HEALTH CARE ANALYTICS

By

Tejveer Singh (MSC IIT Dhanbad)

linkedin.com/in/tejveer-singh-22570418a

github.com/Tej752

ABSTRACT

An insurance policy/plan is an contact between an individual (Policyholder) and an insurance company (Provider). Under the contract, you pay regular amounts of money (as premiums) to the insurer, and they pay you if the sum assured on unfortunate event arises, for example, untimely demise of the life insured, an accident, or damage to a house. Health insurance is a type of insurance that covers medical expenses that arise due to an illness. These expenses could be related to hospitalisation costs, cost of medicines or doctor consultation fees. Anybody who has a health insurance policy will tell you that buying one is one of the smartest financial decisions by any earning individual. Now, that you have decided to buy a health insurance policy, you need to know how to select a good Health Insurance plans that will take care of all your needs. Here is a list of benefits any good health insurance plan should offer you:

- Protection against a large number of critical illnesses
- Flexibility to choose your health cover
- No increase in premiums during the policy term even if your health condition changes
- Long policy term that covers you even in your old age
- Large hospital network for easy access to medical treatment

INTRODUCTION

Med Camp organizes health camps in several cities with low work-life balance. They reach out to working people and ask them to register for these health camps. For those who attend, Med Camp provides them the facility to undergo health checks or increase awareness by visiting various stalls (depending on the format of the camp).

Med Camp has conducted 65 such events over a period of 4 years and they see a high drop off between “Registration” and the Number of people taking tests at the Camps. In the last 4 years, they have stored data of ~110,000 registrations they have done.

One of the huge costs in arranging these camps is the amount of inventory you need to carry. If you carry more than the required inventory, you incur unnecessarily high costs. On the

other hand, if you carry less than the required inventory for conducting these medical checks, people end up having Bab experience.

DATASET-DESCRIPTION

Train.zip contains the following csv alongside the data dictionary that contains definitions for each variable

1. Health Camp Detail.csv – File containing Health Id, Camp Start Date, Camp End Date and Category details of each camp.
2. Train.csv – File containing registration details for all the test camps. This includes Patient ID, Health Camp ID, Registration Date and a few anonymized variables as on registration date.
3. Patient Profile.csv- This file contains Patients' Profile details like ID, Online Followers, Social Media Details, Income, Age, Education, First Interaction Date, City Type and Employer Category
4. First Health Camp Attended.csv- This file contains details about the person who attended health camp of first format. This includes Donations and Health Score of the person.
5. Second health Camp Attended.csv- This file contains detail about people who attended health camp of second format. That includes Health score of the person.
6. Third Health Camp Attended.csv- This file contains detail about people who attended health camp of third format. That includes number of stalls visited.

Test.csv – File containing registration details for all the test camps. This includes Patient ID, Health Camp ID, Registration Date and a few anonymized variables as on registration date.

Camps started on or before 31st March 2006 are considered in Train
Test data is for all camps conducted on or after 1st April 2006.

METHODOLOGY

The project is divided into 3 stages:

1. Visualization -> The data from all the files is properly visualized and the necessary points from the data are taken care of.

2. Merging -> All the files are merged for predictions

3. Prediction -> Predictions are made using Machine Learning techniques.

MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The Ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect. The term 'machine learning' is often, incorrectly, interchanged with Artificial Intelligence, but machine learning is actually a sub field/type of AI. Machine learning is also often referred to as predictive analytics, or predictive modelling. Coined by American computer scientist Arthur Samuel in 1959, the term 'machine learning' is defined as a "computer's ability to learn without being explicitly programmed".

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It's a science that's not new – but one that has gained fresh momentum.

While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development. Here are a few widely publicized examples of machine learning applications you may be familiar with:

- The heavily hyped, self-driving Google car? The essence of machine learning.

- Online recommendation offers such as those from Amazon and Netflix? Machine learning applications for everyday life.
- Knowing what customers are saying about you on Twitter? Machine learning combined with linguistic rule creation.
- Fraud detection? One of the more obvious, important uses in our world today.

The term ‘machine learning’ is often, incorrectly, interchanged with Artificial Intelligence, but machine learning is actually a subfield/type of AI. Machine learning is also often referred to as predictive analytics, or predictive modelling. Coined by American computer scientist Arthur Samuel in 1959, the term ‘machine learning’ is defined as a “computer’s ability to learn without being explicitly programmed”.

At its most basic, machine learning uses programmed algorithms that receive and analyse input data to predict output values within an acceptable range. As new data is fed to these algorithms, they learn and optimise their operations to improve performance, developing ‘intelligence’ over time.

There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement.

DIFFERENT LEARNING METHODS

1.SUPERVISED LEARNING

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

1. Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to
   what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.
2. Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.
3. Forecasting: Forecasting is the process of making predictions about the future based on the past and present data, and is commonly used to analyse trends.

2.SEMI-SUPERVISED LEARNING

Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this
combination, machine learning algorithms can learn to label unlabelled data.

3.UNSUPERVISED LEARNING

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analysing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organise that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organised.

As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined.

Under the umbrella of unsupervised learning, fall:

1. Clustering: Clustering involves grouping sets of similar data (based on defined criteria). It's useful for segmenting data into several groups and performing analysis on each data set to find patterns.
2. Dimension Reduction: Dimension reduction reduces the number of variables being considered to find the exact information required.

4.REINFORCEMENT LEARNING

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

SOME ALGORTHIMS THAT ARE USED

Choosing the right machine learning algorithm depends on several factors, including, but not limited to: data size, quality and diversity, as well as what answers businesses want to derive from that data. Additional considerations include accuracy, training time, parameters, data points and much more. Therefore, choosing the right algorithm is both a combination of business need, specification, experimentation and time available.

1. Logistic Regression (Supervised learning – Classification)

Logistic regression focuses on estimating the probability of an event occurring based on the previous data provided. It is used to cover a binary dependent variable, that is where only two values, 0 and 1, represent outcomes.

2. Artificial Neural Networks (Reinforcement Learning)

An artificial neural network (ANN) comprises 'units' arranged in a series of layers, each of which connects to layers on either side. ANNs are inspired by biological systems, such as the

brain, and how they process information. ANNs are essentially a large number of interconnected processing elements, working in unison to solve specific problems.

ANNs also learn by example and through experience, and they are extremely useful for modelling non-linear relationships in high-dimensional data or where the relationship amongst the input variables is difficult to understand.

3. Random Forests (Supervised Learning – Classification/Regression)

Random forests or ‘random decision forests’ is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a ‘decision tree’ (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.

4. K Means Clustering Algorithm (Unsupervised Learning - Clustering)

The K Means Clustering algorithm is a type of unsupervised learning, which is used to categorise unlabelled data, i.e. data without defined categories or groups. The algorithm works by finding groups within the data, with the number of groups represented by the variable K. It then works iteratively to assign each data point to one of K groups based on the features provided.

RESULTS AND PLOTS

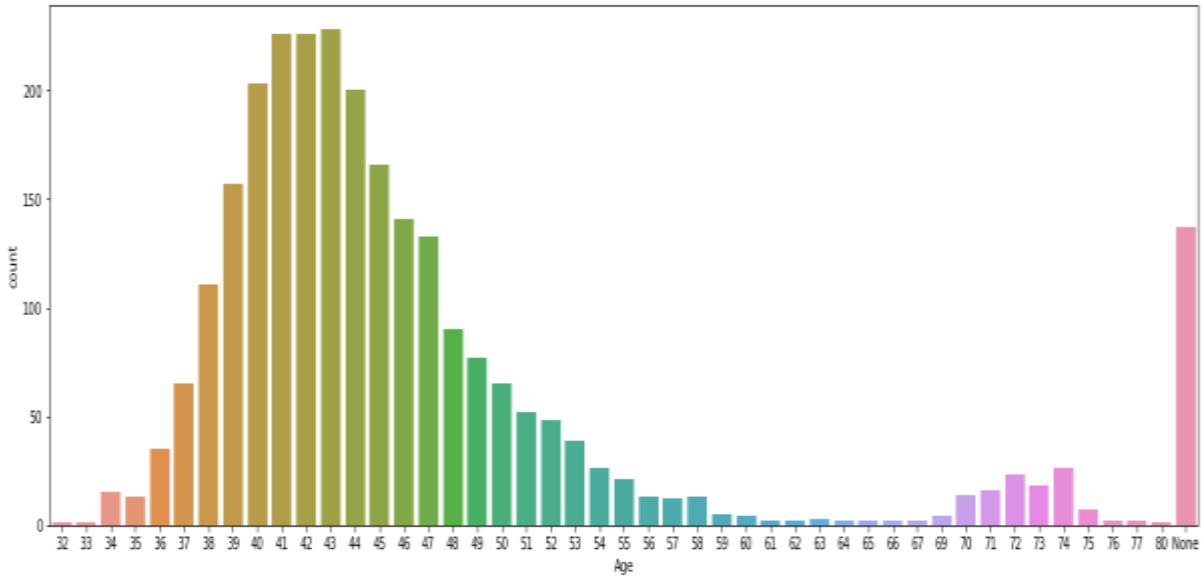

Figure 1 Age wise attendance of people in the camp

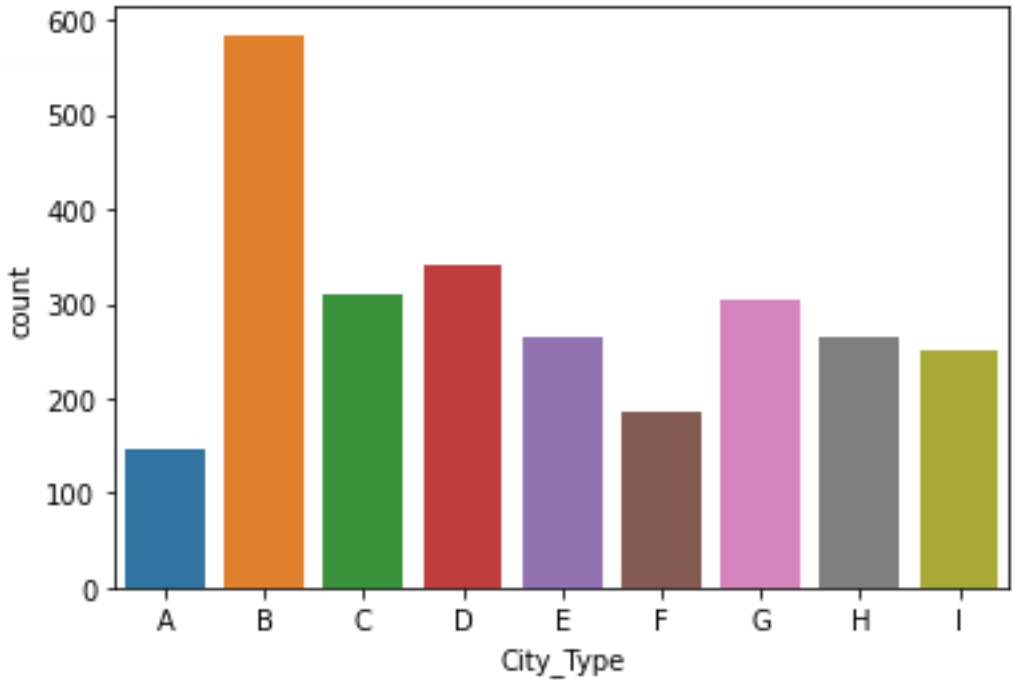

Figure 2 City Wise attendance

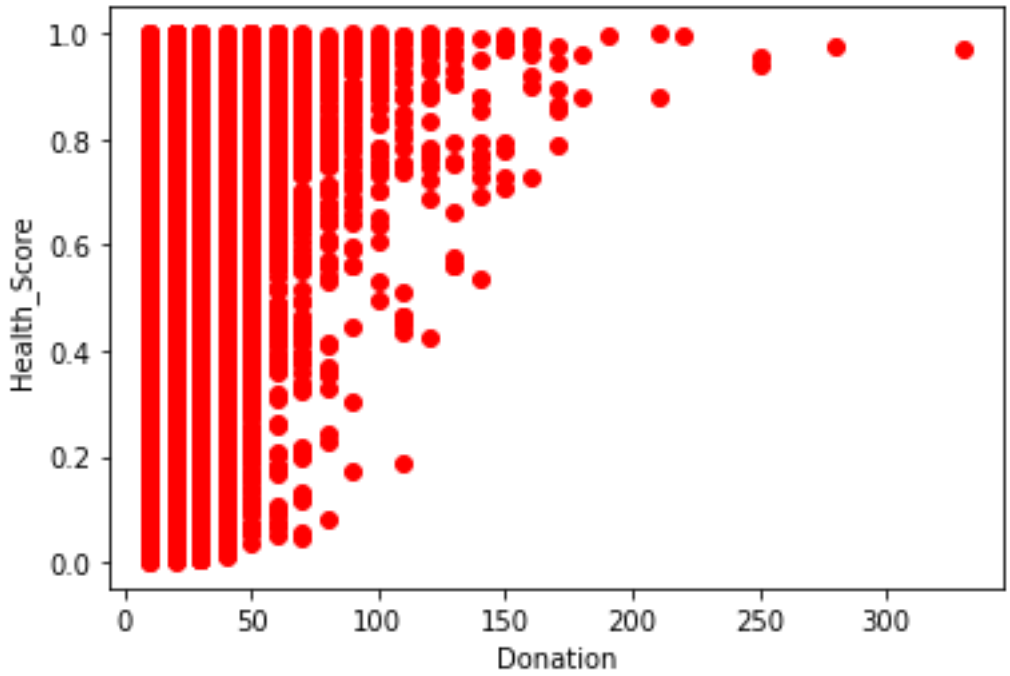

Figure 3 Donation vs Heath Score Distribution

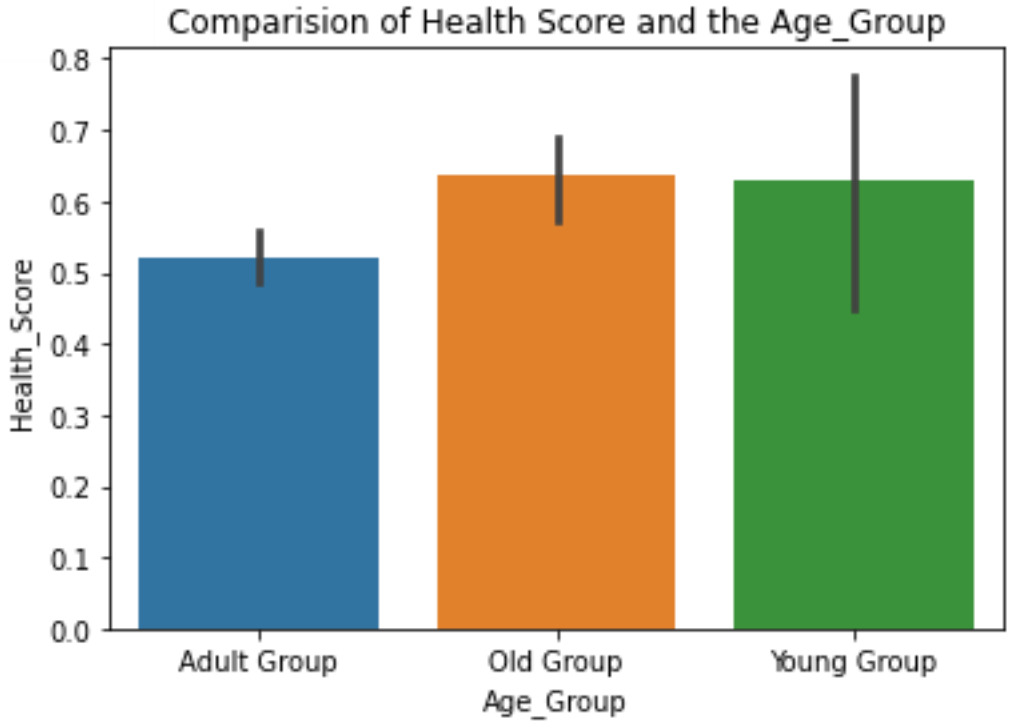

Figure 4 Age Groups vs Health Score

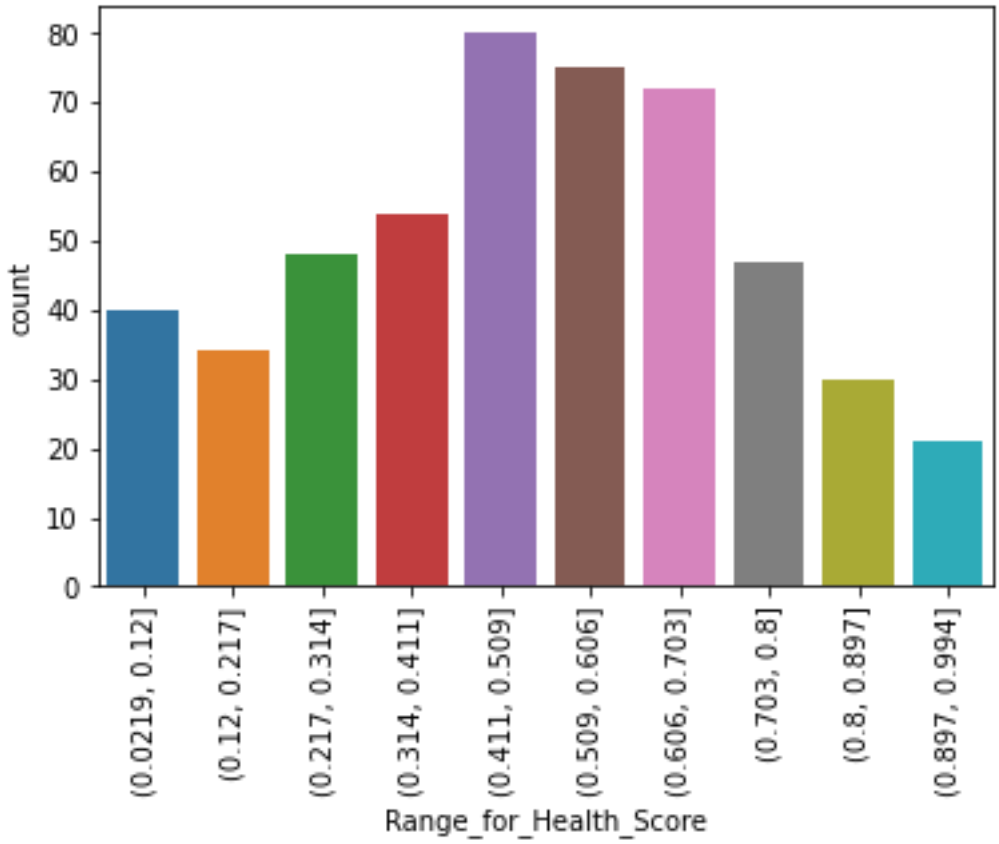

Figure 5 Count of different ranges as per Health Score

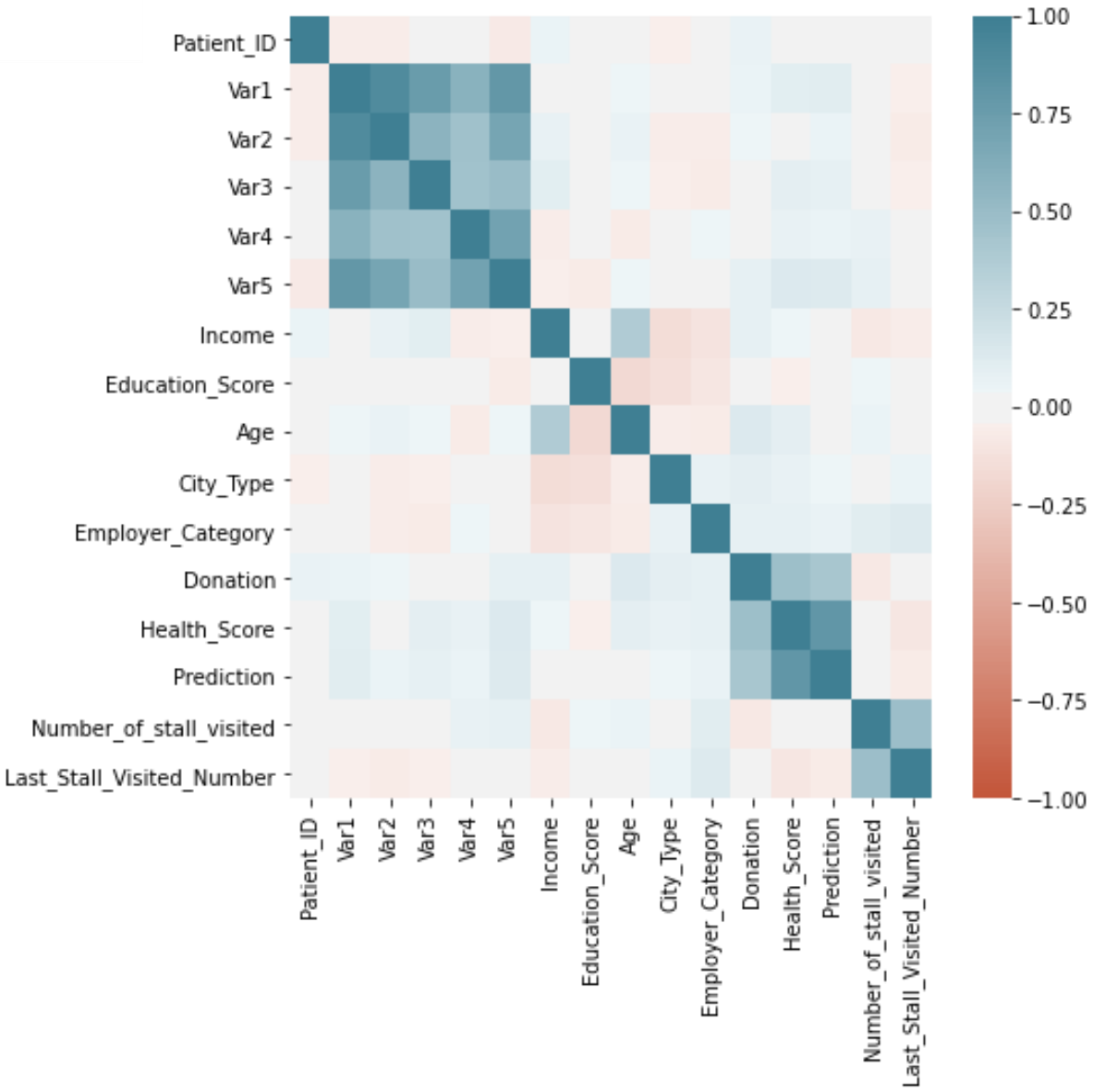

Figure 6 Correlation Map b/w all variables

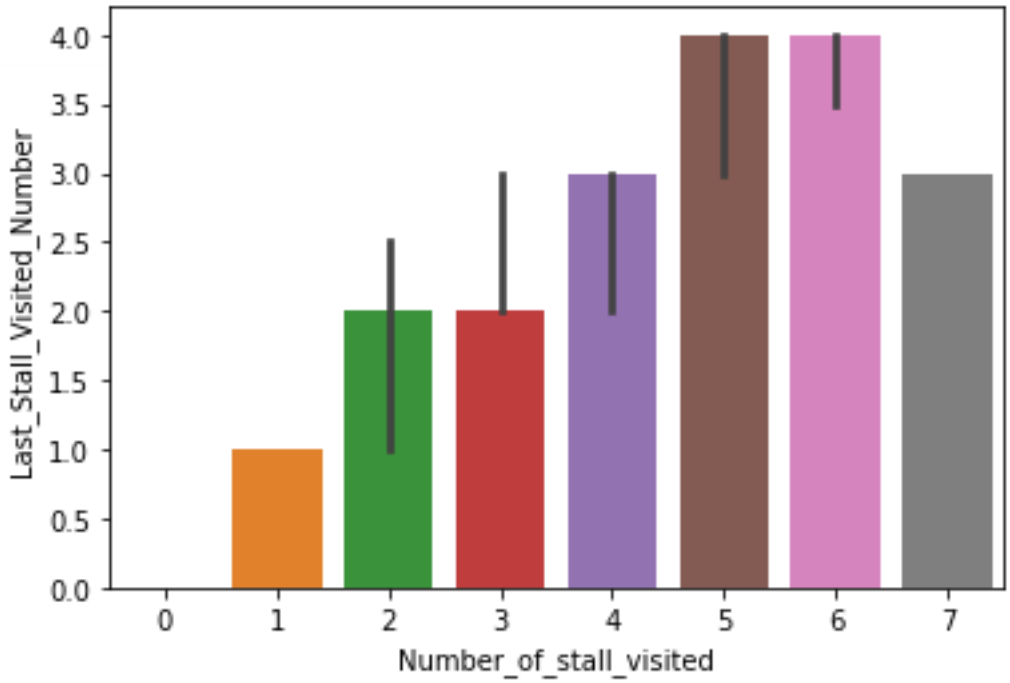

Figure 7 Stalls visited vs the last visited

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.7792	0.8824	0.8276	68
1	0.7500	0.5854	0.6575	41
Accuracy			0.7706	109
Macro Avg	0.7646	0.7339	0.7426	109
Weighted Avg	0.7682	0.7706	0.7636	109

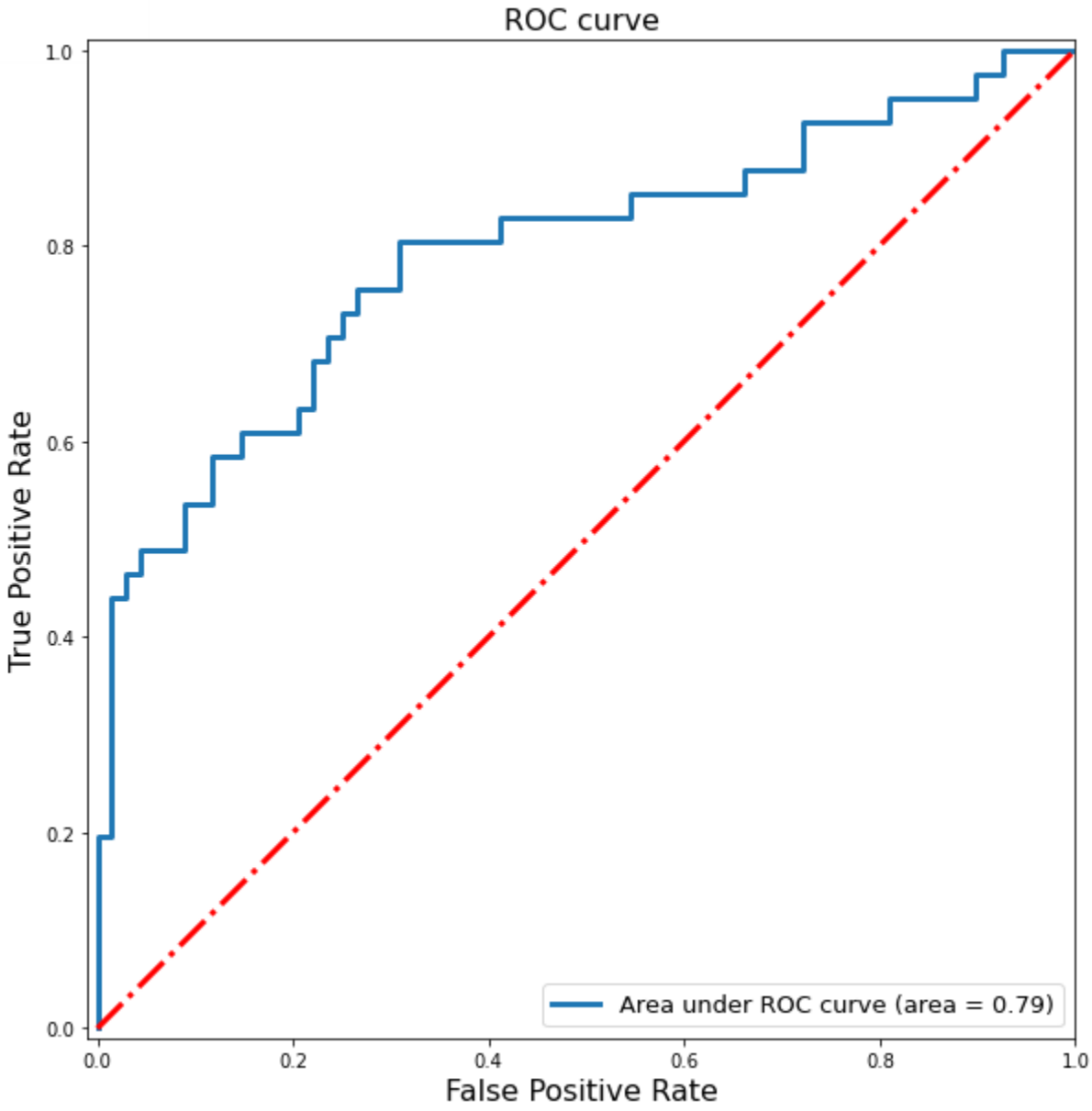

Figure 8 ROC CURVE

CONCLUSION

In this Dataset I have predicted people interested in the insurance as per different variables contributing like Health Scores, Donations, Camp Visited and more. The machine learning method used was Logistic Regression which gave out the best possible results. The Dataset contained different details about all variables concluded. This project helped me understand different ways solving a Machine Learning Projects and also taught me about the insurance and its different ways.

REFERENCE

1. https://www.iciciprulife.com/health-insurance/what-is-health-insurance.html

2. https://www.sas.com/en_in/insights/analytics/machine-learning.html

3. https://machinelearningmastery.com/logistic-regression-for-machine-learning/

4. https://seaborn.pydata.org/

5. https://towardsdatascience.com/life-insurance-risk-prediction-using-machine-learning-algorithms-part-i-data-pre-processing-and-6ca17509c1ef

My Work Datasets

6. https://www.kaggle.com/teju4405/heathcare

7. https://www.kaggle.com/teju4405/book-title-ratings

8. https://www.kaggle.com/teju4405/insurance-prediction

9. https://colab.research.google.com/drive/1cw5pzRIyy88D6ge2al2F84VWVy8wic42