Time Series Forecasting of Daily and Cumulative Cases of Covid-19
using Covid-19 Tweets as an External Regressor

Introduction

The objective of this project is to implement a forecasting model to predict daily and cumulative Covid-19 Cases in the US, using Covid-19 related tweets as an external regressor. Our analysis involves using two time series forecasting models, we use Facebook Prophet to predict the daily Covid-19 Cases in the US using daily tweet counts as the external regressor, and we use SARIMAX to predict the cumulative Covid-19 Cases in the US using the tweet counts of different sentiments (Positive, Negative, and Neutral) as exogenous variables.

Methodology

a) Prediction of Daily Cases using Fb Prophet

Facebook Prophet is an Open-Source Time Series Forecasting Model made by Facebook, which can predict and forecast linear and non-linear trends. It is an additive model which can fit non-linear trends with yearly, weekly and daily seasonality along with Holiday Effects. Prophet works best with time series which has strong seasonal effects, which is why Daily Covid-19 Cases can be implemented better in Facebook Prophet, instead of SARIMAX. Prophet is also Robust to Outliers and works well with missing data.

From the CDC Dataset, we use only the confirmed cases grouped together for the US in our analysis, where the model is input with a 7-day moving average of the daily case counts to smoothen out the fluctuations in daily Covid-19 cases. Here, the model also receives a daily count of Covid-19 related tweets as an additional regressor. The data for Twitter related tweets is sourced from the Panacea Lab which has a global tweets dataset related to Covid-19 (Keywords used: COVD19, CoronavirusPandemic, COVID-19, 2019nCoV, CoronaOutbreak,coronavirus, WuhanVirus, covid19, coronaviruspandemic, covid-19, 2019ncov, coronaoutbreak, wuhanvirus). We have filtered the global dataset to only tweets in the US and in English, giving us a dataset of 2.15 million tweets from the 21st of January 2020 to 17th of October 2021.

b) Prediction of Cumulative Cases using SARIMAX

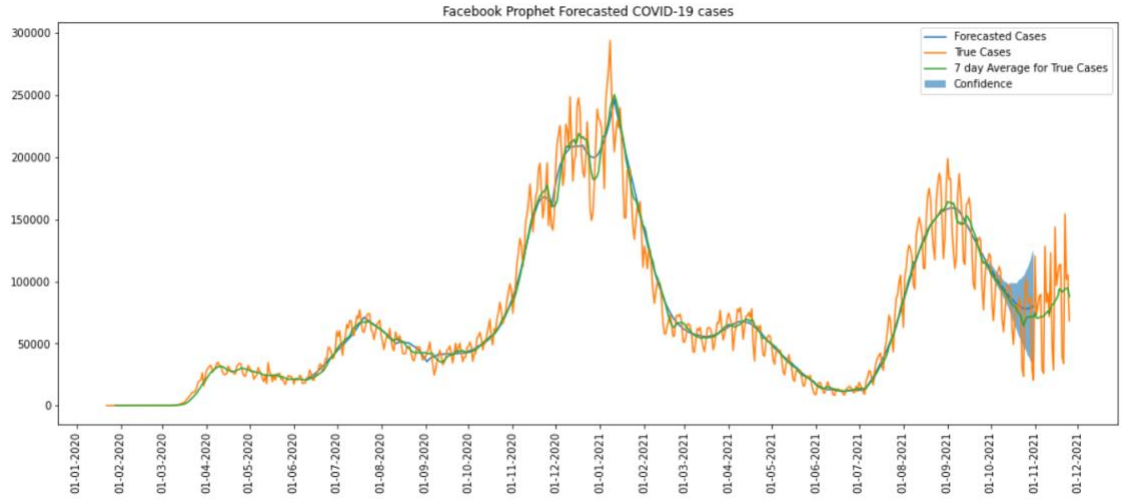
SARIMAX belongs to a class of models that explains a given time series based on its own past values (its own lags and the lagged forecast errors). It adds seasonal components and exogenous variables based on ARIMA. Any time series that exhibits a pattern and is not random white noise can be modeled using this type of model. This type of model is specified by 3 order parameters (p, d and q). In this project, in order to select these values, our model will use the pmdarima library for Python to perform a grid search over multiple values of p, d and q. The model will be trained according to the combination of parameter values, then some performance criteria such as AIC and BIC will be calculated and compared. Finally, the model with the lowest (best) AIC score will be selected for prediction.

In order to further improve the accuracy of the model, we also considered adding Twitter sentiment as exogenous variables to SARIMAX. The dataset from OPENICPSR contains US tweet posts from January 28, 2020 to September 1, 2021, and the keywords covered include “corona”, “wuhan”, “nCov” and “covid”. Topic modeling techniques and pre-trained machine learning-based emotion analytic algorithms are used to label each tweet with qualitative sentiment attributes. By analyzing the relationship between sentiment data and Covid-19 data, we found that the relationship between different sentiments and different types of cases shows different characteristics as the epidemic develops. To assist decision-making, we will use the number of tweets with different sentiments as exogenous variables in our model.

Result

a) FbProphet

Facebook Prophet is initialized with a 7-day moving average of daily covid cases and daily covid related tweet counts (as an external regressor) from the 22nd of January 2020 to 30th of September 2021. The model predicts an output with a given Confidence interval for the entire month of October 2021. We observe that the Confidence interval for predicted Cases increases as the nth day to forecast also increases. The model provides an R^2 value of 0.79 with the following output:



b) SARIMAX

We selected 95% of the data as the training set to predict the cumulative Covid-19 confirmed cases for the next 30 days. The auto_arima is used to quickly perform a grid search of p, d, and q values and select model based on AIC. Compared with ARIMA and SARIMA models, SARIMAX adds the tweet counts with different sentiments as exogenous variables and obtains the smallest RMSE value 379512.41. The prediction results are as follows：