Classification Model for Predicting Whether Individuals took the H1N1 Vaccine or Not

Business Understanding

The public health organization is interested in finding the response of individuals to the H1N1 vaccine in 2010, if a majority took it or not. With the cropping up of global pandemics like COVID-19, public health has become very critical and a priority to the government and health sector. For this reason, they saw it fit to study vaccines, and specifically, whether individuals get them or not. Having been approached with this, we look to build a classification model that will establish whether individuals took the h1n1 vaccine and the factors that were correlated to them and predict an individual's possible reaction to a case like this in the future. This will help the stakeholders to establish a feasible way to approach this issue in future.

We apply machine learning for this project because it will give the stakeholders a view of trends in individuals' behaviors in relation to their response to vaccines, as well as support the development of new vaccines if need be.

Objectives

The main objective of this project is to build a model that will predict the response of individuals to a newly introduced vaccine (in this case H1N1) so it can lay a foundation for what direction of research the health sector should take.

- To identify the most significant features in determining insurance premiums
- To build a linear regression model that can accurately predict insurance premiums based on input features from the Kaggle insurance premium prediction dataset.
- To assess the performance of the predictive model and identify potential areas for improvement.

Data Understanding

Data Source

The dataset used for this project was obtained from DrivenData.

Data Description

The dataframe contains 26707 observations and 35 features. The data on the training features (the input variables that the model will use to predict the probability that people received H1N1 flu vaccines and seasonal flu vaccines) contains 35 feature columns in total, each a response to a survey question.

Data Preparation

Loading the data

At the beginning of the process, the necessary libraries were imported and then the training_features_df dataset was loaded onto the jupyter notebook using pandas.

Reading and checking the data

The data was read and then checked for anomalies, outliers, missing values and duplicates. This was to determine the next course of action that would ensure the data would be set for use. During this process, it was established that the data had a lot of missing values and duplicates .

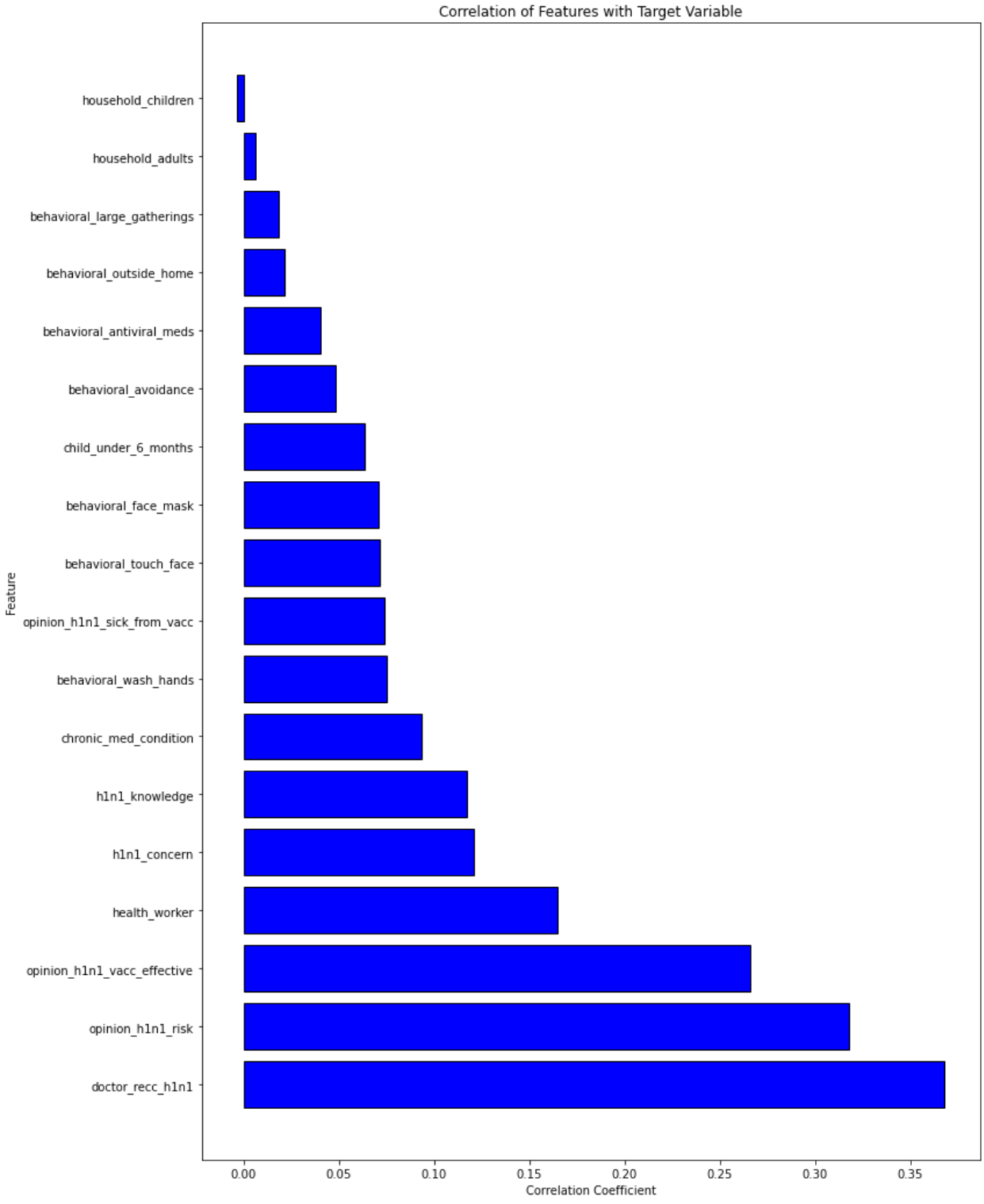
Cleaning the data

The data was then cleaned and pre-processed to ensure that it was in a usable form. This was done by removing duplicates. Some of the missing values after it was established that the model could do without them, were removed. The rest of the missing values were filled using the bfill method. The duplicates were dropped because they would have decreased the accuracy of the model if it was fitted on duplicated data.

Exploratory Data Analysis

The data sets were analyzed and trends found by using statistics and visualizations to aid in comprehending the data set. The proportion of H1N1 vaccine was established and only about 20% of the interviewed population had taken the vaccine.

A stack of horizontal graphs were plotted to study the correlation between the features with the target variable. Among the most correlated variables to the target variable (h1n1_vaccine) were doctor's recommendation, individuals' opinion on the risks posed by the vaccine, opinions and how effective the vaccine is for H1N1 concern. The plot below shows this:


Correlation of Features with Target Variable
household_children
household_adults
behavioral_large_gatherings
behavioral_outside_home
behavioral_antiviral_meds
behavioral_avoidance
child_under_6_months
behavioral_face_mask
behavioral_touch_face
opinion_h1n1_sick_from_vacc
behavioral_wash_hands
chronic_med_condition
h1n1_knowledge
h1n1_concern
health_worker
opinion_h1n1_vacc_effective
opinion_h1n1_risk
doctor_recc_h1n1
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Correlation Coefficient

Modeling

The first was the base model; a simple logistic regression model that was fitted on the training set before any feature selection was done. Fitting this model to the test data resulted in a very weak f1 score of 43%, and regardless of the contrasting higher accuracy score, the model was considered a poor performing model. In an attempt to improve our results, feature selection was performed and the most important features used to fit the second model which was KNN. This, however, proved to be performing even worse than the previous one with an f1 score of 36%. For the third model, we built a decision tree that resulted in an f1 score of 36%, still a poor performance. Following the same kind of results after building three models, there was need to do more to improve the outcome. It was at this point that hyperparameter pruning was carried out on the decision tree model and an improvement in the f1 score was seen at 78%.

Conclusion

The decision tree, which was modified via hyper-parameters, was chosen as the final model after experimenting with many models using various methodologies. This is due to the fact that it recorded an accuracy score of 80% and a f1 score of 78%, which matched the success requirements. Because they were not complicated enough to handle the type of data supplied, the first three models did not perform as well. Because of this, the f1 score remained low even though the accuracy score was high. We were able to identify the ideal complexity that struck a compromise between overfitting and under fitting by changing the hyper-parameters.