# IME692A:Course Project

Submitted by : Group 8
Arvind Singh Yadav (191026)
Sunil Dhaka (17817735)

Instructor : Prof. Shankar Prawesh

October 22, 2021

## Contents

**Abstract**

In this project we will see the different regression models to examine the role of predictors in determining the difference in the Covid-19 vaccination rate (first dose) between White and Black residents in USA on County level. We will also assess the importance of different predictors in our model and see what factors are associated with such disparities.

## 1 Introduction

Now a days data scientist are using various machine learning algorithms to find patterns in data.They use Supervised algorithms or unsupervised algorithms while dealing with regression problem or classification problem. In this project work we used different machine learning techniques to deal with regression problem. In regression the response variable is continuous.In this scenario we have to build a regression model on training data using different regression techniques like Ordinary least squares, LASSO,Random forest etc.

First we have given description about the datasets we have used in this project and the objectives of this project. Then we have given an extensive theory of all the models that we have used in our project.

## 2 Objective

Following are the objectives of this project:

- Built different prediction/regression models and evaluate the performance of our models on the test data.
- Assess the importance of different predictors in our best model.

## 3 Data Description

- The given data set has 19 variables and have 756 rows and the dependent variable is given as **CvdVax_DisparityY** i.e difference in the Covid-19 vaccination rate (first dose) between White and Black residents in a County (in percentage).
- Now we divide the dataset into training and testing set using data labelled as "train" for the model building purpose i.e for training data and rest is for testing purpose.
- Also we have dropped the column **State** and **County**. After splitting the dataset the variable **Test** is also dropped.
- After splitting training datasets has 531 observation and each observation is comprises of 15 predictor variables and continuous response variable as CvdVax_DisparityY. All model training is done on this dataset.
- Test dataset has 225 observation and each observation is comprises of 15 predictor variable and continuous response variable as CvdVax_DisparityY.In this test datset model evaluation will be done.

## 4 Mean Squared Error

MSE is a most used and very simple metric, that we are going to use to evaluate our model. Mean squared error states that finding the squared difference between actual and predicted value.

The mean squared error (MSE) is defined as:

$$\text{MSE} = \frac{\sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2}{n} \tag{1}$$

where $y_i$ is the observed value and $\hat{y}_i$ is the predicted value.

What actually the MSE represents? It represents the squared distance between actual and predicted values. we perform squared to avoid the cancellation of negative terms and it is the benefit of MSE.

# 5 Model Description(Theory)

Let y be the CvdVax_DisparityY and $x_{ij}$ be the ith observation of jth predictor.We have standardized the predictor variables.

After doing analysis we found that following 3 models have performed relatively better than other models on the basis of our model evaluation metric MSE:

- Multiple Linear regression
- Support Vector Regression
- Random Forest

## 5.1 Multiple Linear Regression(MLR)

Multiple linear regression is a linear approach for modelling the relationship between a response and one or more explanatory variables .The response is is known as dependent variable and the explanatory variables are known as independent variables.It is wriiten as follows:

$$y_i = \beta_o + \sum_{j=1}^{p} \beta_i x_{ij} + \epsilon_i \tag{2}$$

Solution to this known as OLS and is given by:

$$\hat{\beta} = \min_{\beta_0,\beta_i} \sum_{i=1}^{n} (y_i - \beta_0 + \sum_{j=1}^{p} \beta_i x_{ij})^2 \tag{3}$$

where

- $\epsilon_i$ is an error term.
- $\beta_0$ is the intercept term
- $\beta_i$'s are coefficient of the predictor variables.

## 5.2 Support Vector Regression(SVR)

SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model. To understand the Support Vector regression we need to understand first parameters of SVR .Following are the parameters in SVR:

- **Hyperplane:**This is basically a separating line between two data classes in SVM. But in Support Vector Regression, this is the line that will be used to predict the continuous output.
- **Kernel:**A kernel helps us find a hyperplane in the higher dimensional space without increasing the computational cost. Usually, the computational cost will increase if the dimension of the data increases. This increase in dimension is required when we are unable to find a separating hyperplane in a given dimension and are required to move in a higher dimension
- **Decision Boundary:**A decision boundary can be thought of as a demarcation line (for simplification) on one side of which lie positive examples and on the other side lie the negative examples. On this very line, the examples may be classified as either positive or negative.

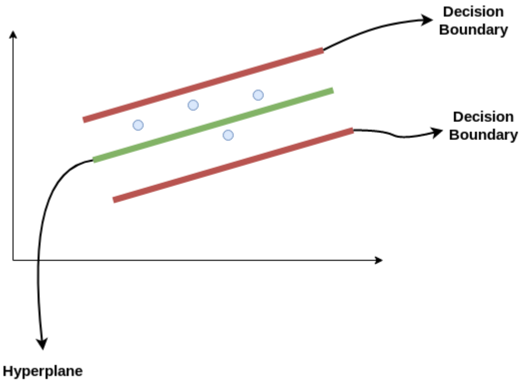


Figure 1: Consider these two red lines as the decision boundary and the green line as the hyperplane.

### 5.2.1 Overview of SVR

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample.

Consider these two red lines as the decision boundary and the green line as the hyperplane. Our objective, when we are moving on with SVR, is to basically consider the points that are within the decision boundary line. Our best fit line is the hyperplane that has a maximum number of points.Our main aim here is to decide a decision boundary at 'a' distance from the original hyperplane such that data points closest to the hyperplane or the support vectors are within that boundary line.
Suppose the training data given are $(x_1, y_1)$ ,$(x_2, y_2)$... $(x_n, y_n)$ where $y_i \in \mathbb{R}$ and and $x_i$ belongs to input space.Then f(x) can be defined as

$$f(x) = \langle w, x \rangle \tag{4}$$

where $\langle w, x \rangle$ is the inner product.
The w can be obtained by minimizing the norm of w. This problem can be written as a convex optimization problem

$$\text{minimize } ||w|| \tag{5}$$

subject to the condition

$$y_i - \langle w, x_i \rangle - b \leq \epsilon \tag{6}$$

$$\langle w, x_i \rangle + b - y_i \leq \epsilon \tag{7}$$

. After the computational process of b and the construction of the regression model, the examples that come with the non-vanishing coefficients are called the support vectors. More number of support vectors explains the relationship more accurately. The support vector method uses the concept of kernel to covert the given data into higher dimension. There are four main types of kernels used, namely linear, polynomial, sigmoid and radial basis function kernel(rbf). Kernel used in this project for the numerical study is radial basis function kernel,

## 5.3 Random Forest

Let's see few terms required to understand ensemble Random Forest:

### 5.3.1 Decision Trees

Decision Trees are predictive models that use a set of binary rules to calculate a target value.Each individual tree is a fairly simple model that has branches, nodes and leaves data. Following are ways to build decision trees:

- We divide the predictor space—that is, the set of possible values for $X_1, X_2, \cdots, X_{q-1}$—into J distinct and non-overlapping regions,$R_1, R_2, \cdots, R_J$ .

- For every observation that falls into the region $R_j$ , we make the same prediction, which is simply the mean of the response values for the training observations in $R_j$ .

The predictor space is divided into high dimensional boxes.the goal is to find boxes $R_1, \ldots, R_J$ that minimize the RSS,given by

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \widehat{y}_{R_j})^2 \tag{8}$$

Where $\widehat{y}_{R_j}$ is the mean response for the training observations within the jth box. . For this , we take a top-down, approach known as **recursive binary splitting**.

For performing recursive binary splitting on the predictor space , first select the predictor $X_j$ and then cut at the point x such that splitting the predictor space into the regions $X|X_j < x$ and $X|X_j \geq x$ that leads to the maximum reduction in RSS. We consider all predictors $X_1, X_2, \cdots, X_{q-1}$, and all possible values of the cutoff point x for each of the predictors, and then choose the predictor and cutoff point such that the resulting tree has the lowest RSS indicated in equation (8). The process continues until a stopping criterion is reached. Once the non-overlapping regions,$R_1, R_2, \cdots, R_J$ has been created, we predict the response for a given test observations.

### 5.3.2 Bagging

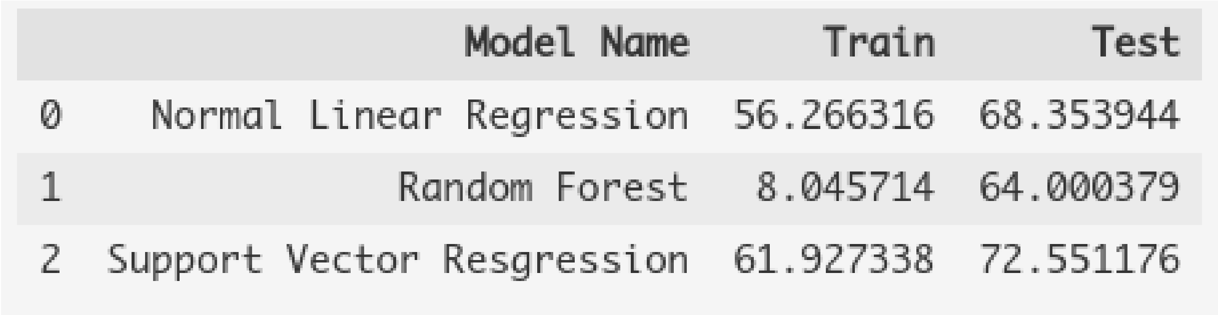
Bagging is used to reduce the variance of a decision tree.Algoritjm create several subsets of data from training sample chosen randomly with replacement.Then, each collection of subset data is used to train their decision trees. As a result, we get an ensemble of different decision tree models. Average of all the predictions from different trees are used which is more robust than a single decision tree.

### 5.3.3 Overview of Random Forest

Random Forest is an extension over bagging.In addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees. It is builds a large collection of de-correlated trees, and then averages them. .Random forest decorrelates the tree since each time a split is performed, a random sample of m predictors is chosen as split candidate from the full set of p predictors.For regression, the default value for m is p/3 . When used for regression, the predictions from each tree at a target point are averaged.

# 6 Model results

Following are the performance of our models:



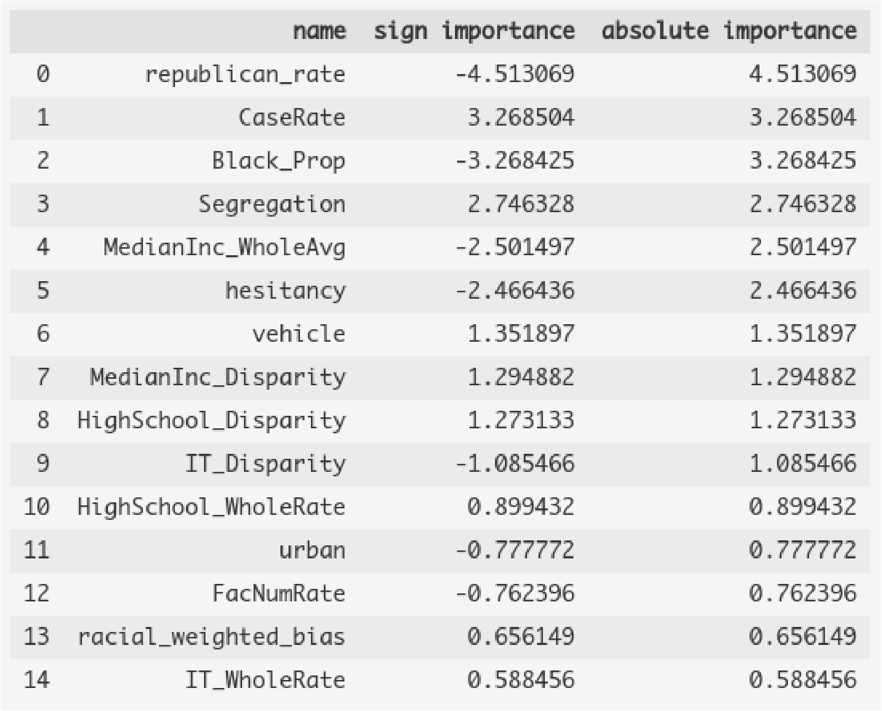
| | Model Name | Train | Test |
|---|---|---|---|
| 0 | Normal Linear Regression | 56.266316 | 68.353944 |
| 1 | Random Forest | 8.045714 | 64.000379 |
| 2 | Support Vector Resgression | 61.927338 | 72.551176 |

Figure 2: Train and Test MSEs for Various Regression Models

### 6.1 Observations

Basic thing that we want in a machine learning model are: it should be simple to implement and more importantly simple to interpret. We also want it to give better predictions about new test points and have better generalization ability from training dataset. Now in all the models that we have tried Normal Linear Regression Model performs better on all the points than other methods. Other methods are more complex than normal linear regression. Some also have overfitting problem, in particular Random Forest and SVR. Relatively, normal linear regression does give lower test mse as well as better generalization(less overfitting).

## 7 Feature Importance



| | name | sign importance | absolute importance |
|---|---|---|---|
| 0 | republican_rate | -4.513069 | 4.513069 |
| 1 | CaseRate | 3.268504 | 3.268504 |
| 2 | Black_Prop | -3.268425 | 3.268425 |
| 3 | Segregation | 2.746328 | 2.746328 |
| 4 | MedianInc_WholeAvg | -2.501497 | 2.501497 |
| 5 | hesitancy | -2.466436 | 2.466436 |
| 6 | vehicle | 1.351897 | 1.351897 |
| 7 | MedianInc_Disparity | 1.294882 | 1.294882 |
| 8 | HighSchool_Disparity | 1.273133 | 1.273133 |
| 9 | IT_Disparity | -1.085466 | 1.085466 |
| 10 | HighSchool_WholeRate | 0.899432 | 0.899432 |
| 11 | urban | -0.777772 | 0.777772 |
| 12 | FacNumRate | -0.762396 | 0.762396 |
| 13 | racial_weighted_bias | 0.656149 | 0.656149 |
| 14 | IT_WholeRate | 0.588456 | 0.588456 |

Figure 3: For normal or multiple linear regression (MLR)feature importance(with sign and without sign)

### 7.1 Observations about predictors

- Our feature importance coefficients are almost same as the authors(of original paper) have reported
- Also our predictor importance ordering is mostly same as the authors.
- Authors also have got a highest negative value for political ideology, we also find same results
- We also found that segregation has a positive role in covid vaccination dis parity
- Median income also plays a quite significant role in vaccination disparity
- Overall socio-economic factors does have a sound effect on covid vaccination disparity between black and white races in USA.

Some features have slightly different importance in our best model, this could be due to following reasons:

- They have used whole dataset (train+test) to report importance of predictors.

- Their model parameters might be different then our in turn that might lead to slight differences in feature importance
- They also have not considered outlier observations

## 8 Conclusions:

- We have fitted different prediction/regression model to examine the role of predictors in determining CvdVax_DisparityY in the train data then checked the model accuracy using MSE of the test data
- Found out that Multiple linear regression works better as it generalises well over both test and train data.
- Assesed the importance of each predictors.
- Found significantly similar results as the author.

## 9 PLEASE NOTE

To avoid repeat of what we have reported in our Jupyter notebook, we are not reporting extensively here.We request you to look at the section-wise markdown report(in detail) that we have created, for the project report.
We strongly believe that to understand code and report they should go hand in hand, so that one can look at both to understand.To do this we have used markdown section of notebook file. This way one can look at the plots, data tables, and observations in one place.
Basically, plots and analysis of this project are attached in the attached code file.

# References

1. An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)
2. Muthukrishnan. R, Maryam Jamila. S ”Predictive Modeling Using Support Vector Regression” international journal of scientific technology research volume9, issue 02, february 2020
3. analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/h2

# List of Figures