شناسایی و ارزیابی دقیق‌ترین مدل سری زمانی برای پیش‌بینی شاخص کیفیت هوا (AQI)

تهیه کننده:

امیرحسین حیدری

دانشگاه شیراز

بخش آمار

استاد راهنما:

دکتر علیرضا نعمت‌اللهی

تاریخ:

۱۴۰۲/۶/۱۳

Shiraz University

تحلیل سری زمانی:

در ریاضیات، **سری زمانی** به یک سری از داده اشاره دارد که به ترتیب زمان فهرست شده‌اند. معمولاً، سری زمانی دنباله‌ای از نقاط .است که در فواصل زمانی مساوی گرفته شده‌اند. بنابراین، این یک دنباله از داده‌های با زمان **گسسته** است

تحلیل سری زمانی روشی خاص برای تحلیل یک دنباله از نقاط داده است که در طول یک بازه زمانی جمع‌آوری شده‌اند. در تحلیل سری زمانی، تحلیلگران نقاط داده را در فواصل منظم و در طول یک دوره زمانی مشخص ثبت می‌کنند، به‌جای اینکه داده‌ها را به‌طور پراکنده یا تصادفی ثبت کنند.

یک سری زمانی دنباله‌ای از نقاط داده است که در زمان‌های مختلف جمع‌آوری شده‌اند. این نقاط اساساً اندازه‌گیری‌های متوالی هستند که از یک منبع داده واحد و در یک بازه زمانی مشخص جمع‌آوری شده‌اند. علاوه بر این، می‌توانیم از این مشاهدات که به‌صورت زمانی جمع‌آوری شده‌اند، برای بررسی روندها و تغییرات در طول زمان استفاده کنیم.

مدل‌های سری زمانی می‌توانند تک‌متغیره یا چندمتغیره باشند. مدل‌های سری زمانی تک‌متغیره زمانی استفاده می‌شوند که متغیر وابسته **یک سری زمانی واحد** باشد، مانند اندازه‌گیری دمای اتاق از یک حسگر واحد. از سوی دیگر، مدل‌های سری زمانی چندمتغیره زمانی استفاده می‌شوند که چندین متغیر وابسته وجود داشته باشد، یعنی خروجی به بیش از یک سری وابسته است. به‌عنوان مثال، می‌توان مدل‌سازی تولید ناخالص داخلی(GDP) ، تورم، و بیکاری را در نظر گرفت، چرا که این متغیرها به هم مرتبط هستند.

- سری زمانی نشان‌دهنده‌ی دنباله‌ای از رویدادهای مبتنی بر زمان است. این دنباله می‌تواند شامل سال‌ها، ماه‌ها، هفته‌ها، روزها، ساعت‌ها، دقیقه‌ها و ثانیه‌ها باشد.
- سری زمانی یک مشاهده از دنباله‌ای از زمان‌های گسسته در فواصل متوالی است.
- یک سری زمانی به عنوان یک نمودار پیوسته یا «نمودار متحرک» در نظر گرفته می‌شود که تغییرات یک متغیر را در طول زمان نمایش می‌دهد.
- متغیر/ویژگی زمانی به‌عنوان متغیر مستقل عمل کرده و به پیش‌بینی نتایج متغیر هدف کمک می‌کند.
- تحلیل سری زمانی (TSA) در زمینه‌های مختلف برای پیش‌بینی‌های مبتنی بر زمان استفاده می‌شود، مانند پیش‌بینی آب‌وهوا، امور مالی، پردازش سیگنال، و حوزه‌های مهندسی مانند سیستم‌های کنترل و سیستم‌های ارتباطی.
- از آنجا که تحلیل سری زمانی (TSA) شامل تولید مجموعه‌ای از اطلاعات در یک دنباله خاص است، این نوع تحلیل را از تحلیل‌های مکانی و دیگر تحلیل‌ها متمایز می‌کند.
- با استفاده از مدل‌های AR، MA، ARMA، و ARIMA می‌توانیم آینده را پیش‌بینی کنیم.

مدل‌های استفاده شده

LSTM - Long short-term memory

حافظه طولانی کوتاه مدت (Long short-term memory یا به اختصار LSTM) یک شبکه عصبی است که در زمینه‌های هوش مصنوعی و یادگیری عمیق استفاده می‌شود. بر خلاف شبکه‌های عصبی استاندارد که به‌طور پیش‌فرض داده‌ها را پردازش می‌کنند، LSTM دارای اتصالات بازخوردی است. چنین شبکه عصبی بازگشتی (RNN) می‌تواند نه تنها نقاط داده‌ای منفرد (مانند تصاویر) را پردازش کند، بلکه دنباله‌های کامل داده (مانند گفتار یا ویدئو) را نیز پردازش نماید. به عنوان مثال، LSTM در وظایفی مانند شناسایی نوشتار متصل بدون تقسیم‌بندی، شناسایی گفتار، ترجمه ماشینی، کنترل ربات، بازی‌های ویدیویی و بهداشت و درمان کاربرد دارد. LSTM به یکی از پراستنادترین شبکه‌های عصبی در قرن بیستم تبدیل شده است.

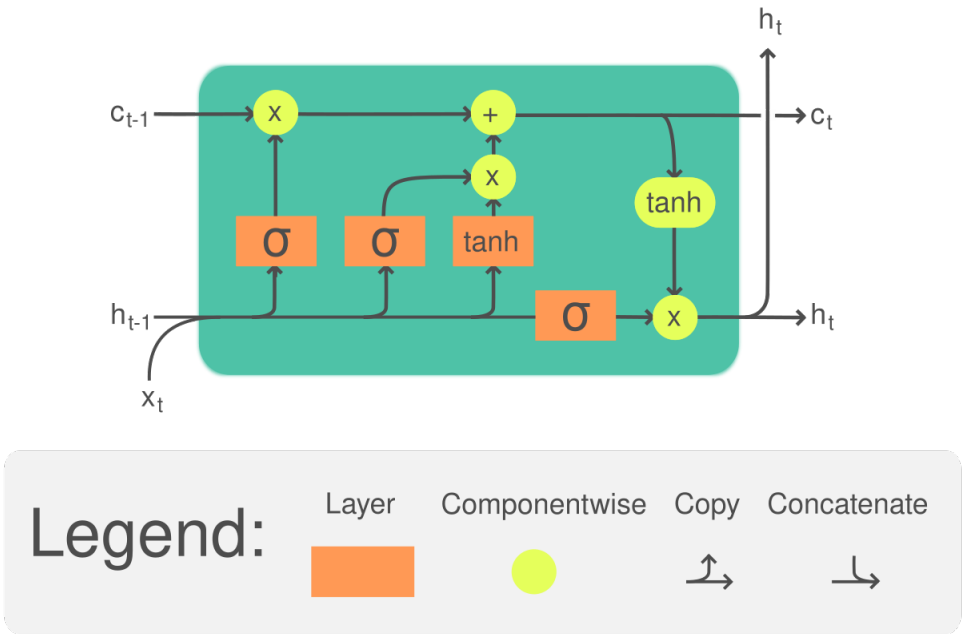
شبکه عصبی بازگشتی (RNN)

انسان‌ها هر ثانیه از ابتدا به تفکر نمی‌پردازند. هنگامی که این مقاله را می‌خوانید، هر کلمه را بر اساس درک قبلی‌تان از کلمات پیشین می‌فهمید. شما همه چیز را دور نمی‌اندازید و دوباره از ابتدا فکر نمی‌کنید. افکار شما دارای استمرار هستند.
شبکه‌های عصبی سنتی نمی‌توانند این کار را انجام دهند که این یک نقص جدی است. به عنوان مثال، تصور کنید که می‌خواهید طبقه‌بندی کنید که چه نوع رویدادی در هر لحظه از یک فیلم در حال اتفاق است. مشخص نیست که چگونه یک شبکه عصبی سنتی می‌تواند از استدلال خود درباره رویدادهای قبلی در فیلم برای اطلاع از رویدادهای بعدی استفاده کند.
شبکه‌های عصبی بازگشتی این مشکل را حل می‌کنند. آنها شبکه‌هایی هستند که دارای حلقه‌هایی درون خود هستند که به اطلاعات اجازه می‌دهند تا ادامه یابند.

نام LSTM به این دلیل انتخاب شده است که یک شبکه عصبی بازگشتی (RNN) دارای "حافظه بلندمدت" و "حافظه کوتاه‌مدت" است. وزن‌ها و بایاس‌های اتصالات در شبکه، پس از هر قسمت از آموزش تغییر می‌کنند که مشابه با تغییرات فیزیولوژیکی در قدرت سیناپس‌ها برای ذخیره حافظه‌های بلندمدت است؛ الگوهای فعال‌سازی در شبکه پس از هر گام زمانی تغییر می‌کنند که مشابه با تغییر لحظه‌به‌لحظه الگوهای شلیک الکتریکی در مغز برای ذخیره حافظه‌های کوتاه‌مدت است. معماری LSTM هدفش ارائه یک حافظه کوتاه‌مدت برای RNN است که می‌تواند هزاران گام زمانی دوام بیاورد، و از این رو به آن "حافظه کوتاه‌مدت بلند" گفته می‌شود.

یک واحد معمولی LSTM از یک سلول، یک دروازه ورودی، یک دروازه خروجی و یک دروازه فراموشی تشکیل شده است. سلول مقادیر را در فواصل زمانی دلخواه به خاطر می‌سپارد و سه دروازه جریان اطلاعات به داخل و خارج از سلول را تنظیم می‌کنند.

شبکه‌های LSTM برای طبقه‌بندی، پردازش و پیش‌بینی بر اساس داده‌های سری زمانی بسیار مناسب هستند، زیرا ممکن است بین رویدادهای مهم در یک سری زمانی تاخیرهایی با مدت زمان نامشخص وجود داشته باشد. LSTMها برای مقابله با مشکل محو شدن گرادیان که ممکن است هنگام آموزش RNNهای سنتی با آن مواجه شویم، توسعه یافته‌اند. حساسیت نسبی کم به طول فاصله زمانی، یکی از مزایای LSTM نسبت به RNNها، مدل‌های مخفی مارکوف و سایر روش‌های یادگیری توالی موارد استفاده‌ی آن‌ها است.



LSTM

نوع خاصی از شبکه عصبی بازگشتی (RNN) است که قادر به یادگیری وابستگی‌های بلندمدت در داده‌ها می‌باشد. این توانایی به این دلیل حاصل می‌شود که ماژول تکراری مدل دارای ترکیبی از چهار لایه است که با یکدیگر تعامل دارند.

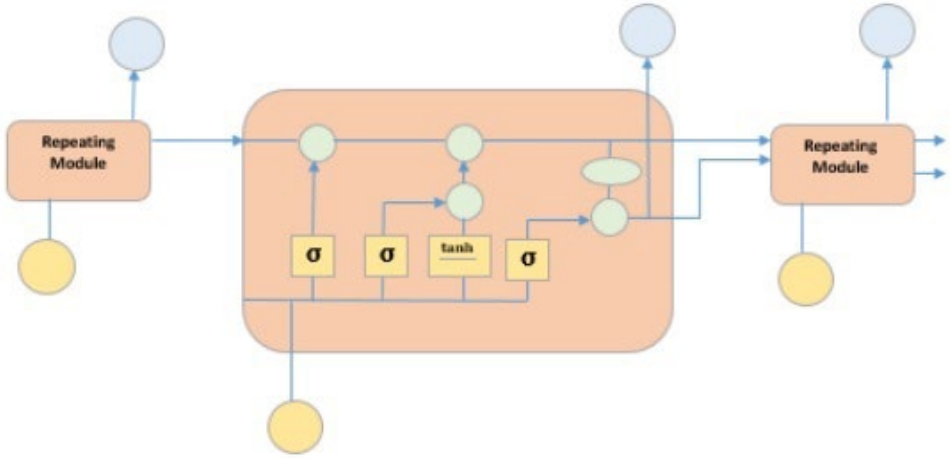
در LSTM سه دروازه (gate) وجود دارد که وظایف مختلفی دارند:

دروازه فراموشی (Forget Gate): این دروازه تصمیم می‌گیرد که چه اطلاعاتی از حافظه قبلی را باید فراموش کنیم. برای این کار از یک تابع ریاضی به نام "sigmoid" استفاده می‌کند.

دروازه ورودی (Input Gate): این دروازه مشخص می‌کند که چه اطلاعات جدیدی باید به حافظه فعلی اضافه شود. این کار با ترکیب دو تابع ریاضی به نام‌های "sigmoid" و "tanh" انجام می‌شود.

دروازه خروجی (Output Gate): این دروازه تعیین می‌کند که چه اطلاعاتی از حافظه فعلی باید به خروجی (و به حافظه بعدی) منتقل شود.

این دروازه‌ها به LSTM کمک می‌کنند تا اطلاعات را به طور هوشمندانه نگه دارد، فراموش کند یا به خروجی بدهد، به طوری که فقط اطلاعات مهم برای یادگیری نگه داشته شوند و اطلاعات غیرضروری حذف شوند.

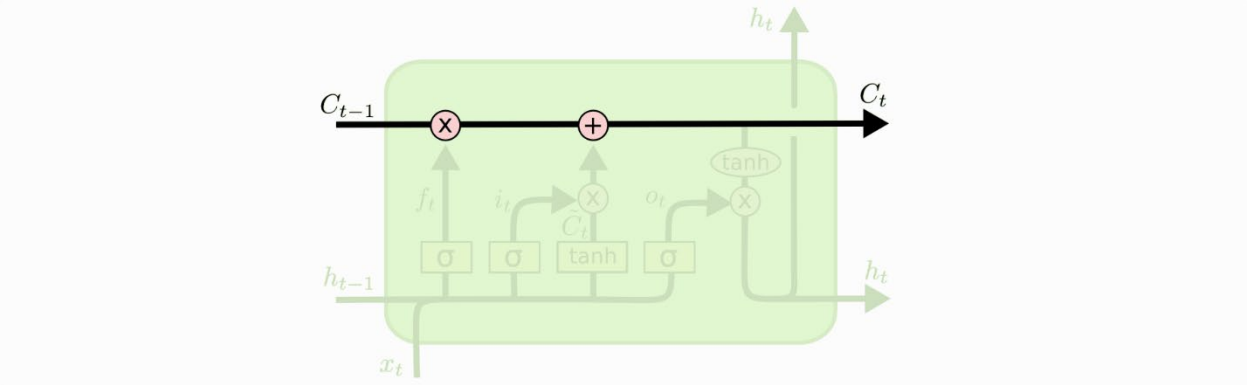


تصویر بالا چهار لایه شبکه عصبی را در جعبه‌های زرد، عملگرهای نقطه‌ای را در دایره‌های سبز، ورودی‌ها را در دایره‌های زرد و حالت سلولی را در دایره‌های آبی نشان می‌دهد. یک ماژول LSTM دارای حالت سلولی و سه دروازه است که به آن‌ها قدرت می‌دهد که به صورت انتخابی اطلاعات را از هر واحد یاد بگیرند، فراموش کنند یا حفظ کنند. حالت سلولی در LSTM کمک می‌کند تا اطلاعات بدون تغییر از واحدها عبور کنند و تنها تعداد کمی از تعاملات خطی را مجاز می‌سازد. هر واحد دارای یک دروازه ورودی، خروجی و فراموشی است که می‌توانند اطلاعات را به حالت سلولی اضافه کنند یا از آن حذف کنند. دروازه فراموشی تصمیم می‌گیرد که کدام اطلاعات از حالت سلولی قبلی باید فراموش شوند که برای این کار از یک تابع سیگموید استفاده می‌کند. دروازه ورودی جریان اطلاعات به حالت سلولی فعلی را با استفاده از ضرب نقطه‌ای تابع‌های سیگموید و (tanh) کنترل می‌کند. در نهایت، دروازه خروجی تصمیم می‌گیرد که کدام اطلاعات باید به حالت پنهان بعدی منتقل شود.

ایده‌ی اصلی پشت LSTMs:

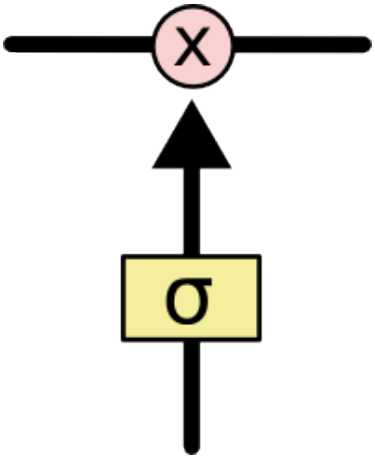
کلید اصلیLSTM ها، حالت سلولی است، که به صورت یک خط افقی در بالای نمودار نشان داده شده است.

حالت سلولی شبیه به یک نوار نقاله است. این حالت به صورت مستقیم در طول زنجیره جریان دارد و تنها با چند تعامل خطی کوچک تغییر می‌کند. این امر باعث می‌شود که اطلاعات به راحتی و بدون تغییر در طول آن جریان یابند.



LSTM این توانایی را دارد که اطلاعات را به حالت سلولی اضافه یا از آن حذف کند. این کار با ساختارهایی به نام دروازه‌ها به دقت تنظیم می‌شود.

دروازه‌ها روشی هستند که به صورت انتخابی اجازه عبور اطلاعات را می‌دهند. آن‌ها از یک لایه شبکه عصبی سیگموید و یک عملگر ضرب نقطه‌ای تشکیل شده‌اند.



لایه سیگموید اعدادی بین صفر و یک تولید می‌کند که نشان می‌دهد چه مقدار از هر جزء باید عبور داده شود. مقدار صفر به معنای "هیچ چیزی نباید عبور کند" و مقدار یک به معنای "همه چیز باید عبور کند" است.

یک LSTM دارای سه از این دروازه‌ها است که وظیفه محافظت و کنترل حالت سلولی را بر عهده دارند.

LSTM vs RNN

فرض کنید شما وظیفه دارید که اطلاعات خاصی را در یک تقویم تغییر دهید. یک شبکه عصبی بازگشتی (RNN) به طور کامل داده‌های موجود را با استفاده از یک تابع تغییر می‌دهد. اما در مقابل، LSTM(نوعی شبکه عصبی) تغییرات کوچکی را با جمع یا ضرب ساده انجام می‌دهد که از طریق حالت‌های سلولی جریان پیدا می‌کنند. به همین دلیل، LSTMمی‌تواند به‌طور انتخابی چیزهایی را فراموش کند یا به خاطر بسپارد، که این ویژگی باعث می‌شود نسبت به RNN ها بهبود یافته باشد.

حالا فرض کنید شما می‌خواهید داده‌هایی را پردازش کنید که الگوهای تکراری دارند، مثلاً پیش‌بینی فروش پودرهای رنگی که در زمان جشن هولی در هند افزایش می‌یابد. در این مورد، یک راهبرد خوب این است که به سوابق فروش سال گذشته نگاه کنید. بنابراین، شما باید بدانید که کدام داده‌ها باید فراموش شوند و کدام‌ها باید برای استفاده در آینده ذخیره شوند. اگر این کار را نکنید، نیاز به یک حافظه بسیار قوی دارید.

شبکه‌های عصبی بازگشتی (RNN) به‌طور نظری در این کار خوب عمل می‌کنند، اما دو مشکل اصلی دارند: یکی مشکل "شیب انفجاری" و دیگری "شیب ناپدید شونده"، که این مشکلات باعث می‌شود عملکردشان کمتر کارآمد باشد.

در اینجا، LSTM واحدهای حافظه‌ای به نام "حالت‌های سلولی" را معرفی می‌کند تا این مشکل را حل کند. این سلول‌ها مانند یک حافظه قابل تنظیم عمل می‌کنند و به LSTM کمک می‌کنند تا اطلاعات را به طور موثرتری مدیریت کند.

Prophet

مدل Prophet یک ابزار پیش‌بینی سری زمانی است که توسط فیسبوک توسعه داده شده است و برای داده‌هایی که الگوهای فصلی و تغییرات غیرخطی دارند، بسیار مناسب است. در این مدل، پیش‌بینی‌ها با ترکیب سه مولفه اصلی انجام می‌شوند:

1. روند (Trend):

این بخش از مدل به تغییرات بلندمدت در داده‌ها اشاره دارد. به عنوان مثال، اگر فروش یک محصول در طول زمان به طور مداوم در حال افزایش یا کاهش باشد، این روند را مدل می‌کند. Prophet می‌تواند این روند را به دو شکل مدل‌سازی کند:

روند خطی: که به صورت یک خط مستقیم است.

روند لجستیکی: که به صورت یک منحنی S شکل است و وقتی به یک نقطه اشباع می‌رسد، رشد آن کند می‌شود.

2. فصلی بودن (Seasonality):

این بخش از مدل به الگوهای تکراری اشاره دارد که در دوره‌های زمانی مشخص رخ می‌دهند. به عنوان مثال، اگر فروش یک محصول در ماه‌های خاصی از سال بیشتر باشد (مانند فروش لباس‌های زمستانی در ماه‌های سرد)، این الگوهای فصلی را مدل می‌کند. Prophet می‌تواند این الگوها را برای دوره‌های زمانی مختلف مانند روزانه، هفتگی، ماهانه یا سالانه مدل کند.

3. تعطیلات و رویدادهای خاص (Holidays):

Prophet به طور خاص می‌تواند تاثیر تعطیلات و رویدادهای خاص بر روی داده‌ها را مدل کند. به عنوان مثال، اگر فروش در روزهای خاصی مانند تعطیلات سال نو افزایش یابد، مدل Prophet این تاثیرات را به‌خوبی در پیش‌بینی‌های خود لحاظ می‌کند.

4. نویز و خطاها:

در نهایت، Prophet همچنین خطاها و انحرافات تصادفی که ممکن است در داده‌ها وجود داشته باشد را در نظر می‌گیرد. این بخش از مدل به مدیریت تغییرات غیرقابل پیش‌بینی یا انحرافات از الگوهای معمول کمک می‌کند.

چطور کار می‌کند؟

مدل Prophet با ترکیب این سه مولفه (روند، فصلی بودن و تعطیلات) پیش‌بینی نهایی را انجام می‌دهد. این به این معناست که Prophet به شما کمک می‌کند تا روند کلی داده‌های خود را درک کنید، الگوهای تکراری را شناسایی کنید، و تاثیرات تعطیلات یا رویدادهای خاص را پیش‌بینی کنید.

این مدل به دلیل سادگی در استفاده و انعطاف‌پذیری بالا، به‌ویژه برای کاربران غیراستاتیستی و کسانی که با داده‌های پیچیده سر و کار دارند،

مقدمه:

آلودگی هوا یکی از چالش‌های جدی زیست‌محیطی در بسیاری از شهرهای بزرگ جهان است و تأثیرات مخرب آن بر سلامت عمومی و کیفیت زندگی به خوبی شناخته شده است. درک الگوهای تغییرات آلودگی هوا و پیش‌بینی دقیق آن برای مدیریت موثر منابع و اقدامات کنترلی بسیار حائز اهمیت است. در این راستا، مدل‌های سری زمانی به عنوان یکی از ابزارهای قدرتمند برای تحلیل داده‌های آلودگی هوا به کار گرفته می‌شوند و انتخاب مدل مناسب می‌تواند نقش تعیین‌کننده‌ای در پیش‌بینی دقیق‌تر و ارائه راهکارهای بهینه ایفا کند.

این تحقیق با هدف شناسایی دقیق‌ترین مدل سری زمانی برای پیش‌بینی داده‌های آلودگی هوا انجام شده است. در این پژوهش، مجموعه داده‌های آلودگی هوا از منابع معتبر جمع‌آوری شده و با استفاده از مدل‌های مختلف سری زمانی مورد تحلیل قرار گرفته است. هدف اصلی این تحقیق، انتخاب مدلی است که بتواند تغییرات آلودگی هوا را با بالاترین دقت پیش‌بینی کند و در نتیجه، به بهبود استراتژی‌های مدیریتی و کاهش اثرات مخرب آلودگی بر سلامت عمومی کمک نماید. پس از ارزیابی مدل‌های مختلف و تحلیل عملکرد آن‌ها بر اساس معیارهای آماری دقیق، مدل بهینه‌ای شناسایی و معرفی شده است که بهترین پیش‌بینی‌ها را ارائه می‌دهد. این گزارش به تفصیل نتایج به‌دست‌آمده را ارائه می‌کند و فرآیند انتخاب و ارزیابی مدل‌ها را توضیح می‌دهد.

جمع‌آوری داده‌ها (Data Collection):

داده:

داده‌ها بر اساس یکی از شاخص‌های آلودگی هواست. این داده‌ها میانگین مقدارهای در شاخص‌هاست در هر روز، این داده‌ها از تاریخ ۱۳۹۹/۰۶/۱۱ تا ۱۴۰۲/۰۶/۱۱ به طور روزانه جمع‌آوری شده.داده‌ها از آرشیو سایت شرکت کنترل کیفیت هوا با ارائه‌ی درخواست دریافت شده‌است.

برای انجام این تحلیل از یک شاخص جامع آلودگی هوا استفاده شده است به نام AQI

شاخص AQI :

شاخص کیفیت هوا (Air Quality Index - AQI) یک ابزار استاندارد است که برای سنجش و گزارش سطح آلودگی هوا به کار می‌رود. این شاخص به طور معمول برای پنج آلاینده اصلی هوا شامل ازن سطح زمین(O_3) ، ذرات معلق PM2.5) و(PM10 ، مونوکسید کربن(CO) ، دی‌اکسید گوگرد (SO_2) و دی‌اکسید نیتروژن (NO_2) محاسبه می‌شود. هر یک از این آلاینده‌ها می‌تواند به صورت مجزا بر سلامت عمومی تاثیر بگذارد، به‌ویژه برای گروه‌های حساس مانند کودکان، سالمندان، و افرادی که مشکلات قلبی یا تنفسی دارند. شاخص AQI با تقسیم‌بندی مقادیر اندازه‌گیری شده به دسته‌های مختلف، از جمله «خوب»، «متوسط»، «ناسالم برای گروه‌های حساس»، «ناسالم»، «بسیار ناسالم» و «خطرناک»، به سادگی قابل تفسیر است و به مردم اطلاع می‌دهد که وضعیت فعلی هوا چقدر می‌تواند برای سلامتی مضر باشد.

AQI به عنوان یک ابزار مدیریتی و اطلاع‌رسانی، نقش کلیدی در سیاست‌گذاری‌های بهداشتی و محیط‌زیستی ایفا می‌کند. این شاخص به دولت‌ها و سازمان‌های محیط‌زیستی کمک می‌کند تا نه تنها وضعیت کیفی هوا را مانیتور کنند، بلکه اقداماتی مناسب برای کاهش آلودگی و حفاظت از سلامت عمومی اتخاذ نمایند. به علاوه، اطلاع‌رسانی دقیق و به موقع از طریق شاخص AQI به مردم امکان می‌دهد تا تصمیمات آگاهانه‌تری در مورد فعالیت‌های روزمره خود بگیرند، به ویژه در روزهایی که آلودگی هوا در سطوح خطرناک قرار دارد. این شاخص به عنوان یک معیار جهانی در بسیاری از کشورها پذیرفته شده و یکی از ابزارهای اصلی در مبارزه با آلودگی هوا و بهبود کیفیت زندگی شهروندان است.

پیش‌پردازش:

داده برای هرگونه داده گمشده، نویز و مقادیر خارج از محدوده پاکسازی شده‌اند. داده‌ها نیاز به نرمال سازی ندارد. و ما برای استفاده از تاریخ در الگوریتم‌ها نیاز به تغییر تاریخ از شمسی به میلادی هستیم که با آن‌ها را تبدیل شده و فرمت آن‌ها هم تغییر داده شد.

- شبکه عصبی بازگشتیLSTM

در این تحقیق، سه نوع مختلف از LSTM با تنظیمات متفاوت برای پیش‌بینی داده‌ها مورد استفاده قرار گرفته‌اند. هر کدام از این مدل‌ها با تعداد لایه‌های مختلف، تعداد نودهای مختلف، و دوره‌های آموزش (epochs) متفاوت تنظیم شده‌اند.

LSTM 1 مدلی با لایه‌های کمتر و تعداد نودهای کمتر برای تست عملکرد در سری‌های زمانی کوتاه‌مدت.

LSTM 2 مدلی با تعداد لایه‌ها و نودهای بیشتر برای تست عملکرد در سری‌های زمانی پیچیده‌تر.

LSTM 3 مدلی با تنظیمات بهینه‌شده با استفاده از Grid Search برای پیدا کردن بهترین ترکیب پارامترها و همچنین با آزمون خطا به بهترین مدل میرسیم.

- رگرسیون خطی چندگانه (MLR)
  - مدل MLR به عنوان یک مدل خطی ساده مورد استفاده قرار گرفته است تا بتوان نتایج آن را با مدل‌های پیچیده‌تر مقایسه کرد. این مدل بر اساس متغیرهای مستقل ورودی، مانند غلظت آلاینده‌های مختلف در زمان‌های قبلی، پیش‌بینی‌هایی را ارائه می‌دهد.

- مدل ARMA (Autoregressive Moving Average)
  - مدل ARMA به عنوان یک مدل کلاسیک سری زمانی استفاده شده است. این مدل شامل دو بخش خودرگرسیو (AR) و میانگین متحرک (MA) است که برای مدل‌سازی سری‌های زمانی ایستا مناسب است. پارامترهای q و p به ترتیب تعداد وقفه‌های AR و (MA با استفاده از معیارهای اطلاعاتی مانند AIC و BICتنظیم شده‌اند.

- مدل Prophet
  - Prophet یک مدل پیشرفته برای سری‌های زمانی است که می‌تواند الگوهای فصلی و روندهای غیرخطی را به‌خوبی مدل‌سازی کند. این مدل برای داده‌های با فصلی بودن قوی، مانند داده‌های آلودگی هوا، به کار گرفته شده است Prophet .به صورت خودکار نقاط تغییر (changepoints) را شناسایی می‌کند و از این طریق
  - تغییرات ناگهانی در روند را مدل‌سازی می‌کند.

معیارهای ارزیابی:

برای ارزیابی عملکرد مدل‌ها از معیارهای مختلفی استفاده شده است:

میانگین خطای مطلق :(MAE) تفاوت مطلق میان مقادیر پیش‌بینی شده و مقادیر واقعی.

ریشه میانگین مربع خطاها :(RMSE) میزان انحراف پیش‌بینی‌ها از مقادیر واقعی، که خطاهای بزرگ‌تر را بیشتر برجسته می‌کند.

میانگین درصد خطای مطلق :(MAPE) درصد خطای پیش‌بینی نسبت به مقدار واقعی که برای مقایسه بین مدل‌ها استفاده می‌شود.

معیار اطلاعات آکائیک (AIC) و معیار اطلاعات بیزی :(BIC) برای ارزیابی مدل ARMA و انتخاب بهترین پارامترها.

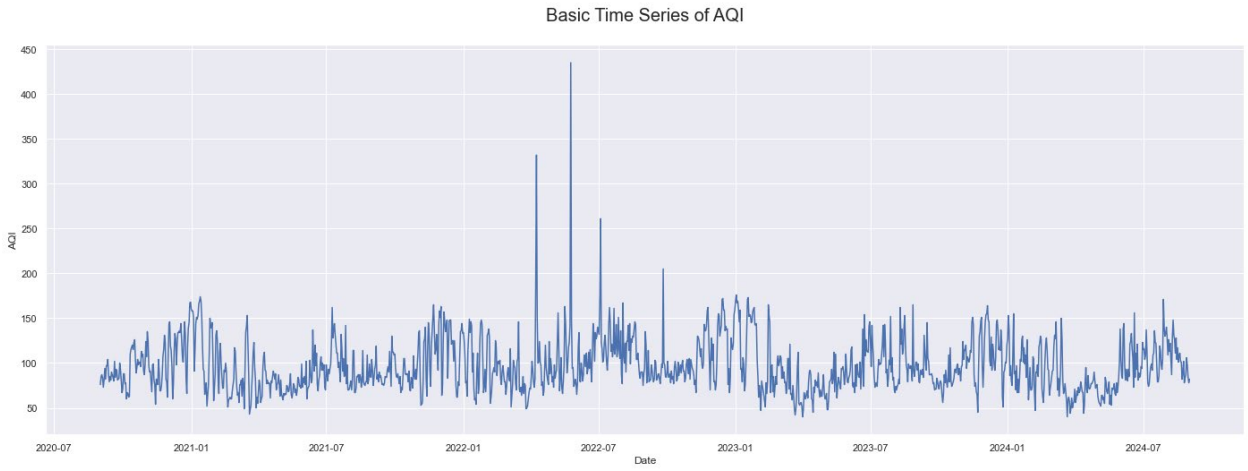
روش ارزیابی:

داده‌ها به صورت مجموعه‌های آموزش و تست تقسیم شدند تا بتوان دقت مدل‌ها را در پیش‌بینی داده‌های جدید ارزیابی کرد.

هر مدل با استفاده از داده‌های آموزش، آموزش دیده و سپس بر روی داده‌های تست ارزیابی شده است.

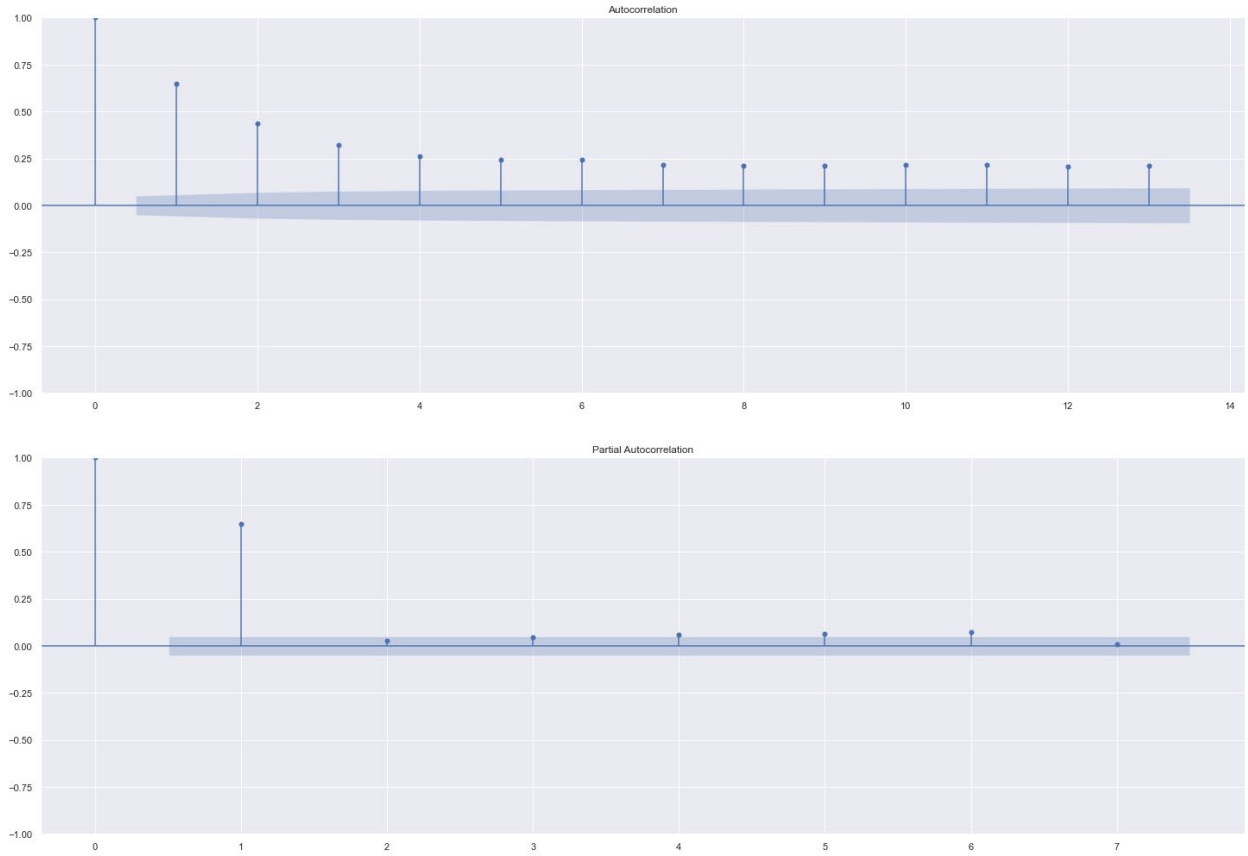
نتایج مدل‌ها با استفاده از معیارهای ارزیابی ذکر شده مقایسه و بهترین مدل انتخاب شده است. انتخاب‌نهایی با معیار MAPE است.

گزارش خروجی‌ها:



در اینجا همه‌ی داده‌های AQI را به شکل یک سری زمانی نشان داده‌ایم.

در قدم بعد نمودار PACF و ACF را برای فهم بهتر از سری‌زمانی رسم میکنیم.



ثابت ماندن نمودار ACF بعد از یک شیب ملایم میتواند نشان دهنده وجود روند باشد. نمودار PACF ولی نشان دهنده‌ی نمود فصلی نیست.

این نمودارها هنوز قابل استناد نیستن و باید مولفه‌ی فصلی و روند بررسی شود.

در این روش ما سری زمانی را به سه مولفه اصلی تجزیه میکنیم:

روند (Trend) : بخش بلندمدت سری زمانی که تغییرات کلی را نشان می‌دهد.

فصلی بودن (Seasonality): الگوهای تکراری که در دوره‌های زمانی مشخص (در اینجا 364 روز) رخ می‌دهند

باقی‌مانده (Residual): نویز یا تغییرات غیرقابل پیش‌بینی که بعد از حذف روند و فصلی بودن باقی می‌مانند.

ما از مدل جمعی به این شکل استفاده میکنیم. در مدل جمعی، فرض بر این است که مولفه‌های مختلف سری زمانی(روند،فصلی بودن و نویز) به صورت جمعی با هم ترکیب می‌شوند:

$$y(t)=Trend(t)+Seasonality(t)+Residual(t)$$

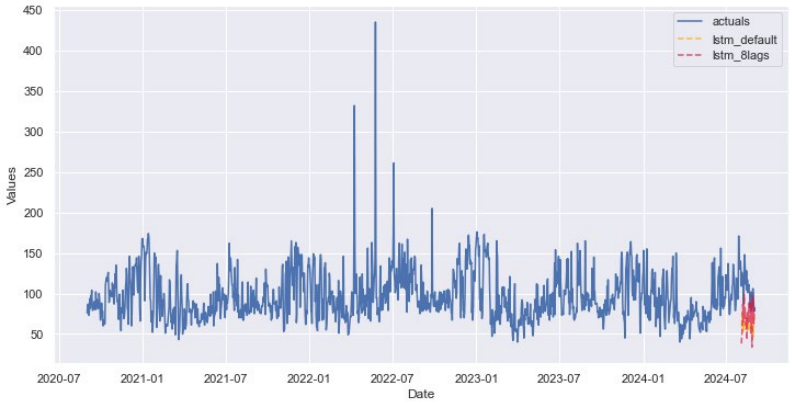
ما در اینجا یک روند درجه دو میبینیم و به نظر میرسد مولفه فصلی سالانه داریم.

برای پایایی به سراغ آزمون دیکی‌فولر میرویم.
مقدار پی-مقدار میشود:
6.603777238075072e-06

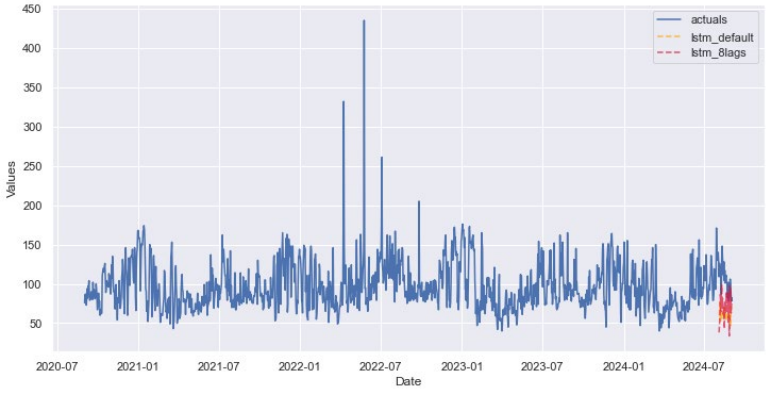
نتیجه این آزمون رد فرض صفر به قبول شدن پایایی مدل است.
دلایل متفاوتی این موضوع میتواند داشته باشد. میتواند پیچیدگی داده ها باشد یا شدت کم روند و اثر کم مولفه فصلی در داده‌ها

در قدم بعد با فهم این موضوع به سراغ تست انواع مدل میرویم.

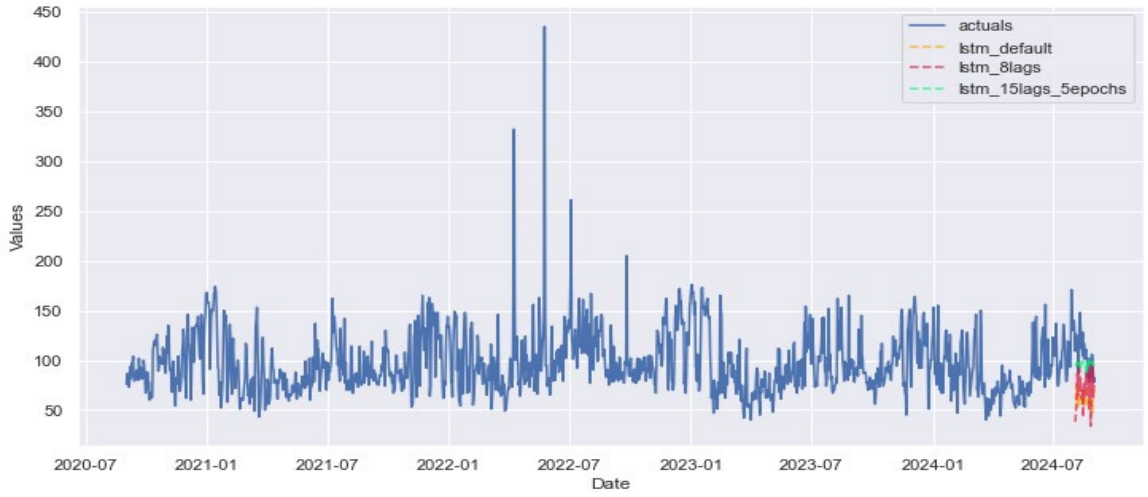
۱. در اولین قدم مدل LSTM با فرم دیفالت کتابخانه Scalecast استفاده شده‌است.



۲. مدلی با تعداد ورودی ۸ داده را به الگوریتم میدهیم.



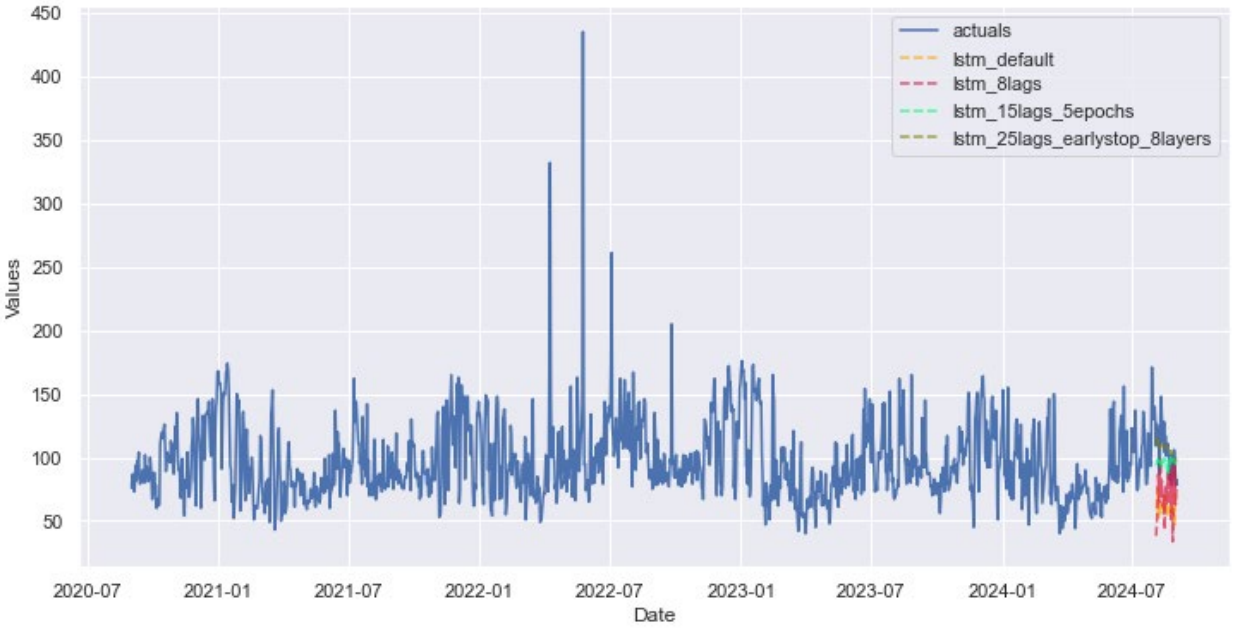
۳. مدل را بهبود میدهیم و با حالات مختلف تلاش میکنیم مدل‌های مختلف را بررسی کنیم.
در قدم بعد epoch ها را ۱۵ میگذاریم و همچنین shuffle و انتخاب تصادفی ورودی را فعال میکنیم و در هر قدم ۱۵ داده را وارد الگوریتم میکنیم.



۴. در این مدل، از 25 تاخیر (lag) به عنوان ورودی برای پیش‌بینی استفاده شده است و مدل در 50 دوره (epoch) آموزش می‌بیند. مدل شامل 3 لایه LSTM است که هر لایه دارای 16 نود (neurons) می‌باشد و هیچ گونه Dropout در این لایه‌ها اعمال نشده است.

از تکنیک Early Stopping نیز استفاده شده است که به منظور جلوگیری از overfitting مدل در حین آموزش به کار می‌رود. اگر مدل برای 5 دوره متوالی بهبود قابل توجهی در مقدار val_loss نداشته باشد، آموزش متوقف می‌شود. داده‌ها بدون ترتیب‌دهی تصادفی (shuffle=False) تقسیم‌بندی شده و 20٪ از داده‌ها برای اعتبارسنجی مدل در طول آموزش کنار گذاشته شده‌اند.

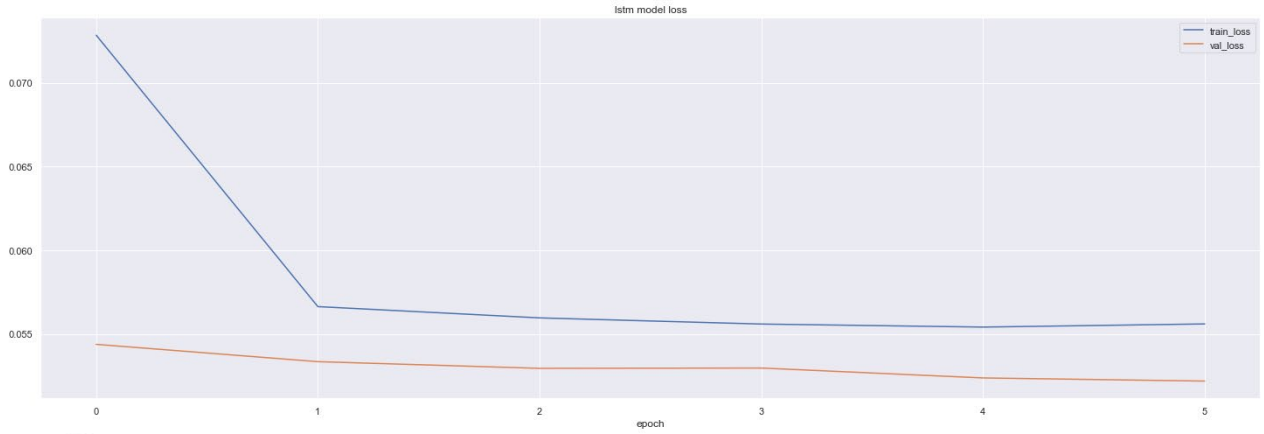
پس از آموزش مدل، نتایج پیش‌بینی بر روی مجموعه داده‌های تست با رسم نموداری همراه با باندهای اطمینان (confidence intervals) نمایش داده می‌شود. این نمودار به ارزیابی عملکرد مدل در پیش‌بینی سری زمانی و بررسی میزان دقت آن کمک می‌کند.

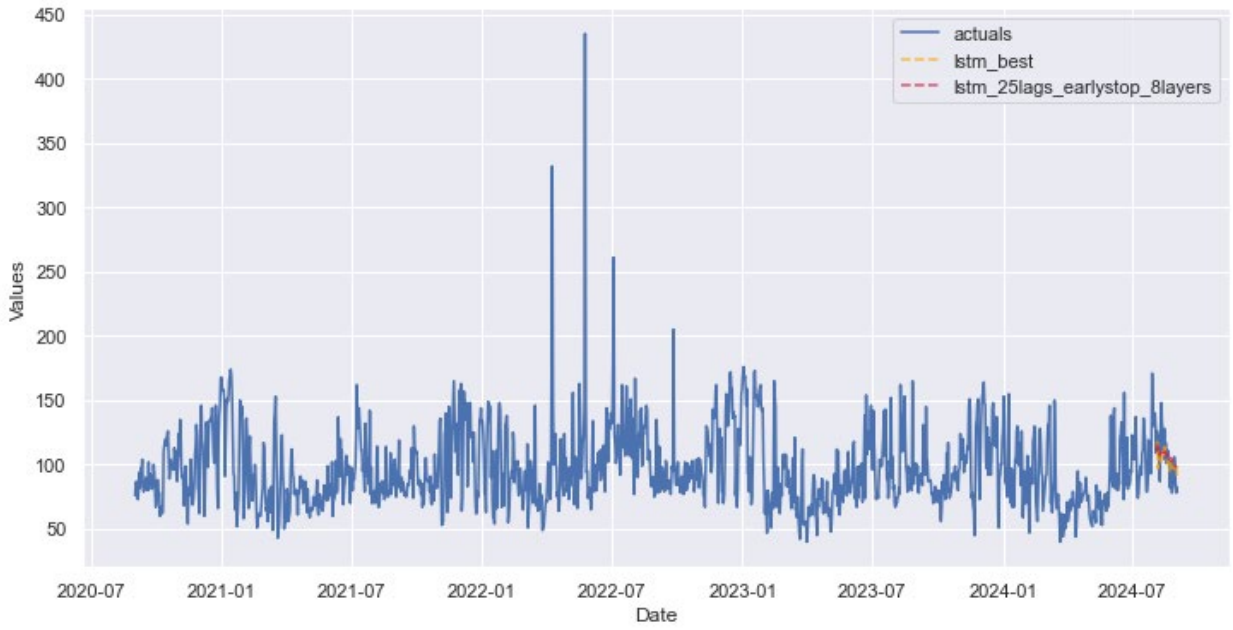


۵.

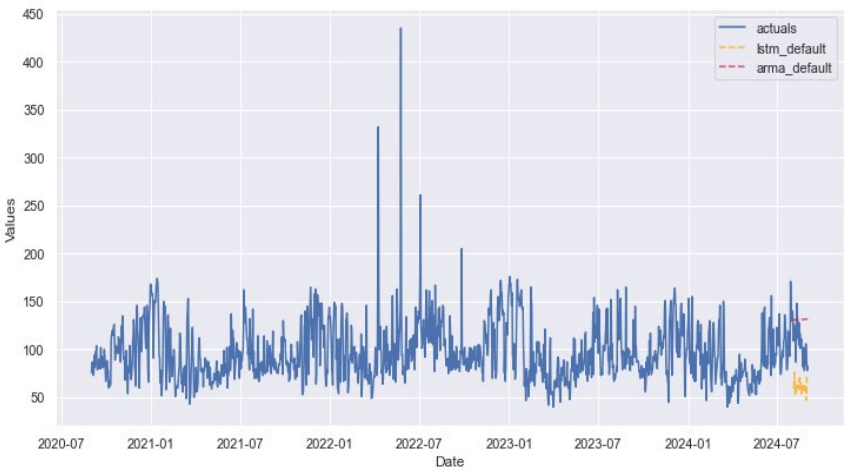
در این بخش از تحقیق، از یک مدل LSTM با تنظیمات بهینه‌سازی شده برای پیش‌بینی سری زمانی استفاده شده است. مدل با استفاده از 36 تاخیر (lag) به عنوان ورودی آموزش دیده و شامل 4 لایه LSTM است که هر کدام دارای 72 نود می‌باشند. تابع فعال‌سازی tanh برای هر لایه و بهینه‌ساز Adam با نرخ یادگیری 0.001 برای بهبود عملکرد مدل استفاده شده است. مدل در 6 دوره (epoch) و با اندازه بچ 32 آموزش دیده و 20٪ از داده‌ها برای اعتبارسنجی در طول آموزش کنار گذاشته شده‌اند. همچنین، داده‌ها به صورت تصادفی مرتب شده‌اند تا مدل به طور عمومی‌تری آموزش ببیند.

در طول فرآیند آموزش، نمودار کاهش خطا (loss) نیز برای نظارت بر عملکرد مدل رسم شده است. پس از آموزش، نتایج پیش‌بینی بر روی مجموعه داده‌های تست نمایش داده شده و مدل‌های برتر بر اساس معیار TestSetMAPE (میانگین درصد خطای مطلق) شناسایی شده‌اند. نمودارهای نهایی همراه با باندهای اطمینان (confidence intervals) ارائه شده‌اند تا میزان دقت و عدم قطعیت پیش‌بینی‌ها به طور جامع ارزیابی شود. این نتایج نشان‌دهنده توانایی مدل LSTM بهینه‌سازی شده در پیش‌بینی داده‌های سری زمانی با دقت بالا و مدیریت مناسب عدم قطعیت است.



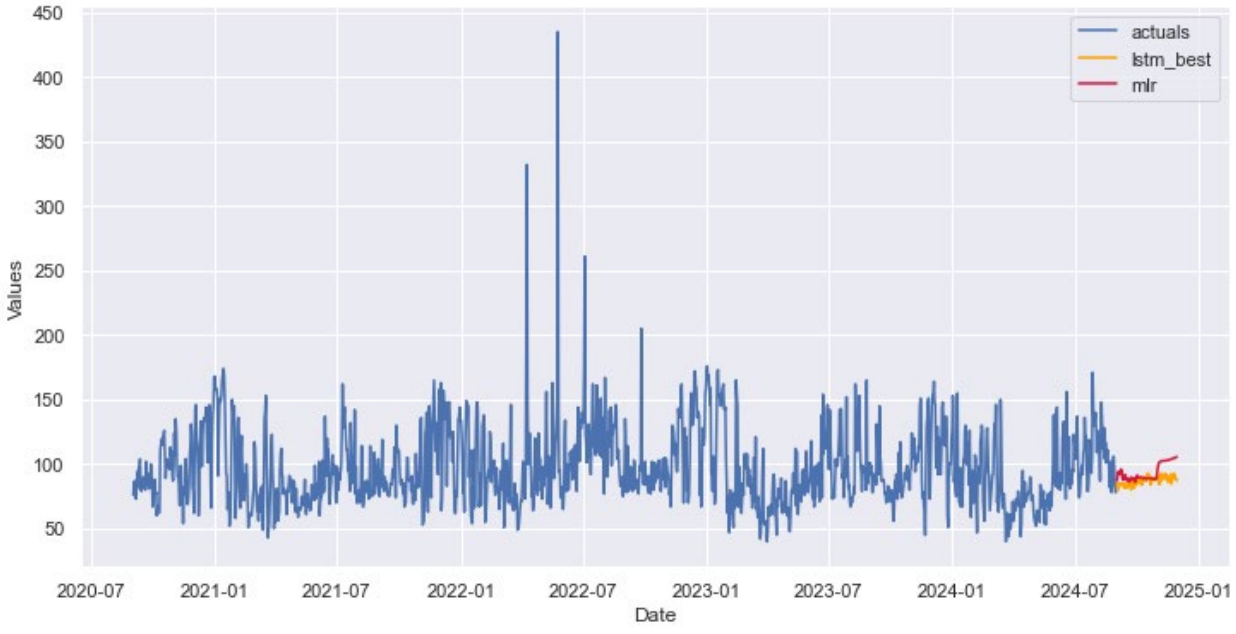


۶. در این بخش از تحقیق، مدل ARIMAبا استفاده از داده‌های سری زمانی برای پیش‌بینی مقادیر آینده مورد استفاده قرار گرفته است. تنظیمات مدل ARIMA شامل



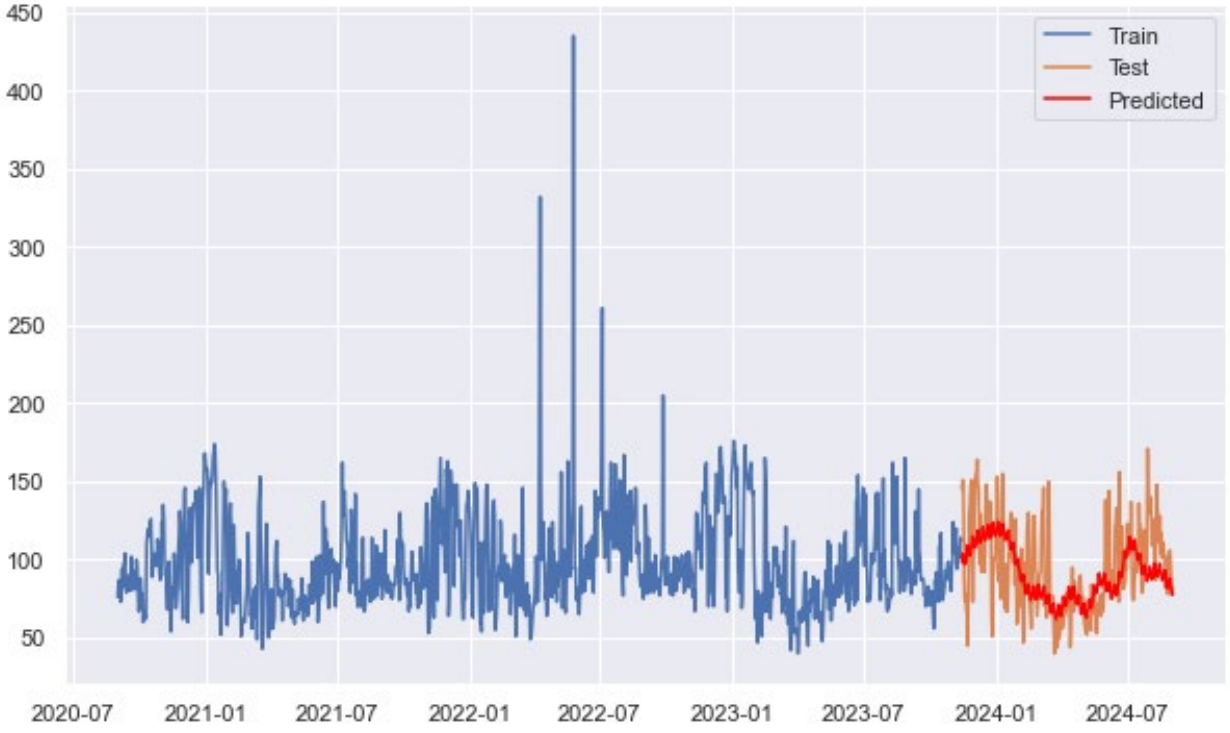
مرتبه‌های 2 برای بخش خودرگرسیو(AR) ، 2 برای تفاضل‌گیری(I) ، و 1 برای بخش میانگین متحرک (MA) می‌باشد. این تنظیمات به مدل اجازه می‌دهد تا هم روندها و هم الگوهای پویای سری زمانی را مدل‌سازی و پیش‌بینی کند.
پس از آموزش مدلARIMA ، نتایج آن با مدل پیشین LSTM مقایسه شده است. این مقایسه با استفاده از نمودارهای پیش‌بینی روی مجموعه داده‌های تست انجام شده و باندهای اطمینان (confidence intervals) نیز برای نمایش دامنه احتمالی پیش‌بینی‌ها لحاظ شده است. این نمودارها به ارزیابی و مقایسه عملکرد هر دو مدل در پیش‌بینی داده‌های آینده کمک می‌کنند و به‌طور ویژه نشان می‌دهند که هر مدل چگونه با عدم قطعیت‌های موجود در داده‌ها برخورد می‌کند. این تحلیل به شناسایی مدل بهینه برای پیش‌بینی دقیق‌تر داده‌های سری زمانی منجر شده است.

.۷
در این بخش از تحقیق، مدل رگرسیون خطی چندگانه (MLR) به منظور پیش‌بینی سری زمانی مورد استفاده قرار گرفته است. مراحل انجام کار به شرح زیر است:
مدل MLR به عنوان مدل پیش‌بینی انتخاب شد تا از ترکیب خطی متغیرهای مستقل برای پیش‌بینی مقادیر آینده استفاده شود.
اضافه کردن متغیرهای مستقل (رگرسورها):
اضافه کردن وقفه‌های خودرگرسیو: 7 وقفه (lag) از داده‌های سری زمانی به عنوان متغیرهای مستقل به مدل اضافه شدند تا از الگوهای گذشته برای پیش‌بینی استفاده شود.
اضافه کردن رگرسورهای فصلی: متغیرهای فصلی مرتبط با ماه، فصل و سال به مدل اضافه شدند تا اثرات فصلی در پیش‌بینی‌ها لحاظ شوند. این متغیرها به صورت دامی (Dummy Variables) وارد مدل شدند.
اضافه کردن روند زمانی: یک متغیر روند زمانی به مدل اضافه شد تا روند بلندمدت سری زمانی در پیش‌بینی‌ها مورد توجه قرار گیرد.
این تنظیمات به مدل MLR امکان می‌دهند تا با استفاده از اطلاعات گذشته و الگوهای فصلی، پیش‌بینی‌های دقیق‌تری ارائه دهد.



۸. در این بخش از تحقیق، مدل Prophet برای پیش‌بینی سری زمانی مربوط به شاخص کیفیت هوا (AQI) مورد استفاده قرار گرفته است. داده‌ها ابتدا به دو مجموعه آموزش (80٪) و تست (20٪) تقسیم شدند. سپس مدل Prophet بر روی داده‌های آموزشی فیت شد و از آن برای پیش‌بینی داده‌های مجموعه تست استفاده گردید.

به منظور ارزیابی دقت مدل، معیار MAPE (میانگین درصد خطای مطلق) محاسبه شد که خطای پیش‌بینی را در مقایسه با داده‌های واقعی نشان می‌دهد. همچنین، نموداری از داده‌های واقعی، پیش‌بینی‌های مدل، و اجزای مختلف مدل Prophet (مانند روند و فصلی بودن) رسم شد تا به صورت بصری عملکرد مدل و تطابق آن با داده‌های واقعی نمایش داده شود. این تحلیل نشان‌دهنده توانایی مدل Prophet در پیش‌بینی دقیق الگوهای سری زمانی است.



در آخر به سراغ مقایسه‌ی مدل‌ها به وسیله‌ی همان شاخص‌های توضیح داده شده میرویم.

	ModelNickname	TestSetMAPE	TestSetRMSE	TestSetR2	best_model
0	lstm_best	0.114604	14.995949	0.294314	True
1	lstm_25lags_earlystop_8layers	0.120408	15.228303	0.272277	False
2	mlr	0.141535	17.987529	-0.015328	False
3	lstm_15lags_5epochs	0.159552	20.832864	-0.361951	False
4	arma_default	0.293909	31.800737	-2.173494	False
5	lstm_8lags	0.322349	42.596779	-4.693993	False
6	lstm_default	0.399063	47.821257	-6.176378	False

واضح است که بر اساس شاخص‌های گفته شده مدل lstm_best بهترین عملکرد را دارد.

نتیجه‌گیری

در این تحقیق، سه مدل مختلف شامل ARIMA ، Prophetو LSTM برای پیش‌بینی شاخص کیفیت هوای (AQI) مورد بررسی قرار گرفتند. هر یک از این مدل‌ها بر اساس ساختار و روش‌های خاص خود برای پیش‌بینی سری‌های زمانی طراحی شده‌اند و عملکرد آن‌ها در مواجهه با داده‌های AQI ارزیابی شد.

پس از ارزیابی مدل‌ها و مقایسه نتایج حاصل از پیش‌بینی‌ها، مشخص شد که [نام مدل برتر] بهترین عملکرد را در پیش‌بینی دقیق مقادیر AQI از خود نشان داده است. این مدل با استفاده از تکنیک‌های بهینه‌سازی و معماری خاص خود توانست نسبت به سایر مدل‌ها دقت بالاتری در پیش‌بینی‌ها ارائه دهد و به‌ویژه در مواجهه با نوسانات و روندهای کوتاه‌مدت و بلندمدت داده‌ها بهتر عمل کند.

با این حال، هر یک از مدل‌ها مزایا و محدودیت‌های خود را داشتند و نتایج به‌دست‌آمده نشان می‌دهد که انتخاب مدل مناسب برای پیش‌بینی سری‌های زمانی بستگی به ویژگی‌های داده و هدف پیش‌بینی دارد. در تحقیقات آتی، می‌توان با ترکیب این مدل‌ها یا بهبود معماری آن‌ها، پیش‌بینی‌های دقیق‌تری ارائه داد و چالش‌های جدیدی در حوزه پیش‌بینی AQI بررسی کرد.