Analyzing the Determinants of H1N1 and Seasonal Flu Vaccine Uptake: Developing Predictive models for Vaccination Probability

Karthikeyan Sivakumar
School of Computer Science
The University of Nottingham
Nottingham, UK
psxks12@nottingham.ac.uk

Adarsh Kizhakoodan
School of Computer Science
The University of Nottingham
Nottingham, UK
psxak17@nottingham.ac.uk

***Abstract*—In 2009 a novel H1N1 virus was detected and quickly spread across the world. The H1N1 flu virus is estimated to have caused 151,700 to 575,400 death around the globe. The World Health Organization (WHO) declared this as a pandemic. The Centers for Disease Control and Prevention (US) surveyed to monitor vaccination coverage. This study used two models that can handle missing values by treating it as NaN and without encoding the categorical data and various machine learning models to predict the probability of how likely an individual will receive vaccinations. XGBoost, CatBoost, Logistic Regression, and SVM models were used for the prediction and explored Sequential Feature Selection. Among all the models XGBoost and CatBoost performed the best, and key determinants of vaccination uptake were discussed. For future recommendations, we recommended using oversampling techniques and meta-classifiers for balancing the data and improving the prediction..**

***Index Terms*—Machine Learning, XGBoost, CatBoost, Logistic Regression, SVM**

I. Introduction

In April 2009, a novel Influence virus was detected in Mexico, and it quickly spread to the US and the rest of North America. The H1N1 strain can be transmitted through coughs and sneezes of infected individuals, similar to the seasonal flu. However, it can also spread through contact with contaminated surfaces and touching one's nose or mouth. This presents a potential source of concern for many people. From April 12, 2009 to April 10, 2010, CDC(Centers for Disease Control and Prevention) estimated there were 60.8 million cases (range: 43.3-89.3 million), 274,304 hospitalizations (range: 195,086-402,719), and 12,469 deaths (range: 8868-18,306) in the United States due to the H1N1 virus[2]. The novel Influence virus was spreading along with the usual seasonal flu during the flu season in North America. Seasonal flu is a common respiratory infection which spreads during the cold half of the year.

Vaccination is the most important thing we can do to protect ourselves and our children against ill health. They prevent up to 3 million deaths worldwide every year[1]. Vaccines help train our immune system, much like machine learning models. We expose the immune system to a antigen which is weakened part of the pathogen that cannot cause infections. This acts as the training set for our immune system, which learns to create antibodies in response to these antigens, better preparing us for the actual virus. In this study we are predicting the probability of how likely an individual will receive vaccinations for H1N1 and seasonal flu and we will determine what are the key determinants of vaccination uptake, and which determinant has the greatest influence.

II. Introduction to the Dataset

When it comes to research, data is crucial, and the data source is even more important for a successful study. The data for our prediction study, A telephonic survey - National 2009 H1N1 Flu Survey (NHFS), was conducted by CDC in the US to monitor vaccination coverage. The CDC is the national public health agency of the United States, which conducts studies each year to determine how well influenza vaccines protect against the flu[3].

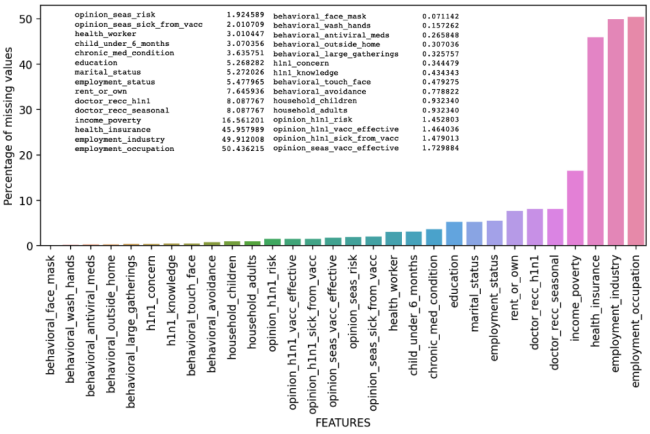

Fig. 1. Missing value percentage

The dataset we are using has 36 features with 26707 entries, and it has two labels known as h1n1 vaccine and seasonal vaccine. Both the training and evaluation sets have 30 missing

values, and the percentage of the missing value is shown in Figure 1

After performing descriptive and exploratory analysis, we found that people are more willing to get vaccinated for seasonal flu than for H1N1. As a result, about half of the people received seasonal vaccination, while only one-fourth received H1N1 vaccination. One-third of all individuals have chronic medical conditions, and the age group above 55 has the highest prevalence of such conditions. A third of the population received seasonal vaccination advice, while about a fourth of respondents received H1N1 vaccination advice from doctors.

Ten percent of respondents have close contact with children under six months old, and except for the age groups of 65 and older and 35 to 44, the remaining groups are evenly distributed. Only 1739 out of 26707 respondents had a very good level of interest in and knowledge about the H1N1 vaccine; surprisingly, more than 95% of respondents did not take antiviral medication. Regarding those with health insurance, the majority are working or are not in the labor force, and most employed people are not below the poverty line. When all the respondent's behaviors were combined, over 35% scored higher than 3 for 6, and more than 70% expressed concern about getting sick after receiving the vaccine. Additionally, 70% of people believe that the H1N1 vaccination is effective, and 80% believe that the seasonal vaccine is effective.

III. Literature Review

There are only few research papers focus on predicting the probability, but some similar researches focus on H1N1 influenza. H.-C. Liu et al. [41] proposed a new classification algorithm for identifying hosts of influenza A viruses based on the physicochemical and fractal properties of hemagglutinin proteins. The authors used SVM, Logistic regression, and a hybrid classifier combining SVM and LR, and the hybrid classifier performed best among all. Inampudi et al. [4] conducted a ""Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines" study. In their paper, the authors prepared regression, classification, and Artificial Neural Network models to predict the likelihood of people getting vaccinated. They used two types of hyper parameter tuning to improve the model's performance. For evaluation, they used ROC AUC score, r2 score, and confusion matrix.

S. S. Ayachit et al. [5] conducted a study titled "Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning". The authors used univariate feature selection to select the features and deployed a probability predicting model using classification, Automated Machine Learning Python Library, and Neural Networks. For evaluation, they used the ROC AUC score.

Both papers [4] and [5] used the most common occurrence to fill in the missing values. Since the nature of the missing values was unknown, it was not an ideal practice for imputation. However, there is no definitive answer to the question of which imputation method is best because the choice of imputation method always depends on the type of data, the proportion of missing values, the missing data mechanism, and the nature of the missing data [6]. Nonetheless, we can identify the nature of the missing values and plan accordingly. If the nature of the missing value is not Missing At Random, mode imputation might cause bias in the model[7].

Moreover, the distribution of the target variables is imbalanced, so there is a high likelihood of getting a skewed model. Therefore, it is better to go for other imputation methods. By knowing the nature of the missing value, we have introduced an imputation method that works well with imbalanced distribution and will be an alternative to mode imputation in the above scenario.

H.-C. Liu et al. [41] proposed a new classification algorithm for identifying hosts of influenza A viruses based on the physicochemical and fractal properties of hemagglutinin proteins. The authors used SVM, Logistic regression, and a hybrid classifier combining SVM and LR, and the hybrid classifier performed best among all.

Methodology

A. Preprocessing

Imputing missing values is a crucial part of preprocessing. Missing values can cause various issues, and the absence of missing values increases statistical power. Missing values may also lead to parameter estimation bias and complicate analysis [8]. Our dataset has 30 features with missing values in the training and evaluation sets, with roughly 20,000 values missing. Therefore, before preprocessing, it is good practice to perform Exploratory Data Analysis (EDA). Data wrangling, which deals with problems in the data before model preparation, is a part of EDA [9]. There are some features with almost 50% missing values. However, we decided not to drop any missing values, as doing so would degrade valuable information [10]. After considering all the points, we decided to create a visualization to check for any patterns in the missing values because "EDA is an important first step in handling missing data because it helps to establish the nature, extent, and pattern of missingness[11]. Moreover, data visualization is crucial in EDA, as humans are better at interpreting visual information than raw numbers[12]. The figure 2 shows the heatmap visualization of missing values. To find the appropriate imputation method for this scenario,

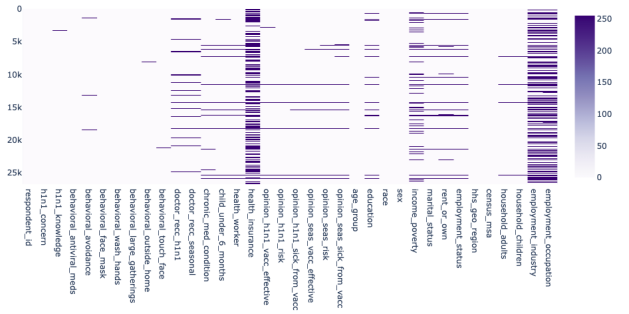

Fig. 2. Heatmap of missing values

we devised two imputation methods for numeric data and a standard method for categorical data. (i) For the numeric data, our first method is to treat the missing values as NaN. This preserves the meaning of the missing value and prevents it from affecting the prediction. (ii) For the second method, we impute -1 for the missing value so that the learning models won't introduce any noise or bias to the data. The model treats this as a separate category. (iii) Since most of the questions in the questionnaire were answered as "unknown," "refused," or "not applicable" by the respondents, we applied accordingly for categorical data such as unknown or refused or unknown or not applicable. (iv) We have used mode imputation as well as it the most commonly used imputation method, however we believe that is not suitable in this scenario as values are not missing at random.

After imputing the missing values, we proceeded to encode the data. Machine learning algorithms only accept numerical inputs, so it is necessary to encode categorical variables into numerical values using encoding techniques[17]. There are many encoding methods, but we used two commonly practised techniques. Label-Encoding or Ordinal Encoding is used for categorical columns with order or ranking in their nature [18]. Numeric or One-Hot encoding is used when the features don't have any order, and for each value in the feature, it will create a new feature of the variable which represents 0 or 1 [19]. By checking whether the data is nominal or ordinal[20], the encoding has done accordingly. After imputing the missing values, we proceeded to encode the data. Machine learning algorithms only accept numerical inputs, so it is necessary to encode categorical variables into numerical values using encoding techniques[17]. There are many encoding methods, but we used two commonly practised techniques. Label-Encoding or Ordinal Encoding is used for categorical columns with order or ranking in their nature [18]. Numeric or One-Hot encoding is used when the features don't have any order, and for each value in the feature, it will create a new feature of the variable which represents 0 or 1 [19]. By checking whether the data is nominal or ordinal[20], the encoding has done accordingly.

B. Model Selection

After all the preprocessing, we started focusing on deploying the model, as many classification models (before applying any ML algorithms) are incapable of directly missing values [21]. However, since most of the time, data scientists or researchers may spend lots of their time in data preprocessing[22], dealing with missing values in the data preprocessing step remains an important step in the classification process before estimation[11]. For the problem statement, we used Logistic Regression, Support Vector Machine, eXtreme Gradient Boosting, and Categorical Boosting models.

Logistic regression is a statistical machine learning model used for binary classification problems. We get a discrete value or categorical variable like 1,0 or True/false as the output of this predictive modelling technique. In Logistic regression, we use the logistic or sigmoid function. The principle of

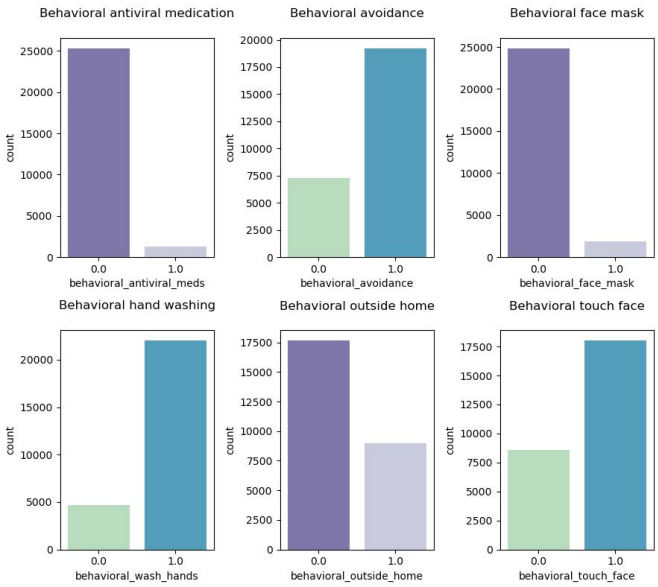

Fig. 3. Behaviour of respondanats

the logistic regression function is to map the result of linear regression to (0,1) by using the characteristic that the value range of the Sigmoid function is (0,1) [32]. This is useful in binary classification problems; it interprets the outputs as probabilities that belong to a specific class. Several logistic regression model approaches can be used for multi-label classification problems; however, this study will use two separate logistic regression models, each predicting one target variable. This is done mainly to eliminate the risk of error propagation from one target variable prediction to the other and to reduce the model's complexity.

Support Vector Machine (SVM) was another approach used in this study for creating a model. It's a supervised machine-learning technique well-known for its high-dimensional capabilities[23].The kernel functions in the SVM allows to map the higher dimensional features and it makes it easier to find a separating hyper plane for classification. It is well-suited for solving classification and regression problems.

XGBoost is built on top of gradient boosting and uses a gradient algorithm as a base learner Classification and Regression Tree (CART). Additionally, XGB uses a gradient descent algorithm to minimize the loss in finding the optimal prediction score[24]. The specialty of the XGB is this algorithm works very well in imbalanced datasets [25], and XGBoost can handle missing values(sparse data)[26]. XGB can handle missing values without imputation preprocessing by treating missing values as NaN[27]. XGBoost offers many regularisation parameters that prevent the model from becoming overfitted.

Similarly, CatBoost is built on gradient boosting and uses a binary decision tree as a base predictor[28]. However, CatBoost has the added advantage of handling categorical datatype columns without encoding, unlike XGBoost. But

still, CatBoost needs help to handle missing values. CatBoost handles the categories without adding bias using the ordered boosting technique[29] and applies a loss function to reduce the residual errors. Additionally, CatBoost has many parameters to prevent the model from overfitting and supports parallel computing. By taking advantage of these boosting models, we trained models without imputing missing values and without encoding the categorical features.

After deciding on the machine learning algorithms, model validation becomes an essential part of building a supervised model. To deploy a model with good generalization performance, it is crucial to have a sensible data-splitting strategy[30], and we decided to use 80 percentage to train the model and 20 percentage for the testing model. Before performing a train-test split, normalization is much more important for models such as Logistic Regression and Support Vector Machines. Normalization is essential for achieving accurate predictive results and reducing training time for the above models[31]. We used StandardScaler to normalize the data before predicting the probability.

C. Hyperparameter Tuning

Selecting the best hyper-parameter configuration for machine learning models directly impacts the model's performance[33]. However, It often requires deep knowledge of machine learning algorithms and appropriate hyper-parameter optimization techniques[34]. For our Hyper-parameter Tuning, we decided to use Randomized_SearchCV[35] and Optuna[36].

Randomized_SearchCV is a method of training and validating a model by a specific number of parameters settings randomly, and it is included in the Scikit Learn package[37]. In addition, it is one of the most commonly-used methods to explore hyper-parameter configuration.

Optuna is a hyper-parameter optimization framework that uses Bayesian Optimization to select the best parameter. Optuna compares all the parameters in probability distribution and searches for the number of iterations we set. Finally, it will evaluate and return the best parameter of the trial. Additionally, Optuna uses a pruning algorithm to prevent adding bias, and we can track the progress by turning on the progress bar.

IV. Evaluation

The Receiver Operating Characteristic curve and accuracy for some models have been used for the evaluation as the dataset is imbalanced. Logistic Regression, Support Vector Classifier, XGB Classifier, and Cat Boost Classifier models were deployed to predict the probability. For Logistic Regression, CatBoost, and SVM, we used -1 for imputing missing value for numeric data, and for XGBoost, the missing value is treated as NaN value. Except for the CatBoost model, label encoding is done for ordinal and categorical columns; pd.getDummies were used for One-Hot Encoding. For the CatBoost model, no columns were encoded as it can handle categories without encoding. After predicting the base model, hyperparameter tuning was performed with appropriate scaling. The categorical Imputation for all the models will remain the same. The model was prepared for evaluation and evaluated using Roc-Auc score. the results of the predictions are shown in the Table 1 below.

TABLE I
Model Performance

Model	H1N1 Prediction	Seasonal Flu	Imputation
LR	0.844187	0.839797	-1
LR tuned	0.844243	0.839806	-1
SVC	0.808089	0.843414	mode
SVC	0.841202	0.844252	-1
XGB	0.857123	0.854778	NaN
XGB tuned	0.867752	0.866857	NaN
CatBoost Tuned	0.868128	0.863517	-1

The Logistic Regression score for H1N1 was 0.844187, and for Seasonal Flu, it was 0.839797, with -1 imputation. SVM scored 0.808089 for H1N1 and 0.843414 for Seasonal Flu with mode imputation. The SVM score for H1N1 was 0.841202, and for Seasonal Flu, it was 0.844252 with -1 imputation. XGB, treating the missing value as NaN, scored 0.857123 for H1N1 and 0.854778 for Seasonal Flu without hyperparameter tuning.

After hyperparameter tuning, Logistic Regression scored 0.844244 for H1N1 and 0.839807 for the seasonal vaccine. XGBoost scored 0.867752 for H1N1 and 0.866857 for the seasonal flu. CatBoost scored 0.868128 for H1N1 and 0.863517 for the seasonal flu. XGBoost and Cat Boost performed the best among all models in the training set.

The results of the Hyperparameter Tuning process revealed the scores for different models. In terms of H1N1. CatBoost achieved a score of 0.868128 for H1N1 vaccine and XGBoost, on the other hand, scored 0.866669 for H1N1 for the seasonal vaccine. Finally, XGB and CatBoost had the overall best performance among all the models.

V. Exploring Feature Selection - Sequential Feature Selection

As part of feature selection, we used Sequential Feature Selection to select the best features and prepare the model for prediction. Sequential Feature Selection learns which features are most informative at each time step, setting the next feature based on the previously chosen features and the classifier's internal belief[38]. Cross-validation is supported, and the scoring parameter refers to the evaluation we want to use. Additionally, SFS supports forward and backward selection methods for selecting and eliminating the features.

XGBoost is used as a model for Forward Feature Selection with 5-fold Cross Validation, and ROC is set for the evaluation. By the end of the trial, the sfs.get_metric_dict() method has been used to get the evaluation metrics and the features used. Then the dictionary is converted into a data frame and plotted the trial is using the SFS inbuild plots.

After exploring the visualization, we decided to select the top 77 features of the H1N1 vaccine and 61 features of the Seasonal Flu Vaccine. After feature selection, XGB was implemented and started predicting the outcomes. Unfortunately,

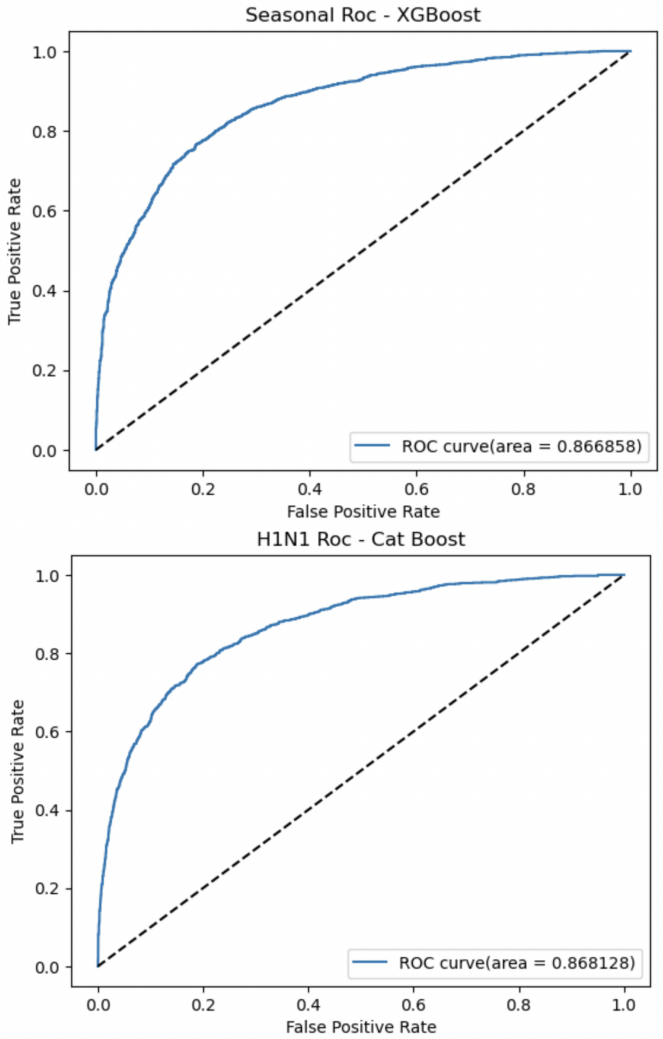

Fig. 4. ROC AUC curves

Sequential Feature Selection didn't give promising results as the H1N1 model got 0.7886 and the Seasonal model 0.7845 in their training set, and in the testing set, the model scored 0.7985. As the ROC score is considerably lower that all the other models used in this study, we recommend not using Sequential Feature Selection for this problem statement.

VI. Discussion

Analysing Determinants of vaccination uptake and what influenced the most for vaccination uptake

After analyzing the data, it is evident that a more significant number of people received the seasonal flu vaccine compared to the H1N1 vaccine. Additionally, females are more likely to get vaccinated and have higher levels of concern and knowledge about H1N1 than males. Individuals with an overall behavior score of 3 or above are more likely to be vaccinated. Doctors are highly influential in encouraging their patients to take vaccines, especially if they have previously recommended H1N1 and seasonal flu vaccines. Respondents with chronic medical conditions tend to receive more seasonal vaccines than

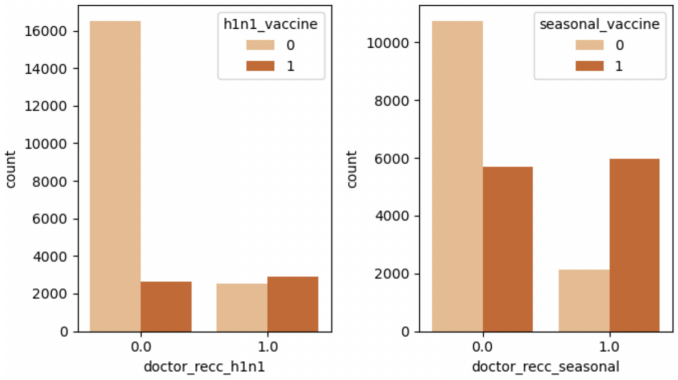

Fig. 5. Doctors recommendation vs vaccination uptake

H1N1 vaccines, and individuals above the age of 55 are the most likely to receive seasonal vaccines.

Individuals who believe that seasonal vaccines are more effective are more likely to get vaccinated. However, less than half of those with the same belief about the H1N1 virus get vaccinated. Additionally, college graduates tend to have higher vaccination rates. Surprisingly, many health workers are not vaccinated. On the other hand, employed individuals and those not in the labor force have the highest vaccination rates. Nearly half of unemployed individuals have also received the vaccine.

Respondents who live in a Metropolitan Statistical Area but not in a major city are more likely to be vaccinated than those living in other regions of the country.

People below the poverty line are less likely to take the vaccine, while people above poverty are more likely to be vaccinated. Additionally, people with health insurance are more likely to be vaccinated than those without. Individuals living below the poverty line are less inclined to get vaccinated, whereas those above the poverty line are more likely to receive the vaccine. Additionally, individuals with health insurance tend to have a higher vaccination rate than those without it.

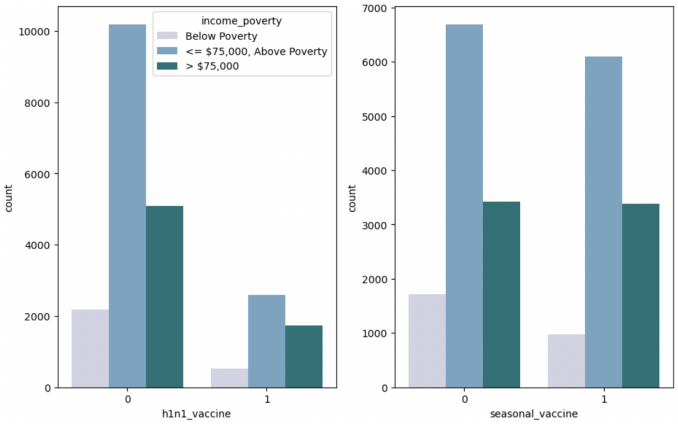

Fig. 6. Income poverty vs vaccination uptake

When comparing the other paper outcomes [5], the XGB model performs better than any different model, and we reached the ROC AUC scores. The highest score they got

for H1N1 were 0.8257 and 0.8601 for seasonal, and we got (‘0.868436‘) for H1N1 and (‘0.867122‘) for seasonal. [4] authors used accuracy for evaluating the results, and accuracy won't be a proper evaluation for the imbalanced classes[39].

A. Critque

This study aims to predict the probability of vaccination uptake for H1N1 and seasonal flu. The study prepared various models for prediction, each with different preprocessing steps. Four models were utilized, and XGBoost and Cat Boost performed the best.

In the preprocessing phase, missing values were imputed using mode imputation and -1 was used for scaling. Scaling was performed perfectly by first fitting and transforming the training data, and then transforming the testing data to prevent data leakage in the testing set. However, mode imputation is not an ideal choice for this problem statement because the nature of the missing values is not Missing Completely At Random. -1 imputation may introduce bias to the model, causing it to perform well in the training set but poorly in the testing set [45].

Logistic Regression is a simple model that can be easily applied. SVMs are good models that can handle a larger number of features. However, these models have limitations in handling large imbalanced classes efficiently and may face difficulties in capturing complex relationships. Furthermore, due to the imbalanced classes, these two models have a high chance of overfitting on unseen data. While they are good at predicting outcomes, they have limitations in predicting probabilities, such as inaccurate probability predictions and SVC not directly providing probabilities.

Random Search CV is used for hyperparameter tuning, and the concept of Random Search CV is to sample hyperparameters randomly. However, it has some limitations, such as using parameters that do not improve performance because it samples hyperparameters in a predefined way. Random Search cannot be efficiently parallelized, and it does not support integration, such as early stopping rounds, to the machine learning model. In the evaluation part, ROC-AUC scores and accuracy has used, and surprisingly, the model performs good results. Even though it performed pretty well, the model might be evaluated to the unseen data to get the actual performance of the data. Accuracy might not be an evaluation of the model's performance as it adds bias to the majority class[46].

As previously concluded, the XGBoost model performed the best. There are several reasons why this is the case. First, we used Optuna for hyperparameter tuning. Optuna has built-in pruning strategies and supports parallelization, which reduces computing time. Second, we performed five-fold cross-validation to ensure that the model could effectively generalize the data and perform well with unseen data. Additionally, since the dataset used in the study is imbalanced, adjusting the class weights can reduce bias and potentially further improve the model.

CatBoost is another model that performs quite well. The tuning of parameters such as learning_rate, l2_leaf_reg, and max_depth using optuna indicates an attempt was made to control overfitting. CatBoost has an advantage when dealing with categorical data, which was utilized. Although we obtained a good AUC ROC score with the model used for H1N1 prediction, we have a considerably low recall and precision score for one class due to class imbalance. Therefore, oversampling should be considered, as it might further improve the model.

Overall, to know the model's actual performance, it is good to test the model with unseen data, and the best results might take some extra time to achieve. In addition, these models can't learn from the previous trial; it might be challenging to increase the performance.

VII. Conclusion

This study was conducted to determine the factors influencing vaccination among the respondents, which we have successfully accomplished. Furthermore, various machine learning models were deployed to predict the probability of vaccination. Among these models, XGBoost had the best performance, with a Receiver Operating Characteristic score of 86.22 percent in the testing set.

Although 86.22 is quite a good score in the test set, we had some limitations in getting the scores for the test set. As our problem statement is to predict the probability of a multi-label, two models will be prepared for each label, reshaping it as two features and checking for the results. In the results, we get a combined score of 2 labels, making it difficult to interpret the score for the appropriate label.

When it comes to the dataset, the distribution of the dataset is imbalanced; the target variable of the model is highly biased to one label, and the missing values add bias to the data, which was confirmed by CDC Dataset User Guide[40]. By considering all of these together, we can only do a preliminary analysis with the given data, and we must note the conclusion that we derived may only partially capture the underline patterns of the data. For the future recommendation in the preprocessing of data, we would recommend to add synthetic data and balance the dataset and do analysis[42].

Since dealing with the missing value is the challenging part of the study, we tried our best to handle the missing values to make better predictions. We recommend improving the survey for the future as addressing the missing value by well-planning the study and collecting the data carefully will be the best possible way to avoid missing value[11]. We also recommend trying Deep learning models and more advanced models such as Bagging Booster and Stacking Classifier for predicting the probability[43][44].

References

[1] "Why vaccination is safe and important," NHS UK. [Online]. Available: https://www.nhs.uk/conditions/vaccinations/why-vaccination-is-safe-and-important/.

[2] "2009 H1N1 Pandemic (H1N1pdm09 virus)," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html.

[3] Centers for Disease Control and Prevention, "General Information — CDC H1N1 Flu," May 1, 2023. [Online]. Available: https://www.cdc.gov/h1n1flu/general_info.htm.

[4] S. Inampudi, G. Johnson, J. Jhaveri, S. Niranjan, K. Chaurasia, and M. Dixit, "Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination," 2021. [Online]. Available: https://doi.org/10.1007/978-981-16-0401-0_11.

[5] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia, and M. Dixit, "Predicting H1N1 and Seasonal Flu: Vaccine Cases using Ensemble Learning approach," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 172-176. [Online]. Available: https://doi.org/10.1109/ICACCCN51052.2020.9362909.

[6] W.-C. Lin et al., "When Should We Ignore Examples with Missing Values?," IJDWM, vol. 13, no. 4, pp. 53-63, 2017. [Online]. Available: https://doi.org/10.4018/IJDWM.2017100104.

[7] D. B. Rubin, "Inference and missing data," Biometrika, vol. 63, issue 3, pp. 581–592, Dec. 1976. [Online]. Available: https://doi.org/10.1093/biomet/63.3.581.

[8] "The prevention and handling of the missing data," Korean J Anesthesiol., vol. 64, no. 5, pp. 402-406, May 24, 2013. [Online]. Available: https://doi.org/10.4097/kjae.2013.64.5.402.

[9] A. A. Dharmasaputro, N. M. Fauzan, M. Kallista, I. P. D. Wibawa, and P. D. Kusuma, "Handling Missing and Imbalanced Data to Improve Generalization Performance of Machine Learning Classifier," in 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jakarta, Indonesia, 2022, pp. 140-145. [Online]. Available: https://doi.org/10.1109/ISMODE53584.2022.9743022.

[10] L. H. Rubin, K. Witkiewitz, J. S. Andre, and S. Reilly, "Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water," J Undergrad Neurosci Educ., vol. 5, no. 2, pp. 71–7, 2007.

[11] H. Kang, "The prevention and handling of the missing data," Korean J Anesthesiol., vol. 64, no. 5, pp. 402-6, May 2013. [Online]. Available: https://doi.org/10.4097/kjae.2013.64.5.402.

[12] J. Schmidt, "Visual Data Analysis," 2023. [Online]. Available: https://doi.org/10.1007/978-3-030-88389-8_25.

[13] Centers for Disease Control and Prevention, "Pandemic flu questionnaire Q1," April 14, 2023. [Online]. Available: https://www.cdc.gov/nchs/data/nis/h1n1/pandemic_flu_questionnaire_q1.pdf.

[14] "The prevention and handling of the missing data," Korean J Anesthesiol., vol. 64, no. 5, pp. 402-406, May 24, 2013. [Online]. Available: https://doi.org/10.4097/kjae.2013.64.5.402.

[15] J. O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," Sociol Methods Res., vol. 6, pp. 215–241, 1977.

[16] J. Sim, J. S. Lee, and O. Kwon, "Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications," Mathematical Problems in Engineering, vol. 2015, Article ID 538613, 14 pages, 2015. [Online]. Available: https://doi.org/10.1155/2015/538613.

[17] K. Potdar, T. Pardawala, and C. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," International Journal of Computer Applications, vol. 175, pp. 7-9, 2017. [Online]. Available: https://doi.org/10.5120/ijca2017915495.

[18] E. Frank and M. Hall, "A Simple Approach to Ordinal Classification," Lecture Notes in Computer Science, vol. 2167, pp. 145-156, 2001. [Online]. Available: https://doi.org/10.1007/3-540-44795-4_13.

[19] I. Ul Haq, I. Gondal, P. Vamplew, and S. Brown, "Categorical Features Transformation with Compact One-Hot Encoder for Fraud Detection in Distributed Environment: 16th Australasian Conference, AusDM 2018, Bahrurst, NSW, Australia, November 28–30, 2018, Revised Selected Papers," 2019. [Online]. Available: https://doi.org/10.1007/978-981-13-6661-1_6.

[20] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," Mach Learn, vol. 107, pp. 1477-1494, 2018. [Online]. Available: https://doi.org/10.1007/s10994-018-5724-2.

[21] F. A. Adnan, K. R. Jamaludin, W. Z. A. Wan Muhamad et al., "A Review of Current Publications Trend on Missing Data Imputation Over Three Decades: Direction and Future Research," October 29, 2021. [Online]. Available: https://doi.org/10.21203/rs.3.rs-996596/v1.

[22] J. Huang, J. W. Keung, F. Sarro, Y. F. Li, Y. T. Yu, W. K. Chan, et al., "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study," J Syst Softw, vol. 132, pp. 226-52, 2017.

[23] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR, vol. 12, pp. 2825-2830, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

[24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in knowledge discovery and data mining ACM SIGKDD International Conference on knowledge discovery and data mining, Washington, DC: University of Washington Vol. 2016, pp. 785–794, 2016.

[25] Q. Gao, X. Jin, E. Xia, X. Wu, L. Gu, H. Yan, Y. Xia, and S. Li, "Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning," 2020. [Online]. Available: https://doi.org/10.3389/fgene.2020.00820.

[26] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," 2020. [Online]. Available: https://doi.org/10.1016/j.patrec.2020.05.035.

[27] D. A. Rusdah and H. Murfi, "XGBoost in handling missing values for life insurance risk prediction," SN Appl. Sci., vol. 2, 1336, 2020. [Online]. Available: https://doi.org/10.1007/s42452-020-3128-y.

[28] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," 2020.

[29] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," J Big Data, vol. 7, 94, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00369-8.

[30] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," J. Anal. Test., vol. 2, pp. 249–262, 2018. [Online]. Available: https://doi.org/10.1007/s41664-018-0068-2.

[31] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 729-735. [Online]. Available: https://doi.org/10.1109/ICSSIT48917.2020.9214160.

[32] H. Ma, P. Leng, Z. Yang, D. Li, and W. Nai, "Logistic Regression Based on t-Distribution Butterfly Optimization Algorithm," in 2021 14th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2021, pp. 372-376. [Online]. Available: https://doi.org/10.1109/ISCID52796.2021.00091.

[33] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," 2020.

[34] R. E. Shawi, M. Maher, and S. Sakr, "Automated machine learning: State-of-the-art and open challenges," arXiv preprint arXiv:1906.02287, 2019. [Online]. Available: http://arxiv.org/abs/1906.02287.

[35] P. Liashchynskyi and P. Liashchynskyi, "Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1912.06059.

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," 2019. [Online]. Available: http://dx.doi.org/10.1145/3292500.3330701.

[37] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," 2011. [Online]. Available: https://doi.org/10.48550/arXiv.1201.0490.

[38] T. Rückstieß, C. Osendorfer, and P. van der Smagt, "Sequential Feature Selection for Classification," 2011. [Online]. Available: https://doi.org/10.1007/978-3-642-25832-9_14.

[39] M. Vejmelka, P. Musílek, M. Palus, and E. Pelikán, "K-Means Clustering for Problems with Periodic Attributes," IJPRAI, vol. 23, no. 4, pp. 721-743, 2009.

[40] Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases, and National Center for Health Statistics. (2012). National 2009 H1N1 Flu Survey (NHFS): A User's Guide for the Public-Use Data File. Presented by NORC at the University of Chicago. Available: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NIS/nhfs/nhfspuf_DUG.PDF.

[41] H.-C. Liu, S.-W. Liu, P.-C. Chang, W.-C. Huang, and C.-H. Liao, "A novel classifier for influenza a viruses based on SVM and logistic regression," in 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, China, 2008, pp. 287-291. doi: 10.1109/ICWAPR.2008.4635791.

[42] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," Jurnal Informatika, vol. 5, pp. 175-185, 2018. doi: 10.31311/ji.v5i2.4158.
[43] P. Bühlmann, "Bagging, Boosting and Ensemble Methods," in Handbook of Computational Statistics, 2012. doi: 10.1007/978-3-642-21551-3_33.
[44] S.-A. Alexandropoulos, C. Aridas, S. Kotsiantis, and M. Vrahatis, "Stacking Strong Ensembles of Classifiers," 2019. doi: 10.1007/978-3-030-19823-7/_46.
[45] B. Akkaya and N. Çolakoğlu, "Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Diseases," in Proceedings of the XYZ Conference, City, Country, Year, pp. 123-135.
[46] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," Pattern Recognit., vol. 41, no. 12, pp. 3692-3705, 2008.