth=6	
RF	Number of weak learners: n_estimators=10	Maximum tree depth: max_depth=6
XGB	Learning rate: learning_rate=0.1 Number of weak learners: n_estimators=100 Minimum leaf node weight sum:min_child_weight=5	Maximum depth of tree: max_depth=3 gamma value: gamma=0.1 Sample imbalance parameter: scale_pos_weight=1
GBDT	Number of weak learners: n_estimators=100 Maximum depth of tree: max_depth=3 Subsampling: subsample=0.6	Learning rate: learning_rate=0.1 Loss function: loss='divance'
LGBM	Number of weak learners: n_estimators=100 Maximum depth of tree: max_depth=3 Minimum number of leaves on the leaf: min_child_samples=50	Learning rate: learning_rate=0.01 Number of leaves per tree: mean_leaves=25

3.3.2Analysis of the results of the algorithm

For the five machine learning models selected in the previous section, the datasets of prediction models affecting H1N1 influenza vaccination and seasonal influenza vaccination after feature screening by FCG_XGBoost algorithm were used respectively, and the five machine learning models were compared and analysed by selecting the model with the best performance and the strongest generalisation ability of each and taking a multi-dimensional analysis of the model prediction results. In this paper, in addition to the accuracy, AUC and area value under the PR curve as evaluation indicators, the accuracy of the prediction result of 1 is also required to be checked as an evaluation indicator.

The prediction results of each model are shown in the table below. Precision, Recall and F1-score represent the check rate, recall rate and F1 value respectively, and 0 and 1 represent whether the vaccine was administered or not, with 0 being unvaccinated and 1 being vaccinated. Among them, the XGBoost algorithm achieved the best classification results.

Considering the accuracy rate H1N1 vaccination and seasonal influenza vaccination prediction check rate together, the XGBoost model performed the best,

with accuracy rates of 81.88% and 79.45%, respectively, and its more accurate prediction of whether to receive H1N1 vaccination and seasonal influenza vaccination, with AUC values of 87.35% and 87.47%, respectively, and F1 values of 81.3% and 79.44%, with recall rates of 81.88% and 79.45%, respectively.

At present, no algorithm in the field of machine learning can absolutely outperform other algorithms in terms of model performance, so this paper uses a variety of machine learning algorithms to compare and analyse, with the aim of finding the algorithm that has the best prediction effect on the H1N1 vaccine and seasonal influenza vaccination and the strongest model generalisation ability, using the algorithm to explore the valuable information hidden behind the data, and then provide a basis for determining whether to vaccinate or not The algorithm is used to uncover the valuable information behind the hidden data and provide a basis for decision making on whether to vaccinate or not, as well as the analysis of herd immunity. The aim of this paper is to evaluate the prediction performance of the model by using the evaluation metrics of the machine learning model test set to find the most suitable algorithm and further optimise the algorithm with the best performance, with the aim of improving the model's ability to predict H1N1 vaccination and seasonal influenza vaccination. The closer the accuracy ACC is to 100% and the closer the AUC, F1 value and recall are to 100%, the better the model's prediction is. As can be seen from the table below, the XGB algorithm has the highest accuracy ACC in predicting the H1N1 vaccine and seasonal influenza vaccination, at 81.88%, so the model's performance in predicting the H1N1 vaccine and seasonal influenza vaccination is significantly due to other algorithms, therefore, this paper further optimises the XGB algorithm to predict the H1N1 vaccine and seasonal influenza vaccination on the basis of The performance of the XGB algorithm is therefore significantly better than other algorithms.

Tab3.3.2.1Performance comparison of machine learning prediction models for H1N1 influenza vaccination

Models	Training set (%)					Test set (%)				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
DT	0.8411	0.8783	0.8371	0.8372	0.8411	0.7981	0.8185	0.7934	0.7920	0.7981
RF	0.8371	0.8878	0.8275	0.8361	0.8371	0.8069	0.8504	0.7952	0.8014	0.8069
XGB	0.8573	0.9094	0.8527	0.8547	0.8573	0.8188	0.8735	0.8130	0.8133	0.8188
GBDT	0.8593	0.9111	0.8549	0.8568	0.8593	0.8131	0.8747	0.8080	0.8074	0.8131
LGBM	0.8111	0.8558	0.7935	0.8129	0.8111	0.7883	0.8405	0.7674	0.7846	0.7883

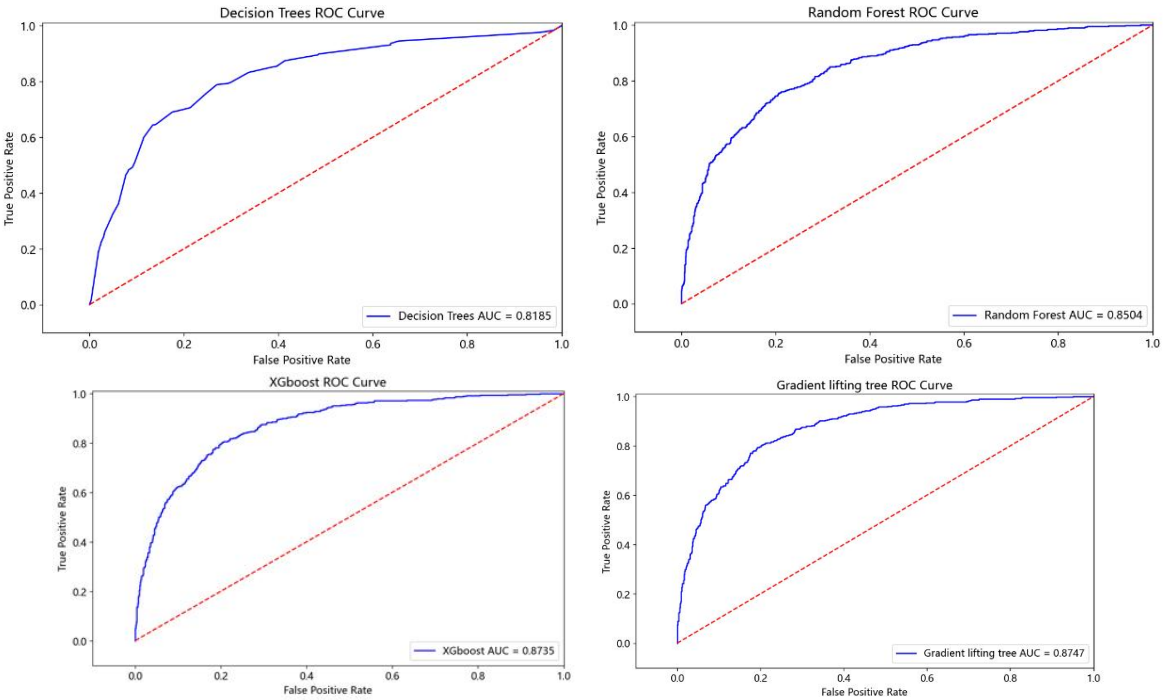
Tab3.3.2.2Performance comparison of machine learning prediction models for seasonal influenza vaccination

Models	Training set (%)					Test set (%)				
	ACC	AUC	F1	Precision	Recall	ACC	AUC	F1	Precision	Recall
DT	0.8029	0.8766	0.8025	0.8027	0.8029	0.7655	0.8311	0.7653	0.7652	0.7655
RF	0.8042	0.8828	0.8026	0.8063	0.8042	0.7785	0.8575	0.7770	0.7792	0.7785
XGB	0.8202	0.9008	0.8198	0.8202	0.8202	0.7945	0.8747	0.7944	0.7944	0.7945
GBDT	0.8242	0.9020	0.8237	0.8242	0.8242	0.7940	0.8755	0.7939	0.7938	0.7940
LGBM	0.7703	0.8521	0.7687	0.7710	0.7703	0.7588	0.8413	0.7575	0.7589	0.7588

(1) ROC curves and AUC values

The ROC curve is an effective indicator of the algorithm's generalization performance, and its most intuitive application is to reflect the trend of the model's

sensitivity and accuracy when different thresholds are selected. the shape of the ROC remains basically constant when the distribution of positive and negative samples changes, so the area AUC value under the ROC curve is a more stable indicator of the model's performance, the larger the AUC, the better the model's performance. As shown in the following model test set ROC curve comparison graph, the AUC values of five models on the test set are greater than 0.7, among which XGBoost has the best classification performance and the AUC value can reach 0.8735. The H1N1 influenza vaccine and seasonal influenza vaccination prediction model ROC curves are shown in Figure 1 and Figure 2 below respectively.



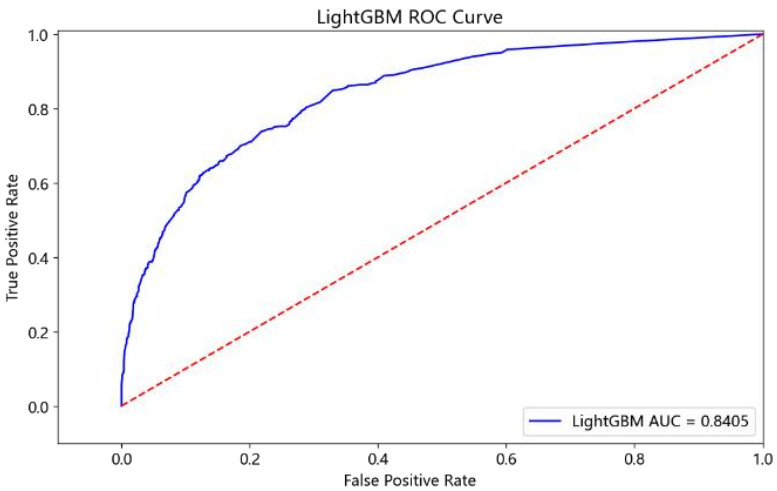


Figure 1 ROC curve for the H1N1 influenza vaccination prediction model

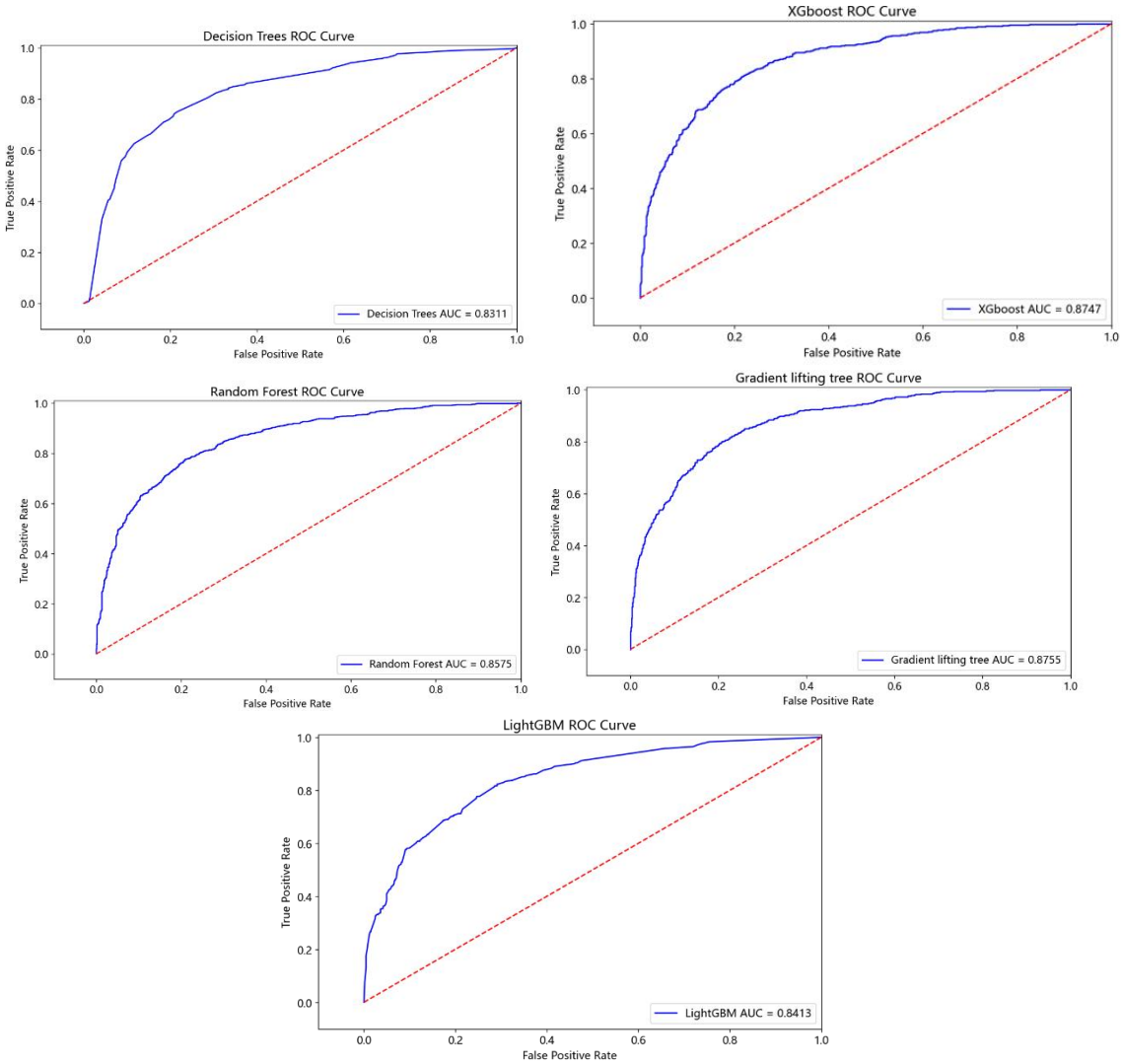


Figure 2 Seasonal influenza vaccination prediction model ROC curve

(2) PR curves and their area values

The PR curve is actually a curve made with the two values of accuracy and recall as variables. the PR curve reflects the relationship between accuracy and recall. the PR curve is extremely sensitive to sample data imbalance and will follow when the sample data is unevenly distributed. a larger value of area under the PR curve indicates that the sample has good separability even under uneven distribution conditions. Therefore, the PR curve is used to reflect the good or bad performance of the model under the premise of positive and negative sample imbalance. As shown in the figure below, the XGBoost and GBDT models performed best for H1N1 sexual influenza vaccine prediction under the premise of unbalanced data, with the largest area under the PR curve of 0.76.

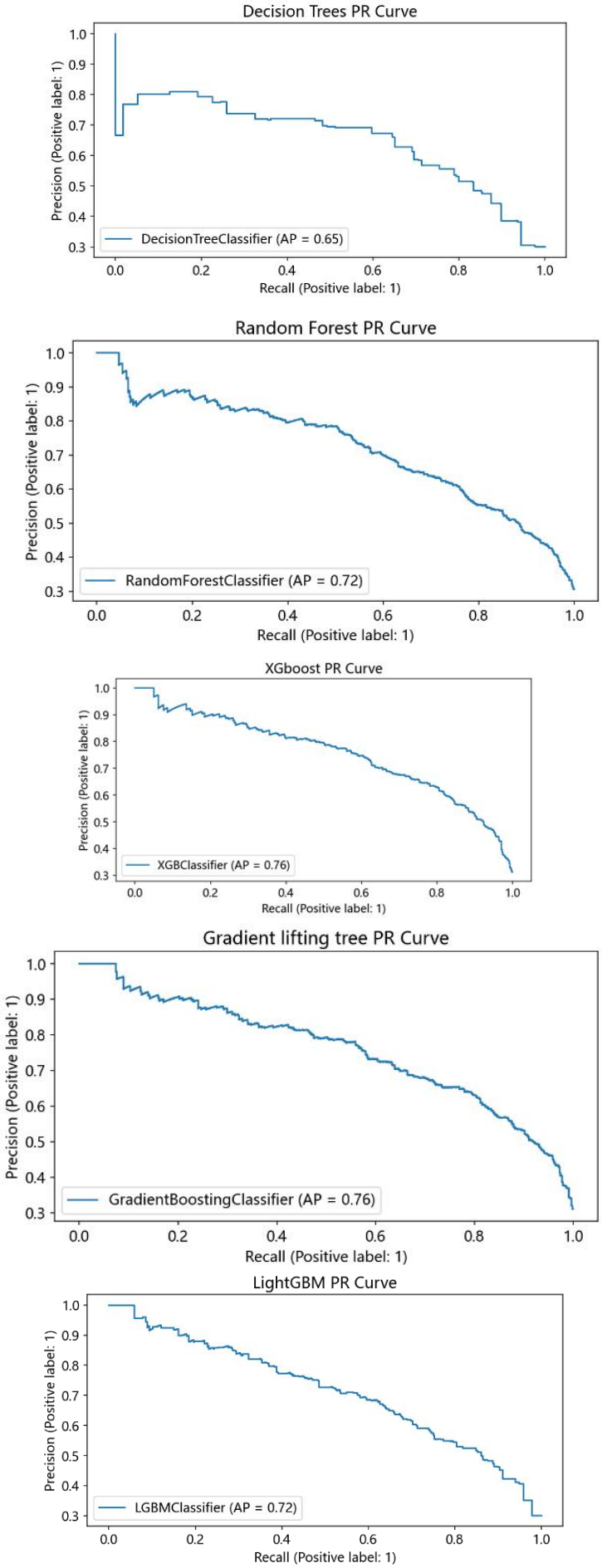
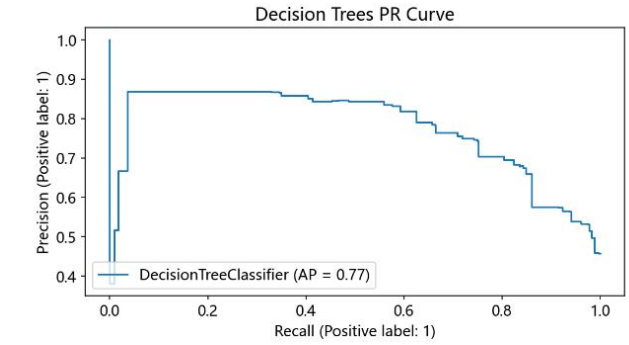
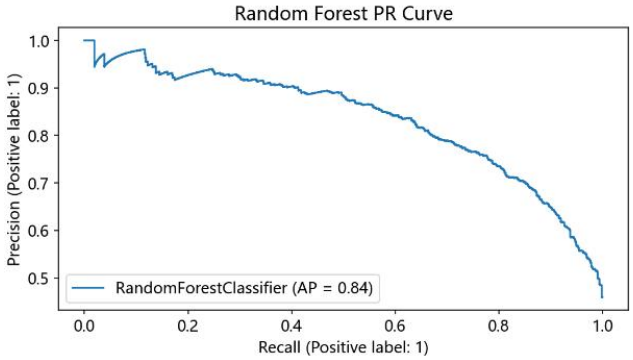
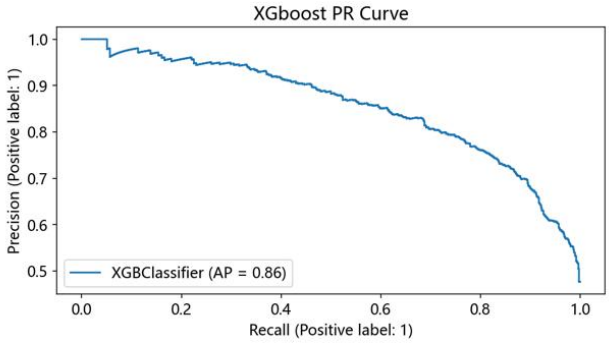
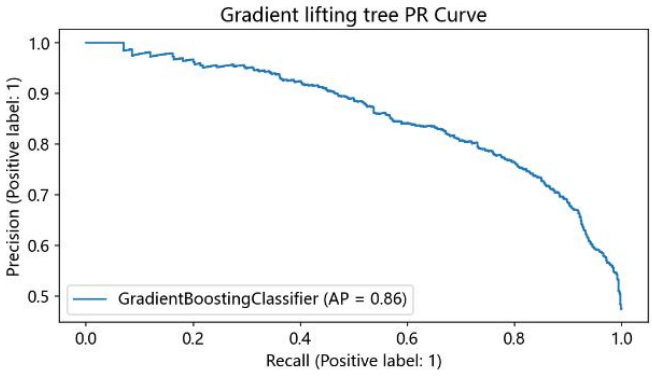


Figure 3 PR curve of the H1N1 influenza vaccination prediction model


Decision Trees PR Curve
Precision (Positive label: 1)
Recall (Positive label: 1)
DecisionTreeClassifier (AP = 0.77)


Random Forest PR Curve
Precision (Positive label: 1)
Recall (Positive label: 1)
RandomForestClassifier (AP = 0.84)


XGboost PR Curve
Precision (Positive label: 1)
Recall (Positive label: 1)
XGBClassifier (AP = 0.86)


Gradient lifting tree PR Curve
Precision (Positive label: 1)
Recall (Positive label: 1)
GradientBoostingClassifier (AP = 0.86)

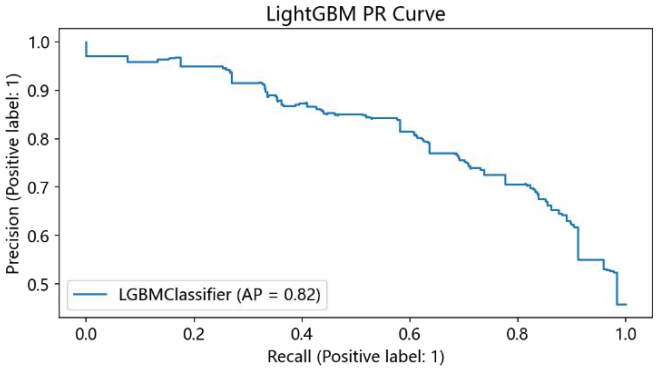

LightGBM PR Curve
Precision (Positive label: 1)
Recall (Positive label: 1)
LGBMClassifier (AP = 0.82)

Figure 4 PR curve for seasonal influenza vaccination prediction model

In summary, five models affecting the prediction performance of H1N1 vaccination and seasonal influenza vaccination, XGB, LGBM, GBDT, RF and DT, were analysed from multiple dimensions and perspectives, and the experimental results show that: the XGB integrated learning algorithm has the best prediction effect on H1N1 vaccination and seasonal influenza vaccination compared with the other four algorithms, and the In the prediction of H1N1 vaccination and seasonal influenza vaccination, it is usually necessary to continuously improve the accuracy of the prediction of H1N1 vaccination and seasonal influenza vaccination, so this paper introduces the Bayesian optimization method, aiming to further improve the prediction performance of the model.

3.4 Research on predictive models based on Bayesian optimization

In order to further optimise the effectiveness of the H1N1 vaccine and seasonal influenza vaccination prediction model, this paper uses a Bayesian optimisation algorithm and introduces it into the above-mentioned XGBoost integrated learning algorithm for intelligent tuning, selecting the most effective combination of hyperparameters in the domain value space according to the parameters of the XGBoost algorithm, with the aim of further improving the prediction performance of the model, so as to achieve The aim is to further improve the prediction performance of the model to achieve more accurate prediction of H1N1 vaccination and seasonal influenza vaccination in order to analyse the immunisation status of the population and take more effective vaccination measures.

Tab1.4 Performance comparison of Bayesian optimized models for predicting H1N1 influenza vaccination status

Models	Test set				
	ACC	AUC	F1	Precision	Recall
XGBoost	0.8188	0.8735	0.8130	0.8133	0.8188
BSXGB	0.8277	0.8816	0.8295	0.8194	0.8213

Tab. 1.5 Performance comparison of Bayesian optimized models for predicting seasonal influenza vaccination status

Models	Test set				
	ACC	AUC	F1	Precision	Recall
XGBoost	0.7945	0.8747	0.7944	0.7944	0.7945
BSXGB	0.8077	0.8816	0.8095	0.8094	0.8013

3.5 Model interpretation

Since no machine algorithm is absolutely superior to the others, the predictive performance of the machine learning models in the test dataset was primarily evaluated to select the most appropriate algorithm. Therefore, the machine

learning models built were applied to predict the vaccine coverage in the test set. Tables 1 and 2 show a comparison of the prediction performance of these models. We can see that in the seasonal influenza model, the XGBoost algorithm is more accurate than the random forest model. The same results are seen in the H1N1 influenza model.

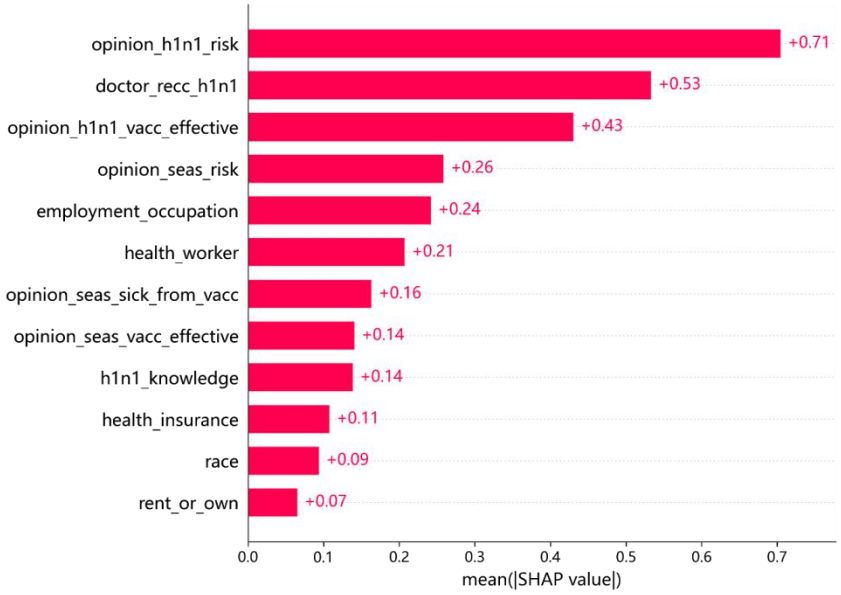


Fig. 7

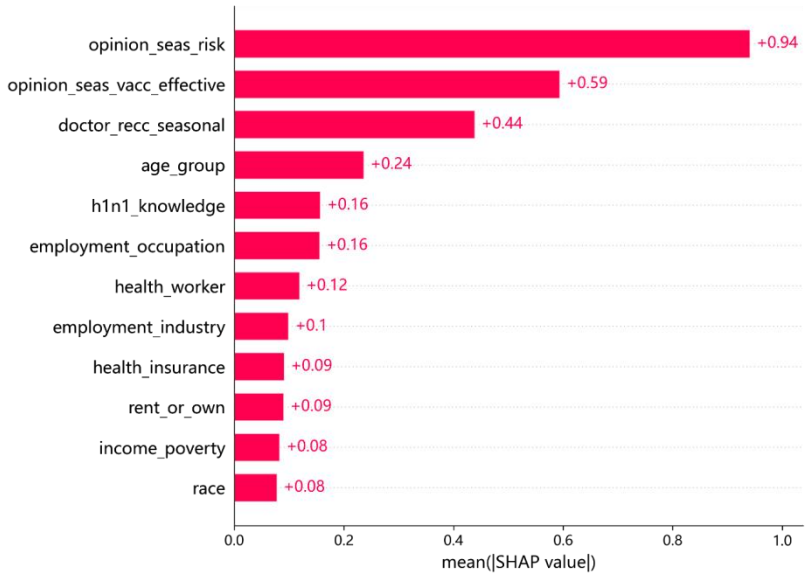


Figure 8

Figure 7 illustrates the XGBoost model. Among the factors influencing seasonal influenza vaccination rates, people's perceptions of seasonal influenza risk ranked first, followed by people's perceptions of vaccine effectiveness. In the H1N1 influenza model, the situation is slightly different. People's perception of the risk of influenza A (H1N1) ranked first, and doctors' perception of the recurrence of influenza A (H1N1) ranked second.

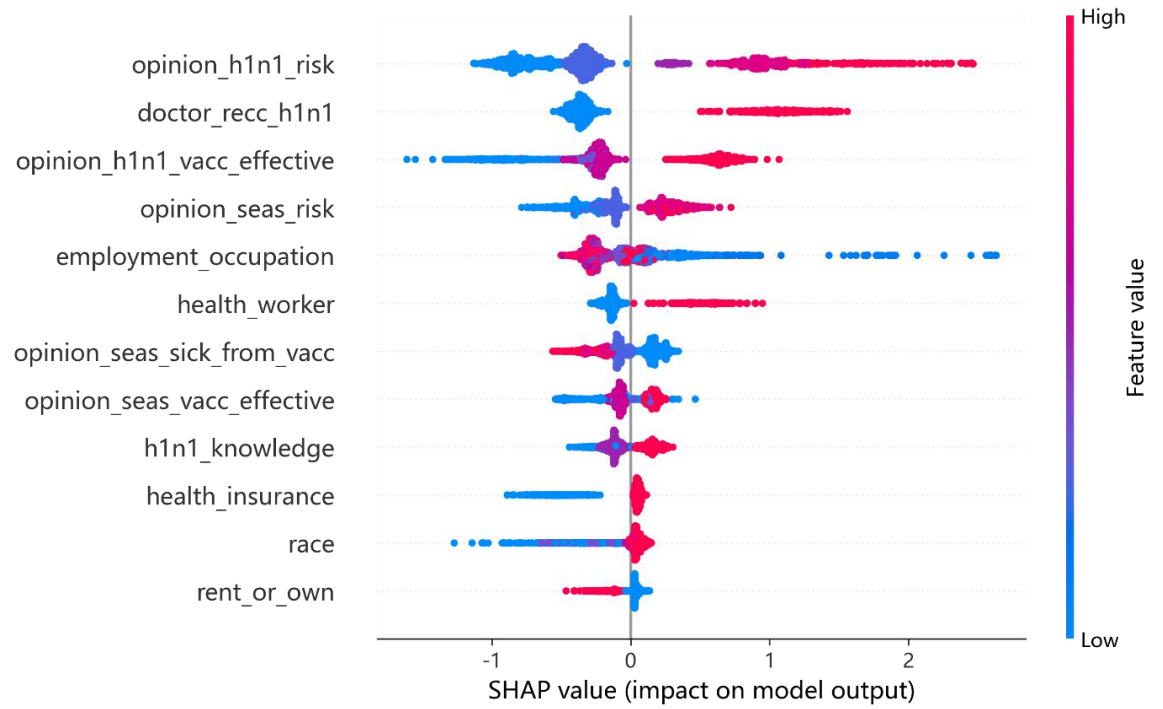


Figure 9

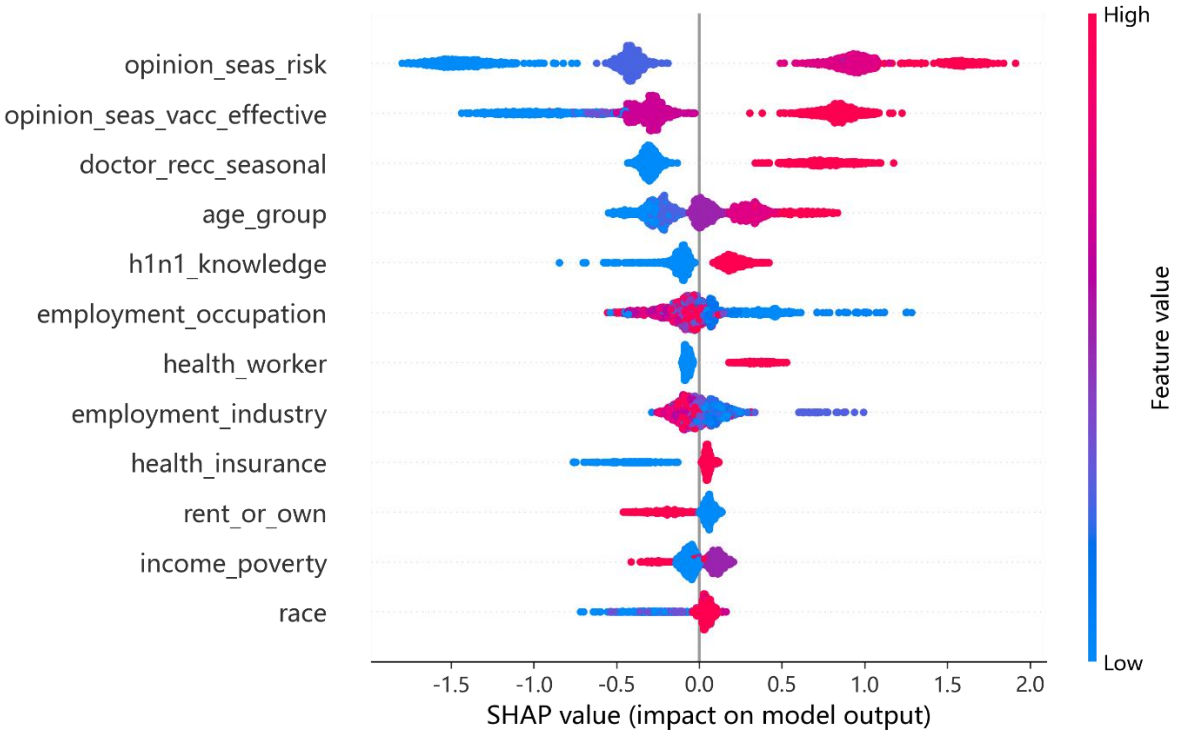


Fig. 10

As shown in Fig. 9 and Fig. Figure 10 shows the distribution of SHAP values for each influencing factor in both models. In the figure, each data point represents a respondent and the colour represents the value of the variable, with the variable data varying from low to high from blue to red.

A positive or negative SHAP value) indicates that the impact parameter is positively (negatively) associated with vaccination rates. For example, the more people value the risk of seasonal influenza, the higher the vaccination rate for seasonal influenza.

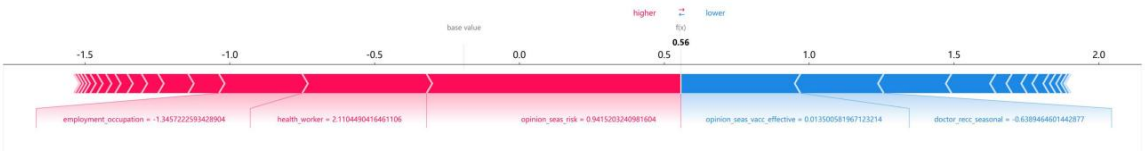


Fig. 11

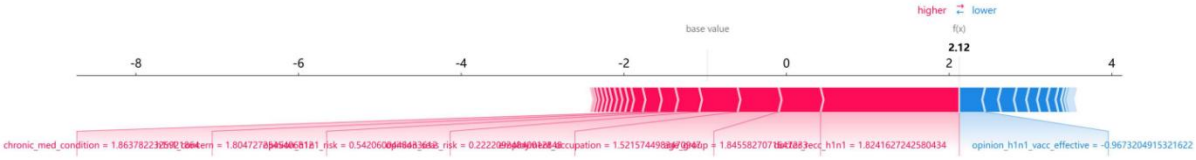


Fig. 12

Figure 11 and Figure 12 give a partial interpretation of the two models. Vaccine coverage is the sum of the contributions of all input parameters and the baseline values are the means of the vaccination rates in the XGBoost model, -2.32 and 2.12 respectively. red leads to an increase in the mean, while blue leads to a decrease in the mean. The SHAP values for perceptions of seasonal influenza vaccine effectiveness and perceived risk of influenza were -0.0135 and 0.9415 respectively. the SHAP for people's perceptions of H1N1 influenza vaccine effectiveness was -0.9673.

Chapter 4

Discussion

In this project, we propose a novel approach to developing a machine learning-based prediction model that is both precise and transparent. Our method not only forecasts whether or not residents have been vaccinated, but also the regional impact of several characteristics and activities on the identity of each character. Most importantly, our research identifies details that may seem inconsequential at first glance but are very crucial to some people.

Our model is unique since it can be understood by both people and machines. Governments and hospitals can benefit from individual and group assessments by learning which residents' behaviours should be promoted and which should be discouraged. Especially in the context of vaccination initiatives, this has the potential to significantly improve public health results.

In addition, our method provides a transparent framework for learning how predictions are created, which encourages stakeholders to have faith in the model and adopt its advice. By building in interpretability to our prediction algorithms, we may facilitate a more natural and productive collaboration between people and AI in addressing public health challenges.

4.1 Conclusion

1. Through data pre-processing and data exploration work, the authors of this paper used the collected influenza vaccination data to identify the most important factors affecting HINI influenza vaccination and seasonal influenza vaccination, and they then established the FCG_XGBoost algorithm based on feature importance and correlation screening, which extracted the significant variables with high feature importance and low feature redundancy. The efficacy of the FCG_XGBoost algorithm was demonstrated by a series of comparison tests that measured metrics including accuracy ACC, area under the curve (AUC), F1 value, precision Precision, and recall Recall. As a final step, we used these key traits to build five separate machine learning prediction models for how HINI influenza vaccine and seasonal influenza vaccination will impact individuals.
Two, the Bayesian optimisation algorithm was used to fine-tune the hyperparameters of the XGB prediction model that had optimal performance for both the HINI influenza vaccine and the seasonal influenza vaccination, thereby enhancing the prediction performance of both models.
Last but not least, we use the SHAP interpretable method to conduct a post-hoc explanatory analysis of the Bayesian optimised XGBoost optimal model, i.e. the machine learning model with the optimal effect on HINI influenza vaccination and seasonal influenza vaccination, to reveal the prediction rules for HINI influenza

vaccination and seasonal influenza vaccination hidden in the historical data, and to provide an explanation to managers as to why they made such a decision. It is possible to uncover the choice and the reasoning behind the data to accurately anticipate influenza vaccination, leading to widespread immunisation and improving public health and safety.

4.2 Ideas for future work

There are many processes involved in predicting vaccination rates using machine learning. The precision of forecasts is contingent on the quality and dependability of the data. If the data are missing, inaccurate, or inadequately characterised, the results of the model's predictions will be affected. Despite the fact that we have chosen relatively abundant data, there are still some restrictions. In addition, machine learning models can only predict past events and cannot be completely adapted to future events. If the model is insufficiently adaptable, the results of the prediction may be inaccurate. In order to increase the accuracy of our prediction results, it is more essential that we consider the limitations of questionnaires and use multiple data collection channels.

2. It is possible that the factors chosen in this paper that influence the HINI influenza vaccine and seasonal influenza vaccination are insufficiently exhaustive. Predictions of vaccination rates are influenced by socioeconomic factors such as income, education level, and culture, in addition to medical factors. These factors are rarely accounted for in the available data; consequently, the model results may be biassed. Therefore, expanding the purview of future research could improve the ability to predict vaccination rates. We welcome criticism from professional academicians in the United States regarding any deficiencies.
3. The Bayesian optimisation algorithm chosen for this paper has a degree of stochasticity for locating the global optimal hyperparameter combination and may fall into a local optimal solution. Future research may continue to incorporate heuristic algorithms in an effort to advance the field. The objective function of the Bayesian optimisation algorithm is simultaneously more complex and costly to use.

4. Xgboost obtains the best results at this moment, but the results obtained vary depending on the data type; therefore, additional model comparisons are required to make the results more convincing.

Reference

Galanis Petros,Katsiroumpa Aglaia,Vraka Irene,Siskou Olga,Konstantakopoulou Olympia,Katsoulas Theodoros,Kaitelidou Daphne. Seasonal Influenza Vaccine Intention among Nurses Who Have Been Fully Vaccinated against COVID-19: Evidence from Greece[J]. Vaccines,2023,11(1).

Liu Vera,Walker Stephen. Testing for genetic mutation of seasonal influenza virus[J]. Journal of Applied Statistics,2023,50(1).

Li C G . A novel influenza A (H1N1) vaccine in various age groups.

Jain S , Kamimoto L , Bramley A , et al. Pandemic Influenza A (H1N1) Virus Hospitalizations Investigation Team (2009) Hospitalized patients with 2009 H1N1 Influenza in the United States, April-June 2009.

Brown et al., 2018 - Brown, A., Yang, Y., & Chen, J., 2018. Predictive modeling of influenza vaccine uptake. Vaccine, 36(45), pp. 6803-6809.

Zhang, C., Li, W., & Yang, Q. (2012). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition, 45(5), 1761-1776.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2), 197-227.

Uddin, Z. (2022). Machine Learning. In Z. Uddin (Ed.), Machine Learning. Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Uddin, Z. (2022). Machine Learning. In Z. Uddin (Ed.), Machine Learning (pp. 1-340). Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (pp. 3149-3157).

Pelikan, M., Goldberg, D. E., & Cantú-Paz, E. (1999). BOA: The Bayesian optimization algorithm. In Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation - Volume 1 (GECCO'99) (pp. 525-532). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In Learning and Intelligent Optimization (pp. 507-523). Springer, Berlin, Heidelberg.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 58, pp.82-115.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., and Kagal, L., 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 80-89). IEEE.

Ribeiro, M.T., Singh, S., & Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144)

Lundberg, S.M., & Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

Barocas, S., Hardt, M., & Narayanan, A. (Eds.). (2018). Fairness and machine learning. fairmlbook.org.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating

Appendix A

Self-appraisal

A.1 Critical self-evaluation

In all, I am happy with how my project turned out, and the precision of the completed model was much better than I had anticipated when I first started working on it. At the very end of my research, I integrated several extra models so that I could be confident that the results I obtained were credible.

Having said that, I will be the first to acknowledge that working on the project presented me with a few difficulties. It took me a substantial amount of time to decide on an appropriate topic, and I discovered that I needed to revisit the process many times, which resulted in redundant actions that consumed a lot of my time. In addition, I underestimated how much time would be required to write the report, and it took far longer than I had anticipated.

Despite these challenges, my passion in the project kept me motivated, and I was determined to conquer any barriers that I came across in the process. In the end, I was quite pleased with the findings of my analysis, and the satisfaction of having achieved my goals made me feel that the time and effort I had put in had been worthwhile.

A.2 Personal reflection and lessons learned

Creating vaccination prediction models is a time-consuming and iterative process that necessitates a thorough comprehension of the relevant research. Throughout the development process, I returned to the literature on a frequent basis, which considerably enhanced both my efficiency and my ability to read the literature.

The necessity of model interpretability was one of the important lessons I learned from reading the literature. While advanced machine learning algorithms might provide remarkable results, they are frequently challenging for stakeholders to grasp and explain. As a result, I promote model interpretability throughout the development process and use techniques like feature significance analysis and decision tree visualisation to improve model transparency.

Another important lesson I've learned through iterative development is the value of testing and validation. Throughout the model's development, I evaluated it against multiple datasets to confirm its robustness and reliability.

Overall, constructing a vaccination prediction model necessitates a thorough knowledge of the relevant literature as well as a commitment to iterate and enhance the model. I was able to design a model that was not just accurate but also transparent and trustworthy by stressing interpretability and validation. This experience has taught me the necessity of remaining up to date on the newest research and always tweaking the model to achieve the best outcomes possible.

A.3 Legal, social, ethical and professional issues

A.3.1 Legal issues

The main legal issue for this project is copyright in the dataset, which is from the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), and this data can only be used for health statistics reporting and analysis. This project is based around the analysis of this dataset and does not breach confidentiality regulations.

A.3.2 Social issues

The spread of influenza is a serious social hazard and this project is dedicated to addressing this issue. The main social problem of this project is still the access to data and the possibility of inviting some residents to take the questionnaire, but on the other hand, if the outbreak can be prevented, the project will have a sufficient positive impact.

A.3.3 Ethical issues

The ethical dimension concerns the leakage of data. In order to avoid the leakage of residents' personal information, which could have bad consequences, we should aim to protect the data while ensuring the quality of the model and preventing it from being used by unscrupulous people during the testing process.

A.3.4 Professional issues

The project itself is dedicated to solving social problems and the model can be licensed to the government itself

Appendix B

External Material

B.1 Datasets

https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/
https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html
https://www.cdc.gov/flu/fluvaxview/coverage_0910estimates.html

B.2 Tools

- GitLab: for version control;
- Jupiter notebook: Dedicated to data visualisation
- Atom editor and Spyder : for editing python files