CS685A Final Project Report

Group 16

06-12-2020

Implications of Various Factors on COVID-19 spread

1 Team Members

Name	Roll No	Email
Aniket Sharma	170111	anikets@iitk.ac.in
Mihir Shashank Jewalikar	170387	mihirsj@iitk.ac.in
Sagnik Bhattacharya	170604	bsagnik@iitk.ac.in
Sahil Agrawal	170607	sahila@iitk.ac.in

2 Abstract

COVID-19 has become a major cause of concern in present times and has necessitated lock-downs all around the world. But blindly shutting everything down is not feasible for a prolonged period and will result in a crash in the economy. So, it is necessary to predict how this virus will progress in the future so that we can take steps to prevent the virus while keeping the economy going. To solve this problem, we test out multiple models for prediction of COVID trends in the future and compare their advantages and disadvantages. We will also rate many factors like population, literacy, net state domestic product, number of hospital beds, homeless population, average age and unemployment rate based on their impact on the number of cases. Using the results from the above methods, we will show the areas that are more likely to be vulnerable in the future and suggest smart lock-downs.

3 Problem Statement

Given the COVID data till date and data of the 2011 census for the factors mentioned above, the problem statement is to predict the trend of the virus in

the future at the district level. Firstly, to get the data, we will used selenium in python for web-scraping and pre-processing of xlsx files. For the prediction, we will use a linear regression, polynomial regression, Random Forest regression, Artificial Neural Network, and Long Short-Term Memory models. We will compare the results of the above models using R2 score and RMS error. From the prediction of the best model, we will suggest areas that are likely to be vulnerable and should go into lock-down and areas that are low risk areas and can reopen the economy. We will assess the advantages and disadvantages of each model and reason out why the model performing the best is doing so. We will also use an SVM model to rank the various factors above based on the amount of weight they carry in determining the number of COVID cases and compare it with the correlation matrix. From this, we will give a ranking to each factor and this can be a guide to the districts on what to do to keep the pandemic in check.

4 Introduction and Motivation

As we all know very well that Covid-19 pandemic has impacted all our lives and has emerged as a topmost concern for everybody and every governing authority. The authorities throughout the world are trying to control the spread of the pandemic, but the situation is not getting any better. In the process the states are spending a lot of bit human and capital resources to fight against the impact the pandemic has caused. In March we witnessed nationwide lockdown which, though was successful in reducing the covid cases, but lead to unemployment rise, business bankruptcies and hit the economy badly such that the GDP of most of the countries were negative. Due to the former factors, imposing lockdowns has become a dilemma as on one side the covid cases needs to be controlled and on the other side the businesses and employment needs to be ensured. In a developing country such as India, where the resources are already limited, lockdown can cause sever dent in our fight against unemployment and poverty.

Looking at this problem, we notice that a main reason of why the pandemic is not being controlled is that the governments are in a dilemma about at what precise point the lockdowns should be imposed optimally so that neither the covid cases spread like fire nor the economy and employment suffers. If we could predict the rate/trend which the covid curve is going to follow, then the problem may be eliminated, as then we can predict exactly that which district/state will be heading towards being the next hotspot and thus government can timely impose lockdowns only on that district/state, while others can thrive economically.

Thus we got the idea and motivation to build a predictor that predicts the future trends of daily covid-cases of a district, using the past data and various district attributes such as population density, average age, literacy rate etc. Also analysis of the previous data and peaks in various districts/states will help us determine the future trend of covid-cases in the districts.

5 Datasets Used

For each district we needed the Covid-19 data of the districts and various attributes of the districts which may be the deterministic factors of how quickly or slowly the pandemic would spread in the region. Also since the effect of the urban and rural regions may be different, we took the values of the district attributes differently for urban regions and rural regions, whenever available. The details of the datasets, there sources and their respective methods of extraction are given bellow:

1. **Population Density for urban and rural regions of the district.**
   Population density plays a key role in determining the spreading rate of the pandemic, for instance we can see that the situation in Dharavi, Mumbai and Sikkim are quite different. For obtaining the dataset we used the 2011 Census data.

2. **Literacy Rate for Urban and Rural regions of the district.**
   Literacy rate can be considered as a measure to determine how seriously and cautiously people are treating the pandemic concerns. As the rural literacy rate is often low as compared to the urban regions, but still significant awareness is possible, hence it is better to take different data of urban and rural regions. The data was collected through the 2011 census data. For downloading of individual files for states we used the selenium web-automation library.

3. **Net state domestic product (NSDP)**
   The more the people are well off the more the people will follow the guidelines to prevent covid. Net state domestic product (NSDP) gives us a fair estimate of how well-off the people are in the area. As NSDP is calculated statewise, so we assumed uniform distribution of NSDP throughout all the districts in the state, proportional to the population.

4. **Number of hospital beds per 10,000 people**
   A fair estimate of the capacities of all the hospitals in an area can be measured through the number of hospital beds. As the data was available statewise, so we assumed uniform distribution of hospital beds throughout all the districts of the state, with respect to their population.

5. **Percentage of homeless population in Urban and Rural regions of the district.**
   As the homeless people have to live in unhygienic conditions, so their chances of catching the virus is significantly higher. We collected this data from the 2011 census data website, and used selenium web-scraper library to download the individual datafile for all the 632 district.

6. **Average Age in Urban and Rural regions of the districts.**
   As various type of social interactions are age dependent, so we can incorporate a fair estimate of them through average age. As the tasks, interactions

in urban and rural regions are vastly different , so we consider them as different attributes for a district. We collected this data from the 2011 census data website, and used selenium web-scraper library to download all the datafiles.

7. **Unemployed Population Percentage for Urban and Rural regions of the districts.**
   As unemployed population can give us a fair estimate of the number of people finding jobs/daily wage workers, so they shall be traveling to find jobs or be working in unhygienic environments in case of daily wage workers, so this is another factor we should consider that impacts the number of covid cases. We take the data separately for urban and rural regions. We collected this data from the 2011 census data website, and used selenium web-scraper library to download all the datafiles.

8. **COVID-19 data for all the districts.**
   We used the past data from *covid19india.org* to download the daily number of covid cases in all the districts throughout India, from Mid April to December start. We downloaded the data in json format and used it for further processing.

6 Methodology

1. Data Cleaning and Understanding:

   (a) Correlation between input attributes and possible reduction:
   A heatmap of the correlation matrix of all the input features (average age rural, . . . ,). There emerges a strong correlation between average age rural and average age urban but the correlation (90%) is not strong enough to remove one of these attributes (required correlation ¿= 95%) from the input features. Thus none of the initial attributes are removed.

   (b) Missing values :
   there were 53 missing values in the columns "population density rural and urban". Replacing them by the population density values of other rows with similar values of other attributes doesn't make sense because of the dissimilarity and disparity in the number of people living in some districts as opposed to others (e.g. Delhi population - 1,67,87,941, Dindori population - 7,04,524). Hence to smoothen the effects of population density on the covid cases prediction of these districts, they have been replaced by the mean population density urban and rural.

2. Data Visualization:

(a) Histograms:
The histograms of the input features are plotted. The data is more or less uniformly distributed for each attribute.

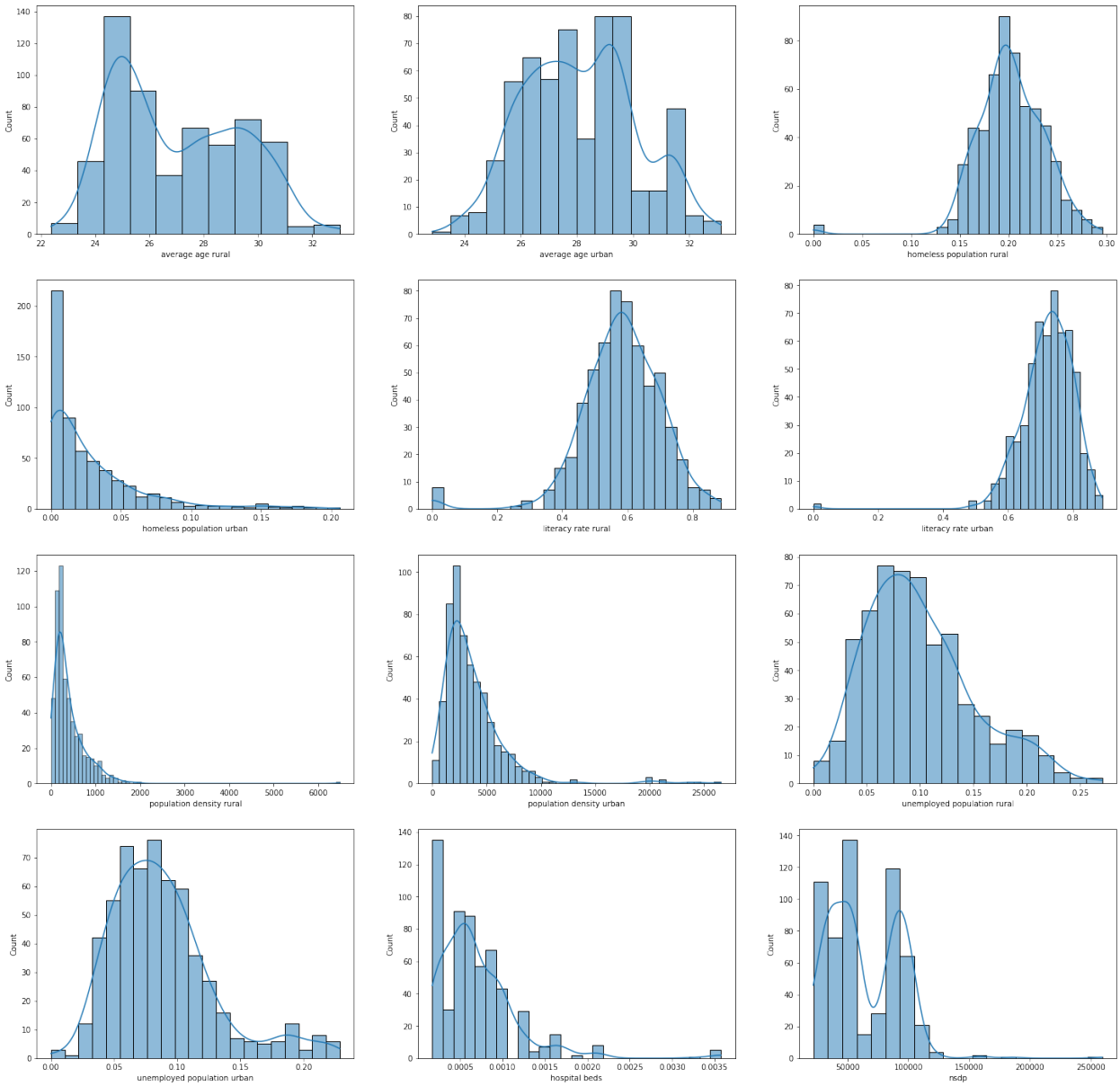

Figure 1: Histogram of number of COVID cases versus attributes

(b) Correlation heat map:
A heat map of the correlation between all the input features is plotted. We see a strong correlation between average age urban and average age rural (90% rural), apart from that none of the pairs of features are closely correlated, nsdp and population have little to no correlation at all

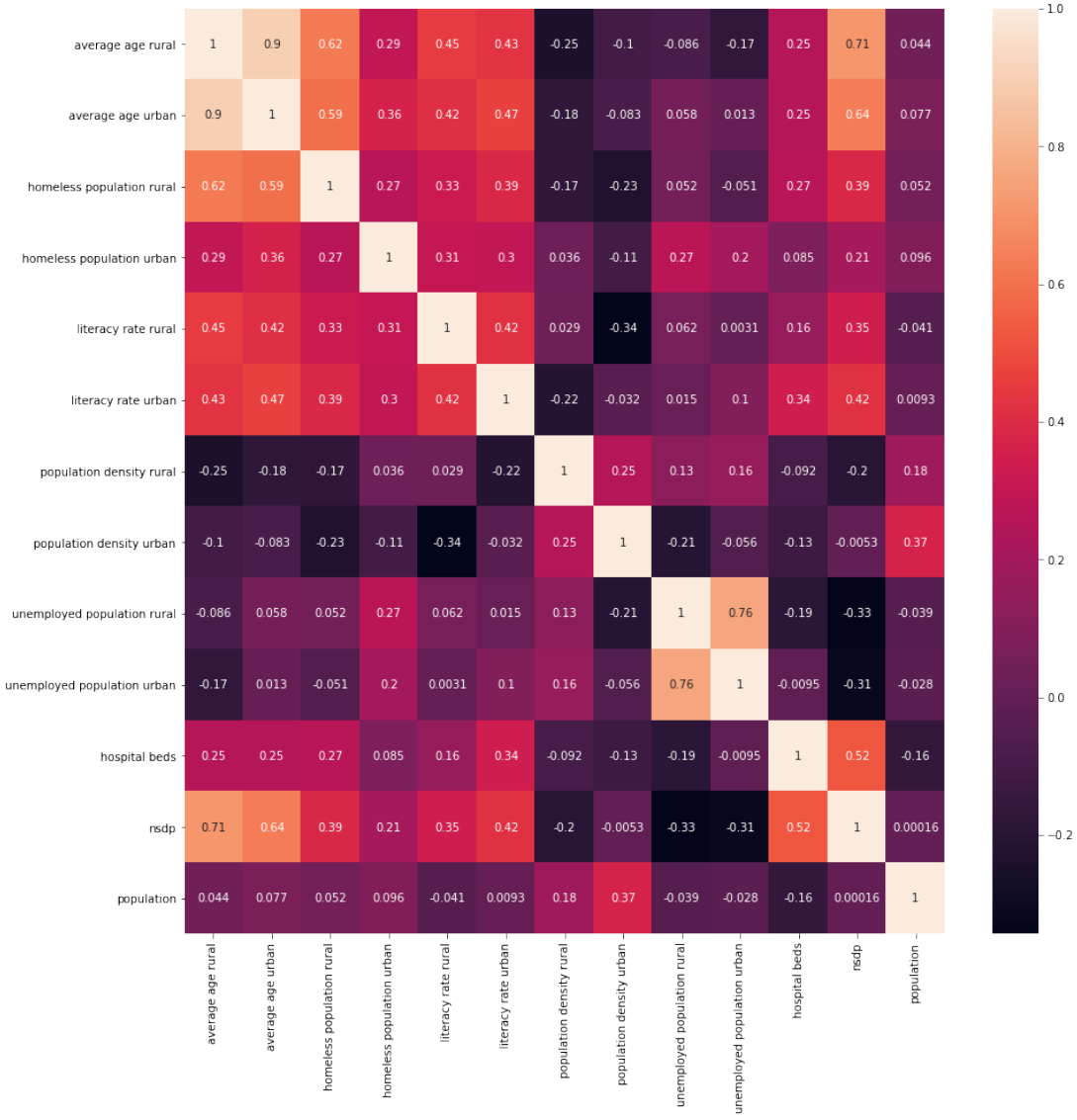

Figure 2: Heat map of Correlation Matrix

(c) Scatter plots of number of covid cases versus input features

Figure 3: Plot of confirmed number of COVID cases versus features

Figure 4: Plot of peak of daily cases versus features

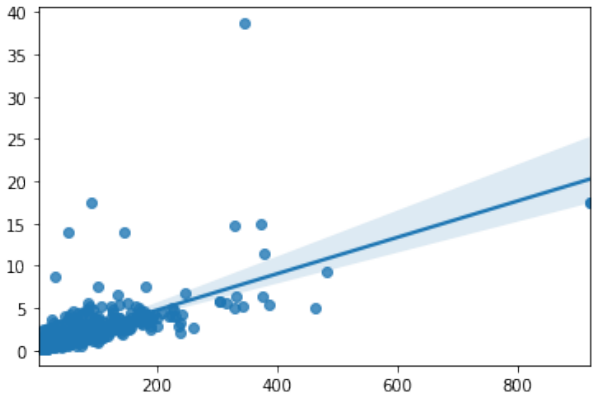

Figure 5: Plot of confirmed number of COVID cases versus peak of daily cases

(d) Heat Maps of COVID cases by state

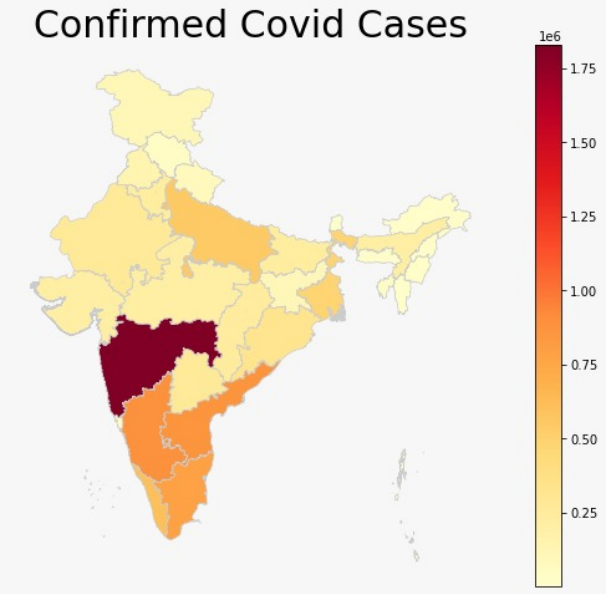

Figure 6: Heat map of confirmed number of COVID cases by state

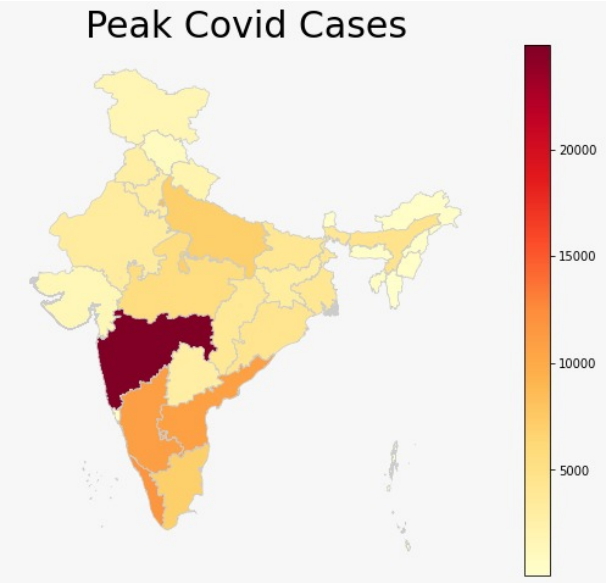

Figure 7: Heat map of peak of daily cases

3. Models:

   Data Standardization:
   We normalize the dataset to zero mean unity standard deviation. The values of covid cases for districts mostly range from -1 to 1 with outliers present occasionally

   (a) Regression:

      i. Linear Regression:
      We started with linear regression but it did not perform well even in training data and performed worse on test data. We tried different linear regressions like lasso regression and Ridge regression but they did not perform well either.

      ii. Polynomial Regression:
      We tried polynomial regression with degree 2. It performed well in training data but diverged in testing data. Increasing the degree(3,4,5,10) did not change the outcome of training data by much but diverged more and more in testing data
      It was understandable that a simple regression won't work as a polynomial will move to infinity for large values of input but that is not the case for cases(input data).

   (b) Random Forest Regression:
   We tried to fit a random forest of decision trees to the data and make

predictions based on that.

(c) Artificial Neural Network (with Keras using Tensorflow backend):
We run a four layered dense artificial neural network. The input is passed as the set of dates and the output is the number of covid cases for each district for that particular date. Performance is better than that of regression but predictions are still highly inaccurate.

(d) LSTM (Long Short Term Memory):
We tried to predict the future values based on given time-series values using LSTM. We tried for different look-backs (1,5,10,12,15) . We chose look-back 5 as increasing it anymore caused the results to diverge.

(e) Feature importance assigning by Support Vector Regression:
To determine the importance of features on number of covid cases, we use Support Vector Regression. The coefficients of the features give us the importance or the correlation with the number of covid cases

7 Results

We have tried and tested a total of 3 methods to predict the future number of covid cases of the districts, based upon the districts historical Covid data and the distrcits general attributes. We shall now look at their results and their advantages and disadvantages individually.
NOTE: All graphs in the results section have been shown for the DELHI district

1. **Regression**

   Linear Regression Root Mean Square error(RMS) on training data : 0.67
   Linear Regression Root Mean Square error(RMS) on test data : 1.73
   Polynomial Regression Root Mean Square error(RMS) on training data : 0.60
   Polynomial Regression Root Mean Square error(RMS) on test data : 3.04

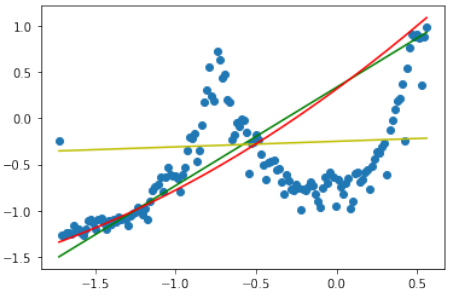

Figure 8: Regression on Training Data

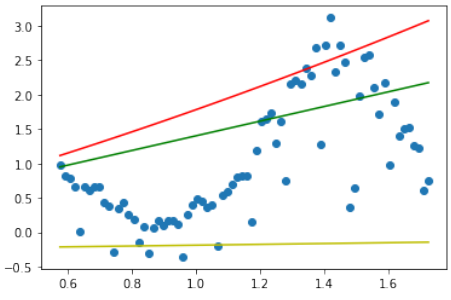

Figure 9: Regression on Test Data

Advantages:

(a) Easy to train and takes less computational power

(b) Fits the covid curve perfectly, so can be used to get estimate of the trends at various times, and predict missing values of training set.

Disadvantages

(a) Predicts the future trend of the covid cases poorly.

(b) Diverges more and more from actual value as time passes

2. **Random Forest Regression**

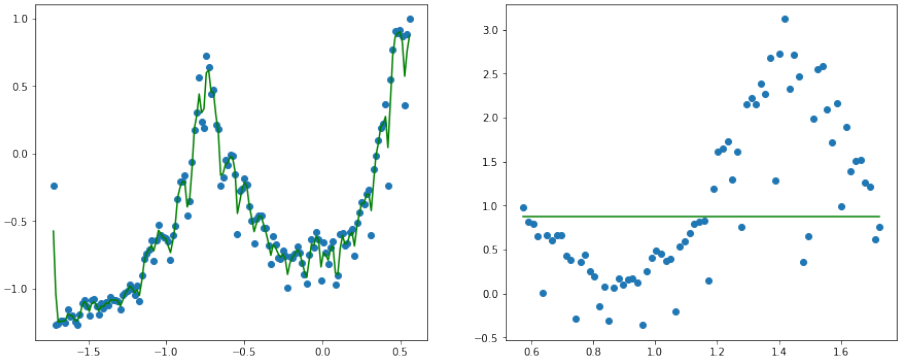

Figure 10: Random Forest Regression on train and test set respectively

3. **Artificial Neural Network**
Root Mean Square error(RMS) on training data : 0.53
Root Mean Square error(RMS) on test data : 0.55

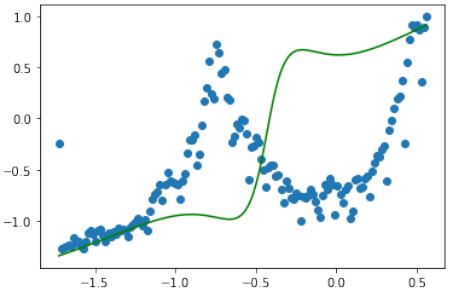

Figure 11: ANN on Training Data

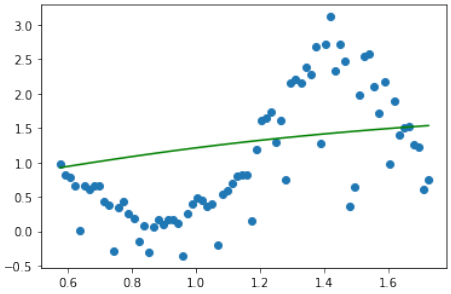

Figure 12: ANN on Test Data

Advantages:

(a) Easy to program and analyse and takes moderate computational power.

(b) It is a nonlinear model and so performs better than regression.

Disadvantages

(a) Has to be trained separately for each district.

(b) The district attributes being same upon the training for a particular district, get embedded in the weights, and hence they have no long term effect on the predictions.

(c) Can not predict a descend in the graph if the recent close values are in increasing arrangement.

4. **LSTM**
Root Mean Square error(RMS) on training data : 0.51
Root Mean Square error(RMS) on test data : 0.71

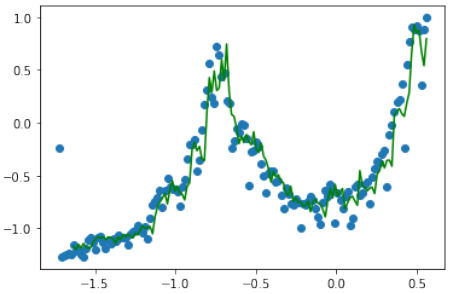

Figure 13: LSTM on Training Data

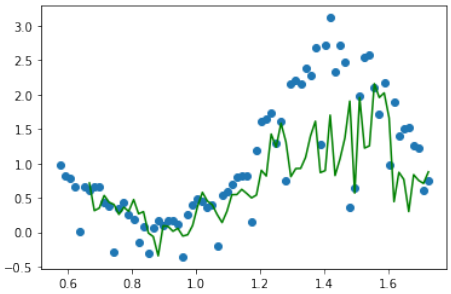

Figure 14: LSTM on Test Data

Advantages:

(a) This is a time series model so the historic results/predictions have significant effect on the future predictions.

(b) Requires the covid data of only previous 10 days to predict the future trends

(c) Has the ability to predict a descent in the number of cases as it can learn to predict wave like curves.

(d) Unlike models like ANN or Regression, which try to fit a global model on the training data, this model takes the advantage of local temporal information and hence not only are its predictions more accurate, considering some of the factors (e.g. testing done per day etc.) which are unknown to us as this stage, remain roughly constant for a span of 5-6 days, and then they change, but also the fact that LSTM can also detect a second peak and hence can be used to accurately predict the rough time and duration of district based smart lockdowns.

Disadvantages

(a) It can't be used for long term predictions, due to insufficient data for large scale training.

(b) Computationally expensive compared to other methods

8 Discussion

First let us look at how the general attributes of a place affect the spread of pandemic, and also let us see the relative contribution of the factors in the number spread of Covid cases.
We ran an SVM model to see how much a feature contributes to the covid cases. The results of that in decreasing contribution towards the pandemic spread are:

1. Average Age Rural
   It is surprising that the average age in rural is affecting the districts covid spread the most based on the available data. A possible reason for this may be that as village is a very social region, so as the average age increases then the retired people visit each other more often, which directly increases the chances of caching COVID-19.

2. Hospital Beds Per 10,000 people
   As can be expected the healthcare system directly impacts the pandemic effect in a region. Since Hospital beds can be considered as the capacity/extent of the healthcare system, so we can see that the healthcare is the 2nd most influential attribute to determine the COVID-19 spread.

3. Population Density Urban
   As urban regions generally have a high population density, which is a serious problem, as it will lead to lack of social distancing and high number of unavoidable interactions with others, which in turn increases the chances of catching the virus. So as expected the population density plays a key role, and is the 3rd most influential attribute to determine the COVID-19 spread.

4. Unemployed Population Rural
   In rural regions farming is a wide spread job, so most of the daily wage workers have to do labour in farms, which will lead to more chances of getting covid -19 hence it is a major factor.

5. NSDP
   Since in a region where people are well-off, the guidelines to prevent covid will be followed more cautiously, and they can afford to take leave from work so they have less chance of catching the virus as opposed to the region where people earn less and can't afford to take leave from work.

6. Homeless Population Urban
   The urban regions are mostly unhygienic and so a homeless person is most likely to be living in a unhygienic environment and so is more likely to get corona, on the other hand they may have developed enough immunity, hence this factor occurs in the middle of the list, as it does not affect too much.

7. Literacy Rate Urban
   The literacy rate in urban area's may not play such an important role as for a person to be called literate, he needs to know how to read and write. Whereas for taking precautions and knowing about COVID-19, takes only common sense and not the ability to read/write.

8. Population Density Rural
   As the population density in rural areas are already too low, so it does not affect the corona cases much.

9. Homeless Population Rural
   As compared to urban areas, the rural areas are more hygienic and so it comes lower in ranking than urban homeless population.

10. Literacy Rate Rural
    Same as what we observed in the literacy rate of urban areas, we can say that for a person to be called literate, he needs to know how to read and write. Whereas for taking precautions and knowing about COVID-19, takes only common sense and not the ability to read/write.

11. Unemployed Population Urban
    This is surprising that this factor is not having much effect on the corona spread. The reason might be that the people in urban regions usually live in a joint family, so the other family members may provide for the unemployed member.

12. Average Age Urban
    As this does not affect the covid-spread much, so we can conclude that the people of all age groups in urban area sociallize almost equally.

Now, let us look at the various models tried and their results.

1. Linear Regression:
   As this model can only form linear models, it is understandable that it will not perform well as the trend of COVID cases is of at least degree 2. This model is just predicting the best fit line of the data and will keep increasing the daily cases as the data it is trained on follows that trend. This is visible in the graphs.

2. Polynomial Regression:
   This is a non linear model. It did better than Linear regression during training but there were a few problems. Since, we are expecting 2 peaks(or

more) the degree of polynomial should be atleast 3. But, higher degree polynomials diverge quickly. So, we used a polynomial of degree 2.

3. Random Forest Regression:
   This model is able to follow the training curve, but is not able to generalise to the test data and is predicting a flat line. The reason for this is that there is no clear correlation between dates and daily covid cases which can be identified by a line. Another factor is that the values of the features like population density, number of hospital beds, etc. don't change over the dates and hence don't add any value leaving the prediction to be made based on just the date (which is not working due to reasons mentioned above).

4. Artificial Neural Network:
   Being a non-linear model with multiple feature extraction layers, it is performing better than all regression models which is as expected. However, a limitation of this model is that it can't predict a second peak and will just predict the cases in the direction that the graph was going previously. For now, its test error is coming better than expected as most districts have not experienced a second peak yet. But when cases start rising again, this model will fail. So, a better model which takes into account the data of just the previous few days is necessary.

5. LSTM:
   This model takes into account the data of the previous 10 days to make the prediction for the next day. Currently, it is performing almost same as ANN as the second peak has not started yet. But when it does, this model is expected to perform much better than ANN as it can use the previous few days data to predict that the cases will rise again while ANN will predict that cases will continue to fall as they were before.

   Thus, most of the results are coming up to be as we expected, with a few surprise results. Linear regression models just don't have the means to predict such a graph. Polynomial regression is not performing as expected and can be improved. ANN is fine for now, but will fail in the future if cases increase again while LSTM should give good results. But this needs to be verified as the situation progresses.

9 Future Direction

The avenues left to be explored in the project are the following.

1. Lack of proper data
   The data we have is a set of district based attributes which essentially are constants and don't change on a day to day basis. Thus there is no day to day varying data on the input that can be used as a predictor for the

output, i.e. the number of daily covid cases. We require some daily level data, e.g. district level daily testing data for better predictions.

2. Tuning of the hyperparameters
   The models vary greatly in their rms errors or predictive capability based on the hyperparameters, e.g. number of layers in neural networks, or number of neurons in each layer or the window of past values taken for predicting the current value in LSTM. Hence a more rigorous hyperparameter tuning might produce better results.

3. Better result from Polynomial Regression
   Polynomial Regression ended up giving results similar to Linear Regression. On increasing the degree, it resulted in overfitting on the training data which resulted in even worse predictions on the test data. This can be avoided by extensive 5-fold cross validation on the data. This will be facilitated again by the presence of larger amounts of data.

4. Difference between LSTM and ANN predictions
   Both fail to predict very accurate values of covid cases based on the current data. But while the ANN fits a non linear model which will predict cases, LSTM also takes into account temporal memory data. Hence with higher amounts of data, the second peak predicted by LSTM will be more accurate compared to that of ANN and there will emerge a distinctive difference between the performance of the two models.