

Final Report

|  |
| --- |
| School of Computing  Faculty of Engineering AND PHYSICAL SCIENCES |

<Full Title of Project>

<Full Name of Author>

Submitted in accordance with the requirements for the degree of  
<Name of Degree> (*e.g.* BSc Computer Science)

**<Session>** (*e.g.* 2021/22)

**<Module code and name>** (*e.g.* *COMP3931 Individual Project*)

The candidate confirms that the following have been submitted*:*

*<As an example>*

|  |  |  |
| --- | --- | --- |
| **Items** | **Format** | **Recipient(s) and Date** |
| *Final Report* | *PDF file* | *Uploaded to Minerva (DD/MM/YY)* |
| *Scanned participant consent forms* | *PDF file / file archive* | *Uploaded to Minerva (DD/MM/YY)* |
| *Link to online code repository* | *URL* | *Sent to supervisor and assessor (DD/MM/YY)* |
| *User manuals* | *PDF* | *Sent to client and supervisor (DD/MM/YY)* |

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.

(Signature of student)

© <Year of Submission> The University of Leeds and <full name of candidate>

# Summary

*<Concise statement of the problem you intended to solve and main achievements (no more than one A4 page)>*

Influenza is a highly contagious respiratory disease that can cause mild to severe symptoms including fever, cough, sore throat, nasal congestion, headache, muscle pain and fatigue. Influenza can cause serious complications such as pneumonia, heart disease, encephalitis and pulmonary embolism, and can even lead to death. Influenza viruses mutate constantly, making it difficult for the human immune system to mount an effective defense against them, making the pandemic unpredictable and uncertain. Influenza virus is also easily spread through coughing, sneezing and other ways, so it is very important to control and prevent influenza.

H1N1 influenza, also known as swine influenza, human swine influenza, is a viral infection caused by H1N1 subtype influenza virus. In 2009, H1N1 flu broke out in Mexico and then quickly spread around the globe, causing a severe global pandemic. The symptoms of H1N1 flu are similar to those of the regular flu and include fever, cough, sore throat, stuffy nose, headache, muscle aches and fatigue. Still, H1N1 flu is more contagious and sicker than normal flu, and can even cause death. H1N1 influenza can be transmitted through airborne transmission, direct contact and survival on contaminated surfaces for a period of time, so it is important to prevent and control H1N1 influenza. The spread of H1N1 flu can be effectively reduced through vaccination, frequent hand washing and avoiding contact with patients.

The flu vaccine is one of the most effective ways to protect against seasonal flu. Flu vaccines are usually made from inactivated influenza viruses and are administered by injection or nasal spray. Influenza vaccines usually include multiple strains of influenza virus to cover the different virus subtypes and variants that may emerge.

Flu vaccination helps the body's immune system produce antibodies to respond quickly to the flu virus when it is infected, reducing symptoms and reducing the risk of complications. The flu vaccine is usually given each year before the flu season, because the strain of the flu virus changes from year to year, so it needs to be re-vaccinated every year.

# Acknowledgements

*<This page should contain any acknowledgements to those who have assisted with your work. Where you have worked as part of a team, you should, where appropriate, reference to any contribution made by others to the project.>*

*Note that it is not acceptable to solicit assistance on ‘proof reading’ which is defined as “the systematic checking and identification of errors in spelling, punctuation, grammar and sentence construction, formatting and layout in the text”; see*

https:://www.leeds.ac.uk/secretariat/documents/proof\_reading\_policy.pdf

# Table of Contents (example of how to format)

Summary iii

Acknowledgements iv

Table of Contents v

Chapter 1 Level 1 Heading with ‘heading 1’ Style 1

1.1 Level 2 Heading with ‘heading 2’ Style 1

1.1.1 Level 3 Heading with ‘heading 3’ Style 1

1.1.1.1 Level 4 Heading with ‘heading 4’ Style 1

Chapter 2 Tables and Figures 2

2.1 Tables using the ‘table caption’ and ‘table description’ Styles 2

2.2 Figures using the ‘figure caption’ and ‘figure description’ Styles 2

List of References 3

Appendix A Self-appraisal 4

A.1 Level 2 Heading with ‘heading 2’ Style 4

A.1.1 Level 3 Heading with ‘heading 3’ Style 4

A.1.1.1 Level 4 Heading with ‘heading 4’ Style 4

Appendix B External materials 4

B.1 Level 2 Heading with ‘heading 2’ Style 4

# Chapter 1 Introduction and Background Research

## 1.1 Introduction

The flu shot is one of the most effective ways to protect against seasonal flu. Influenza vaccines are usually made from inactivated influenza viruses and are administered by injection or nasal spray. Influenza vaccines usually include multiple strains of influenza virus to cover the different subtypes and variants of the virus that may appear[1].

Influenza vaccination helps the immune system produce antibodies to respond quickly to the flu virus when it is infected, reducing symptoms and reducing the risk of complications. The flu shot is usually given each year before the flu season, because the strain of the flu virus changes from year to year, so it must be revaccinated every year[2].

Machine learning is an artificial intelligence technique that USES algorithms and statistical models so that computer systems automatically learn from the data and use the knowledge to accomplish specific tasks[3]. Machine learning can be used for classification, regression, clustering, dimension reduction, recommendation and other tasks, and is widely used in many fields, such as medical care, finance, logistics, image recognition, voice recognition, etc. The core of machine learning is to realize prediction and decision of unknown data by learning and summarizing the data[4]. Common machine learning algorithms include decision tree, support vector machine, naive Bayes, neural network and so on[5].

In this report, we trained the model using machine learning based on XGBoost to predict the uptake of H1N1 and seasonal influenza vaccines.

## 1.2 influenza

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[6]. 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[7].

The most effective way to prevent seasonal flu is to get the flu vaccine every year. In addition, good personal hygiene, such as washing hands after meals, minimizing contact with sick people and avoiding places where people gather, can effectively reduce the spread of influenza virus[8]. If you develop flu symptoms, seek medical attention immediately, take protective measures on the way to the doctor to reduce the possibility of infecting others, follow your doctor's advice, and treat the flu scientifically.

The seasonal flu vaccine is one of the most effective ways to prevent seasonal flu. It stimulates the body's immune response by injecting a vaccine containing influenza virus antigens, and produces specific antibodies to kill the virus specifically, thus reducing the risk of getting infected with influenza virus. The seasonal flu vaccine is updated annually based on changes in the seasonal flu virus to ensure its effectiveness[9].

The seasonal flu vaccine is best received starting in the fall, and is recommended every year for those who can afford it. Getting the seasonal flu vaccine does not cause influenza virus infection, although sometimes mild side effects, such as soreness at the injection site, redness and fever, can be manageable. Seasonal influenza vaccination is more important for certain groups of people, such as pregnant women, children, the elderly and people with chronic diseases, because they are susceptible to influenza viruses and suffer more severe illness when they do get sick[10].

People infected with seasonal influenza virus will appear fever, cough, sore throat, headache, muscle pain and other symptoms, seriously affect the life and work status. In certain populations, such as the elderly, children, pregnant women and people with chronic diseases, influenza virus infection can lead to more serious complications and even life-threatening conditions[11].

Seasonal influenza is not only damaging to individuals, both large and small, but also has serious consequences for society. Seasonal influenza often causes large numbers of workers to take sick leave, which can affect social production and operations. In addition, an increase in influenza cases may lead to increased stress on the health care system, increasing the consumption of health care resources, thereby crowding out health care resources and increasing mortality from other conditions[12]. During seasonal influenza outbreaks, when a large number of employees are infected with the virus, they need to take leave or be quarantined, which will affect production and operations. This will lead to the reduction of production efficiency of enterprises and the increase of production costs, which will lead to the economic downturn of enterprises and even the society.

In addition, the strain on health systems will increase dramatically, requiring significant additional resources to treat and control the spread of disease. This will lead to increased consumption of medical resources and increased medical costs, which will have an impact on the social economy. Medical staff also have to deal with a large number of patients with the flu virus during the peak flu season, which adds to the stress of their work. Medical staff need to work overtime and longer hours to respond to patients' demands, which affects their health and productivity and thus the treatment of other conditions[13]. It would also lead to increased demand for medical equipment and drugs, as well as the large number of beds needed to accommodate patients, potentially creating shortages that would affect patient treatment and rehabilitation. In addition, treating and controlling the spread of the disease will consume a large amount of medical resources, which will lead to increased consumption of medical resources and medical expenses, thus putting pressure on the medical system. This pressure will affect the way health care systems focus on other diseases, leading to indirect deaths from influenza[14].

1.2.2 H1N1 influenza

H1N1 influenza, also known as swine flu, is a disease caused by the H1N1 influenza A virus. H1N1 is an RNA virus that belongs to the orthomyxovirus family. Its hosts are birds and some mammals[15]. Almost all the A (H1N1) viruses have been isolated, and disease in wild birds is rare. Some H1N1 viruses cause severe illness mostly in poultry and pets, but rarely in humans. However, by spreading and mutating in birds and some mammals, the virus can also have a dramatic effect on humans, which can be more severe than seasonal influenza, which could lead to outbreaks or widespread spread of human influenza[16]. The novel influenza A (H1N1) virus was first identified in case reports in Mexico and the United States in April 2009. Since then, the virus has spread rapidly around the world, causing influenza pandemics in many countries, generating enormous and irreversible economic losses, and becoming a formidable public health challenge[17].

The H1N1 flu virus is very similar to seasonal flu, but it is more contagious, but the virus mutates more quickly and the symptoms are more severe. The virus can also be spread through air, contact and droplets. The H1N1 flu virus is transmitted from group to group mainly through the coughing and sneezing of infected people, so wearing masks is an effective preventive measure[18]. Swine flu virus infections are more likely to occur in crowded environments, and there is growing evidence that trace amounts of the virus can remain on desktops, phones or other surfaces and be transmitted through finger-to-eye, nose and mouth contact. Therefore, physical contact, including shaking hands, kissing and eating together, should be avoided during an epidemic. People are also most likely to become infected if they touch an object carrying the influenza A (H1N1) virus and then touch their nose or mouth. An infected person is likely to infect others before showing symptoms and often becomes ill a week or more after infection, making it more difficult to prevent outbreaks[19].

Like seasonal flu, swine flu is highly damaging to individuals and society, and more damaging because it is more contagious, mutates more quickly and causes more severe symptoms.

Among the swine influenza vaccines that have been developed and distributed, the most technologically mature ones that have been put into production are mainly monovalent or bivalent h1n1 subtype and H3N2 subtype swine influenza whole virus inactivated vaccines[20]. Most vaccines are oil adjuvant vaccines, and inactivated agents are usually formaldehyde or BEI. Inactivated swine influenza vaccine has been reported to be effective in protecting weaned piglets and breeding pigs from SI infection, reducing morbidity by 30 to 70 percent and mortality by 60 to 87 percent. These vaccines are already on the market in many countries, Examples include Intervet's End-Fluence (including Imugen) and End FLUence(including Microsol Diluvac Forte), Pfizer's FluSure and Schering-Plough Animal Health's MaxiVac fiu and MaxiVac Excell 3 Wait. South China Agricultural University, Harbin Veterinary Research Institute and other institutions have also made major breakthroughs in the development of swine influenza oil emulsion inactivated vaccine. Their results showed that the H1 subtype was prepared by breeding the domestic isolates. The titer of Hl antibody could reach more than 1:160 at 3 weeks after single immunization, the level of HI antibody could reach more than 1:40 at 4 weeks after immunization, and the titer of 12NHI antibody could remain at more than 1:100, exceeding the antibody positive standard of 1:80[21]. If two exemptions are made, the HI antibody level is high, and the effective protection period can be as long as 6 months or more, which can fully meet the epidemic prevention needs of various day-old pigs. These vaccines have passed clinical trials and are in the process of being certified as new veterinary drugs. It is believed that in the near future, the domestic commercial vaccine will be put on the market and occupy a certain market share.

Although swine influenza whole virus inactivated vaccine has the advantages of safety, efficiency, low production cost and long duration of effective antibodies, it also has some disadvantages such as slow production of antibodies, weak ability to induce cellular immunity and strong stress response. In particular, most vaccine strains come from circulating strains. Listen, if the virus is leaked during vaccine preparation, it could easily cause environmental contamination and trigger a larger outbreak[22]. To this end, scientists are still exploring the development of a new swine flu vaccine.

**1.3 XGBoost**

1.3.1 CART decision tree algorithm

CART decision tree algorithm (classification and regression tree) is a classical decision tree learning method. CART decision tree algorithm can solve both classification problem and regression problem[23]. Therefore, XGBoost model based on CART regression tree is used as one of the base models in this paper.

The CART decision tree is a widely used decision tree algorithm. Its construction algorithm consists of three steps :(1) selection of characteristics, (2) generation of the decision tree and (3) pruning of the decision tree. The CART decision tree is a binary tree. Discrete values are classified directly according to the classification rules and continuous values are divided into ranges. The generation of a CART decision tree is done by a characteristic based recursive binary operation that results in an ideal decision tree. Splitting a decision tree is an operation that optimizes a decision tree. In recursive binary operations ina CART decision tree, overshooting can occur if the segmentation is successive, creating a tree with a very complex structure. Cutting optimizes the structure of the CART decision tree and reduces overrides for better results[24].

CART decision tree algorithm can solve both classification problem and regression problem. CART decision tree algorithms also use different strategies for different problems. For regression problem, CART decision tree algorithm uses minimum square error criterion. For classification problem, CART decision tree algorithm chooses Gini index minimization criterion. Since the XGBoost model used in this article uses the CART regression tree as the base learner, only the regression tree generation strategy is described here[25].

The generation algorithm of CART classification tree is to select the feature variable with minimum Gini coefficient when splitting nodes

To split a node, perform the following steps.

1. Firstly, the Gini coefficient of the total data set is calculated. For the data set D, there are K classifications. The sample sets after classification are C1 , C2,...Ck respectively, and their Gini coefficient is defined as:



(2) Calculate the Gini coefficient of each feature variable. The total data set D is divided into D1 and D2 parts according to the possible value of each attribute of feature A, then the Gini coefficient of set D under the condition of feature A is defined as:



(3) The characteristic variable with the minimum Gini coefficient is selected as the optimal splitting node, and the value of the branch is the attribute classification that can obtain the minimum Gini coefficient, and two branches are generated.

(4) Repeat the above steps until the category contained by the split node is unique, and CART can be generated

Classification tree.

**1.3.2 XGBoost model**

XGBoost (Extreme Gradient Boosting), Boosting is an ensemble learning method based on residuals. Ensemble learning, also known as multi-classifier system and committee-based learning method, is to train multiple weak learners by gathering them together and combining them through certain strategies to finally get a strong learner[26]. Integrated learning is generally divided into two categories: homogeneity and heterogeneity. As the name implies, homogeneity refers to the integration of only the same learning algorithm, such as the inclusion of decision tree algorithm, while heterogeneous refers to the inclusion of different learning algorithms. In general, homogeneous learning algorithms are used more often[27]. In the homogeneous learner, according to the dependence between algorithms can be divided into two kinds. Boosting class integration algorithm with strong dependency, while bagging class integration algorithm with less strong dependency.

The following figure shows the process of boosting algorithm:

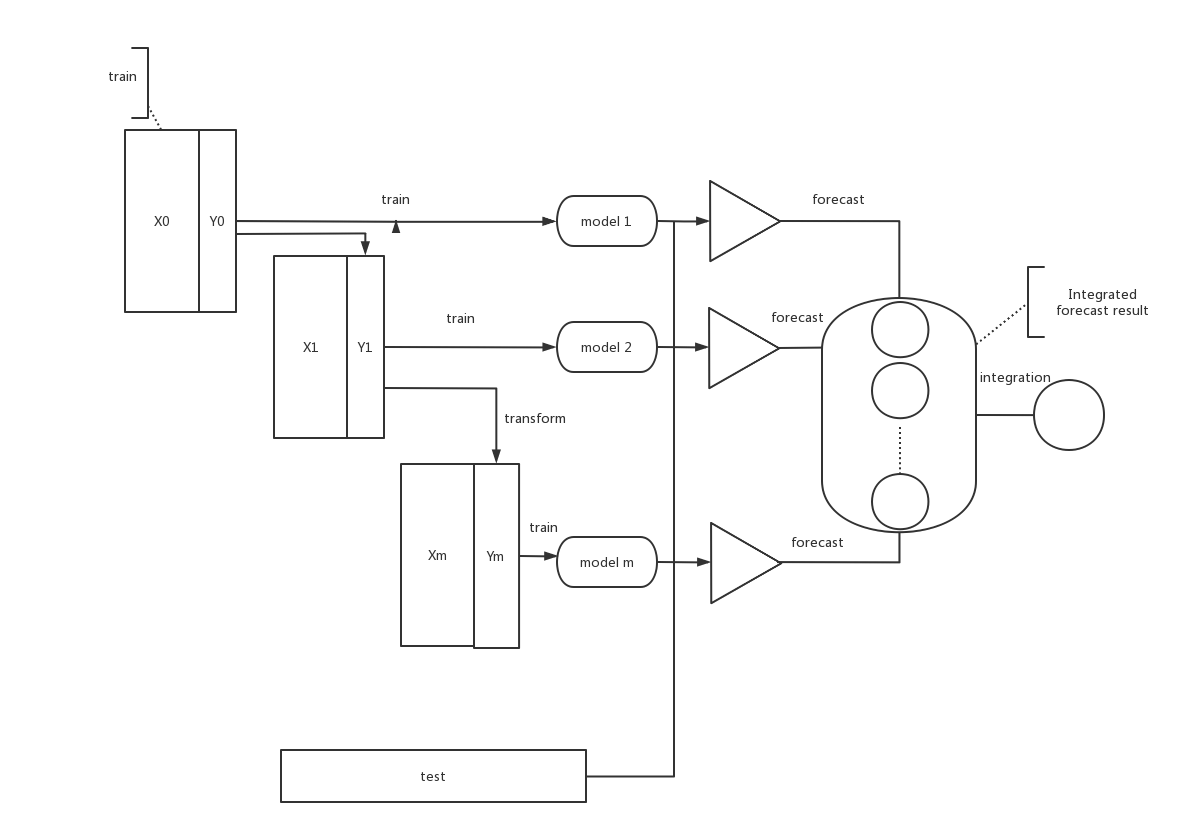


Figure1

According to Figure1, Boosting ensemble algorithm will learn the training residuals of previous weak learners during the cycle process, so as to make the loss function drop along the gradient and finally synthesize the predicted result.

XGBoost model is a model using Boosting method, which is a concrete implementation of GBDT algorithm. It is an addition model composed of k CART regression tree models, and the result obtained at last can be relatively close to the real value.

The loss function of XGBoost model is defined as follows:



The above equation consists of a loss function l and a regular term Ω that prevents overfitting by controlling the complexity of the model. Where, Ω is the complexity of the model by adding the complexity of all the trees t in the model.

The XGBoost model, as a model that uses the Boosting integration method, naturally follows the rules of Boosting integration method, which uses forward distribution addition. When performing the t iteration, the model will obtain the result of the t step according to the predicted result given in the previous step and the predicted value calculated in this iteration[28]. The formula is as follows:



**1.4 Random forest**

Random Forest (RF), an emerging and highly flexible machine learning algorithm, has a wide range of applications, from marketing to health care insurance. It can be used to model marketing simulations, count customer origin, retention and loss, and predict disease risk and patient susceptibility[29].

Random forest is an algorithm that integrates multiple trees through the idea of Ensemble Learning. Its basic unit is decision tree, and its essence belongs to ensemble learning, a branch of machine learning. The name of random forest has two keywords, one is "random", the other is "forest". "Forest" is very well understood. If one tree is called a tree, then hundreds of trees can be called a forest. This analogy is quite appropriate, and in fact, it is the embodiment of the main idea of random forest -- the integration idea[30].

Random forest is a very flexible and practical method, which has the following characteristics:

1. It is unexcelled in accuracy among current algorithms；
2. It runs efficiently on large data bases；
3. It can handle thousands of input variables without variable deletion；
4. It gives estimates of what variables are important in the classification；
5. It generates an internal unbiased estimate of the generalization error as the forest building progresses；
6. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

In fact, random Forest has more than just these six features. It's the Leatherman of machine learning. You can throw almost anything into it, and it's basically usable. It is particularly useful in estimating inferential maps, so that you do not need to do as much parameter debugging as SVM[31].

**1.5 SHAP**

In today's financial, medical and other data mining applications, the interpretability and precision of the model are equally important. As we all know, models with high precision, such as integration model and deep learning model, have complex and changeable internal structures, which cannot be intuitively understood. SHAP, as a classical post-interpretation framework, can calculate the importance value of each characteristic variable in each sample to achieve the effect of interpretation. This Value is specifically called the Shapley Value in SHAP[32]. Therefore, Shapley Value is the core of the SHAP method. Understanding the meaning behind the value will greatly help us to understand the idea of SHAP. Shapley value was first proposed by Lloyd Shapley, a professor from UCLA, and is mainly used to solve the distributive equilibrium problem in cooperative game theory[33].

SHAP is the full name of SHapley Additive exPlanation, which belongs to the method of after-the-fact model explanation, and can explain complex machine learning models. It's derived from game theory, but it's based on that idea. For local interpretation, the core of SHAP is to calculate the Shapley Value of each of the characteristic variables.

When we do post-shap interpretation of the model, we need to be explicit. Given the data set (with M characteristic variables and n samples), the original model f, and all the predicted values of the original model f on the data set. g is the model used in SHAP to explain f.

First, f is used to predict the data set, and the average value of the model prediction is 90. A single sample is expressed as x=(x1,x2...xM).f(x) is the predicted value under the original model. g(x) is the predicted value of the post-interpretation model, satisfying . Where  represents the Shapley Value of the ith characteristic variable, which is the value to be calculated by the core in SHAP and needs to satisfy uniqueness.

1.6 Purpose

We found that using SHAP to explain the black box model Xgboost has become quite popular recently. For example, parsa et al. used XGBoost to robustly detect traffic accident-related features on Chicago Metropolitan Expressway, and then used SHAP to interpret the results and analyze the importance of individual features. By constructing a comprehensive interpretation framework of landslide susceptibility evaluation model based on SHAP-XGBoost algorithm, Zhang et al. analyzed the regional characteristics and spatial heterogeneity of landslide influencing factors, and discussed the model's generalization heterogeneity under different landscapes. Maciej et al. established the prediction model of HGI using XGBoost, used SHAP additive interpretation technology to explain the relationship between the predicted value and the input data, and found the influence of various coal parameters on the prediction of grindability of mixed coal. Therefore, we also used the SHAP-XGBoost algorithm model to predict the effectiveness of H1N1 and seasonal influenza. In addition, building a random forest model is more difficult to learn bagging and boosting.

## 

# Chapter 2 Methods

2.1 Principle of machine learning

2.1.1 Machine learning method

Machine learning is a multi - domain cross technology, its essence is a special algorithm. By analyzing big data, we find potential patterns within the data and apply these patterns to make predictions[34]. Suppose that a training data set D contains n learning samples, and each sample has m influence parameters and an output parameter,

,

I=1,2,...,n and n≥m

Select the corresponding machine learning algorithm to build a prediction mode  in this data set，The hyperparameter  of the algorithm is optimized to minimize the following error function values



Where: f -- the model established by a machine learning algorithm;  -- the hyperparameter of the algorithm; -- the error function.

2.1.2 Interpretability of machine learning models

Machine learning model interpretability provides users with a channel to understand machine learning model, which is not only the agent of machine learning model, but also a method to explain the model.

Model interpretability is divided into two categories: ex ante interpretability and ex post interpretability. Ex ante explainability means that the model with good explainability is adopted or the model with explanatory ability is designed so that the model itself has explanatory ability. Post-interpretation is the interpretation of the established machine learning model, independent of the machine learning modeling process, strong flexibility[35]. According to the scope of explainability, post-explainability can be divided into global and local. Global explainability is used to understand the internal working principle of the model and explain the overall capacity of the model. Local explainability is used to understand the prediction process and basis of machine learning model for a single sample, and explain the prediction results of the sample[36].In this paper, the prediction results of H1N1 vaccine and influenza vaccine effectiveness model are explained, which belongs to post-explanatory, using SHAP explainable technology.SHAP method is an additive method for model interpretation by calculating the predicted contribution value of each input variable. For example, when using SHAP method to interpret the machine model predicted value  of sample x\*, the predicted value  can be decomposed into



Where:  -- the average predicted value of the prediction model f (x) on the data set;  -- contribution value of the Jth input variable to the prediction of sample x\*, namely, SHAP value; M -- the number of input variables.

# Chapter 3 Results

## 3.1 Dataset acquisition

需要你自己写

## 3.2 Preparing the Dataset

The large data volume of the influenza vaccine data set collected in this article has led to some problems, such as collection errors, large format differences, null values, and wide differences in data ranges. Therefore, there is no way to train the model without processing the original data, so a series of processing should be carried out on the original data first. The purpose of data preprocessing is to remove the worthless information and process the data format according to the standard, so as to obtain the effective data set.

Here are the steps for preprocessing raw data.

（1）Missing value processing

In the data set we collected, there were no missing values, so all variables could be considered during modeling. See Figure 2 and Figure 3 for details.

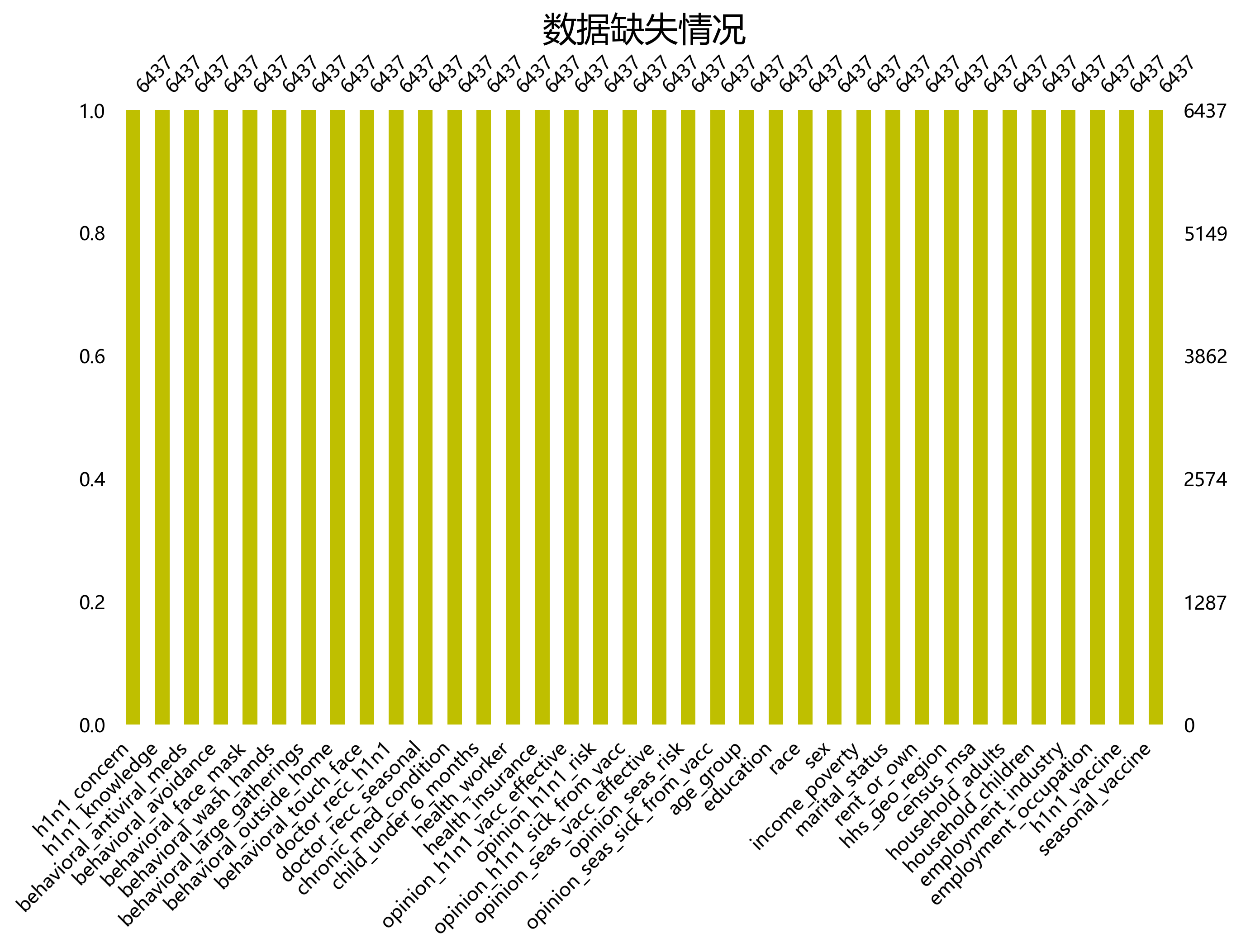


图2

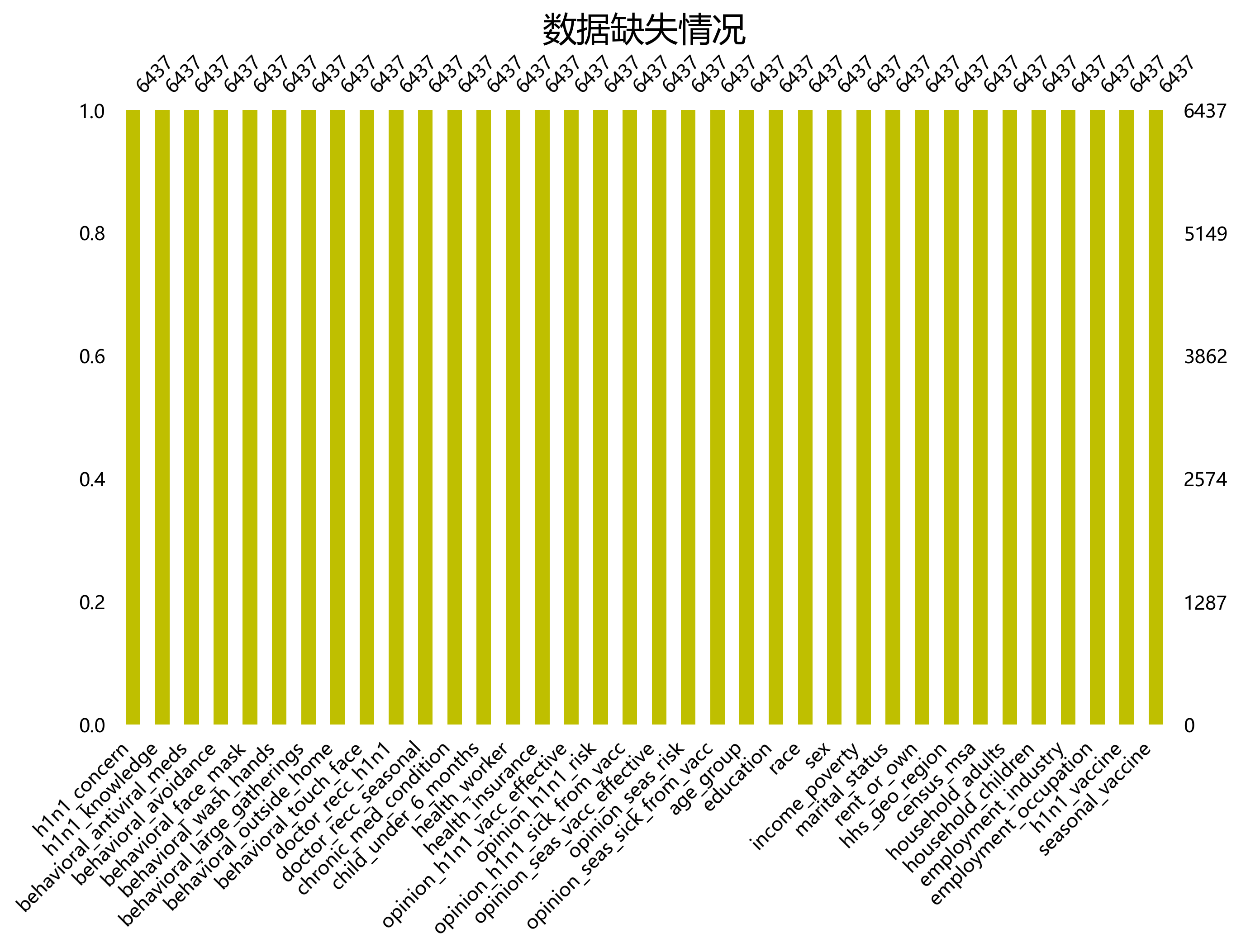


图3

（2）Outlier processing

There are unreasonable values in the original data, which will affect the prediction accuracy of the model. The outlier detection method based on boxplot is adopted to use the upper and lower quartiles (Q3, Q1) and interquartile distance (IQR) in the data. The interquartile distance is defined as the difference between the upper quartile and the lower quartile, that is, IQR=Q3-Q1, and the variable value greater than Q3+1.5IQR and less than q1-1.5iQR are the outliers.

1. Correlation analysis

If the input parameters with strong correlation are used, the modeling will not only increase the training time of the model, but also affect the interpretability of the model.

Figure 4 shows the Pearson correlation coefficient between the variables of H1N1 influenza and seasonal influenza. We find that opinion\_seas\_risk has the strongest correlation with opinion\_h1n1\_risk, which reaches 0.6. In addition, the correlation coefficient between opinion\_seas\_risk\_from\_vacc and opinion\_h1n1\_risk\_from\_vacc reached 0.53. The correlation coefficient between opinion\_seas\_vacc\_effective and opinion\_h1n1\_vacc\_effective is 0.51. This indicates that the above three groups of variables have a high degree of internal inertia, so a stepwise backward regression model is adopted to select the input parameters. The results show that opinion\_h1n1\_risk, opinion\_h1n1\_risk\_from\_vacc and opinion\_h1n1\_vacc\_effective can be deleted.

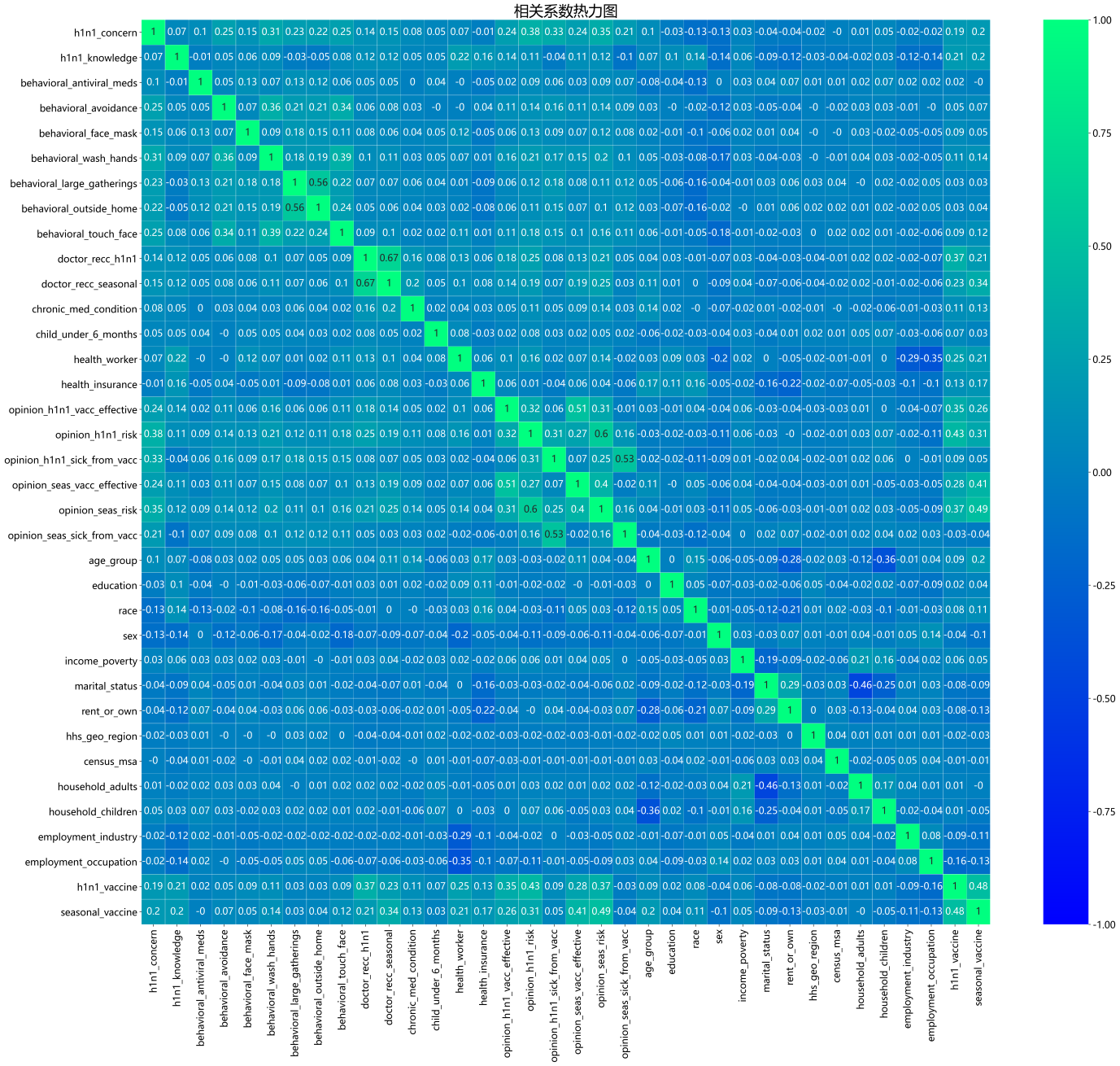


Figure 4

3.3 Construction of prediction models

The tree model is used to screen characteristic variables. The tree model can give the importance ranking of variables and finally specify the top several variables as the result of variable selection. Such tree models include decision tree model, random forest model and XGBoost model. XGBoost is divided according to structure

The gain situation of the number can be calculated to select which feature variable as the segmentation point. Therefore, the more a feature variable is used to construct the decision tree in the model, the more important it is.

The pre-processed data set was randomly divided into training set and test set according to the proportion of 80%~20%. XGBoost and random forest algorithm were used to learn from the training set to build the prediction model.

The ROC curve of the model is as follows:

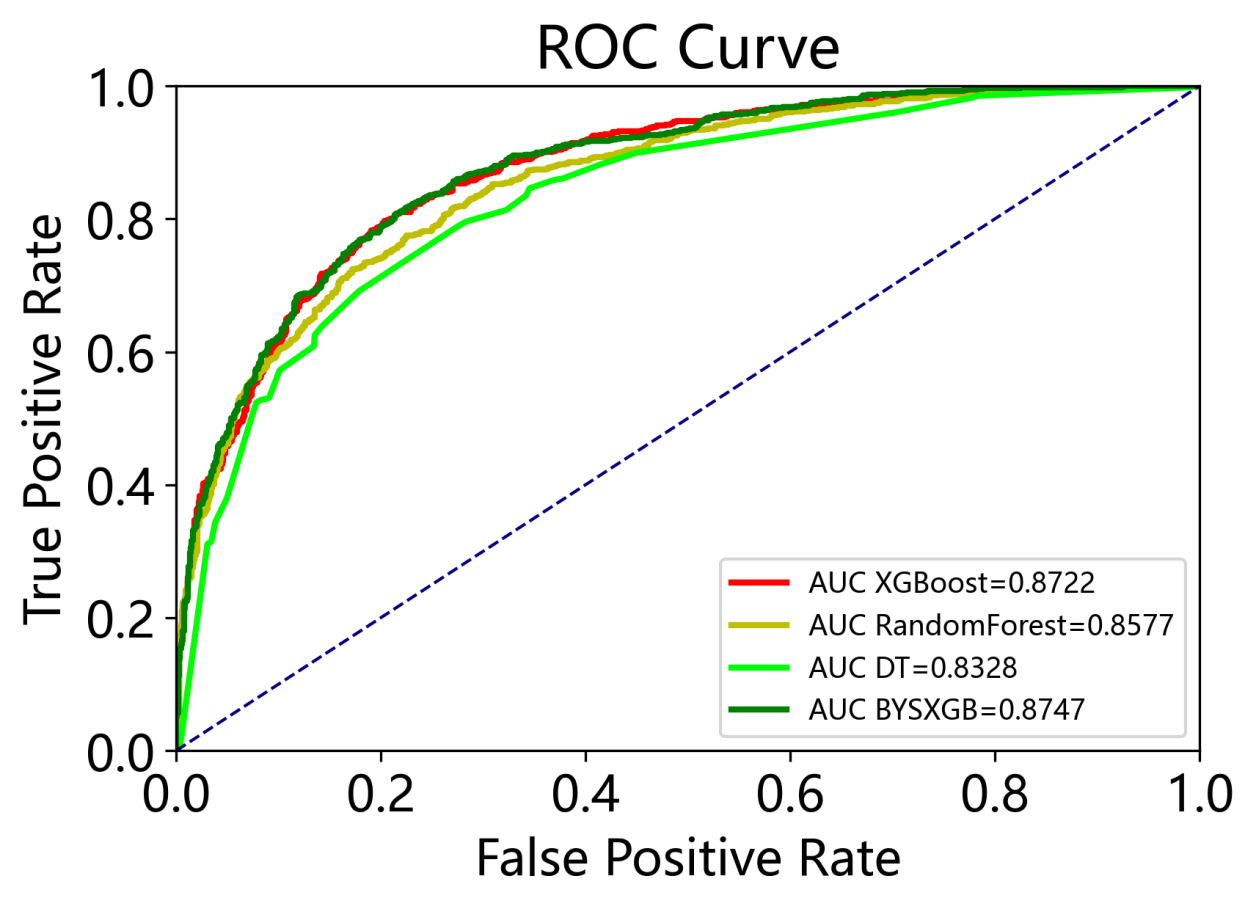


Figure 5

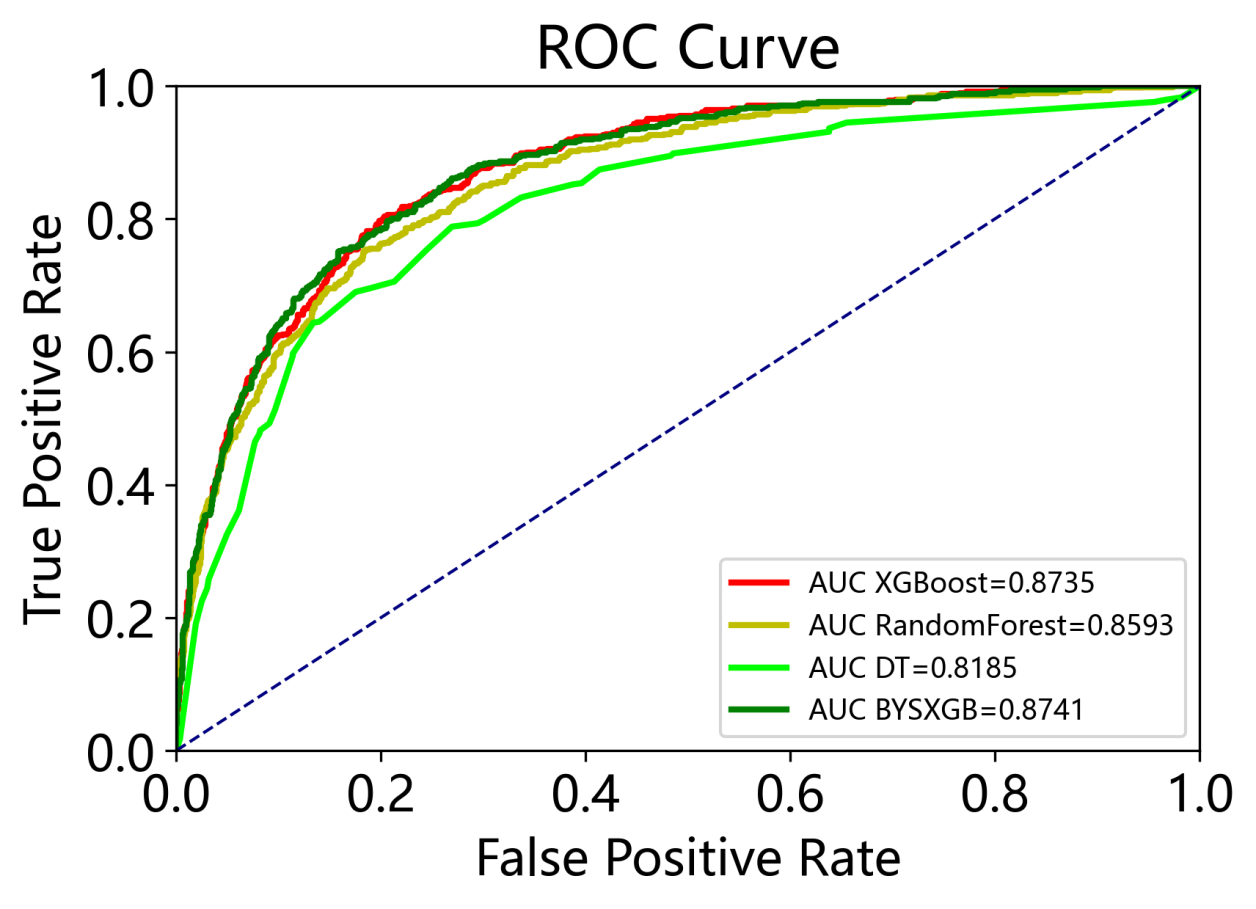


Figure 6

As can be seen from the figure, the curve area of XGBoost model for seasonal influenza is 0.8722, and that of random forest model is 0.8577. The XGBoost model of H1N1 influenza had a curve area of 0.8735 and the random forest model had a curve area of 0.8593.

3.4 Model interpretation

Table 1 Seasonal influenza

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Train | | | | | Test | | | | |
| AUC | accuracy | F1\_score | precision | recall | AUC | accuracy | F1\_score | precision | recall |
| XGBoost | 0.932 | 0.854 | 0.838 | 0.856 | 0.820 | 0.872 | 0.795 | 0.774 | 0.782 | 0.765 |
| Random forest | 0.872 | 0.778 | 0.716 | 0.864 | 0.705 | 0.833 | 0.762 | 0.727 | 0.766 | 0.691 |

Table 1 H1N1 influenza

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Train | | | | | Test | | | | |
| AUC | accuracy | F1\_score | precision | recall | AUC | accuracy | F1\_score | precision | recall |
| XGBoost | 0.909 | 0.857 | 0.738 | 0.822 | 0.670 | 0.874 | 0.819 | 0.668 | 0.744 | 0.606 |
| Random forest | 0.893 | 0.824 | 0.612 | 0.902 | 0.463 | 0.859 | 0.792 | 0.539 | 0.813 | 0.403 |

Since no machine algorithm is absolutely superior to other algorithms, the prediction performance of the machine learning model in the test data set is mainly evaluated to select the most suitable algorithm. Therefore, the established machine learning model is applied to predict the vaccine coverage rate in the test set. Table 1 and Table 2 are the comparison of the prediction performance of these models.

We can see that in the seasonal influenza model, the accuracy of XGBoost algorithm is higher than that of the random forest model. The same results were seen in the H1N1 influenza model.

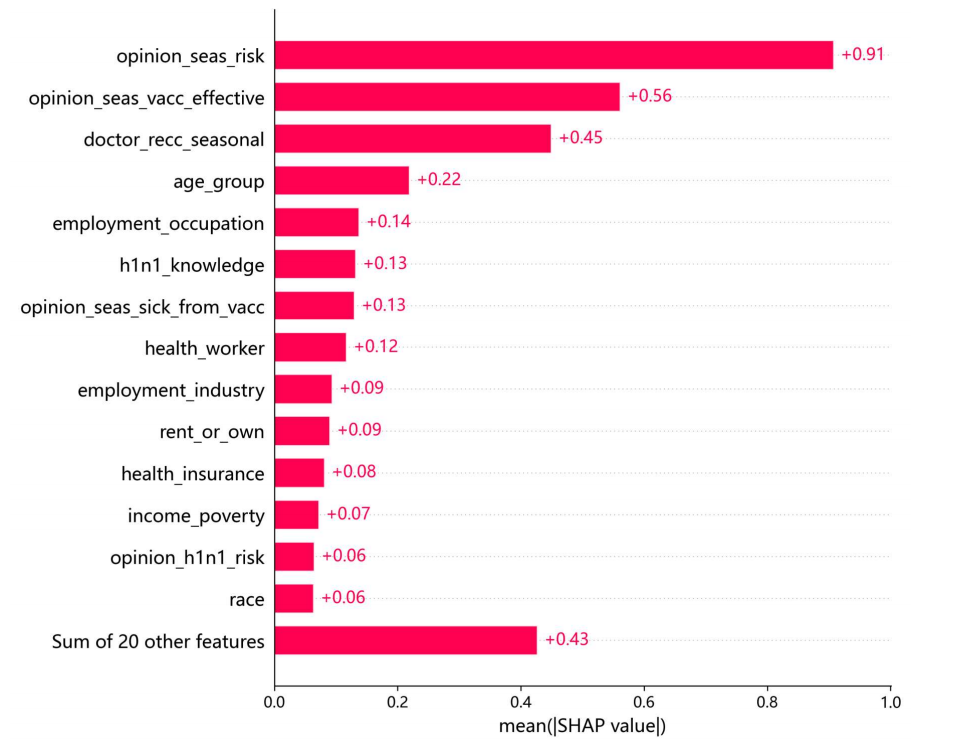


Figure 7

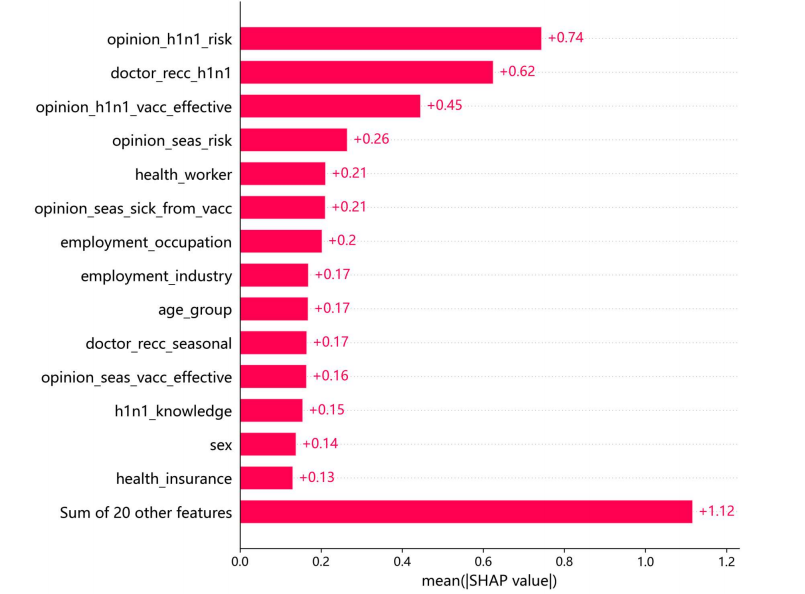


Figure 8

Figure 7 illustrates the XGBoost model. Among the factors influencing vaccination rates for seasonal influenza, people's perceptions of the risk of seasonal influenza ranked first, followed by people's perceptions of vaccine effectiveness. In the H1N1 flu model, it's a little different. People's perceptions of the dangers of H1N1 flu come first, but doctors' reccurence of H1N1 flu takes second place.

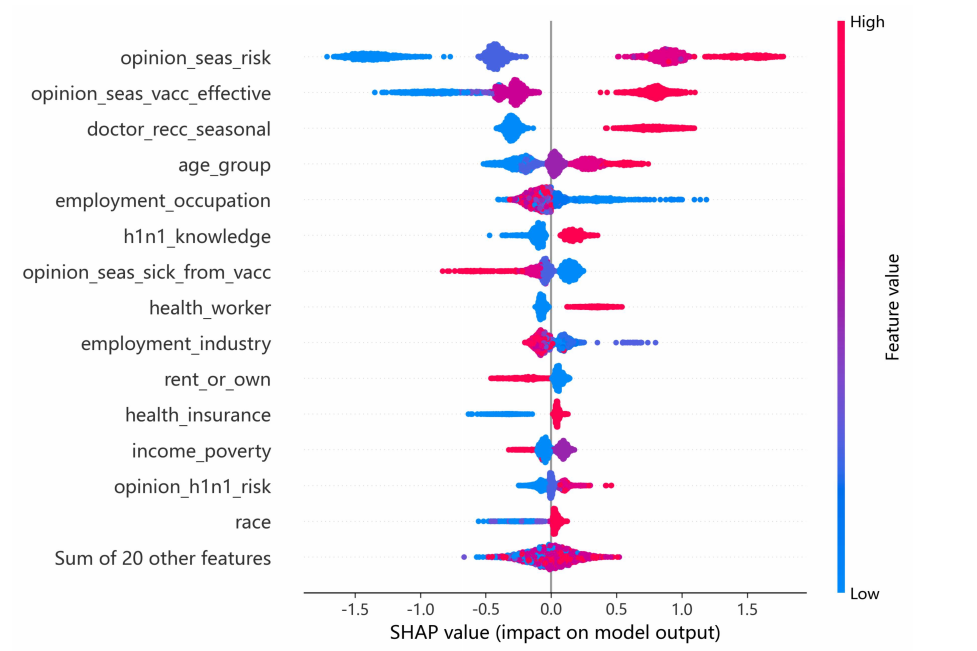


Figure 9

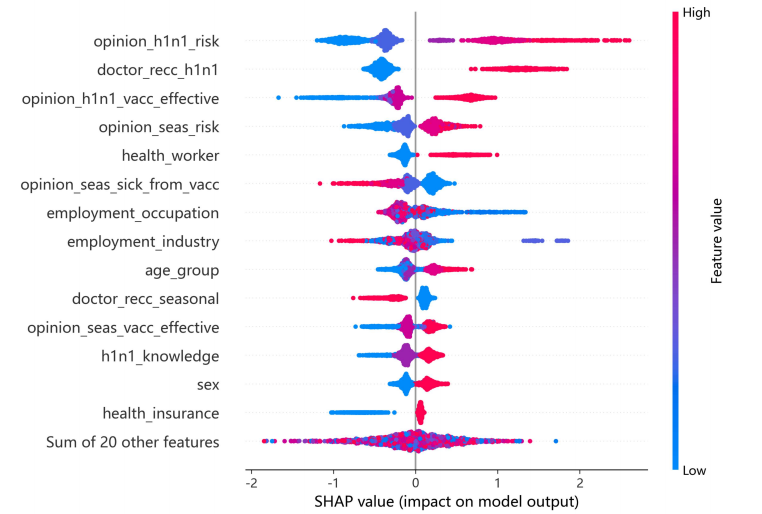


Figure 10

FIG. 9 and FIG. 10 show the distribution of SHAP values of each influencing factor in the two models. In the figure, each data point represents a respondent, the color represents the value of the variable, and the change of variable data from low to high from blue to red.

SHAP value of positive or negative) indicates that the influence parameters are positively (negatively) correlated with vaccination rates. For example, the more seriously people perceive the risk of seasonal influenza, the higher the vaccination rate for seasonal influenza.

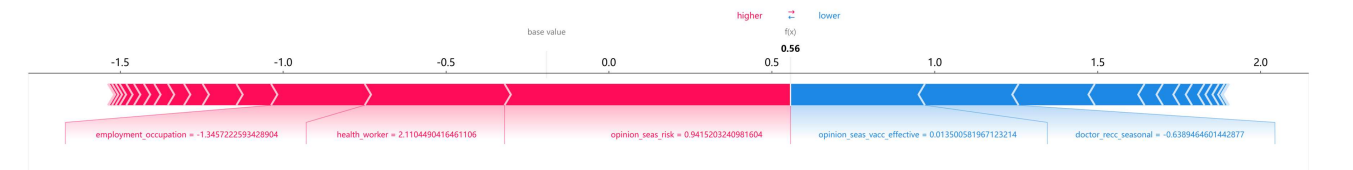


Figure 11

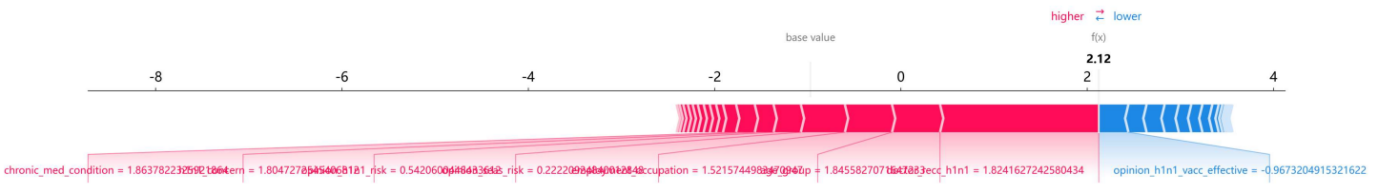


Figure 12

Figure 11 and Figure 12 give a partial explanation of the two models. The vaccine coverage rate was the sum of contributions of all input parameters, and the baseline value was the average of the vaccination rates in the XGBoost model, which were -2.32 and 2.12, respectively. Red causes the average to increase, while blue causes it to decrease. SHAP values for perceptions of the effectiveness of seasonal influenza vaccine and perceptions of the danger of influenza were -0.0135 and 0.9415, respectively. The SHAP for people's perceptions of the effectiveness of the H1N1 flu vaccine was -0.9673.

# Chapter 4 Discussion

## 4.1 Conclusions

(1) A method based on machine learning algorithm for prediction and model interpretation of seasonal influenza vaccine and H1N1 influenza vaccine coverage was proposed. This method is an effective supplement to the existing prediction methods of vaccination coverage. It has multi-type data integration and strong prediction ability, and can accurately improve the control of the epidemic situation by medical institutions or the government, with strong practicability.

(2) Learning the global interpretability of the model helps to understand the coverage prediction mechanism of the XGBoost algorithm, and make it clear that the main parameters affecting the coverage rate of seasonal influenza vaccine and H1N1 influenza vaccine are people's perceptions of vaccine effectiveness and safety.

(3) SHAP value can not only help to analyze the factors affecting the vaccination rate of seasonal influenza vaccine and H1N1 influenza vaccine, but also understand the influence of interaction between input variables on the vaccination rate.

(4) For the first time, interpretable machine learning algorithm was applied to predict the vaccination rate of seasonal influenza vaccine and H1N1 influenza vaccine. The two machine learning algorithms were applied, and the comparison between the models showed that XGBoost algorithm was more suitable for predicting the vaccination rate. This can provide a basis for future projections of vaccination rates against infectious diseases.

## 4.2 Ideas for future work

There are still many steps to using machine learning to predict vaccination rates. The accuracy of the prediction depends on the quality and reliability of the data. If the data is missing, inaccurate, or not sufficiently characterized, the results predicted by the model will be affected. Although we have selected relatively rich data, there are still some limitations. Moreover, machine learning models can only predict what has already happened, not fully adapt to what might happen in the future. If the model is not adaptive enough, the prediction results may not be accurate.

The prediction of vaccination rates is influenced not only by medical factors, but also by socioeconomic factors, such as income, education level and cultural background. These aspects are rarely considered in the current data, so the results of the model may be biased. Therefore, in future work, expanding the scope of consideration can better predict the vaccination rate.

# List of References

1. Longini I M , Elizabeth H M . Strategy for Distribution of Influenza Vaccine to High-Risk Groups and Children[J]. American Journal of Epidemiology, 2005(4):303-306.
2. Nichol K L , Nordin J D , Nelson D B , et al. Effectiveness of influenza vaccine in the community-dwelling elderly.[J]. New England Journal of Medicine, 2007, 357(14):1373-81.
3. Goldberg D E . Genetic Algorithm in Search, Optimization, and Machine Learning[M]. Addison-Wesley Pub. Co, 1989.
4. Fulkerson, Bill. Machine Learning, Neural and Statistical Classification[J]. Technometrics, 1994, 37(4):459-459.
5. Bradley P . The use of the area under the ROC curve in the evaluation of machine learning algorithms[J]. Pattern Recognition, 1997, 30( 7):1145-1159.
6. 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).
7. Liu Vera,Walker Stephen. Testing for genetic mutation of seasonal influenza virus[J]. Journal of Applied Statistics,2023,50(1).
8. Huong Q. McLean,Joshua G. Petrie,Kayla E. Hanson,Jennifer K. Meece,Melissa A. Rolfes,Gregg C. Sylvester,Gabriele Neumann,Yoshihiro Kawaoka,Edward A. Belongia. Interim Estimates of 2022–23 Seasonal Influenza Vaccine Effectiveness — Wisconsin, October 2022–February 2023[J]. MMWR. Morbidity and Mortality Weekly Report,2023,72(8).
9. Aoun Tia,Borrow Ray,Arkwright Peter D. Immunogenicity and safety of seasonal influenza vaccines in children under 3 years of age.[J]. Expert review of vaccines,2023.
10. de Lusignan Simon,Ashraf Mansoor,Ferreira Filipa,Tripathy Manasa,Yonova Ivelina,Rafi Imran,Kassianos George,Joy Mark. Impact of General Practitioner Education on Acceptance of an Adjuvanted Seasonal Influenza Vaccine among Older Adults in England[J]. Behavioral Sciences,2023,13(2).
11. Fadl Noha,Al Awaidy Salah T,Elshabrawy Abdelhamid,Makhlouf Mona Sayed Aly Hassan,Ibrahim Sarah Assem,AbdelRahman Suzan,Tookhy Nazir Ahmad,Alsalmani Abdullah,AlSaeedi Mays,AlSawalha Ibrahim,ElDin Mohammad Amin Aly,Saad Janet,Ayoob Zainab,Rourou Mohamed Khalil,Ali Manahil,Tawati Salha M,Gadain Yahia Marwan Ahmed,AlSaidi Sara Yunis,Hassan Ghadeer Ali,Alsanafi Mariam,Sandouk Leen,Youssef Naglaa,Alothman Shaykhah,Yazbek Saja,AlAnsi Khlood Saleh,Mehdad Slimane,Adam Mohammed Fathelrahman,Gebreal Assem,Ghazy Ramy Mohamed. Determinants of parental seasonal influenza vaccine hesitancy in the Eastern Mediterranean region: A cross-sectional study.[J]. Frontiers in public health,2023,11.
12. SamelKowalik Piotr,Jankowski Mateusz,LisieckaBiełanowicz Mira,Ostrowska Aurelia,Gujski Mariusz,Kobuszewski Bartosz,Pinkas Jarosław,Raciborski Filip. Factors Associated with Attitudes towards Seasonal Influenza Vaccination in Poland: A Nationwide Cross-Sectional Survey in 2020[J]. Vaccines,2021,9(11).
13. Kuliese Monika,Mickiene Aukse,Jancoriene Ligita,Zablockiene Birute,Gefenaite Giedre. Age-Specific Seasonal Influenza Vaccine Effectiveness against Different Influenza Subtypes in the Hospitalized Population in Lithuania during the 2015–2019 Influenza Seasons[J]. Vaccines,2021,9(5).
14. Shields Gemma E,Elvidge Jamie,Davies Linda M. A systematic review of economic evaluations of seasonal influenza vaccination for the elderly population in the European Union.[J]. BMJ open,2017,7(6).
15. Fairchok M P . A Novel Influenza A (H1N1) Vaccine in Various Age Groups. 2009.
16. Newman A P , Reisdorf E , Beinemann J , et al. Human Case of Swine Influenza A (H1N1) Triple Reassortant Virus Infection, Wisconsin[J]. Emerging Infectious Diseases, 2008, 14(9):1470-2.
17. Hatano B , Goto M , Fukumoto H , et al. Mobile and accurate detection system for infection by the 2009 pandemic influenza A (H1N1) virus with a pocket‐warmer reverse‐transcriptase loop‐mediated isothermal amplification[J]. Journal of Medical Virology, 2011, 83(4).
18. Yang S G , Cao B , Liang L R , et al. Antiviral Therapy and Outcomes of Patients with Pneumonia Caused by Influenza A Pandemic (H1N1) Virus[J]. Plos One, 1932, 7.
19. Vellozzi C , Broder K R , Haber P , et al. Adverse events following influenza A (H1N1) 2009 monovalent vaccines reported to the Vaccine Adverse Event Reporting System, United States, October 1, 2009-January 31, 2010.[J]. Vaccine, 2010, 28(45):7248-7255.
20. Li C G . A novel influenza A (H1N1) vaccine in various age groups.
21. De L . Clinical Features of the Initial Cases of 2009 Pandemic Influenza A (H1N1) Virus Infection in China — NEJM[J]. New England Journal of Medicine.
22. 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.
23. Rutkowski L , Jaworski M , Pietruczuk L , et al. The CART decision tree for mining data streams[J]. Information Sciences, 2014, 266:1-15.
24. Na, Zhang, Yongsheng, et al. Detection of sea ice in sediment laden water using MODIS in the Bohai Sea: a CART decision tree method[J]. International Journal of Remote Sensing, 2015.
25. Hari B , Vijaya. Empirical Investigation of CART and Decision Tree Extraction from Neural Networks[J]. 2009.
26. Chen T , Tong H , Benesty M . xgboost: Extreme Gradient Boosting[J]. 2016.
27. Chen W , Fu K , Zuo J , et al. Radar emitter classification for large data set based on weighted-xgboost[J]. IET Radar Sonar ? Navigation, 2017, 11(8):1203-1207.
28. Zhang D , Qian L , Mao B , et al. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGBoost[J]. IEEE Access, 2018:1-1.
29. Liaw A , Wiener M . Classification and Regression by randomForest[J]. R News, 2002, 23(23).
30. Ham J , Chen Y , Crawford M M , et al. Investigation of the random forest framework for classification of hyperspectral data[J]. IEEE Transactions on Geoscience & Remote Sensing, 2005, 43(3):492-501.
31. Bart H , Stefan B , Frans V R , et al. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites[J]. Nucleic Acids Research, 2012, 40(14):e106-e106.
32. U, FaigleW, Kern. The Shapley value for cooperative games under precedence constraints[J]. International Journal of Game Theory, 1992.
33. Dutta R , Roy S , Borkotokey S , et al. The Generalized Shapley Value of Cooperative Games as a Social Preference Function. 2023.
34. Papernot N , Mcdaniel P , Goodfellow I , et al. Practical Black-Box Attacks against Machine Learning[J]. ACM, 2016.
35. Hall M A . Practical feature subset selection for machine learning[J]. springer, 1998.
36. Ganapathi A , Kuno H A , Dayal U , et al. Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning[C]// IEEE International Conference on Data Engineering. IEEE, 2009.

# Appendix A Self-appraisal

<This appendix must contain everything covered under the ’self-appraisal’ criterion in the mark scheme. Although there is no length limit for this section, 2-4 pages will normally be suﬃcient. The format of this section is not prescribed, but you may like to consider the following sections and subsections.>

## A.1 Critical self-evaluation

## A.2 Personal reﬂection and lessons learned

## A.3 Legal, social, ethical and professional issues

<Refer to each of these issues in turn. If one or more is not relevant to your project, you should still explain *why* you think it was not relevant.>

### A.3.1 Legal issues

<Discussion of legal issues>

### A.3.2 Social issues

### <Discussion of social issues>

### A.3.3 Ethical issues

### <Discussion of ethical issues>

### A.3.4 Professional issues

<Discussion of professional Issues>

# Appendix B External Materials

<This appendix should provide a brief record of materials used in the solution that are not the student's own work. Such materials might be pieces of codes made available from a research group/company or from the internet, datasets prepared by external users or any preliminary materials/drafts/notes provided by a supervisor. It should be clear what was used as ready-made components and what was developed as part of the project. This appendix should be included even if no external materials were used, in which case a statement to that effect is all that is required.>