TEAM 3 PHASE 3 PROJECT CRISP - DM REPORT

PROJECT:

A CLASSIFICATION MODEL TO PREDICT WHETHER AN INDIVIDUAL WILL RECEIVE A VACCINE OR NOT BASED ON PREVIOUS H1N1 VACCINATION RECORDS.

TEAM MEMBERS:

- **Anthony Ngatia**
- **Jessyca Aperi**
- **Joy Chepchumba**
- **Naomi Rotich**

DATE: 10TH MAY 2023

BUSINESS UNDERSTANDING

BUSINESS OVERVIEW

As the world struggles to vaccinate the global population against COVID-19, an understanding of how people's backgrounds, opinions, and health behaviors are related to their personal vaccination patterns can provide guidance for future public health efforts. We aim to predict whether people got H1N1 and seasonal flu vaccines using data from the National 2009 H1N1 Flu Survey.

This is a binary classification problem, but there are two potential targets: whether the survey respondent received the seasonal flu vaccine, or whether the respondent received the H1N1 flu vaccine.

We chose to work with one target i.e. whether the respondent received the H1N1 flu vaccine, and hence this is a single classification problem.

BUSINESS OBJECTIVES

Our main objective is to build a classification model to predict whether an individual will presumably get an H1N1 vaccine or not.

Specific Objectives

- To import and clean the dataset inorder to prepare the data for analysis and modeling.
- To model the data using three techniques namely: Decision trees, Random Forest, and Logistic Regression.
- To perform feature selection of our dataset.
- To validate our model using different metrics.

Business Success Criteria

To be able to predict whether an individual will receive an H1N1 vaccination or not,. This will assist public health efforts by health institutions to classify the likelihood of an individual receiving a vaccine or not in the future.

ASSESSING THE SITUATION

Resource Inventory

Datasets

- H1N1 and Seasonal Flu Vaccines

Software Used

- Google Colab
- Pandas
- Numpy
- GitHub

ASSUMPTIONS

The data provided is correct and up to date.

CONSTRAINTS

Selection of one target to use for the prediction model i.e H1N1 vaccination

PROJECT GOALS

The goal of this project was to build a classification model to predict the probability of a patient receiving a H1N1 vaccine based on different underlying factors such as health behaviors, opinion, and individual background. The parametric used adhered to the project constraints i.e. time bound, cost-effectiveness, and scope.

PROJECT PLAN

The project was conducted by a team of four members. The project was divided into major sections i.e. Data importation and analysis, Data Understanding and cleaning, Data Modelling and deployment, Non-technical presentation and reporting. The teammates were allocated tasks and expected to deliver within the timelines specified.

DATA MINING GOALS

Our data mining goals for this project are as follows;

- To import and download the dataset from Driven Data: - The datasets were already provided and hence we did not perform data scrapping.

DATA MINING SUCCESS CRITERIA

Our success will be measured by the following criteria.

- Downloading the datasets from the data sources i.e. Driven Data
- Loading the datasets successfully onto the Google Colab workspace

DATA UNDERSTANDING

Overview

For this project, we are using the available dataset in Kaggle.

The dataset is the H1N1 and seasonal flu vaccines. The link is attached below;

https://www.drivendata.org/competitions/66/flu-shot-learning/

DATA DESCRIPTION

There are three datasets provided for this project namely:

- Training features - containing the training dataset
- Test features - containing the test dataset
- Training labels - containing the set labels of the H1N1 vaccines

The dataset available for this project contains information about various respondents. The column definitions are presented below:

- `h1n1_concern` - Level of concern about the H1N1 flu.

    - 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- h1n1_knowledge - Level of knowledge about H1N1 flu.
    - 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- 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 own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or an other 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)
- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)

- `opinion_h1n1_vacc_effective` - Respondent's opinion about H1N1 vaccine effectiveness.
    - `1` = Not at all effective; `2` = Not very effective; `3` = Don't know; `4` = Somewhat effective; `5` = Very effective.
- `opinion_h1n1_risk` - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine.
    - `1` = Very Low; `2` = Somewhat low; `3` = Don't know; `4` = Somewhat high; `5` = Very high.
- `opinion_h1n1_sick_from_vacc` - Respondent's worry of getting sick from taking H1N1 vaccine.
    - `1` = Not at all worried; `2` = Not very worried; `3` = Don't know; `4` = Somewhat worried; `5` = Very worried.
- `opinion_seas_vacc_effective` - Respondent's opinion about seasonal flu vaccine effectiveness.
    - `1` = Not at all effective; `2` = Not very effective; `3` = Don't know; `4` = Somewhat effective; `5` = Very effective.
- `opinion_seas_risk` - Respondent's opinion about risk of getting sick with seasonal flu without vaccine.
    - `1` = Very Low; `2` = Somewhat low; `3` = Don't know; `4` = Somewhat high; `5` = Very high.
- `opinion_seas_sick_from_vacc` - Respondent's worry of getting sick from taking seasonal flu vaccine.
    - `1` = Not at all worried; `2` = Not very worried; `3` = Don't know; `4` = Somewhat worried; `5` = Very worried.
- `age_group` - Age group of respondent.
- `education` - Self-reported education level.
- `race` - Race of respondent.

- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of 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.
- household_adults - Number of *other* adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.
- employment_occupation - Type of occupation of respondent. Values are represented as short random character strings.

VERIFYING DATA QUALITY

The data did not meet our required quality because after checking through we found that the dataset had missing values but we replaced the missing values with the mode. Moreover, we dropped some columns that were not relevant to the case study.

DATA PREPARATION

In this phase, we conducted data munging which involved preparing the final dataset for modeling.

Data Selection - The dataset was provided on the Kaggle website namely: H1N1 and seasonal flu vaccines

Data Cleaning - The dataset was corrected by removing missing values, dealing with duplicated records, and conducting univariate, bivariate, and multivariate analysis. Other operations such as checking for multicollinearity were performed to ensure that the variables in the dataset were non-biased.

Data Integration - The training features dataset was combined with the testing features dataset and training set labels in order to clean all the datasets and use them together to perform modeling.

Data Formatting - we performed label encoding and one hot encoding to enable conversion of the categorical datasets into numerical columns for easier mathematical operations.

DATA MODELLING

The tasks involved in this phase were as follows:

- Selecting modeling techniques
- Generating test design
- Building the model
- Model Assessment

As instructed, we were to use three modelling techniques and choose the model with the best metrics of performance.

Baseline Classifier:

Logistic Regression

As the baseline classifier, we chose the logistic regression model. The model achieved the following results:

Training Precision: 0.6816595579682048
Testing Precision: 0.6998784933171325

Training Recall: 0.41296687808315713
Testing Recall: 0.40649258997882853

Training Accuracy: 0.8342486270594108
Testing Accuracy: 0.8370525685187958

This recall for the training and testing dataset was low and hence we proceeded to tune our model using class imbalance improvement and hyperparameter tuning, where we observed an improvement in the metrics as shown below:

Test accuracy: 0.775797513853527
Test precision: 0.4810964083175803
Test recall: 0.7184191954834157
Test f1 score: 0.5762807812057741

Decision Trees

We proceeded to perform decision tree classification where we were able to achieve the following results:

Train Accuracy: 0.9999366005198758
Test Accuracy: 0.7534821027407518

Train Recall_score: 0.9998732010397515
Test Recall_score: 0.47071277346506707

Train Precision_score: 1.0
Test Precision_score: 0.4267434420985285

We observed that our model was overfitting due to the big difference between the metric scores and proceeded to perform hyper parameter tuning and pruning to improve on the figures. We also used the mean train and mean test scores to train the model. The results improved as shown below:

Train Accuracy: 0.8241932416154187

Test Accuracy: 0.8147371574060207

Train Recall_score: 0.7804476003296773
Test Recall_score: 0.5977417078334509

Train Precision_score: 0.8552768707010352
Test Precision_score: 0.559445178335535

We further incorporated the use of k features to train the model and the model performance improved as shown below:

Train Accuracy: 0.8380777277626323
Test Accuracy: 0.8124906395087614

Train Recall_score: 0.811957141951436
Test Recall_score: 0.6005645730416372

Train Precision_score: 0.8567128236002408
Test precision_score : 0.5536759921925829

Random Forest

We performed Random Forest Modelling as our third model and obtained the following initial results:

Train Accuracy: 0.8219171243135297
Test Accuracy: 0.8241725325745095

Train Recall_score: 0.22903453136011276

The model performed well on both the train and test data. The recall score was quite low which indicated a lot of false positives. We proceeded to perform cross-validation and hyperparameter tuning to overcome the initial underfitting, and were able to obtain the following results:

Mean Training Score: 82.51%
Mean Test Score: 83.45%

Train Accuracy: 0.8440838741887169

DATA EVALUATION

Results Evaluation:

We observed that the model with the highest performance metrics of the three, was the Decision Tree modeling technique. It was noted that the initial metric performance of all three datasets was not optimal nor favorable for deployment and thus further tuning was required.

Two of the models were initially overfitting i.e. Decision Trees and Random Forest whereas the Logistic Regression model had a very low Recall Score. The training set labels contained a class imbalance where the number of people who did not receive a vaccine was way higher than the number of people who received the vaccine. This was corrected using the smote technique and the results improved.

We performed various tuning techniques to improve the accuracy of the individual models such as; hyperparameter tuning, gradient boosting, and KNN classifiers, and the performance metrics were observed to improve significantly on all three models, with the Decision Trees obtaining the highest metric performance.

CONCLUSION

Based on these findings, we can conclude that the Decision Tree is the most appropriate to proceed with for model deployment. This project was however restricted to only three modeling techniques and should incorporate further techniques such as; Neural Networks, Bayesian Networks etc. would further compare the performance metrics, to be able to achieve the model with the best performance metrics.

REFERENCES:

1. https://www.drivendata.org/competitions/66/flu-shot-learning/ accessed on 2nd May 2023 at 0900hrs.

2. https://www.dcc.fc.up.pt/~ltorgo/Papers/DFRBR/DFRBR-4.html accessed on 10th May 2023at 1000 hrs.
3. Canvas Content.