RESEARCH

Introduction to The Profile Areas in Data Science: Project 6

Esra Gücükbel*, Emilio Kuhlmann and Weronika Kłos

*Correspondence:
esrag95@zedat.fu-berlin.de
Full list of author information is available at the end of the article

Abstract

Goal: The goal of this project is to implement the decomposition model and auto-regressive models and by using these models perform the forecasting for last two weeks.

Main results: We achieved best results with auto-regressive model which has 0.79% mean absolute percentage error.

Key learnings: We learned how to analyze time series and make prediction with 3 different models such as decomposition, auto-regressive and lstm networks.

Estimated working hours: 6 per person

Project evaluation: 1.5

Number of words: 1622

Keywords: covid-19; time-series analysis; decomposition

1 Scientific Background

COVID-19 which is from Orthocoronavirinae subfamily, different from both MERS-CoV and SARS-CoV, 2019-nCoV is the seventh member of the family of corona viruses that infect humans is a contagious disease. In December 2019, the first case was diagnosed in Wuhan, China. It spreads mainly close contact with infected person. Air particles and aerosols containing virus from an infected person's nose, mouth, breath carry disease and cause infection. The mode of transmission accelerates the increasing number of patients. The incubation period of COVID-19, which is the time between exposure to the virus and symptom onset, is on average 5-6 days, but can be as long as 14 days. The ongoing disease, in other words global pandemic which was declared by WHO in March, was spread to whole world and currently 69 million of people became infected and 1.5 million of people died due to this disease.[1]

In Germany, the first case was confirmed near Munich, Bavaria on 27 January 2020.[2] In Baden-Württemberg, many cases related to outbreak in Italy were detected on 25 and 26 February 2020. As of this date German states impose restrictions, lock downs and curfews. By mid-April, these restrictions were gradually eased. The succeeding summer and holiday season caused increment of cases. By mid-October, second wave of pandemic began. In order to prevent spreading the virus again, partial lockdown rules such as allowing only take-away services, closing clubs and entertainment facilities, continuing education in school by following the distance rules will be extended twice and will be applied until January 10, 2021.

As of 9 December 2020, the RKI has officially reported 1,218,524 cases, 19,932 deaths.[3]

In the present circumstances, making projections over the number of cases for next weeks is crucial to be prepared the needs of bed, respiratory device, medicine, etc. Constructing model by using statistical and machine learning methods and providing forecast of possible numbers for next days can assist the medical systems and governments.

2 Goal

The goal of this project is to implement the decomposition model of Facebook's Prophet library and an auto-regressive model on the dataset of Germany's covid-19 case numbers and by using these models to perform forecasting for last two weeks in the dataset.

3 Data

In this project 'cases-rki-by-state.csv' file was used. [4] The file consists state by state accumulated daily case numbers of Germany on the date range 02/03/2020 and 08/12/2020. There are 282 rows and 18 columns, first one is date, last one is sum of the case number for all states, and others are the case number for each state. The data was provided by Robert Koch Institute in every day. We downloaded the file on 9^{th} of December and all operations were applied to this data. Figure 1 shows the daily new cases state by state. All analysis and implementations were applied to 'sum_cases' column which consists the sum of the case number for all over Germany.

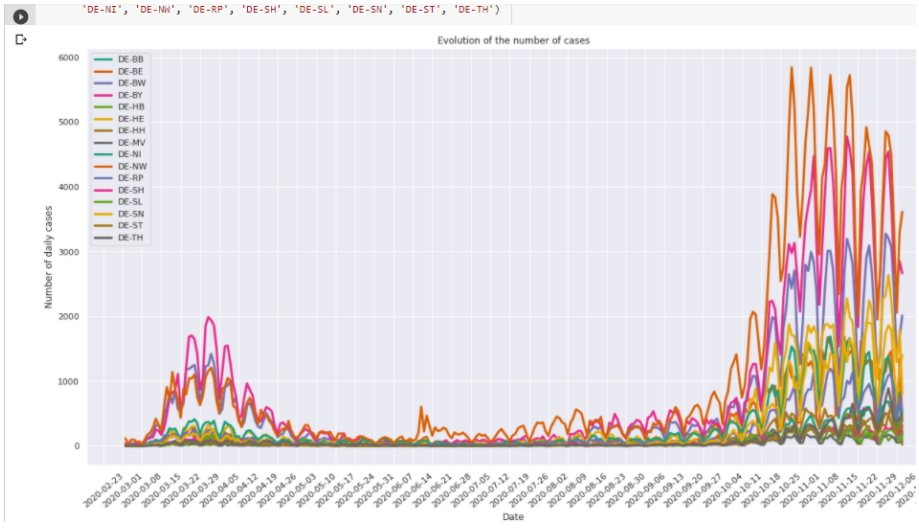

Figure 1 Screenshot of the number of daily cases.

4 Methods and Results

We used two different approaches to perform time-series analysis of the given dataset: an additive decomposition model, which is the default in facebook's *Prophet* library and an auto-regressive model. A first look at the data in figure 2 already

makes clear that there is a strong trend which can be exploited for forecasting. However, there seem to be no seasonal aspects in the data, at least not in this time-window. Before starting the analysis we have to split it in a training and testing set. While the whole dataset contains data from 282 days, we chose the last two weeks (14 days) to be the test set, leaving 268 days for the training set.

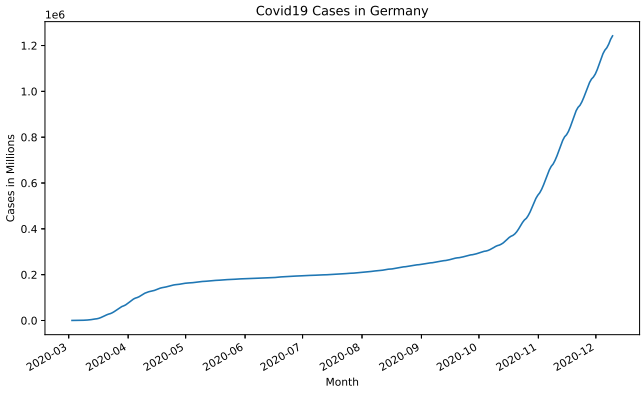

Figure 2 Number of Covid-19 cases in Germany.

Decomposition Model

The additive decomposition model assumes that the observed data can be described by a sum of three different components: a trend, a seasonal component and an unpredictable white noise component. This model, that we fitted to the training set, can then be used for future prediction. The three components of the fitted model can be seen below in Figure 3. There is a clear upward trend, with a strong and abrupt increase in slope in October 2020. There also seems to be a weekly seasonality, showing fewest cases on Tuesdays, as well as a daily seasonality with a peak of cases around 3pm.

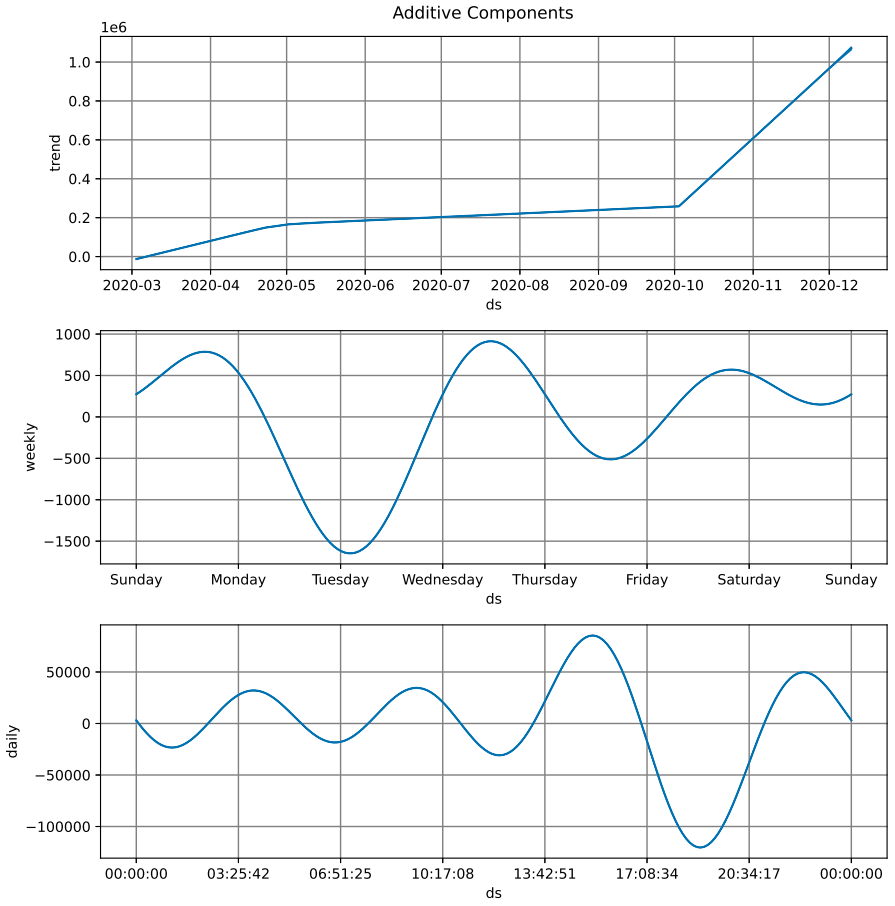

Figure 3 Components of the Additive Model.

The result of fitting the model to the Covid-19 data and predicting the number of cases for the last two weeks is depicted below, in Figure 4. The mean squared error on the test set that we get by this technique is 20184421559 (around 20 billion).

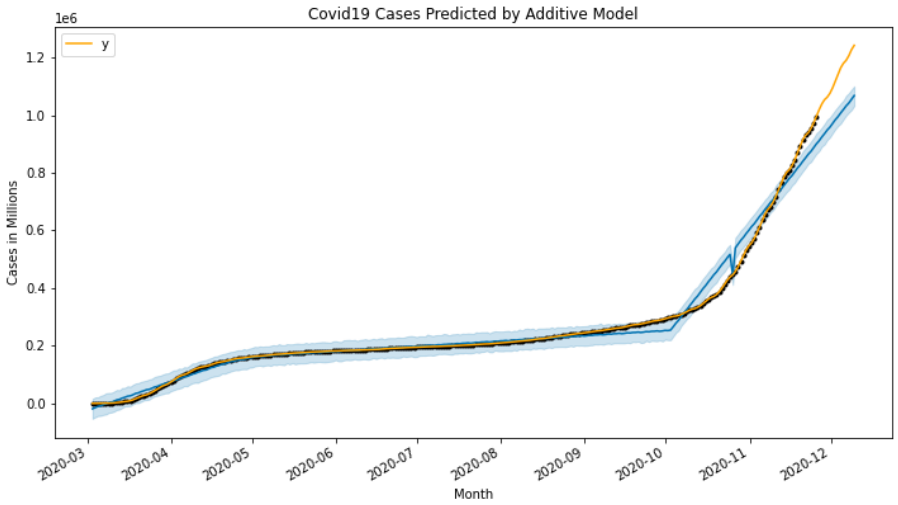

Figure 4 Prediction of the Additive Model.

Auto-Regressive Model

An auto-regressive model of order p (often denoted as AR(p)) assumes that the next value x_t of a given time-series can be expressed as a weighted sum of p previous values $x_{t-p}, ..., x_{t-1}$ (p is the number of lags). The weights are usually estimated by simple linear regression. We trained the AR(p) model with p-values ranging from 1 to 30, and observed that a p-value of 18 produces the best results - a mean-squared error of 37390542 (around 37 million) on the test set to be precise. An implementation from the *statsmodels.tsa.ar_model* package was used. A visualization of the fit and the prediction is depicted below in Figure 5 and 6.

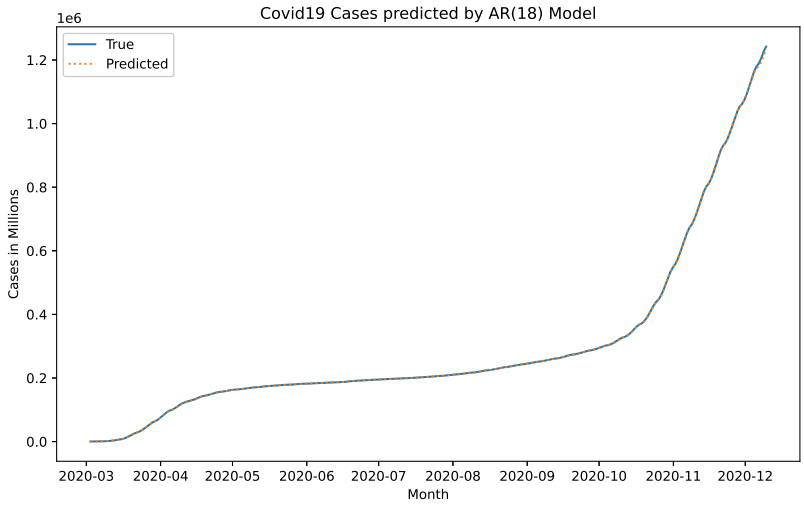

Figure 5 Prediction of the AR(18) Model.

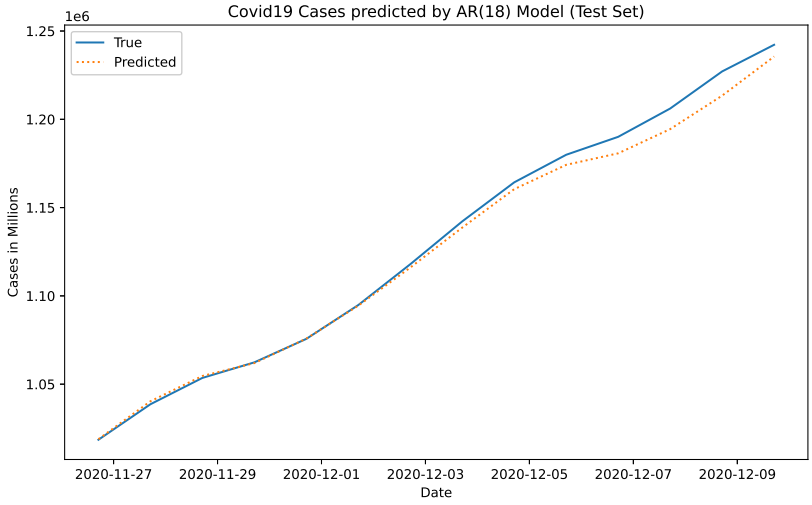

Figure 6 Prediction of the AR(18) Model on the Test Set.

Comparison of Decomposition and Auto-Regressive Model

First we need to note that the daily seasonality doesn't make much sense in this case as the new data carries the daily timestamp at 17:00:00. The Prophet library introduces this daily seasonality with something similar to a sin function, this could be seen as white-noise. It looks very regular except around 17:00:00. Upon closer inspection of the data, we noticed that there is a single timestamp that carries the time 18:00:00, namely the 25^{th} of October. This was likely a mistake, however it

clearly has an impact as the seasonality clearly goes down at 18:00. The Prophet library is trying to interpolate the mistake leading to some unwanted consequences.

The weekly seasonality doesn't make that much sense either, given that the Coronavirus doesn't care about weekdays. A factor could be that people are more likely to get tested on some weekdays than others, for example on the weekend they might have more time to go get tested. Why, however, cases seem to be lowest on Tuesday and highest on Wednesday we cannot explain.

The trend looks like what one has come to expect from the news. Initial steep incline, followed by a long plateau after the successful lockdown measures, followed by a worryingly steep incline after a high percentage of the population stopped caring about the virus.

On to the comparison: The Auto-Regressive model is much much better. This is hardly surprising given that predicting a pandemic is a textbook regression example. The Decomposition model is very squared off and the only usable data is the trend, as we already explained that the seasonality doesn't make much sense here. Perhaps if we wait a few more years we will have seasonality data on the actual seasons of the year and it might show that there are more cases in winter.

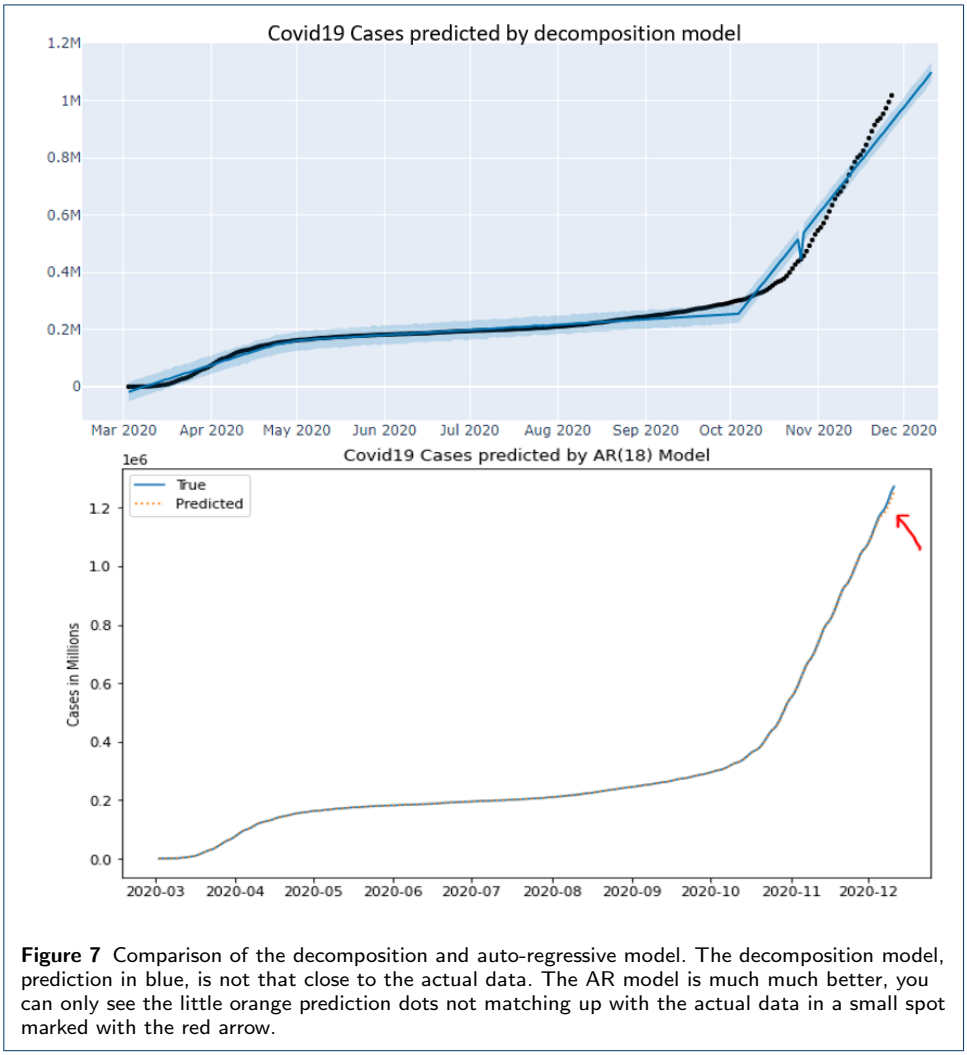

Figure 7 Comparison of the decomposition and auto-regressive model. The decomposition model, prediction in blue, is not that close to the actual data. The AR model is much much better, you can only see the little orange prediction dots not matching up with the actual data in a small spot marked with the red arrow.

Change-point Analysis of Decomposition model

Let us now look at the change-points of the data. First, however, let us set some expectations based on intuition. We expect there to be a change-point in May, after the successful lockdown measures and a change-point in October, when the numbers exploded again.

Looking at Figure 7, our expectations are confirmed. We get the two main change-points in May and October plus a few minor ones depending on the number of change-points we required from the Prophet library.

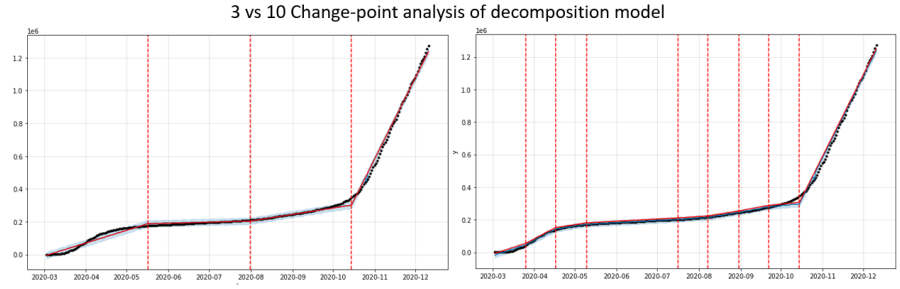

Figure 8 The change-points indeed line up with expectations.

LSTM(Long Short Term Memory) Model

Apart from auto-regressive and decomposition models, we used the data(sum of the case number - sum_cases column) at daily level. By subtracting the accumulated values of each day from previous day's value, daily new cases were calculated for each day and the model was built on daily new cases. First 254 row were used in training, since before applying LSTM, time series were created on previous and next 14 days time windows for each day. Test set includes the data between 255^{th} and 282^{th} row. By dividing the data with this method, we could predict last 14 days. LSTM model includes one node layer with 512 nodes with and 1 dense layer. Adam was used optimizer and ReLu as activation function on the node layer. The mean absolute percentage error of train set is 32.63% and test set is 14.37%. R^2 value of test set is 0.73. This implementation was performed by inspiring the notebook in Kaggle platform.[5]

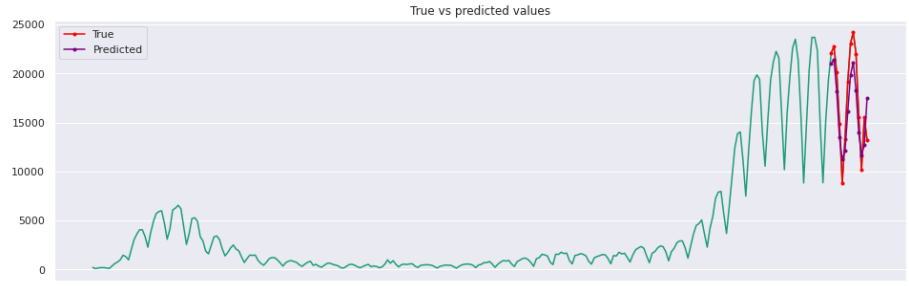

Figure 9 Prediction of the LSTM Model

5 Discussion

We analyzed the data and applied 3 different methods. The achieved the results is given below. It shows us the most successful model in terms of MAPE metric is auto-regressive(18).

Model	MAPE
Decomposition	10.44%
Auto-Regressive	0.79%
LSTM	14.37%

This project is a typical project for a data scientist because numerous real world problems can be formulated as time series data, and it's useful to have at least the basics (ARIMA models, some of machine learning models). In finance, marketing, and other time-series based applications, it's critical. The data that we used was easy to analyze and implement over models. Moreover this project is a real-world scenario for nowadays. We were pleased with carrying out this project.

Appendix

Authors' contributions

Esra Gücükbel performed LSTM model and wrote the scientific background, goal, data sections. Weronika Kłos implemented the time-series analysis, decomposition model of prophet library and performed the forecasting for the decomposition and auto-regressive models and wrote the respective parts. Emilio Kuhlmann implemented the change point analysis and wrote the comparison of the models required in task 3.

Author details

References

1. Worldometer, Covid-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/
2. Wikipedia, Covid-19 Pandemic in Germany. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Germany
3. RKI, Covid-19 Dashboard. https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4
4. Github, Covid-19 Germany GAE. https://github.com/jgehrcke/covid-19-germany-gae/blob/master/cases-rki-by-state.csv
5. Kaggle Notebook, LSTM Method. https://www.kaggle.com/lagors/time-series-analysis-on-cov19-outbreak