In the Name of Allah

Candidate: Fatemeh Vakili **Job Position: Machine Learning Engineer**

هدف از این پروژه، بررسی انواع الگوریتم‌های مبتنی بر سری زمانی بر اساس مسائل مربوطه است. در این گزارش، به سه سوال اصلی پاسخ داده می‌شود که سوالات را می‌توانید در این مسیر مشاهده کنید:

https://github.com/FATEMEHVAKILI/TimeSeriesForecasting/blob/main/Report/Questions.docx

در سوال اول، تمامی مراحلی که در علم داده بررسی می‌شود، انجام شد. در ابتدا شاخص " Exports of goods and services (constant ۲۰۱۵ US$)" از داده‌های اصلی استخراج شد که کد آن در فایل Data قرار دارد. این کد مبتنی بر اصول SOLID پیاده سازی شده است که هدف آن استخراج داده‌های مربوط به هر شاخص است. با تامل در داده‌ها متوجه می‌شویم داده‌ها، سری زمانی هستند.

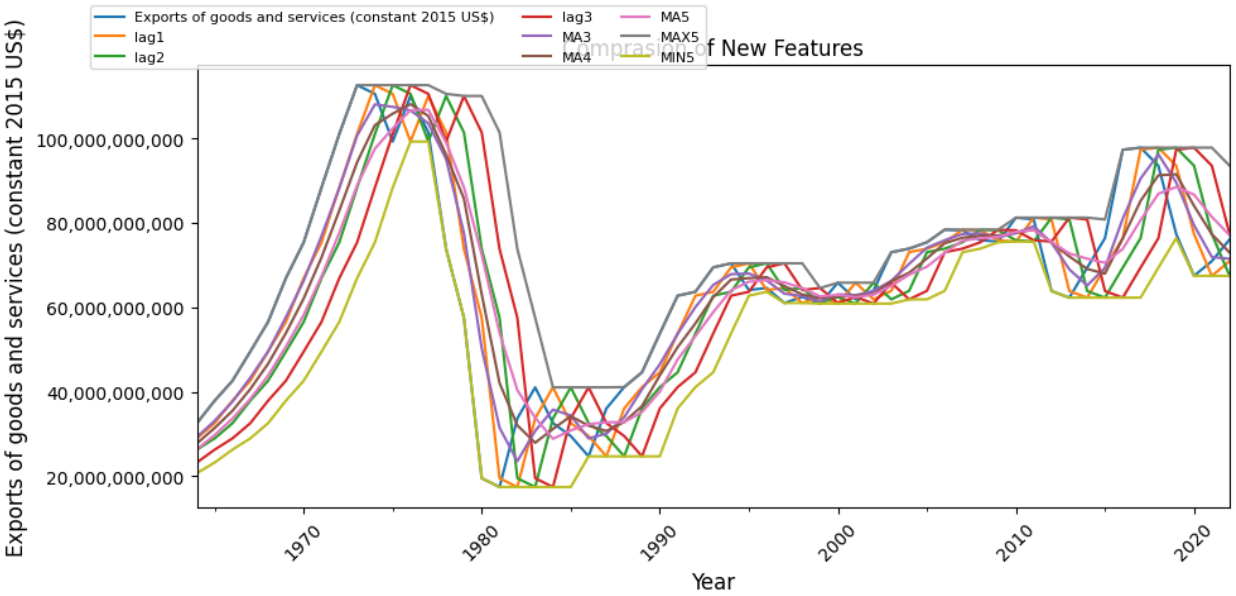
برای پاسخ به سوال اول، الگوریتم‌های ساده‌تر مانند ARIMA و SARIMA بر روی داده‌ها اعمال شده‌اند و پس از آن الگوریتم‌های مبتنی بر یادگیری ماشین مانند XGBoost استفاده شده است. لازم به ذکر است که این الگوریتم‌ها تک متغیره هستند. به همین ترتیب به تفکیک هر کدام از فایل‌ها توضیح داده می‌شود. توصیه می‌شود برای بررسی بیشتر حتما فایل کدها را مشاهده فرمائید.

در فایل ۱.ipynb داده‌ها ابتدا خوانده می‌شوند و در مرحله بعدی که پیش پردازش داده‌ها است؛ داده‌ها با دو روش نرمال سازی و استانداردسازی محاسبه می‌شوند. نرمال سازی در داده‌های با scale متفاوت استفاده می‌شود و معمولا در الگوریتم‌های مبتنی بر رگرسیون مانند RandomForestRegressior نتایج بهتری ارائه می‌دهد. همچنین در داده‌های غیر خطی مناسب‌تر است. هدف از استانداردسازی، در داده‌های با توزیع خطی است و مبتنی بر میانگین و standard deviation محاسبه می‌شود.

از مهندسی داده برای تولید ویژگی‌های جدید استفاده می‌شود که این ویژگی‌ها می‌توانند در پیش نگری یا Forecasting موثر باشند. بنابراین Lagged و Rolling محاسبه شدند و برای تعیین بهترین ویژگی از معیار MSE استفاده می‌شود که کمترین مقدار آن به معنای ویژگی بهتر است و نتایج آن به شرح ذیل است:

	Model	Mean Squared Error
0	MA 3	7.436002e+19
1	Lag 1	9.966655e+19
2	MA 4	1.336058e+20
3	MA 5	1.962835e+20
4	Lag 2	2.846223e+20
5	MIN 5	3.356244e+20
6	MAX 5	4.334514e+20
7	Lag 3	4.891217e+20

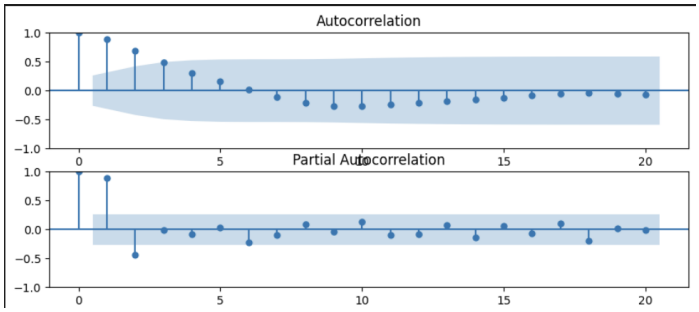
برای واضح‌تر شدن از نمودار استفاده شده است که در نمودار نیز ویژگی‌های MA3 نسبه به دیگر ویژگی‌ها مناسب‌تر است.

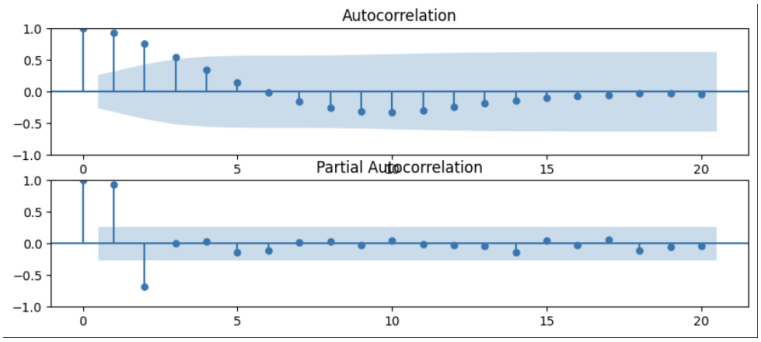


در داده‌های سری زمانی، داده‌ها باید دارای شرایطی مانند Stationary, Trend و Seasonality باشند که بهتر است داده‌ها دارای این شرایط باشند تا مدل‌ها بهتر فیت شوند. بنابراین برای فرض Stationary از آزمون فرض H0 و H1 استفاده می‌کنیم. نتایج آن نشان می‌دهد که داده‌ها از حالت Stationary برخوردار هستند.

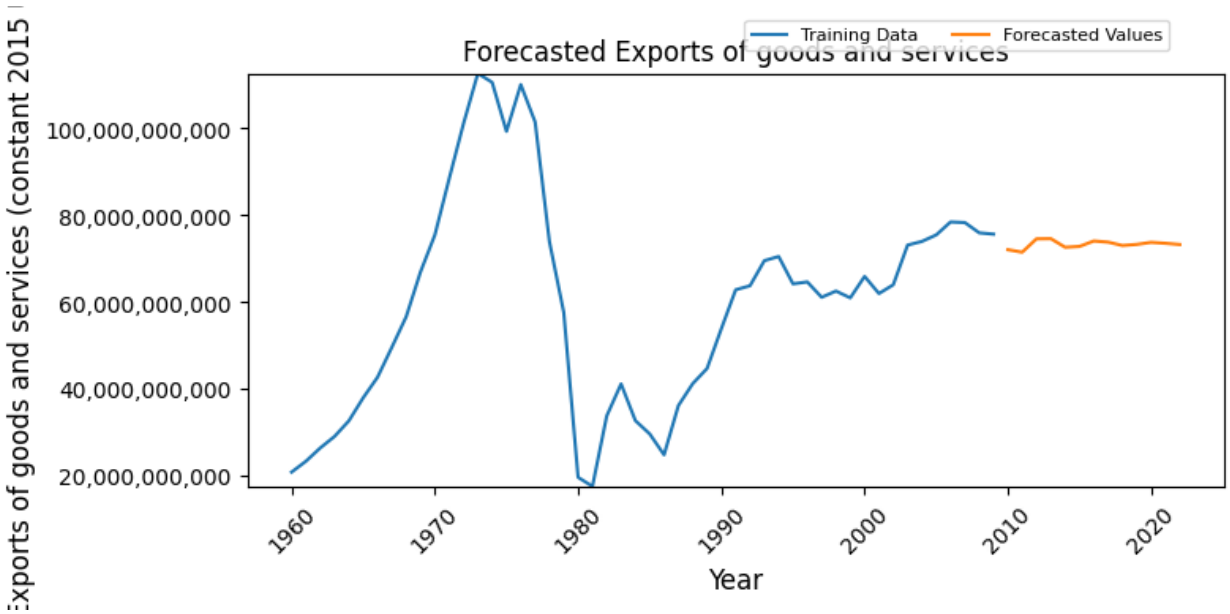
```
ADF Test Statistic : -2.9388346680956707
p-value : 0.041014851385688O2
#Lags Used : 1
Number of Observations Used : 57
stationary
```

همچنین می‌توان از مدل‌های آماری مثل Autocorrelation and Partial Autocorrelation Functions (ACF & PACF) استفاده کرد که آنها نیز نشان دهنده Stationary هستند. لذا به ترتیب برای ویژگی اصلی یا Actual و MA3 که کمترین MSE را دارد؛ ترسیم می‌شوند.





حالا در فایل بعدی که 11.ipynb است به مدل‌های ARIMA و SARIMA پرداخته می‌شود. برای مدل اول، هر دو روش پیش پردازش نرمال سازی و استانداردسازی انجام می‌شود که نتایج در هر دو یکسان است. لذا فقط نرمال سازی توصیح داده می‌شود. داده‌ها نیز به نسبت ۷۰ درصد برای مجموعه آموزشی و ۳۰ درصد برای مجموعه تست استفاده می‌شوند. نمودار زیر مدل پیش نگری داده‌ها را مبتنی بر مدل ARIMA نشان می‌دهد.



همچنین میزان MSE در این مدل

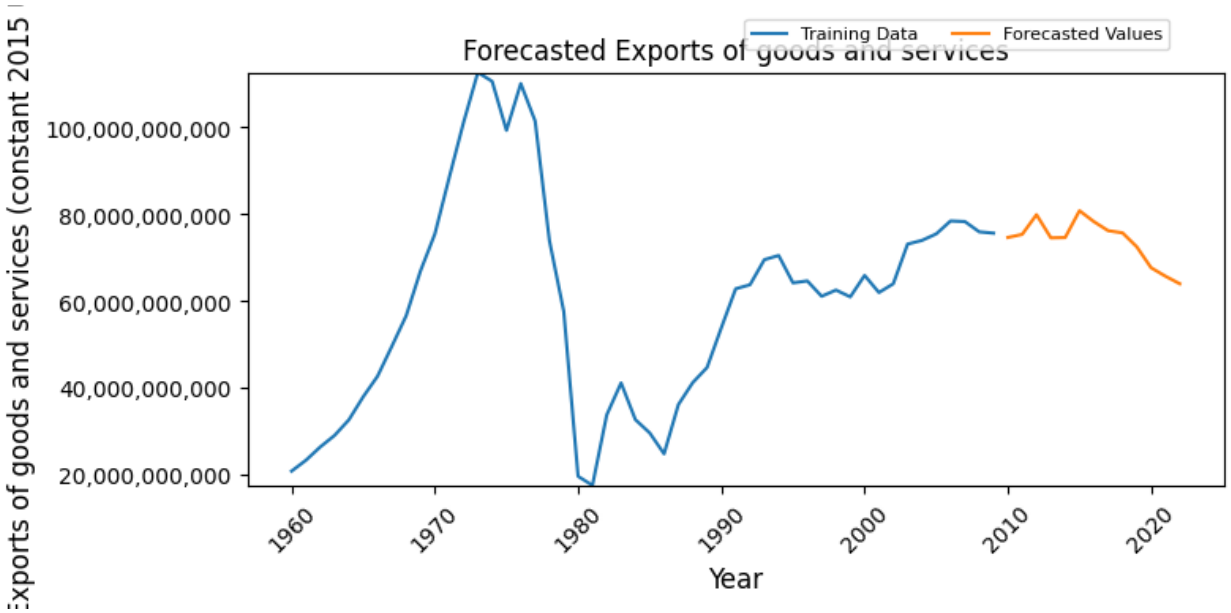
1.6021126418533155e+20

است که مقدار آن کم است و نشان می‌دهد داده‌ها به خوبی روی مدل فیت شده‌اند.

مدل بعدی SARIMA است که دقیقا مراحل قبل بر روی آن انجام شده‌اند و نمودار زیر نتایج این مدل را نشان می‌دهد.

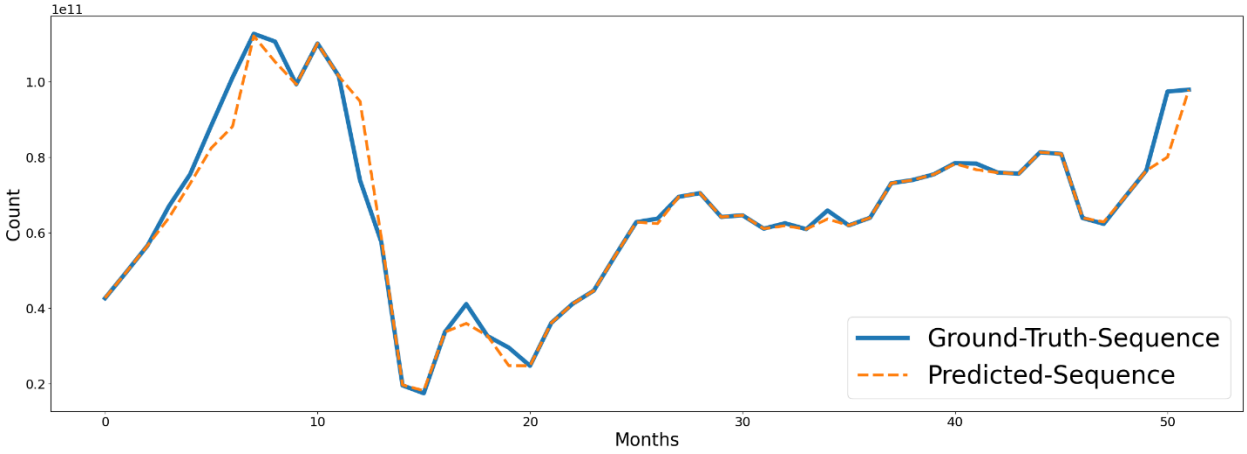
همچنین مقدار MSE در این مدل نسبت به مدل ARIMA کمتر است:

1.4486908283426279e+20

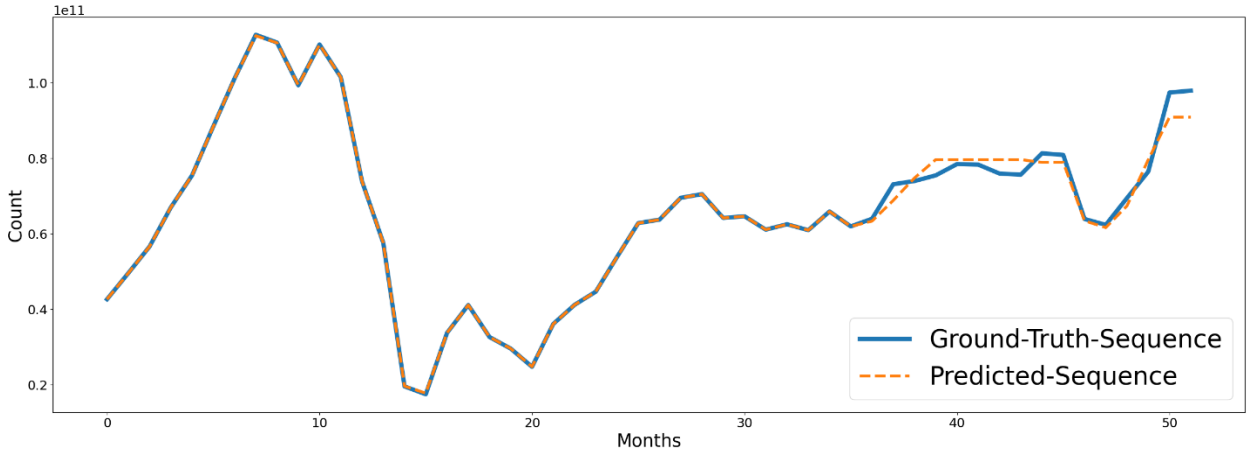


بنابراین می‌توان نتیجه گرفت مدل SARIMA پیش نگری بهتری داشته است و وقتی که آن را با نمودار اصلی مقایسه می‌کنیم، داده‌ها به مقادیر اصلی شبیه‌تر هستند.

در مدل بعدی (فایل 12.ipynb) از الگوریتم XGBoost با روش Random Splitting استفاده شده است. تمامی مراحل پیش پردازش در این الگوریتم نیز انجام شده‌اند و نتایج این الگوریتم در نمودار ذیل قابل مشاهده است:



در فایل 13.ipynb نیز از همان الگوریتم قبلی استفاده شده است اما با این تفاوت که در این مدل داده‌ها بر اساس زمان تقسیم بندی شده‌اند. لذا نتایج این مدل به شرح ذیل است:

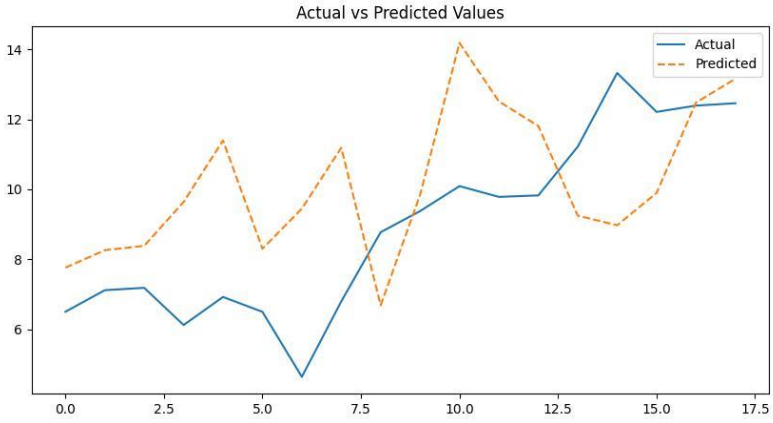


هر دو الگوریتم نتایج تقریبا مشابهی دارند و قطعا نسبت به مدل‌های آماری مثل ARIMA و SARIMA نتایج بهتری در پیش نگری داشتند.

بنابراین مدل‌های یادگیری ماشین نسبت به مدل‌های آماری پیچیده‌تر هستند اما نتایج بهتری را ارائه می‌دهند اما مدل‌های اماری در شرایطی که سادگی مدل مطرح باشد، قابل استفاده هستند.

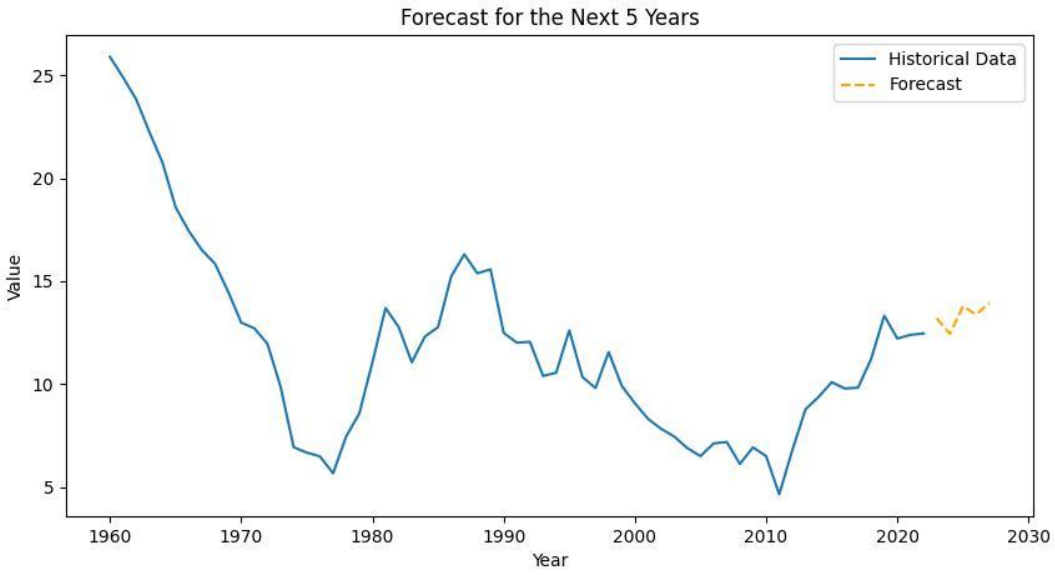
در فایل 21.py از الگوریتم‌های یادگیری عمیق برای پیش بینی ۵ سال آینده در داده‌های سری زمانی استفاده شده است. برای اینکه از مدل‌های مختلفی استفاده شده باشد و نحوه کدنویسی نیز بررسی گردد؛ لذا در این سوال از یادگیری عمیق استفاده شده است و در فرمت .py قرار دارد.

الگوریتم اول CNN-LSTM است و وظیفه آن پیش نگری شاخص مد نظر در ۵ سال آینده است. لذا این الگوریتم با ۱۰۰۰ EPOCH آموزش دیده است و مقادیر پیش نگری آن به شرح ذیل است:

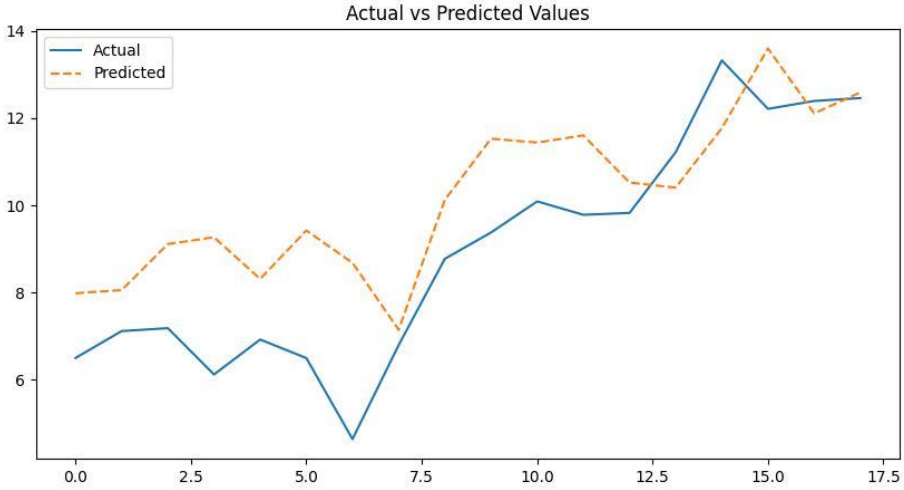


Forecasted Value
13.142433
14.215651
14.866517
16.013350
15.815674

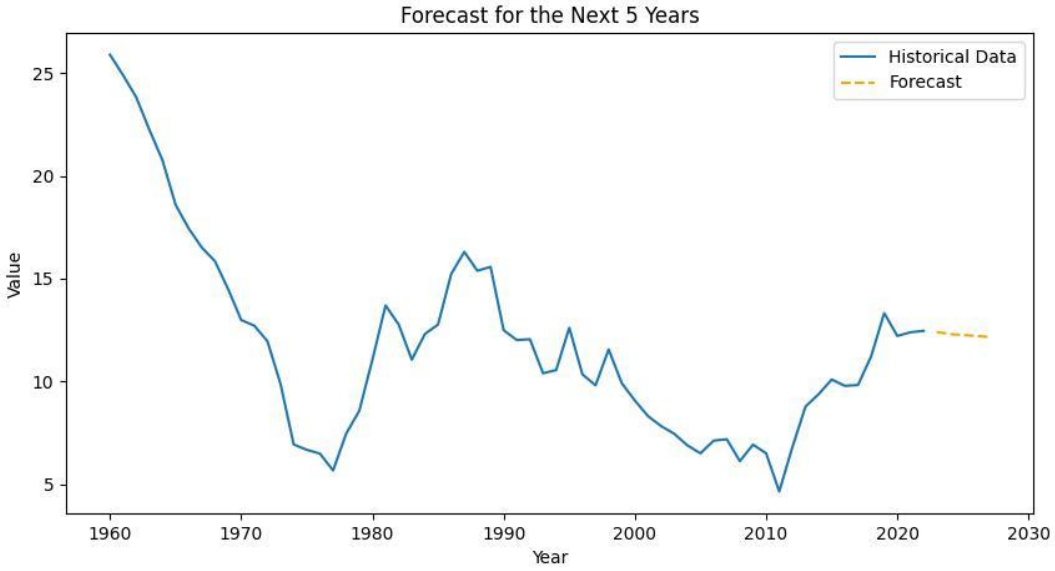
نمودار بالا داده‌های واقعی و داده‌های آموزش دیده را به همراه مقادیر پیش نگری شده نشان می‌دهد و در نمودار بعدی پیش نگری ۵ سال آینده به تصویر کشیده شده است:



در فایل 22.py نیز الگوریتم LSTM با شرایط پیش پردازش یکسان با الگوریتم قبلی محاسبه شده است. بنابراین ابتدا نمودار داده‌های آموزش دیده و سپس نمودار پیش نگری ۵ سال آینده به همراه مقادیر پیش نگری مشاهده می‌شود:

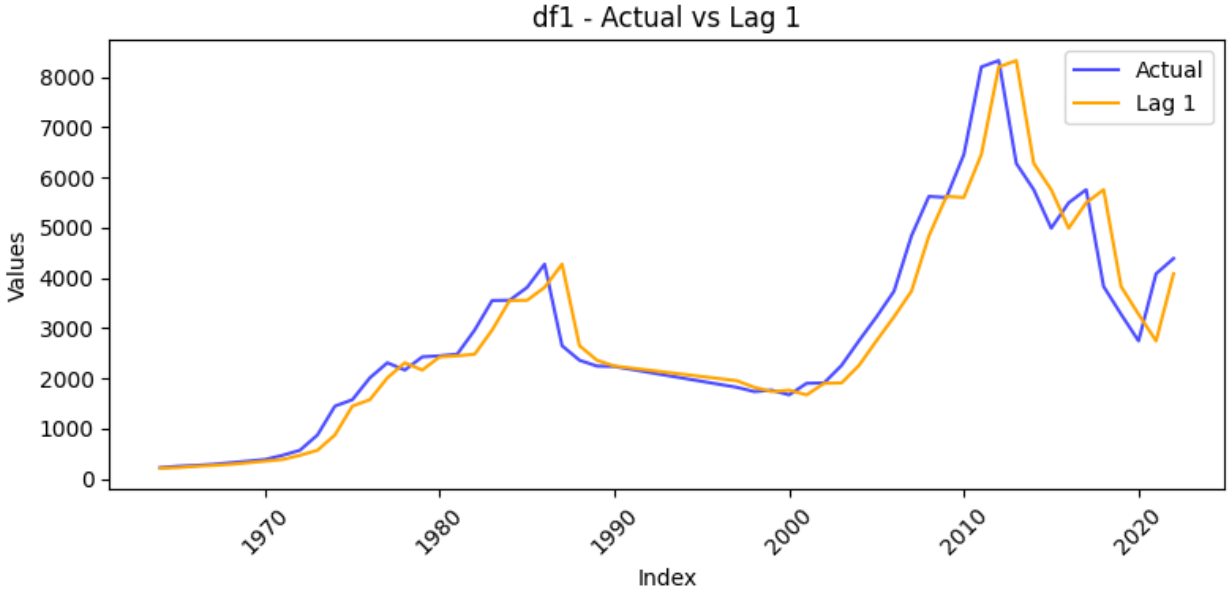


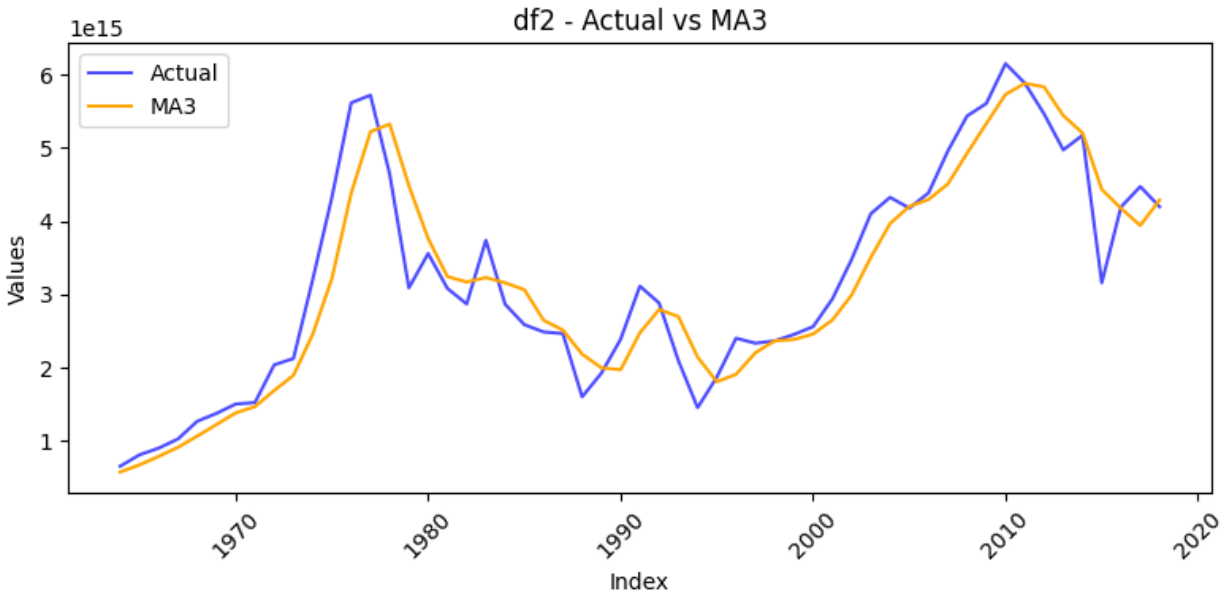
2023-01-01	12.402226
2024-01-01	12.304695
2025-01-01	12.262155
2026-01-01	12.211852
2027-01-01	12.164143

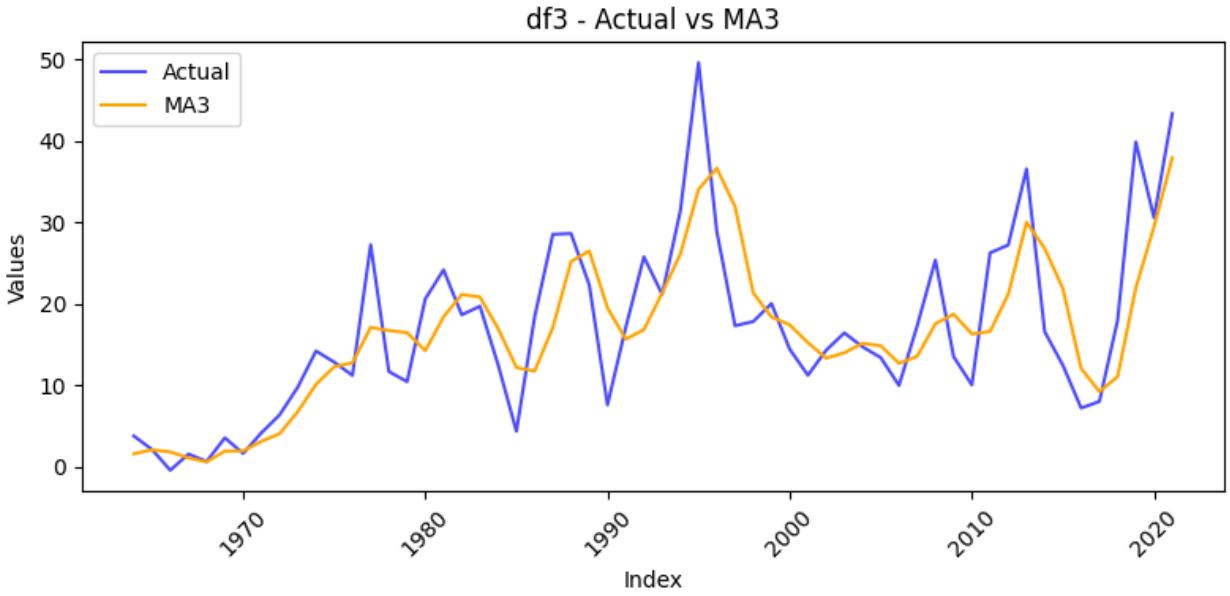


با مقایسه این دو الگوریتم یادگیری عمیق، مدل LSTM بهتر توانسته است آموزش ببیند و نتایج را پیش نگری کند.

در سوال آخر نیز، هدف بررسی نظرات دو کارشناس در حوزه‌ی تحریم‌های اقتصادی علیه ایران که از سال ۲۰۱۲ اعمال شده است، صحبت می‌شود. این سوال در واقع Time Series Analysis است و از روش‌های تحلیلی استفاده می‌شود. مانند روش‌های پیش پردازش در سوال اول، در اینجا نیز از Data Transformation استفاده می‌کنیم. بنابراین روش‌های مانند .Feature Eng، lagged variables و Rolling استفاده کرده‌ایم. در نهایت با استفاده از معیار MSE بهترین متغیر انتخاب می‌شود و در آخر ویژگی‌ها با هم مقایسه می‌شوند. بنابراین، طبق نظر کارشناسان؛ شاخص‌های درآمد سرانه، سرمایه گذاری و تورم بررسی می‌شوند. ابتدا سه نمودار زیر به ترتیب برای درآمد سرانه، سرمایه گذاری و تورم هستند:







در نمودار اول که مربوط به درآمد سرانه است؛ پس از سال ۲۰۱۲ افت زیادی در نمدار اتفاق افتاده است اما در حدود سال ۲۰۲۰ این نزول در نمدار در حال رسد و بهبود است. لذا می‌توان نتیجه گرفت در ابتدای تحریم‌ها با افت نسبی روبرو شدیم اما پس از سال ۲۰۲۰ این شرایط بهبود یافته است. در حالیکه ما هنوز در شرایط تحریم قرار داریم.

در نمودار دوم که سرمایه گذاری مطرح شده است، در ابتدای سال ۲۰۱۲ نزول بسیار کمی را در نمودار شاهد هستیم که این شرایط در حدود سال‌های ۲۰۱۵ تا ۲۰۱۸ در حال بهبود است.

در نمودار آخر نیز، تورم بررسی شده است. در سال ۲۰۱۲، تورم افزایش یافته است که پس از گذشت ۴ الی ۵ سال کاهش مناسبی را داشته است اما در حدود سال ۲۰۱۹ و ۲۰۲۰، مجددا تورم افزایش یافته است.

بطور کلی، تحریم‌ها در سرمایه گذاری تاثیر چندانی نداشته است ولی در تورم، دچار نزول تورم در بازه کوتاهی شدیم که پس از دوره حدودا ۴ ساله مجددا افزایش یافته است. شاخص درآمد سرانه در بازه تحریم تا بحال، همواره در حالت نزول نمودار قرار داشته است.

لینک ریپازیتوری در گیت هاب برای بررسی کدها:

https://github.com/FATEMEHVAKILI/TimeSeriesForecasting

Date: 8/19/2024-Monday

Author: Fatemeh Vakili