تعریف پروژه و گزارش فعالیت ها (تا به این نقطه)

سید سروش مرتضوی مقدم – محمد مظفری

زیر نظر : مهندس اسدی

تعریف مسئله :

به طور خیلی خلاصه ، این پروژه در نظر دارد تا قیمت سهام های بازار بورس ایران را برای روز بعد پیش بینی کند. این پیش بینی توسط سری های زمانی مربوط به ریز قیمت های هر سهام و شاخص های بازار صورت خواهد گرفت.

پیاده سازی این پروژه در قالب یک شبکه عصبی تشکیل شده از لایه های LSTM و Fully connected می باشد که به سه قسمت کلی تقسیم شده است:

۱ـ ورودی شبکه اطلاعات مربوط به : CLOSE , OPEN , HIGH , LOW , VOL یک سهام است و خروجی آن CLOSE برای روز بعد

۲ـ ورودی شبکه یک ماتریس شامل : correlation های میان همه سهام ها و همه شاخص های بازار + سری زمانی مربوط به قیمت CLOSE سهام مد نظر + سری زمانی شاخص های بازار . خروجی شبکه همچنان CLOSE سهام مد نظر برای روز بعد است.

۳ـ ورودی شبکه مورد بالا اما پیاده سازی آن با استفاده از لایه های corrLSTM می باشد.

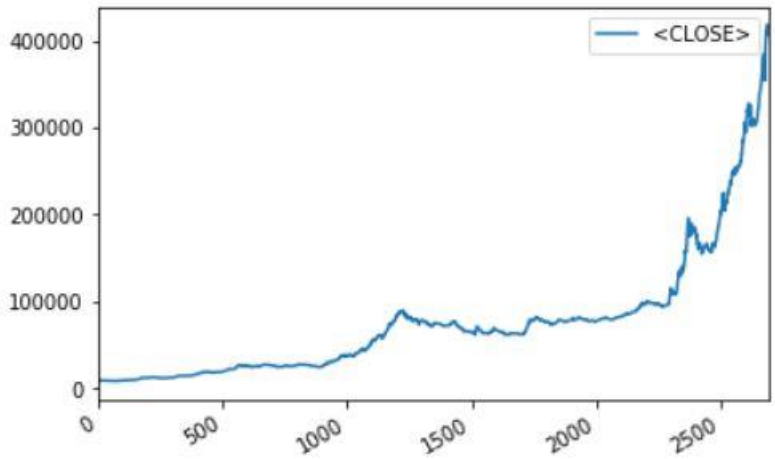
قسمت اول :

برای پیاده سازی قسمت اول پروژه شرایط و ساختار های مختلفی برای شبکه در نظر گرفته شد. شبکه به صورت لایه Dense تنها ، به صورت یک لایه LSTM و Dense خروجی ، همچنین به صورت دو لایه LSTM و Dense خروجی ، برای دو حالت کلی : ۱) ورودی: فقط CLOSE و خروجی CLOSE روز بعد و ۲) ورودی ۵ مورد ذکر شده در بالا (CLOSE,OPEN,HIGH,LOW,VOL) پیاده سازی شد.

همچنین لازم به ذکر است که داده مورد استفاده در این قسمت **شاخص کل** بوده و همچنین درین بخش بیشتر تمرکز پیش بینی های به دست آمده برای ۲۰ روز بعد بودند.

نتایج در ادامه قابل روئیت می باشد.

شاخص کل



بخش ۱-۱ :

پیش بینی شاخص کل برای یک روز بعد و استفاده از همان مدل برای پیش بینی ۲۰ روز بعد.

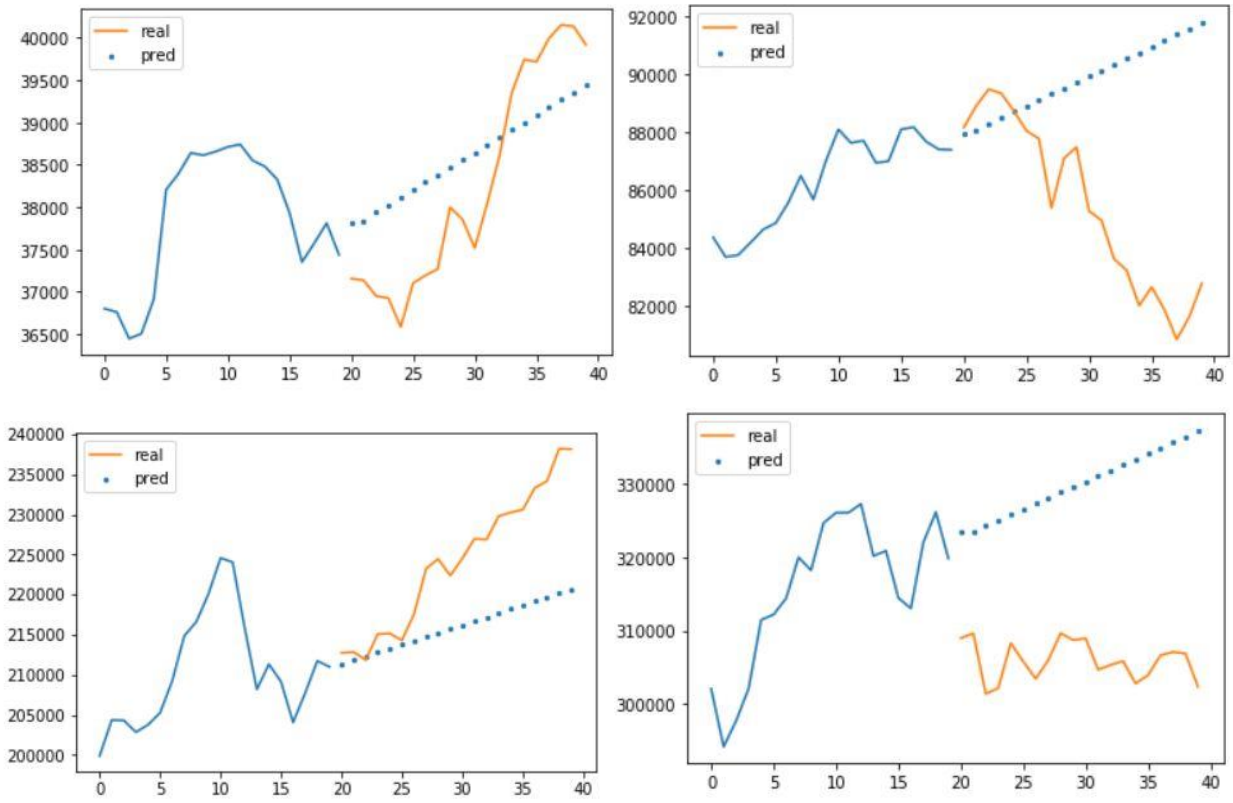
شبکه : 1 layer LSTM(20) + 1 layer Dense(1)

با استفاده از شبکه داده ها در بازه های **۲۰ تایی** به شبکه داده شده و پیش بینی در **۱ نقطه** بعد به دست آمده.

با نقطه به دست آمده اخیر + ۱۹ نقطه قبلی ، برای عدد بعدی پیش بینی کردیم

با دو نقطه به دست آمده اخیر + ۱۸ نقطه قبل برای عدد بعدی پیش بینی کردیم

و... در ادامه نمودار نتایج (تا ۲۰ روز آینده پیش بینی شده) برای ۴ قسمت از شاخص آمده است:



نظر شخصی در مورد پیش بینی : به نظر می رسد مدل این نکته را متوجه شده است که نمودار در کل به سمت بالا می رود برای همین همه پیش بینی ها به سمت بالا هستند . اما به هر حال به نظر نمی رسد که مدل ، الگو های دقیق تری را learn کرده باشد. **و این مدل برای پیش بینی ۲۰ روزی خوب نیست.**

باید خود مدل ۲۰ روز را پیش بینی کند (یعنی خروجی آن ۲۰ باشد و نه ۱)

نکات پایانی : پیاده سازی شبکه به صورت Dense تنها خروجی های بسیار ضعیفی می دهد.

و پیاده سازی شبکه به صورت LSTM ۲ خروجی بسیار مشابه بالا می دهد (۲ لایه غیر ضروری است)

بخش ۲-۱ :

پیش بینی ۲۰ روز آینده شاخص کل با استفاده از سه شبکه.

شمای کلی کار مشابه قسمت قبل است. اما سه شبکه امتحان کردیم:

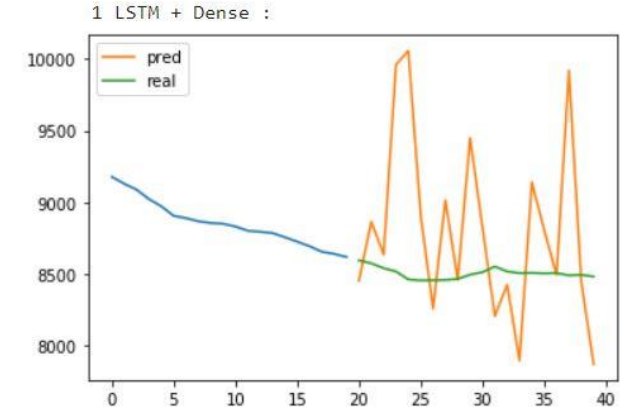
شبکه اول : Lstm (20) + Lstm (15) + Dense(20)

شبکه دوم : Lstm(20) + Dense(20)

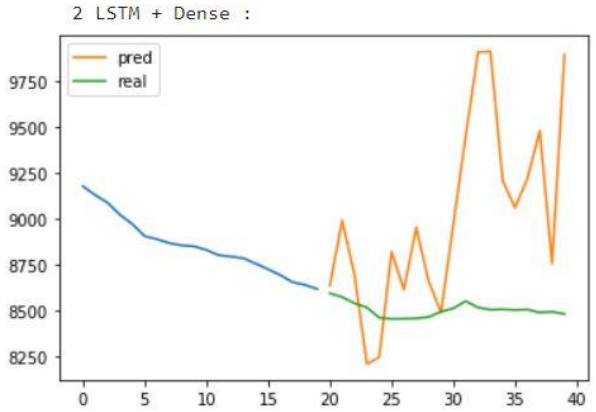
(شبکه سوم در انتهای این بخش بررسی می شود)

پس از آموزش ، برای هر دوی این شبکه ها چند نمونه از پیش بینی ها در بازه های ۲۰ روزه را مشاهده می کنیم:

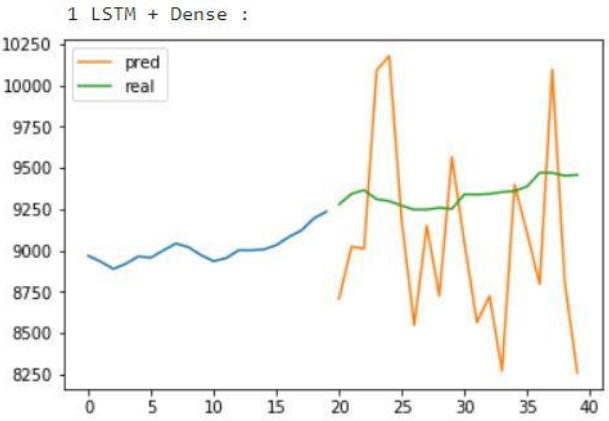
1 LSTM + Dense :



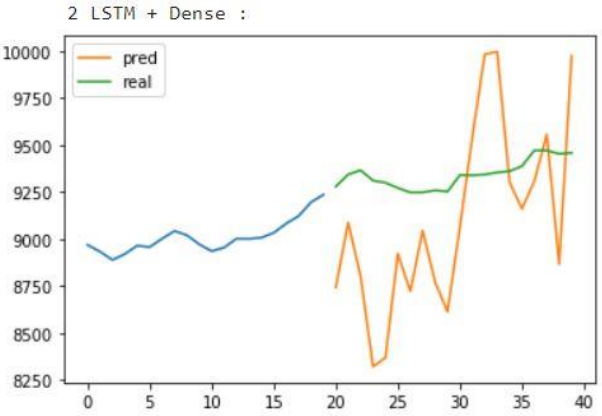
2 LSTM + Dense :



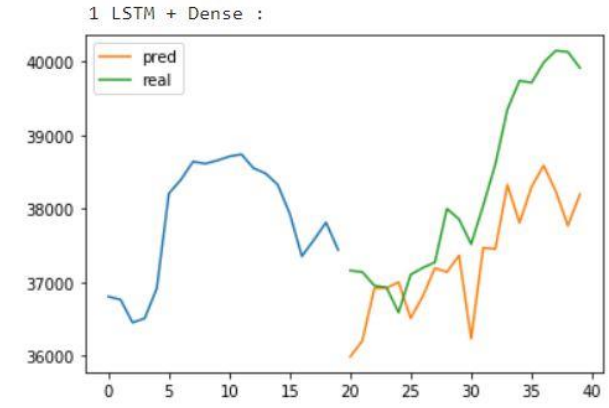
1 LSTM + Dense :



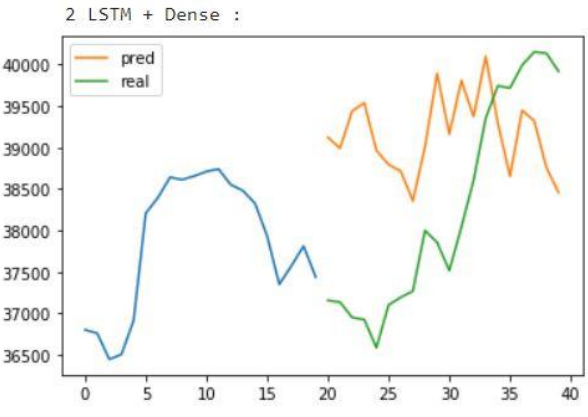
2 LSTM + Dense :

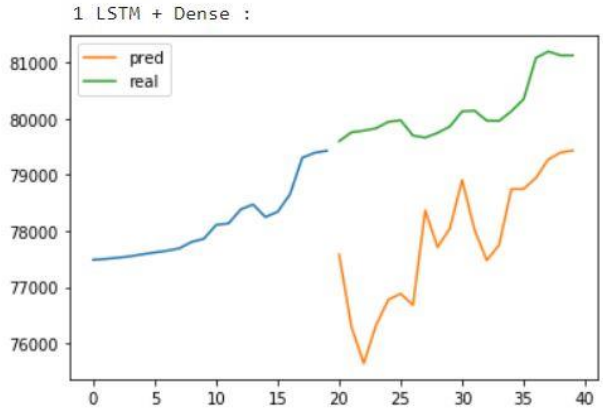


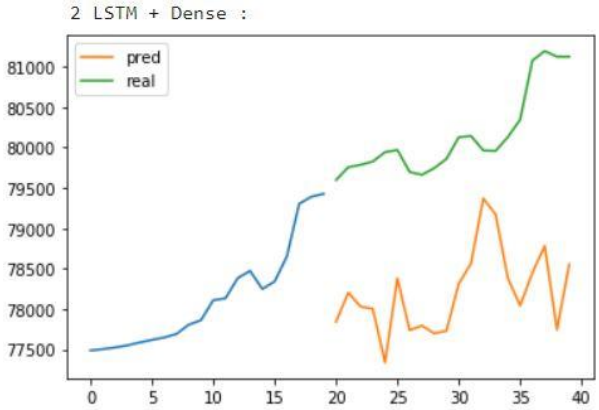
1 LSTM + Dense :



2 LSTM + Dense :

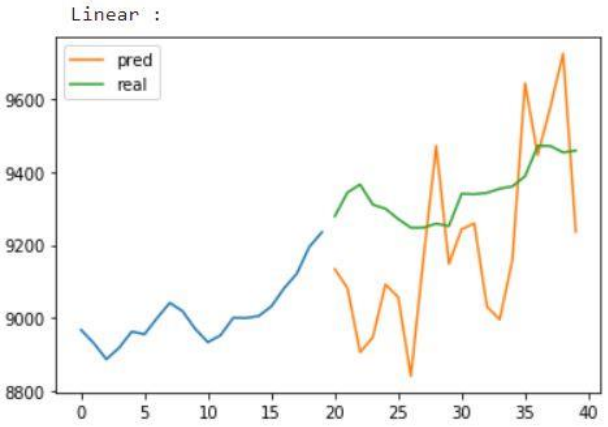


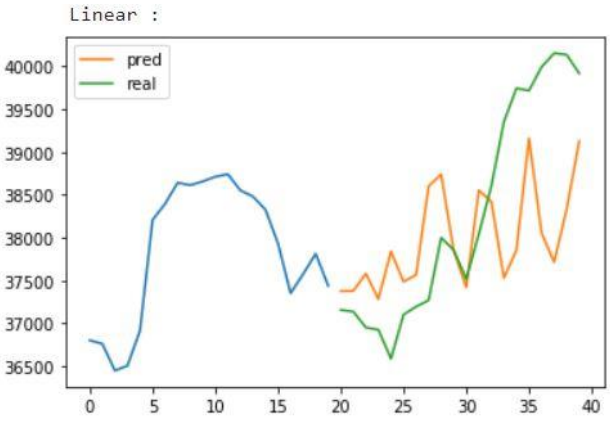


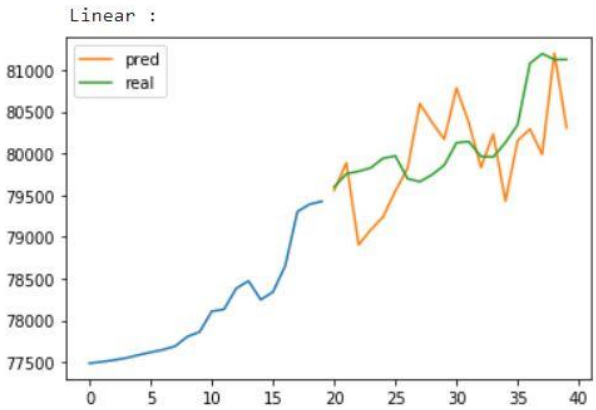


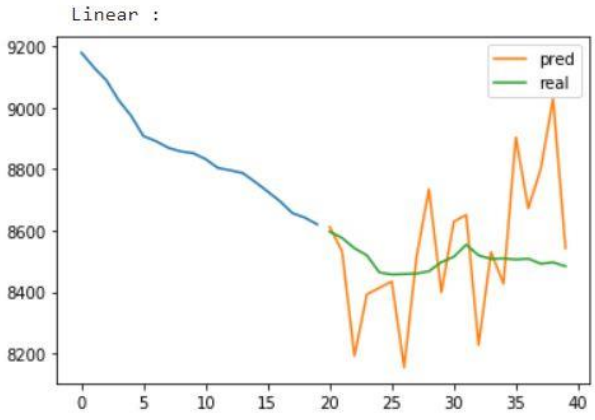
در این قسمت چون loss شبکه تک لایه کمتر است ، پس شبکه تک lstm بهتر عمل می کند. (در مقایسه با دو لایه)

شبکه سوم مورد استفاده ، یک شبکه با وزن دهی linear بود یعنی تشکیل شده از یک لایه Dense(20) تنها. در ادامه ۴ نمودار بالا را برای این شبکه هم رسم می کنیم:



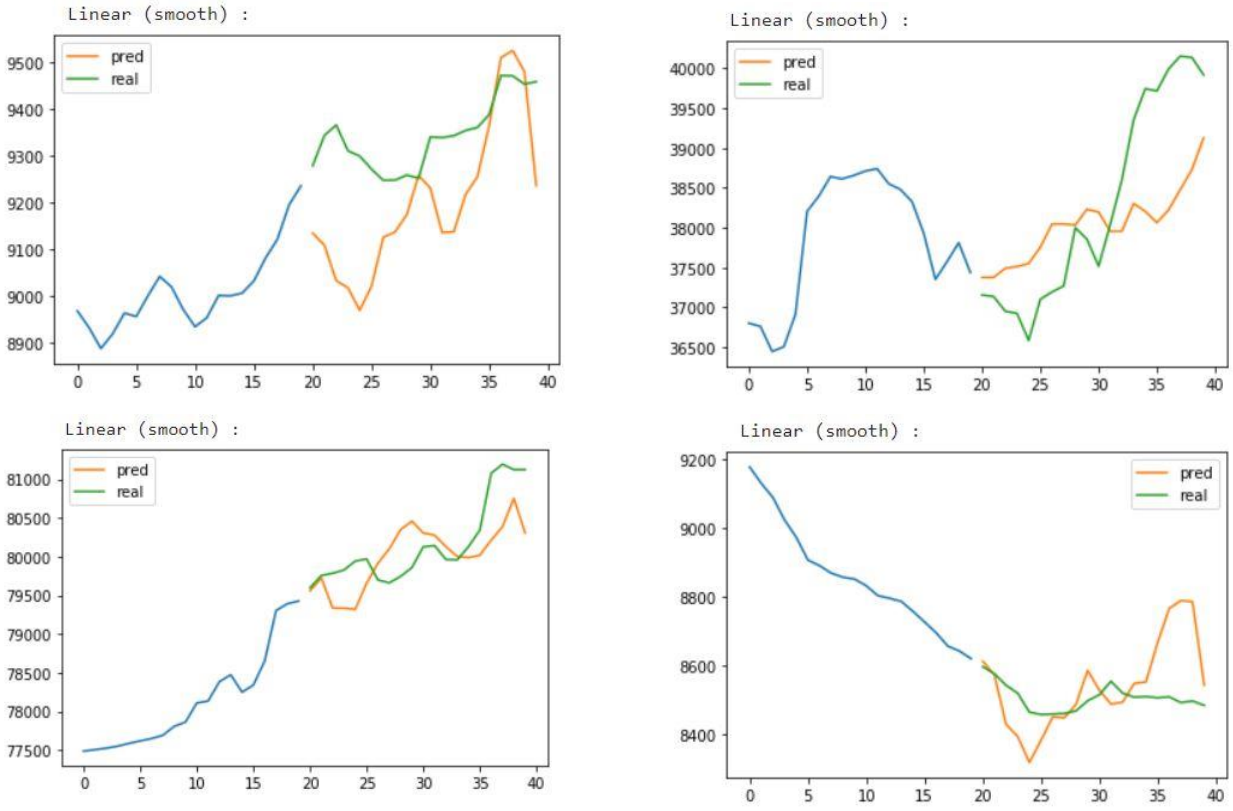






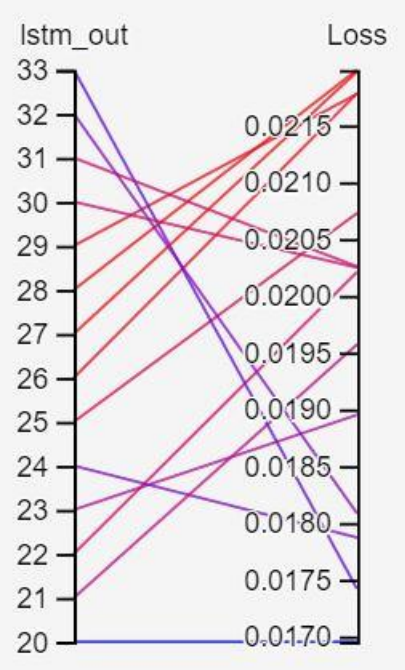
نکته قابل توجه این است که loss نهایی برای این شبکه در مقایسه با دو مورد بالا بهتر بود و پس این شبکه عملکر بهتری دارد . (هرچند در ادامه خواهیم دید که هنگامی که ورودی ها ۵ feature شوند عملکرد LSTM بهتر است)

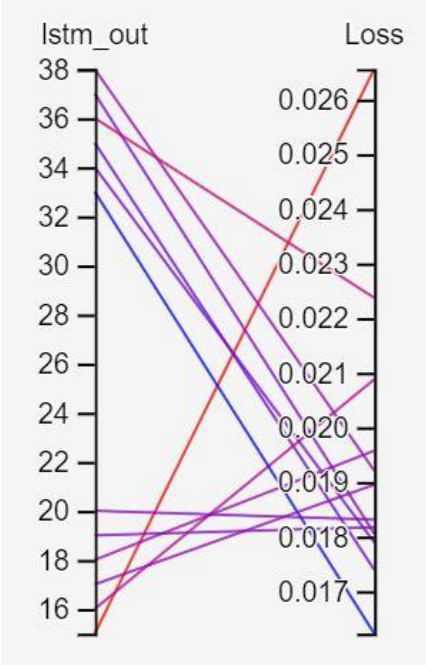
در اینجا این ایده به ذهنمان زد که اگر پیش بینی های همین مدل را کمی smooth کنیم شاید نتایج بهتری هم بگیریم:



نکته پایانی هم آنکه برای تعیین پارامتر های Lstm از hyperparameter tuning استفاده کردیم .

هرچند این بخش در قسمت های بعدی مهم تر است (چرا که نتایج بهتر در پیش بینی ۲۰ رزو بعد به دست آمد) اما برای مثال برای شبکه تک لایه نتایج در تصویر زیر قابل مشاهده است . طبق نتایج بهترین انتخاب ها یا ۳۲ و یا ۲۰ است. **ما هم با ۲۰ پیش رفتیم.**





بخش ۳-۱ :

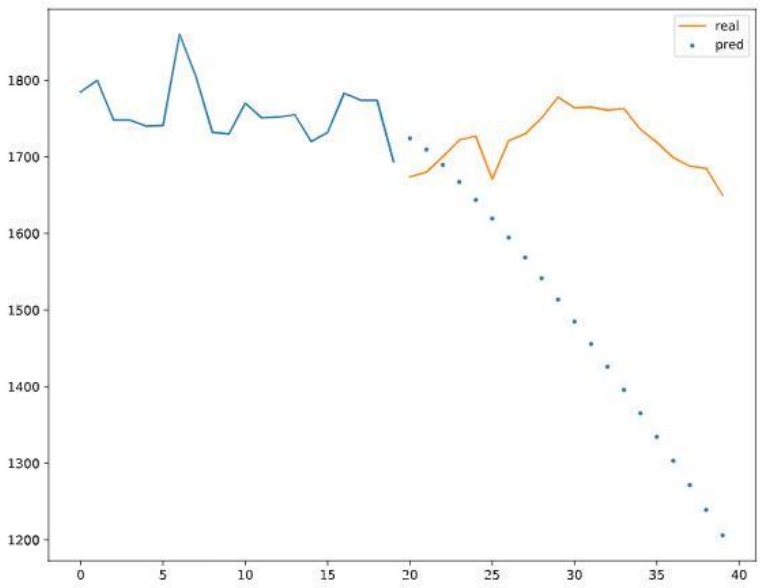
دو فرایند بالا را تکرار می کنیم اما با این تفاوت که ورودی های بجای فقط close پنج feature هستند :

```
'<OPEN>', '<CLOSE>', '<HIGH>', '<LOW>', '<VOL>'
```

ابتدا مدلی که فقط یک روز را پیش بینی می کند را پیاده سازی کردیم و سعی کردیم با استفاده از آن و پیش بینی روز به روز ، مقدار close تا ۲۰ روز بعدی را پیش بینی کنیم.

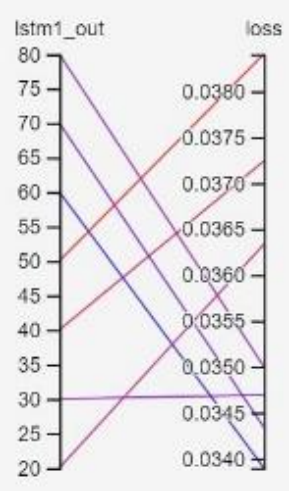
اما نتایج خیلی شبیه بخش ۱-۱ شدند

(یعنی پیش بینی ها به صورت خطی در آمده و آینده را خوب نشان نمی دادند.) برای مثال نمودار مقابل نتایج پیش بینی در یک مورد را نشان می دهد :



پس به سمت پیش بینی ۲۰ روز به صورت یکجا آمدیم.

سه مدل را استفاده می کنیم . اولی تک lstm و دومی دو لایه lstm (در انتهای هردو یک Dense تک خروجی هم است) و اخری هم dense تنها با استفاده از hyperparameter tuning مقادیر مناسب خروجی های این لایه هارا مشخص می کنیم.



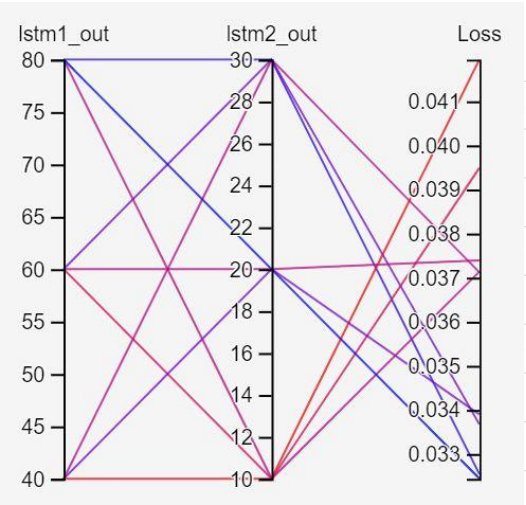
مدل تک لایه:

مطابق نتایج حاصل از این قسمت (که در مقابل آمده) برای lstm تعداد ۶۰ خروجی در نظر گرفتیم. (نتایج پیش بینی را بعدتر همزمان با مدل ۲ لایه ای می آوریم تا مقایسه راحت تر باشد.)

مدل دو لایه ای:

بنا بر نتایج حاصل (زیر) برای مدل دو لایه ای :

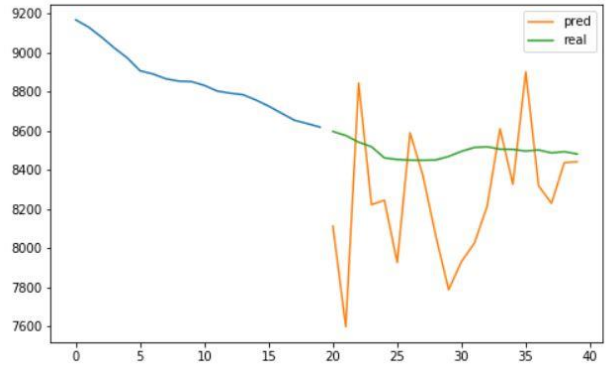
ما خروجی لایه اول را ۸۰ و لایه دوم را ۲۰ را در نظر گرفتیم .



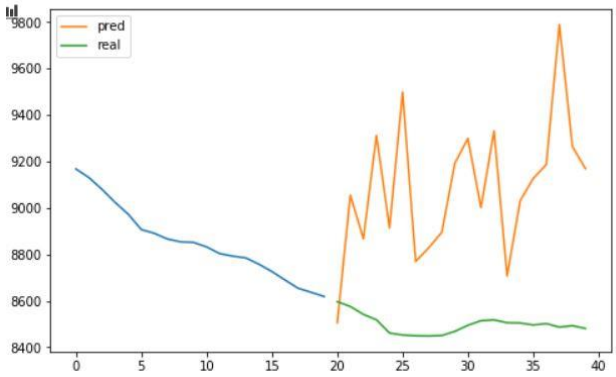
lstm1_out	lstm2_out	Loss
40.000	30.000	0.037118
80.000	30.000	0.032542
60.000	30.000	0.033670
40.000	10.000	0.041946
60.000	10.000	0.039490
40.000	20.000	0.033895
80.000	10.000	0.037135
60.000	20.000	0.037393
80.000	20.000	0.032437

در ادامه نتایج هر دو مدل را به روی همان داده هایی که در قسمت اول و دوم پیش بینی کرده بودیم را مشاهده می کنیم :

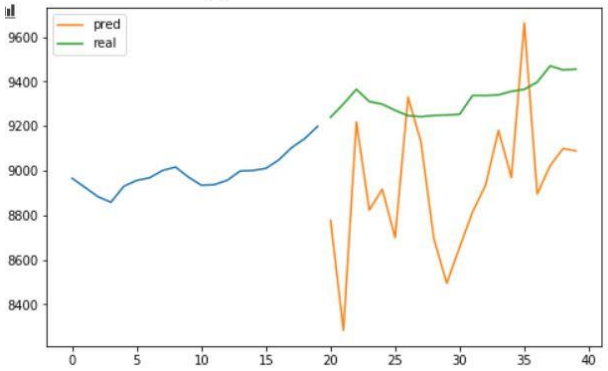
2 Lstms + Dense (1)



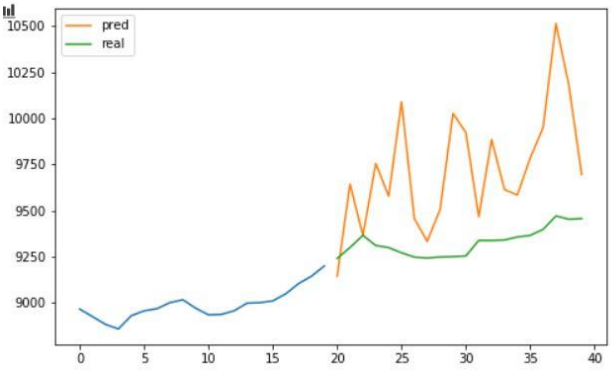
1 Lstm + Dense (1)



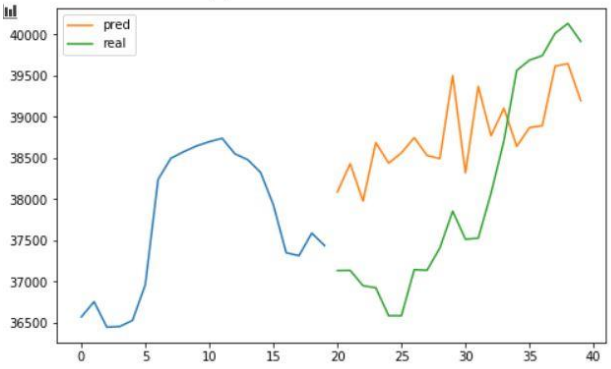
2 Lstms + Dense (1)



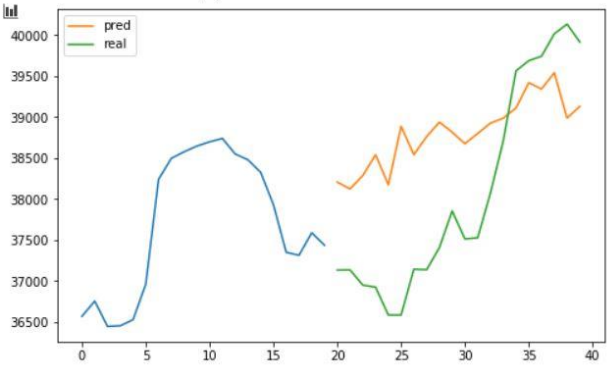
1 Lstm + Dense (1)



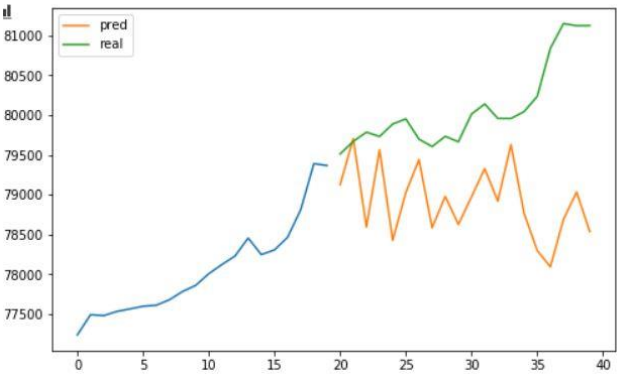
2 Lstms + Dense (1)



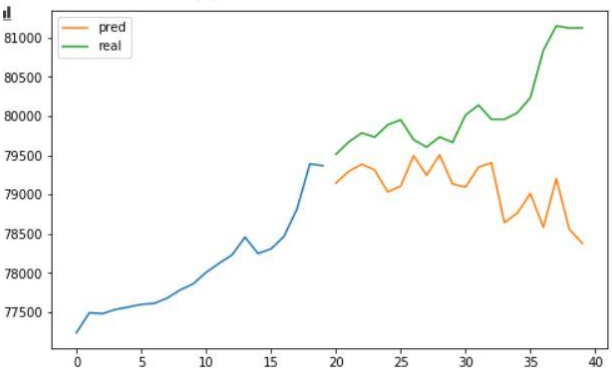
1 Lstm + Dense (1)



2 Lstms + Dense (1)



1 Lstm + Dense (1)



مقدار loss برای مدل **دو لایه** کمتر است. (پس در واقع مدل دو لایه بهتر عمل کرده است)

همچنین ، چون در قسمت های قبل دیدیم که یک مدل با dense تنها هم می تواند عملکرد خیلی خوبی داشته باشد آن را هم امتحان کردیم.

اما نتایج حاکی از آن بود که با دادن ۵ feature به عنوان ورودی ، مدل dense تنها ، عملکرد مناسبی ندارد (loss=0.04) که در مقابل دو مدل بالا که تک لایه برابر 0,038 و مدل دو لایه 0,032 بود بیشتر است.

حال این سوال پیش می آید که مدل دو لایه بالا که ۵ feature می گیرید بهتر است یا مدل Dense تنهای قسمت دوم که فقط روی CLOSE عمل می کرد. برای همین خاطر loss این دو مدل را با هم مقایسه می کنیم.

مدل دو لایه : `0.031251611916670 84`

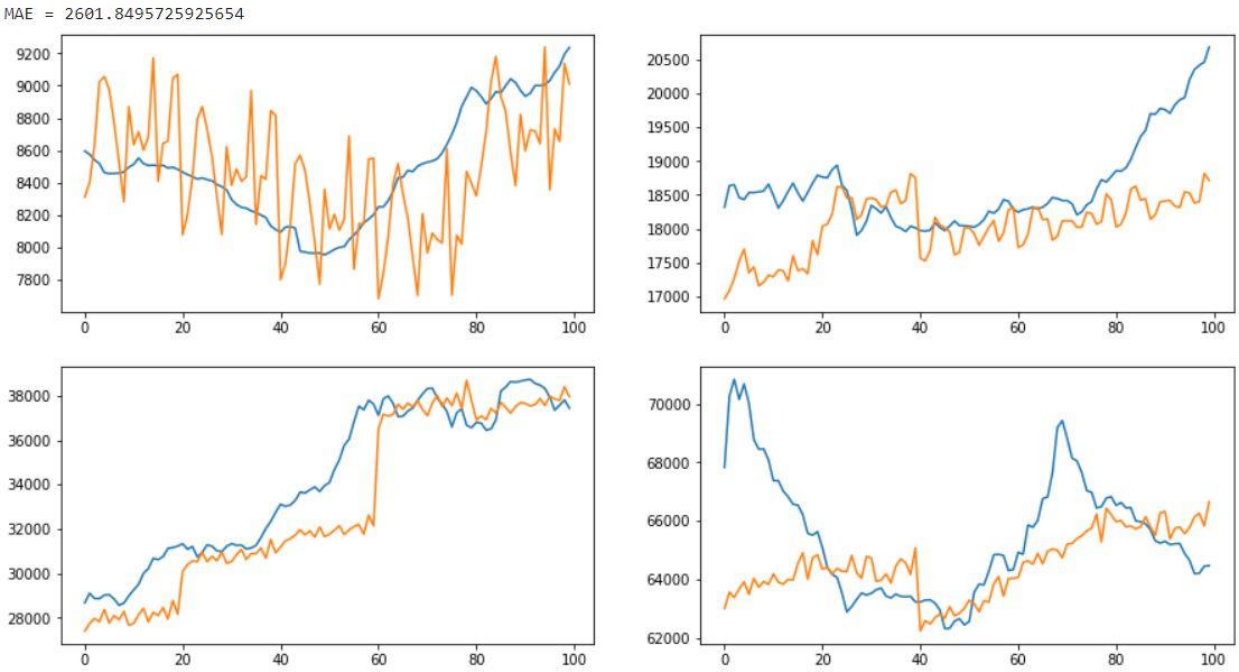
مدل dense تنها : `0.037747752369686 78`

پس مشخص شد که مدل دو لایه از Dense اولیه هم بهتر است.

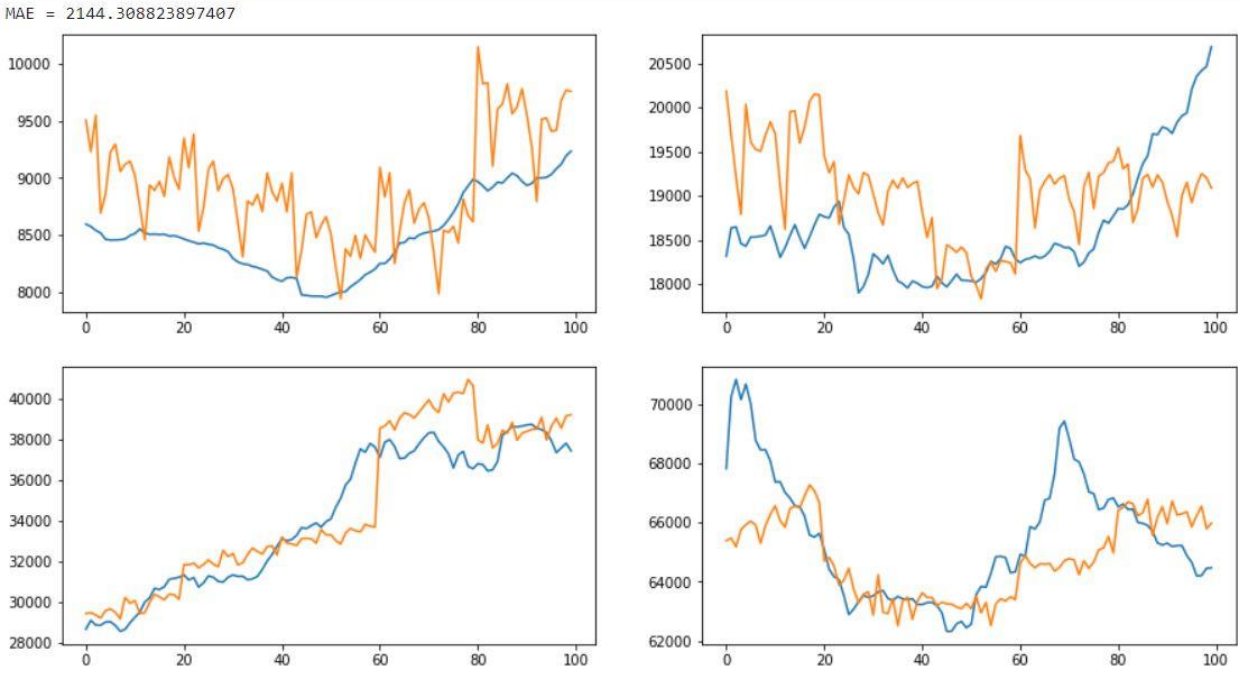
برای جمع بندی عمل کرد همه مدل ها را برای پیش بینی ۲۰ روز در کنار هم می آوریم . برای هر مدل هم پیش بینی با ۵ feature و هم پیش بینی با ورودی CLOSE را نمایش می دهیم. برای اینکه دید بهتری داشته باشیم **بازه های بزرگ تری** را رسم خواهیم کرد و برای هر مدل ، میزان mean absolute error را هم به ازای مقادیر واقعی پیش بینی محاسبه می کنیم:

نکته : خطوط نارنجی پیش بینی هستند ـ خطوط آبی مقادیر واقعی

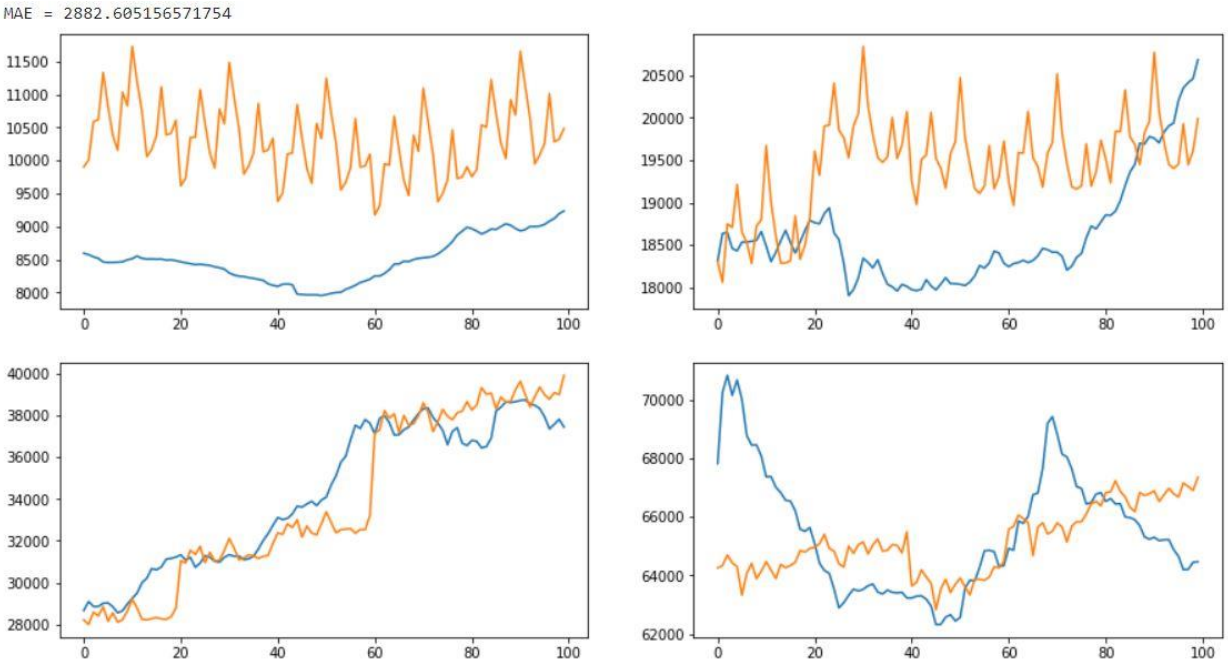
مدل تک لایه Lstm (60) با خروجی Dense (20) : ورودی فقط CLOSE



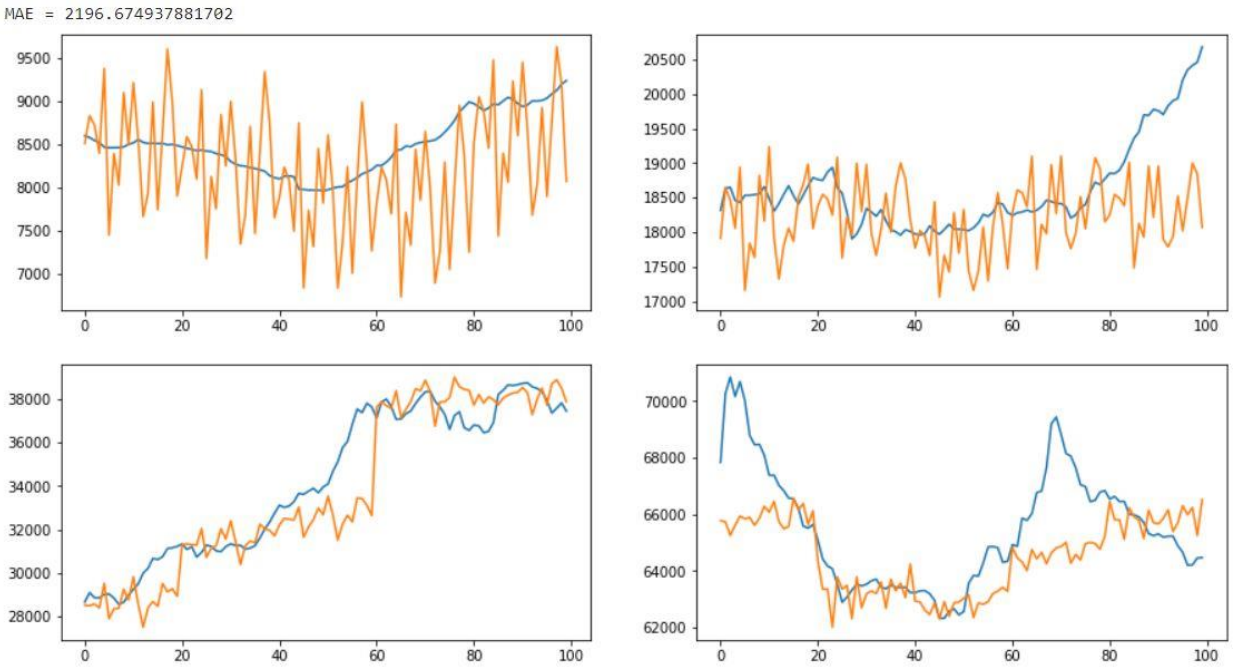
ورودی ۵ feature :



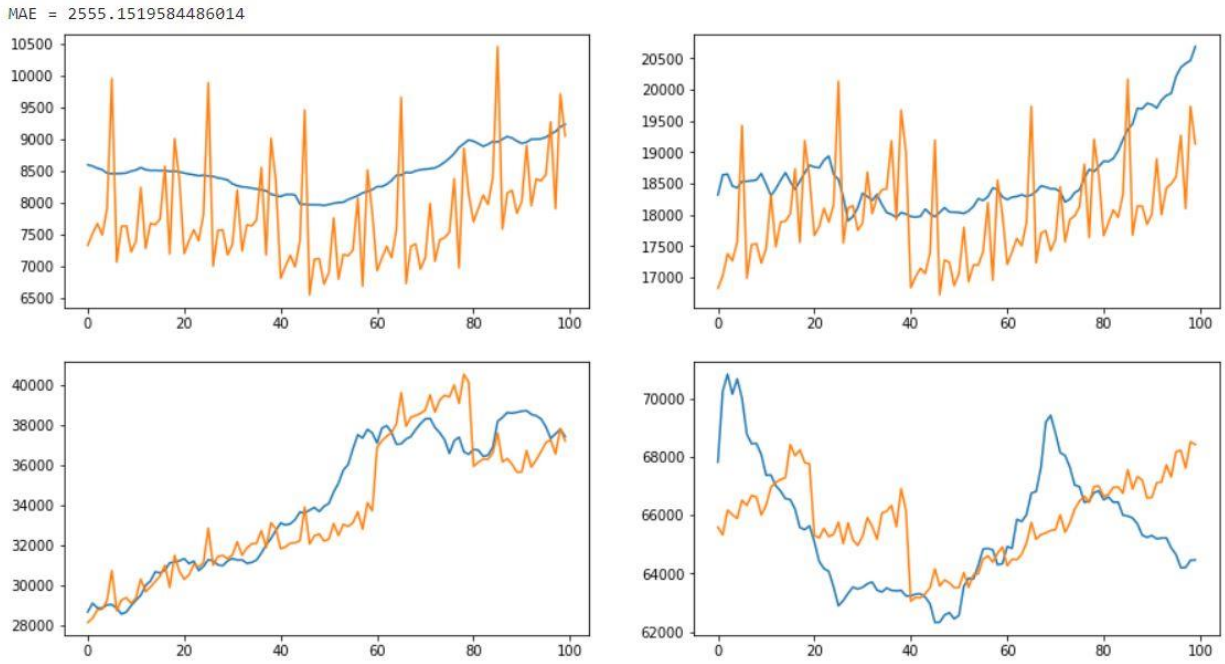
مدل دو لایه Lstm (80) + Lstm (20) و خروجی Dense (20) : با ورودی CLOSE



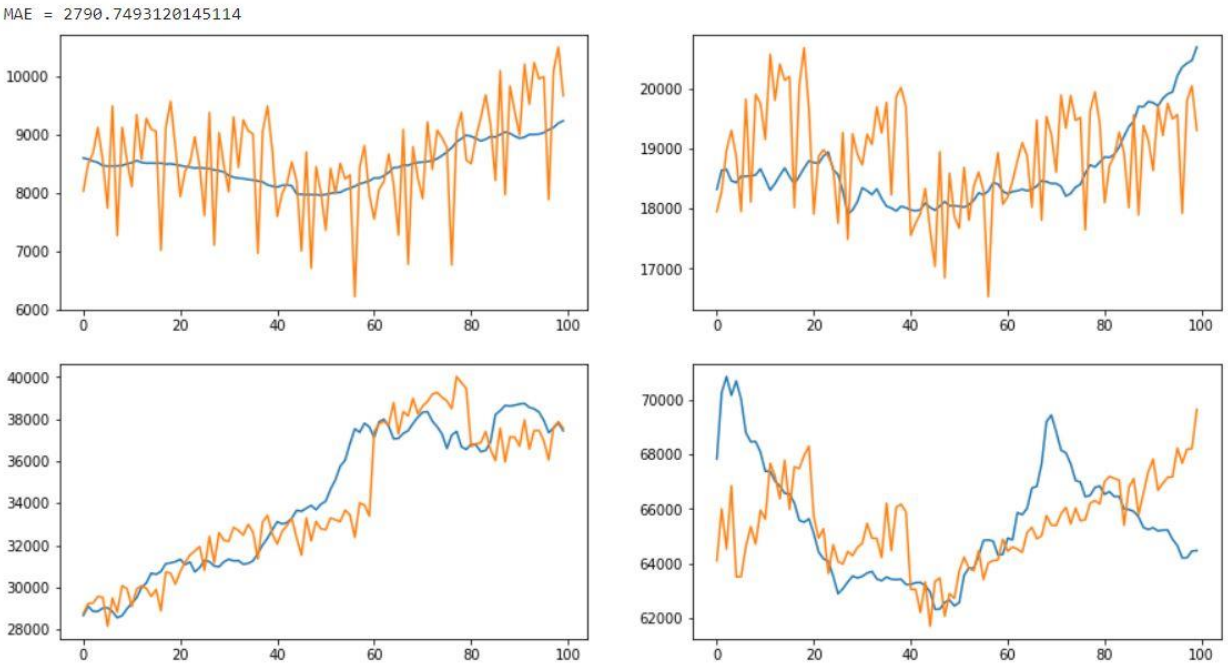
ورودی ۵ feature :



مدل Linearیعنی فقط Dense (20) : ورودی CLOSE



ورودی ۵ feature :

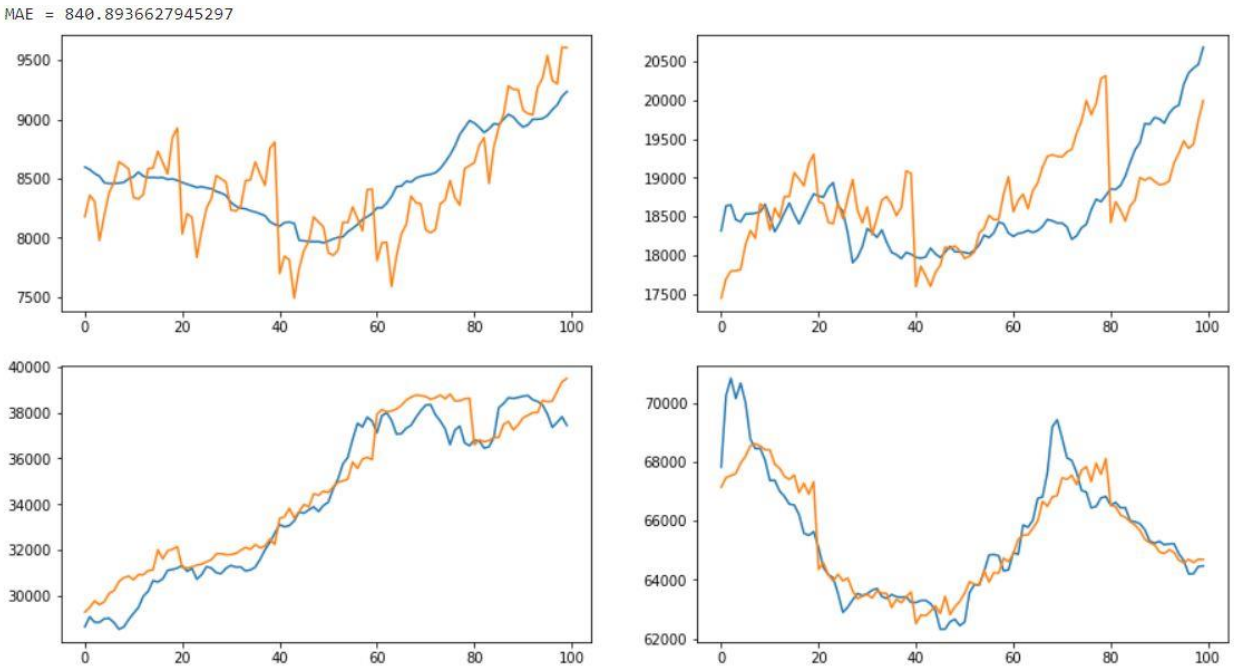


با توجه به برسی ها ، بهترین مدل ها به ترتیب :

2Lstm_5features و سپس **1Lstm_5features** بودند.

به همین منظور مدل **دو لایه** را این بار به مدت طولانی تری train کردیم (دفعات پیش 25 epoch بود صرفا برای به دست آوردن نسبت عملکر مدل ها به هم اما این بار 180 epoch) و توانستیم به loss= 0.012904731089531003 برسیم.

و این کار باعث شد mas در نمودار اعداد واقعی از ۲۱۹۶ به ۸۴۰ تقلیل پیدا کند. هرچند شبکه با این تعداد بالای آموزش over fit می شود. نمودار چند نمونه قبل پیش بینی شده با مدل اخیر :



ذکر دو نکته لازم است. یک مشکلی که در داده های ما وجود دارد این است که قیمت شاخص کل در ابتدا در حدود 8,000 بوده و رشد آن خیلی ارام است ولی در اواخر به حدود 400,000 می رسد و رشد نمایی پیدا می کند. این تغییر مقیاس داده ها و رشد آنها باعث میشود پیش بینی ها خیلی دقیق نباشند برای همین بعدا شاید بهتر باشد که برای پیشبینی از داده های به روزتر استفاده شود تا اینکه همه داده موجود را به شبکه بدهیم. همچنین ما تا اینجای کار روی داده های شاخص کل کار کردیم با این امید که مدل به دست آمده بروی نماد های دیگر هم خوب عمل کند.

قسمت دوم :

در قسمت بعد ما correlation میان سهام ها و شاخص هارا محسابه کرده و به عنوان ورودی به شبکه می دهیم با این امید که نتایج بهتری بگیریم. ورودی شبکه یک ماتریس شامل :

correlation های میان همه سهام ها و همه شاخص های بازار + سری زمانی مربوط به قیمت CLOSE سهام مد نظر + سری زمانی شاخص های بازار . خروجی شبکه همچنان CLOSE سهام مد نظر برای روز بعد است.