­

Contents

[Introduction 2](#_Toc75876519)

[Data Source And Tool Selection 2](#_Toc75876520)

[Modelling Process 3](#_Toc75876521)

[Business Understanding 3](#_Toc75876522)

[Data Understanding 3](#_Toc75876523)

[Data Initial Review 4](#_Toc75876524)

[Data Preparation 6](#_Toc75876525)

[Identification of Null values across the dataset 6](#_Toc75876526)

[Age Bucket Attribute 7](#_Toc75876527)

[Categorical Attributes 7](#_Toc75876528)

[Removal of Categorical Attributes 7](#_Toc75876529)

[Modelling 7](#_Toc75876530)

[Model Evaluation 8](#_Toc75876531)

[Logistic Regression Model 8](#_Toc75876532)

[K-Nearest Neighbour Model 9](#_Toc75876533)

[Random Forest Model 10](#_Toc75876534)

[Model Comparison 11](#_Toc75876535)

[Further modelling with Logistic Regression 11](#_Toc75876536)

[Supporting Documents 13](#_Toc75876537)

[Project Code 13](#_Toc75876538)

[Datasource 13](#_Toc75876539)

[Description of dataset attributes 13](#_Toc75876540)

# Introduction

The world is currently faced with the gravest epidemic in living memory. Since early March 2020 normal social interactions have been drastically curtailed either by choice or through newly introduced legislations. As a result business have closed, and for some this will be permanent. To-date almost 4 million people in the world have died.

Science now offers us hope, vaccines have been developed and although the future of this epidemic is not fully known what is known is that the vaccine offers protection against the current mutations. The biggest simultaneous vaccine rollout in our history has begun. To-date 22.6% of the world population have received at least one dose of a Covid-19 vaccine.

However globally people are resisting the vaccine. I wondered if it was at all possible to predict the percentage of vaccine adaption within a population and if there is a pattern to the resistance.

Furthermore the logistical challenges involved in a global vaccine are immense. Thus, I believe that a successful model has potential application in vaccine purchases, in selection of vaccine centres and staffing of these centres. It could also have useful applications in identifying demographics where vaccine adoption is low, enable channelling of education and media towards these target these demographics.

With Covid-19 being so new I was not able to model on this particular epidemic, I was however able to find data relating to the N1H1 pandemic. (Data Source: <https://www.kaggle.com/c/prediction-of-h1n1-vaccination>)

The N1H1 pandemic occurred in 2009 and lasted 19 months. Possible infection rates was as high as 1.4 billion globally with an estimated death rate between 150,000 – 575,000.

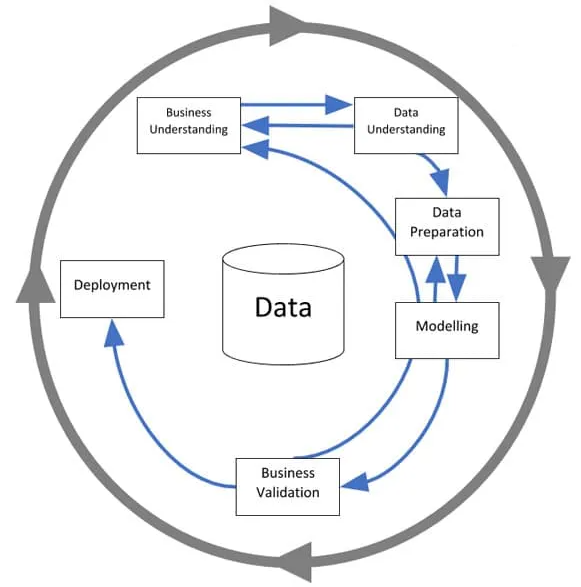
# Data Source And Tool Selection

As mentioned above this data source was obtained from <https://www.kaggle.com/c/prediction-of-h1n1-vaccination>.  
Initial review was done in MS Excel. This resulted in the following observations of the data.  
 - 38 columns  
 - 26707 Rows of data and 1 Row with column headers  
 - Data was not completely filled for all cells  
 - Data quality was high with cells populated cell values either being of type binary, integers or text  
 - Some cells were in obfuscated form

As the level of data quality was high I felt it was possible to achieve the project objective using Python and Jupyter Notebooks.

# Modelling Process

In order to build the model I will utilise the Cross Industry Standard Process for Data Mining (CRISP-DM). This is an iterative process consisting of 6 phases and is the most common data analytic model.



## Business Understanding

As stated above in the introduction the aim of this model is to attempt to predict the rate of N1H1 vaccine adaption. The obtained dataset consists of a 36 attribute column questionnaire divided into sections such as attitudes and behaviours, medical history and social demographics with a binary target label “h1n1\_vaccine”.

## Data Understanding

The data was loaded into Notebooks, stored in a data frame and the head inspected.   
The dataset consisted of 37 attribute columns and 1 identifier column. The data can be classified into 3 distinct groupings; **Behavioural; Medical; Social**

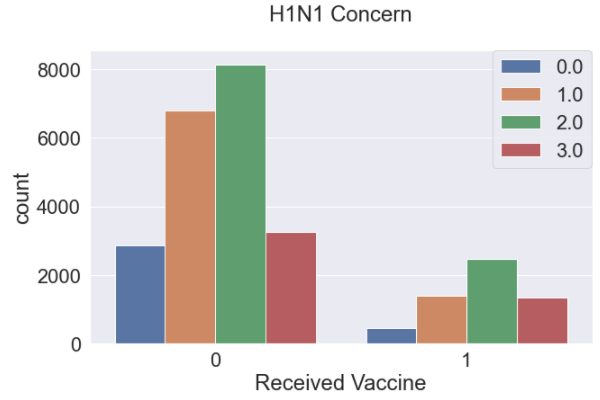
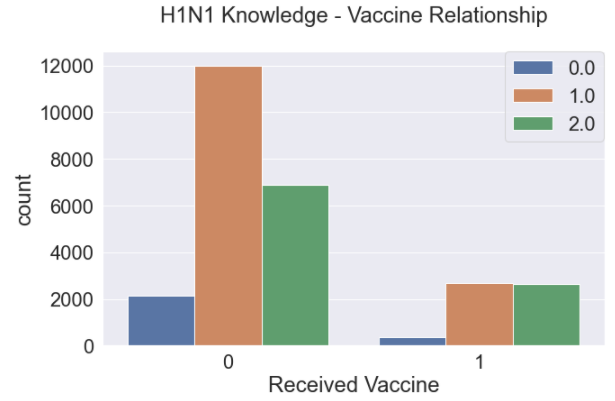
|  |  |  |
| --- | --- | --- |
| **Behavioural Group Attributes** | **Medical Group Attributes** | **Social Group Attributes** |
| h1n1\_concern  h1n1\_knowledge  behavioral\_antiviral\_meds  behavioral\_avoidance  behavioral\_face\_mask  behavioral\_wash\_hands  behavioral\_large\_gatherings  behavioral\_outside\_home  behavioral\_touch\_face  seasonal\_vaccine | child\_under\_6\_months  chronic\_med\_condition  doctor\_recc\_h1n1  doctor\_recc\_seasonal  h1n1\_concern  h1n1\_knowledge  health\_insurance  health\_worker  opinion\_h1n1\_risk  opinion\_h1n1\_sick\_from\_vacc  opinion\_h1n1\_vacc\_effective  opinion\_seas\_risk  opinion\_seas\_sick\_from\_vacc  opinion\_seas\_vacc\_effective  seasonal\_vaccine | child\_under\_6\_months  health\_insurance  age\_group  education  race  sex  income\_poverty  marital\_status  rent\_or\_own  employment\_status  hhs\_geo\_region  census\_msa  household\_adults  household\_children  employment\_industry  employment\_occupation |

A brief description of each of the data attributes can be viewed in the in the supporting documents section.

### Data Initial Review

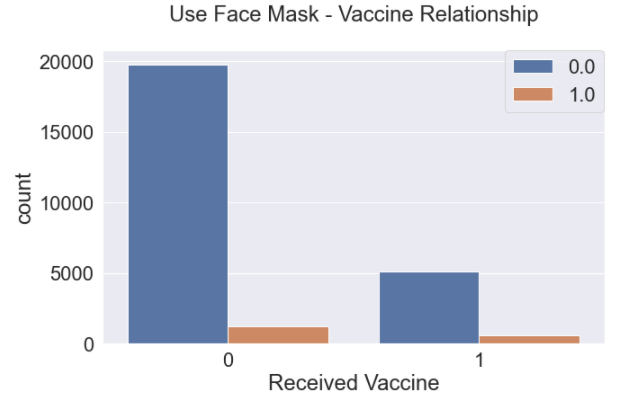
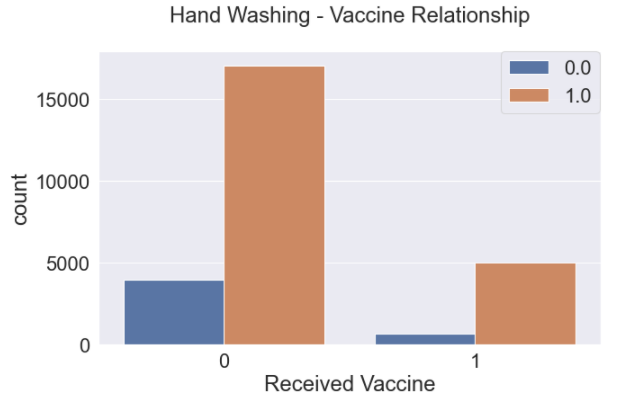
High level data inspection observed some key points about the data some of which would have to be addressed before modelling could commence. Please see the section; Data Preparation for more detailed information on this process.

Data with Integer Values  
A number of columns in the dataset were represented in categorical form using integers to represent a known category.



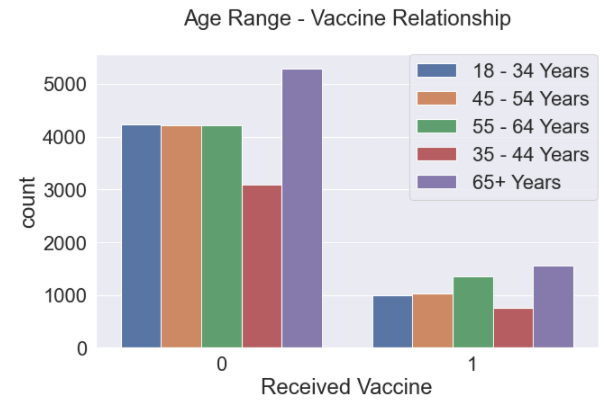
|  |  |
| --- | --- |
| **h1n1\_knowledge** (Level of knowledge about H1N1) | **h1n1\_concern**  (Level of knowledge about H1N1) |
| 0 = No Knowledge  1 = Not very concerned  2 = A Lot of Knowledge | 0 = Not at all concerned  1 = Not very concerned  2 = Somewhat concerned  3 = Very Concerned |

Data with Binary Values

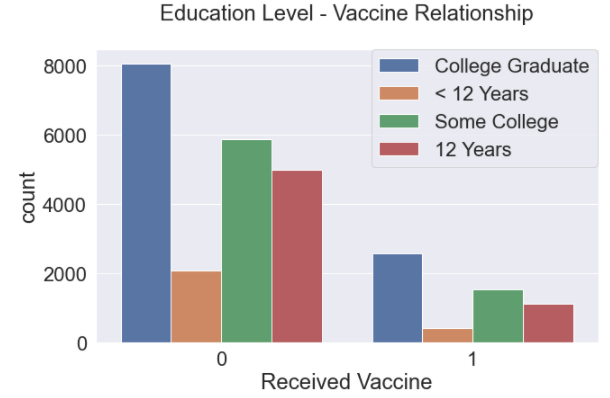
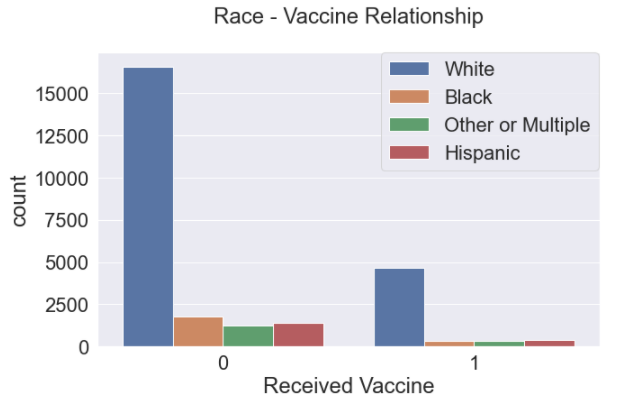
 

|  |  |
| --- | --- |
| **behavioral\_face\_mask** (Level of knowledge about H1N1) | **behavioral\_wash\_hands** (Level of knowledge about H1N1) |
| 0 = False  1=True | 0 = False  1=True |

Data with Bucket Values

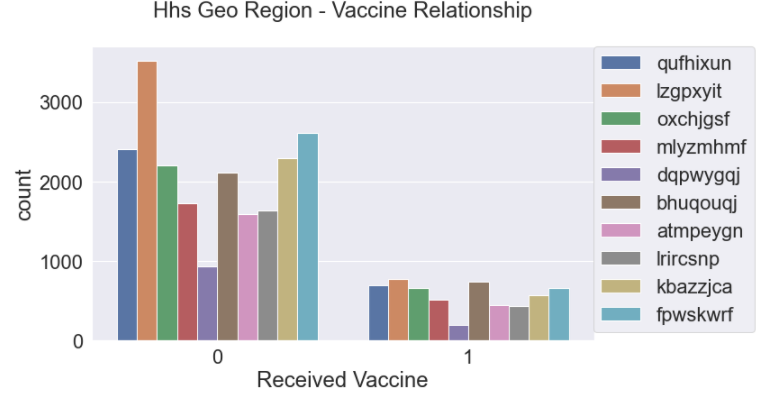


|  |
| --- |
| **Age\_range** |
| The model can not function with this bucket format. The average age of each bucket will be calculated and this value will be populated to a new column. |

|  |  |
| --- | --- |
| **Education Level** | **Race** |
| These bucket attributes will be turned binary by splitting the values across several new binary columns | |

Data with Obfuscated String Values



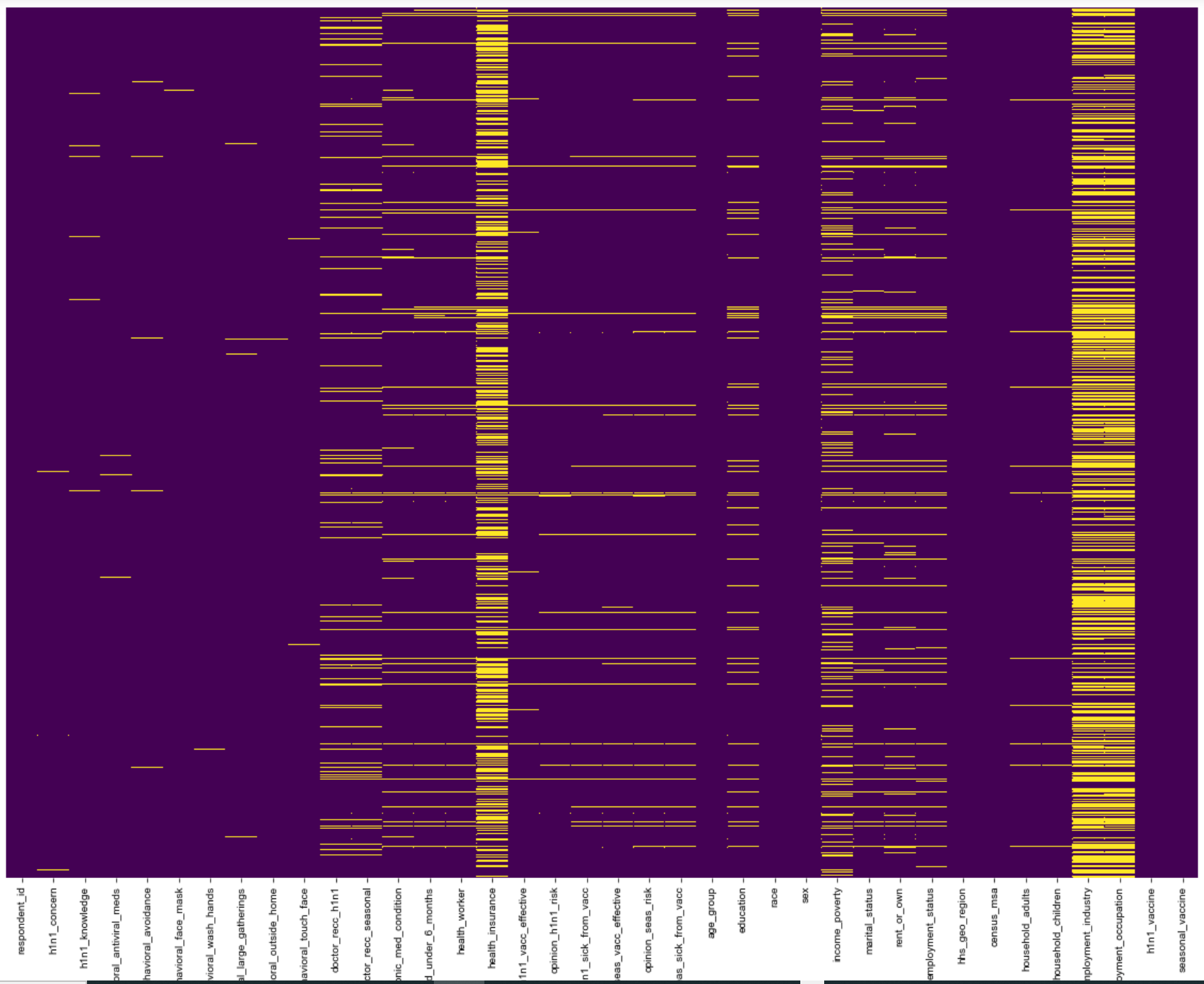
|  |
| --- |
| **Hss\_geo\_region** |
| These Obfuscated string values will remain and form new column names to hold binary values |

## Data Preparation

A number of steps were involved in getting the data into a state in which modelling result can successfully be obtained. These steps can be observed below.

### Identification of Null values across the dataset

Using Seaborn heatmap a high level overview across the dataset was obtained. From this it was obvious the very few columns were derived of Null values.   
*(see code labelled: # Analyse the dataset for Null Values by creating a heatmap)*



Before resolving the Null values on the attribute columns identified any of these attribute columns with a Null row count >= 50% of the entire data set were considered to have too many Null values and were dropped from the data set.  
  
This resulted in the attribute column ‘employment\_occupation’ removal from the dataset with a with a null record count of 13470 within the 26707 total rows.  
*(see code labelled: # Checking Columns for Number of Null Records and Dropping where it is >=50%)*

All remaining rows were filled using a frequency distribution method.

*(see code labelled: #For Each of Column with Nulls calculate the distribution freq to populate null values.)*

### Age Bucket Attribute

Observed within the dataset was the ‘Age’ attribute. Rather than store this value as an integer it was stored in bucket format with the values:

|  |
| --- |
| **Age Bucket Values** |
| 18 - 34 Years |
| 35 - 44 Years |
| 45 - 54 Years |
| 55 - 64 Years |
| 65+ Years |

These values were converted to an interger by calculating the midpoint of the the bucket value of each recipient, this calculated value was then stored in a newly created column ‘AverageAge’.   
Once done, the average age across the data can be inspected to be a value of 50.  
*(see code labelled: #Remove the Age Buckets)*

### Categorical Attributes

A number of categorical attributes still remained in the dataset.

|  |
| --- |
| **Categorical Attributes** |
| education |
| race |
| sex |
| income\_poverty |
| marital\_status |
| rent\_or\_own |
| employment\_status |
| hhs\_geo\_region |
| census\_msa |
| employment\_industry |

For simplicity of model selection and function with sklearn the categorical attributes were all converted to numeric attributes and stored in new columns using pandas.get\_dummies with drop\_first either applied as part of this method or applied retrospectively on a targeted column. Targeted columns were used to ensure the logical value in a sequence was dropped.  
*(see code labelled: #Convert Categoricials To Binary and related cells)*

### Removal of Categorical Attributes

Once all of the categorical attributes were converted to integer values all of the categorical attributes were removed from the dataset. The Age attribute utilising the ‘Age Buckets’ was also removed as part of this process.  
*(see code labelled: #Add the new binary columns and drop the original columns and the identifier column)*

## Modelling

With the data cleaned the Train-Test split was extracted from the data. 30% of the data was randomly selected for the test split with the remaining used for training the models.  
The X\_test.head was inspected after this process to ensure the target role label ‘h1n1\_vaccine’ was not included in the data set.*(see code labelled: #Split the dataset by X and y, with y to hold h1n1\_vaccine values)*

I focused on models supporting integer values as multiple independent values capable of determining a binary classification. The convention for binary classification is to have two classes 0 and 1 (matching the ‘h1n1\_vaccine’ label values in the dataset)  
I did not utilise any advanced selection criteria and picked the models randomly.  
With this in mind my chosen models were: Logistic Regression, K Nearest Neighbours and Random Forest.

### Model Evaluation

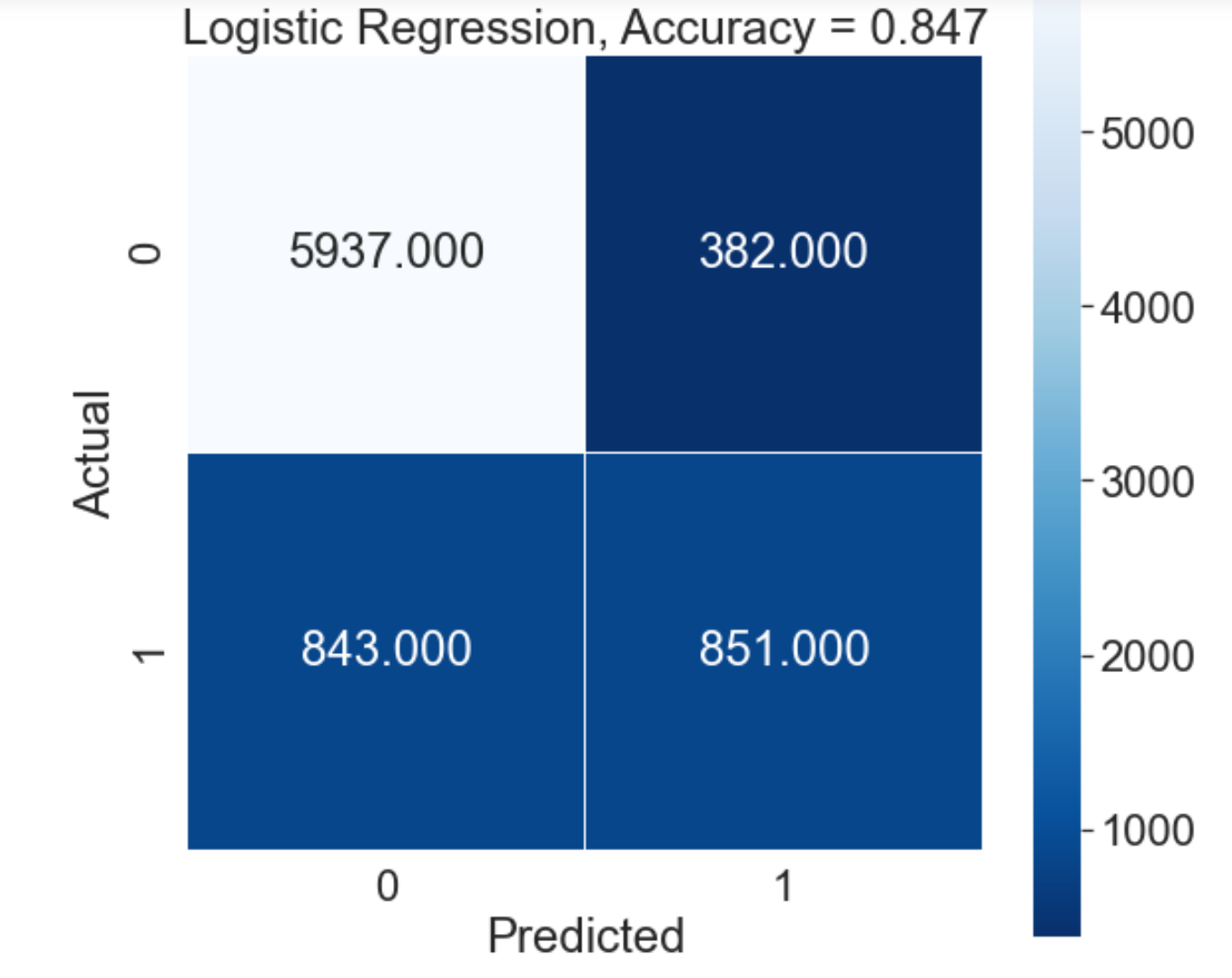
The confusion Matrix is a table that is used to describe the performance of a classification model.  
Confusion Matrix was used to evaluate each of the models. In addition to this the Accuracy of each model was obtained as well as a Classification Report.

All of these evaluations were also carried out in Python, although in addition I calculated the precision and recall for both 0 and 1 manually for each model.   
*(Each of the models build, run and analysis is clearly marked in the code)*

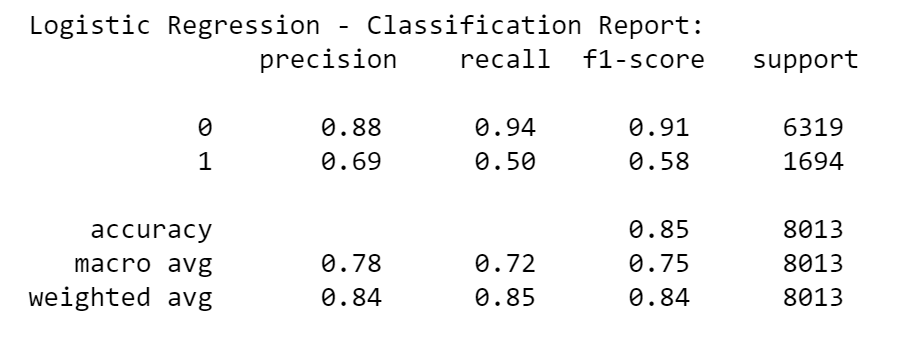
### Logistic Regression Model

The Logistic Regression Model allows solving of classification problems when trying to predict binary classifications.

This model produced the following Confusion Matrix and Accuracy Score



Classification Report



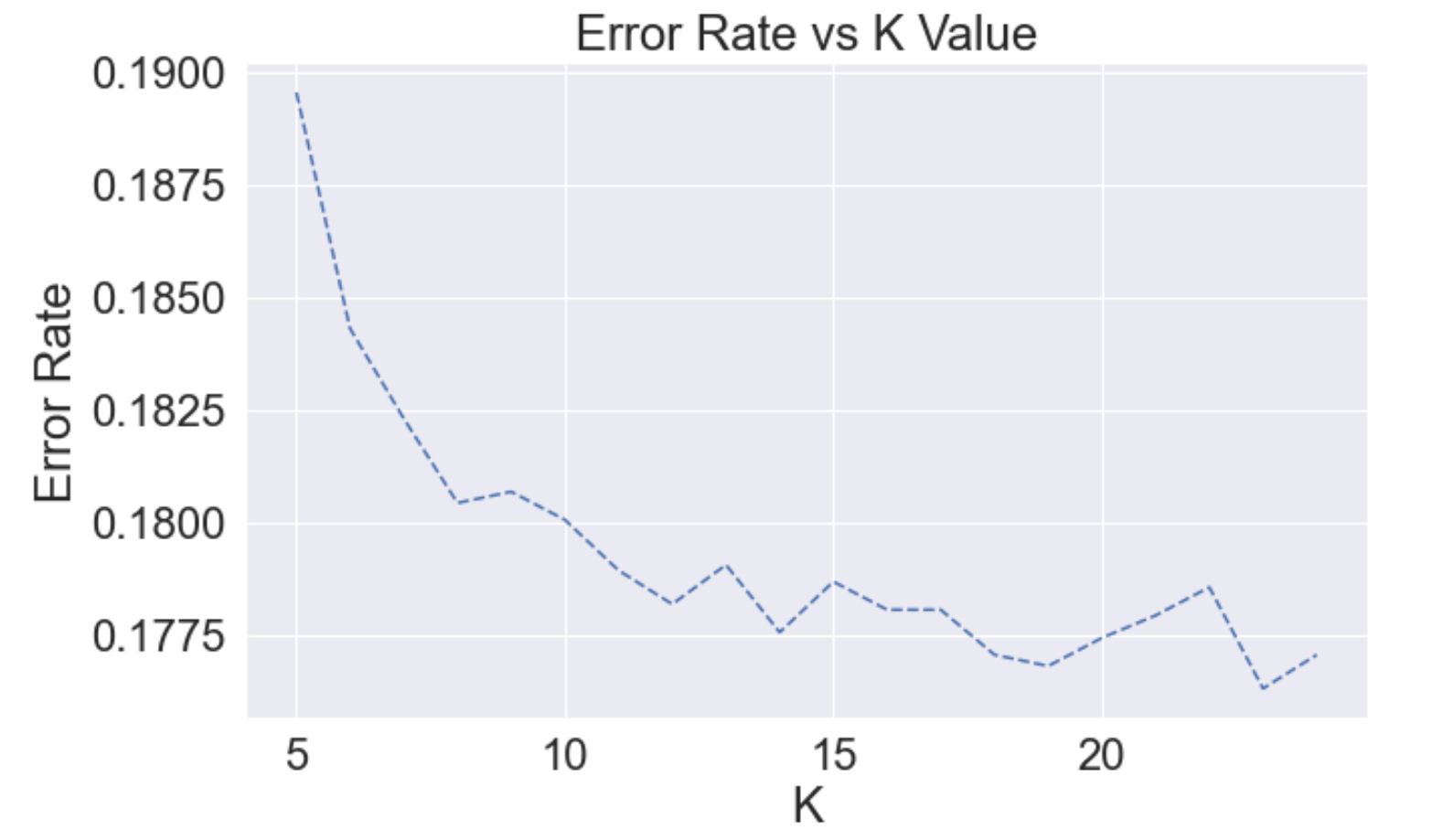
|  |  |  |  |
| --- | --- | --- | --- |
| True Positives | 5937 | False Positives | 843 |
| True Negatives | 843 | False Negatives | 382 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Precision 0 | TP/(TP+FP) | 0.88 | Precision 1 | TN/(TN+FP\_2) | 0.69 |
| Recall 0 | TP/(TP+FN) | 0.94 | Recall 1 | TN/(FP\_1+TN) | 0.50 |

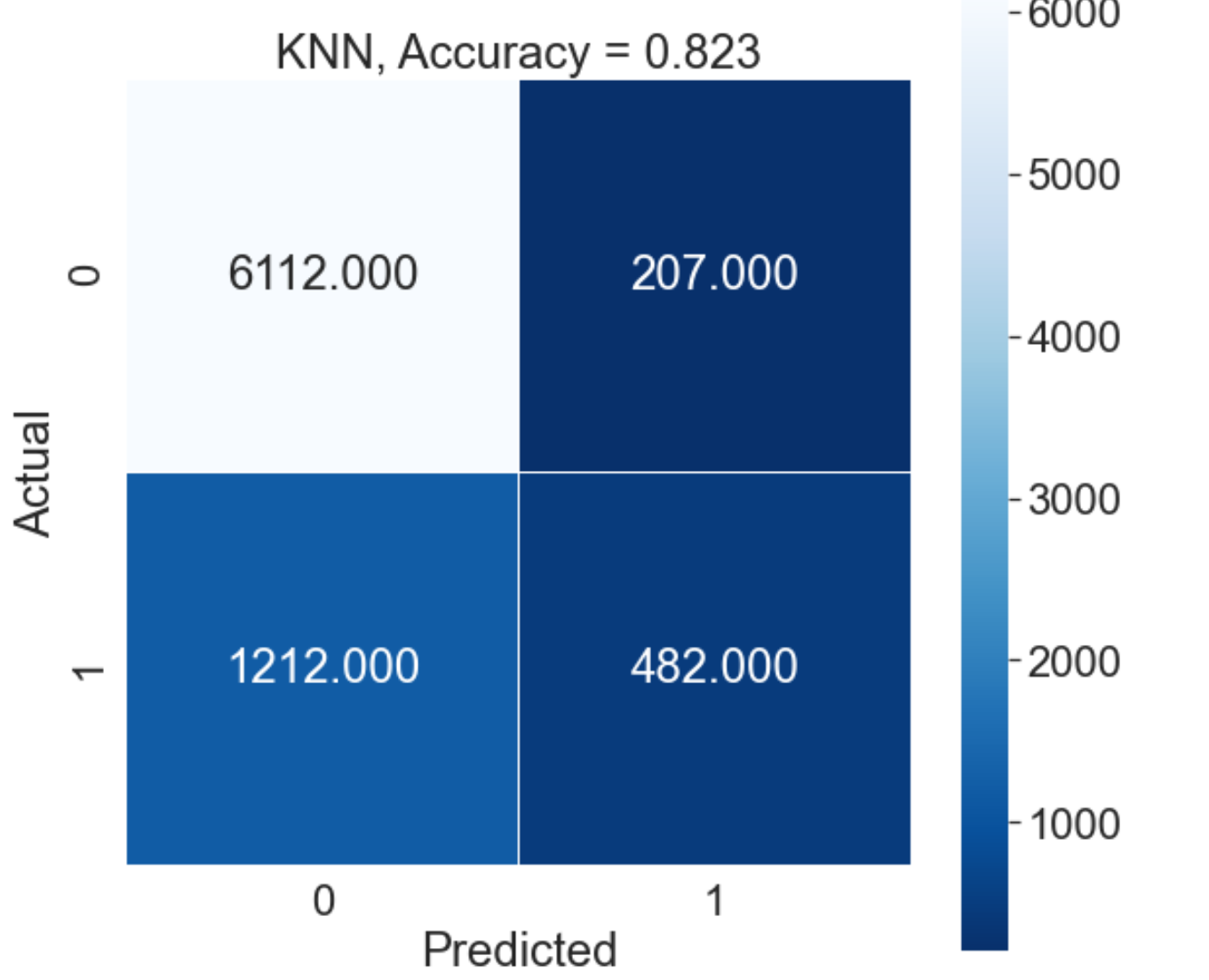
### K-Nearest Neighbour Model

The K-Nearest Neighbour Model is used to solve both classification and regression problems. This model is not limited to binary classifications.

Before this model was run (execution controlled via flag) analysis was performed in order to discover the n\_neighbours value with the lowest incorrect score. The model was run for multiple n\_neighbours values in the range of 5-25. The mean error rate was recorded for each iteration.

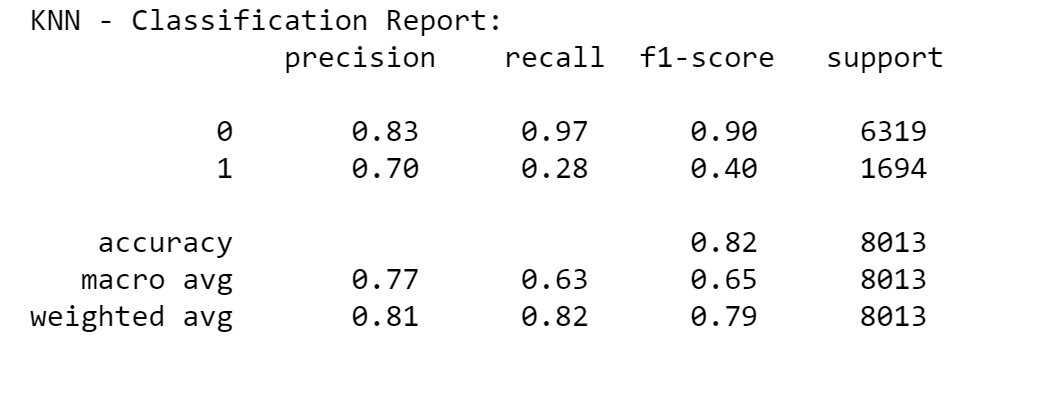


24 was identified as the most successful neighbours value and this the value used for the model. Below is the Confusion Matrix and Accuracy Score for this model.



Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives | 6112 | False Positives | 1212 |
| True Negatives | 482 | False Negatives | 207 |

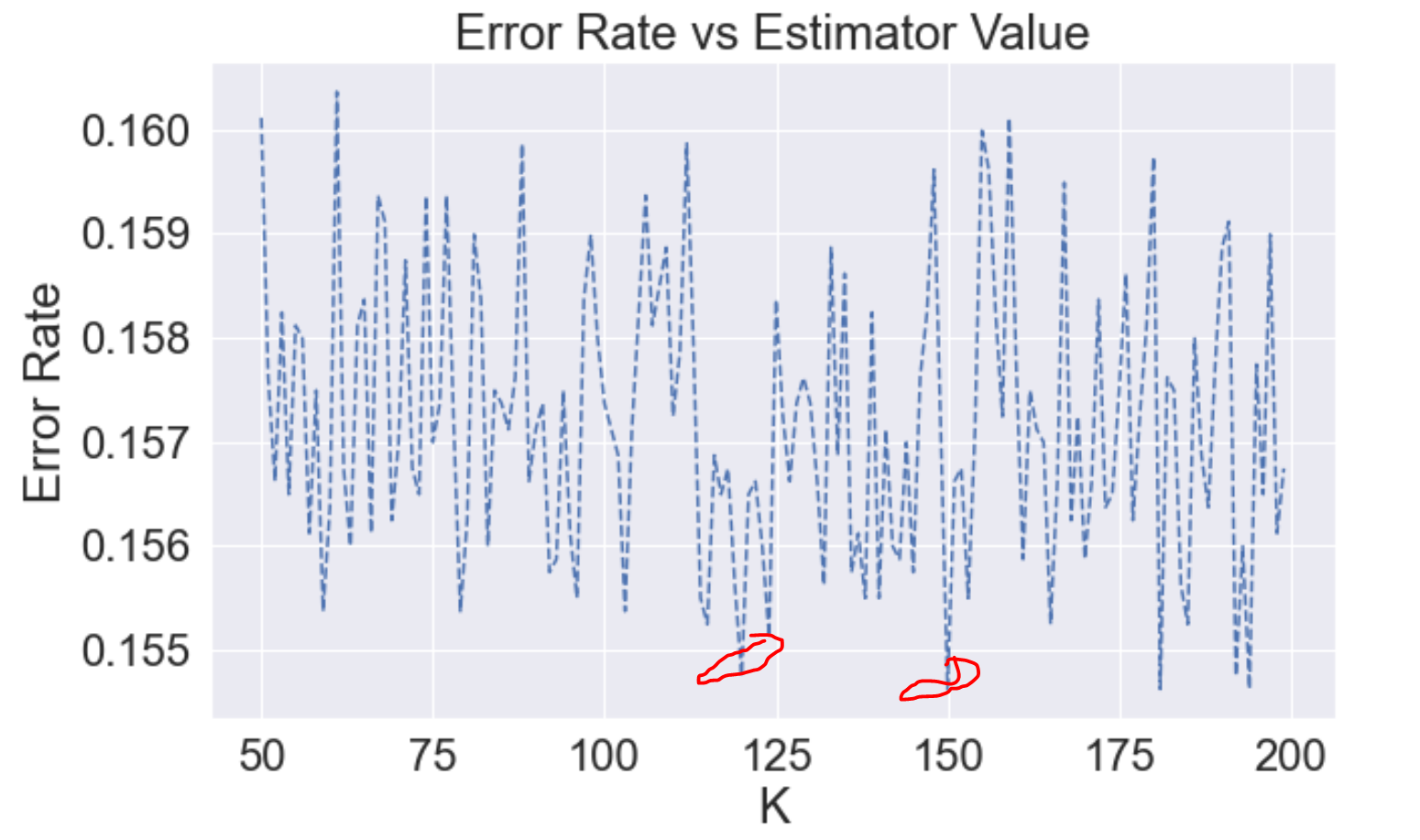


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Precision 0 | TP/(TP+FP) | 0.83 | Precision 1 | TN/(TN+FP\_2) | 0.70 |
| Recall 0 | TP/(TP+FN) | 0.97 | Recall 1 | TN/(FP\_1+TN) | 0.28 |

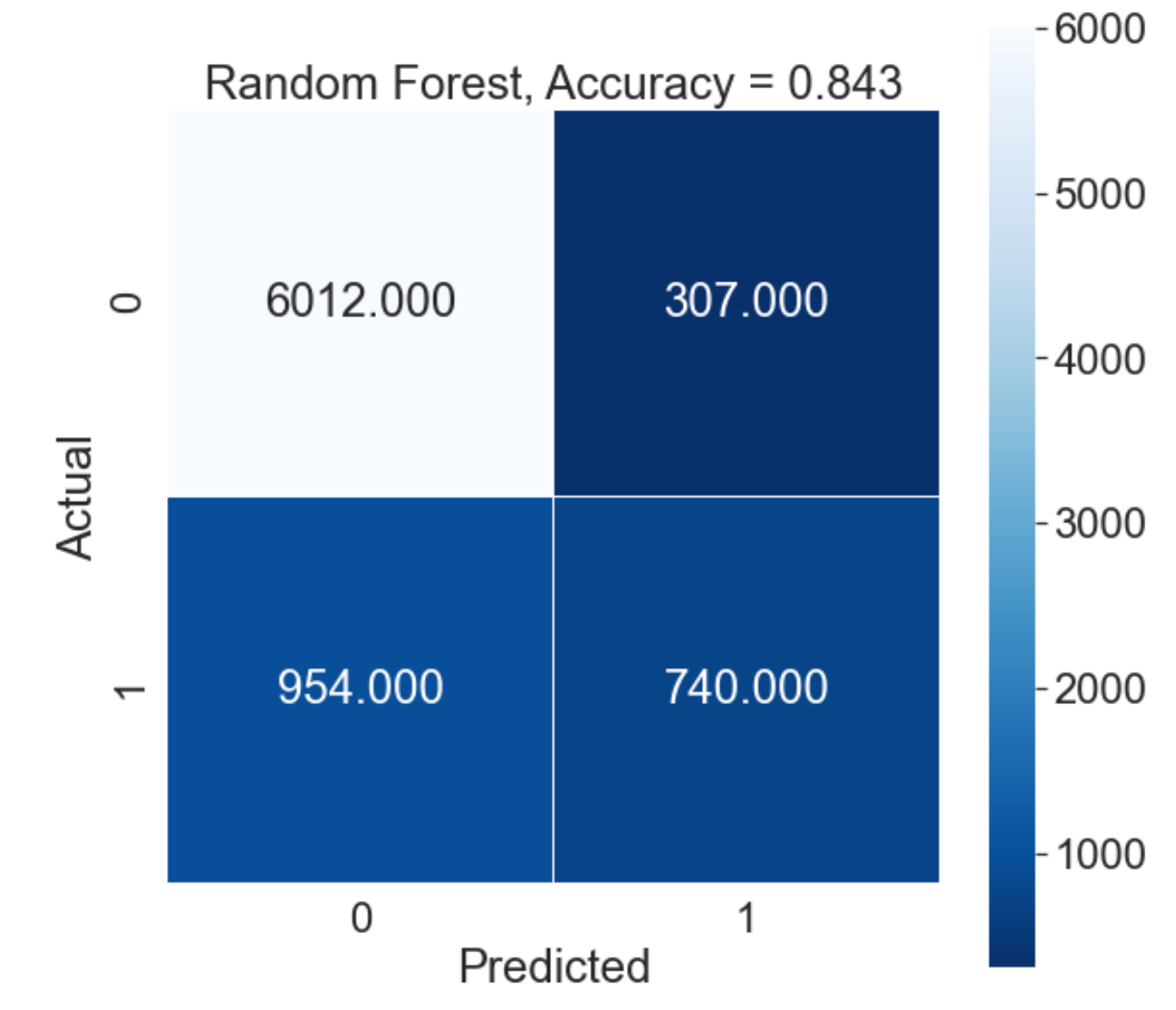
### Random Forest Model

The Random Forest Model is used to solve both classification and regression problems. This model is not limited to binary classifications.

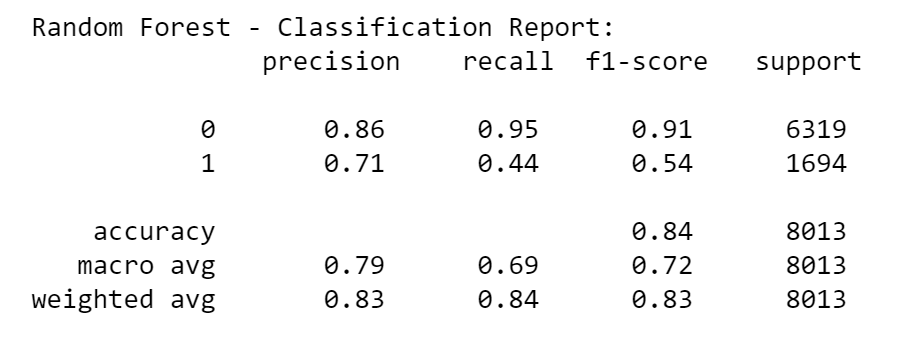
Before this model was run analysis was performed in order to discover the n\_estimators value with the lowest incorrect score. The model was run for multiple n\_ estimators values in the range of 50-200. The mean error rate was recorded for each iteration.



Values were returned with a large deviation from other values close to their range. These were identified as possible outliers and were discarded from selection. A value of 95 was selected for the model. Below is the Confusion Matrix and Accuracy Score for this model.



Classification Report



|  |  |  |  |
| --- | --- | --- | --- |
| True Positives | 6012 | False Positives | 954 |
| True Negatives | 740 | False Negatives | 307 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Precision 0 | TP/(TP+FP) | 0.86 | Precision 1 | TN/(TN+FP\_2) | 0.71 |
| Recall 0 | TP/(TP+FN) | 0.95 | Recall 1 | TN/(FP\_1+TN) | 0.44 |

## Model Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 |
| Logistic Regression | .847 | .88 | .69 | .94 | .50 |
| K-Nearest Neighbour | .823 | .83 | .70 | .97 | . |
| Random Forest | .843 | .86 | .71 | .95 | .44 |

K-Nearest Neighbour performed the poorest out of the three models tested when predicting vaccine uptake. It failed to identify a high number of recipients who would take the vaccine. This high error rate can be seen in the high number of False Positive Type 1 Errors and poor Recall 1 score.

Overall:  
Random Forest and Logistic are almost equally performant at identifying vaccine adoption figures.  
The Logistic regression returns a better vaccine adoption number; correctly predicting 851 out of a total number of 1,694 records, therefore this could be considered a better model for predicting vaccine purchases. When identifying demographics adopting a vaccine however it performed poorly with a recall for 1 of 50%, therefore there is not much reliable information to be extracted from the model from this objective.  
  
All models perform almost equally well at predicting the number of people who would not get a vaccine. KNN was the best predictor here having a higher recall score for 0 at 97%

## Further modelling with Logistic Regression

In an attempt to explore the dataset a little more 3 additional datasets were selected. As mentioned at the start of this document these were the 3 logical data divisions were observed.   
The Behavioural Group was possibly highly correlated but it resulted in a poor model result so the results were simply discounted.  
The Social Group resulted in a really poor predictor for vaccine adoption only predicting 10 records.

|  |  |  |
| --- | --- | --- |
| **Behavioural Group Attributes** | **Medical Group Attributes** | **Social Group Attributes** |
| h1n1\_concern  h1n1\_knowledge  behavioral\_antiviral\_meds  behavioral\_avoidance  behavioral\_face\_mask  behavioral\_wash\_hands  behavioral\_large\_gatherings  behavioral\_outside\_home  behavioral\_touch\_face  seasonal\_vaccine | child\_under\_6\_months  chronic\_med\_condition  doctor\_recc\_h1n1  doctor\_recc\_seasonal  h1n1\_concern  h1n1\_knowledge  health\_insurance  health\_worker  opinion\_h1n1\_risk  opinion\_h1n1\_sick\_from\_vacc  opinion\_h1n1\_vacc\_effective  opinion\_seas\_risk  opinion\_seas\_sick\_from\_vacc  opinion\_seas\_vacc\_effective  seasonal\_vaccine | child\_under\_6\_months  health\_insurance  age\_group  education  race  sex  income\_poverty  marital\_status  rent\_or\_own  employment\_status  hhs\_geo\_region  census\_msa  household\_adults  household\_children  employment\_industry  employment\_occupation |

The fourth dataset was an experiment in correlations.  
I identified the fields with the highest correlation to the target label and stored these values in a new Dataframe.

|  |
| --- |
| 'seasonal\_vaccine','doctor\_recc\_h1n1','opinion\_h1n1\_risk','opinion\_h1n1\_vacc\_effective','opinion\_seas\_risk','doctor\_recc\_seasonal','opinion\_seas\_vacc\_effective','health\_worker','h1n1\_concern','h1n1\_knowledge' |

Then for each column in the dataframe I inspected the correlation to the other column members. I did not see any highly correlated columns and ran the linear regression model.

The results of the model on these datasets are as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision 0 | Precision 1 | Recall 0 | Recall 1 |
| Logistic Regression | .847 | .88 | .69 | .94 | .50 |
| Behavioural Group | .789 | .80 | .50 | .98 | .06 |
| Medical Group | .848 | .88 | .70 | .94 | .50 |
| Social Group | .788 | .79 | .45 | 1 | .01 |
| Soft Correlations | .837 | .86 | .68 | .94 | .44 |

We can see from the above table modelling results from the Medical Group Dataset resulted in improved accuracy compared to the Logistic Regression result on the full dataset.

# Supporting Documents

## Project Code

https://github.com/Hammerman-69/n1h1/blob/main/h1n1Project.ipynb

## Datasource

<https://www.kaggle.com/c/prediction-of-h1n1-vaccination>

Data is provided courtesy of the United States National Center for Health Statistics.

U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. The National 2009 H1N1 Flu Survey. Hyattsville, MD: Centers for Disease Control and Prevention, 2012.

## Description of dataset attributes

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Attribute Values** |
| h1n1\_concern | Level of concern about H1N1 flu | 0 - Not at all concerned 1 - Not very concerned  2 - Somewhat concerned 3 - Very Concerned |
| h1n1\_knowledge | Level of knowledge about H1N1 | 0 - No Knowledge  1 - Not very concerned  2 - A Lot of Knowledge |
| behavioral\_antiviral\_meds | Has taken antiviral medications | 0 - No 1 - Yes |
| behavioral\_avoidance | Has avoided close contact with others with flu-like symptoms | 0 - No 1 - Yes |
| behavioral\_face\_mask | Has bought a face mask | 0 - No 1 - Yes |
| behavioral\_wash\_hands | Has frequently washed hands or used hand sanitizer | 0 - No 1 - Yes |
| behavioral\_large\_gatherings | Has reduced time at large gatherings | 0 - No 1 – Yes |
| behavioral\_outside\_home | Has reduced contact with people outside of own household | 0 - No 1 – Yes |
| behavioral\_touch\_face | Has avoided touching eyes, nose, or mouth | 0 - No 1 – Yes |
| doctor\_recc\_h1n1 | H1N1 flu vaccine was recommended by doctor | 0 - No 1 – Yes |
| doctor\_recc\_seasonal | Seasonal flu vaccine was recommended by doctor | 0 - No 1 – Yes |
| 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 * neurological or neuromuscular condition * liver condition * weakened immune system caused by a chronic illness or by medicines taken for a chronic illness. | 0 - No 1 – Yes |
| child\_under\_6\_months | Has regular close contact with a child under the age of six months | 0 - No 1 – Yes |
| health\_worker | Is a healthcare worker | 0 - No 1 – Yes |
| health\_insurance | Has health insurance | 0 - No 1 – Yes |
| opinion\_h1n1\_vacc\_effective | 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 | 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 | 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 | 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 | 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 | 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 | Bucket Values of:  18 - 34 Years 45 - 54 Years,  55 - 64 Years 35 - 44 Years 65+ Years |
| education | Self-reported education level | Values are represented as short category character strings   * < 12 Years * 12 Years * Some College * College Graduate |
| race | Race of respondent | Values are represented as short category character strings   * White * Black * Hispanic * Other or Multiple |
| sex | Sex of respondent | Values are represented as short category character strings   * Male * Female |
| income\_poverty | Household annual income of respondent with respect to 2008 Census poverty thresholds | Values are represented as short category character strings   * Below Poverty * <=$75,000, Above Poverty * > $75,000 |
| marital\_status | Marital status of respondent | Values are represented as short category character strings   * Not Married * Married |
| rent\_or\_own | Housing situation of respondent | Values are represented as short category character strings   * Own * Rent |
| employment\_status | Employment status of respondent | Values are represented as short category character strings   * Not in Labor Force * Employed * Unemployed |
| hhs\_geo\_region | 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 | Residence within metropolitan statistical areas (MSA) as defined by the U.S. Census | Values are represented as short category character strings   * Non-MSA * MSA, Not Principle City * MSA, Principle City |
| household\_adults | Number of other adults in household | Integer 0-3 |
| household\_children | Number of other children in household, top-coded to 3 | Integer 0-3 |
| 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 |
| seasonal\_vaccine | Whether respondent received seasonal flu vaccine | 0 - No 1 – Yes |