به نام خدا

محمد زیاری – (97222047)

تاریخ: 1400 /11/5

گزارش تمرین امتیازی-سری زمانی **درس: مبانی علوم داده** **استاد درس: دکتر خردپیشه**

در این تمرین دیتاست سومی که معرفی شد که لینکش را در زیر مشاهده می کنیم انتخاب شد.

https://www.kaggle.com/selfishgene/historical-hourly-weather-data

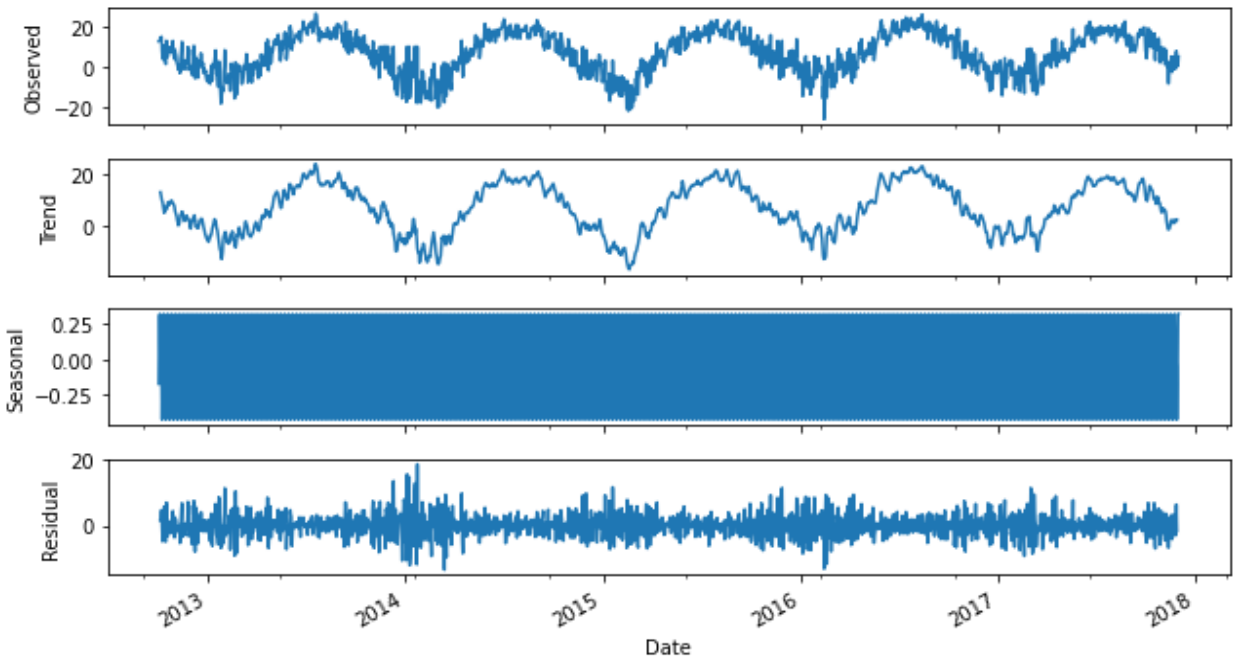
در این دیتاست باید با داده های زمانی کار می کردیم. در ابتدا باید تارگت یا در واقع فیچری که میخواهیم روی آن پیشبینی را انجام دهیم انتخاب میکنیم که فیچر Temperature است. سپس در 3 شهری که مقدار null کمتری دارند این ویژگی را بررسی می کنیم.

شهر هایی که انتخاب کردم عبارتند از Las Vegas , Toronto , Detroit.

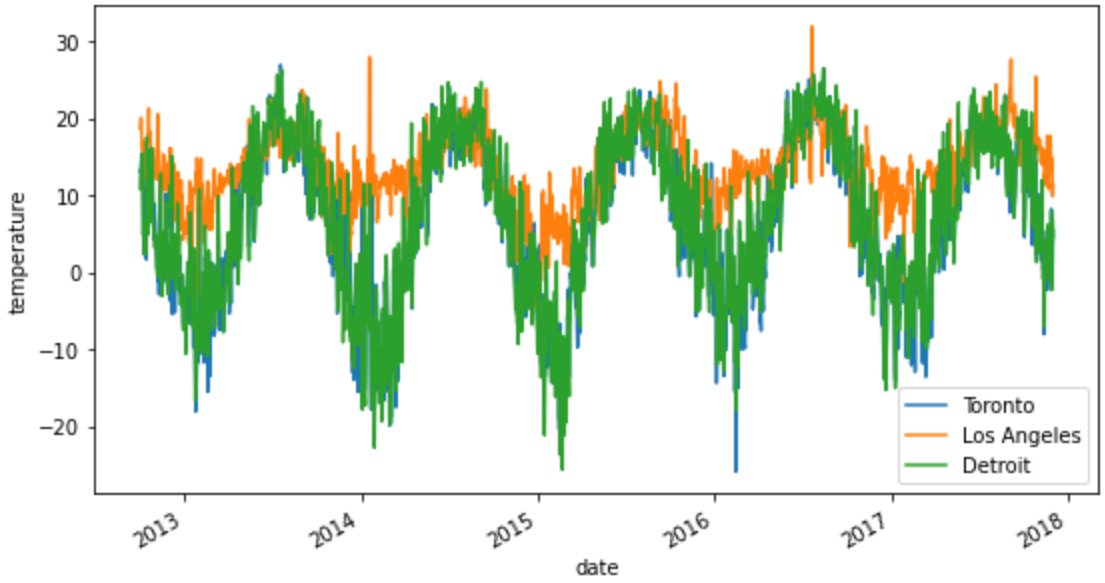
پس از دراپ کردن ستون مربوط به بقیه شهرها مواردی که در صورت سوال خواسته شده را چک می کنیم تا مشاهده کنیم داده های ما در طی گذر زمان چه رفتاری از خود نشان می دهند.

همانطور که گفته شد ابتدا تک متغیره یا univariate مدلمان را تست می کنیم و پس از گرفتن داده های دمای 3 شهر انتخاب شده ، داده هایی که مقدار null دارند را با متد bfill یعنی دمای روز بعدی پر می کنیم. دمای مورد نظر به صورت کلوین بیان شده است و ما با تغییر پارامتر آن را به سلسیوس تبدیل می کنیم. در اخر فیچر های جدید تحت عنوان hour , season , year , month اضافه می کنیم که شاید بعدا در مدل ما و بررسی تغییرات طی زمان موثر واقع شوند.

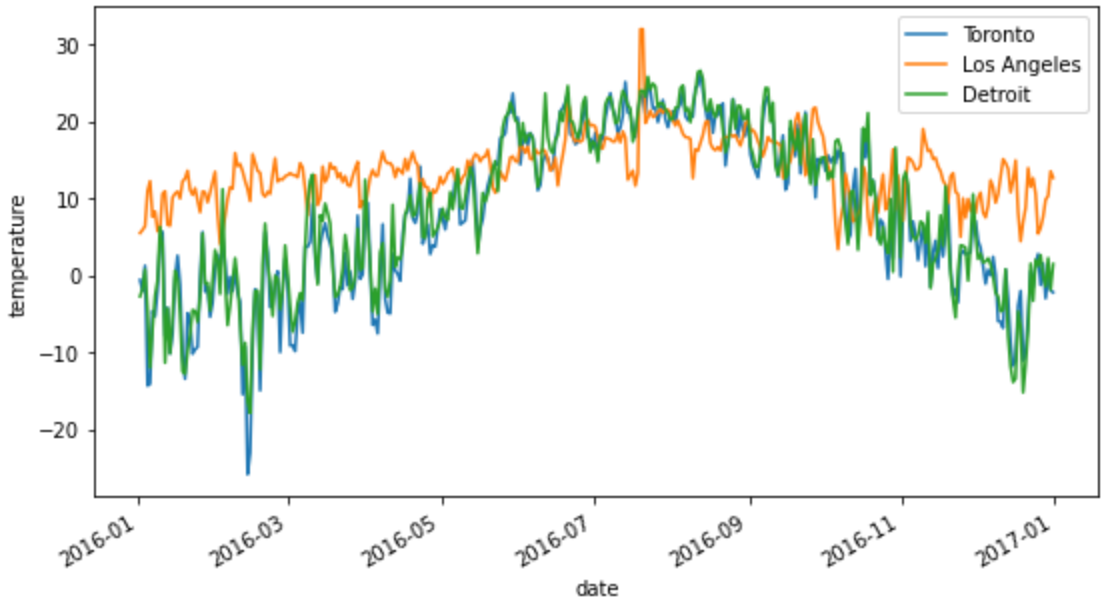
ابتدا نموداری آماده که پیدا کردم را روی داده های تورنتو تست می کنیم تا ببینیم چه تغییراتی در روی دمای این شهر در گذر زمان اتفاق می افتد.



ترند یا تغییرات دما در این چند سال به طور واضحی در شکل آمده است. هرچند نمودار seasonal به علت تغییرات مکرر در دما به خوبی نمایش داده نشده است اما با استفاده از همان نمودار ترند می توان تا حدودی میزان تغییرات در طول ماه یا فصل ها را مشاهده کرد. مشخص است در اوایل سال ها میزان دما پایین تر است و هر چه به اواسط سال که تابستان است نزدیک می شویم مقدار دما بالاتر می رود. نمودار های زیر نمایانگر تغییرات در شهر های تورنتو و دیترویت و لس آنجلس می باشد.

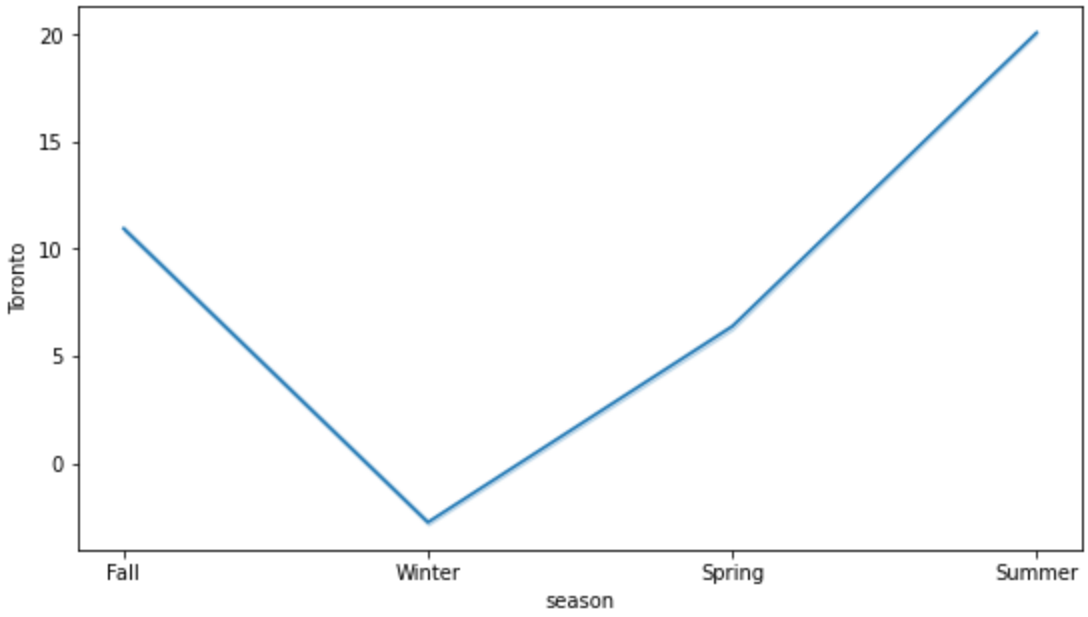


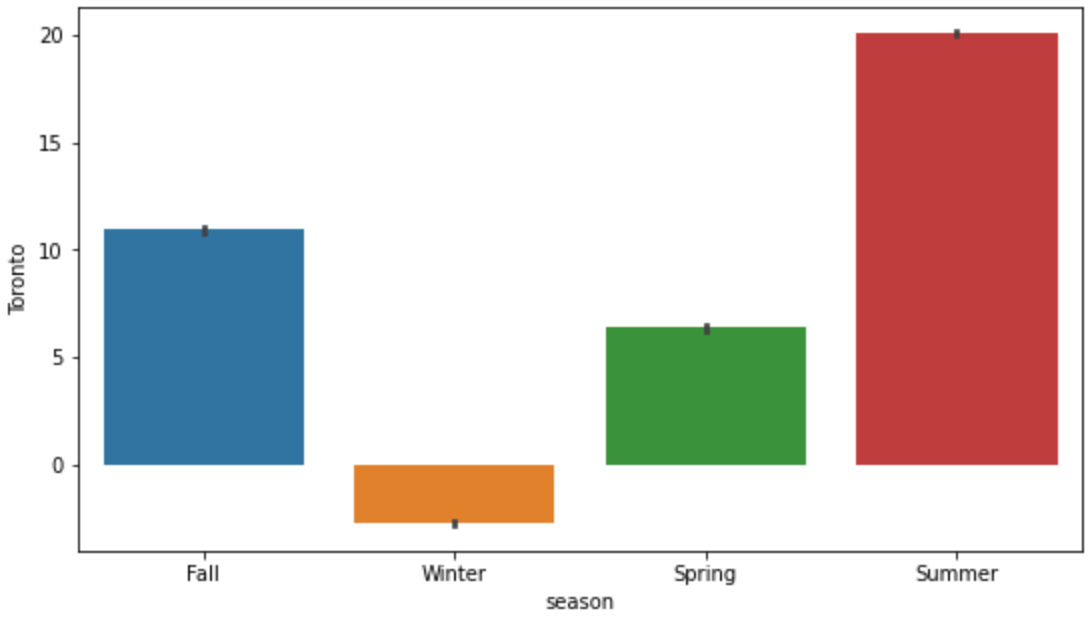
چون دامنه تغییرات بالاست به خوبی متوجه تغییرات نمی شویم. از این رو نمودار دمای این 3 شهر را در یک سال به خصوص 2016 مشاهده می کنیم.



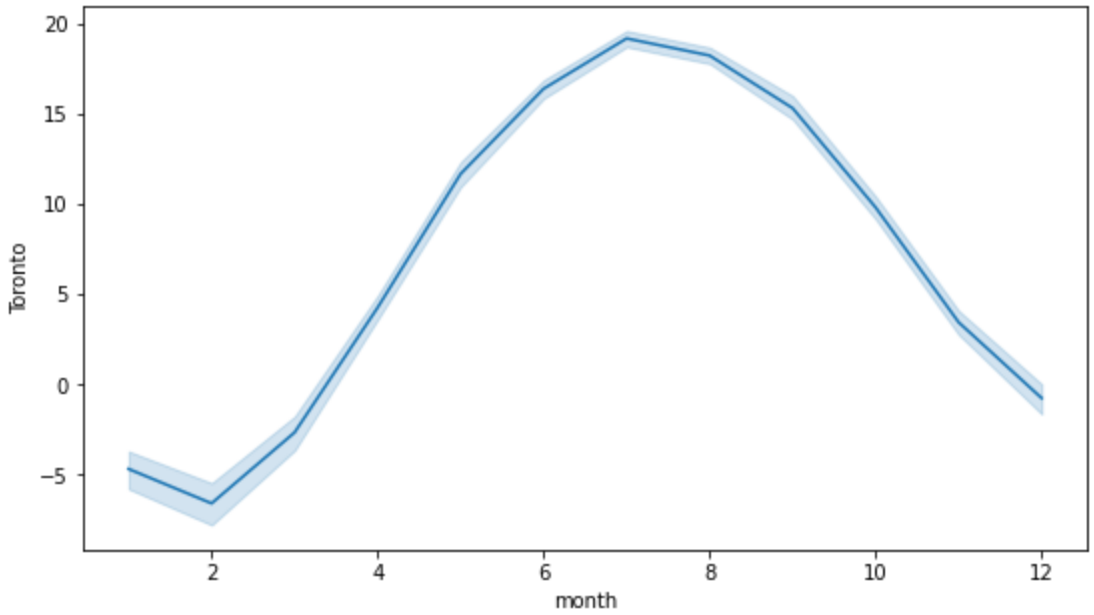
اما در خصوص تغییرات ماهانه یا فصلی یک شهر به خصوص مثلا تورنتو را انتخاب می کنیم و میزان دمای آن را در ماه و فصل های متفاوت می سنجیم. نمودار های زیر نمایانگر تغییرات در طی فصل های این شهر است که به

صورت barplot و lineplot کشیده شده است. مشخصا دما در تابستان از همه فصل ها بیشتر است و در زمستان از همه کمتر. موردی که در نمودار ترند نیز تا حدودی آن را حدس زدیم.





نمودار ماهانه نیز روایتگر یک تغییر با نظم است که تا ماه جولای که در تابستان است دما به طور پیوسته در حال افزایش است اما و پس از آن تا ماه فوریه که ماه سوم زمستان است دائما در حال کاهش است و کمترین مقدار را در آن دارد.



پس از بررسی نمودار ها نوبت به بررسی stationary بودن داده ها می رسد. با استفاده از روش ADF این کار را انجام می دهیم. پس از اعمال این روش بر روی داده ها مشاهده می کنیم که P-value برای تمامی شهر ها برابر با 0 بوده است. از این رو تا چند رقم اعشار آن ها را بررسی کردیم تا مشخص شود چه مقداری دارند.

برای انکه اولین عدد نمایش داده شود در شهر های تورنتو و دیترویت تا 9 رقم اعشار و در شهر لس آنجلس تا 14 رقم اعشار مجبور به نمایش P-value شدیم. مقادیر آماره به دست آمده را در زیر برای هر شهر مشاهده میکنیم. از آنجا که معمولا P-value کوچکتر از 0.01 stationary محسوب می شود با توجه به مقدار این اماره و همانطور که از نمودار داده ها نیز مشخص بود داده ها stationary می باشند.

```
Detroit
ADF Statistic: -7.027062
p-value: 0.000000001
Critical Values:
        1%: -3.430
        5%: -2.862
        10%: -2.567
```

```
Toronto
ADF Statistic: -6.702604
p-value: 0.000000004
Critical Values:
        1%: -3.430
        5%: -2.862
        10%: -2.567
```

```
Los Angeles
ADF Statistic: -9.100713
p-value: 0.000000000000004
Critical Values:
        1%: -3.430
        5%: -2.862
        10%: -2.567
```

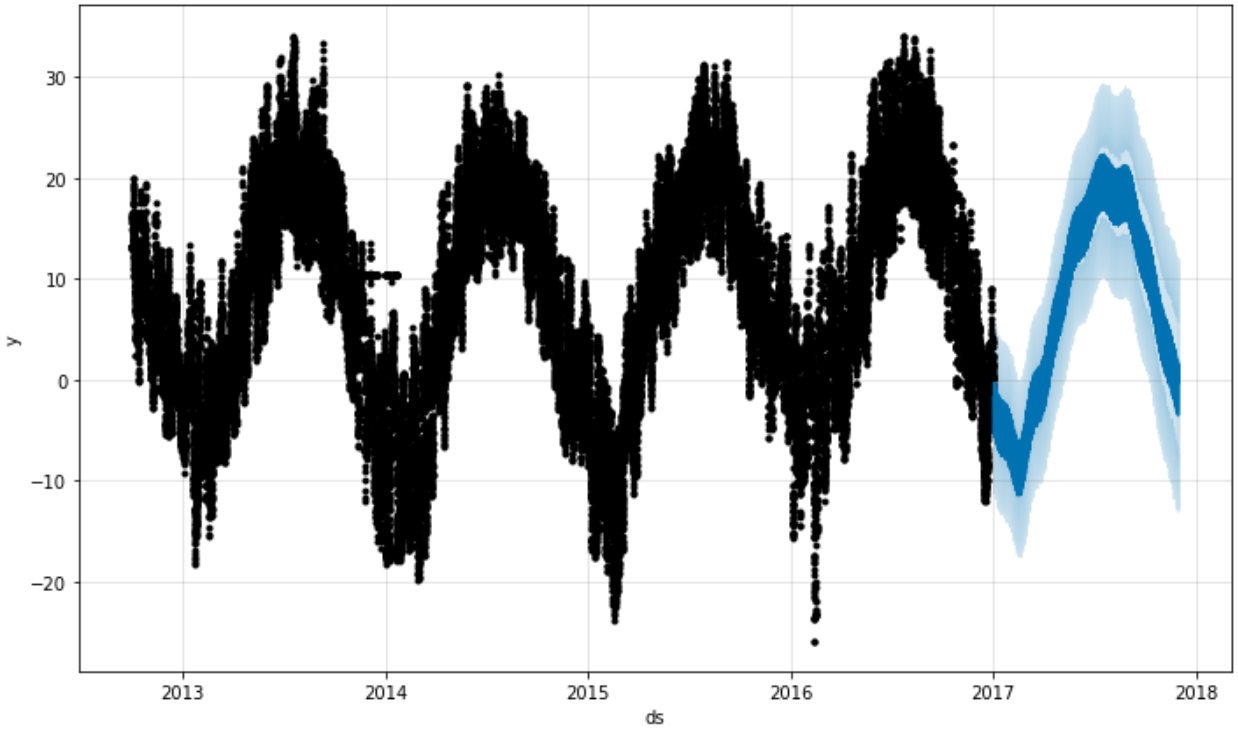
پس از بررسی موارد مربوط نوبت به دادن این داده ها به مدل می رسد.

برای هریک از مدل ها به طور جداگانه عملیات پیش پردازش و جدا کردن داده های اموزشی و تست را انجام داده ایم که همه آنها را با یکدیگر بررسی خواهیم کرد.

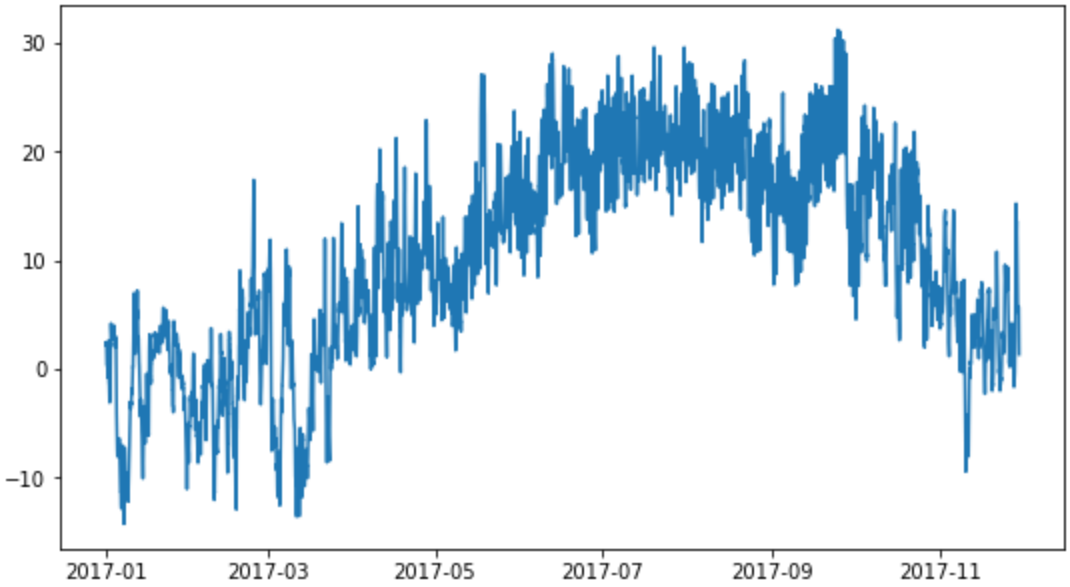
- **مدل prophet :** در این مدل datetime را تبدیل به index نمیکنیم. چرا که دو فیچری که این مدل برای فیت شدن می خواهد ستون های 'y' , 'ds' می باشند. پس از کارهای پیش پردازش گفته شده برای جدا کردن داده های train , test داده های مربوط به سال 2017 به قبل را آموزشی و داده های سال 2017 را تست در نظر می گیریم.

سپس ستون های datetime و toronto را در این دو جدا می کنیم تا به مدل بدهیم. پس از تغییر نام این ستون ها به 'y' , 'ds' مدل را روی داده های train فیت می کنیم. سپس تاریخ های مربوط به داده های تست را به صورت

جداگانه به مدل می دهیم تا برایمان میزان دما را در شهر تورنتو در سال 2017 حدس بزند. نمودار زیر مربوط به همین موضوع است. رنگ سیاه مربوط به داده هایی که از قبل فیت شده اند و رنگ ابی که برای سال 2017 است داده هایی است که به تازگی پیشبینی شده اند.



نمودار دمای اصلی این سال به شکل زیر است.



مشخصا این پیشبینی توانسته الگوی کلی تغییر دما را در این شهر با توجه به داده هایی که از سال های گذشته داشته است پیدا کند اما نمی تواند مقدار دقیقی را گزارش دهد. زیرا صرفا دارد بر اساس داده های سال های پیشین پیشبینی

می کند و داده های این سال را ندارد. من از عمد اینطور این مدل را فیت کردم تا تفاوت این مورد را با اینکه دمای روزهای قبل را بدانیم بررسی نمایم.

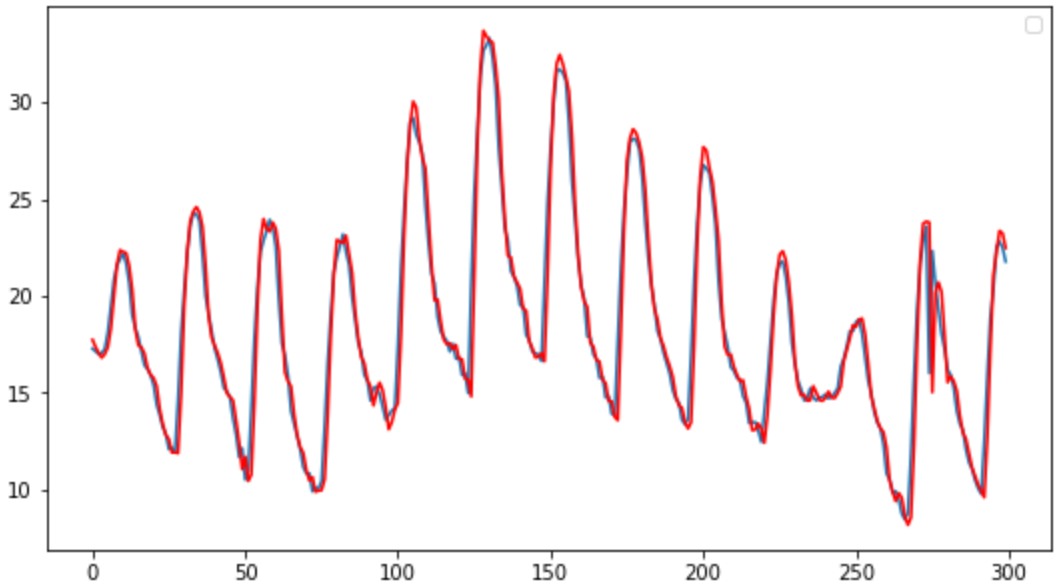
در آخر از با استفاده از root mean squared error میزان خطای مان بر روی داده های تست

Test RMSE: 5.674 بود.

- **مدل Arima :** در این مدل ما داده هایی که ایندکسشان زمان یا datetime است را به مدل می دهیم. پس از نسبت دادن داده های مربوط به سال 2017 به تست و بقیه داده ها به train تنها فیچری را که می خواهیم بررسی کنیم که این بار دمای los angeles است را در train , test ای که گفته شد نگه می داریم.

پس از انجام این کار مدل را روی داده های train فیت می کنیم. سپس مدل دمای روز بعد را پیشبینی می کند و به ما می دهد. تفاوتی که این مدل با مدل قبلی دارد این است که این بار هر یک از داده های test را که ازمایش کردیم به history یا همان مقدار دماهای قبلی اضافه می کنیم تا برای پیشبینی دمای روز بعدی مورد استفاده قرار بگیرد. از آنجایی که این کار بسیار زمان بر است نتوانستم تمام داده های مربوط به سال 2017 را تست کنم در نتیجه 300 داده اخر را برای تست در نظر گرفتم و بقیه را به train دادم.

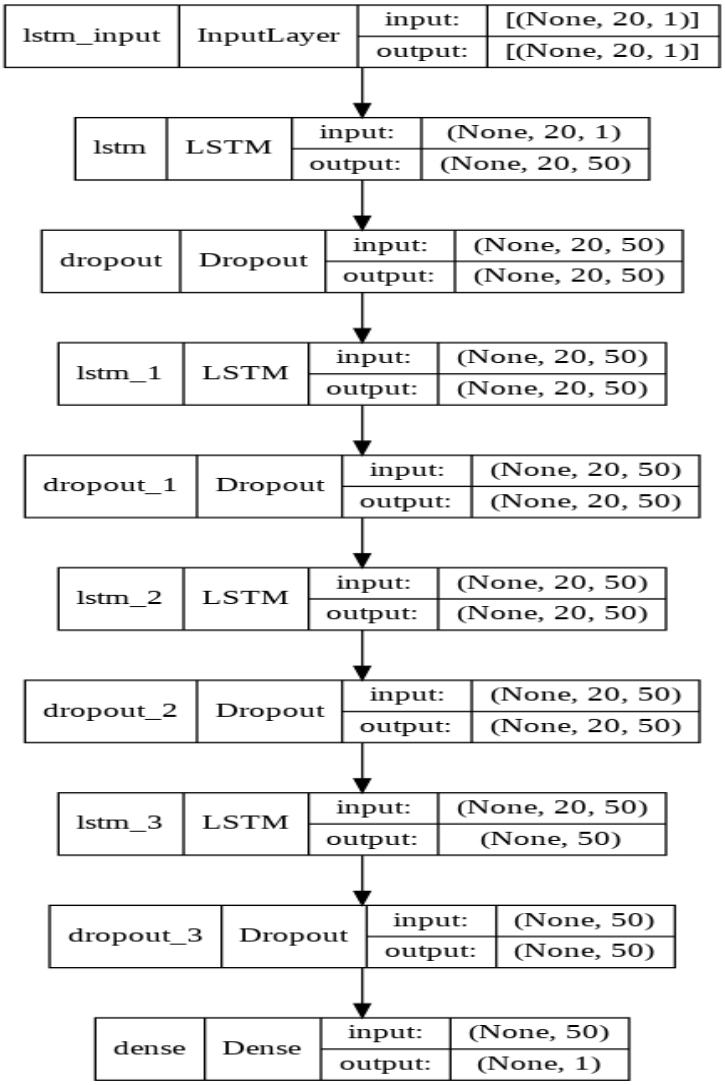
نمودار زیر نمایانگر مقدار دقیق و مقدار حدس زده شده برای دما در هر روز می باشد. رنگ آبی مقدار دقیق و رنگ قرمز مقداری است که مدل تخمین زده است که به نظر روی هم تطابق خوبی دارند.



میزان خطا که با root mean squared error اندازه گیری کرده ایم نیز برای این داده ها برابر با Test RMSE: 1.315 بود که به نسبت مدل قبلی پیشرفت قابل توجه ای به حساب می اید که البته قابل حدس بود. چون در این داده ها ما میزان دما در داده قبلی را میدانیم ولی در قبلی تنها باید از الگو پیروی می کردیم.

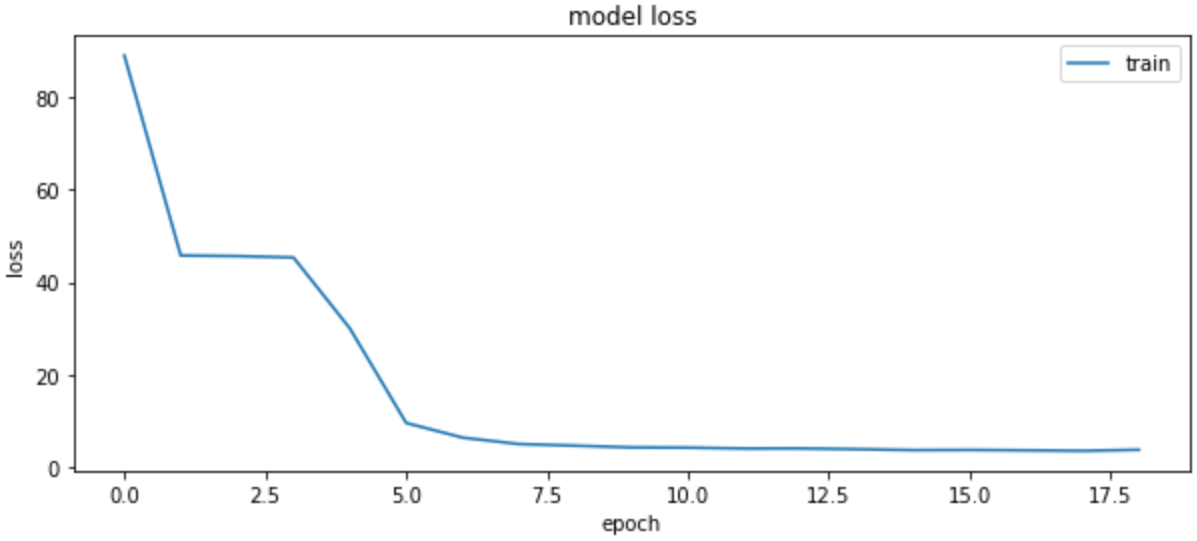
- **مدل های CNN و RNN :** در این مدل ها باید x و y را به صورت جداگانه به مدل بدهیم. لذا برای اینکه فیچری برای حدس دما داشته باشیم ، دمای 20 خانه قبل را به عنوان x و دمای داده فعلی را y یا همان target گرفتیم. چون بیش از این مقدار برای مدل های rnn خیلی قابل درک و ماندن در حافظه اش نیست. سپس از آنجا که نمیتوان با روال معمولی و رندم داده های تست و آموزشی را از یکدیگر جدا کرد با استفاده از TimeSeriesSplit داده ها را به 10000 داده آموزشی و 2000 داده تست تقسیم می کنیم تا به مدل هایمان بدهیم.

ابتدا داده هایمان را به مدل lstm که نوعی از RNN محسوب می شود می دهیم. روند کلی مدل ما به شکل زیر خواهد بود.



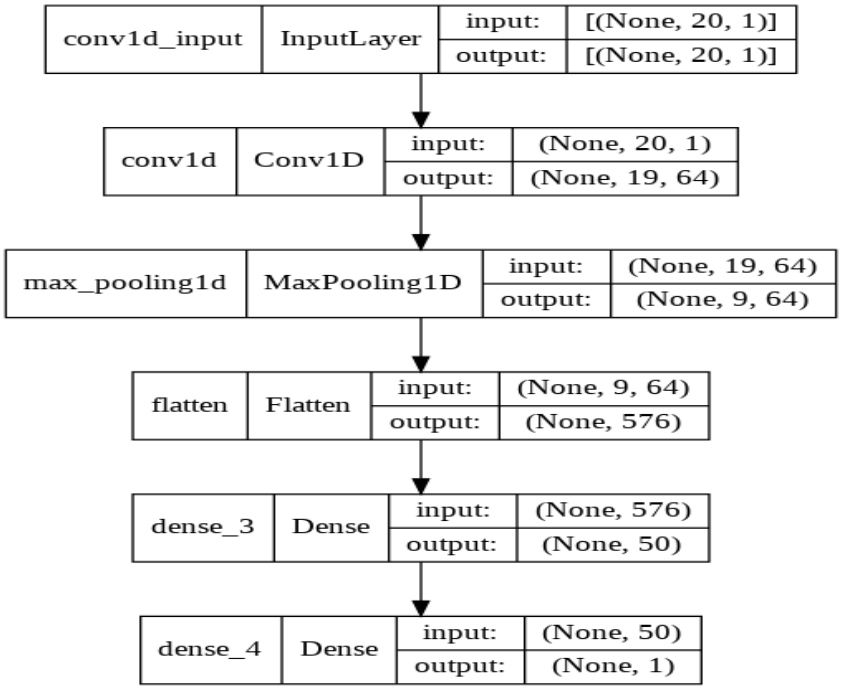
مدل مان را با اپتیمایز Adam فیت می کنیم و داده های آموزشی را به آن می دهیم. دو مورد در فیت کردن مدل استفاده شده است اولی استفاده از early stopping برای جلوگیری از overfitting و دومی استفاده از validation یا معیاری برای سنجش که این مورد هم به overfit نشدن مدل کمک خواهد کرد.

نمودار زیر نمایانگر خطای مدل بر حسب mean squared error بر حسب ایپاک ها می باشد که مشخص است روندی نزولی دارند و فرایند لرنینگ در حال انجام است.



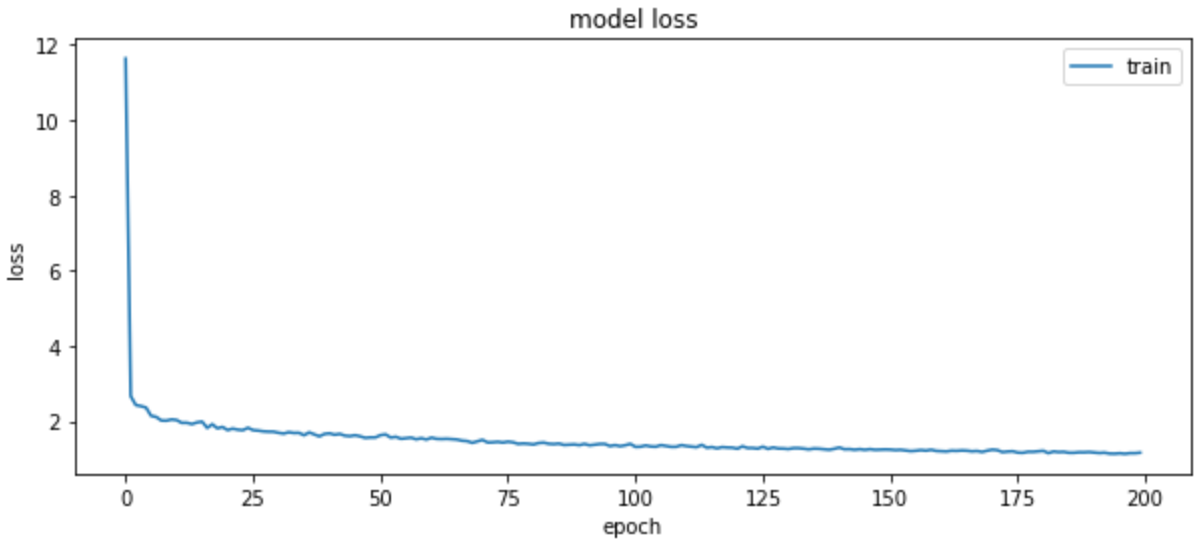
در انتها مانند مدل های قبلی خطای rmse را می سنجیم که مقدار 0.8315475925069586 را داشت که به نسبت مدل های قبلی پیشرفت داشت و می توان این مدل را کارآمد تر از مدل های قبلی نامید.

در اخر برای مدل cnn داده ها را مشابه با lstm پیش پردازش می کنیم و مدلی مطابق با شکل زیر می سازیم.



این بار هم مدل مان را با اپتیمایز Adam فیت می کنیم و داده های آموزشی را به آن می دهیم.
این بار تنها از validation استفاده شد.

نمودار زیر نمایانگر خطای مدل بر حسب mean squared error بر حسب ایپاک ها می باشد که مشخص است روندی نزولی دارند و فرایند لرنینگ در حال انجام است.



پس از آن نیز مقدار rmse = 0.9250944819515146 شد که با توجه به اینکه داده ها اسکیل نشده اند میزان قابل قبولی محسوب میشود.

همه مدل های خواسته شده به صورت univariate یا در واقع تنها با استفاده از تغییرات دما در روزهای قبلی پیشبینی شده اند و با توجه به زمان محدود از ستون های دیگر در این پیشبینی استفاده نشده است. قاعدتا با استفاده از مقادیر دیگر هم چون رطوبت و وزش باد و ... می توان دقیقتر میزان دما را پیشبینی کرد.