Covid 19 Confirmed Cases 90 Day Forecast

Welamaza M

1/18/2021

In this project we use a model to predict how covid confirmed cases will look for next 3 months in the US. According to the Centers for Disease Control and Prevention (CDC) , a confirmed case “is an individual who had a confirmatory viral test performed by was of throat swab, nose swab or saliva and that specimen tested positive for SARS-CoV-2, which is the virus that casues COVID-19”. We explore different R packages including: covid19.analytics, dplyr, prophet, lubridate, ggplot and more packages to forcast how the curve will look.

Dataset: We will be working at live data from reported covid19 cases specifically the confirmed cases, the data is collected by John Hopkins University. The dataset can be found in the “covid19.analytics” package

Load required R packages:

```
library(covid19.analytics)
library(dplyr)
library(prophet)
library(lubridate)
library(ggplot2)
library(sna)
library(openair)
library(kableExtra)
```

Read data

```
tsc <- covid19.data(case = 'ts-confirmed')
```

```
## ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## --------------------------------------------------------------------------------
```

```
dl <- tail(tsc[,1:10])
```

```
kable(dl)%>%
  kable_styling(bootstrap_options = "striped", font_size = 15)
```

	Province.State	Country.Region	Lat	Long	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	2020-01-27
267		Venezuela	6.42380	-66.58970	0	0	0	0	0	0
268		Vietnam	14.05832	108.27720	0	2	2	2	2	2
269		West Bank and Gaza	31.95220	35.23320	0	0	0	0	0	0
270		Yemen	15.55273	48.51639	0	0	0	0	0	0
271		Zambia	-13.13390	27.84933	0	0	0	0	0	0
272		Zimbabwe	-19.01544	29.15486	0	0	0	0	0	0

Filter for the US Data

```
tsc <- tsc %>% filter(Country.Region == 'US')
```

Transpose, bind columns, rename headers and remove first 4 rows - they do no conatin useful data

```
tsc <- data.frame(t(tsc)) # to transpose the data
tsc <- cbind(rownames(tsc), data.frame(tsc, row.names = NULL))
colnames(tsc) <- c('Date', 'Confirmed')
tsc <- tsc[-c(1:4),] # remove first 4 rows

dh <- head(tsc)
kable(dh)%>%
  kable_styling(bootstrap_options = "striped", font_size = 15)
```

	Date	Confirmed
5	2020-01-22	1
6	2020-01-23	1
7	2020-01-24	2
8	2020-01-25	2
9	2020-01-26	5
10	2020-01-27	5

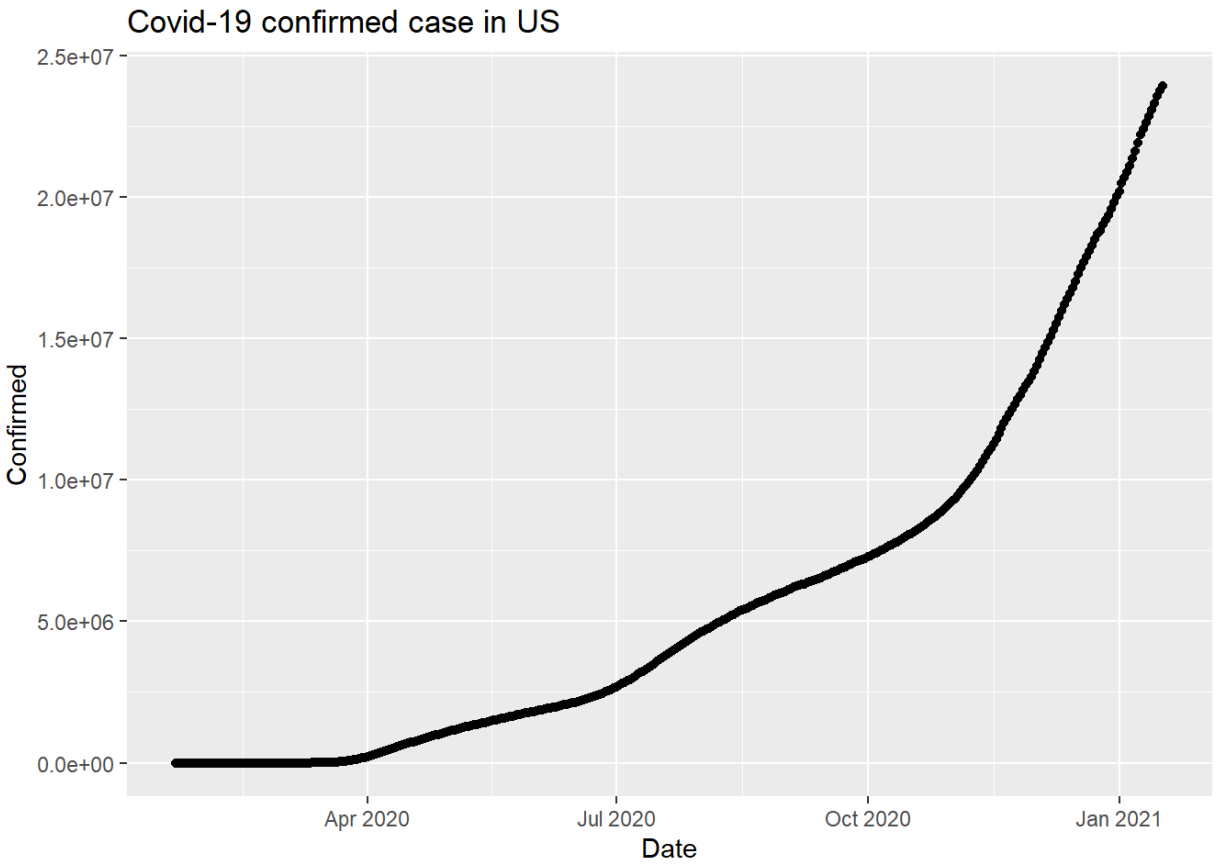
In order to convert a factor to numeric, convert to character first

```
tsc$Confirmed <- as.numeric(as.character(tsc$Confirmed))
tsc$Date <- ymd(tsc$Date) #ensure date is in correct format
str(tsc)
```

```
## 'data.frame':    362 obs. of  2 variables:
##  $ Date     : Date, format: "2020-01-22" "2020-01-23" ...
##  $ Confirmed: num  1 1 2 2 5 5 5 6 6 8 ...
```

Plot line graph showing confirmed cases, the y-axis are cumulative

```
qplot(Date, Confirmed, data = tsc,
      main = 'Covid-19 confirmed case in US')
```

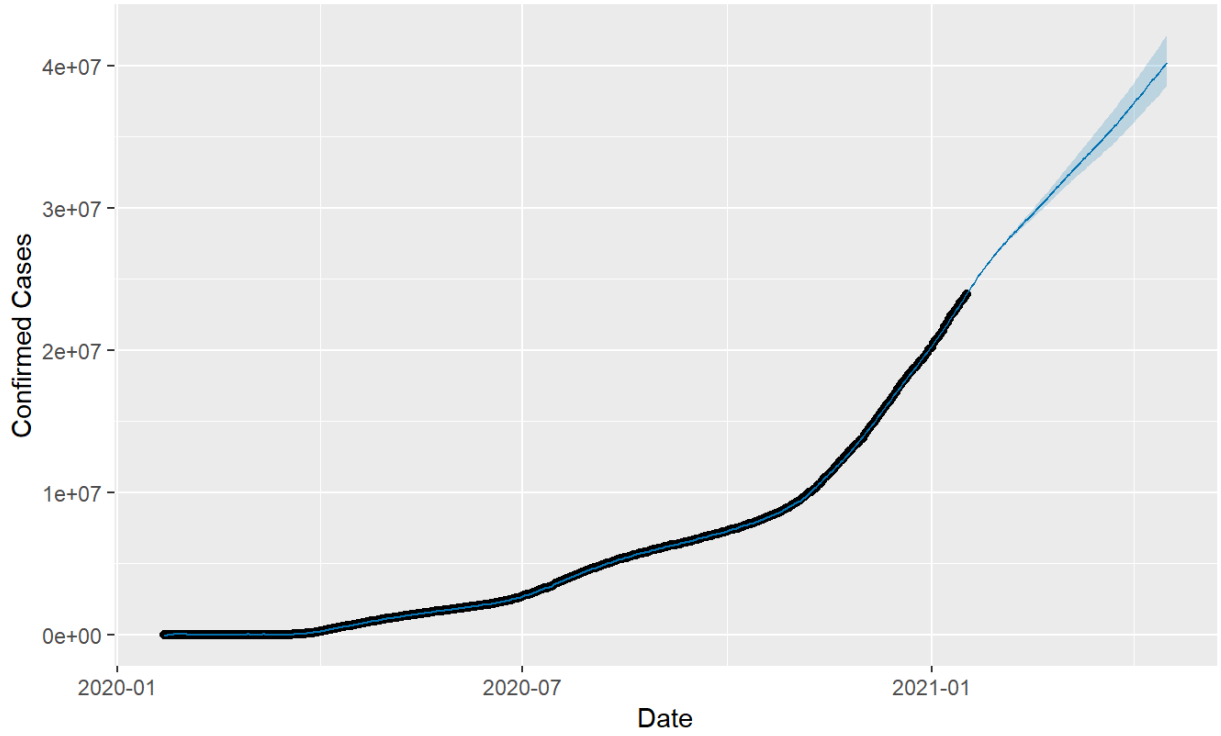


Prepare for forecasting: store dates in ds, cases confirmed in y and make date frame with both

```
ds <- tsc$Date
y <- tsc$Confirmed
df <- data.frame(ds, y)
```

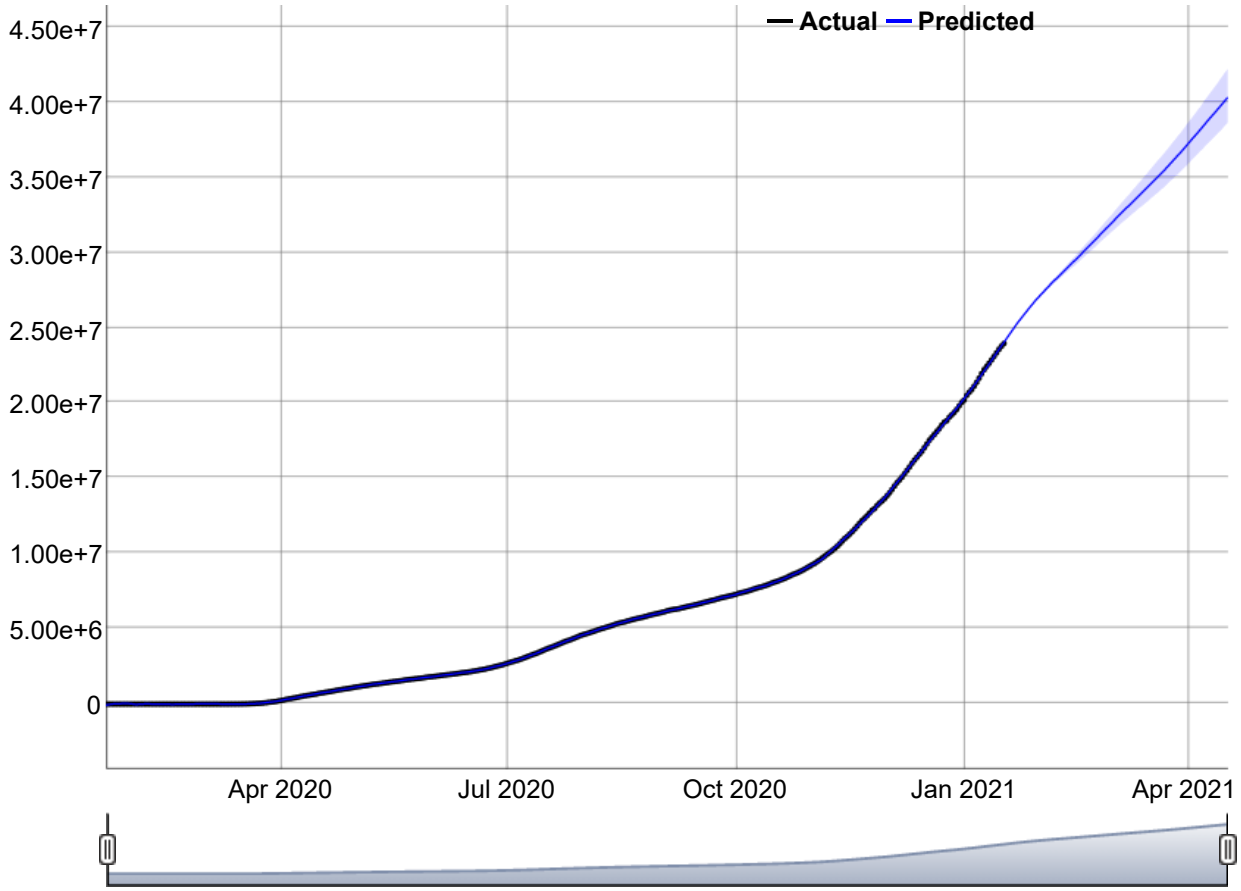
Use function prophet to forecast and make a new data frame that spans 3 months into the future. Make new variable with predictions and plot.The shaded blue around the line shows the confidence interval, if nothing changes confirmed cases should lie in between the interval. With vaccine in action we should expect to see the curve flatten steadily.

```
# Forcasting
m <- prophet(df, yearly.seasonality = TRUE)

# Prediction
future <- make_future_dataframe(m, periods = 90) #for future dates

forecast <- predict(m, future)

plot(m, forecast,
     xlab = "Date",
     ylab =  'Confirmed Cases')
```



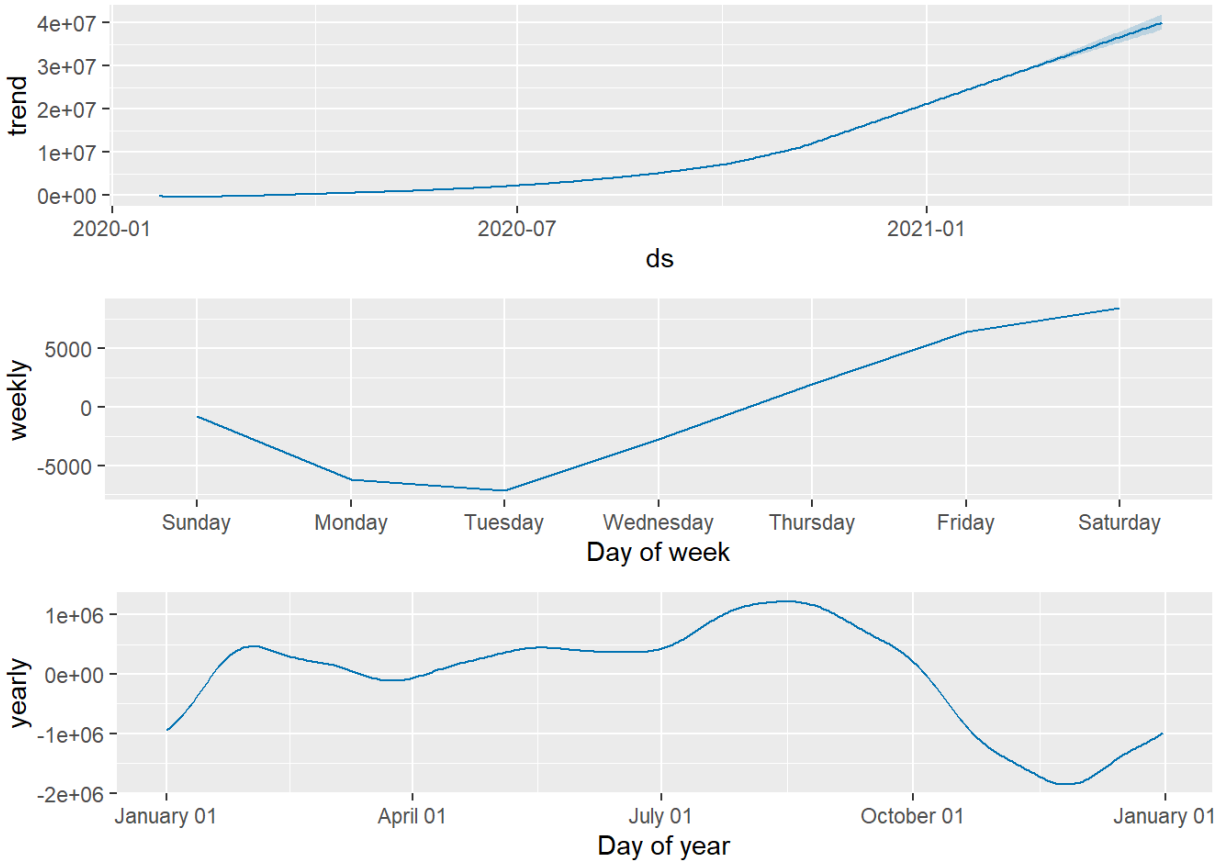
For interactive plot. Predictions show we can reach as high as 4 million cases in April.

```
dyplot.prophet(m, forecast)
```



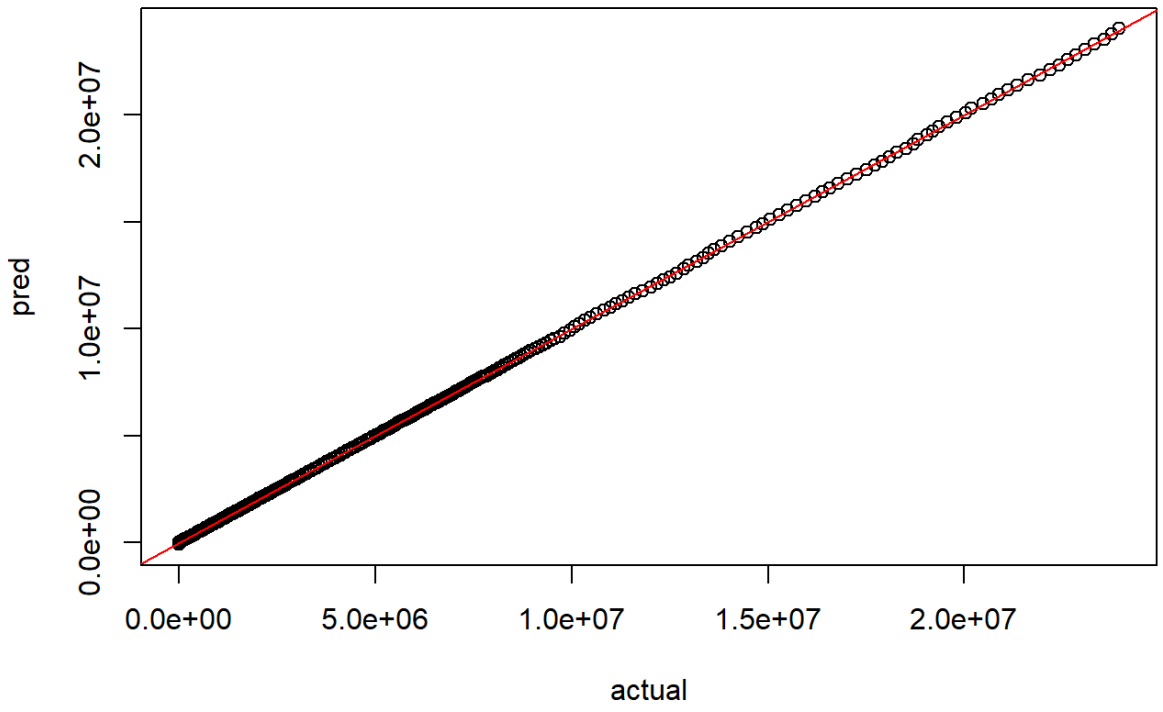
Graphs below show trends of average confirmed case yearly, weekly and daily

```
prophet_plot_components(m, forecast)
```



Next evaluate the accuracy of model.

```
pred <- forecast$yhat[1:362]
actual <- m$history$y
plot(actual,pred)
model1 <- lm(pred ~ actual)

abline(lm(model1),col= 'red')
```



R is squared equal to 1 and p-value is very small thus we have high confidence that the model is statistically significant.

```
summary(model1)
```

```
##
## Call:
## lm(formula = pred ~ actual)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93546  -6682     49   7435 107607
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.124e+02  1.827e+03    0.062    0.951
## actual      1.000e+00  2.148e-04 4654.822   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25410 on 360 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.167e+07 on 1 and 360 DF,  p-value: < 2.2e-16
```