BIRLA INSTITUTE OF TECHNOLOGY, MESRA

CA590 - MASSIVE OPEN ONLINE COURSE PROJECT

TOPIC :

Using polynomial regression for predicting the total imminent coronavirus cases in India

SUBMITTED TO:

RASHMI RATHI UPADHYAY
ASST. PROFESSOR (CSE)

SUBMITTED BY:

INDRANIL SARMAH
MCA/10011/19

`CONTENTS

Abstract

Corona virus disease or COVID-19 is a novel virus disease that started in the last year of 2019. The World Health Organization (WHO), on 11th March 2020, stated the outbreak of COVID-19 as a pandemic. Recently, especially during the last year, while this worldwide pandemic has persistently continued to affect the lives of millions, several countries have no other solution but to resort to total lockdown.

In the wake of COVID-19 disease, in this project based learning on MOOC. Looking forward to the circumstances I was obliged to design and develop a real time predictive model based on Machine Learning algorithm of regression to determine the upcoming affection risk of COVID-19. In this project, I have used a dataset of 'Our world in data , by OXFORD ' where it contains daily record of laboratory-confirmed COVID-19 patients from 207 countries around the world. But I have confined my project study to carry out a short-term projection of new cases and forecast the maximum number of active cases for India.

Related Study

Many Projects has already been done by using various artificial intelligence and machine learning model for diagnosing and predicting COVID-19 infection and recovery.

- In the work of data mining predictive model for COVID-19 patients recovery were developed. In the work of convolution neural networks that predict novel corona virus with x-ray images were developed.

- The deep learning technique, which is one of the sub-branches of ML, inspired by the structure of the human brain is used for the automatic prediction of 2019-nCoV patients.

- In the work of machine learning studies were carried out to predict 2019- nCoV incidence by levering Google trend data in India. Linear Regression Models were used to estimate the number of 2019-nCoV positive cases.

The related works that have been reviewed so far indicate that ML techniques and other artificial intelligence techniques have played important roles in prediction, diagnosis and containment of the COVID-19 pandemic, which can help reduce the huge burden on limited health care systems.

Introduction

This project is about developing a predictive model for total upcoming corona virus cases in India. As, SARS-Cov-2, emerged on the horizon in late December 2019 when few local health authorities in China reported clusters of patients with pneumonia of unknown cause. The surveillance mechanism established during 2003 SARS outbreak helped in identification of the pathogen (SARS-CoV-2).

The SARS-CoV-2 infection has spread in 210 countries as of 24 April 2020 with 2,697,316 cases and 188,857 fatalities.

In India, the first case was reported on 30 January 2020, and the numbers gradually increased till 03 March after which the per day increase has been faster.

So it was an essential task to prevent the spread of Corona virus further. To stop or reduce the spread of Corona virus, across the nation various steps have been taken like announcing the lockdown, isolation, maintain distancing etc.

Researchers had done several studies with the available set of data of COVID-19 spread in India and other countries. Some of the notable works have used several other statistical approaches while some have used compartmental studies to model and predict the viral spread.

Regarding to it, this project is also an approach for developing a Machine learning mathematical model based on regression where it allows to predict a continuous outcome variable of upcoming corona virus cases based on daily affected cases.

This project is divided into three phases :

a) Data collection

b) Prepossessing the data

c) Building the model for prediction.

d) Testing

Dataset

Effectiveness of forecasting is based upon the quality of data source used for forecasting. For getting a better result for the model data is been collected from the authentic source of '**Our world in data by OXFORD Martin school, university of OXFORD , Global change data lab**'.

As For a comprehensive assessment, this dataset track the impact of the pandemic across for 207 countries to study in depth the statistics on the coronavirus pandemic for every country in the world including India**.**

Methodology

Polynomial Regression

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$\mathbf{y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+...... b_nx_1^n}$$

- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

- It is a linear model with some modification in order to increase the accuracy.

- The dataset used in Polynomial regression for training is of non-linear nature.

- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.

- Hence, **"In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."**

Need for Polynomial Regression

The need of Polynomial Regression for this project can be understood in the below points:

- If a linear model on a **linear dataset** is applied, then it provides a good result in Simple Linear Regression, but if the same model is applied without any modification on a **non-linear dataset**, then it produces a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.

- So for such cases, **where data points are arranged in a non-linear fashion, there is a need of Polynomial Regression model**. To obtain a clear view the of below diagram is been depicted.

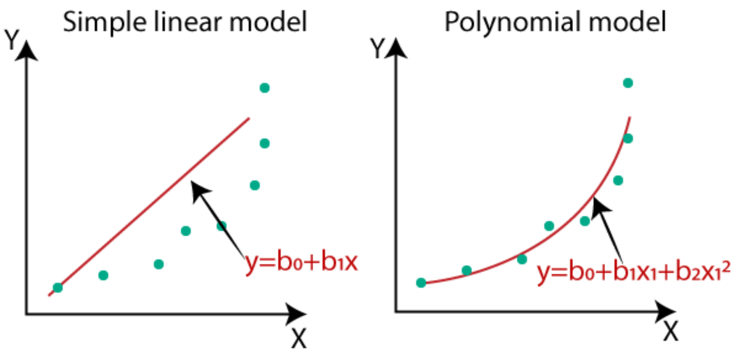


Figure : Comparison between linear and Polynomial regression

- In the above image, a dataset is taken which is arranged non-linearly. So to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.

Referring to this concept , In this project also all the data related to total cases of affected coronavirus patients are in non-linear fashion that is why polynomial regression suits the best to predict the upcoming total cases.

Workflow and Implementation

- **Data collection and pre-processing**

The (.csv) file containing all the related data has been downloaded and imported project using pandas

```
import pandas as pd
df = pd.read_csv("owid-covid-data.csv")
```

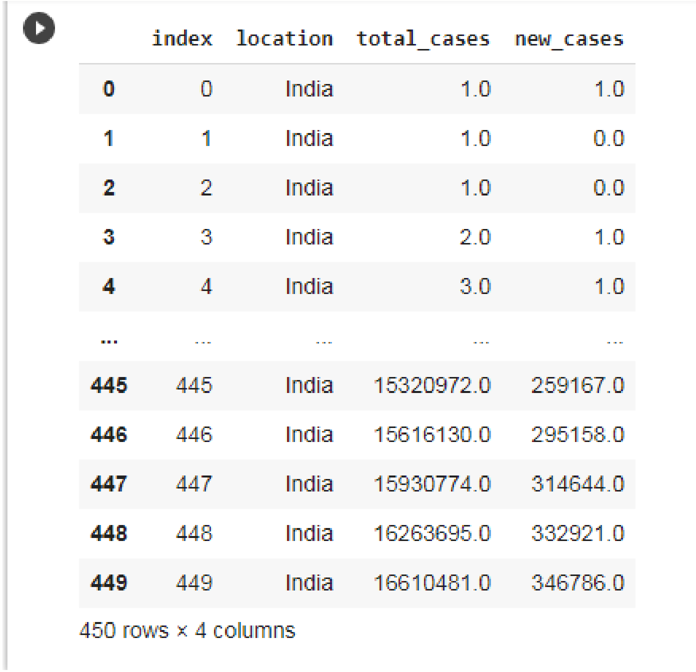
The dataframe (df) contains all the data of all the countries, but this project confined to India So, total cases with respect to number of days is given by ,

```
df = df.loc[df["location"].isin(["India"])]
```

```
df = df[["location","total_cases","new_cases"]]
```

Resetting , the index as number of days

```
df = df.reset_index()
df["index"]=df.index
```

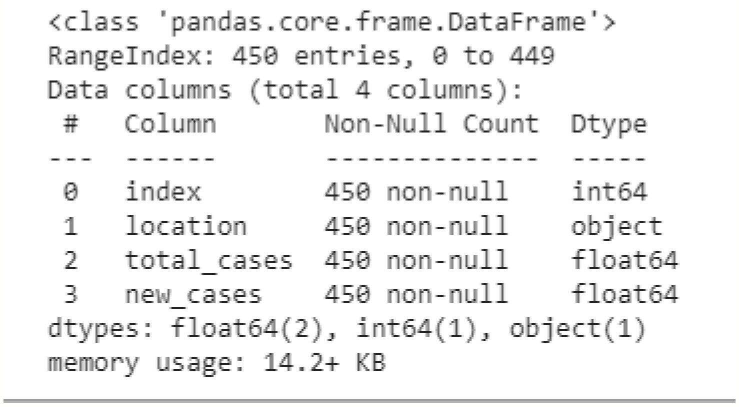


	index	location	total_cases	new_cases
0	0	India	1.0	1.0
1	1	India	1.0	0.0
2	2	India	1.0	0.0
3	3	India	2.0	1.0
4	4	India	3.0	1.0
...	...	...	...	...
445	445	India	15320972.0	259167.0
446	446	India	15616130.0	295158.0
447	447	India	15930774.0	314644.0
448	448	India	16263695.0	332921.0
449	449	India	16610481.0	346786.0

450 rows × 4 columns

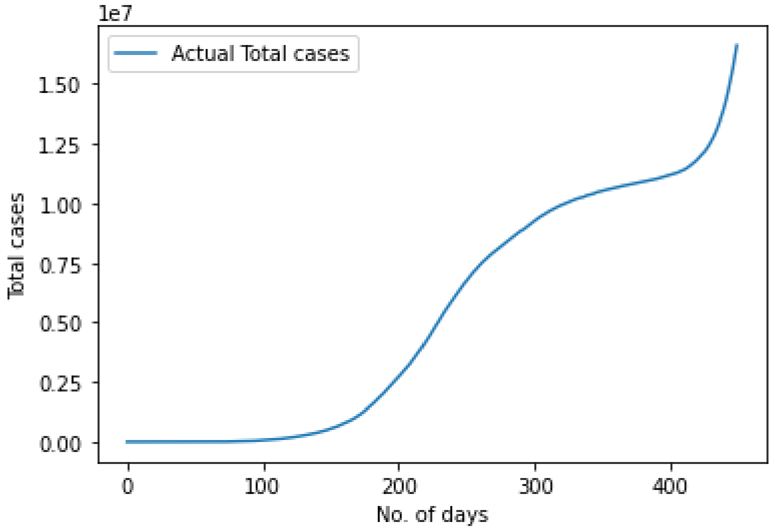
Checking for any missing data and then Plotting(using matplotlib) with : no of days vs total cases

```
df.info()
```

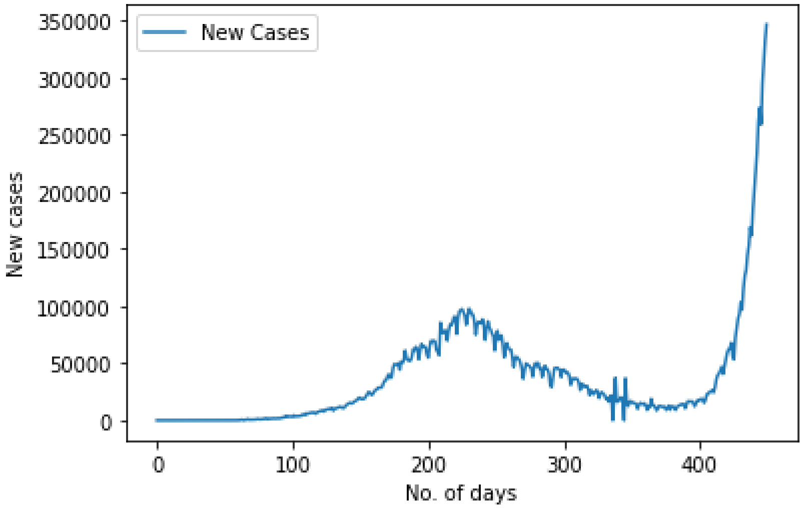


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 450 entries, 0 to 449
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   index        450 non-null    int64  
 1   location     450 non-null    object 
 2   total_cases  450 non-null    float64
 3   new_cases    450 non-null    float64
dtypes: float64(2), int64(1), object(1)
memory usage: 14.2+ KB
```

```
import matplotlib.pyplot as plt
plt.plot(df["total_cases"],label = "Actual Total cases")
plt.xlabel("No. of days")
plt.ylabel("Total cases")
plt.legend()
plt.show()
```



Also for reference plotting the daily new cases



- **Building the model**

First the Simple Linear model is created using lin_reg object of **LinearRegression** class and then for generating Polynomial Regression model, **PolynomialFeatures** class is been used of **preprocessing** library.

Moreover , X is number of days (index)

y is total_cases

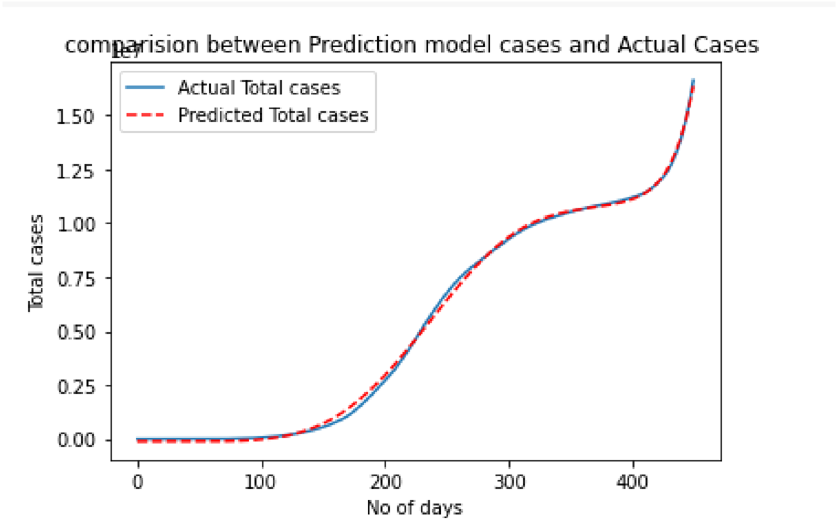
```
from sklearn.preprocessing import  PolynomialFeatures
from sklearn.linear_model import LinearRegression
import numpy as np

X = np.array(df["index"]).reshape(-1,1)
y = np.array(df["total_cases"]).reshape(-1,1)

poly = PolynomialFeatures(degree = 6)
X = poly.fit_transform(X)

lin_reg = LinearRegression()
lin_reg.fit(X,y)
lin_reg.score(X,y)
```

In the above lines of code, **poly.fit_transform(x)** , is used because first we are converting our feature matrix into polynomial feature matrix, and then fitting it to the Polynomial regression model. The parameter value(degree = 11) depends on our choice. It can be choosen according to requirement of our Polynomial features.

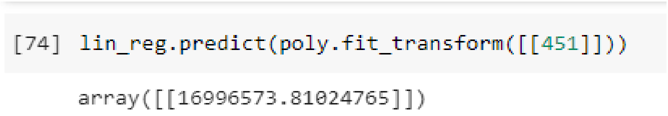
At last visualizing ,

```
plt.plot(df["total_cases"],label = "Actual Total cases")
plt.plot(lin_reg.predict(X),"r--",label = "Predicted Total cases")
plt.title("comparision between Prediction model cases and Actual Cases ")
plt.xlabel("No of days")
plt.ylabel("Total cases")
plt.legend()
plt.show()
```

Final result

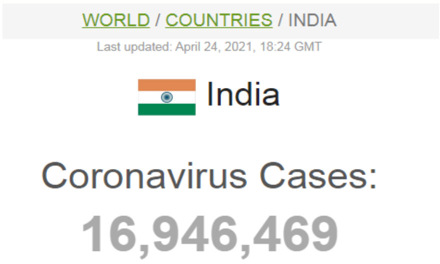


Tested for data for 451th day 24-April 2021,



```
[74] lin_reg.predict(poly.fit_transform([[451]]))

     array([[16996573.81024765]])
```

And match with live corona virus status of world meter which yields approximately same value



WORLD / COUNTRIES / INDIA

Last updated: April 24, 2021, 18:24 GMT

India

Coronavirus Cases:

16,946,469

Applications of this model

Although prediction of the COVID-19 pandemic may be inevitably accompanied by uncertainty, but it may be useful for

- Health care system for future management
- Government decision-makers to plan and manage the outbreak of COVID-19
- Economist to maintain and hold economy of the country during crisis.
- Business makers to foresee and plan for the upcoming risks.

Tools and Libraries used for the project

Language : python

Platform : Google colaboratory , Jupyter notebook

Version control : GitHub

Libraries : pandas, matplotlib , scikit-learn , numpy

References

NPTEL :

Machine Learning, ML By Prof. Carl Gustaf Jansson
https://onlinecourses.nptel.ac.in/noc21_cs51/preview

Research Paper :

Predictive models of COVID-19 in India: A rapid review :
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7298493/

Forecasting COVID-19 epidemic in
Indiahttps://www.sciencedirect.com/science/article/pii/S2213398420301639

Website reference :

https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

****