Predict H1N1 and Seasonal Flu Vaccines

Project , Research , Competition

Youssef Mahmoud Farouk
Department of Computer Engineering
Arab Academy for Science and Technology
Alexandria, Egypt
y.m.youssef@student.aast.edu

Farida Sabry
Department of Computer Engineering
Arab Academy for Science and Technology
Alexandria, Egypt
f.m.abdelaziz1@student.aast.edu

***Abstract*— Our reason ,goal and purpose of this research project and study is to find ,identify and forecast the chance of persons and individuals and groups in obtaining the H1N1 and seasonal flu vaccines using data collected from the National 2009 H1N1 Flu Survey. In this paper research work, we carefully used a huge different variety of an advanced machine learning methods, such as Gradient Boosting Classifier, Decision Trees Classifier, LightGBM(LGB) , and CatBoost Classifier and using grid search , crossvalidtion(CV) and ensembling techniques to examine thirty-five different variables , in order to ultimately produce a good reliable predictions regarding these vaccination probabilities. Additionally, we carefully evaluate our models using the (ROC AUC) metric so we perform a thorough analysis of the performance demonstrated by the different models used in this project or research, concentrating on identifying the most efficient method.**

Keywords—component, formatting, style, styling, insert (*key words*)

I. Introduction

The 2009 H1N1 flu pandemic health problem exposed how it's a very and critical important health issue it is to understand and studyd the public health lifestyle and routines, particularly with regard to vaccination point and if the person is vaccinated or not. One from the essential public health measures is "vaccination".. In addition to protecting people , groups, and individuals from health problem and illness, it also helps with the development and the help of having a good immune systems, which can obviously decrease and slow the spread speed of the disease within groups of people. To ensure that an important percentage of the population is protected and vaccinated and to plan effective health programs, it is necessary to predict whether people will decide to get vaccinated or not ..

The motivation and the purpose behind this project and research are a competition held by Driven Data and a project to my college focusing on the prediction of the seasonal and H1N1 flu vaccinations. This project's dataset, which is derived and collected from the National 2009 H1N1 Flu Survey, contains an immense variety of data, including survey replies, demographic information, and actions taken to prevent the flu, such as wearing face masks, washing hands, and reducing social interactions without protection...

The fact that this dataset has more than 1 target variable it has two(2) target variables—one measuring whether an individual received the seasonal flu vaccine or the H1N1 vaccine—makes it even more and more interesting. Our job and goal is to estimate and predict the likelihood of each of the two(2) binary (yes or no) targets. This two-target combination is quite uncommon and offers an unusual special interesting challenging challenge.

The predictions we make may not be as reliable and accurate as we would like due to a huge significant amount of missing data in many several features. Developing accurate models requires dealing with these missing information in an good effective way.

In this research paper, project and study, we investigate and check a number of variables that may change and affect people decision to receive the vaccinee of a seasonal flu shot and the H1N1 vaccine. To create prediction models to see the prediction of H1N1 SEASANOAL VACCINE , we use a different and variety types of machine learning models approaches, including Gradient Boosting Classifier, Decision Trees, LightGBM(LGB) , and CatBoost Classifier and others. We also employ and use ensemble techniques approaches, grid search, and cross-validation(CV) and sometimes a mix of them to make sure the models are as accurate as possible..

We use the ROC AUC for both vaccinations to predict ,see check and assess how well these models function. This metric allows us to check and study how effectively our models differentiate between those who received the vaccination(1 yes) and those who did not (0 no) and for the final test we join the competition in driven data and submit in a submission format to test the accuracy on real hidden unknown test set and to see where is my place in the leaderboard among other ..

II. RELEATED WORK

The prediction of taking the vaccine or not has attracted a lot of interest from many organizations and people around the world. .The Use of machine learning methods and Ai models approaches for analyzing survey data and discover significant variables of vaccination behavior, many numerous attempts and studies in data mining and health operations and communities have investigated in this field as its one of the important thing , We will discuss and see some of them .

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

First lets discuss the GitHub repository which was done by adalseno under the name "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines" is a one interesting project that offers a deep perfect good analysis of the DrivenData competition. This repository contains multiple notebooks that use many libraries like pandas , CatBoost, dtale, profiling, mlxtend , SweetWiz, , and Optuna to illustrate and show the different AI modelling approaches and the exploratory data analysis (EDA).In order to Predict whether participants or the person received the seasonal flu and H1N1 vaccines from survey replies, personal history, and health related behaviors is the competition's main objective.[1]

Another GitHub repository which was done by someone called emykes that set all of its focus exclusively on predicting H1N1 flu vaccination status(0 not vaccinated and 1 vaccinated) with the National 2009 H1N1 Flu Survey data .A good number of Ai model were used, Six machine learning models were used in this project to predict H1N1 and seasonal vaccine which are : K-Nearest Neighbors, XGBoost, Decision Tree, Random Forest , Gradient Boosting and Logistic Regression. The best result was with the Gradient Boosting Classifier model which gave the highest accuracy and was the most accurate/ precise of the other models. Columns like health insurance , Doctor recommendations for vaccine were among the top highest major predictors found in this study.[2]

To put it in a more simple way, A huge and a many numerous number of research papers and data mining projects and notebooks have used the data collected from the National 2009 H1N1 Flu Survey to Predict H1N1 and Seasonal Flu Vaccines , if the person has taken the vaccine or not. These research studies and projects goal is to seek and to determine important indicators to guide public health safety measures. Doctor recommendations, vaccine effectiveness, insurance status, and Health related behaviors have been consistently demonstrated to be from the most top influential factors. These researches and projects that use machine learning techniques, have shown a good promising improving in public health tactics to increase vaccine coverage around the world , especially in times of pandemics such as the 2009 health issue pandemic.

III. PROPOSED MODEL

i. Proposed Model

In this project and research , We trained many Ai models and classifier for example decision tree , random forest, Catboost,gradient classifier , KVM ,XGBoost ,logistic regression and others. Every time we trained each more than 2 or times as each was trained and tested using different parameters and also for more training to get the best output we also used Grid search alone , cross validation alone, and ensemble techniques . It didn't stop there we also tested using a mix sometimes a mix between CV and grid search ,also a mix between ensemble technique and grid search and ensembling with CV also finally a mix with more than Ai model classifier using ensemble technique with grid search and CV .

IV. EXPERIMENTAL WORK

i. Dataset

This research study's project and notebook uses a dataset its size is 26707 rows and 36 columns originates from the National 2009 H1N1 Flu Survey data , this survey was carried out and done in in late 2009 . The main goal and purpose of this survey was to gather as much as possible and collect as much data as possible about people, group s and individual health safety measures and habits, vaccination status ,if he is vaccinated or not, and other features and demo. location .. There are 36 columns in the dataset: one identification (ID) column which is mainly dropped and the rest 35 feature columns. The dataset is divided into 3 data frames one for the train, one for the train labels and the third is for the testing .The dataset contains some object type data and a numeric object data. The features address and show us a significant important number of topics for example the vaccination status if he/she is vaccinated or not, the healthr related behaviors, opinions on vaccine efficacy, and other .With the two binary target variables indicating whether or not they received the seasonal and H1N1 flu vaccines, each row represents a one single person t.[3]

- Target Variables:

H1N1 vaccine received is indicated and identified by the variable h1n1_vaccine (0 = No, 1 = Yes). And Seasonal vaccine is identified by (0 = No, 1 = Yes): Indicates whether the respondent has had the seasonal flu shot or not.

Features :There are thirty-five(35) features in the dataset, which are divided into many primary category's demographic factors' , 'Health behaviors' & 'Vaccine Efficacy' :

First demographic factors like Race, income ,education, Age, sex, race,eduction, marital status, housing situation, employment status, and other are all considered 'demographic factors.

Second 'Health behaviors' include things like the use of the face masks, washing and cleaning hands, avoiding large crowds' area places, minimizing, and decreasing the outside contact with other, taking antiviral drugs or medicine, and avoiding contact with sick people.

Third 'Vaccine Efficacy' and in this section it discusses the knowledge & concerns that respondents and people had on regarding H1N1, as well as their evaluations and opinions of the advantages and disadvantages of seasonal and H1N1 vaccinations...

Meanwhile not all the respondent has taken the vaccine and not all of them has taken both some take one type only , in the below table it will show it more clearly :

No. of Resp.	1 of 2 vaccines	Both types	Vacc. by 1 or 2	Did not take vaccine.
26707	13412	4697	18109	8598

Other way to learn more about the dataset features we plot some Stacked bar chart so we able to visualize the vaccination rate for H1N1 vaccine and SEASONAL vaccine :

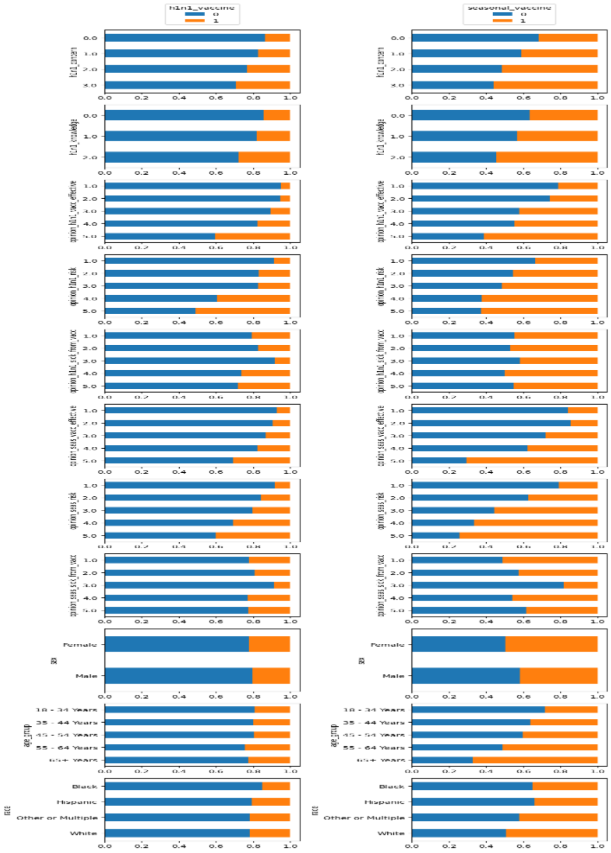


Figure 1.Visualize the vaccination rate for some features.

Other way to learn more about the dataset is to plot the frequencies for the features , the histogram diagram is for the numerical columns and the pie charts are for object type columns , let see an example on the train dataset:

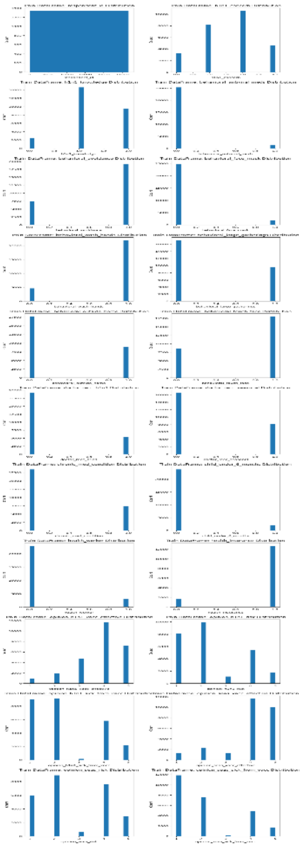


Figure 3.Histogram for Numerical type.

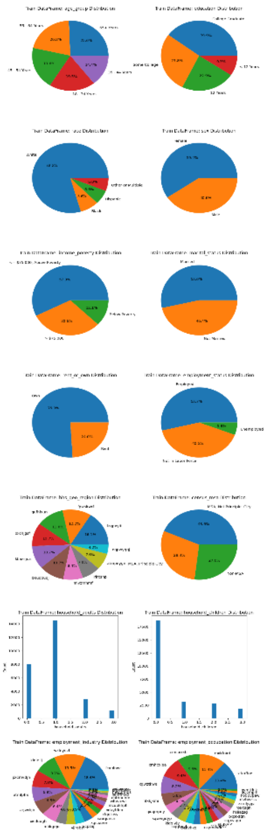
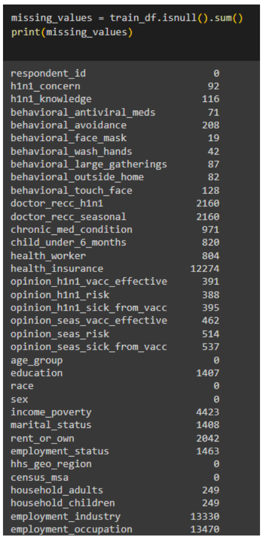


Figure 2.Pie charts for object type.

Handling missing values:

Many features in the dataset have a huge significant number of missing values; for certain features, the missing value data percentage is over and exceeded the 60%. This means that in order to guarantee the consistent of the model performance, and. Preserve and save data integrity then a careful data treatment is a preferably required to be done and maybe involving the imputation of some attributes or their deletion And dropping.

In figure 1 it shows the number of missing values in each feature .



```
missing_values = train_df.isnull().sum()
print(missing_values)
```

```
respondent_id                      0
h1n1_concern                      92
h1n1_knowledge                   116
behavioral_antiviral_meds         71
behavioral_avoidance             208
behavioral_face_mask              19
behavioral_wash_hands             42
behavioral_large_gatherings       87
behavioral_outside_home           82
behavioral_touch_face            128
doctor_recc_h1n1                2160
doctor_recc_seasonal            2160
chronic_med_condition            971
child_under_6_months             820
health_worker                    804
health_insurance               12274
opinion_h1n1_vacc_effective      391
opinion_h1n1_risk                388
opinion_h1n1_sick_from_vacc      395
opinion_seas_vacc_effective      462
opinion_seas_risk                514
opinion_seas_sick_from_vacc      537
age_group                          0
education                       1407
race                               0
sex                                0
income_poverty                  4423
marital_status                  1408
rent_or_own                     2042
employment_status               1463
hhs_geo_region                     0
census_msa                         0
household_adults                 249
household_children               249
employment_industry            13330
employment_occupation          13470
```

Figure 4.No of missing values in train.

ii. *Evalution mertircs*

The prediction models' efficacy will be tested and predicted by calculating the (ROC AUC) curves for both target variables. The average of these two ROC AUC curves scores determines the final prediction result. The best and better Stronger model performance is identified and indicated by a the bigger larger value. This multilabel classification metric can be computed in Python using the sklearn.metrics.roc_auc_score function..

iii. *Sumbsion format*

The three columns in the submission file should be resp_id, h1n1_vaccine, and seasonal vaccine in a CSV format. There should be float probabilities between 0.0 and 1.0 in the predictions for the two target variables. The values supplied must represent the probabilities of vaccination rather than binary labels because the evaluation use ROC AUC.

iv. *Results*

The best performance was achieved using training/validation split of 0.05. for preprocessing, filling empty values of categorical columns using “none” rather than median and numerical columns were filled using mode. Dropping columns using multiple techniques decreased the model’s testing accuracy of the model. Additionally, handling the imbalanced classes did not affect testing accuracy by much.

The models tested and their corresponding performance metrics are as follows:

- **Bagging classifier**

H1N1 Vaccine - AUC: 0.8152841274595293, Log Loss: 1.045307199464847

Seasonal Vaccine - AUC: 0.8190505500743375, Log Loss: 1.1013754069497217

- **Extra Tree**

H1N1 Vaccine - AUC: 0.8450901340814876, Log Loss: 0.370765516961414

Seasonal Vaccine - AUC: 0.8540668071898447, Log Loss: 0.48049679630808645

- **KNN**

H1N1 Vaccine - AUC: 0.7319428690258367, Log Loss: 2.050766522123509

Seasonal Vaccine - AUC: 0.7799667203739744, Log Loss: 1.7694958600779065

- **Naïve bayes**

H1N1 Vaccine - AUC: 0.7831933564428637, Log Loss: 0.8204487280965175

Seasonal Vaccine - AUC: 0.8033276112736913, Log Loss: 0.913730536156911

- **Random forest**

H1N1 Vaccine - AUC: 0.8582691711603077, Log Loss: 0.38289202516640986

Seasonal Vaccine - AUC: 0.8559410211699476, Log Loss: 0.4768174041375467

- **SVM**

H1N1 Vaccine - AUC: 0.8361092201800167, Log Loss: 0.3746919845753487

Seasonal Vaccine - AUC: 0.8546477404728671, Log Loss: 0.4732390301912313

- **SVM with cross validation**

H1N1 Vaccine - AUC: 0.8312463757154025, Log Loss: 0.3822685804082766

Seasonal Vaccine - AUC: 0.8508738765095069, Log Loss: 0.4802438587058179

- **NN**

H1N1 Vaccine - AUC: 0.8613209087244342, Log Loss: 0.354745538647374

Seasonal Vaccine - AUC: 0.8613828412086453, Log Loss: 0.46589431379335655

- **CatBoost model**

used grid search and cross-validation, had the best accuracy, with a test accuracy of 0.8618.

H1N1 Vaccine - AUC: 0.8905079299972656, Log Loss: 0.3190256337837003

Seasonal Vaccine - AUC: 0.8703356123078173, Log Loss: 0.4511298861211398

V. Conclusion and future Work

i. Conclusion

In this paper, we used data of H1N1 to explore several machine learning algorithms to estimate the probability of people obtaining seasonal and H1N1 flu vaccines. We tried to find the best way to predict vaccination uptake by using models like support vector machine(SVM), Decision Trees, K-nearest neighbor (KNN), CatBoost, and other algorithms. For both the H1N1 and seasonal flu vaccinations, the CatBoost model performed better than the others, displaying the best accuracy and AUC scores after being optimized by grid search and cross-validation.

The accuracy of the model relies on handling the missing values and the use of several preprocessing techniques, such as replacing empty categorical entries with "none" and the mode in numerical columns. Further, there was little effect of correcting class imbalances on the model's accuracy. The results highlight how important it is for public health to use machine learning techniques to improve vaccination plans and uptake estimates.

ii. Future work

We'll explore the following important topics to improve our prediction models:

Feature Engineering: We'll go further into designing and picking features that more accurately represent the variables affecting vaccination rates.

Model Transparency: By using tools to better understand how each variable affects model projections, we will be able to make more informed judgments about public health.

Real-Time Updates: In order to facilitate quicker public health responses, we will be concentrating on creating systems that continuously update predictions as new data becomes available.

Additional Datasets: To keep our models accurate and up to date, we'll incorporate more recent data.

Better Imputation: To create more dependable models, we'll explore with more advanced methods for addressing missing data.

Real-World Evaluation: We'll work with healthcare institutions to put our models to the test in real-world scenarios, and we'll use the feedback we receive to improve them even more.

VI. REFERENCES

[1] adalseno, “Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines,” GitHub repository, 2023. Available: https://github.com/adalseno/Flu-Shot-Learning-Predict-H1N1-and-Seasonal-Flu-Vaccines.)

[2] emykes, “Flu Vaccination ML,” GitHub repository, 2023. Available: https://github.com/emykes/Flu_Vaccination_ML.I. S. Jacobs and C. P.

[3] DrivenData, “Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines,” DrivenData competition, 2023. Available: https://www.drivendata.org/competitions/66/flu-shot-learning/.

[4] Project: Predict H1N1 and Seasonal Flu Vaccines. Available: https://www.researchgate.net/publication/346061469_Project_Predict_H1N1_and_Seasonal_Flu_Vaccines,