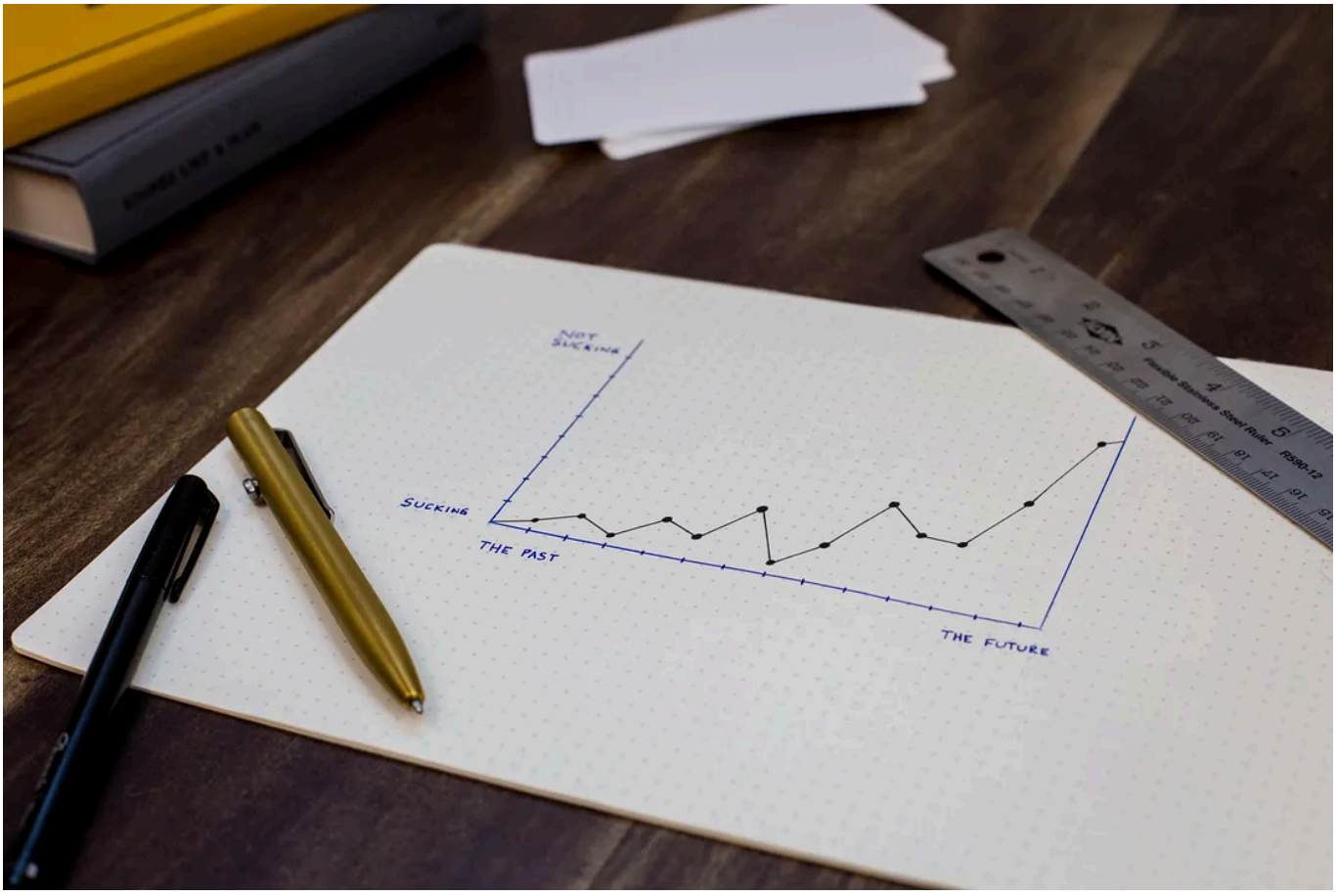
Time Series Forecasting with ARIMA , SARIMA and SARIMAX

A deep-dive on the gold standard of time series forecasting

Brendan Artley · Follow

Published in Towards Data Science · 10 min read · Apr 26, 2022



Time Series Graph — By Isaac Smith

Time series forecasting is a difficult problem with no easy answer. There are countless statistical models that claim to outperform each other, yet it is never clear which model is best.

That being said, ARMA-based models are often a good model to start with. They can achieve decent scores on most time-series problems and are well-suited as a baseline model in any time series problem.

This article is a comprehensive, beginner-friendly guide to help you understand ARIMA-based models.

Introduction

The ARIMA model acronym stands for "Auto-Regressive Integrated Moving Average" and for this article we will will break it down into AR, I, and MA.

Autoregressive Component — AR(p)

The autoregressive component of the ARIMA model is represented by AR(p), with the p parameter determining the number of lagged series that we use.

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \epsilon_t$$

AR Formula — By Author

AR(0): White Noise

If we set the p parameter as zero (AR(0)), with no autoregressive terms. This time series is just white noise. Each data point is sampled from a distribution with a mean of 0 and a variance of sigma-squared. This results in a sequence of random numbers that can't be predicted. This is really useful as it can serve as a null hypothesis, and protect our analyses from accepting false-positive patterns.

AR(1): Random Walks and Oscillations

With the p parameter set to 1, we are taking into account the previous timestamp adjusted by a multiplier, and then adding white noise. If the multiplier is 0 then we get white noise, and if the multiplier is 1 we get a random walk. If the multiplier is between $0 < \alpha_1 < 1$, then the time series will exhibit mean reversion. This means that the values tend to hover around 0 and revert to the mean after regressing from it.

AR(p): Higher-order terms

Increasing the p parameter even further is just means going further back and adding more timestamps adjusted by their own multipliers. We can go as far back as we want, but as we get further back it is more likely that we should use additional parameters such as the moving average (MA(q)).

Moving Average — MA(q)

"This component is not a rolling average, but rather the lags in the white noise." — Matt Sosna

MA(q)

MA(q) is the moving average model and q is the number of lagged forecasting error terms in the prediction. In an MA(1) model, our forecast is a constant term plus the previous white noise term times a multiplier, added with the current white noise term. This is just simple probability + statistics, as we are adjusting our forecast based on previous white noise terms.

ARMA and ARIMA Models

ARMA and ARIMA architectures are just the AR (Autoregressive) and MA (Moving Average) components put together.

ARMA

The ARMA model is a constant plus the sum of AR lags and their multipliers, plus the sum of the MA lags and their multipliers plus white noise. This equation is the basis of all the models that come next and is a framework for many forecasting models across different domains.

ARIMA

$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \ldots + \omega_q \varepsilon_{t-q} + \varepsilon_t$$

ARIMA Formula — By Author

The ARIMA model is an ARMA model yet with a preprocessing step included in the model that we represent using I(d). I(d) is the difference order, which is the number of transformations needed to make the data stationary. So, an ARIMA model is simply an ARMA model on the differenced time series.

SARIMA, ARIMAX, SARIMAX Models

The ARIMA model is great, but to include seasonality and exogenous variables in the model can be extremely powerful. Since the ARIMA model assumes that the time series is stationary, we need to use a different model.

SARIMA

$$y_t = c + \sum_{n=1}^{p} \alpha_n y_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{P} \phi_n y_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

SARIMA Formula — By Author

Enter SARIMA (Seasonal ARIMA). This model is very similar to the ARIMA model, except that there is an additional set of autoregressive and moving average components.The additional lags are offset by the frequency of seasonality (ex. 12 — monthly, 24 — hourly).

SARIMA models allow for differencing data by seasonal frequency, yet also by non-seasonal differencing. Knowing which parameters are best can be made easier through automatic parameter search frameworks such as pmdarina.

ARIMAX and SARIMAX

$$d_t = c + \sum_{n=1}^{p} \alpha_n d_{t-n} + \sum_{n=1}^{q} \theta_n \epsilon_{t-n} + \sum_{n=1}^{r} \beta_n x_{n_t} + \sum_{n=1}^{P} \phi_n d_{t-sn} + \sum_{n=1}^{Q} \eta_n \epsilon_{t-sn} + \epsilon_t$$

Sarimax Formula — By Author

Above is the the of the SARIMAX model. This model takes into account exogenous variables, or in other words, use external data in our forecast. Some real-world examples of exogenous variables include gold price, oil price, outdoor temperature, exchange rate.

It is interesting to think that all exogenous factors are still technically indirectly modeled in the historical model forecast. That being said, if we include external data, the model will respond much quicker to its affect than if we rely on the influence of lagging terms.

Code Example

Lets look at these models in actions through a simple code example in Python.

```
from datetime import datetime
import numpy as np
import pandas as pd
import matplotlib.pylab as plt
%matplotlib inline
from matplotlib.pylab import rcParams

from statsmodels.tsa.stattools import adfuller
!pip install pmdarima --quiet
import pmdarima as pm
```

arima imports hosted with ❤ by GitHub view raw

Loading Data

For this example, we are going to use the Air Passengers Dataset. This dataset contains the number of air travel passengers from the start of 1949 to the end of 1960.

This dataset has a positive trend and annual seasonality.

As soon as the dataset is read, the index is set to the date. This is standard practice when working with time-series data in Pandas, and makes it easier to implement ARIMA, SARIMA, and SARIMAX.

	#Passengers
Month	
1949-01-01	112
1949-02-01	118
1949-03-01	132
1949-04-01	129
1949-05-01	121

Cell output — By Author

Trend

The general direction of the data over time. For example, if we are looking at the height of a newborn baby, their height will follow an upward trend into their youth. On the other hand, someone on a successful weight loss program will see their weight follow a downward trend over time.

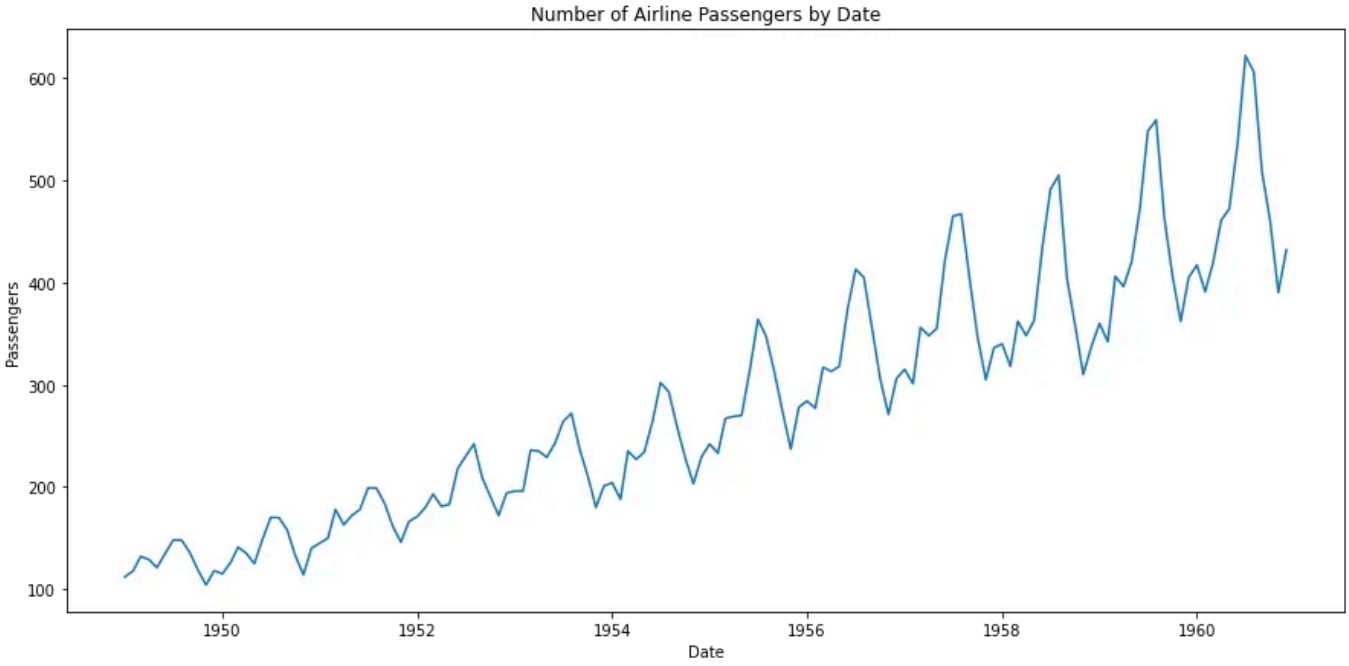
Seasonality + Cycles

Any seasonal or repeating patterns with a fixed frequency. Could be hourly, monthly, daily, annually, etc. One example of this is that Winter Jacket sales increase in the winter months and decrease in the summer months. Another example of this could be the balance of your bank account. In the 10 days at the start of every month, your balance follows a downward trend as you pay monthly rent, utilities, and other bill payments.

Irregularities + Noise

This is any large spikes or troughs in the data. One example of this could be your heart rate when you run the 400-meter dash. When you start the race your heart rate is similar to what it has been throughout the day, but during the race, it spikes to a much higher level for a small period of time before returning to a normal level.

In the visualization of the airline passenger data below, we can look for these components. At first glance, there looks to be a positive trend and some sort of seasonality or cyclicity in the dataset. There does not appear to be any major irregularities or noise in the data.

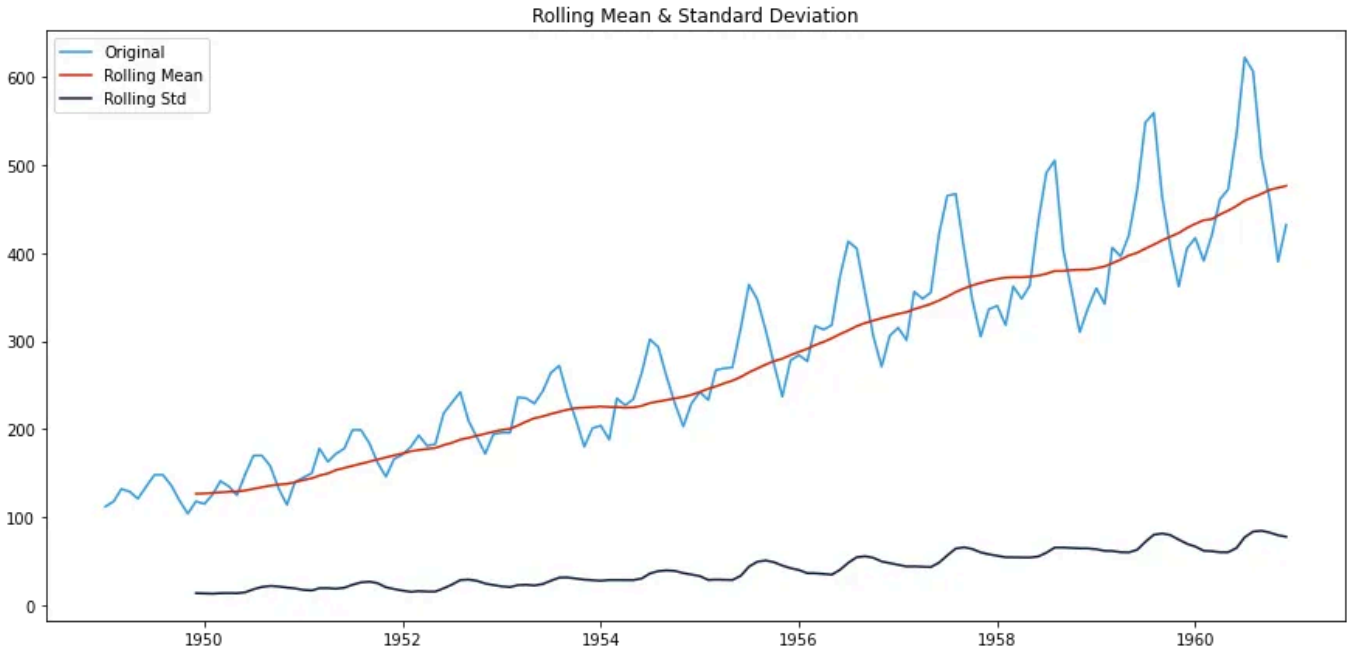


Airline Passengers Line Plot — By Author

Rolling Statistics

A rolling average is a great way to visualize how the dataset is trending. As the dataset provides counts by month, a window size of 12 will give us the annual rolling average.

We will also include the rolling standard deviation to see how much the data varies from the rolling average.



Rolling Statistics Plot — By Author

Augmented Dickey–Fuller Test

The Augmented Dickey-Fuller Test is used to determine if time-series data is stationary or not. Similar to a t-test, we set a significance level before the test and make conclusions on the hypothesis based on the resulting p-value.

Null Hypothesis: The data is not stationary.

Alternative Hypothesis: The data is stationary.

For the data to be stationary (ie. reject the null hypothesis), the ADF test should have:

- p-value <= significance level (0.01, 0.05, 0.10, etc.)

If the p-value is greater than the significance level then we can say that it is likely that the data is not stationary.

We can see in the ADF test below that the p-value is 0.991880, meaning that it is very likely that the data is not stationary.

```
Results of Dickey Fuller Test:
Test Statistic                  0.815369
p-value                         0.991880
#Lags Used                     13.000000
Number of Observations Used    130.000000
Critical Value (1%)             -3.481682
Critical Value (5%)             -2.884042
Critical Value (10%)            -2.578770
dtype: float64
```

ADF Test Output — By Author

ARIMA Model Selection w/ Auto-ARIMA

Although our data is almost certainly not stationary (p-value = 0.991), let's see how well a standard ARIMA model performs on the time series

Using the auto_arima() function from the `pmdarima` package, we can perform a parameter search for the optimal values of the model.

Model Diagnostics

Four plots result from the plot_diagnostics function. The Standardized residual, Histogram plus KDE estimate, Normal q-q, and the correlogram.

We can interpret the model as a good fit based on the following conditions.

Standardized residual

There are no obvious patterns in the residuals, with values having a mean of zero and having a uniform variance.

Histogram plus KDE estimate

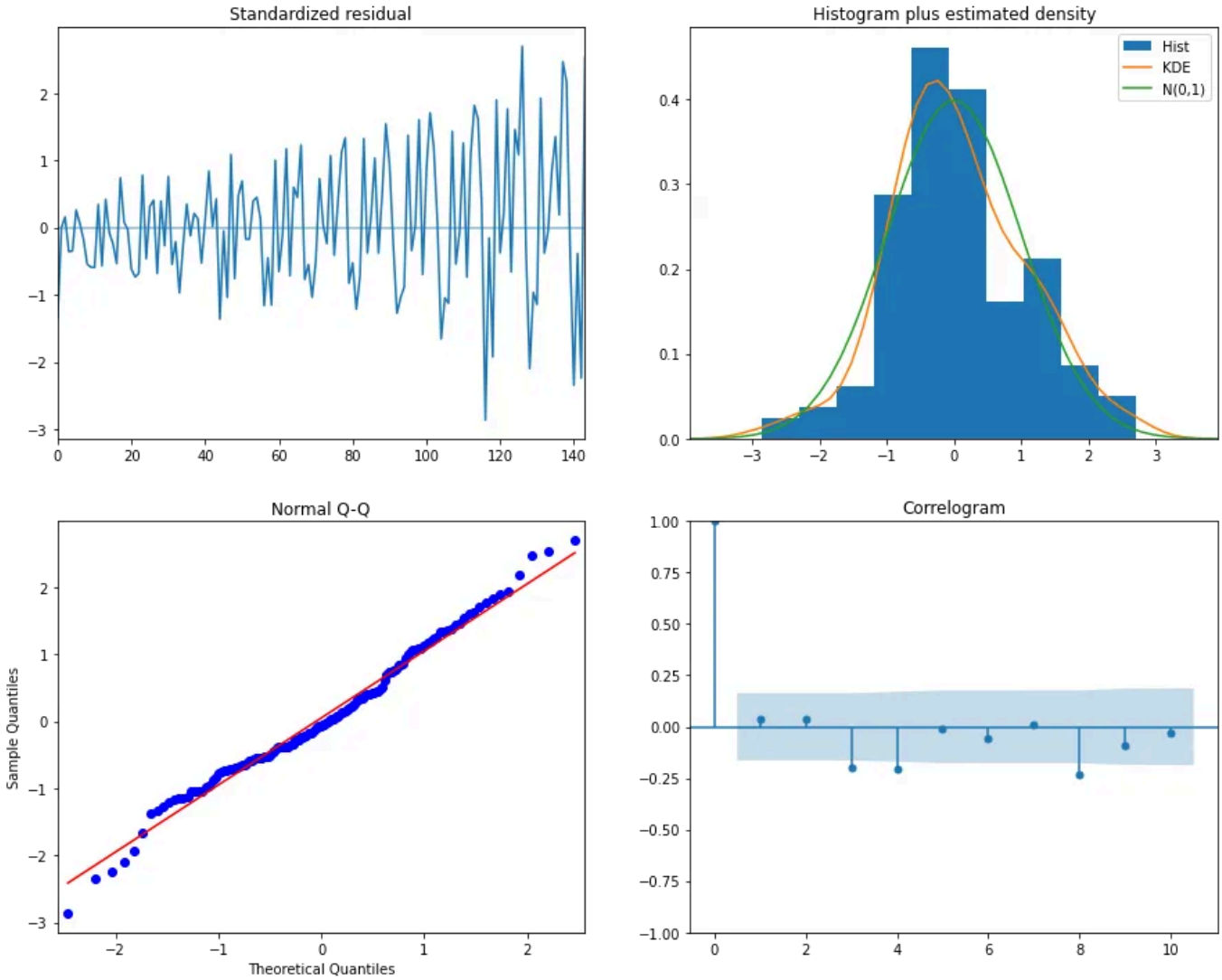
The KDE curve should be very similar to the normal distribution (labeled as N(0,1) in the plot)

Normal Q-Q

Most of the data points should lie on the straight line

Correlogram (ACF plot)

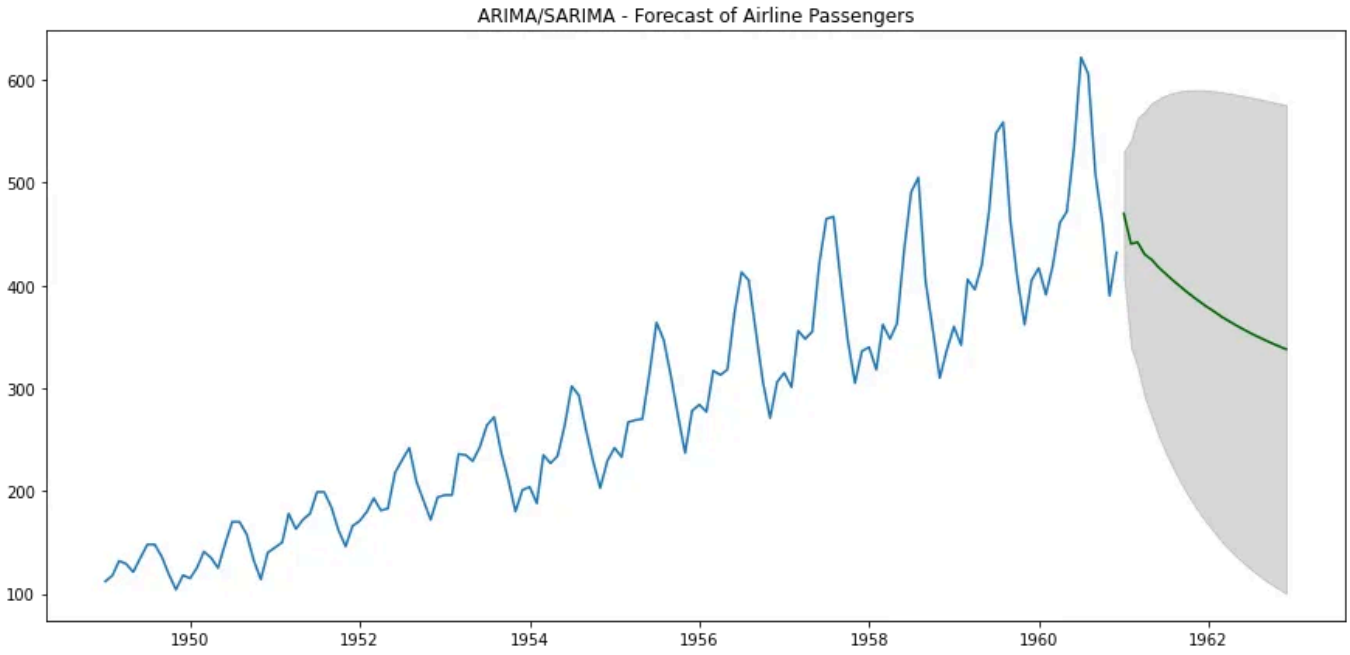
95% of correlations for lag greater than zero should not be significant. The grey area is the confidence band, and if values fall outside of this then they are statistically significant. In our case, there are a few values outside of this area, and therefore we may need to add more predictors to make the model more accurate



Arima Diagnostics — By Author

We can then use the model to forecast airline passenger counts over the next 24 months.

As we can see from the plot below, this doesn't seem to be a very accurate forecast. Maybe we need to change the model structure so that it takes into account seasonality?



ARIMA Forecast — By Author

SARIMA Model

Now let's try the same strategy as we did above, except let's use a SARIMA model so that we can account for seasonality.

Taking a look at the model diagnostics, we can see some significant differences when compared with the standard ARIMA model.

Standardized residual

The Standardized residual is much more consistent across the graph, meaning that the data is closer to being stationary.

Histogram plus KDE estimate

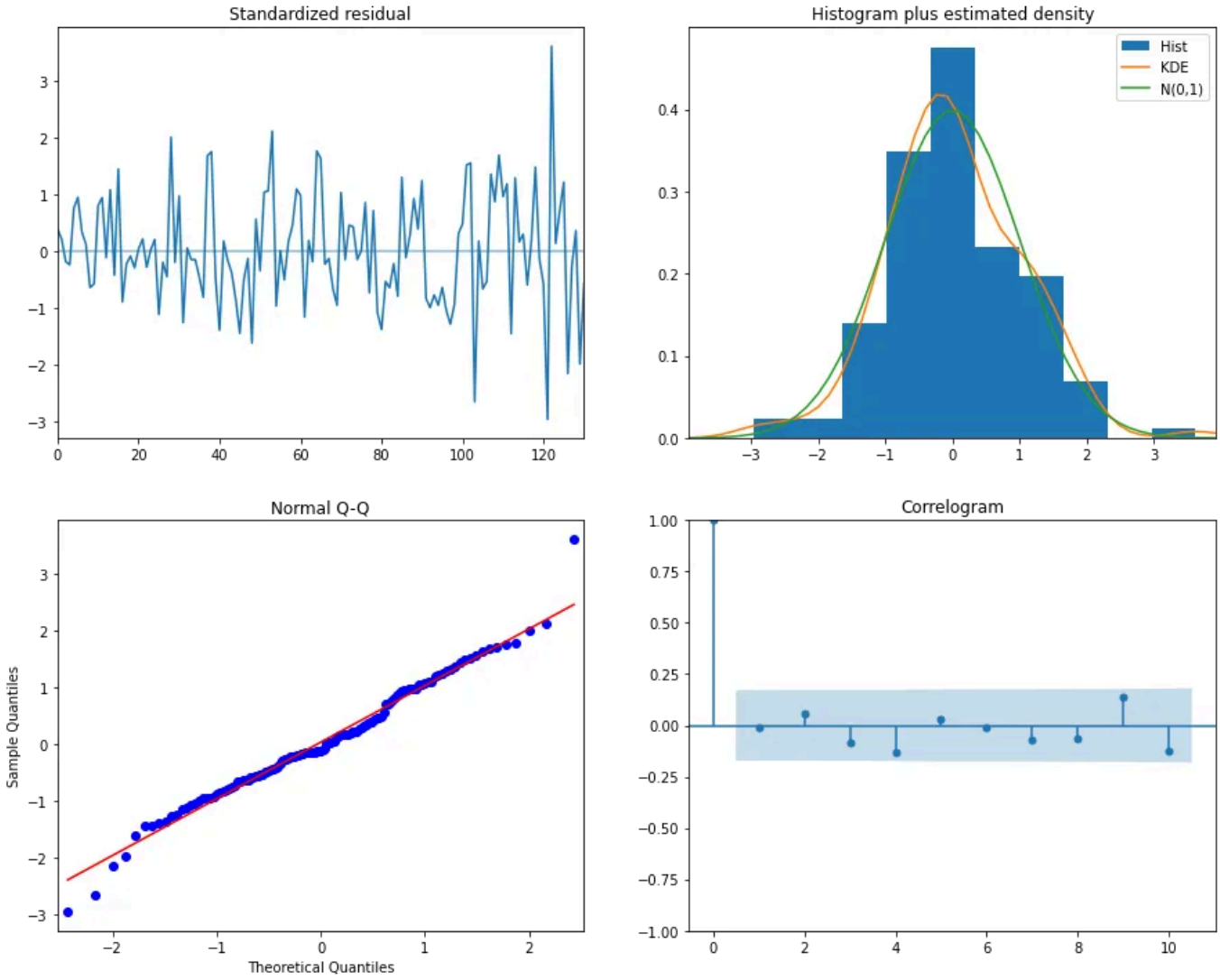
The KDE curve is similar to the normal distribution (not much changed here).

Normal Q-Q

The data points are clustered much closer to the line than in the ARIMA diagnostic plot.

Correlogram (ACF plot)

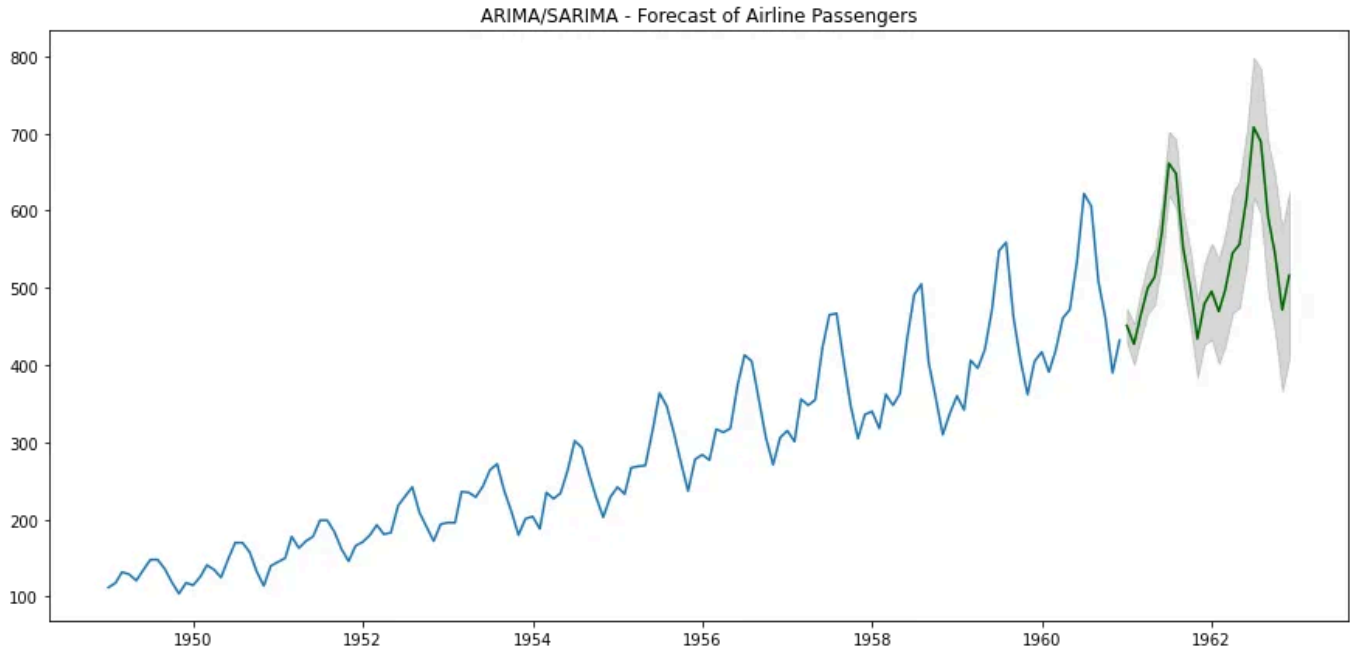
The grey area is the confidence band, and if values fall outside of this then they are statistically significant. We want all values inside this area. Adding the seasonality component did this! All the points now fall within the 95% confidence interval.



Sarima Diagnostics

We can then use the model to forecast airline passenger counts over the next 24 months as we did before.

As we can see from the plot below, this seems to be much more accurate than the standard ARIMA model!



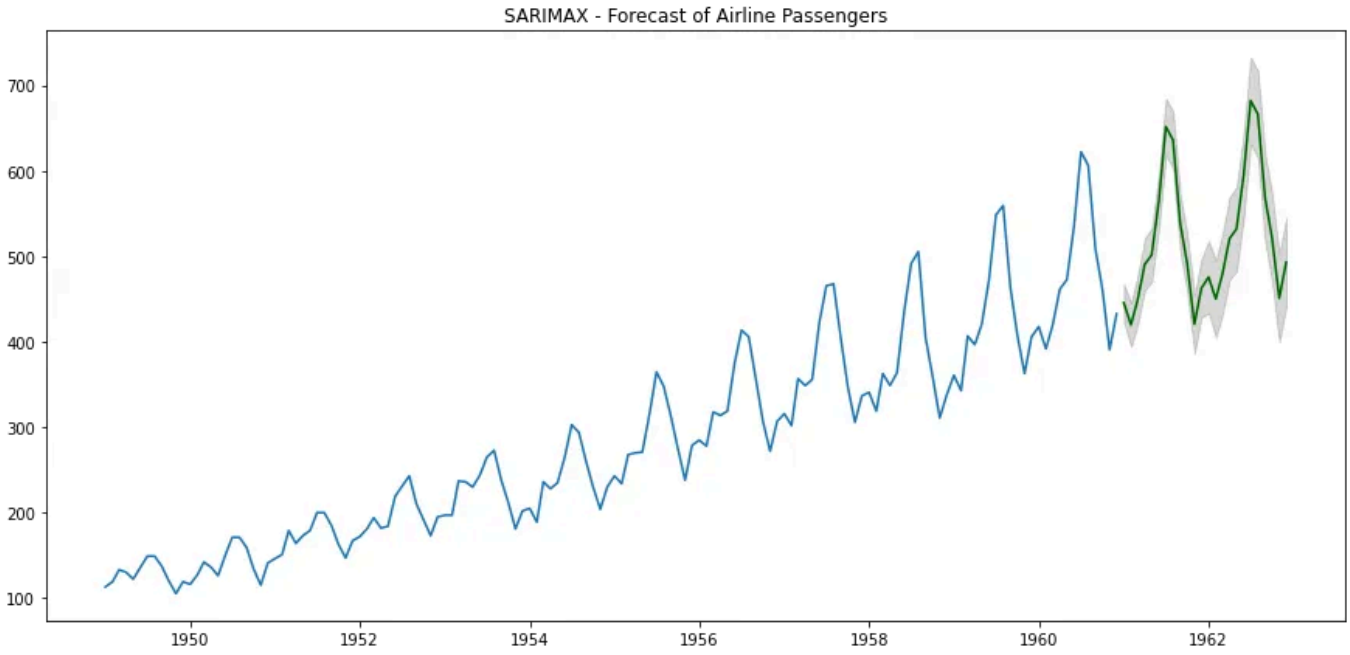
SARIMA Forecast — By Author

SARIMAX Model Selection

Now let's practice adding in an exogenous variable. In this example, I am simply going to add the month number as an exogenous variable, but this is not super useful as this is already conveyed through the seasonality.

Note that we are adding additional square brackets around the data being passed into the SARIMAX model.

We can see from the following predictions that we are getting some pretty good-looking predictions and the width of the forecasted confidence interval has decreased. This means that the model is more certain of its predictions.



SARIMAX Forecast — By Author

Closing Thoughts

Please find the code for this article here.

Putting ideas into my own words and implementing ARIMA models hands-on is the best way to learn. Hopefully this article can motivate others to do the same.

ARIMA model architectures provide more explainability than RNN's, yet RNN's are known to generate more accurate predictions. Now I have a good grasp on the ARIMA model architecture, I need to look into LSTM and RNN deep learning models for forecasting time series data!

Further Reading

Throughout the notebook I implement and reword ideas from the following sources. Thank you to all for sharing!

A Deep Dive on Arima Models — by Matt Sosna ← MUST READ!

Time Series For beginners with ARIMA — by @Arindam Chatterjee

Arima Model for Time Series Forecasting — by @Prashant Banerjee

StatsModels ADF Documentation

Removing Trends and Seasonality Article — by Jason Brownlee

A Gentle Introduction to SARIMA — by Jason Brownlee

Time Series Forecasting | Arima | Machine Learning | Statistics | Time Series Analysis

Responses (8)

What are your thoughts?

Respond

Matias
over 2 years ago

Hey, nice Article!, helped me alot.

About the Forecasting function definition shouldnt the forecasts start at df.index[-1] + 1 month?, instead of just df.index[-1]?

pd.date_range((df.index[-1]) + relativedelta.relativedelta(months=1) , periods = n_periods, freq='MS').

12 1 reply Reply

Matthias Wiedemann
over 2 years ago (edited)

I like your articles. How about combining the SARIMA with LightGBM?

10 1 reply Reply

Matheus Vizzotto
over 2 years ago

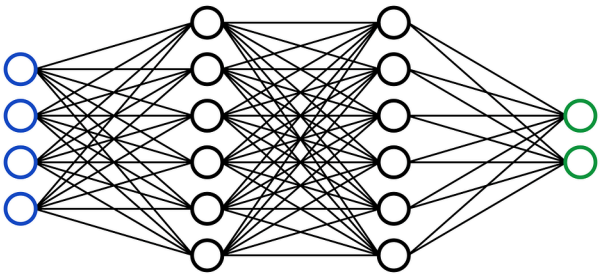
Awesome! You deserve more views in your article :)

Maybe in the SARIMAX section you could use some monthly economic indicator instead of "month", since you also said that the seasonality you used is already captured in the SARIMA model.

10 1 reply Reply

See all responses

More from Brendan Artley and Towards Data Science

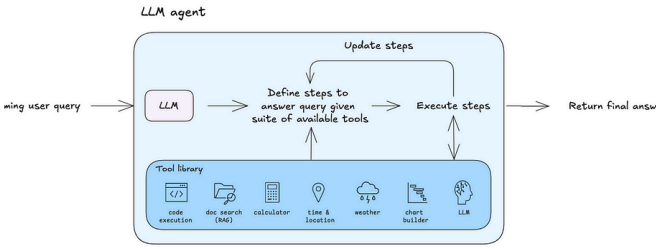


Brendan Artley

MNIST: Keras Simple CNN (99.6%)

Neural Networks in computer science are modelled after the biological neural network...

Apr 27, 2022 61

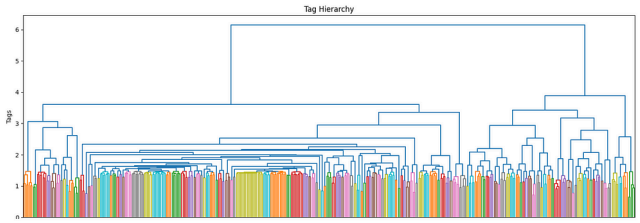


tds In Towards Data Science by Maya Murad

How to Build a General-Purpose LLM Agent

A Step-by-Step Guide

17h ago 116 4



tds In Towards Data Scien... by Michelangiolo Mazzes...

Introducing Univariate Exemplar Recommenders: how to profile...

Surveying and improving the current methodologies for customer profiling

1d ago 112 1

tds In Towards Data Science by Brendan Artley

Time Series Forecasting: Prediction Intervals

Estimate the range of a future observation with confidence.

Jun 14, 2022 185 1

See all from Brendan Artley

See all from Towards Data Science

Recommended from Medium

Kishan A

ARIMA vs SARIMA vs SARIMAX vs Prophet for Time Series...

Time series forecasting is a crucial tool in various industries like retail, finance, and...

Oct 3 58

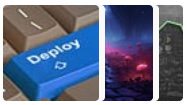
Irina (Xinli) Yu, Ph.D.

Mastering Time Series Forecasting with ARIMA Models

Time series forecasting is a critical component in many domains, including...

Jun 6 2

Lists



Predictive Modeling w/ Python

20 stories · 1701 saves

Practical Guides to Machine Learning

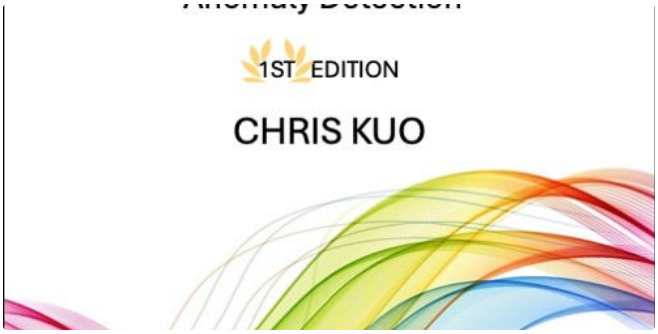
10 stories · 2071 saves

Natural Language Processing

1844 stories · 1468 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 518 saves

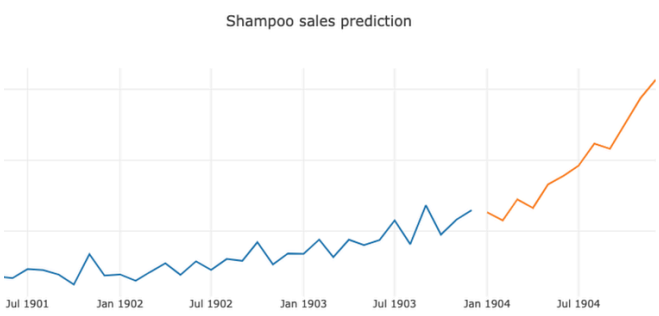


In Dataman in AI by Chris Kuo/Dr. Dataman

Temporal Fusion Transformer for Interpretable Time Series...

Sample eBook chapters (free): https://github.com/dataman-git/modern-...

Apr 18 381 3

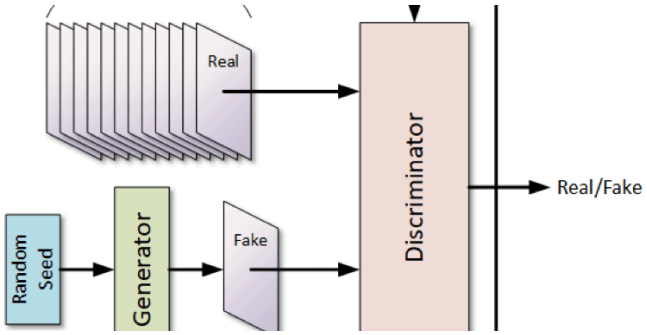


Chris Yan

Understanding SARIMAX: An Seasonal Time Series Forecasting...

SARIMAX, or Seasonal AutoRegressive Integrated Moving Average with eXogenous...

Aug 5 55



Pape

Time Series Forecasting with GANs: A Comprehensive Guide

Time series forecasting is essential in various fields such as finance, weather prediction, a...

Jun 8 114

date	store_nbr	item_nbr	unit_sales	onp
2014-01-02	1	108786	5.0	
2014-01-02	1	115847	4.0	
2014-01-02	1	115894	13.0	
2014-01-02	1	116018	4.0	
2014-01-02	1	122095	3.0	

In Towards AI by Alexandre Warembourg

The Endless Possibilities of Forecasting in Data Science

Discover the numerous methods available for forecasting in data science through practical...

Dec 15, 2023 547 1

See more recommendations