Predictive modelling of H1N1 Vaccine uptake

INTRODUCTION

- The 2019 Covid-19 pandemic really shook the world and also shown light on some of the loop holes in the health sector, some of which are vaccination trends and uptake. The prevention of infectious diseases depends heavily on vaccination which is a critical component of public health. But not everyone reacts to immunization in the same way. People react differently, they may experience different side effects such as fever, feel dizzy or soreness. Our objective is to create a reliable classification model that can correctly evaluate a person's response to h1n1 vaccine based on specific characteristics which are but not limited to age, education and race among others. The end result of this initiative will help healthcare practitioners to make decisions about the delivery of the vaccine.

CONTENT

i. Business Understanding
ii. Data Understanding
iii. Data Preparation
iv. Modelling
v. Model Evaluation
vi. Findings and Conclusion
vii. Recommendations

1.BUSINESS UNDERSTANDING

- The National 2009 H1N1 Flu Survey was carried out to establish the uptake of the vaccine. This provided an understanding of how people's opinions, age, demographic factors and health behaviors are related to the vaccination patterns.
- The aim of this project is to examine the data collected and develop a predictive model that can pinpoint certain characteristics of individuals who are likely to partake or not partake the vaccine.
- This model will serve as a valuable tool for shaping and predicting what type, i.e characteristics of individuals likely to get the vaccines now and in future and will enable in resource allocation in distribution of vaccines, education and outreach programs.

OBJECTIVES

- To build a classification model that can predict the reaction of individuals to a vaccine based on certain factors.
- Identify which factors affects individual's vaccine uptake.
- Build a classification model that accurately predicts the response of individuals to new vaccines.

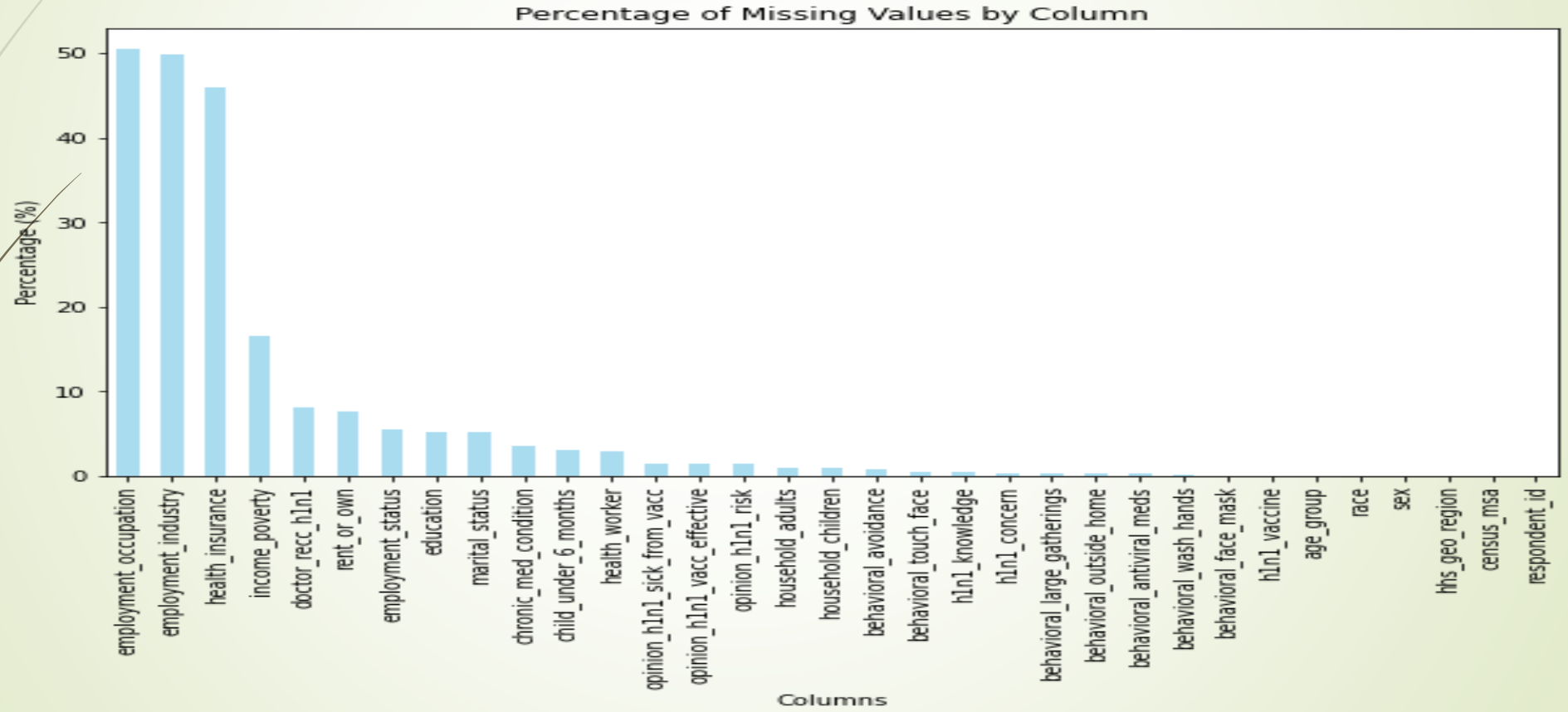
2. DATA UNDERSTANDING

- The data used in this project was retrieved from DrivenData. The dataset contains data on:
- i) training features ii) training labels iii) test 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. These questions cover several different topics, such as whether people observed safe behavioral practices, their opinions about the diseases and the vaccines, and their demographics.
- The data on training labels (the labels corresponding to the observations in the training features) contains two target variables : h1n1_vaccine and seasonal_vaccine. Both are binary variables, with 1 indicating that a person received the respective flu vaccine and 0 indicating that a person did not receive the respective flu vaccine.
- For this project the focus will be on one target variable h1n1_vaccine.

3. DATA PREPARATION

- At this stage data is cleaned and this includes identifying missing values and duplicates.
- Missing values in this dataset were filled with appropriate values in this case for numeric, mode was use and for non-numerical values, “unknown” was used to fill the null values.
- EDA was also done at this point.

Percentage of missing values

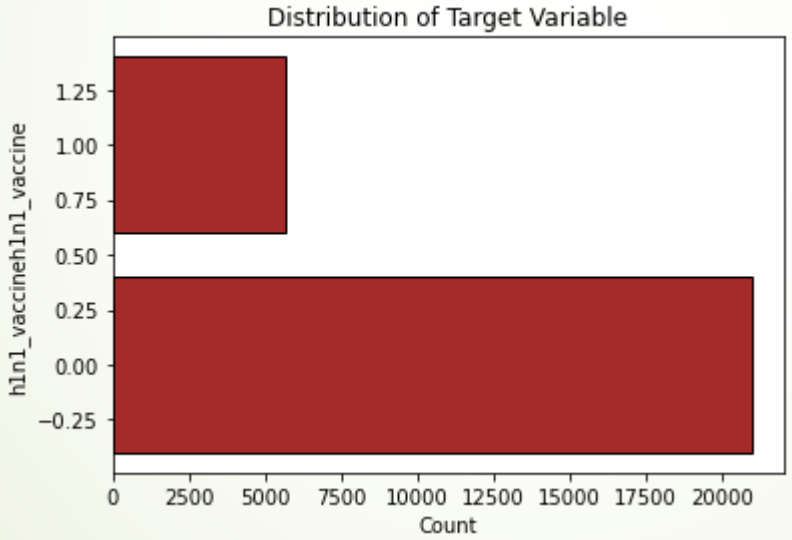


Exploratory Data Analysis

- Looking at the data before and after cleaning is important to understand the structure of the data, identify missing values, outliers, inconsistencies, and any other data quality issues. This involved Univariate and Bivariate analysis.

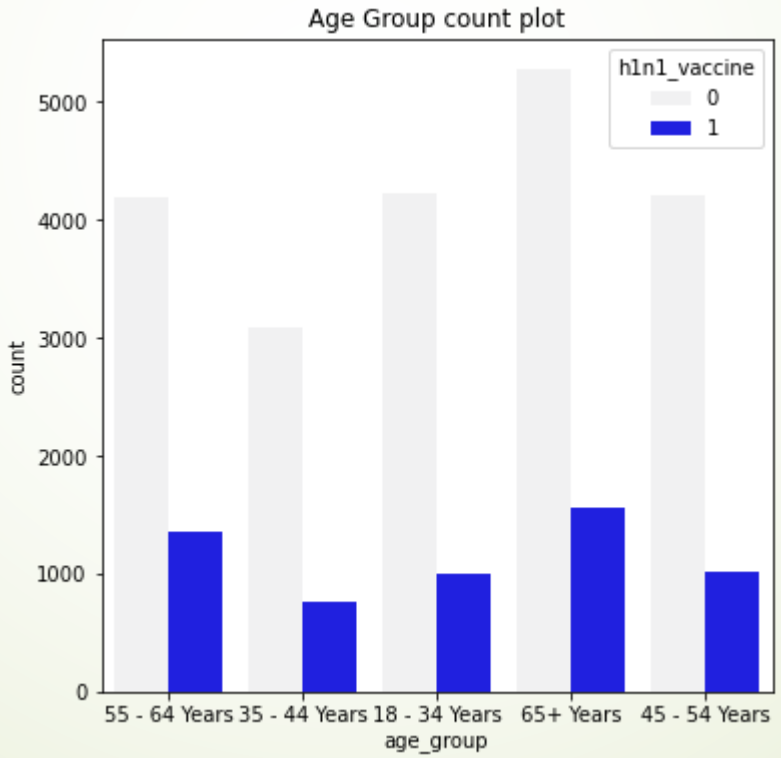
Univariate Analysis

- The goal of this analysis is to establish the proportion tar of target variable h1n1_vaccine. From the graph only 5000 people out of 20000 took the vaccine



Bivariate Analysis

- Age group seemed like a good feature to work with for bivariate analysis and gave a clear picture of how the age groups responded to vaccination.



4. MODELLING

- Modelling for this project will be done using three models to ascertain which model best predicts future uptake of the H1N1 vaccine.
- The three models are

1) Logistic Regression Model
2) Decision Trees Model
3) Random Forests Models

- Before modelling we also did Label encoding to convert the non-numerical labels into numerical for easier classification by our models.
- Also Scaling was done to normalize the features; the train and test data so that they have a standardized range of values.

Model 1 - Logistic Regression

- Logistic regression model was fitted on the training dataset and this resulted in an accuracy score of 72% and an f1-score of 54%.

- Accuracy: 0.728978978978979
- F1-Score: 0.5492430154518495
- Recall-Score: 0.776478375992939 1
- Precision-Score: 0.4248973677855590 3

Model 2 - Decision Tree

- The decision tree was fitted to the dataset and the outcome was accuracy score of 99% and an f1-score of 99%. This model was also not doing well because it clearly indicated overfitting.
- Hyperparameter Tuning was carried out in order to improve the model performance by introducing hyperparameters and tuning them until the best combination was found and applied to the model. Grid search cv was used to carry this out on the decision tree model and the outcome was a model with 83% accuracy score. The model improved with hyperparameter tuning.

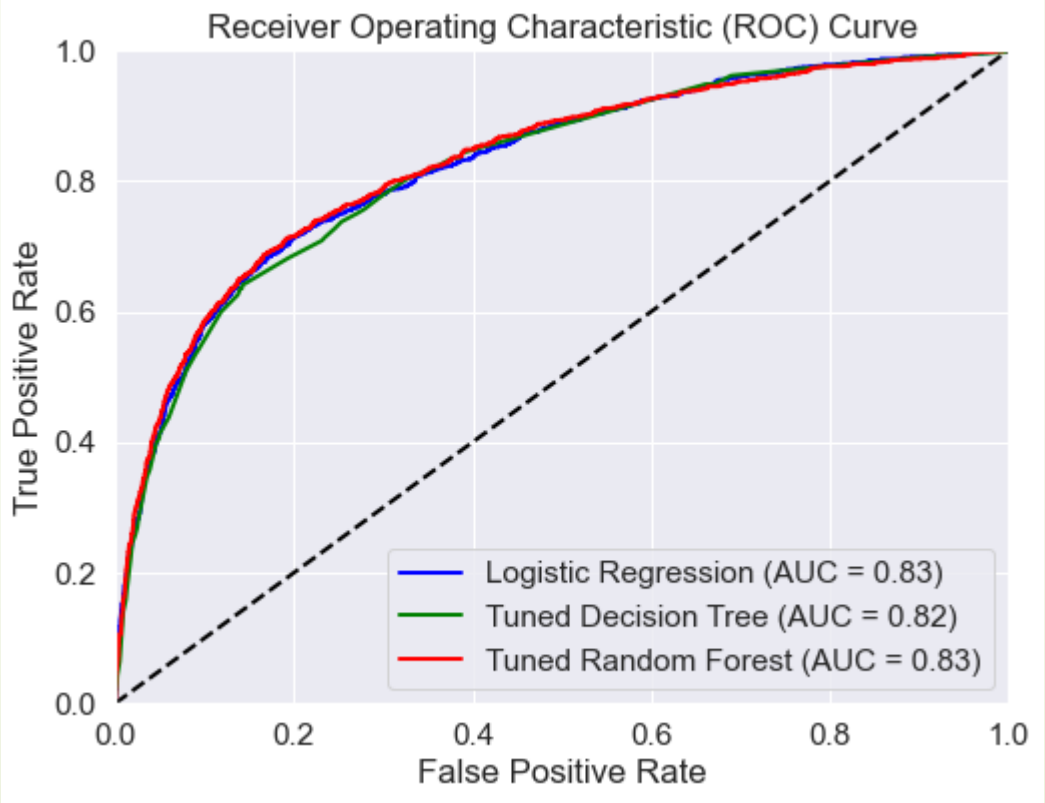
Model 3 – Random Forest

- The random forest was fitted to the dataset and the outcome was accuracy score of 82% .
- Hyperparameter Tuning was also carried out on this model in order to improve the model performance by introducing hyperparameters and tuning them until the best combination was found and applied to the model. Grid search cv was used to carry this out on the decision tree model and the outcome was a model with 84% accuracy score. The model improved with hyperparameter tuning.

5. MODEL EVALUATION

- Here, the model is evaluated using different metrics to see how well it performed.

ROC Curve for further analysis



ROC Curve for further analysis Cont….

- Both Tuned Random Forest Model and Logistic Regression model achieve an accuracy of 83%, which is slightly higher than the Decision Trees model (82%). The hyperparameter tuning has led to a modest improvement in overall accuracy.
- The Random Forest model however, is successfull in identifying a higher proportion of actual positives.
- The Tuned Random Forest Model exhibits fewer false positives but higher false negatives compared to other models. This suggests trade offs will need to be carried out.
- In summary, the Tuned Random Forest Model demonstrates a subtle enhancement in key metrics, providing a more refined and optimized solution for predicting H1N1 vaccination status. The adjustments made during hyperparameter tuning contribute to improved model performance, particularly in capturing positive instances.

Confusion Matrix analysis

Metric	Decision Tree	Random Forest	Logistic Regression
False Positives	713	89	1157
False Negatives	638	854	264
Actual Positives	3476	4100	3032
Actual Negatives	502	286	876

- The Random Forest model is successfull in identifying a higher proportion of actual positives.
- The Tuned Random Forest Model exhibits fewer false positives but higher false negatives compared to other models. This means, resource allocation and necessary interventions will be done by the relevant authorities.

CONCLUSION

- Our model(Random Forest) has a great performance of 83%. This means that, in a 100 test samples, 83 of them will be classified correctly. From the Data Exploratory Analysis we have the following findings:
- Most of those who received the vaccine were male.
- Most of the health workers received the vaccine.
- The older people(65 Years and above) received the vaccine in high numbers compared to the young people.
- The number of those who received the vaccination through recommendation by a doctor was also high. People with chronic medical conditions did not receive the vaccine.
- Factors like Education level, employment status and race rarely affected the vaccination rate.

RECOMMENDATIONS

- The public health sector should prioritize educational and awareness campaigns to debunk vaccine-related myths and misconceptions. This approach has the potential to boost vaccine acceptance rates.
- Vaccinate more older people than already vaccinated because they are willing to partake the vaccine.
- Doctors' recommendation helps really helped so the authority should engage more doctors in their educational awareness campaigns
- Avail more ways to ensure the vaccines reaches the poverty stricken society who had a very low uptake