CS4642 - Data Mining & Information Retrieval

Flu Shot Learning Competition
Final Report

Team : uomcse_u2js
160046C - A.U.S. Athukorala
160224V - M.T.U. Isuranga
160555K - T.H.J. Sandaruwan
160571F - S. Senanayake

Table of Contents

1. Introduction

The H1N1 virus was first founded in 2009, in Mexico and then this influenza spread across the world. Due to the high spreading speed and the powerful impact of this contagious novel virus, it was declared as a pandemic situation by the World Health Organization. A vaccine for the H1N1 virus was found in late 2009 and also the researchers found that yearly flu vaccination as a foremost step to protect from this virus.

The United States conducted the 'National 2009 H1N1 Flu Survey' to track the levels of disease activity and the concern of the public towards this pandemic. This phone survey asked respondents whether they had received the H1N1 and seasonal flu vaccines, and some additional questions covered their social, economic, and demographic background, opinions on risks of illness and vaccine effectiveness, and their health behaviors. By analyzing and having a better understanding of how these characteristics are associated with personal vaccination patterns can provide better guidance to improve public health in the future.

'Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines' is a data science competition held by 'drivendata.org' to predict whether people got H1N1 and seasonal flu vaccines using information they shared about their backgrounds, opinions, and health behaviors. This report contains details about the methods and techniques that we used to predict.

2. Problem Statement

In this project, we address the problem of predicting how likely individuals are to receive their H1N1 and seasonal flu vaccines based on the provided individuals data. Those data contain details of an individual's social, economic and demographic background, opinion, and knowledge about the risks of illness and vaccine effectiveness, and their health behaviors.

3. Objectives

The data science competition "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines" hosted by DrivenData are about predicting whether people got H1N1 and seasonal flu vaccines using data they shared about their backgrounds, opinions, and health behaviors.

The goal of this project is to research and build a model for the above mentioned "Flu Shot Learning" competition which can predict whether people got H1N1 and seasonal flu vaccines. The main aim of our project is to find the probability that a person will receive H1N1 and seasonal Flu vaccination based on different attributes of a given person in the society.

4. Methodology

The following are the major steps followed to achieve a good score in predicting whether persons got h1n1 vaccine and seasonal vaccine.

1. Gathering data
2. Perform exploratory data analysis
3. Data preprocessing
4. Choosing a model
5. Training and evaluation
6. Parameter tuning
7. Prediction

4.1 Dataset

The dataset consists of data that people share about their backgrounds, opinions, and health behaviors. Data is provided courtesy of the United States National Center for Health Statistics.

There are mainly three data files provided by the DrivenData competition.

1. Training Data Features
    - Contains the features and the relevant attribute values of the training dataset
    - Contains the same 35 features and attribute values for 26707 persons.
2. Training Data Labels
    - Contains two target variables
        - h1n1_vaccine - Whether the respondent received the H1N1 flu vaccine.
        - seasonal vaccine - Whether the respondent received the seasonal flu vaccine.

3. Test Data Features
    - Contains the features and the relevant attribute values of the testing data set
    - Contains the same 35 features like the training feature set
    - There are data related to 26708 persons who need to predict whether they got seasonal flu and h1n1 vaccines.

There are 33 categorical features in the data as follows.

- h1n1_concern - Level of concern about the H1N1 flu.
- h1n1_knowledge - Level of knowledge about H1N1 flu.
- behavioral_antiviral_meds - Has taken antiviral medications. (binary)

- behavioral_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of their own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by the doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by the doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or another lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)
- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- opinion_h1n1_vacc_effective - Respondent's opinion about H1N1 vaccine effectiveness.
- opinion_h1n1_risk - Respondent's opinion about the risk of getting sick with H1N1 flu without the vaccine.
- opinion_h1n1_sick_from_vacc - Respondent's worry about getting sick from taking the H1N1 vaccine.
- opinion_seas_vacc_effective - Respondent's opinion about seasonal flu vaccine effectiveness.
- opinion_seas_risk - Respondent's opinion about the risk of getting sick with seasonal flu without the vaccine.
- opinion_seas_sick_from_vacc - Respondent's worry of getting sick from taking the seasonal flu vaccine.
- age_group - Age group of respondents.
- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - the household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of the respondent.
- rent_or_own - Housing situation of the respondent.
- employment_status - Employment status of the respondent.
- hhs_geo_region - Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.

- census_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
- employment_occupation - Type of occupation of the respondent. Values are represented as short random character strings.

Two of the features are discrete numerical.

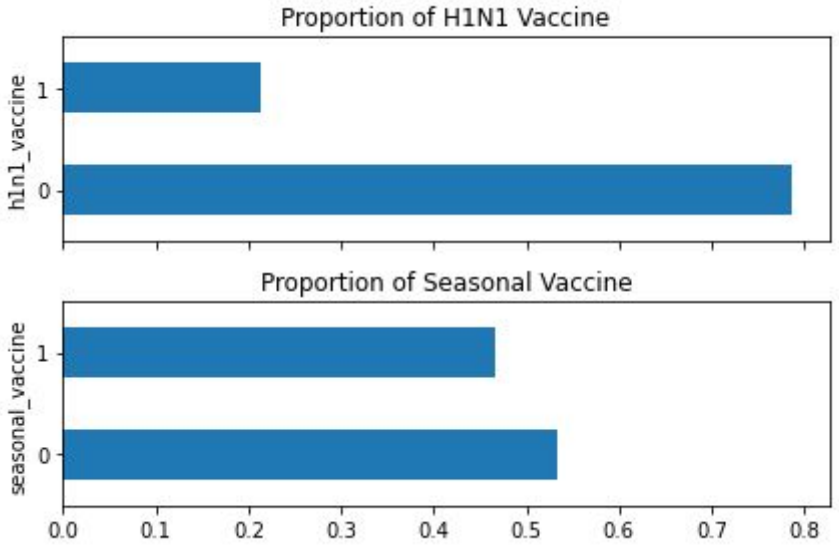
- household_adults - Number of other adults in the household, top-coded to 3.
- household_children - Number of children in the household, top-coded to 3.

4.2 Data Analysis

4.2.1 Label Analysing

The dataset provided by DrivenData for this project consists of whether respondents had received the H1N1 and seasonal flu vaccines, and some additional questions covered their background, opinions on risks of illness and vaccine effectiveness, and health behaviors.

The following diagram shows the distribution of the two target variables.

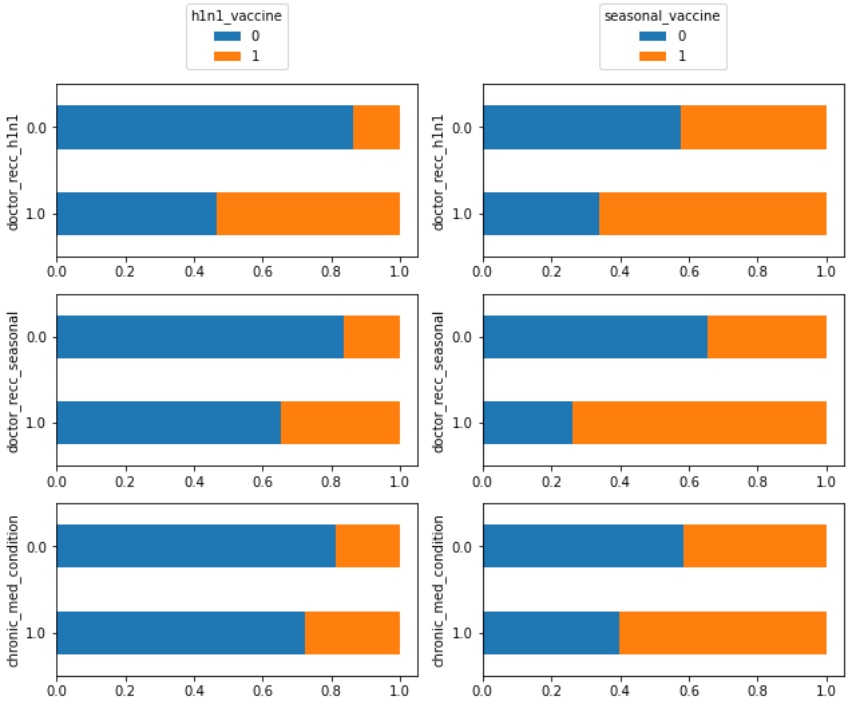


According to the graph, it looks like roughly half of people received the seasonal flu vaccine, but only about 20% of people received the H1N1 flu vaccine. So the seasonal flu vaccine target has balanced classes, but the H1N1 flu vaccine target has moderately imbalanced classes.

4.2.2 Feature Analysing

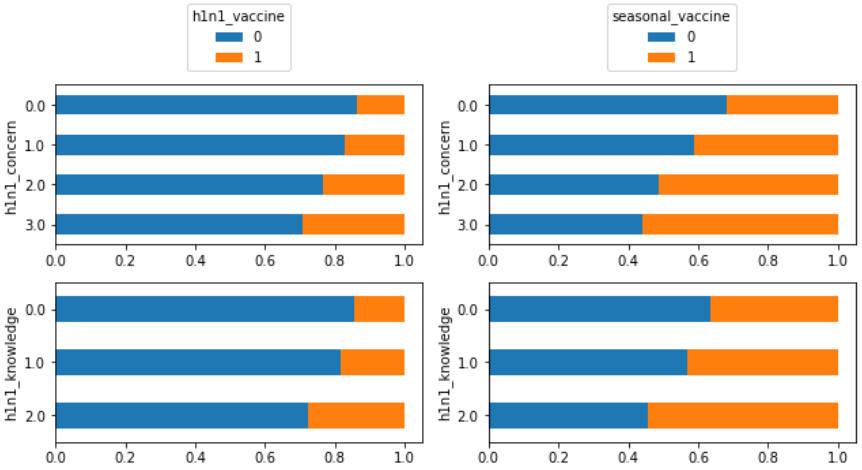
If people haven't a better idea or knowledge about h1n1 influenza or the vaccine, they are refraining from taking the vaccinations. The following diagrams show how the features are correlated with the target variables.

4.2.2.1 Correlation between the health-history based questions with targets



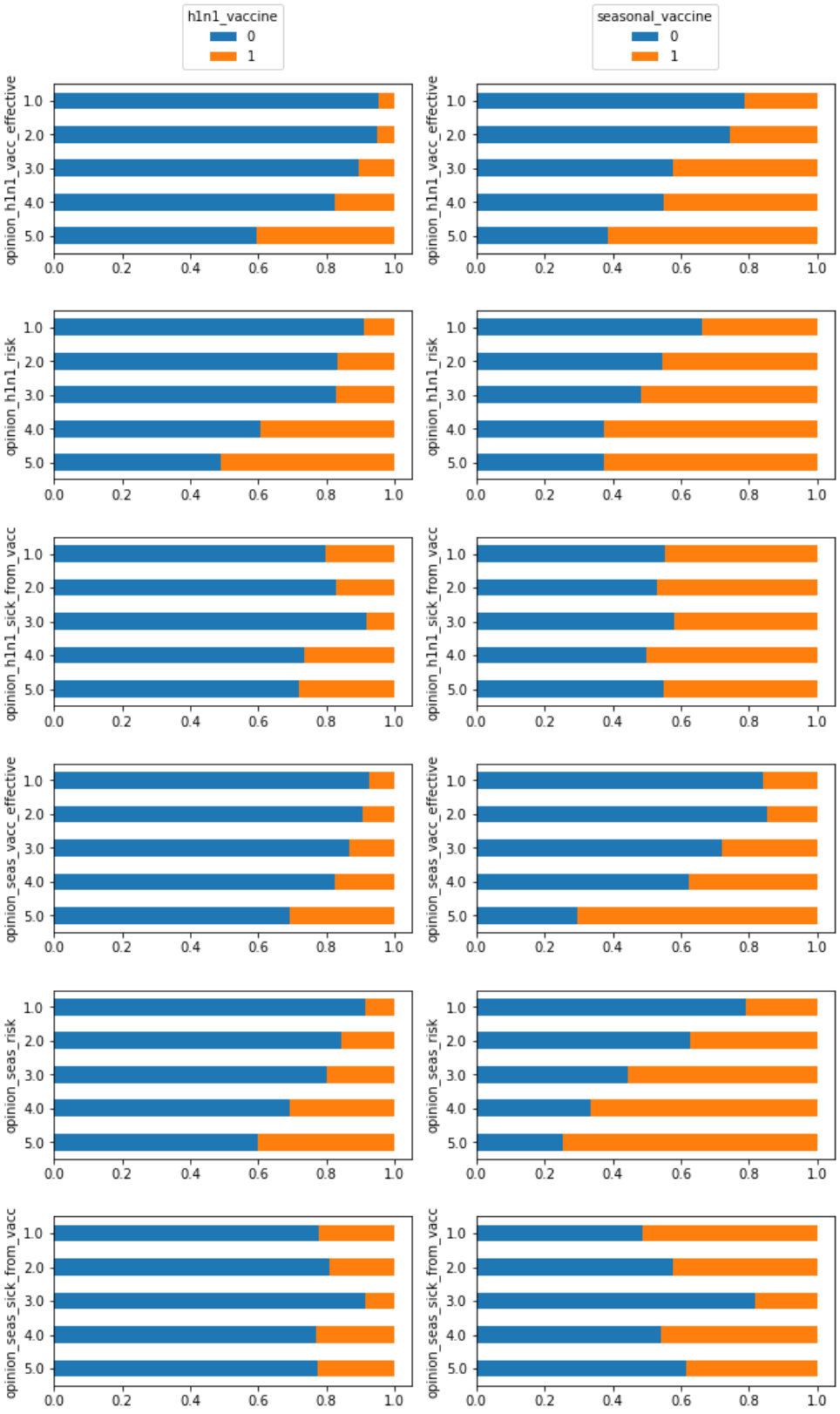
There is a strong correlation for both target variables with the health-history based questions.

4.2.2.2 Correlation between the h1n1 knowledge-based questions with targets



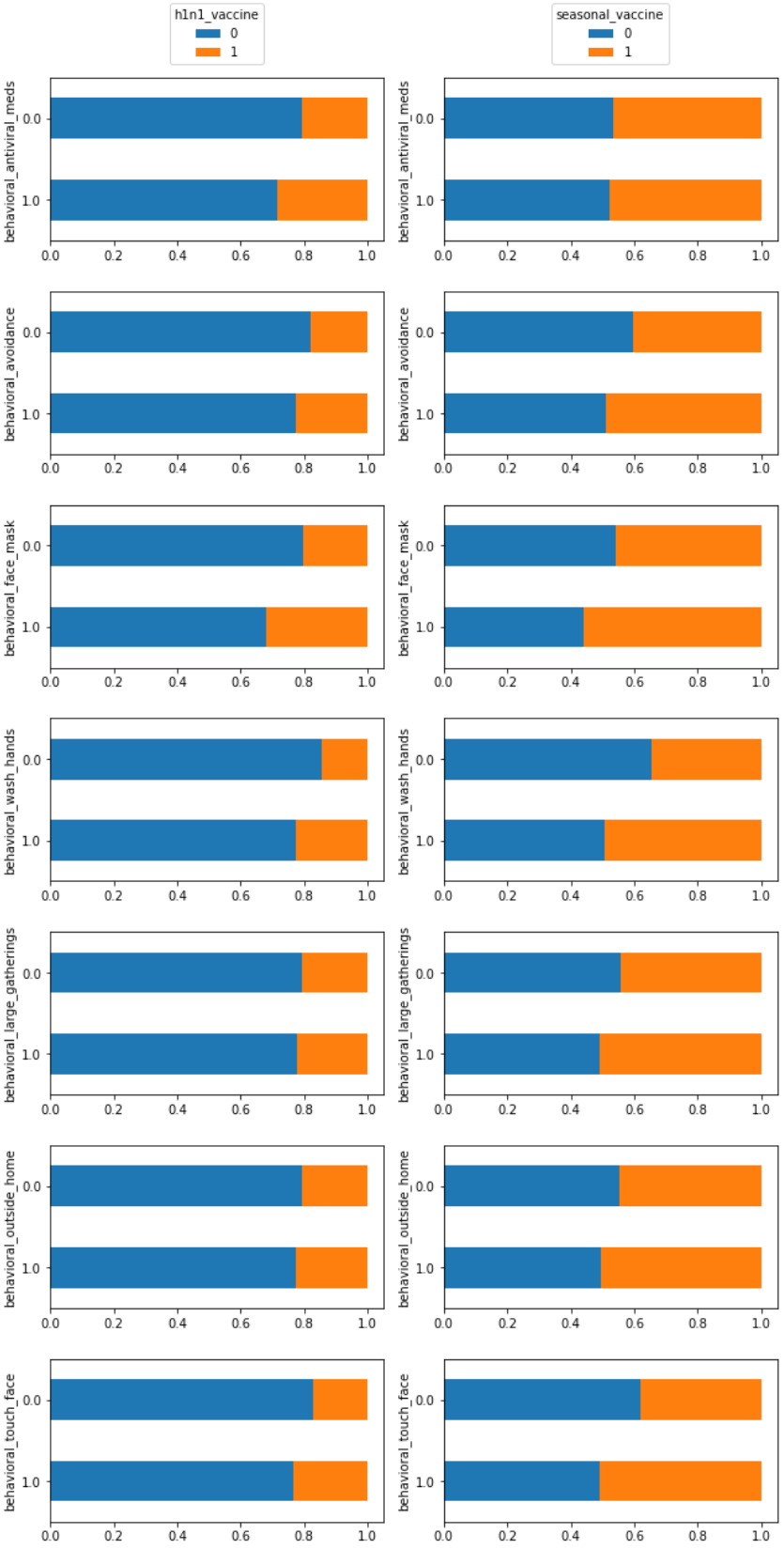
H1N1 Knowledge-based questions and H1N1 concern questions also have a large correlation for both target variables.

4.2.2.2 Correlation between the opinion-based questions with target variables



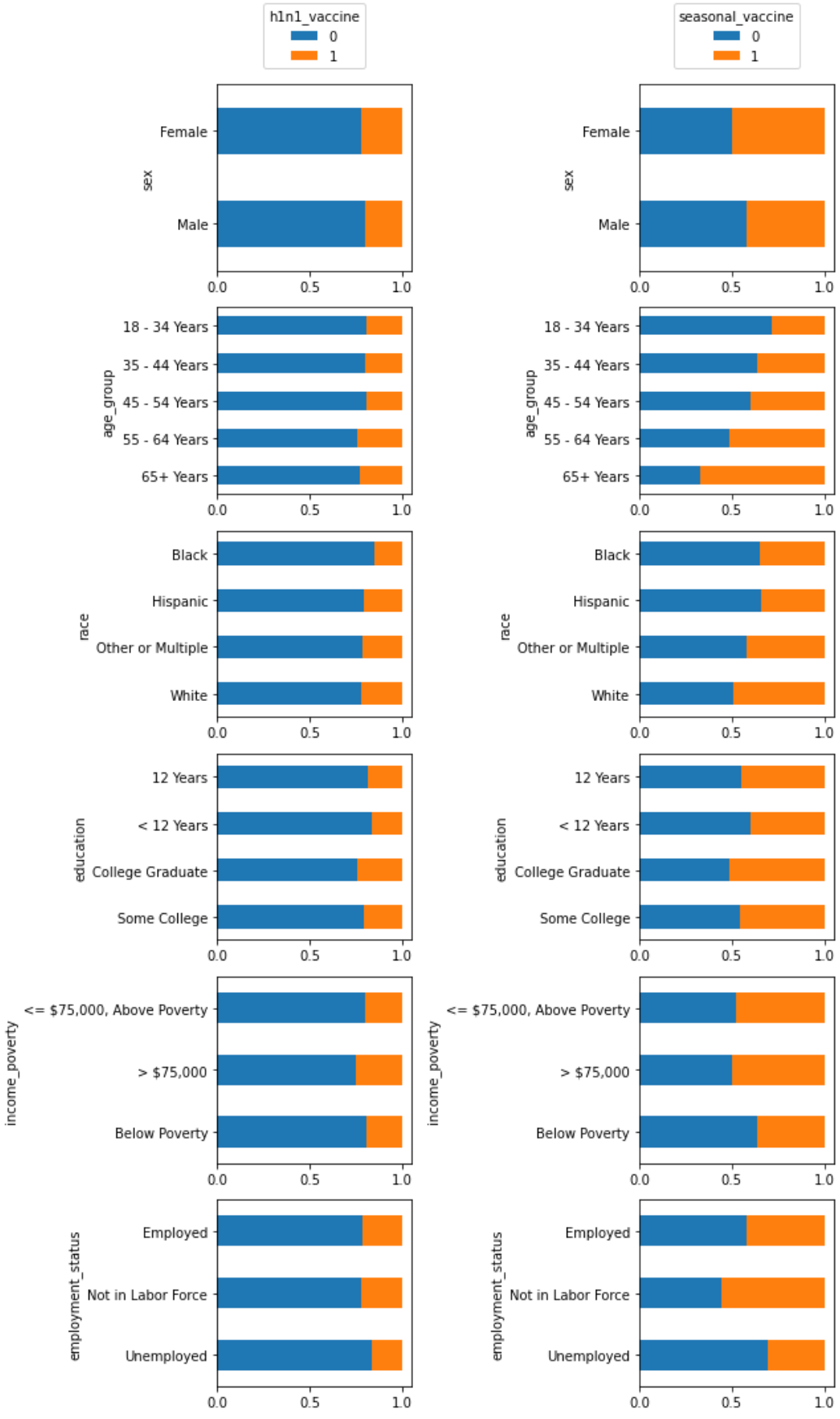
It looks like the opinion-based questions have a pretty strong correlation for both target variables.

4.2.2.2 Correlation between the behavior-based questions with target variables



Behavioral-based questions also have a considerable correlation for both target variable

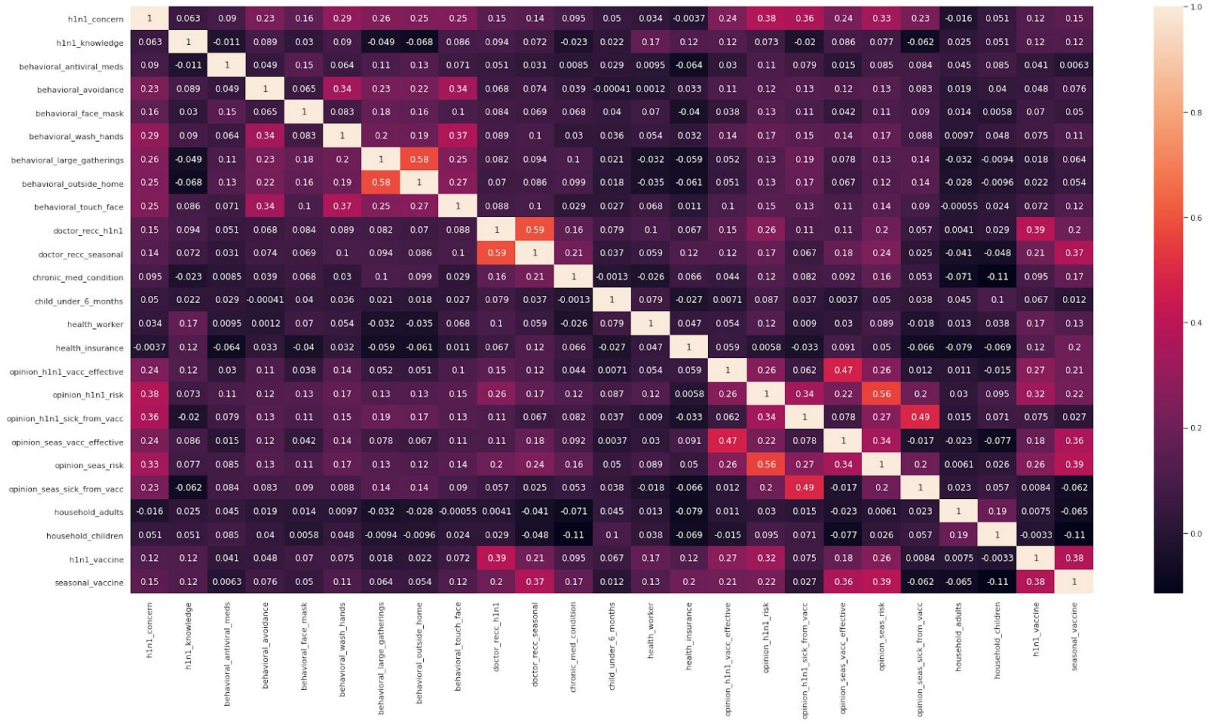
4.2.2.3 Correlation between the demographic-based questions with targets



According to the above diagrams, it seems that the demographic features like age group, education, race, income_poverty, and employment_status features had a stronger correlation with ‘seasonal_vaccine’ but less with the ‘h1n1_vaccine’. We can see a strong correlation with age_group with the seasonal_vaccine but not with h1n1_vaccine. So shows that higher-risk adults are taking the flu vaccine when compared to youngers.

4.2.2.3 Cross-correlation among features

We tested the cross-correlation among features. The following diagram represents the cross-correlation of the features and the heat map used Pearson’s correlation coefficients. This analysis can be used to identify the features that are repeated and what are the features that can be reduced. By observing the heatmap we identified that there are no high correlation features in the given dataset.



4.3 Data Preprocessing

4.3.1 Data Cleaning

4.3.1.1 Imputation of missing values

Most of the features of the training dataset and test dataset have some missing values. Some features like health insurance, employment_industry, and employment_occupation have about half of missing values in both train and test datasets.

There are several methods for imputing missing values. Among them, we used the following techniques to fill the missing values.

Missing values in some features like employment_industry are filled with global constants like “other”, “unknown”.

ex: -

```
dataset['employment_industry'] =
dataset['employment_industry'].fillna('other')
```

Some features which are highly correlated with other features are filled according to the correlated feature.

We tried several other techniques to fill other features. Filling values with mean gave the best performance among them.

ex:-

```
X_train= X_train.fillna(X_train.mean())
```

In addition to these, we tried filling missing values with mode and k nearest neighbor imputer with different k values.

```
imputer = KNNImputer(n_neighbors=10)
dataset = imputer.fit_transform(dataset)
```

4.3.2 Data Transformation

Data transformation is a process that converts data from one format to another format. Some examples of the data transformation techniques are smoothing, aggregation, generalization, normalization, attribute construction. Among them, the techniques we used are described below.[1]

4.3.2.1 Normalization

We used data normalization to reduce high scaled attribute data into a low specified range. Standard scalar is used to transform values into the 0-1 range for the normalization.

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
```

4.3.2.2 Categorical Encoding

All the features in the dataset are categorical variables. Among them, a lot of features are binary or have discrete numeric values. However, the values of features like sex, race, age-group are in texts. Hence they are transformed into numerical values. Two techniques are used for this.

- Label encoding
- One hot encoding

When the different categorical values of the feature have an ordering, then label encoding is used. For example, the values of *education* have a clear order. Hence it is encoded as follows.

```
dataset['education'] = dataset['education'].map({"< 12 Years": 1, "12 Years":2, "Some College":3, "College Graduate":4})
```

One-hot encoding is used if there is no clear ordering between values. For example, the ”race” feature is encoded as follows.

```
dataset["race"] = dataset["race"].astype("category")
dataset = pd.get_dummies(dataset, columns = ["race"],prefix="R")
```

4.3.2.3 Attribute Construction

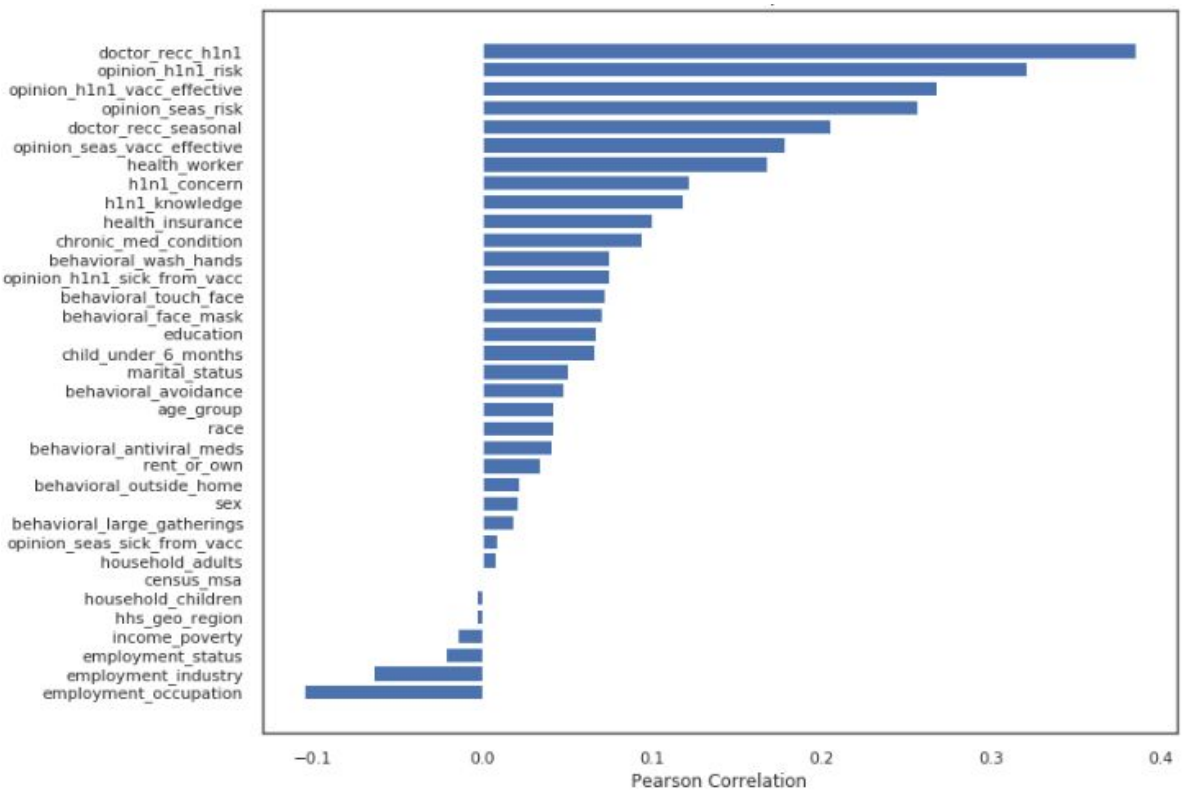
We created a new feature called *household_tot* by combining *household_child* and *household_adult*. However, that didn't give much accuracy.

4.3.3 Data Reduction

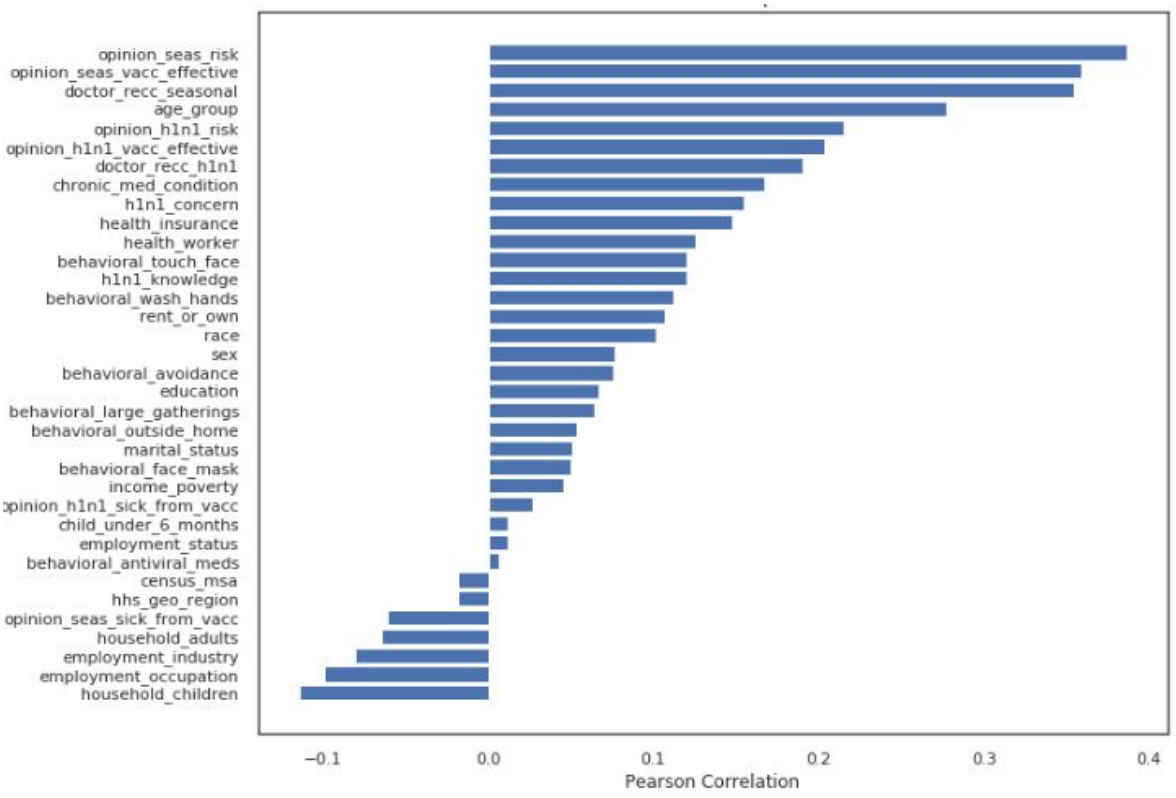
4.3.3.1 Reducing number of features

The original dataset consists of 35 features. However, all features are not used for the models. A selected subset is used for the models.

We removed the set of features by analyzing the correlation of features with *h1n1_vaccine* and *seasonal vaccine* labels.



feature correlation with h1n1_vaccine



feature correlation with seasonal flu

In addition to the feature correlation, we used a univariate selection method for feature selection. SelectKbest scikit-learn library is used with chi-squared statistical tests to find the best k features.

Specs	Score
doctor_recc_h1n1	2831.893378
opinion_h1n1_risk	1911.128957
opinion_seas_risk	1212.823595
doctor_recc_seasonal	691.076127
employment_occupation	660.721299
health_worker	645.327072
opinion_h1n1_vacc_effective	496.280005
employment_industry	275.070380
opinion_seas_vacc_effective	243.255983
h1n1_concern	201.974270
chronic_med_condition	161.644925
behavioral_face_mask	123.368060
opinion_h1n1_sick_from_vacc	115.176420
h1n1_knowledge	111.511825
child_under_6_months	103.590619
behavioral_touch_face	43.875377
behavioral_antiviral_meds	41.591712
education	38.151527
age_group	30.031988
marital_status	29.348784
behavioral_wash_hands	25.921642
health_insurance	17.196478
behavioral_avoidance	16.431639
race	14.931364
employment_status	12.763574
behavioral_outside_home	8.340735
rent_or_own	6.604882
behavioral_large_gatherings	5.408120
sex	4.636082
opinion_seas_sick_from_vacc	1.506381
income_poverty	1.014393
household_adults	0.957152
hhs_geo_region	0.481817
household_children	0.466320
census_msa	0.003839

selectkbest scores with h1n1

Specs	Score
opinion_seas_risk	2759.080914
doctor_recc_seasonal	2066.049373
age_group	1370.370317
opinion_seas_vacc_effective	990.781133
opinion_h1n1_risk	858.904336
doctor_recc_h1n1	695.512437
employment_occupation	580.437781
household_children	555.157437
chronic_med_condition	515.154497
employment_industry	418.406205
health_worker	361.872933
h1n1_concern	325.523263
opinion_h1n1_vacc_effective	287.520373
behavioral_touch_face	123.394755
h1n1_knowledge	115.670404
race	91.674338
opinion_seas_sick_from_vacc	81.429407
household_adults	70.595766
behavioral_large_gatherings	69.756457
rent_or_own	68.115572
sex	64.475384
behavioral_face_mask	62.279296
behavioral_wash_hands	58.670400
behavioral_outside_home	50.361891
behavioral_avoidance	42.103587
education	38.204047
health_insurance	38.029430
marital_status	30.593912
opinion_h1n1_sick_from_vacc	15.340242
hhs_geo_region	14.872176
income_poverty	9.101169
census_msa	7.253763
employment_status	3.684082
child_under_6_months	3.371014
behavioral_antiviral_meds	0.995289

selectkbest scores with seasonal flu

We dropped the below set of features from the dataset and other features are selected to use in models.

- employment_status
- census_msa
- hhs_geo_region

4.4 Modeling

4.4.1 Simple Modelling

4.4.1.1 Cross validate models

We compared 16 popular classifiers and evaluated the mean accuracy of each of them by a stratified k fold cross-validation procedure.

- XGBoost Classifier
- Support Vector Classifier
- Decision Tree Classifier
- AdaBoost Classifier
- Random Forest Classifier
- Extra Trees Classifier
- Ridge Classifier
- Gradient Boosting Classifier
- Multiple Layer Perceptron (neural network) Classifier
- KNN Classifier
- Logistic regression
- Linear Discriminant Analysis
- Naive Bayesian (BernoulliNB)
- Naive Bayesian (GaussianNB)
- CatBoost Classifier
- LightGBM Classifier

Catboost and LightGBM classifiers are different from other classifiers as they can handle categorical features without one-hot encoding or any other encoding. So we did the modeling for those two classifiers separately.

	CrossValMeans	CrossValerrors	Algorithm
0	0.859521	0.009114	XGB
1	0.825145	0.010496	SVC
2	0.683764	0.011340	DecisionTree
3	0.797626	0.055283	AdaBoost
4	0.858096	0.008986	RandomForest
5	0.849472	0.007530	ExtraTrees
6	0.831726	0.009785	RidgeRegression
7	0.867893	0.008188	GradientBoosting
8	0.820554	0.010182	MultipleLayerPerceptron
9	0.736587	0.016406	KNeighboors
10	0.835068	0.009479	LogisticRegression
11	0.831715	0.009780	LinearDiscriminantAnalysis
12	0.713239	0.007748	BNB
13	0.737868	0.010368	GNB

	CrossValMeans	CrossValerrors	Algorithm
0	0.854192	0.004744	XGB
1	0.857985	0.005424	SVC
2	0.683702	0.007821	DecisionTree
3	0.768396	0.042765	AdaBoost
4	0.852786	0.005791	RandomForest
5	0.849305	0.005293	ExtraTrees
6	0.850093	0.004292	RidgeRegression
7	0.861738	0.004642	GradientBoosting
8	0.831501	0.008402	MultipleLayerPerceptron
9	0.787669	0.006507	KNeighboors
10	0.850793	0.004353	LogisticRegression
11	0.850094	0.004291	LinearDiscriminantAnalysis
12	0.720376	0.009724	BNB
13	0.753067	0.007219	GNB

	CrossValMeans	CrossValerrors	Algorithm
0	0.863333	0.005316	LGBM
1	0.865709	0.005511	CatBoost

H1N1 vaccine

	CrossValMeans	CrossValerrors	Algorithm
0	0.868750	0.007477	LGBM
1	0.871701	0.008122	CatBoost

Seasonal flu vaccine

We decided to choose the GradientBoost, XgBoost, RandomForest, and the ExtraTrees classifiers for the ensemble modeling for both h1n1 and seasonal flu as they gave the best cross-validation scores. As mentioned earlier, ensembling Catboost and LightGBM classifiers done separately.

4.1.1.1.1 Hyperparameter tuning for best models

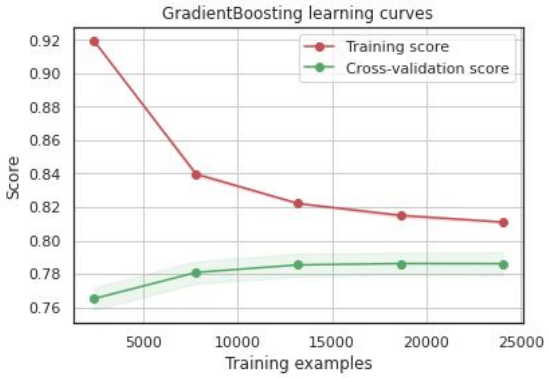
We performed a grid search optimization for GradientBoost, ExtraTrees, RandomForest, XgBoost, Catboost, and LightBGM classifiers.

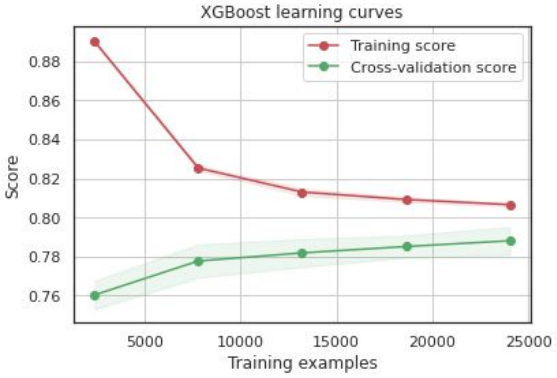
The following hyperparameters were tuned in 4 classifiers.

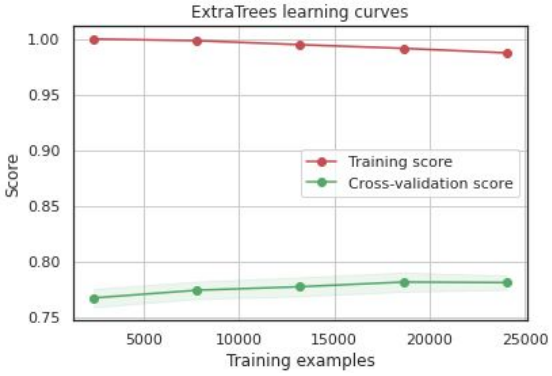
XGBoost	GradientBoost	RandomForest	ExtraTrees
Learning rate	Learning rate	Max depth	Max depth
Max depth	Max depth	Max features	Max features

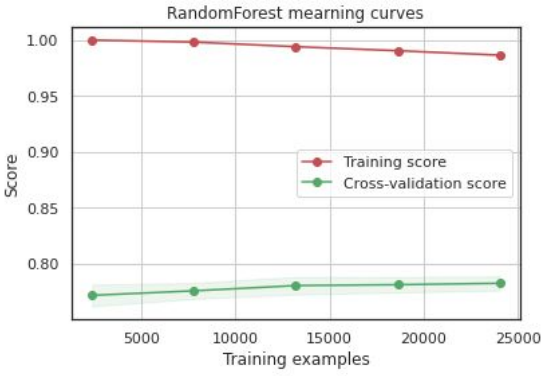
Number of estimators	Number of estimators	Number of estimators	Number of estimators
Min child weight	Max features	Min sample split	Min sample split
Gamma	Subsample	Min sample leaf	Min sample leaf
Booster	Min samples split		
Colsample by tree	Min samples leaf		

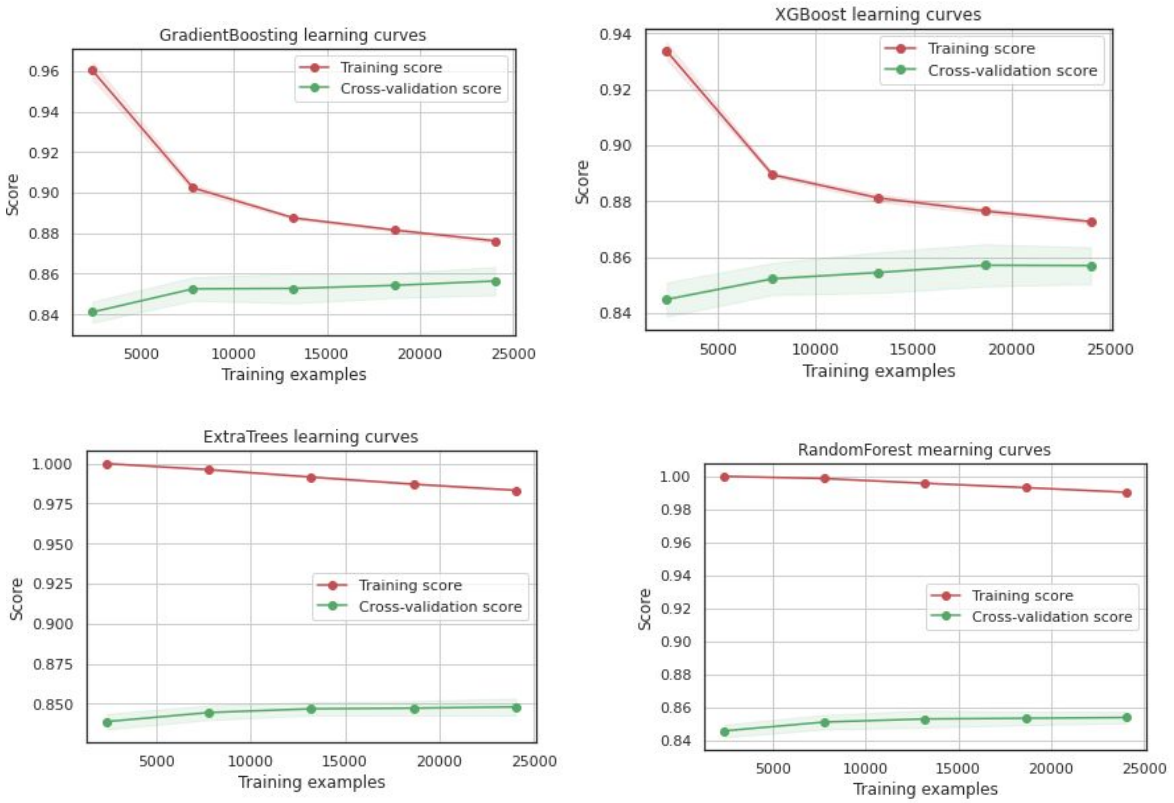
Then plot the learning curves to see the overfitting effect on the training set and the effect of the training size on the accuracy









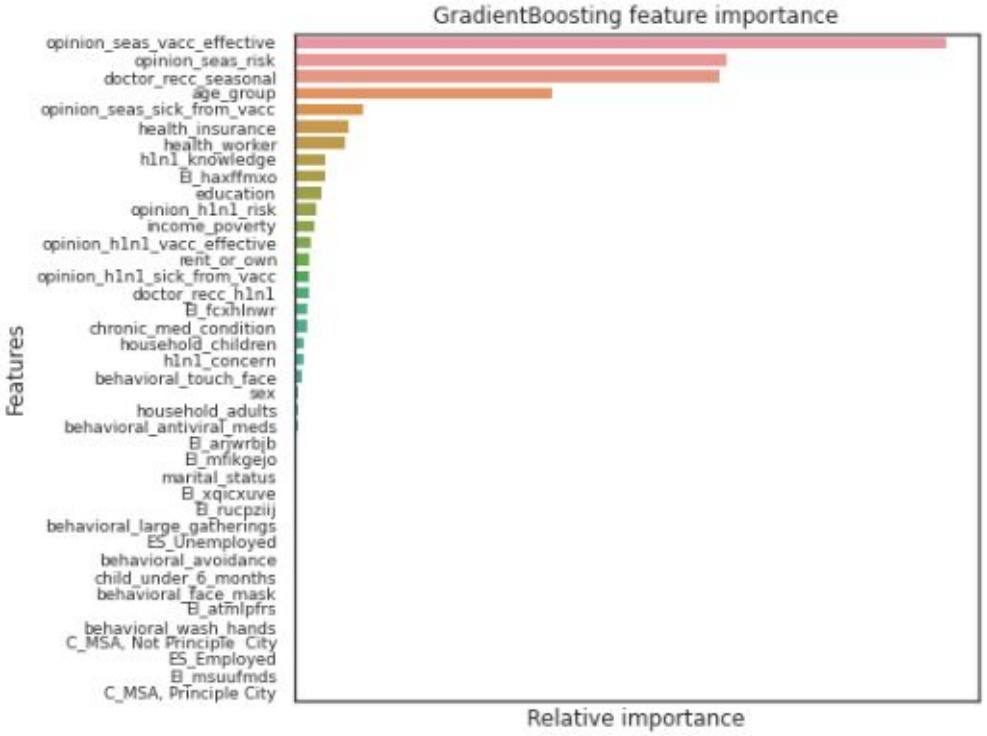


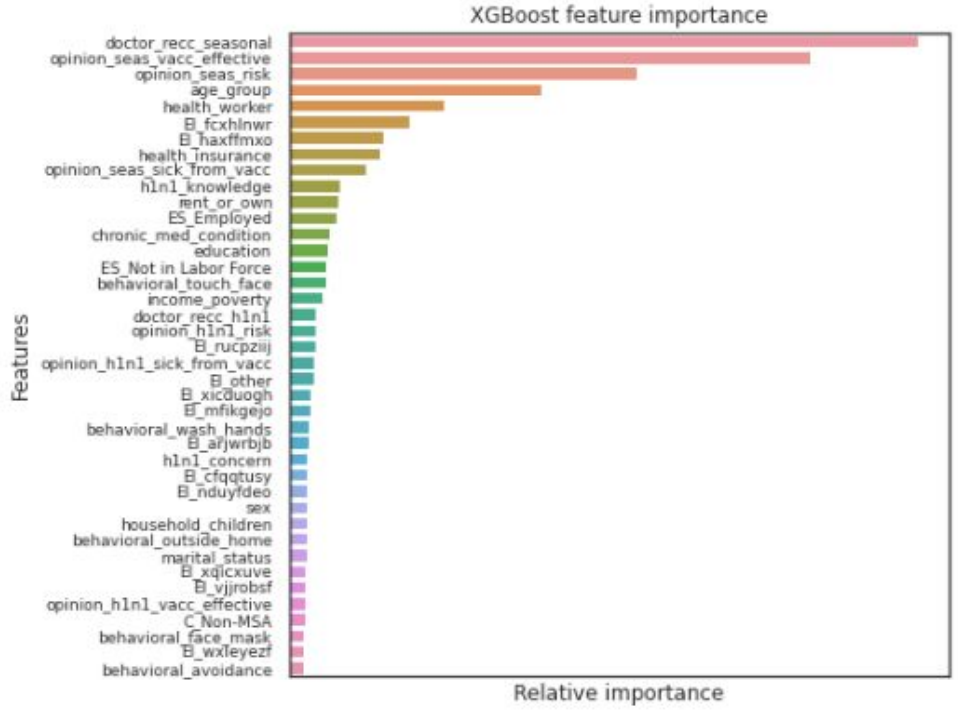
H1N1 Vaccine

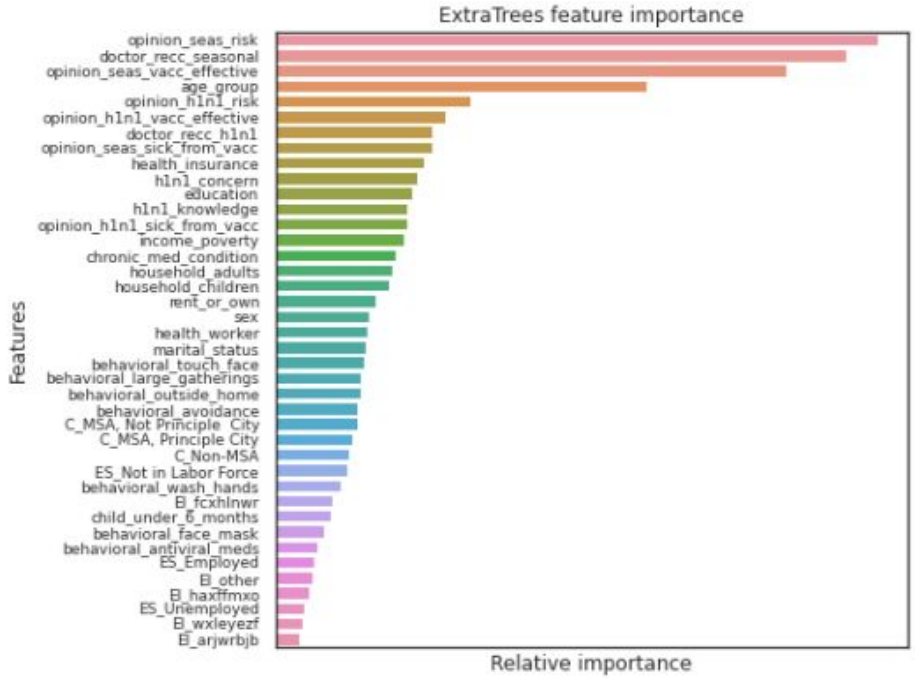
For both targets seasonal flu and H1N1, all 4 classifiers show similar learning curves. XGBoost and GradientBoost classifiers seem to better generalize the prediction since the training and cross-validation curves are close together. According to the graphs, RandomForest and ExtraTrees classifiers tend to overfit the training set. According to the growing cross-validation curves, GradientBoosting could perform better with more training examples in the seasonal flu case while xgboost could perform better with more training examples in the h1n1 case.

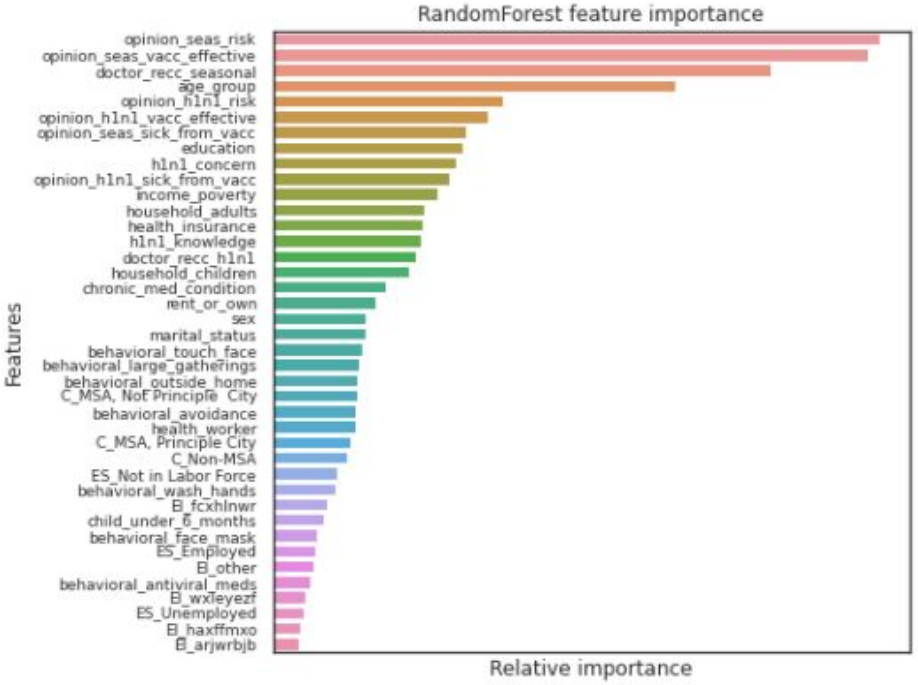
4.1.1.1.2 Feature importance of tree-based classifiers

In order to see the most informative features for the prediction, we displayed the feature importance for the 4 tree-based classifiers.

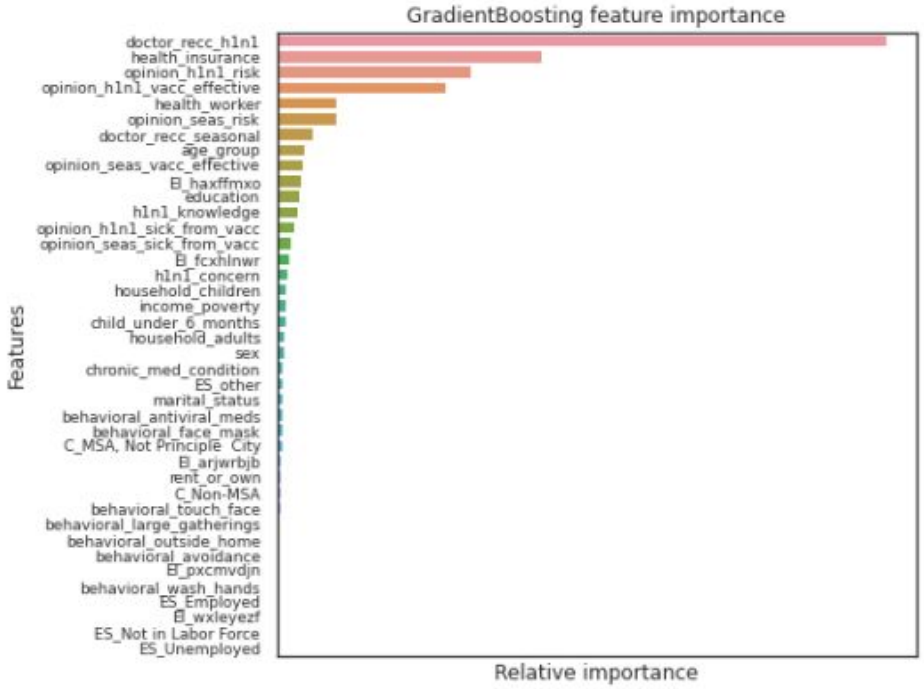

GradientBoosting feature importance
opinion_seas_vacc_effective
opinion_seas_risk
doctor_recc_seasonal
age_group
opinion_seas_sick_from_vacc
health_insurance
health_worker
h1n1_knowledge
B_haxffmxo
education
opinion_h1n1_risk
income_poverty
opinion_h1n1_vacc_effective
rent_or_own
opinion_h1n1_sick_from_vacc
doctor_recc_h1n1
B_fcxhlnwr
chronic_med_condition
household_children
h1n1_concern
behavioral_touch_face
sex
household_adults
behavioral_antiviral_meds
B_arjwrbjb
B_mfikgejo
marital_status
B_xqicxuve
B_rucpziij
behavioral_large_gatherings
ES_Unemployed
behavioral_avoidance
child_under_6_months
behavioral_face_mask
B_atmlpfrs
behavioral_wash_hands
C_MSA, Not Principle City
ES_Employed
B_msuufmds
C_MSA, Principle City
Features
Relative importance

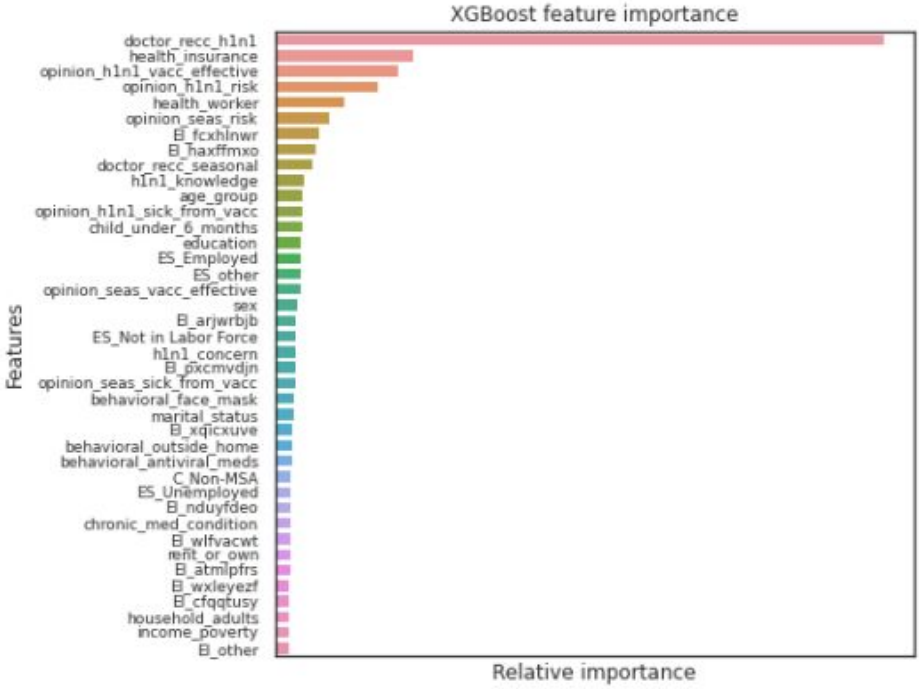

XGBoost feature importance
doctor_recc_seasonal
opinion_seas_vacc_effective
opinion_seas_risk
age_group
health_worker
B_fcxhlnwr
B_haxffmxo
health_insurance
opinion_seas_sick_from_vacc
h1n1_knowledge
rent_or_own
ES_Employed
chronic_med_condition
education
ES_Not in Labor Force
behavioral_touch_face
income_poverty
doctor_recc_h1n1
opinion_h1n1_risk
B_rucpziij
opinion_h1n1_sick_from_vacc
B_other
B_xicduogh
B_mfikgejo
behavioral_wash_hands
B_arjwrbjb
h1n1_concern
B_cfqqtusy
B_nduyfdeo
sex
household_children
behavioral_outside_home
marital_status
B_xqicxuve
B_vjjrobsf
opinion_h1n1_vacc_effective
C_Non-MSA
behavioral_face_mask
B_wxleyezf
behavioral_avoidance
Features
Relative importance

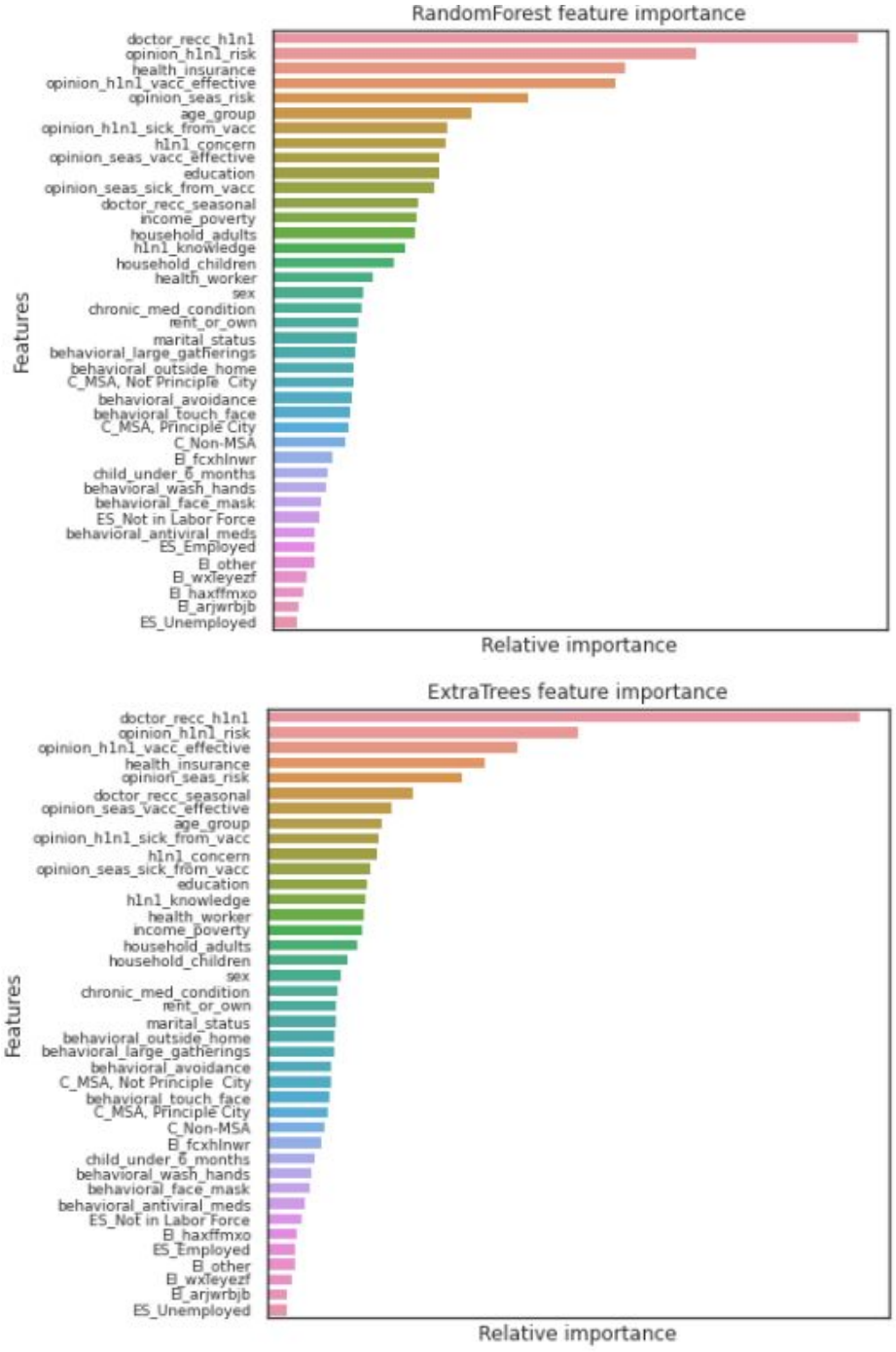




Seasonal fluVaccine


GradientBoosting feature importance
doctor_recc_h1n1
health_insurance
opinion_h1n1_risk
opinion_h1n1_vacc_effective
health_worker
opinion_seas_risk
doctor_recc_seasonal
age_group
opinion_seas_vacc_effective
EI_haxffmxo
education
h1n1_knowledge
opinion_h1n1_sick_from_vacc
opinion_seas_sick_from_vacc
EI_fcxhlnwr
h1n1_concern
household_children
income_poverty
child_under_6_months
household_adults
sex
chronic_med_condition
ES_other
marital_status
behavioral_antiviral_meds
behavioral_face_mask
C_MSA, Not Principle City
EI_arjwrbjb
rent_or_own
C_Non-MSA
behavioral_touch_face
behavioral_large_gatherings
behavioral_outside_home
behavioral_avoidance
EI_pxcmvdjn
behavioral_wash_hands
ES_Employed
EI_wxleyezf
ES_Not in Labor Force
ES_Unemployed
Features
Relative importance


XGBoost feature importance
doctor_recc_h1n1
health_insurance
opinion_h1n1_vacc_effective
opinion_h1n1_risk
health_worker
opinion_seas_risk
EI_fcxhlnwr
EI_haxffmxo
doctor_recc_seasonal
h1n1_knowledge
age_group
opinion_h1n1_sick_from_vacc
child_under_6_months
education
ES_Employed
ES_other
opinion_seas_vacc_effective
sex
EI_arjwrbjb
ES_Not in Labor Force
h1n1_concern
EI_pxcmvdjn
opinion_seas_sick_from_vacc
behavioral_face_mask
marital_status
EI_xqicxuve
behavioral_outside_home
behavioral_antiviral_meds
C_Non-MSA
ES_Unemployed
EI_nduyfdeo
chronic_med_condition
EI_wlfvacwt
rent_or_own
EI_atmlpfrs
EI_wxleyezf
EI_cfqqtusy
household_adults
income_poverty
EI_other
Features
Relative importance



H1N1 Vaccine

We note that the 4 classifiers have different top features according to the relative importance. It means that their predictions are not based on the same features. Nevertheless, they share some common important features for the classification, for example,

Seasonal flu - opinion_seas_risk, doctor_recc_seasonal, opinion_seas_vacc_effective, and age group

H1N1 - opinion_h1n1_risk, doctor_recc_seasonal and opinion_h1n1_vacc_effective.

4.1.1.2 CatBoost and LightGBM Classifiers

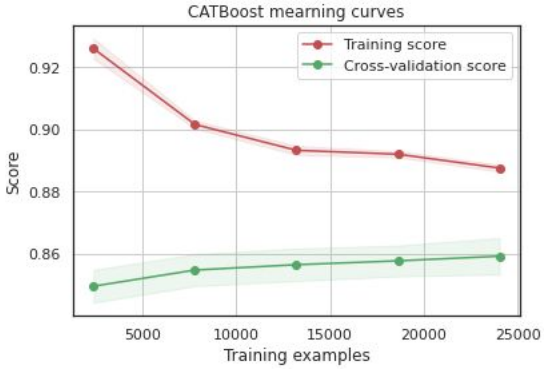
Modeling for CatBoost and LightGBM classifiers done separately as those classifiers handle categorical features separately. Categorical features were used without encoding with these classifiers. Grid search optimization was done for these two classifiers as well to tune their hyper-parameters.

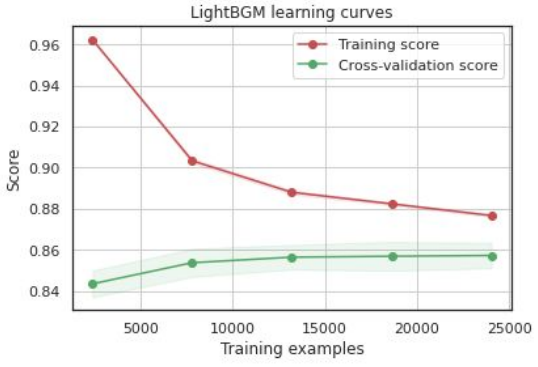
Hyperparameter Tuning

Following hyperparameters were tuned in these two classifiers using gridsearch optimizations.

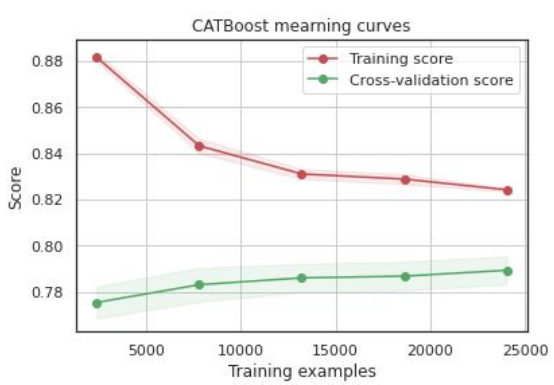
CatBoost	LightGBM
Depth	Number of estimators
Learning rate	Learning rate
Iterations	Number of iterations
l2_leaf_reg	Number of leaves

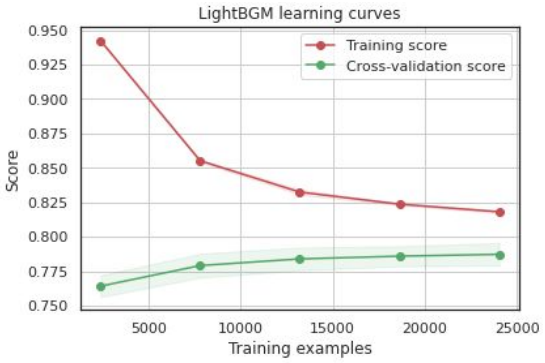
Learning Curves





H1N1 Vaccine





Seasonal flu Vaccine

Both CatBoost and LightGBM classifiers seem to better generalize the prediction since the training and cross-validation curves are close together.

4.4.2 Ensemble Modelling

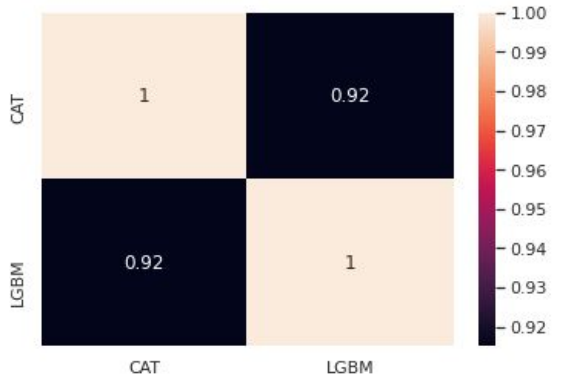
The following figures show the correlation between the predictions on the test by 4 classifiers(XGBoost, RandomForest, GradientBoost, ExtraTrees) for each of the two cases.

Seasonal flu Vaccine

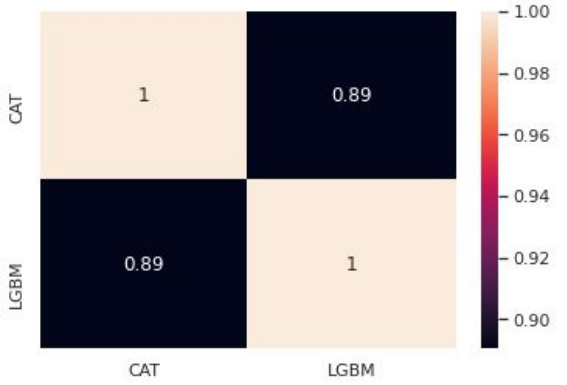
H1N1 Vaccine

The prediction seems to be quite similar for the 4 classifiers. The 4 classifiers give more or less the same prediction but there are some differences. These differences between the 4 classifier predictions are sufficient to consider an ensemble.

The following figures show the correlation between the predictions on the test by CatBoost and LightGBM classifiers for each of the two cases.



Seasonal flu Vaccine



H1N1 Vaccine

When ensembling also, CatBoost and LightGBM classifiers were ensembled separately.

For ensembling, we used two approaches: voting and stacking.

4.4.2.1 Voting

We chose a voting classifier to combine the predictions coming from the 5 classifiers and passed the argument "soft" to the voting parameter to take into account the probability of each vote.

4.4.2.2 Stacking

When stacking, we used RandomForest, ExtraTrees, and GradientBoost classifiers as base first-level models. Then an XGBoost classifier was trained as the second level on the first level outputs.

4.4.3 Deep Neural Network Modelling

Neural network-based model with a few hidden layers was trained to predict the h1n1_vaccine and seasonal_flue_vaccine. Neural networks with 2 hidden layers gave the best scores among trained neural networks. Unlike the above classification algorithms used, this approach doesn't need two separate models to predict both h1n1 and seasonal_flue.

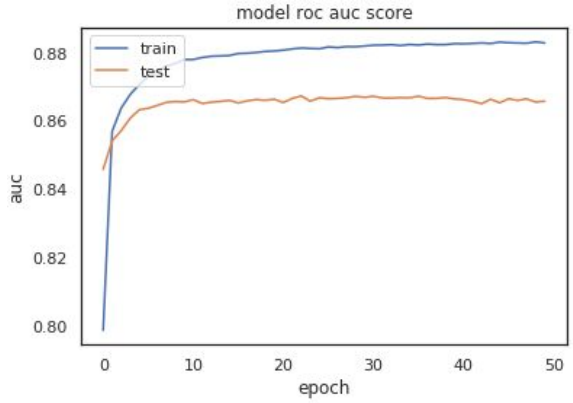
Hyperparameter Tuning

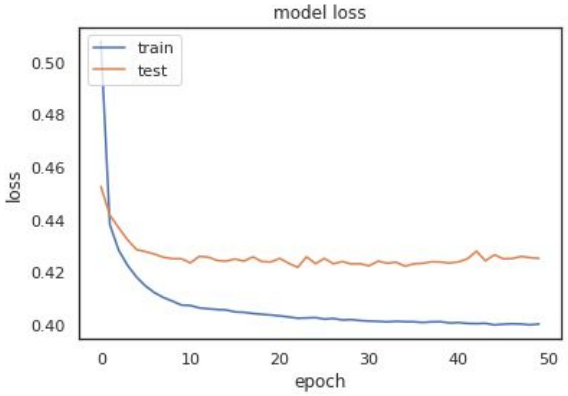
Following hyperparameters were tuned in the model.

- epoch
- batch size
- optimizers

- kernal_initializer
- activation
- no of hidden layers
- no of hidden nodes

Learning Curves





4.5 Prediction

Following is a summary of predictions done using different models. Best scores received from each model type is displayed.

Models	ROC_AUC score
XGBoost	0.8605
Random Forest	0.8564
CatBoost	0.8616
Voting with XGBoost, GradientBoost, ExtraTrees and Random Forest	0.8615
Voting with XGBoost, GradientBoost, and RandomForest	0.8624

Stacking with XGBoost as second-level model and GradientBoost, RandomForest as base level models	0.8612
Stacking with RandomForest as second-level model and GradientBoost, XGBoost as base level models	0.8598
Voting with CatBoost and LightGBM	**0.8629**
Deep Neural Network model	0.8445

.

Overall best leaderboard score was achieved by the model with the voting ensemble on Catboost and LightGBM classifiers.

5. Discussion

For the competition, we have submitted 21 submissions. These submissions are related to different models that are described above. Most of them are voting classifiers with xgboost, random forest, gradient boost, extra trees, etc.

As the first step, we identified this project as a classification problem and we gathered data with 35 features and analyzed the dataset. Then we performed Label analysis, Feature analysis and after that identified the correlations between features and the target variables and cross-correlation of the features.

Data preprocessing is one of the most important steps in data mining as incomplete, noisy, and inconsistent data present in the dataset affect the quality of the prediction. We had to preprocess the noisy dataset before training the models with the dataset. We followed three major steps in the data preprocessing. i.e. data cleaning, data transformation, and data reduction. We used several data preprocessing approaches such as imputing missing values, normalization, categorical encoding, feature selection with correlation analysis, etc. All techniques that were used are described in the above sections.

In the modeling part, we created separate two models for predicting h1n1 and seasonal flu. First, we compared 16 popular classifiers and evaluated the mean accuracy of each of them by a stratified k fold cross-validation procedure and found out the best models. Then we implemented those models and evaluated the models. The implemented models are XGboost, Random forest, extra tree, and gradient boost. We evaluated these models as simple models as well as using ensembling techniques such as voting and stacking with different combinations of models. The ensembling models gave better scores compared to simple models. Later we tried CatBoost and LightGBM which gave better scores. We also tried a neural network-based model to predict the h1n1 and seasonal flu. However, that performed poorly relative to other models. The best score was given by the voting classifier

with Catboost and LightGBM. It is noticed that in addition to ensembling, hyperparameter tuning increased the score significantly.

6. Conclusion

In our project, we created predictive models of H1N1 and Seasonal flu vaccines in connection with the 'Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines' competition hosted by drivenaata.org.
We used several classification models as well as deep neural network models. After initial submissions, we recognized that neural network-based methods perform poorly relative to other classification models. So we mainly used XGBoost, Random Forest, Gradient Boost, Extra Trees, CatBoost, and LightBGM classifiers for further optimization.
Most of those models predicted whether the people got the H1N1 and Seasonal flu vaccine with good enough accuracy, even with minimal human intervention. We obtained our best score for the ensemble model of Catboost and LightGBM classifiers which has given a ROC_AUC score of 0.8629. The importance of Data Mining and Machine Learning is emphasized by a competition like this.

7. References

[1] lastnightstudy. (2019). Data Transformation In Data Mining - Last Night Study. [online] Available at: http://www.lastnightstudy.com/Show?id=42/Data-Transformation-In-Data-Mining [Accessed: 06- Jul- 2020]

[2] "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines", *DrivenData*, 2020. [Online]. Available: https://www.drivendata.org/competitions/66/flu-shot-learning/. [Accessed: 06- Jul- 2020].