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 a County.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.

2 Objective

Following are the objectives of this project:

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

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.

5 Model Description

Let y be the CvdVax_DisparityY and X be the matrix of predictors including the intercept term.We have standardized the predictor variables.

After doing analysis we found that follwing 3 models have performed relatively better than other models on the basis of MSE:

- Lasso
- Support Vector Regression
- Random Forest

5.1 Lasso

Lasso (least absolute shrinkage and selection operator; also Lasso or LASSO)is a regression analysis method that performs both variable selection and regularization in order to improve the prediction, accuracy and interpretability of the resulting statistical model.

5.1.1 Basic Form

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

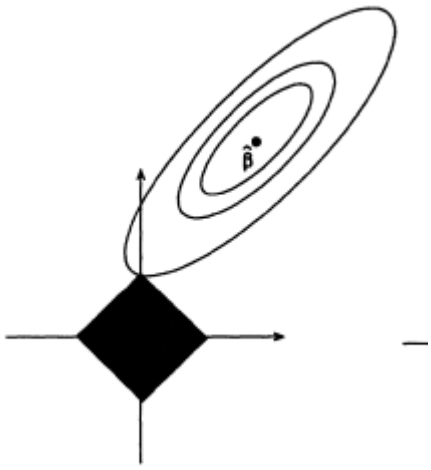
subject to condition that

$$\sum_{j=1}^{p} |beta_j| \leq \text{B} \tag{3}$$

where

- ϵ_i is an error term.
- β_0 is the intercept term
- β_i's are coefficient of the predictor variables.
- A tight bound B, corresponds to a large penalty *lambda* and vice-versa where *lambda* is the hyperparameter.

Lasso can set coefficients to zero due to the the shape of its constraint boundaries as shown in the figure below



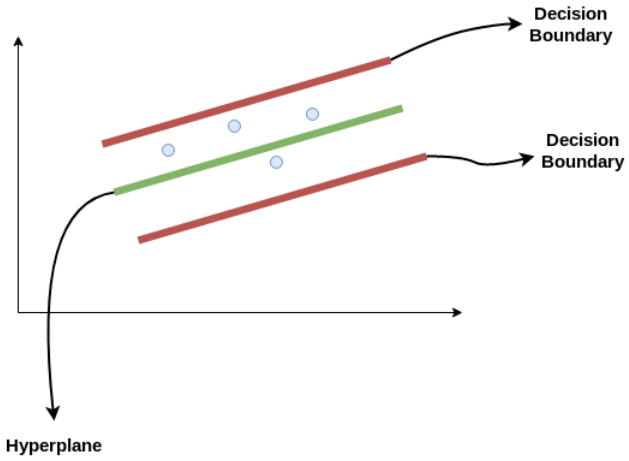
5.2 Support Vector Regression(SVR)

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.

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 can be summarized with the below points:

i. Decision trees are predictive models that use a set of binary rules to calculate a target value.
ii. Each individual tree is a fairly simple model that has branches, nodes and leaves.data.

Overview of the Regression Tree Algorithm: There are two steps for building a regression tree. They are:

- 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**.

The recursive binary splitting approach is top-down since it begins at the top of the tree at which point all observations belong to a single region and then splits the predictor space. Each new split is indicated by two new branches further down on the tree. 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 (6.1). 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 or Bootstrap Aggregation is used when our goal is to reduce the variance of a decision tree. In this we 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 end up with 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. It takes one extra step where 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 a substantial modification of bagging that 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 simply averaged.

6 Comparison of different models

Model	**MSE(Train)**	**MSE (Test)**
LASSO	56.27	68.31
Support Vector Regression	30.85	61.41
Random Forest	8.42	61.02

Observations:

- We have tried different multiple linear regression models but LASSO outperformed them due to the fact that it shrinks the non important coefficents to zero.
- SVR performs better than Lasso in both cases of test and train MSE.
- Random forest has best performance among all , but it seems to overfit the data due to the difference between train and test MSE.

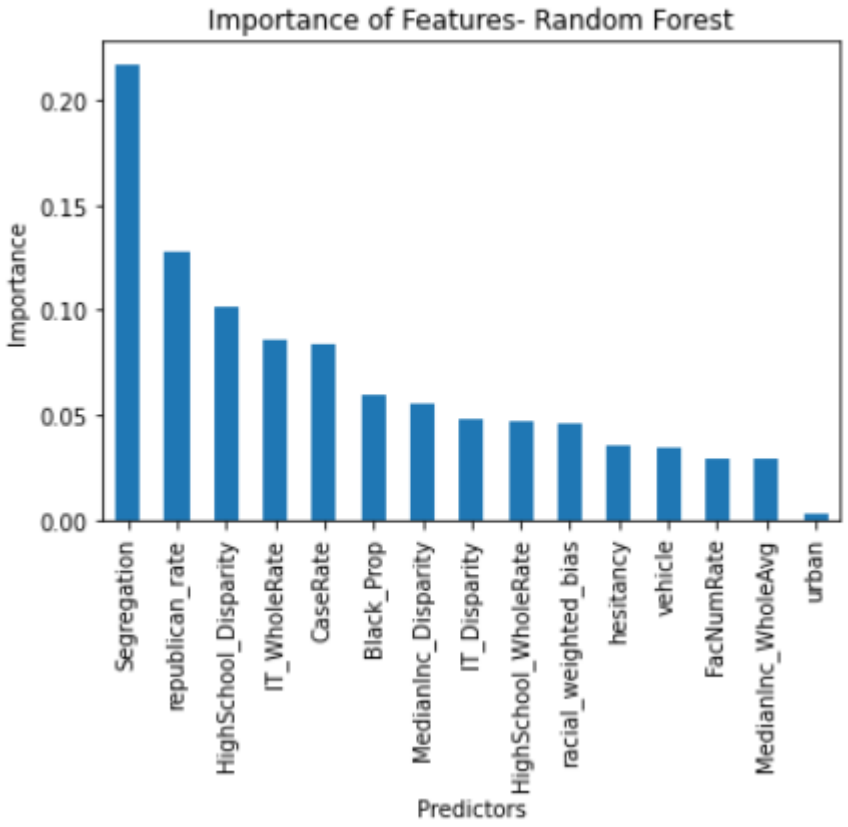
7 Explanation for the best model

From our analysis we have decided that Random forest is the best algorithm for prediction in the given data set due to the following reasons:

- It has the lowest test and train MSE.
- It is easy to know the relative importance of the each predictor as shown in next section.

8 Assessing the importance of different predictors

Unfortunately it can be difficult to interpret the resulting model in case of Random forest algorithm because when we bag a large number of trees, it is no longer possible to represent the resulting statistical learning procedure using a single tree, and it is no longer clear which variables are most important to the procedure. Thus, random forest improves prediction accuracy at the expense of interpretability. But one can obtain an overall summary of the importance of each predictor using the residual sum of squares. We can record the total amount that the RSS is decreased due to splits over a given predictor, averaged over all the trees. A large value indicates an important predictor.In the case of Random Forest regression , the node impurity is measured by the training RSS. A graphical representation of the variable importances in the calculated on the train data set is shown below.



The results indicate that across all of the trees considered in the random forest, the **Segeration** , **republican_rate** and **HighSchool_Disparity** are the most important predictors. Following are the exact information about that important features.

- **Segeration:** Black-White segregation index measures. This index ranges from 0 (complete integration) to 100 complete segregation.
- **republican_rate:** The share of votes cast for the Republican candidate in the 2020 election

- **HighSchool_Disparity:** Difference in county level high school education between White and Black population

9 Findings

Our findings was different from the paper given in the hyperlink
https://www.pnas.org/content/118/33/e2107873118 due to the following reasons:

- In paper they have used regression analysis as predictive and in our report we have found Random forest as one of the best model on test data.
- The authors have winsorized all variables at 5th and 95th percentiles.
- The authors have used all 756 observation for the modelling and fitted the model after pre-processing. Also the metric reported was on same data they have modelled, not on test data.
- Lastly we have used an ensemble method for quantifying the importance of predictors, and in the paper they have used standard regression methods using OLS.

10 Conclusion

- We have fitted 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.
- In cases of Multiple linear regression LASSO performed better in terms of low test MSE.
- Random forest have minimum Test and Train MSE of 61.02 and 8.42 respectively .
- Visualised the importance of predictors using Random forest algorithm and found that **Segeration** , **republican_rate** and **HighSchool_Disparity** are the important predictors .

11 References

- An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)
- Muthukrishnan. R, Maryam Jamila. S "Predictive Modeling Using Support Vector Regression" INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020
- analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/h2