Flu Shot Learning: Predicting Whether Individuals got their H1N1 Vaccines

Business Understanding

The spread of infectious diseases is a significant public health concern, and vaccines are seen as a key tool to control their spread. However, vaccine hesitancy is a growing concern, and understanding the factors that influence people's vaccination behavior is essential to designing effective vaccination campaigns. The purpose of this project is to develop a classification model to predict the likelihood of individuals receiving the H1N1 and seasonal flu vaccines, based on demographic, social, economic, and health-related data. The project will use the National 2009 H1N1 Flu Survey data to understand the public's behavior towards vaccines and identify the factors that influence their decision-making process. The target audience for this project is public health officials, policymakers, and healthcare providers responsible for designing and implementing vaccination campaigns. The findings of this study can help them understand the factors that influence vaccine hesitancy and design targeted interventions to improve vaccine uptake. The performance of the classification model will be evaluated using accuracy, recall, and precision scores, and the results of the study will be communicated transparently and in compliance with ethical guidelines for data handling and analysis.

Objectives

The main objective of this project is to develop a classification model to predict the response of individuals to H1N1 and seasonal flu vaccines.

To achieve the main objectives, the following additional objectives will be pursued:

- Explore and visualize the data to gain a better understanding of the relationship between the variables and the target variable.
- Select and train different classification models, including logistic regression, decision trees, and random forest and evaluate their performance.
- To provide actionable insights for public health officials and policymakers to reduce the spread of contagious infections and promote herd immunity.

Data Understanding

Data Source

The dataset used for this project was obtained from Datadriven. The datasets we named training_set_features , training_set_labels and test_set_features

Data Description

The dataset provided contains 36 columns where the first column is respondent_id, which is a unique and random identifier. The remaining 35 features provide information about the level of concern and knowledge about the H1N1 flu, behavior towards the flu, the recommendation of the H1N1 and seasonal flu vaccines by doctors, chronic medical conditions, contact with a child under six months, health worker status, health insurance, opinions about the H1N1 and seasonal flu vaccines, age group, education level, race, sex, household annual income, marital status, housing situation, employment status, respondent's residence using a 10-region geographic classification and respondent's residence within metropolitan statistical areas, the number of adults and children in the household, and the type of industry and occupation the respondent is employed in. Most of the features are binary variables indicating yes or no, while some are multi-level variables representing different opinions, levels of concern, or knowledge.

Data Preparation

Loading the data

At the beginning of the process, the necessary libraries were imported and then the datasets were imported into to Jupyternotebook

Reading, checking the data and cleaning

After reading the data, it was examined for anomalies, outliers, and 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 datasets had missing values, especially the training_set_features which had a lot of missing values. The training_set_labels had no missing values while test_set_features also had missing values The data was then cleaned and pre-processed to ensure that it was in a usable form. This was done by filling in some for example on the training_set_features I replaced missing values "Not Available".

Exploratory Data Analysis

Utilizing statistics and visualizations, the data sets were examined to identify trends that would make the data easier to understand. There were several questions that were answered in this step by comparing the predictor variables with the target variable which was expense (premium) using data visualization tools. The questions answered and variable relationships established include the following:

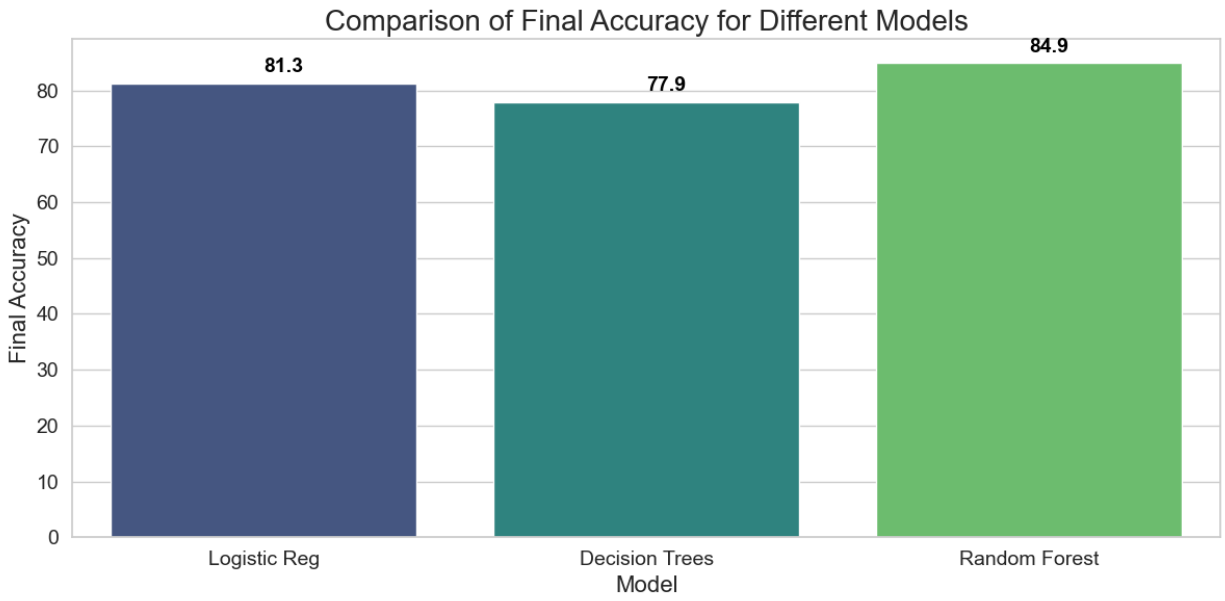
- Does age_group affect the intake of H1N1 vaccine and seasonal_vaccine?
- Does education affect the in_take of H1N1 vaccine and seasonal_vaccine?
- Does Insurance affect the intake of H1N1 vaccine and seasonal_vaccine?
- Does a region affect the intake of H1N1 vaccine and seasonal_vaccine ?

- Does employment_status affect the intake of H1N1 vaccine and seasonal_vaccine?
- Does Race affect the intake of H1N1 vaccine and seasonal_vaccine ?

Modeling

I started my commerce modeling by preparing the data for machine learning models. This involved splitting the data into training and testing sets, scaling the numerical features using the StandardScaler method, and encoding the categorical features using the OrdinalEncoder method. The same steps were applied to the testing set, resulting in the variables X_train, X_test, y_train, and y_test.

Next, I tested three machine learning algorithms: logistic regression, decision trees, and random forest. After evaluating the accuracy, precision, recall, and final accuracy of each model, I chose Random Forest as the best performer, with an accuracy of 84.9% for H1N1 and 78.5% for the seasonal flu vaccine. Logistic regression had an accuracy of 81.3% and decisioni tree had an accuracy of 77.9%. The overall entropy accuracy for both datasets was 0.78
After evaluation, the Random Forest model was chosen for final evaluation on the test dataset as it achieved the highest accuracy of 84.9% for H1N1 and 78.5% for the seasonal flu vaccine. The Logistic Regression model had an accuracy of 81.3% for both datasets and the Decision Tree model had an accuracy of 77.9%. The overall entropy accuracy for both datasets combined was 0.78.



Conclusion

This analysis is deemed important in achieving the objectives described above and would be helpful in the campaign towards vaccination and immunization