Boston Crime Forecasting: Seasonal ARIMA and Prophet models

Natalia Gust-Bardon

October 2019

Preparing Our Data

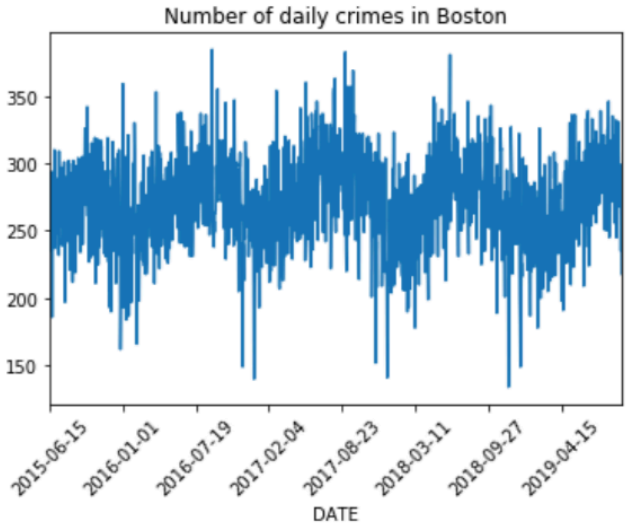
- We will use the Crime Incident Reports from Boston Police Department (Download)
- The data set requires transformation
- We will prepare the training and test sets

Goal of the Project

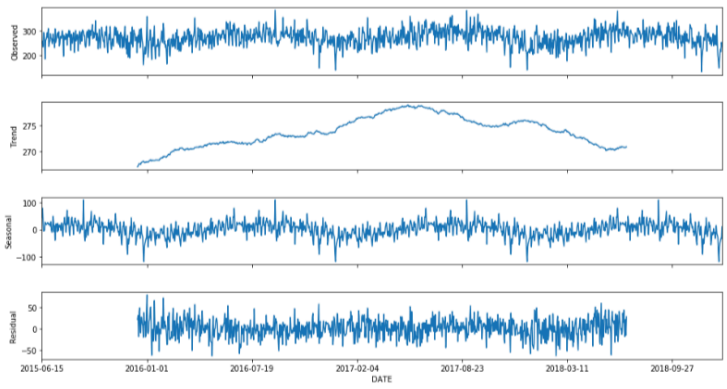
- Our goal is to find the most accurate model predicting the number of daily crimes in Boston

Time Series Data

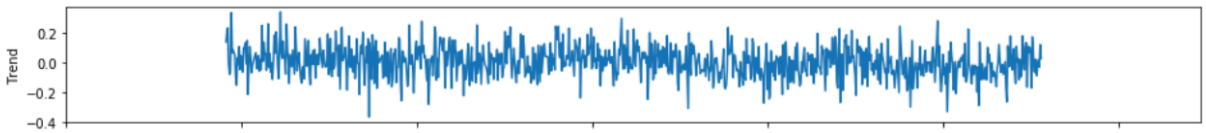
- We plot our data set (2015-06-15, 2019-09-29)



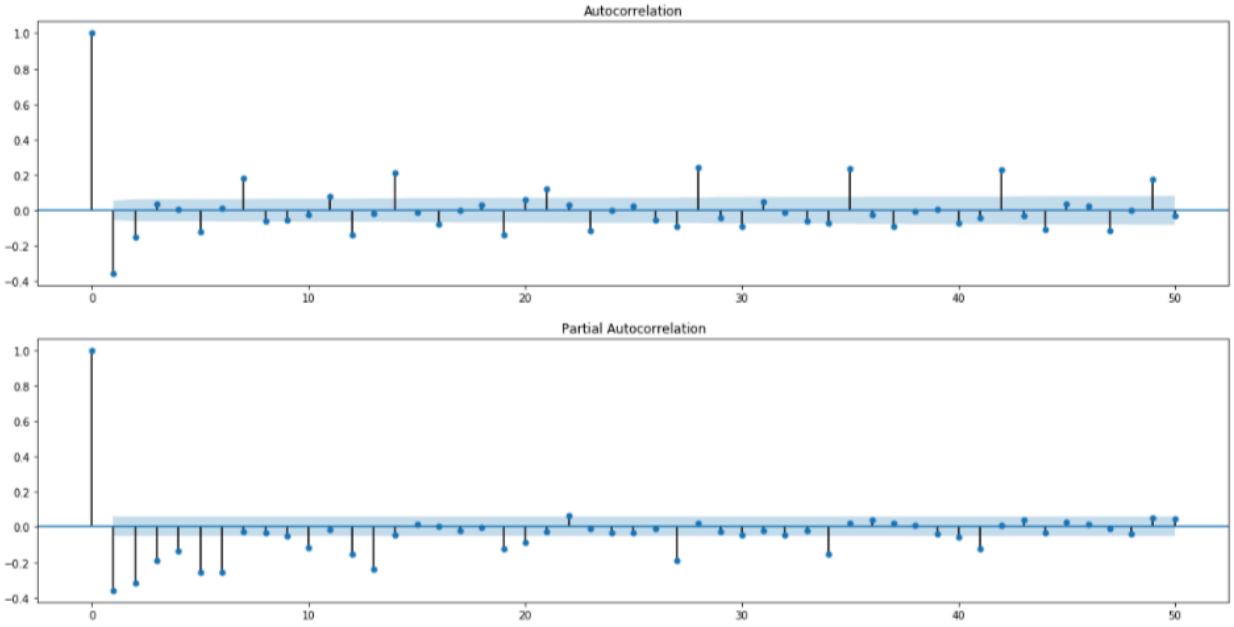
Stationarity test and parameters checking



- Based on the plot and the ADF Test (p-value 0.31), we conclude that the data is non-stationary. Hence, we take the first difference to make data stationary and check which parameters suit our model best.
- Now, the data is stationary



- We can observe seasonality every 7 days. That is why we will run the seasonal ARIMA model



- We run the simulation showing us the most appropriate parameters. We will continue with the parameters associated with the lowest AIC. Out model: ARIMA$(1,1,1) \times (1,1,1)_7$

	order	seasonal_order	AIC
63	(1, 1, 1)	(1, 1, 1, 7)	12199.278553
59	(1, 1, 1)	(0, 1, 1, 7)	12201.915155
47	(1, 0, 1)	(1, 1, 1, 7)	12228.598751
31	(0, 1, 1)	(1, 1, 1, 7)	12234.637956
43	(1, 0, 1)	(0, 1, 1, 7)	12236.709335

Model fitting and diagnostic checking

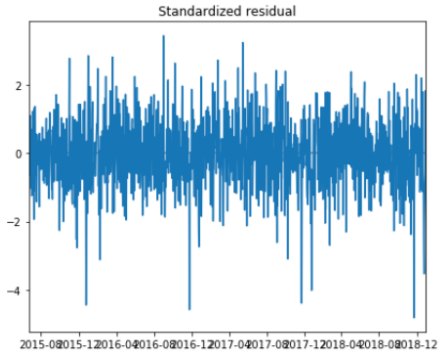
Statespace Model Results

Dep. Variable:	Number_of_Crimes	No. Observations:	1296
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 7)	Log Likelihood	-6094.639
Date:	Wed, 23 Oct 2019	AIC	12199.279
Time:	16:45:49	BIC	12225.083
Sample:	06-15-2015	HQIC	12208.965
	- 12-31-2018		
Covariance Type:	opg		

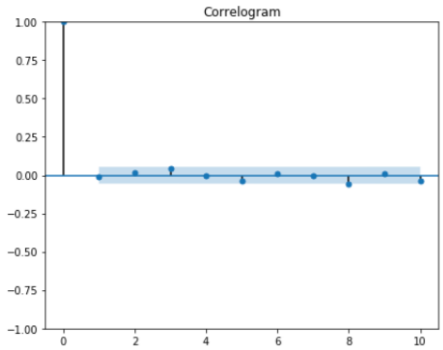
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1957	0.028	6.893	0.000	0.140	0.251
ma.L1	-0.9301	0.012	-75.803	0.000	-0.954	-0.906
ar.S.L7	-0.0644	0.027	-2.412	0.016	-0.117	-0.012
ma.S.L7	-0.9964	0.031	-31.950	0.000	-1.057	-0.935
sigma2	732.7346	30.017	24.411	0.000	673.902	791.567

Ljung-Box (Q):	64.14	Jarque-Bera (JB):	104.07
Prob(Q):	0.01	Prob(JB):	0.00
Heteroskedasticity (H):	1.18	Skew:	-0.25
Prob(H) (two-sided):	0.09	Kurtosis:	4.30

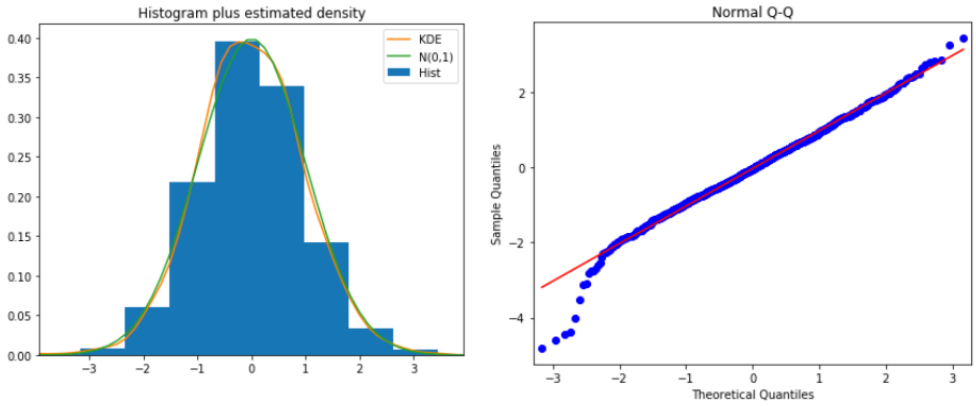
- The plot for standardized residuals suggests that we might violate the constant variance assumption



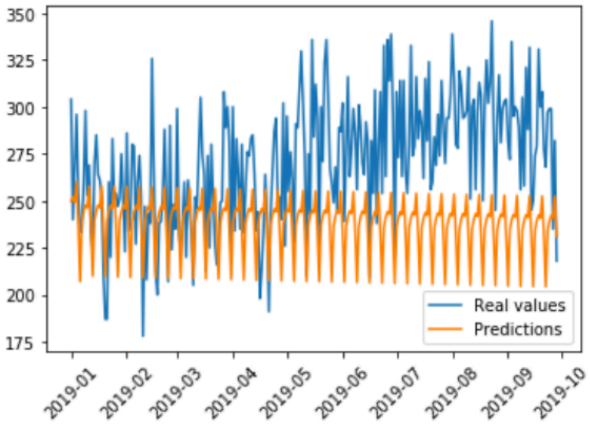
- The ACF (a correlogram) of the residuals does not show any statistically significant correlation. The model seems to have captures the essence of the dependence in the series



- The shape of The histogram of the residuals is "bell-shaped". The QQ plot suggests that the distribution may have a tail thicker than that of a normal distribution. The Shapiro-Wilk test statistic for normality equals 0.967 with p-value 2.07e-16. We reject the null hypothesis that the data is normally distributed.



- The MSE of this model equals 2071.46

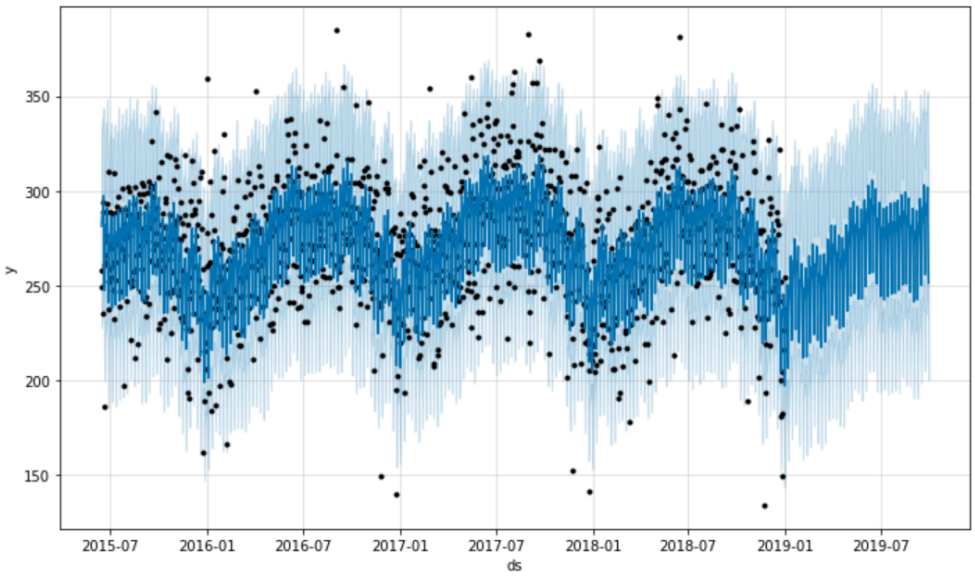


Conclusions

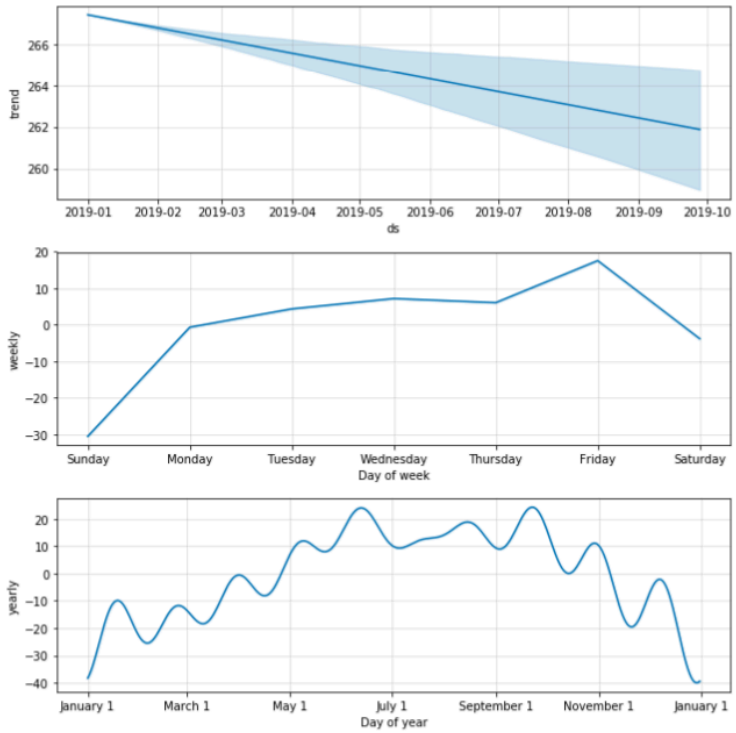
- ARIMA$(1,1,1) \times (1,1,1)_7$ is not a good model to predict daily crime in Boston because we violate the constant variance assumption

Prophet modeling

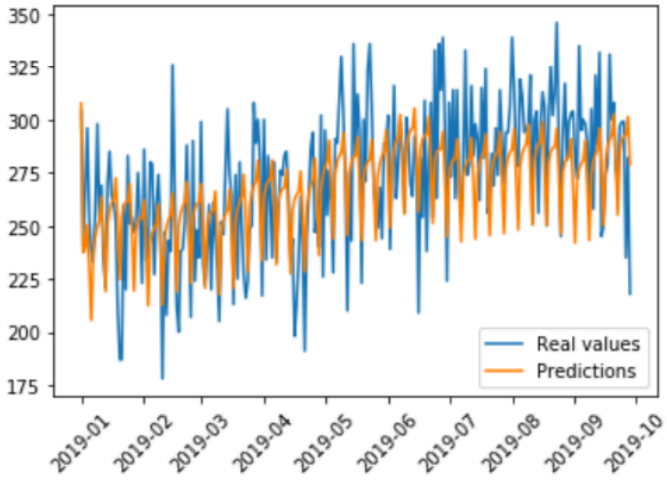
- We can see that Prophet's predictions are more accurate



- Prophet gives us the forecast components

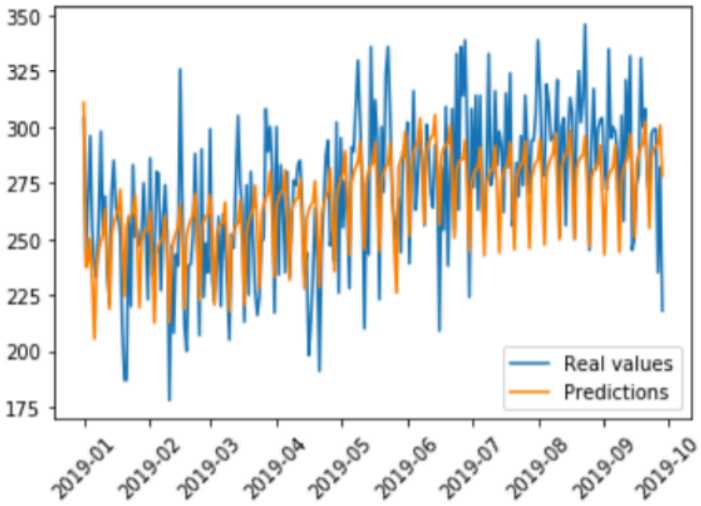


- The MSE of this model is 684.37



Prophet modeling with holidays

- We add 8 holidays to the model
- The MSE of this model is 652.44



Conclusions

- The Prophet model that includes public holidays gives us the most accurate predictions of daily crime in Boston