ISEN 613

PROJECT REPORT - PHASE 1

COVID-19 HOSPITALIZATION PREDICTION USING FOOT TRAFFIC AND MOBILITY DATA

YADHU KRISHNAN 632005640

PRATIK LOKHANDE 132005562

SHRADDHA PHALKE 932008341

Contents

Executive Summary

COVID-19 has disrupted healthcare systems in terms of shortages in beds and equipment for treatment of patients, and non-availability of healthcare workers. Therefore, it is important to predict the number of hospitalizations so that the above requirements for treating the patients are fulfilled in time. The model built is based on an available data set that consists of number of individuals hospitalized on a given day due to COVID-19 for eight months of the pandemic (11^{th} April 2020 to 11^{th} January 2021). The region considered is Bryan/College Station consisting of 27 different localities within the town.

The available data set is pre-processed and cleaned by replacing missing data by optimally estimating the variable values. Lag variables up to 7 days is added to the total influx population at every zip code. This is necessary because people entering the locality 11 days to at least 5 days before the baseline date can transmit the disease and influence hospitalization. Further, after trying multiple regression models like Linear Regression, Ridge & Lasso Regression, SVM, etc., it was found that predicting the number of hospitalizations was best explained by forward subset selection method.

Post model analysis, (1) Mobility changes in places of activity and (2) Foot Traffic per zip code were found to be the most important attributes that significantly affected the number of hospitalizations on a given day. (1) Age demographics per zip code, (2) Average Income per zip code and (3) Percentage of people using public transportation per zip code had no critical impact on the prediction.

Interpretation of Best Model:

- The number of people hospitalized is directly influenced by the influx of people in every zip code, in general, prior to 9-7 days. **Less people entering a zip code is inferred to be an increase in hospitalization.** A similar trend was observed during peak's of Covid when people visited areas of red zone/hot spots less frequently.
- **Zip codes 77852 and 77805 are considered hotspots** due to their change in population influx's directly indicating a strong change in hospitalizations.
- **A 20% decline in mobility in zip code 77805 over 11 days can indicate the hospitalizations to increase by 33**.
- Mobility changes in residential and workspaces are most significant in predicting hospitalizations. **A 10% decline in mobility in residential areas and workplaces over 11 days suggest and increase in hospitalization by 50 and 10 respectively**. This adds to the common rationale that people tend to work less from workplaces when hospitalization are high in general.
- **A 15% increase in mobility across restaurants, groceries, retail and movie theatres suggest an increase in hospitalization by 8.**
- **A 10% decrease in mobility across public transport hubs indicate and increase in hospitalization by 5.**

All team members have contributed evenly to performing different phases of the project. Data manipulations, modelling, performing the analysis, drafting of report, and interpretations of all stated models were prepared and assessed together equally.

Technical Summary

Model 1 – Random Forests Regression

Random Forests is a nonlinear model that consists of aggregating the results of an ensemble of decision trees, each created using a subsample of the data. It is like bagging but with an additional layer of randomness. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis and support vector machines, and is robust against overfitting.

The accuracy of random forest in predicting hospitalizations to other non-linear models like SVM, Principal Component Regression and subset selection were compared. It is observed that random forests run quite favourably on the given data as compared to the other. Given data was split into samples for training (70%) and the remaining for testing (30%). Different models with 500, 1000, 2000, and 3000 trees were tested. The number of variables tried at each split (mtry) were considered equal to 86 (p/3), 129 (p/2) and 16 (p^0.5), and the node-size was set to default. After doing numerous iterations of random forests on the training data, model with 1000 trees for the mtry values has maximum variability (53.46%) explained as compared to the others.

Further, the random forest for different mtry values (p/2, p/3 and p/4) were run and number of hospitalizations on the test data split were predicted from the original data. Root Test Mean Square Error (RMSE) is used to compare the models and conclude on the one which has the least Test RMSE.

Table below shows the comparison of the root mean square errors (RMSE):

Sr. No.	Value of mtry	Training RMSE	Test RMSE
1	p/3 = 86	8.91	11.06
2	p/2 = 129	6.83	10.94
3	p/4 ≈ 60	7.06	13.89

Figure 1

The model with mtry = 60 has comparatively the least Test RMSE value. Hence, this model is further analysed, and important predictors were identified. Figure 2 displays the graph between number of trees and MSE. It is observed that mean square error is least for number of trees 1000.

Figure 3 is the variable importance plot with reference to average increase in Node Purity. The x-axis displays the average increase in node purity of the regression trees based on splitting on the various predictors displayed on the y-axis. It can be inferred that variable “X76629.t.influx.change_

from_baseline_11" i.e., number of people entering the zip code 76629, 11 days before the baseline is most critical in predicting number of hospitalizations followed by "workplaces_percent_change_ from_baseline_11" i.e., mobility changes at work places 11 days before and "X77840.t.influx.change_ from_baseline_10", number of people entering zip code 77840, 10 days prior.

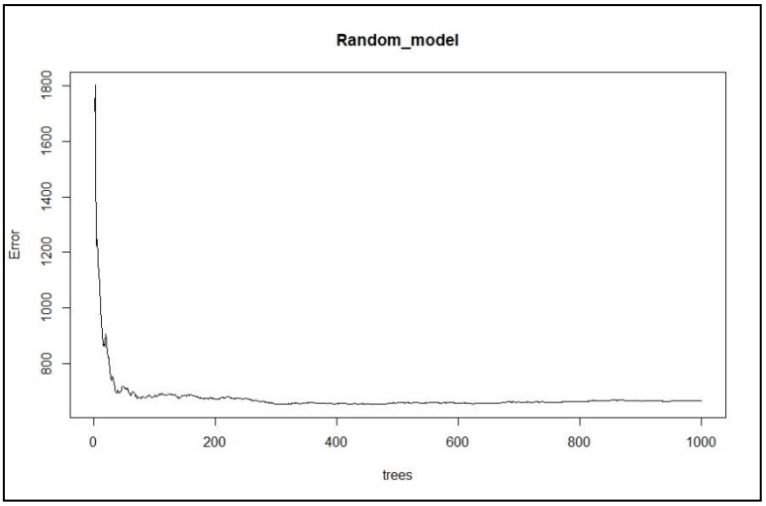

Figure 2

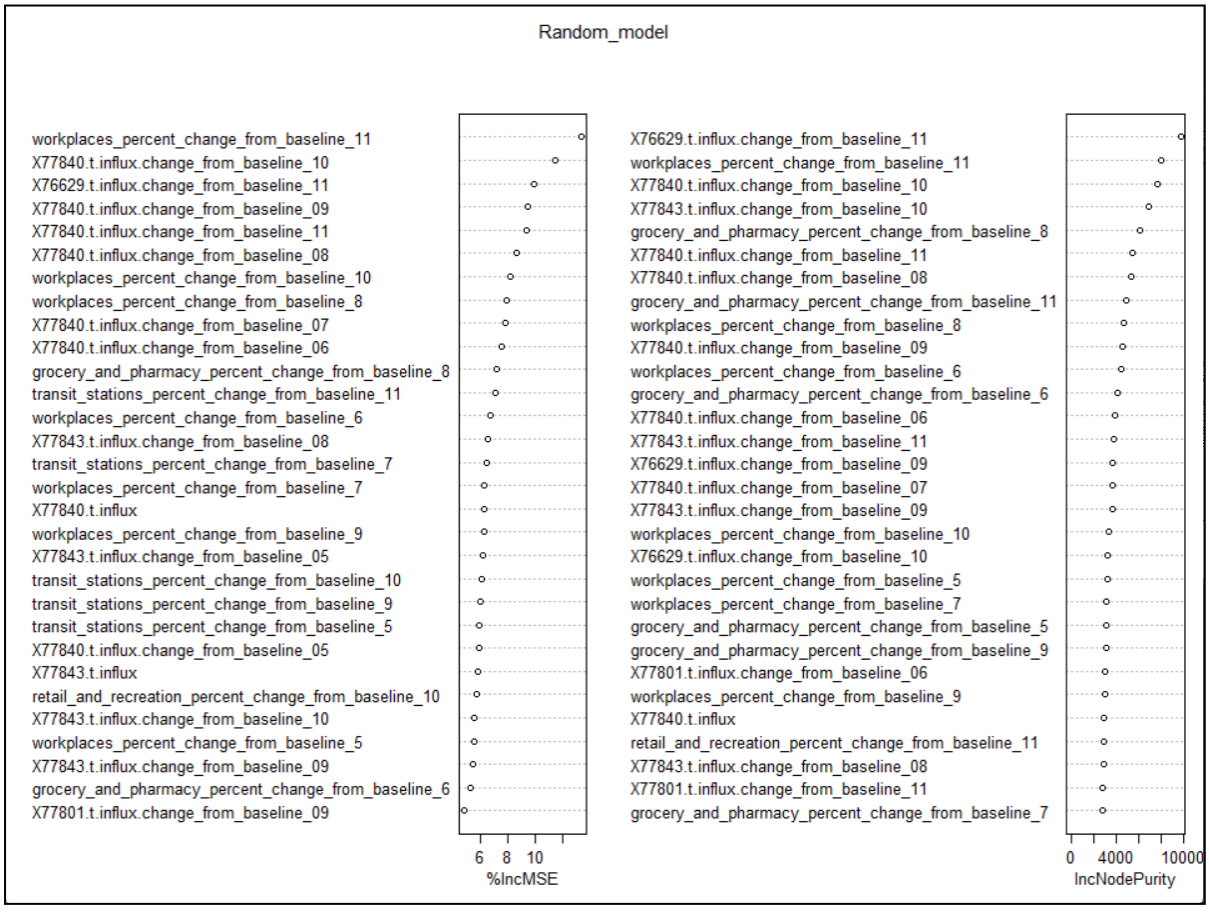

Figure 3

It is also examined that our Random Forest Regression Model has the least root mean square error as compared to the boosting approach with a reduction of 9% in the error. However, the models discussed further in this report reflect more accuracy.

Model 2 – Forward Subset Selection

The rationale behind choosing forward selection approach is due to its ability to deal with huge number of predictors unlike best subset method which would get computationally cumbersome. Forward selection is tractable and gives a good sequence of models. Thus, for the cleaned data set of 258 predictors, forward selection is a good possible candidate.

To ensure reliability, the dataset is split into 70-30 for training and testing respectively. The training observations alone are considered since it will yield a more accurate estimate of the test error. In the iterative process the initial active null model is updated by one variable at each iteration instead of completely re-optimizing the model. The final best model upon completing all the iterations include 40 predictors. These 40 predictors in the model, according to forward selection, are the most significant in predicting hospitalizations.

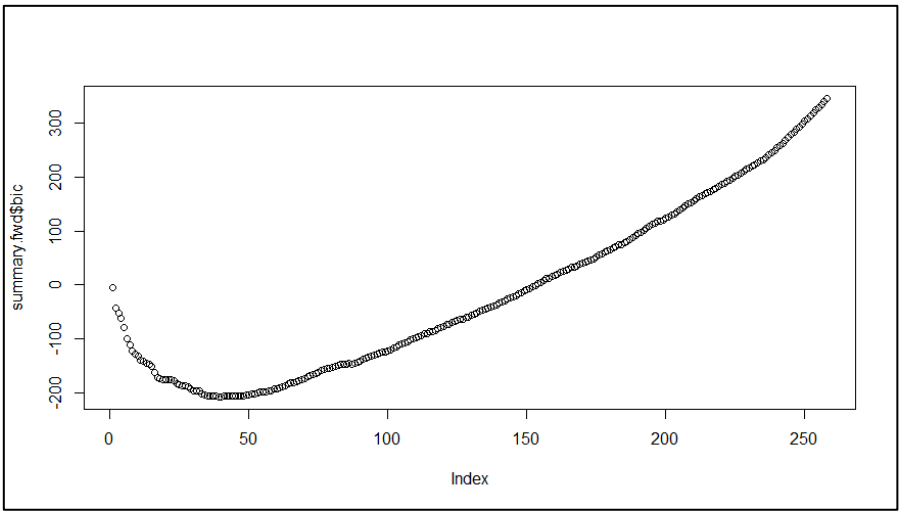

Figure 4

Figure 4 plots the respective BIC obtained for all the iterative models generated with forward selection approach up to 258 predictors. A clear depiction of high penalty on increasing number of variables is observed for the models with more than half the initial number of predictors. The final best model with the lowest BIC is observed to have 40 predictors. The coefficients of the 40 predictors in the best model is displayed in Figure 5.

Mobility changes in residential area from baselines 11, 10 and 8 seem to have highly significant effect on hospitalizations which is indicated by their highly negative coefficients -5.01, -1.77 and -2.03. Workplace mobility changes from baseline 11 also has a similar significance owing to a -1.14 coefficient. This adds to the general rationale that residential areas and workplaces being highly interactive, influences covid transmissions and hospitalizations significantly. Thus, as mobility changes at work and residential areas increase over a stretch of 11 to 8 days it could be a result of a decrease in hospitalization and a higher decrease in mobility would directly indicate an increase in hospitalization. Although not as highly impactful as mobility changes in residential and workplace areas, retail & recreation and transit stations also seem to be capable of influencing hospitalization predictions. This is evident from their slightly lower

in magnitude coefficients 0.57, -0.3. A higher decrease in mobility change in transit stations with respect to baseline 11 could directly indicate an increase in hospitalization.

Population influx changes in zip codes 77852 & 77805 from baseline 7 with coefficients -1.67 and -0.79 indicate a strong influence on hospitalizations. Population influx changes across other zip codes also seem to have a decent effect on hospitalization but not as influential as zip codes 77852 and 77805; indicating these areas to be considered a likely hotspot. Population influx changes regarding most zip codes seem to be significant for baselines 9 and 7 when compared to other baselines potentially due to the reason that quarantine zones are declared usually these many days.

(Intercept)	retail_and_recreation_percent_change_from_baseline_11
116.903001371	0.570934125
parks_percent_change_from_baseline_11	transit_stations_percent_change_from_baseline_11
-0.239102773	-0.376369158
workplaces_percent_change_from_baseline_11	residential_percent_change_from_baseline_11
-1.142245257	-5.011303664
residential_percent_change_from_baseline_10	parks_percent_change_from_baseline_9
-1.770340533	-0.369190654
transit_stations_percent_change_from_baseline_9	residential_percent_change_from_baseline_9
-0.433831855	-0.729320749
residential_percent_change_from_baseline_8	parks_percent_change_from_baseline_5
-2.038632651	-0.278013515
workplaces_percent_change_from_baseline_5	X77801.t.influx
-0.236274006	-0.048907421
X77882.t.influx	X77840.t.influx
-0.054073547	-0.002048627
X77807.t.influx	X77879.t.influx
0.069202808	-0.109258819
X77802.t.influx.change_from_baseline_11	X77801.t.influx.change_from_baseline_11
0.014131645	-0.058997028
X77805.t.influx.change_from_baseline_11	X77805.t.influx.change_from_baseline_09
-0.790172438	-0.793495057
X77879.t.influx.change_from_baseline_09	X77808.t.influx.change_from_baseline_08
-0.127881633	0.045872929
X77805.t.influx.change_from_baseline_08	X77863.t.influx.change_from_baseline_08
-0.478284573	0.581640881
X77808.t.influx.change_from_baseline_07	X77802.t.influx.change_from_baseline_07
0.031166095	0.007011061
X77801.t.influx.change_from_baseline_07	X77845.t.influx.change_from_baseline_07
-0.019671264	-0.006420207
X77852.t.influx.change_from_baseline_07	X77863.t.influx.change_from_baseline_07
-1.671268698	0.706928331
X76629.t.influx.change_from_baseline_07	X77808.t.influx.change_from_baseline_06
-0.066504355	0.044683434
X77801.t.influx.change_from_baseline_06	X77805.t.influx.change_from_baseline_06
-0.038575355	-0.423031250
X77863.t.influx.change_from_baseline_06	X76629.t.influx.change_from_baseline_06
0.643339811	-0.064627439
X77807.t.influx.change_from_baseline_05	X77844.t.influx.change_from_baseline_05
0.018423523	-0.057088128
X77879.t.influx.change_from_baseline_05	
-0.108840909	

Figure 5

BIC statistic was used as the criterion on determining the best model among all the models generated from the forward selection. The BIC statistic enforces a heavy penalty with increasing number of variables; hence a concise number of effective predictors are observed in the final model. The final model with the lowest BIC also tends to have the lowest test error and hence BIC is chosen as the selection criterion over other indirect statistics such as C_p, AIC and Adjusted R^2.

Cp, AIC and Adjusted R^2 generates their best model with 58, 40 and 184 predictors respectively. Hence it is evident that BIC is the best criterion for better interpretability.

Direct methods such as Validation Set Approaches were also considered, however a minimum validation set error of 20.9 resulted in a model with 255 predictors. This being evidently higher than the number of predictors in the best model, put forth by BIC, is discarded.

Model 3 – Principal Components Regression

Principal component regression (PCR) is a regression analysis technique that is based on principal component analysis. Some of the most notable advantages of performing PCR are the following:

i. Dimensionality reduction
ii. Avoidance of multicollinearity between predictors
iii. Overfitting mitigation

The first principal component direction of the data is along which the observations vary the most. Principal components for every predictor are obtained in a new feature space with main objective of maximizing variation or information of data subjects to non-collinearity within variables. By doing so, any high-dimensional data can be reduced to low-dimensional dataset. Severe collinearity can be observed in the given dataset since some of the variables are obtained by doing simple arithmetic calculation. Also, large number of variables makes it more difficult to understand the collinearity between each of these variables. By using principal component regression, the least square model can be fit on non-collinear principal components.

The PCR method generates principal components for every predictor in the data and then fit least-squared linear regression model. The model returns the information about how many numbers of principal components are considered. The intuition behind selecting first M principal components lies in the objective of their creation. Small number of principal components explains most of the variation in the data. Using small number of principal components is a reasonable approximation for obtaining good results.

The input for PCR consists of response variable, prediction variable, scale, and validation. Scaling decides whether the principal components are based on correlation matrix. In matrix algebra, correlation matrix is a scaled version of covariance matrix. If the relative magnitude of deviation of all predictors is not significant, then correlation matrix is used to calculate principal component. Another important aspect of using correlation matrix is that the relative importance of PCs reduces as compared to PCs calculated with correlation. PCR then computes the ten-fold cross-validation error for each possible value of M, the number of principal components used.

After executing the model, the variance explained is examined and training error and test error are estimated. It is observed that cross-validation error is lowest at M=37. Since, there is no independent test dataset to compute test MSE, the data is split into train set and test set using validation set approach. 70% of the total number of observations are considered for training and fitting the model whereas 30% held for testing the model predictability. After splitting the data, principal component regression is performed using only first 37 principal components.

The training root mean-squared error of 23.01 is the lowest with 37 principal components. The test RMSE is also estimated to be 21.53. This is the average deviation between the predicted value for hospitalizations and the observed value for the same in the testing set. The U-shaped MSE curve is shown in Figure 6. It is clearly depicted that the adjusted cross-validation error and cross validation error almost overlap in the graph. It can also be inferred that the RMSE drastically increases after certain number of components.

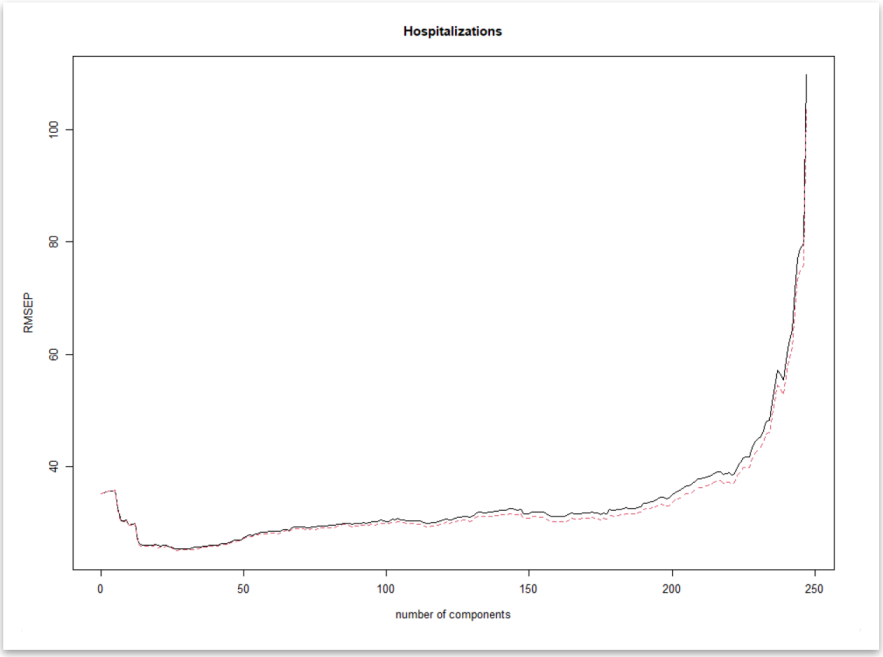

Figure 6

Root mean squared-error vs number of principal components

Even though the test MSE is as reasonable as other models used for the dataset, the final model is very difficult to interpret since it does not give you information about the effect of any variable for the prediction. Information leakage is one of the disadvantages of using pc-based regression model. Even if the first few PCs are considered for prediction, the information or variability that is explained by remaining PCs won't be used. For the new dataset, if those unused PCs are relevant, then prediction accuracy would be compromised.

Comparison of Models

The primary factors considered for choosing the best model are prediction accuracy and model interpretability. The parametric statistical learning methods - Principal Component Regression, Random Forest Regression and Forward Subset Selection are the extensions of linear regression. By observation of the test mean-square error, competitive results are obtained by each model.

Model		**Train RMSE**	**Test RMSE**
Principal Component Regression (PCR)		23.01	21.53
Forward Subset Selection		25.68	20.90
Random Forest Regression	m=p/4	10.10	21.70
	m=p/2	09.83	21.94
	m=p/3	09.91	22.06

The test RMSE ranges from 20-23 for all the methods whereas train MSE is relatively high for forward subset selection and comparatively low for random forest regression. PCR and Forward selection have a higher training MSE than test MSE. This is likely due to sampling bias or the distribution of the data in training and testing dataset. However, the difference between these two quantities is not significant. Considering prediction accuracy, similar test results were obtained with minor difference between the quantity. The lowest test RMSE is achieved by forward subset selection. However, considering only prediction accuracy it may not be the best option while choosing the best model.

On the grounds of model interpretability, the forward subset selection is the most interpretable among the three learning methods. Principal component regression does not interpret results based on original variables but the principal components. Thus, there exists no clear information about any important variable which aids in predicting the response variable correctly. Random Forest is an ensemble method which is mainly used to reduce the variance in the regression tree method by decorrelating the variables and thereby increasing the model accuracy. However, random forest grows very large trees in each iteration making it computationally complex and less interpretable. Another important observation is the significant difference between test RMSE and train RMSE for different m values. The method may have to overfit the model which resulted to smaller train RMSE. Overfitting of the model causes a reduction in the predictive power.

Test RMSE is the lowest for forward subset selection method which also indicates that the prediction accuracy of this method is better than PCR and Random Forest regression. Unlike these two methods, forward subset selection is reasonably more interpretable. With a 40-variable model, selected by BIC criterion, one can also plot the importance of each variable for predicting the response. The method gives insight about the relationship between the response and the predictors

References

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani -An Introduction to Statistical Learning - Springer

2. https://quantifyinghealth.com/report-random-forest/

3. https://www.geeksforgeeks.org/random-forest-approach-for-regression-in-r-programming/