به نام خدا

گزارش عملکرد مسابقه ی داده کاوی فناورد

اعضا :

مریم ابراهیم زاده

منیره صفری

سپیده ملانوروزی

یاسمن میرمحمد

فاطمه سادات هاشمی گلپایگانی

خرداد ۱۳۹۷

فهرست مطالب صفحه

مقدمه :

پیش بینی قیمت طلا یکی از مسائل پیچیده می باشد و عامل های مختلفی روی ان تاثیر می گذارد. عواملی چون تورم، سیاست خارجی ، میزان تقاضا، درآمد حاصل از فروش نفت و ....

بعد از مطالعات انجام شده به این مسئله به یک مسئله سری زمانی مدل شد و مدل هایی روی آن امتحان شد. با تصویر سازی داده های مربوط به طلا به نظر می امد که قیمت آن در هر مقطع زمانی تابعی از زمان نبوده و به صورت شهودی تابعی از مقدار قیمت در مقطع زمانی قبلی است. ازاین جهت قیمت در هر لحظه را تابعی از قیمت در زمان قبلی در نظر گرفته شد. با این روش این مسئله به یک مسئله supervised **تبدیل شد. بخشی از داده ها به عنوان داده تست جدا شدند و مدل بر اساس بقیه ی داده ها اموزش داده شد. خطای هر مدل با مقایسه مقدار تخمین زده شده توسط مدل آموزش داده شده برای داده های تست و مقدار واقعی آن ها اندازه گیری شد.**

خلاصه ی مختصری از روند کار :

ابتدا فایل صورت سوال خوانده شد و پس از بررسی کلی اطلاعات داده شده و صورت مسئله اعضا به مطالعه ی مدل های مناسب برای این نوع مسائل پرداختند.
مدل های مطالعه شده شامل مدل شبکه عصبی lstm، مدل svm، مدل anfis، مدل normal و مدل خطی بود که تصمیم بر این شد ابتدا از مدل های ساده شروع کرده و در هر مرحله مدل قبلی بهبود داده شود و یا از مدل پیچیده تری استفاده شود .
برای زبان برنامه نویسی نیز با توجه به مهارت اعضا و کتابخانه ها امکانات مورد نیاز زبان پایتون انتخاب شده است .

مراحل بعدی کار را می توان به بخش های زیر تقسیم کرد:
۱ . بررسی فایل داده ها
۲. تبدیل داده ها به داده های نظارت شده (supervised)
۳. به دست آوردن مدل اولیه baseline و محاسبه ی خطای آن
۴ . اجرای مدل svm و محاسبه ی خطا و تصویرسازی
۶. اجرای مدل خطی بر روی کل داده ها
۷. اجرای مدل خطی بر روی داده های بازه بندی شده
۸. مقایسه مدل های مختلف روی موضاعات مختلف و انتخاب بهترین مدل برای هر ارز یا فلز گرانبها

بخش اول :

پردازش داده ها :

ابتدا داده های فایل ticker بررسی شده اند .

تمامی ویژگی های داده شده در این فایل (ستون ها) از a0 تا a8 نام گذاری شده اند.

a0: زمان ثبت اطلاعات که با رزولوشن ۵ دقیقه جمع آوری شده است

a1: قیمت حداکثر سفارش ثبت شده برای فروش

a2: حجم حداکثر سفارش ثبت شده برای فروش.

a3: قیمت حداکثر سفارش ثبت شده برای خرید

a4: حجم حداکثر سفارش ثبت شده برای خرید.

a5: قیمت آخرین معامله صورت گرفته در بازار

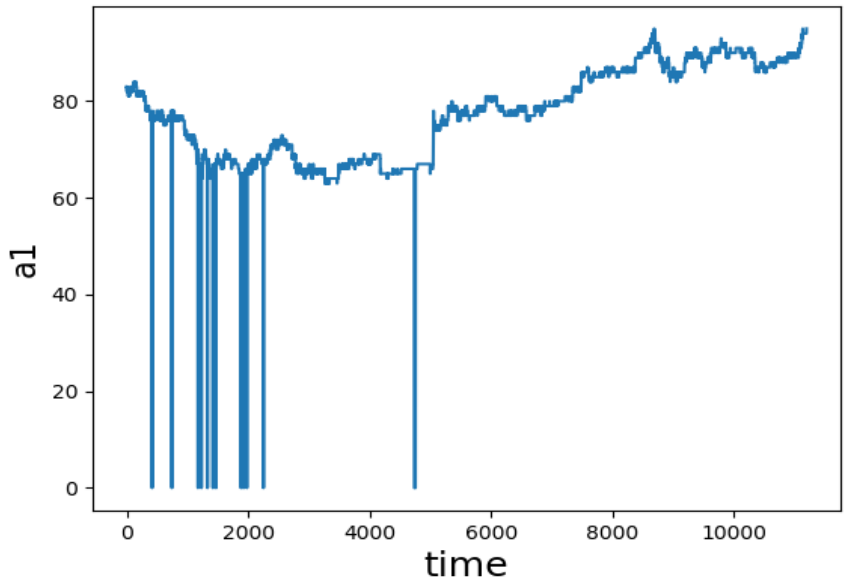
a6: حجم معاملات روزانه.

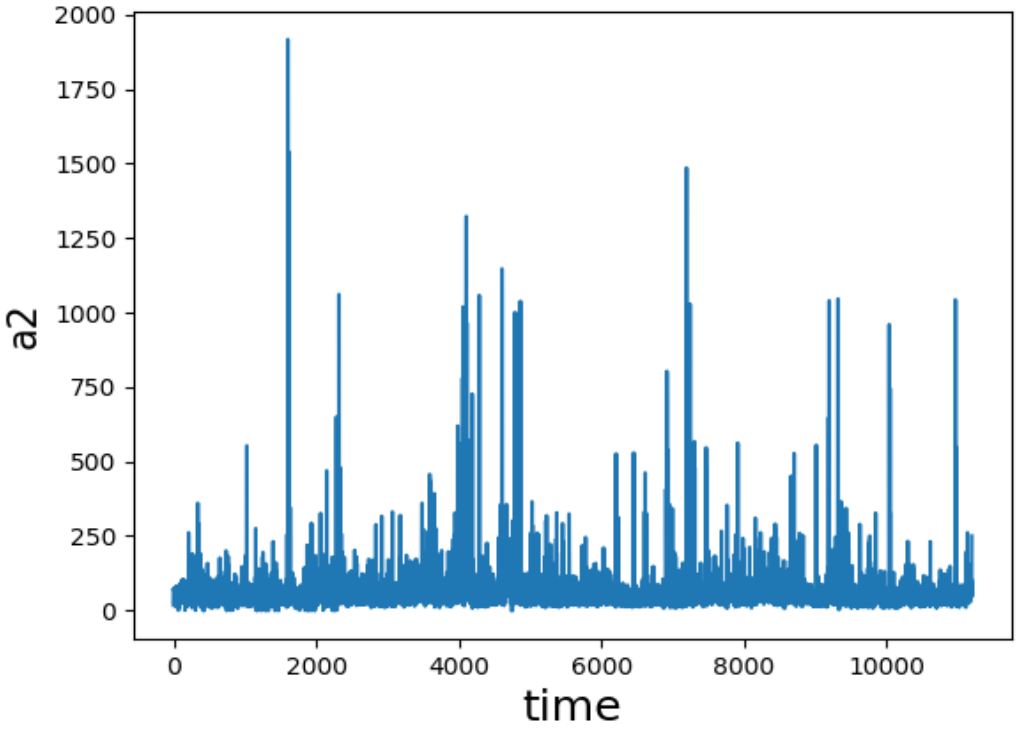
a7: حداکثر قیمت روزانه.

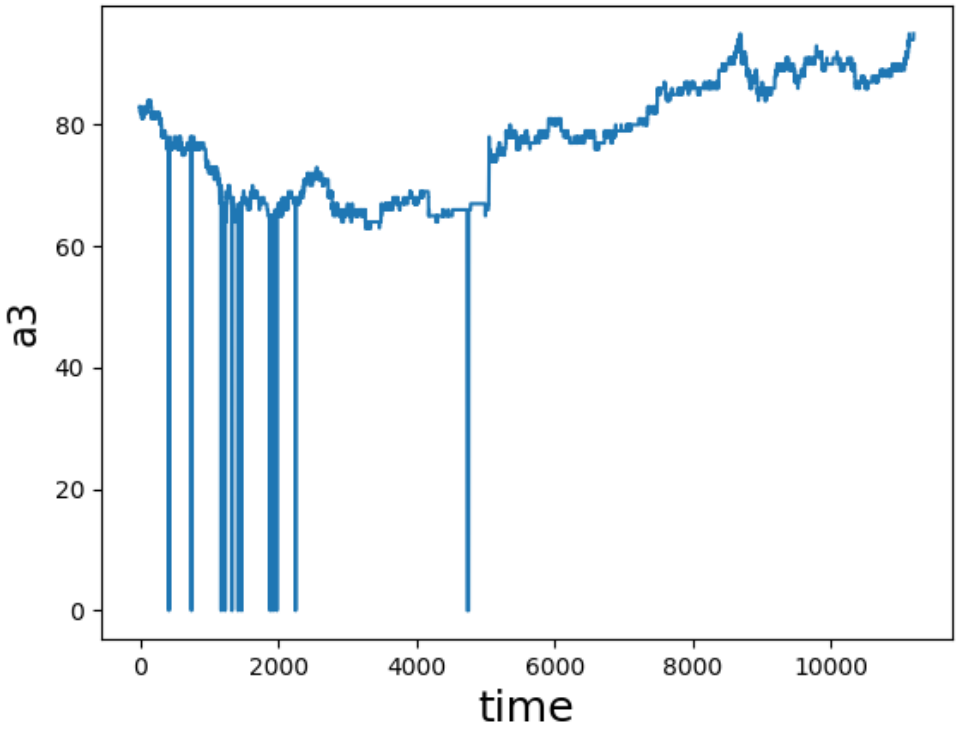
a8: حداقل قیمت روزانه

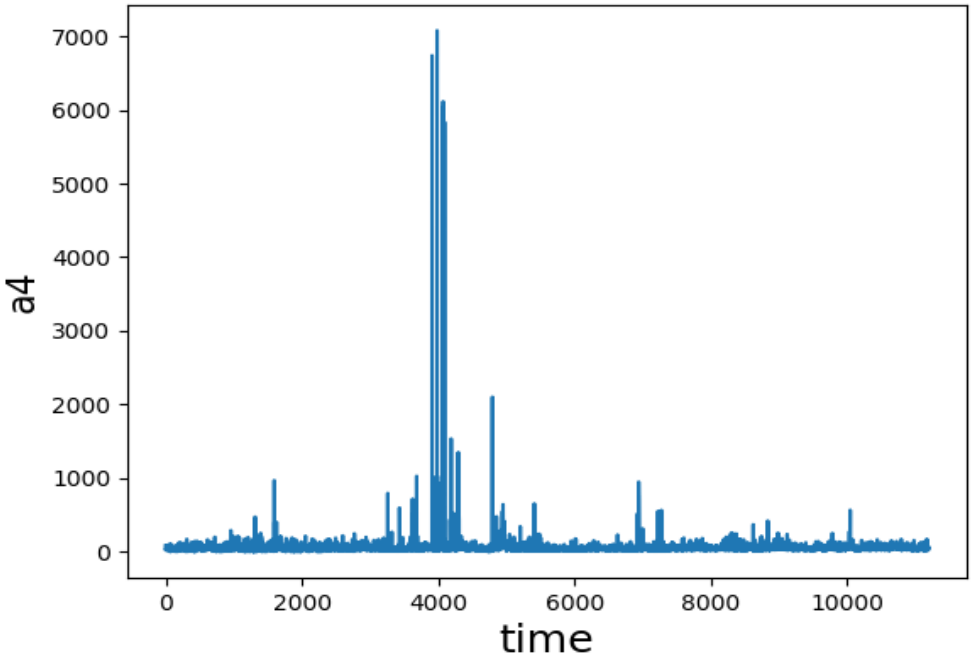
همه داده ها بر اساس زمان(a0) کشیده شده اند .

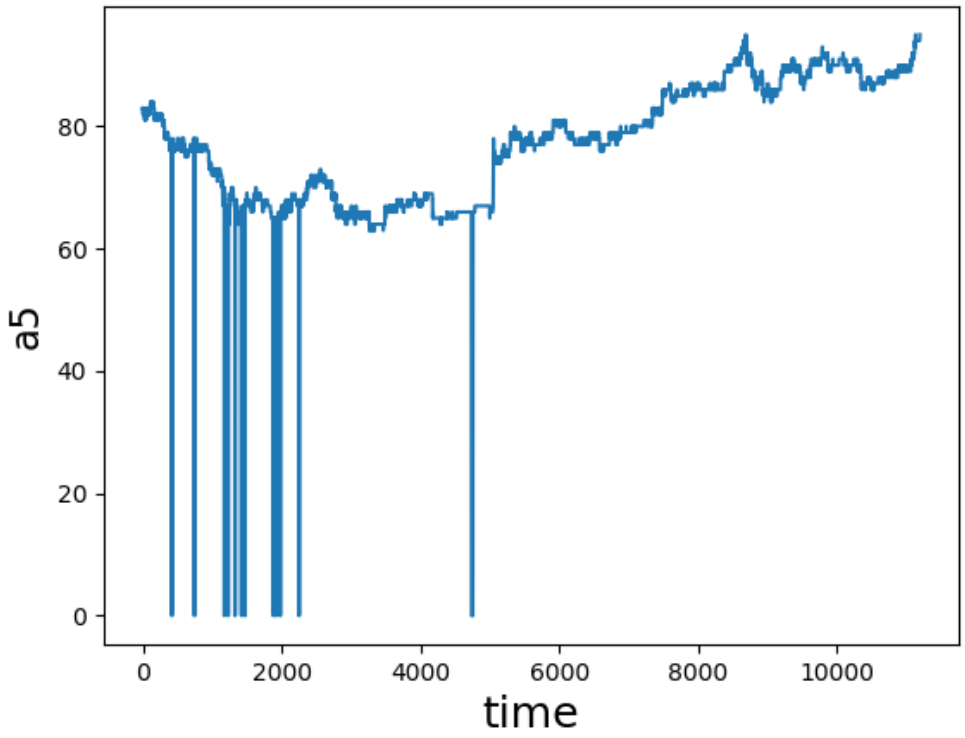
(منظور از b1 هم همین ویژگی a1 است ولی برای جنس b و به همین ترتیب)

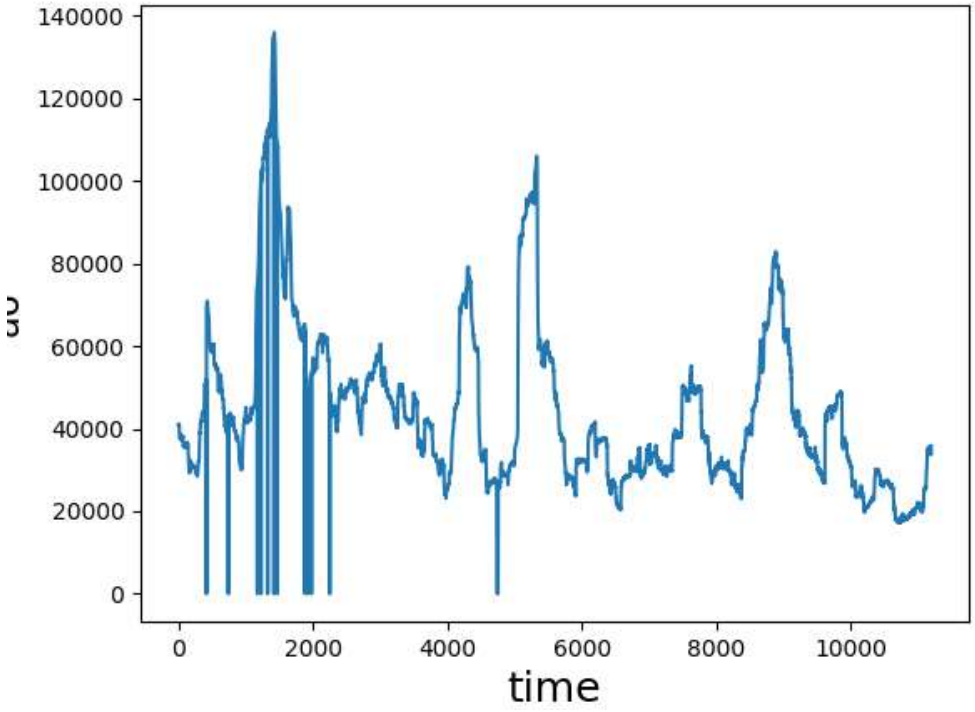


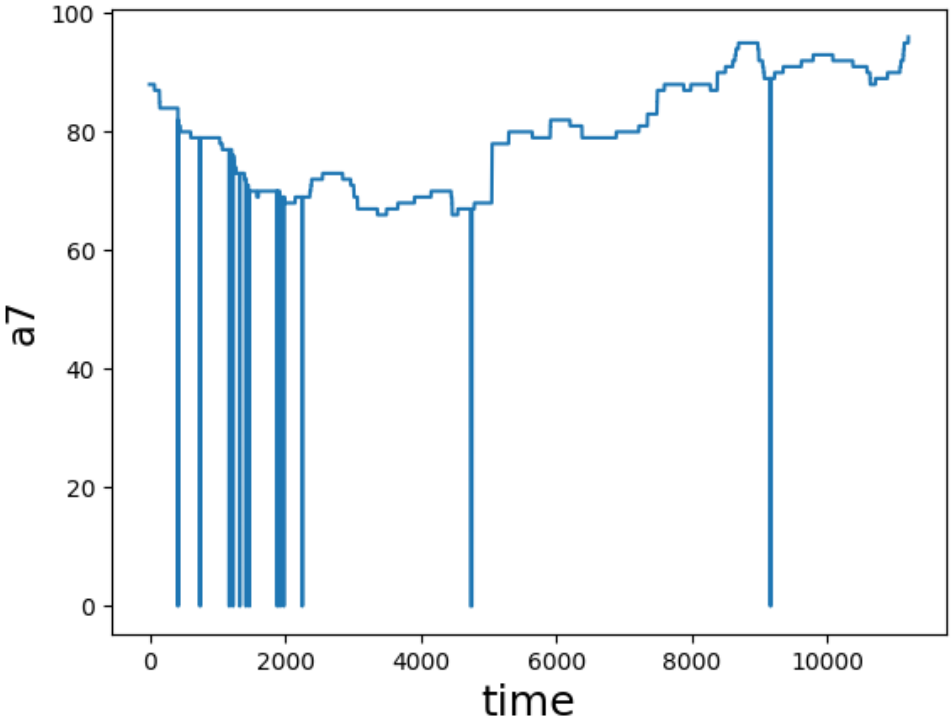

2000
1750
1500
1250
1000
750
500
250
0
a2
0
2000
4000
6000
8000
10000
time

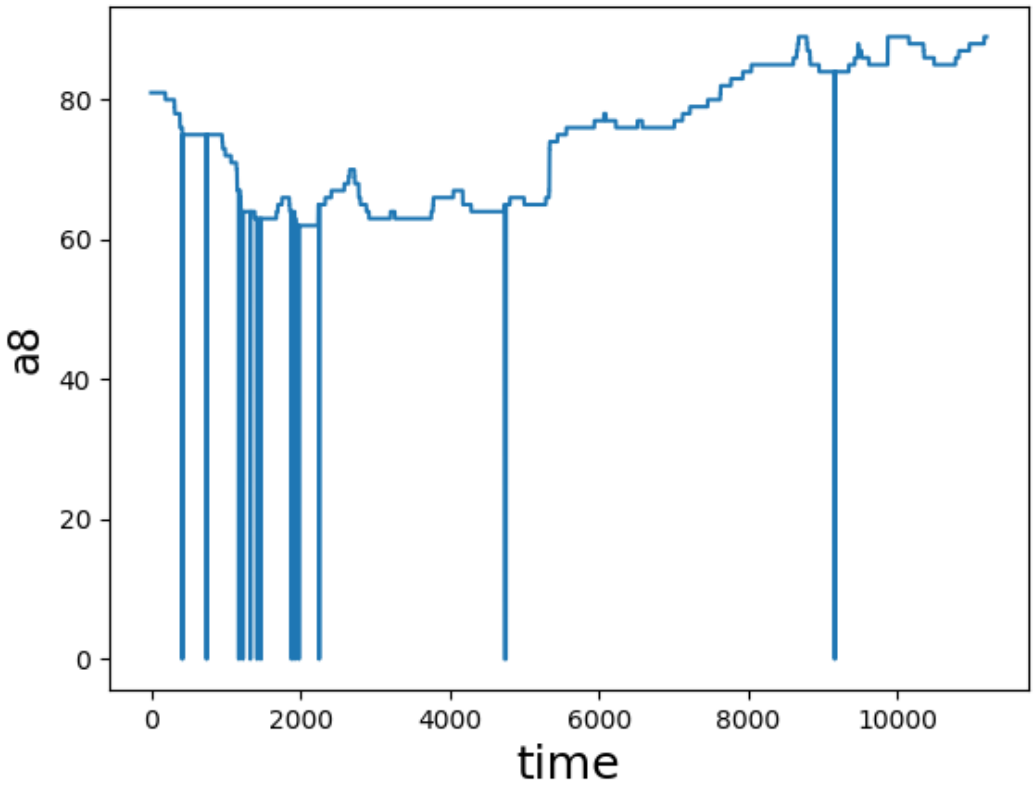

80
60
40
20
0
a3
0
2000
4000
6000
8000
10000
time


7000
6000
5000
4000
3000
2000
1000
0
a4
0
2000
4000
6000
8000
10000
time


80
60
40
20
0
a5
0
2000
4000
6000
8000
10000
time

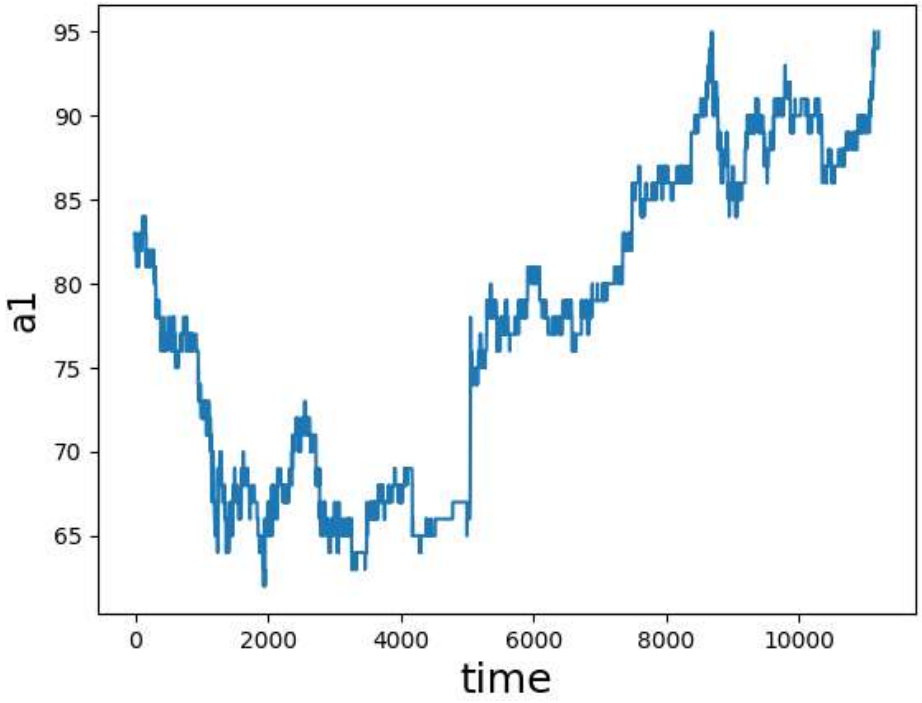

140000
120000
100000
80000
60000
40000
20000
0
0
2000
4000
6000
8000
10000
time

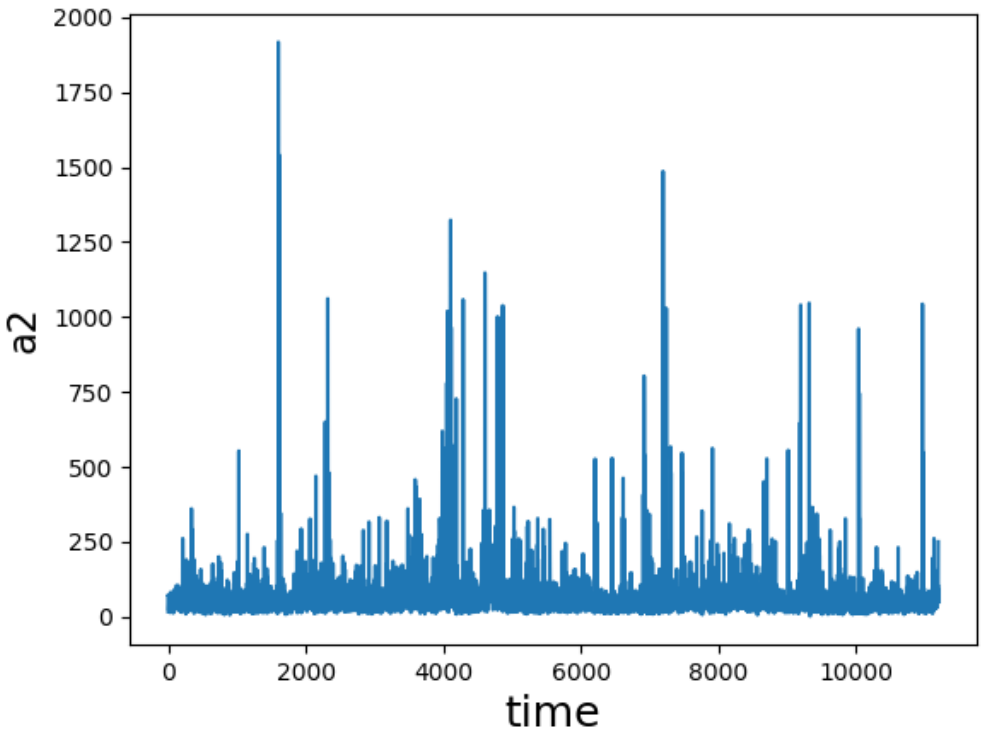

100
80
60
40
20
0
a7
0
2000
4000
6000
8000
10000
time

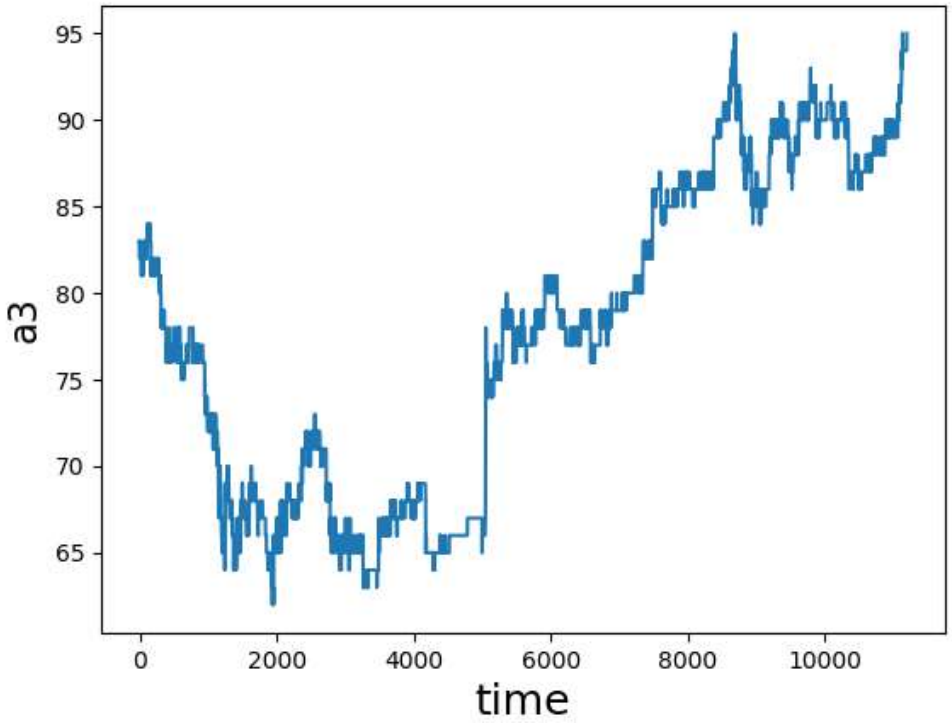


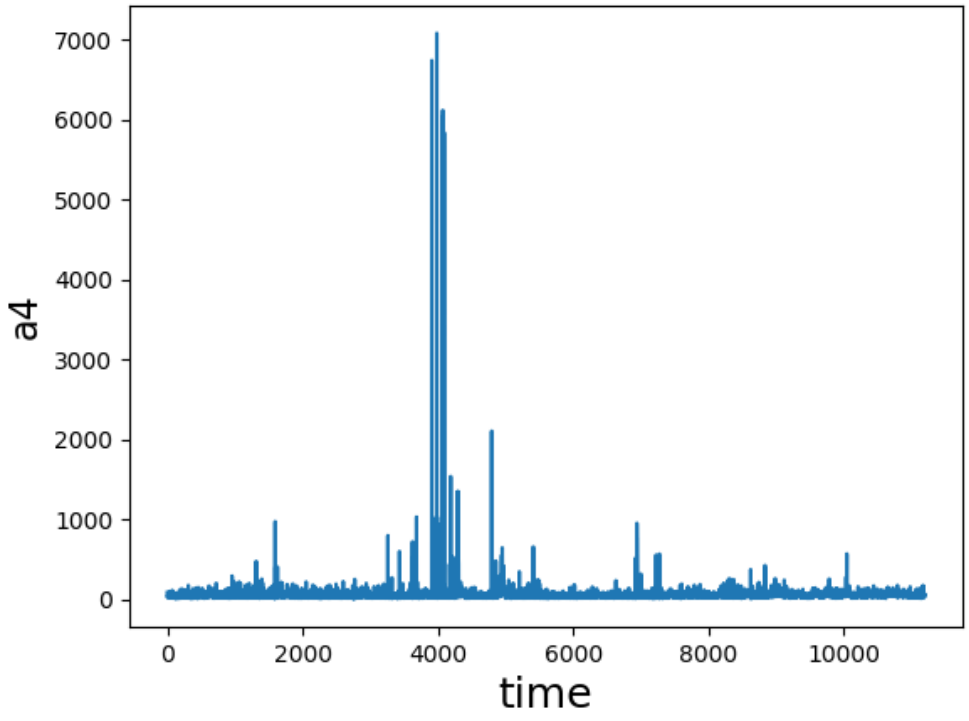
همانطور که از شکل های کشیده شده نتیجه گیری می شود نمودار تغییرات a۱ و a۳ و a۵ و a۷ و a۸ بر حسب زمان مشابه هم هستند.

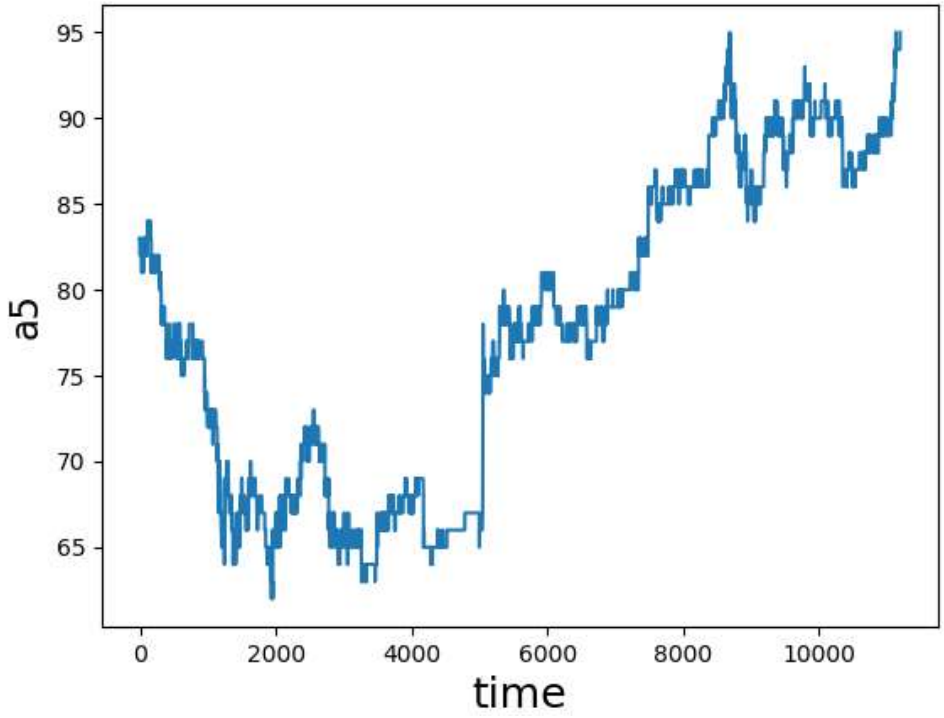
همچنین از شکل های کشیده شده نتیجه گیری می شود که فایل داده ها دارای داده های پرتی یا دادهای اشتباهی هستند که باعث ایجاد خطا در مراحل بعدی میشوند در نتیجه آن ها را از اطلاعات حذف کردیم و همانطور که در شکل های زیر مشاهده می شود تغییرات ناگهانی داده ها از بین رفت.

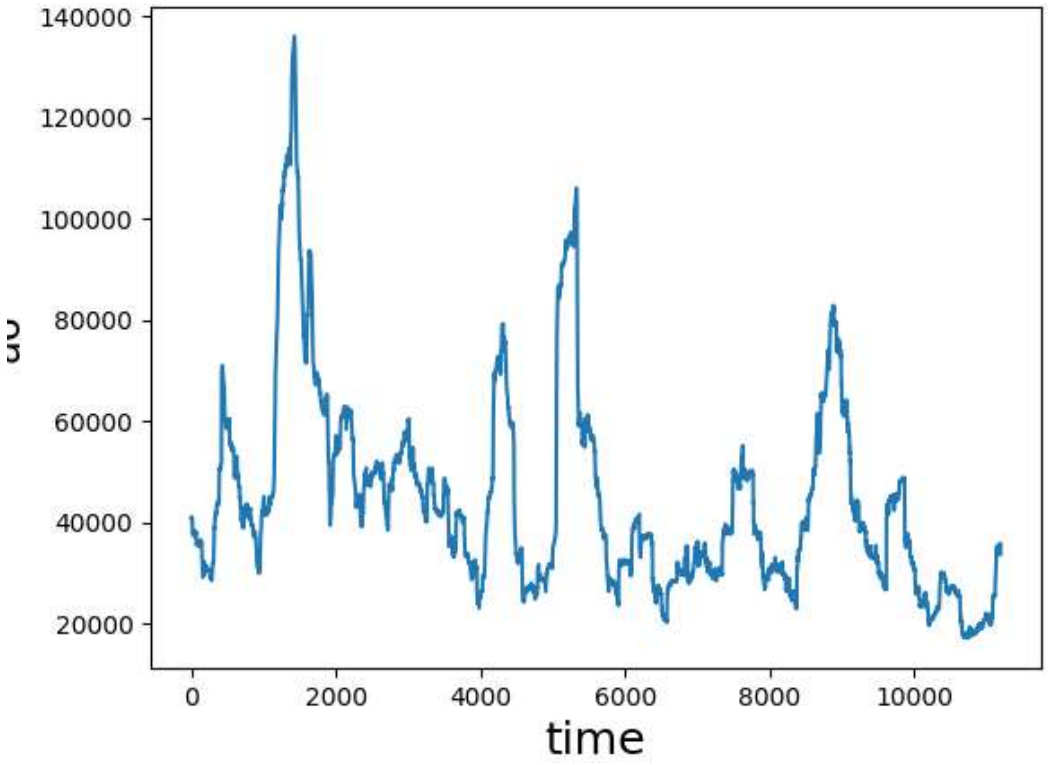

95
90
85
80
75
70
65
a1
0
2000
4000
6000
8000
10000
time

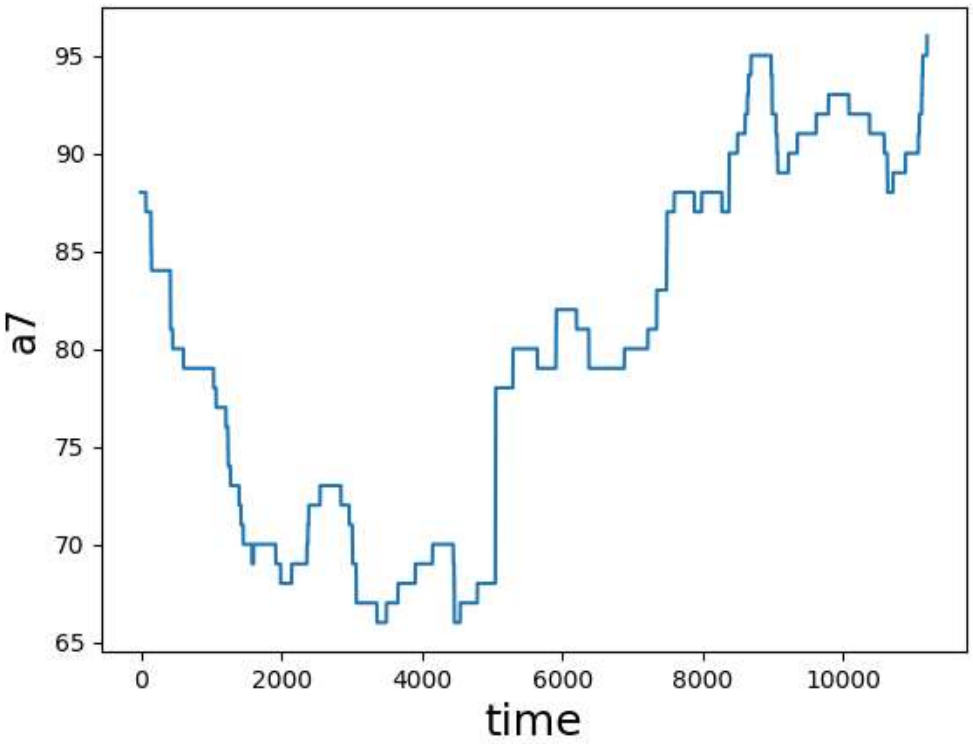

2000
1750
1500
1250
1000
750
500
250
0
a2
0
2000
4000
6000
8000
10000
time

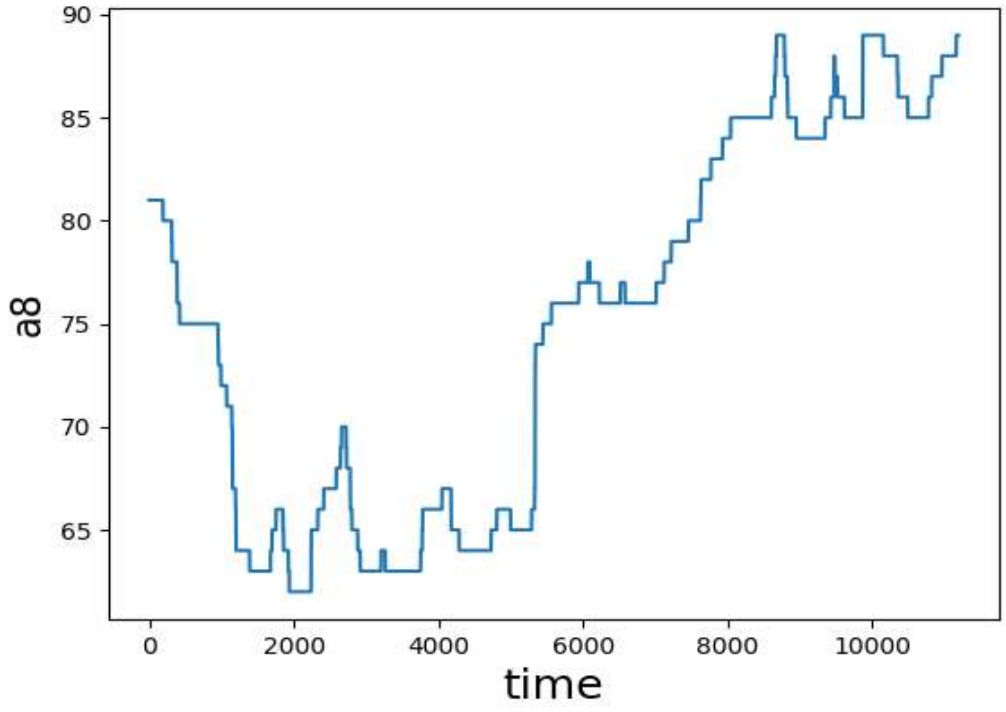

a3
95
90
85
80
75
70
65
0
2000
4000
6000
8000
10000
time


a4
7000
6000
5000
4000
3000
2000
1000
0
0
2000
4000
6000
8000
10000
time


95
90
85
80
75
70
65
a5
0
2000
4000
6000
8000
10000
time


140000
120000
100000
80000
60000
40000
20000
0
2000
4000
6000
8000
10000
time

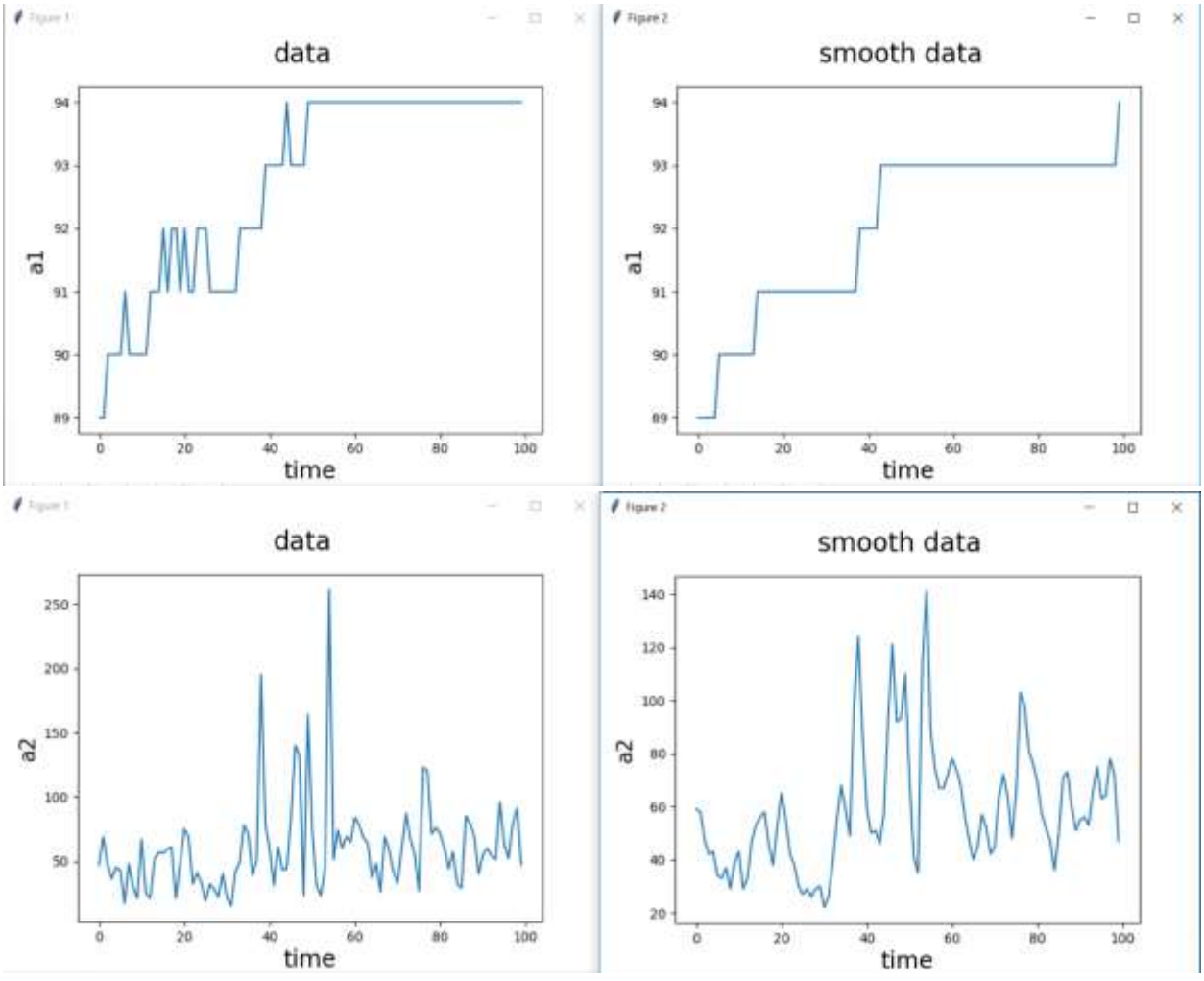

a7
95
90
85
80
75
70
65
0
2000
4000
6000
8000
10000
time

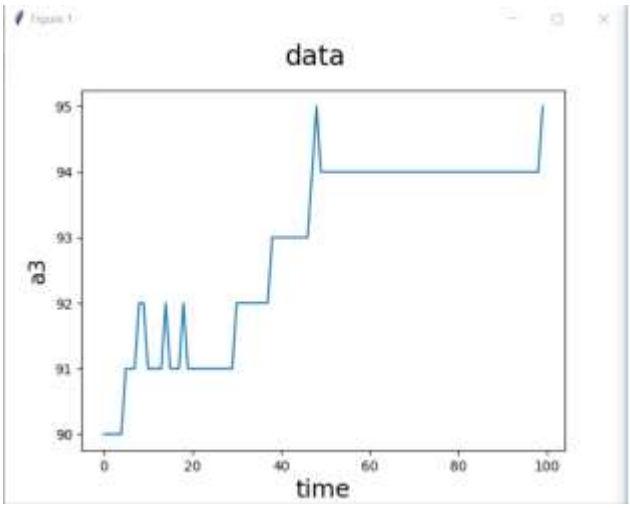

a8
90
85
80
75
70
65
0
2000
4000
6000
8000
10000
time

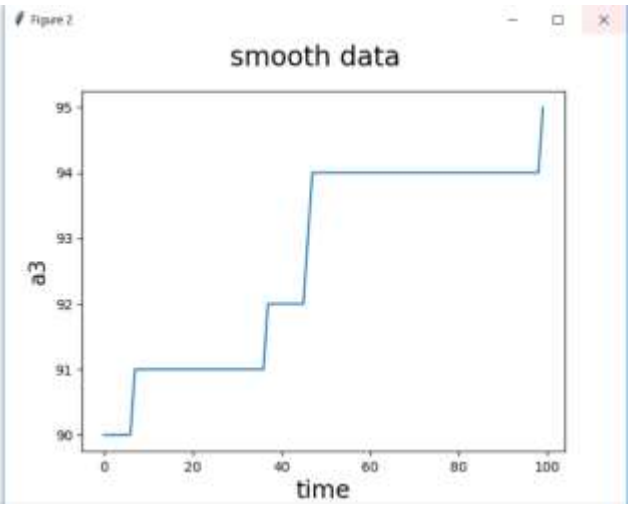
هموارسازی داده ها :

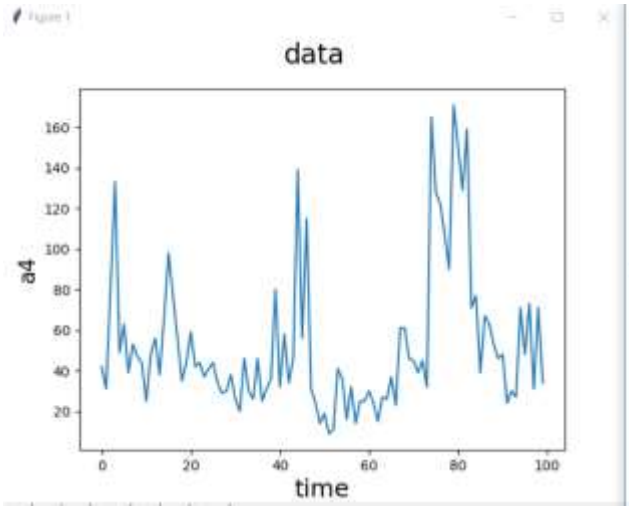
یکی دیگر از کار های انجام شده روی داده ها هموارسازی داده ها بود که به این صورت انجام شد که مقدار هر نقطه برابر میانگین مقدار آن نقطه و یک مقدار قبل آن و یک مقدار بعد آن قرار می دهید.

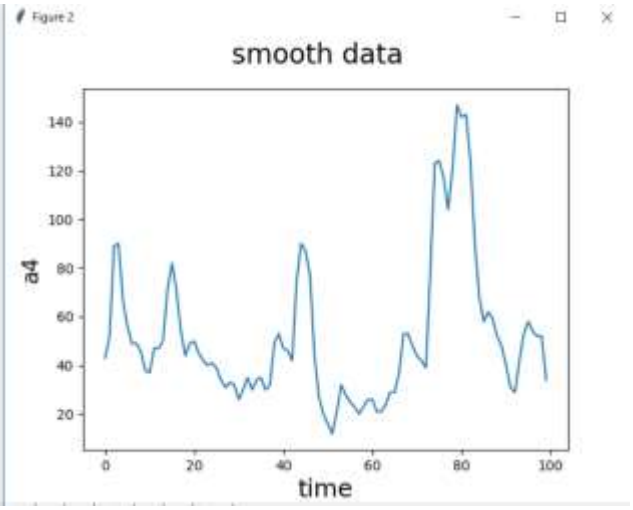
شکل ۱۰۰ داده ی آخر پس از هموار سازی(برای بهتر نشان داده شدن تغییرات پس هموارسازی همه ی دادها رسم نشده است) :

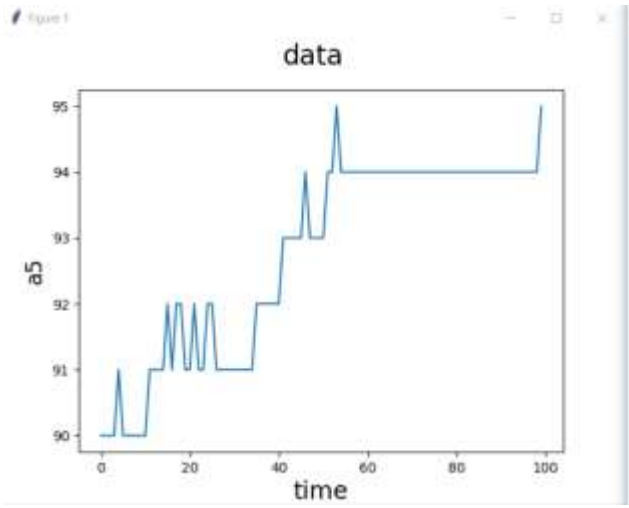


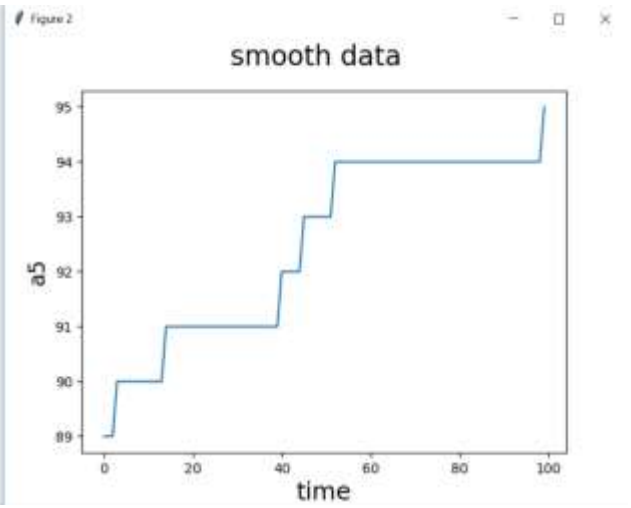

data
a3
time

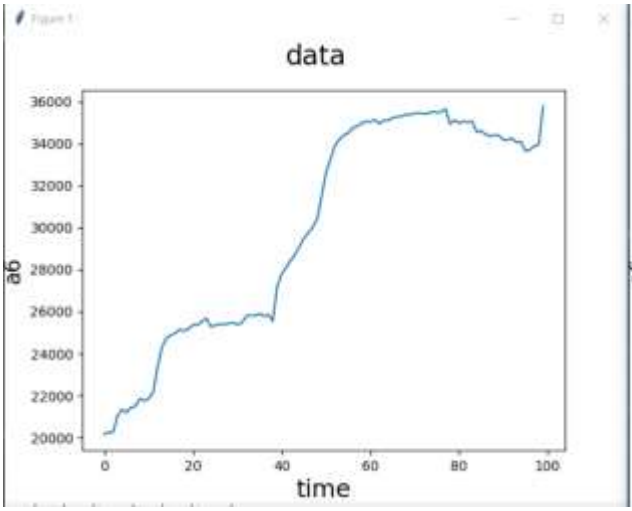

smooth data
a3
time

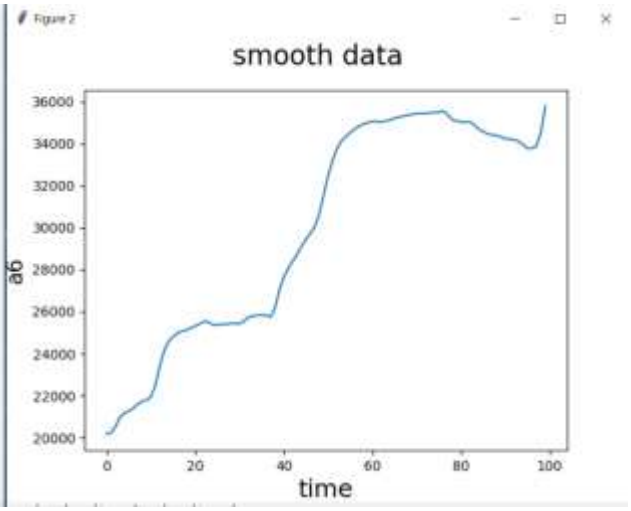

data
a4
time

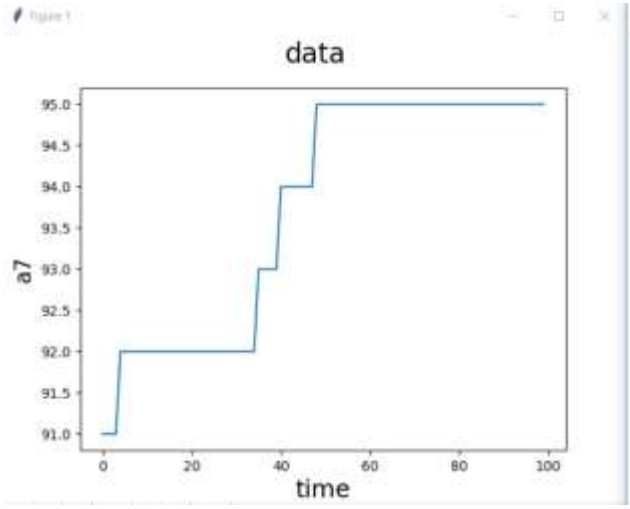

smooth data
a4
time

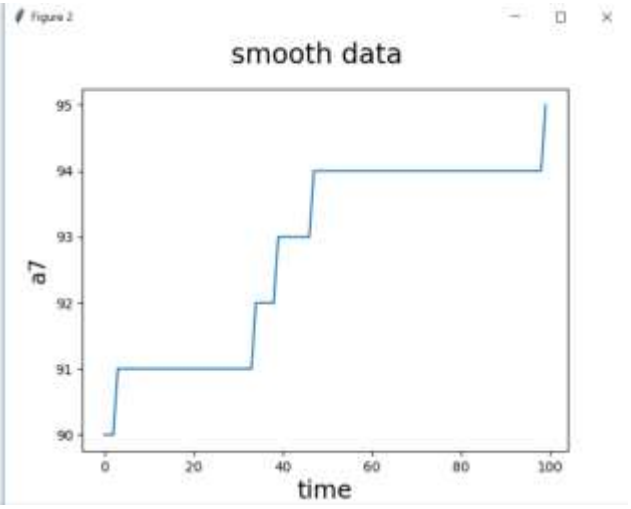

data
a5
time

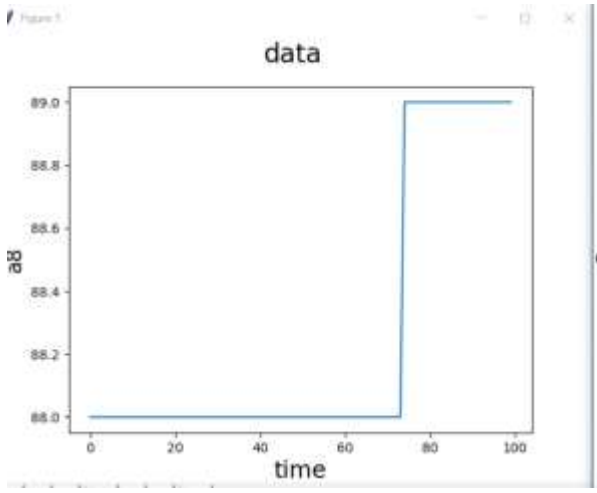

smooth data
a5
time

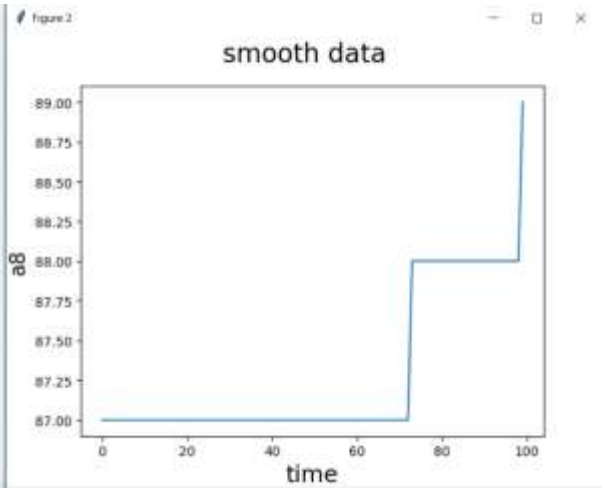

data
a6
time


smooth data
a6
time


data
a7
time


smooth data
a7
time

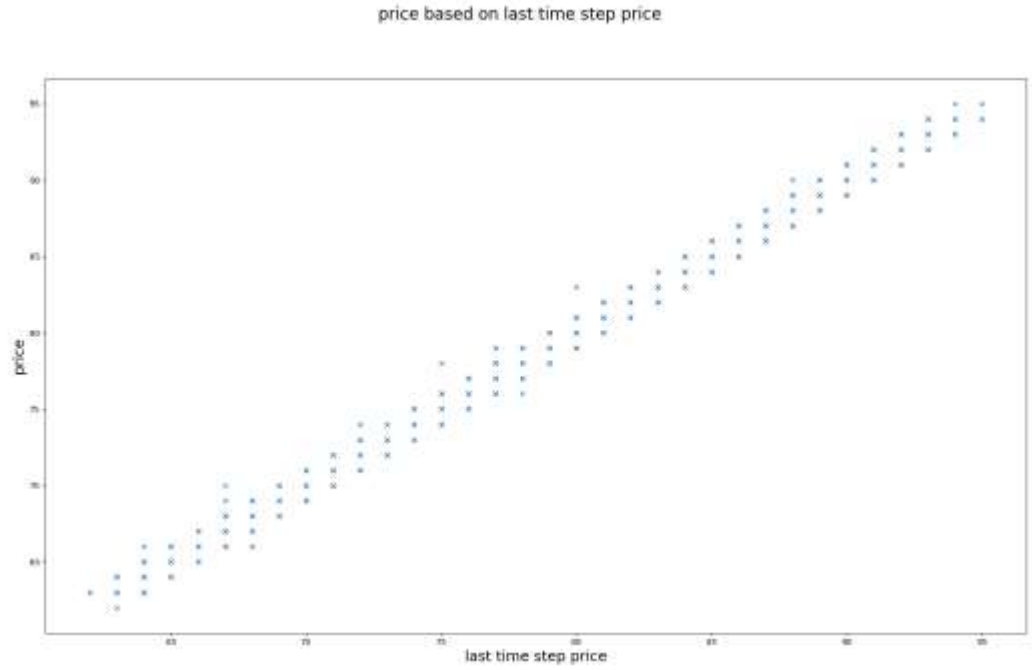

data
a8
time


smooth data
a8
time

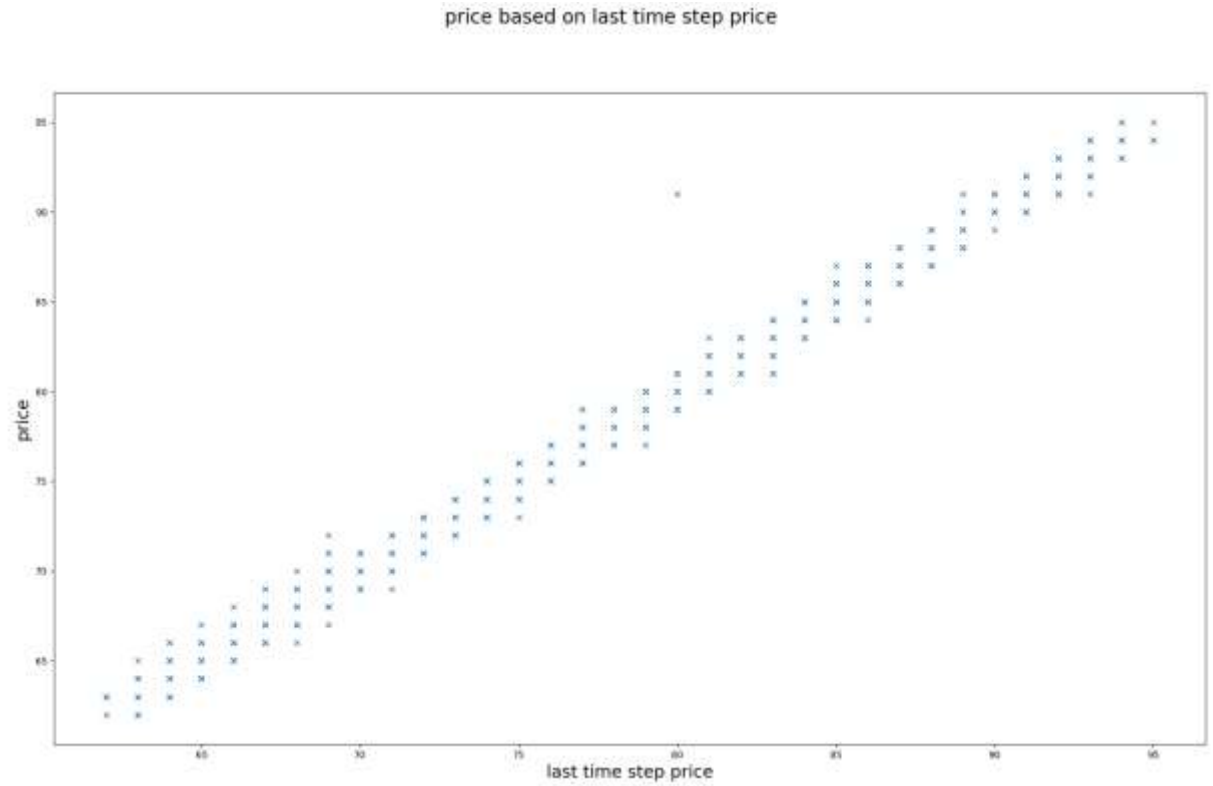
تصویر سازی داده ها:

همان طور که پیش تر گفته شد داده ها به این صورت مدل شدند که مقدار قیمت در هر لحظه تابع مقدار قیمت در لحظه قبلی در نظر گرفته شد. بر همین اساس داده ها رسم کردیم و به شکل های زیر رسیدیم:

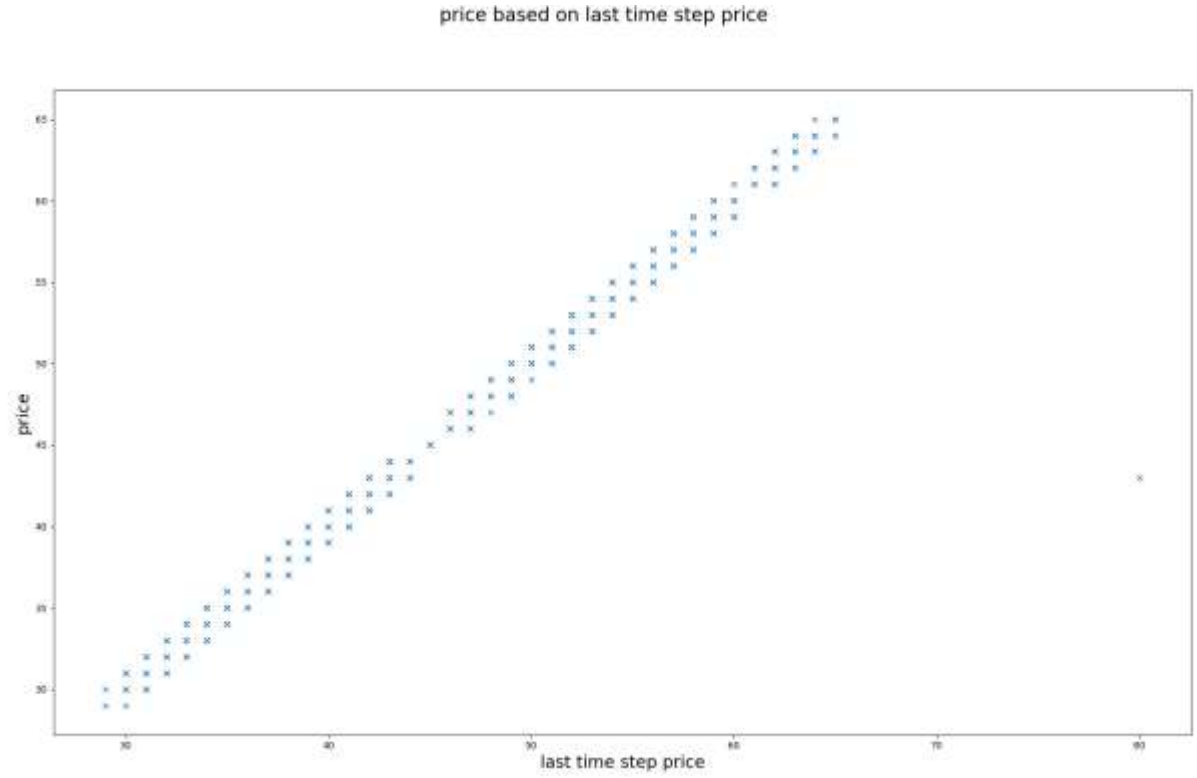
A_ticke (a1):



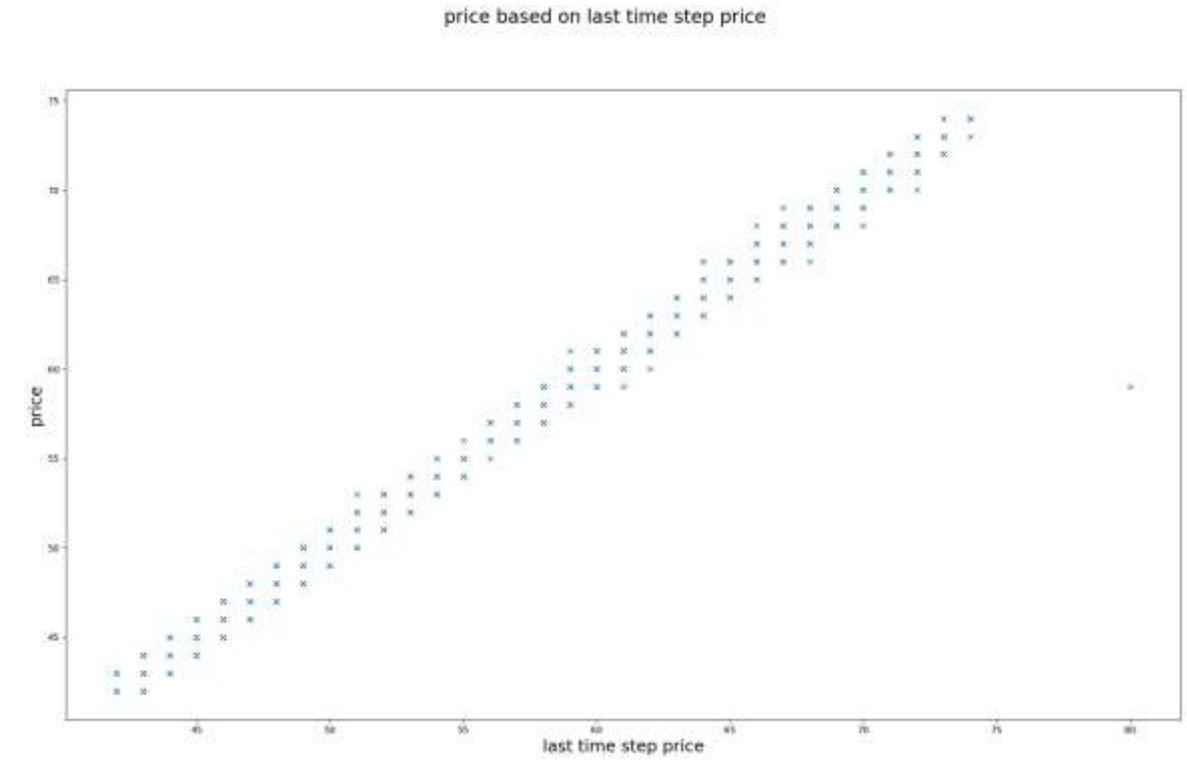
B_ticker(b1):



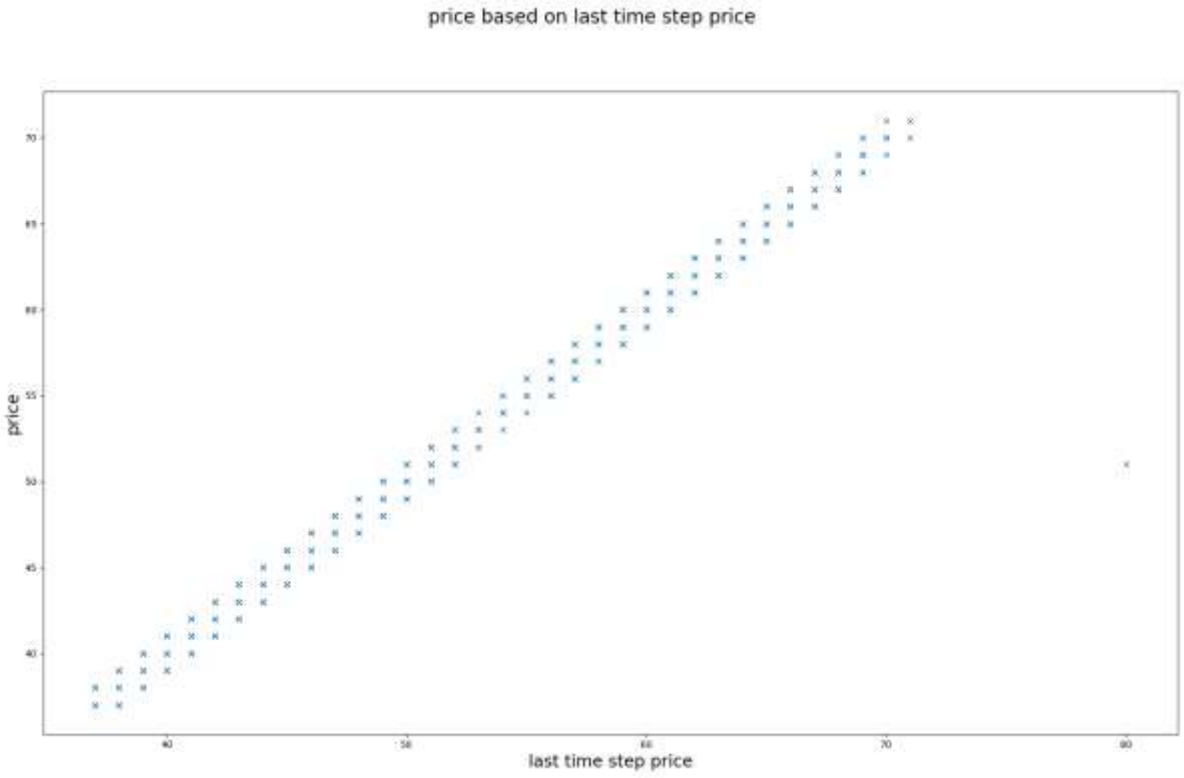
C_ticker(c1):



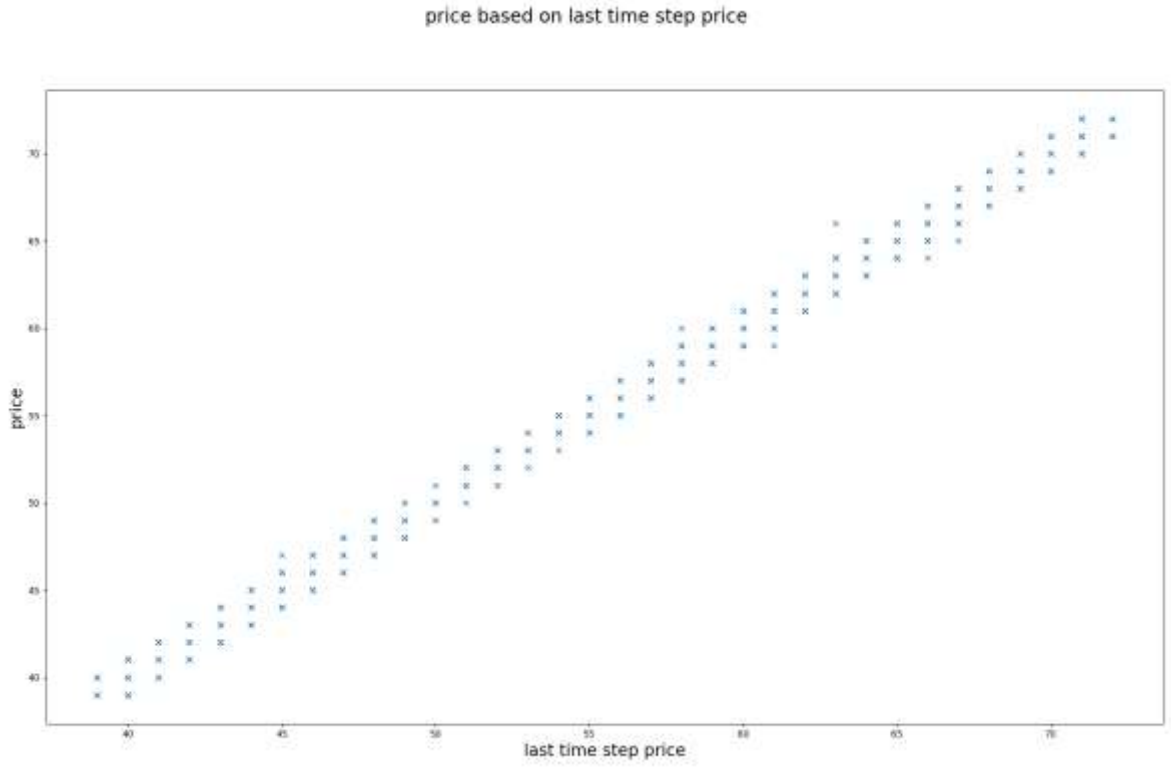
D_ticker(d1):



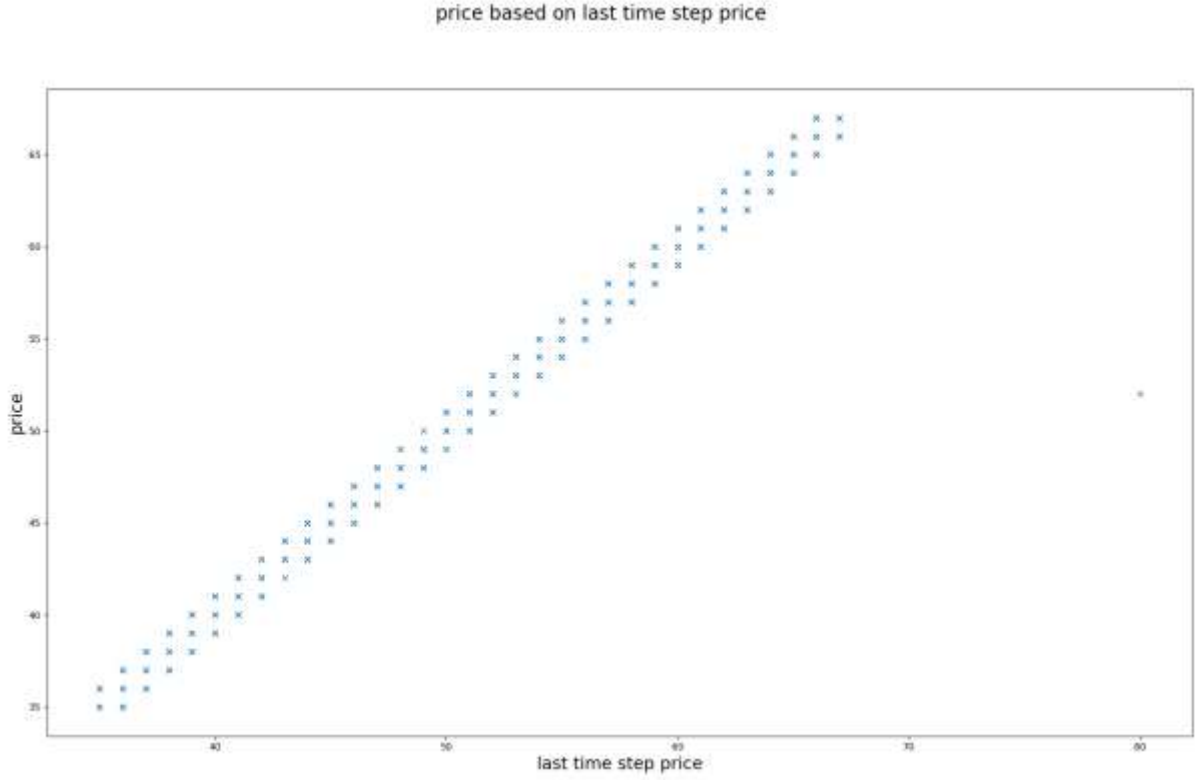
E_ticker(e1):



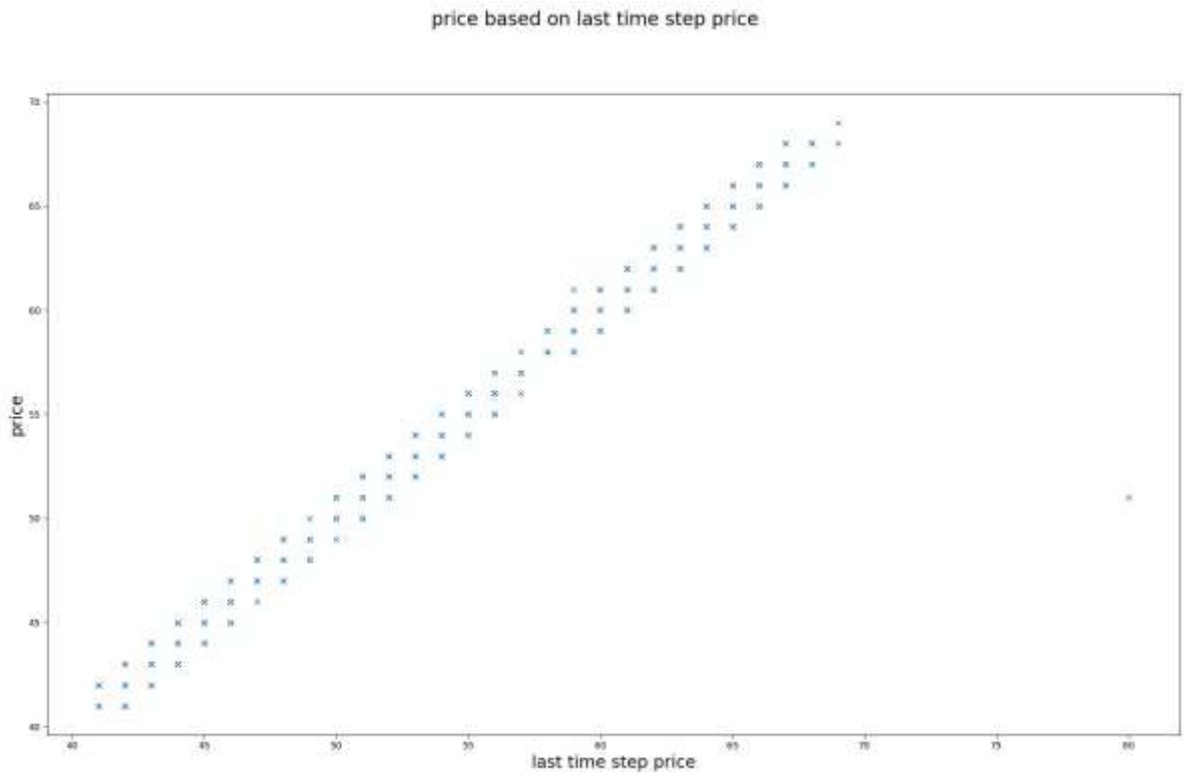
F_ticker(f1):



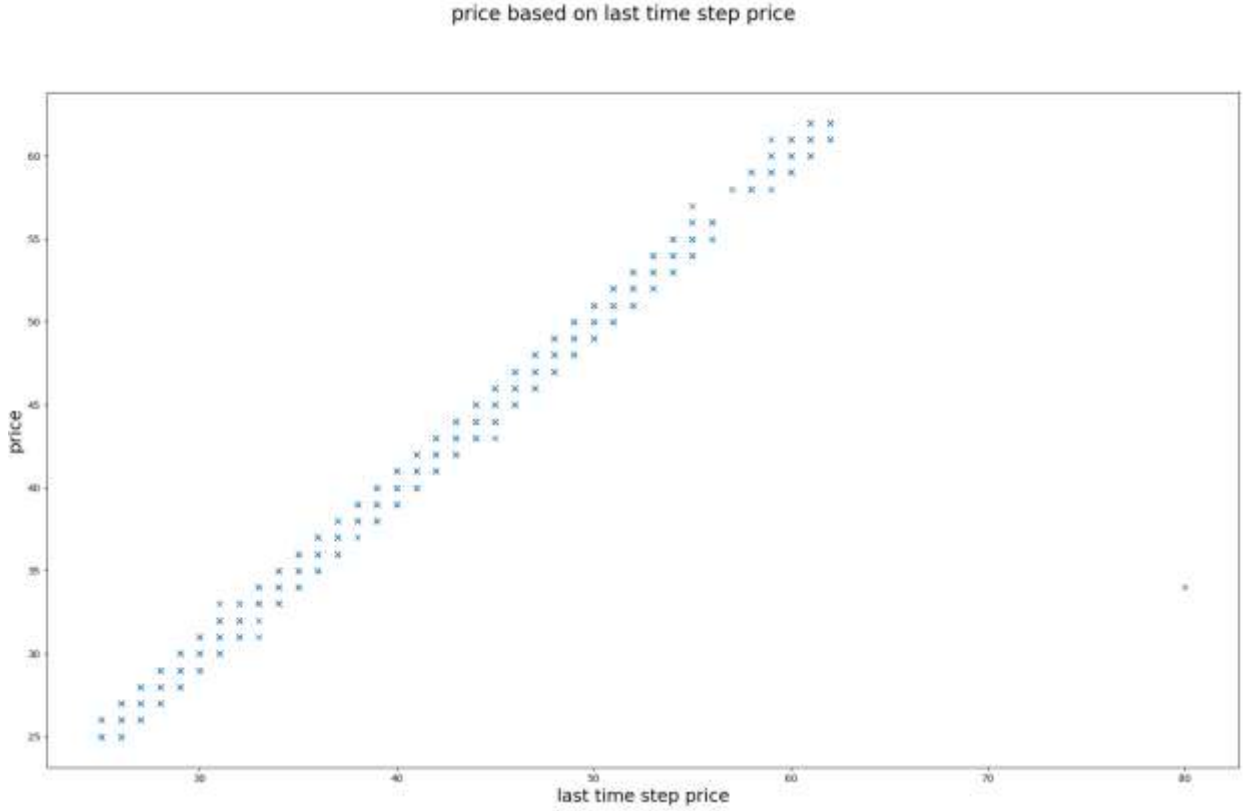
G_ticker(g1):



H_ticker(h1):


price based on last time step price
price
last time step price

I_ticker(i1):


price based on last time step price
price
last time step price

همان طور که مشاهده می شود داده ها حالت خطی دارند و برای همین پیش بینی می شود که مدل های خطی نتیجه خوبی داشته باشند.

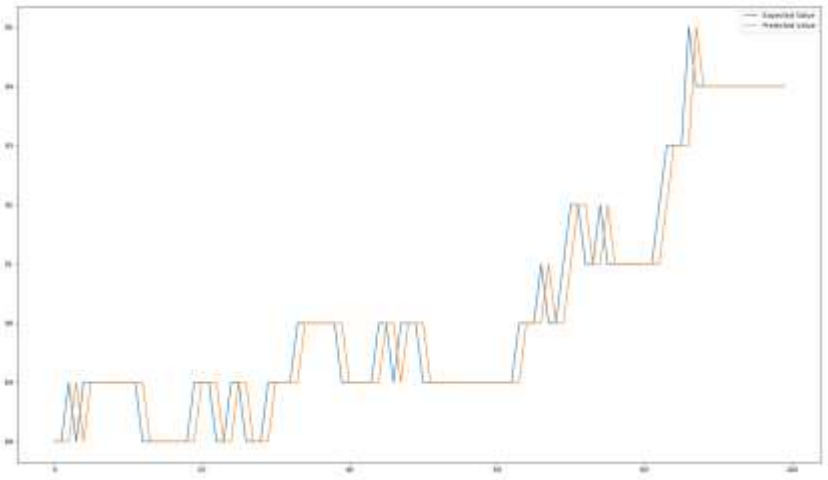
بخش دوم :

به دست آوردن مدل اولیه baseline و محاسبه ی خطای آن :

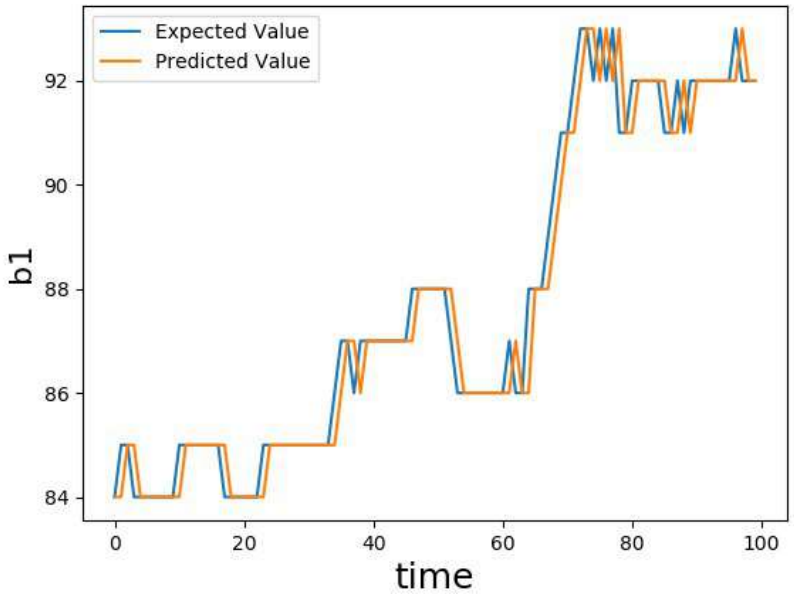
به طور کلی برای اجرای این الگوریتم ها و پیدا کردن مدلی برای تخمین قیمت آینده نیاز به ایجاد داده های test و train داشتیم که برای این کار بعد از خواندن اطلاعات فایل اول به صورت رندوم ۲۰ درصد از داده ها به عنوان test و ۸۰ درصد دیگر به عنوان train انتخاب شدند.

قبل از شروع تست کردن مدل های مختلف روی داده نیاز بود که یک baseline برای تخمین خود داشته باشیم. بدین منظور که بتوانیم دیگر مدل ها را با ان مقایسه کنیم و از مقایسه آن ها با این مدل میزان کارایی آن ها را مورد بررسی قرار دهیم .
برای این منظور از یک مدل ساده که برای تخمین داده های سری زمانی استفاده می شود و مقدار قیمت در هر گام زمانی را برابر قیمت در گام قبلی در نظر می گیرد استفاده کرده ایم.
برای بررسی مدل تصویر داده های اصلی و مقدار تخمین زده شده کشیده شد و همچنین میزان خطایrmse مدل محاسبه شده است

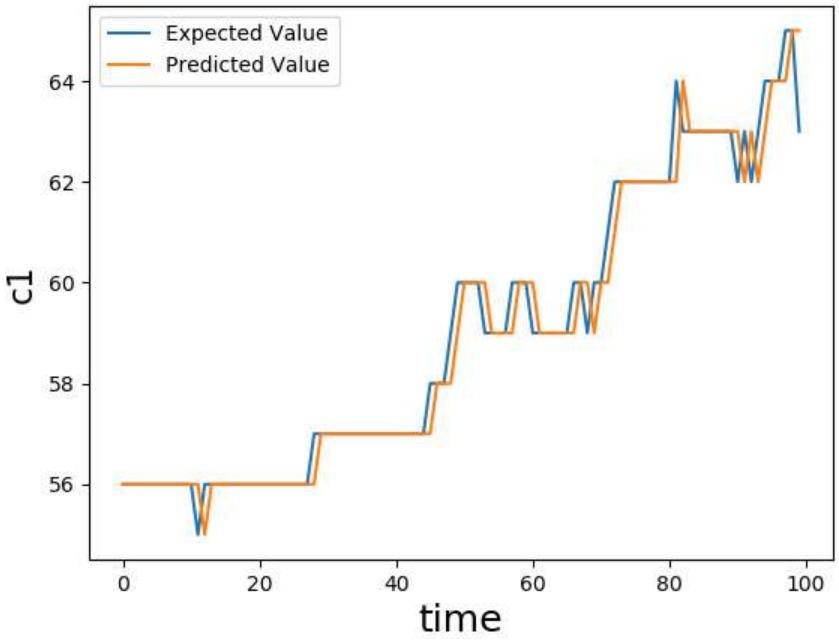
تصویر ۱۰۰ تا از داده های تست فایل A_ticker برای baseline :



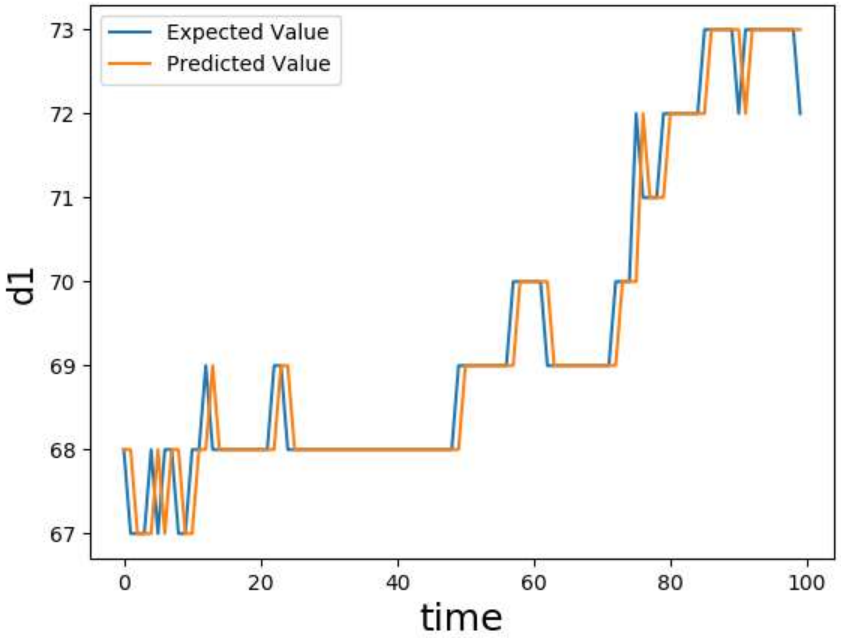
تصویر ۱۰۰ تا از داده های تست فایل B_ticker برای baseline :



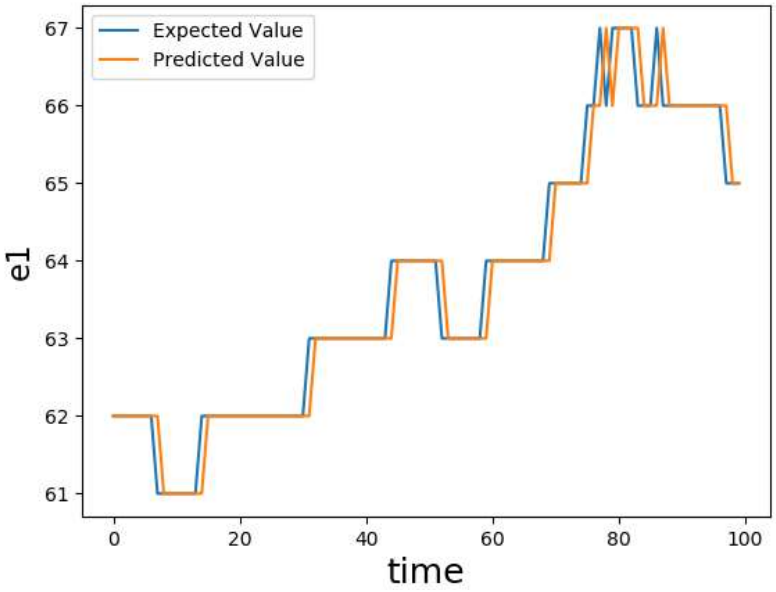
تصویر ۱۰۰ تا از داده های تست فایل C_ticker برای baseline :



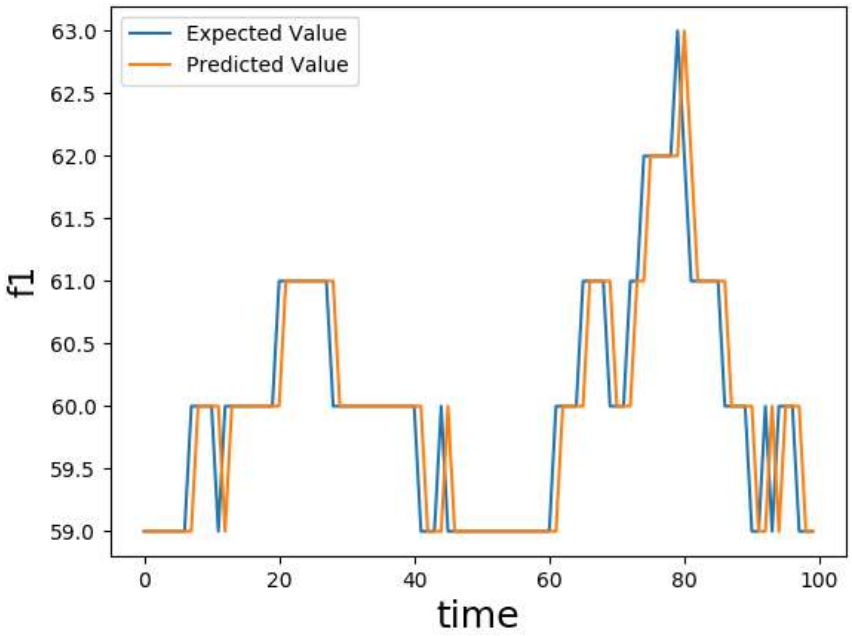
تصویر ۱۰۰ تا از داده های تست فایل D_ticker برای baseline :



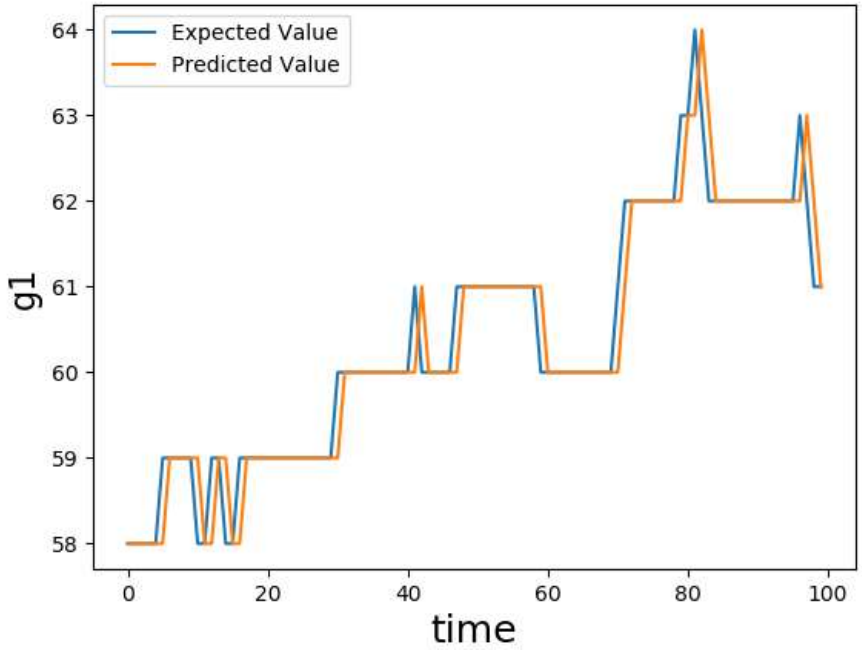
تصویر ۱۰۰ تا از داده های تست فایل E_ticker برای baseline :



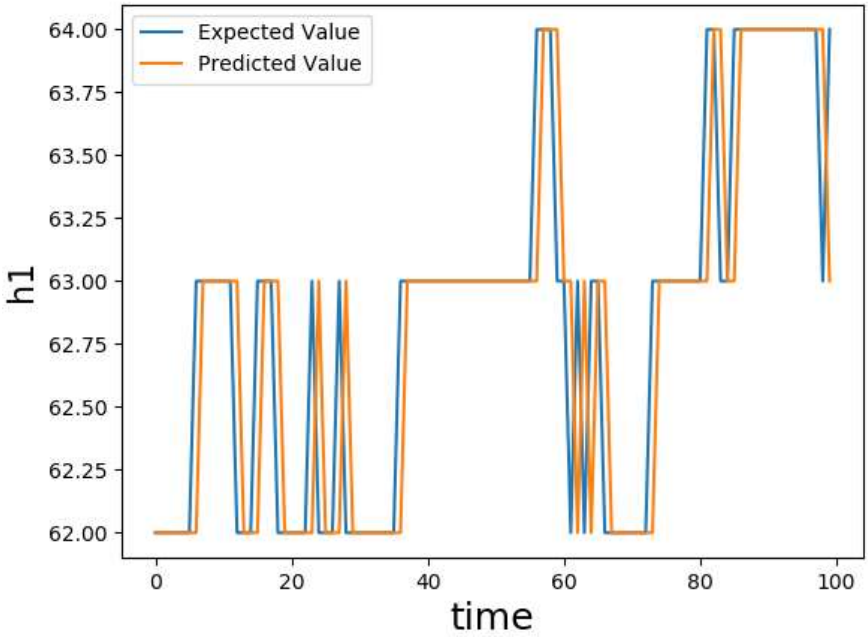
تصویر ۱۰۰ تا از داده های تست فایل F_ticker برای baseline :



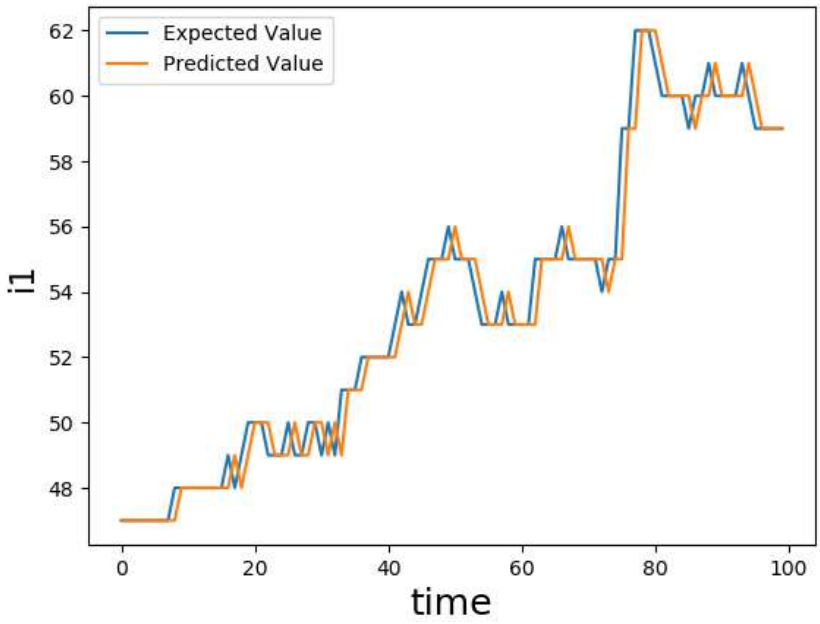
تصویر ۱۰۰ تا از داده های تست فایل G_ticker برای baseline :



تصویر ۱۰۰ تا از داده های تست فایل H_ticker برای baseline :



تصویر ۱۰۰ تا از داده های تست فایل I_ticker برای baseline :



میانگین خطای RMSE و واریانس بعد از ۱۰ بار اجرای مدل baseline روی فایل a_ticker تا i_ticker :

I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.733	0.551	0.525	0.588	0.582	0.623	0.634	0.664	0.625	RMSE
0.0001	0.0002	0.0001	0.0001	0.0000	0.0001	0.0000	0.0002	0.0002	VAR

همچنین از داده های smooth شده که در قسمت پردازش داده توضیح داده شد نیز استفاده کردیم ولی به دلیل این که تغییر چندانی در خطاها ایجاد نکرد کارایی زیادی نداشت.

نکته ی قابل توجه این است که فقط داده های بخش train را هموار کردیم و برای تخمین از داده های هموار نشده ی test استفاده کردیم.

بخش سوم :

پیاده سازی مدل svm با کرنل rbf :

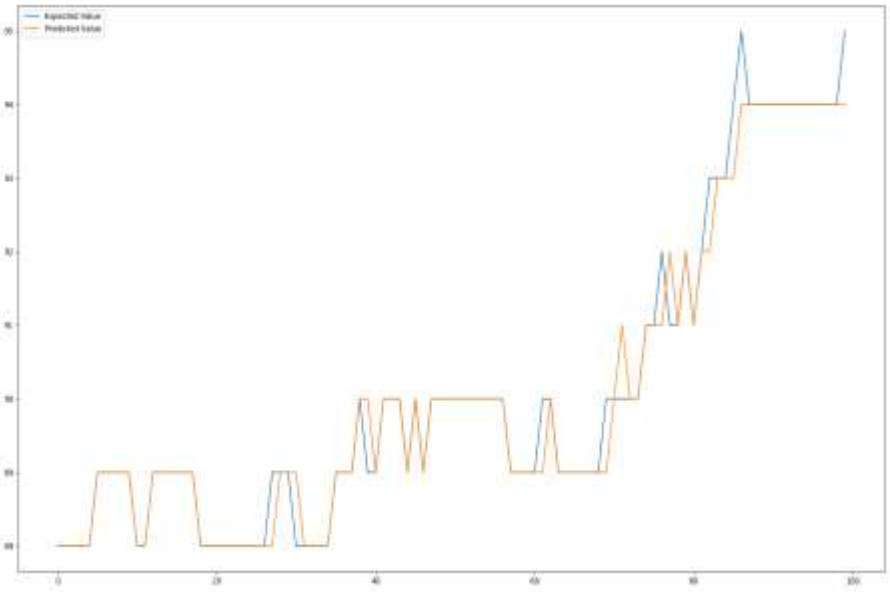
بعد از به دست آمدن مدل اولیه از یک الگوریتم پیچیده تر به نام svm برای ایجاد مدل و تخمین داده ها استفاده کردیم .

همانطور که در بخش قبلی هم گفته شد به طور کلی برای اجرای این الگوریتم ها و پیدا کردن مدلی برای تخمین قیمت آینده نیاز به ایجاد داده های test و train داشتیم که برای این کار بعد از خواندن اطلاعات فایل اول به صورت رندوم ۲۰ درصد از داده ها به عنوان test و ۸۰ درصد دیگر به عنوان train انتخاب شدند.

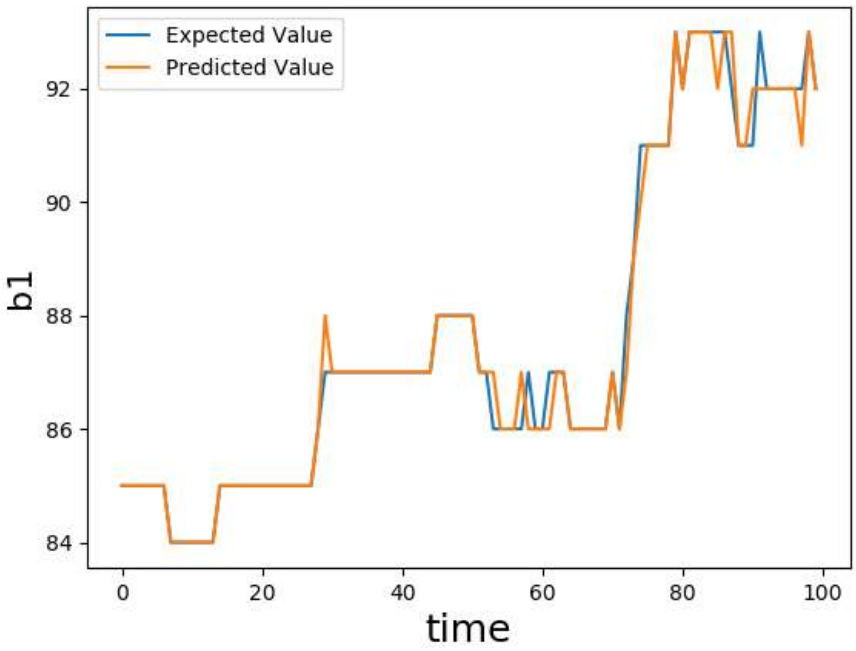
حال به بررسی شکل داده های تخمین زده شده و خطای مربوط به الگوریتم های پیاده سازی شده می پردازیم :

الگوریتم svm با کرنل rbf :

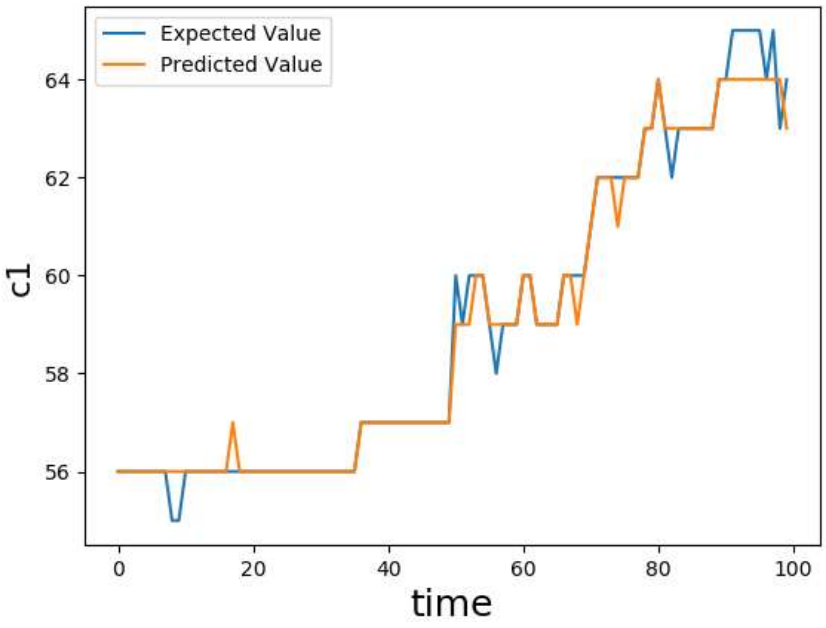
تصویر ۱۰۰ تا از داده های تست فایل A_ticker برای svm :



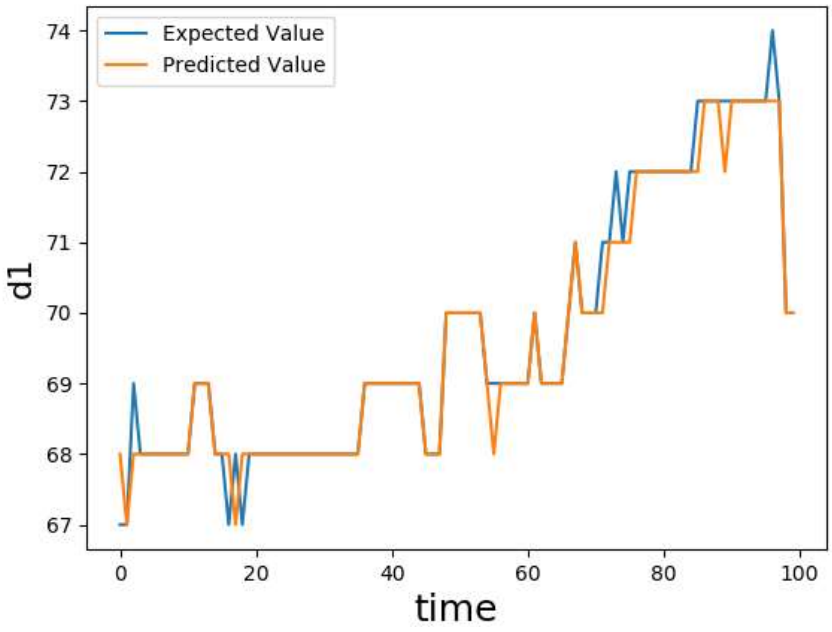
تصویر ۱۰۰ تا از داده های تست فایل B_ticker برای svm :



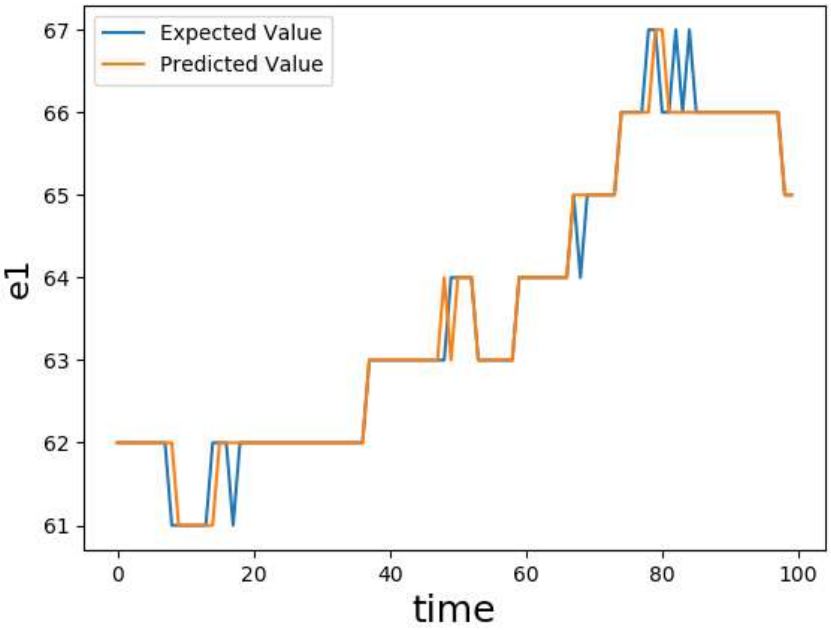
تصویر ۱۰۰ تا از داده های تست فایل C_ticker برای svm :



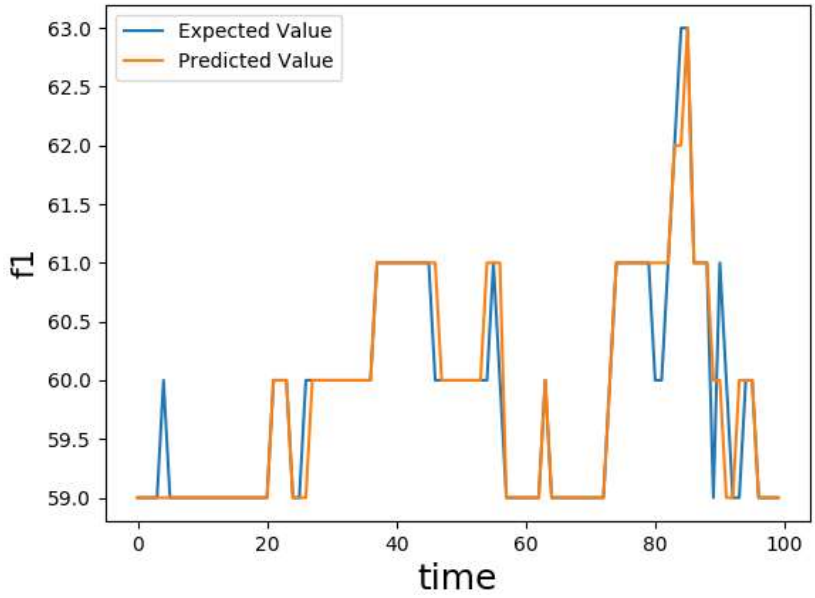
تصویر ۱۰۰ تا از داده های تست فایل D_ticker برای svm :



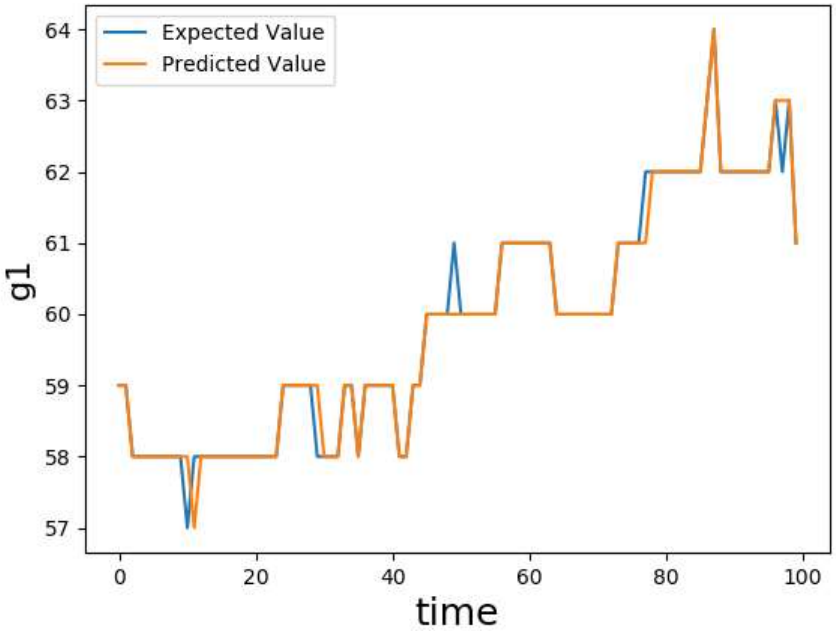
تصویر ۱۰۰ تا از داده های تست فایل E_ticker برای svm :



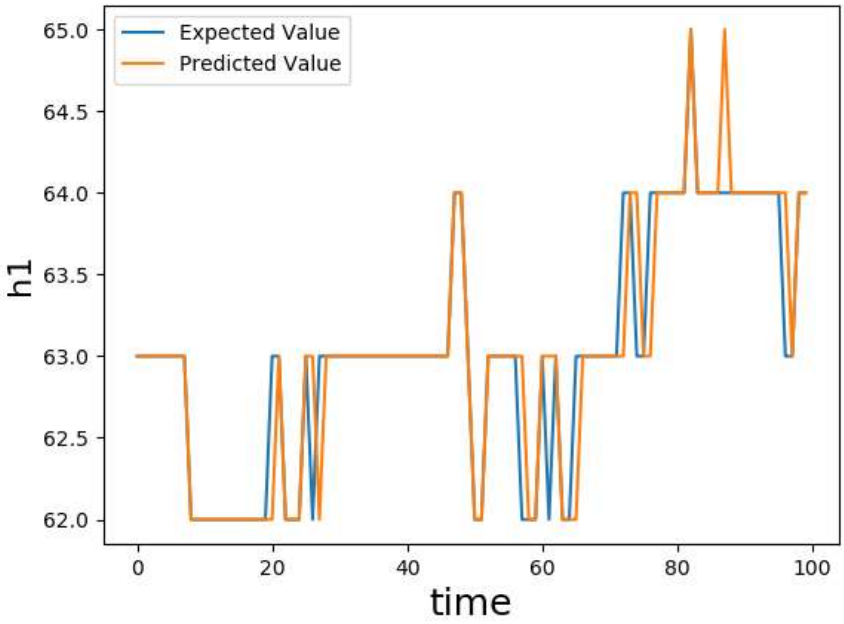
تصویر ۱۰۰ تا از داده های تست فایل F_ticker برای svm :



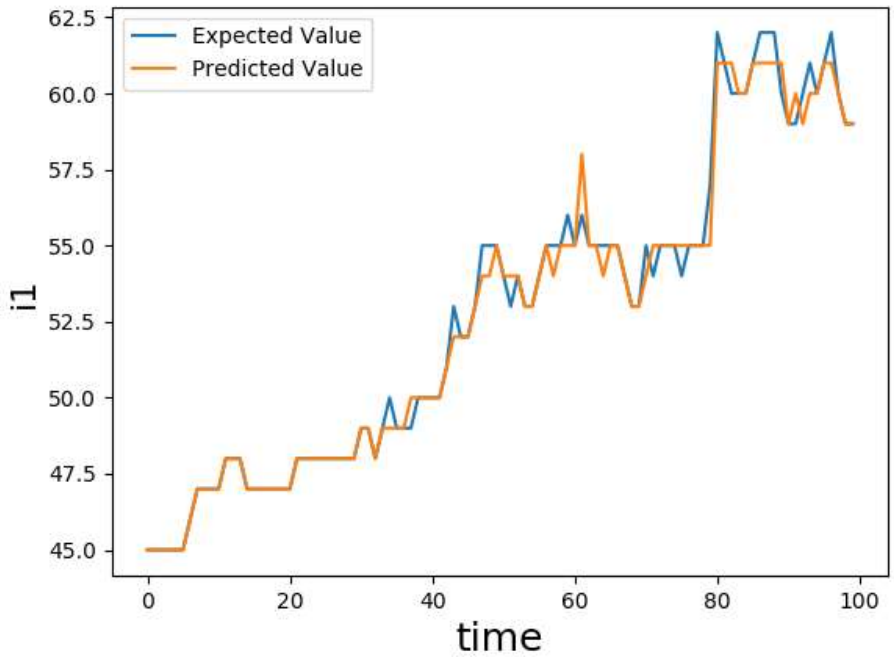
تصویر ۱۰۰ تا از داده های تست فایل G_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل H_ticker برای svm :



تصویر ۱۰۰ تا از داده های تست فایل I_ticker برای svm :



میانگین خطای RMSE و واریانس بعد از ۱۰ بار اجرای مدل SVM روی فایل a_ticker تا i_ticker :

I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.323	0.325	0.312	0.361	0.344	0.355	0.303	0.436	0.370	RMSE
0.0000	0.0008	0.0014	0.0022	0.0065	0.0001	0.0001	0.0020	0.0000	VAR

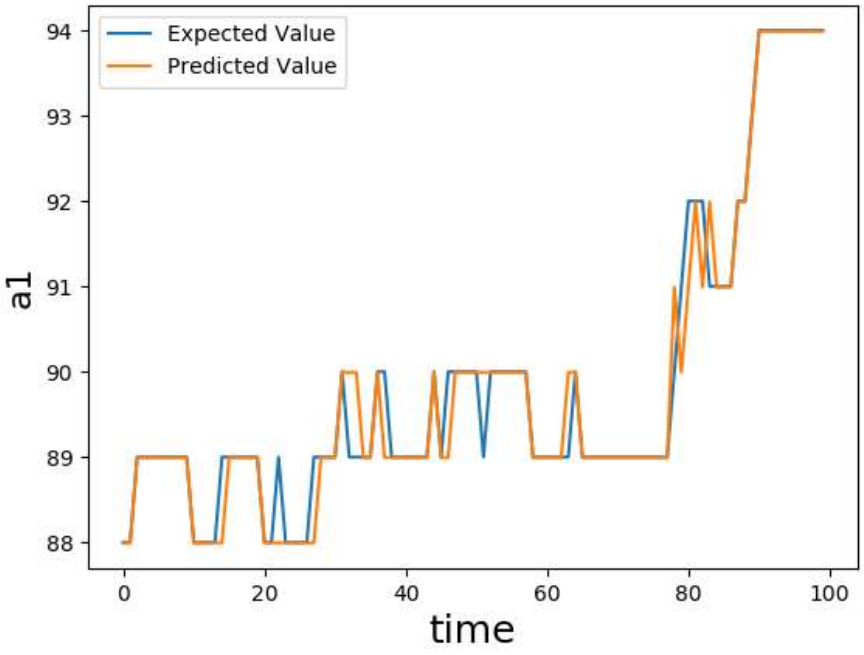
همچنین از داده های smooth شده که در قسمت پردازش داده توضیح داده شد نیز استفاده کردیم ولی به دلیل این که تغییر چندانی در خطاها ایجاد نکرد کارای زیادی نداشت.

بخش چهارم :

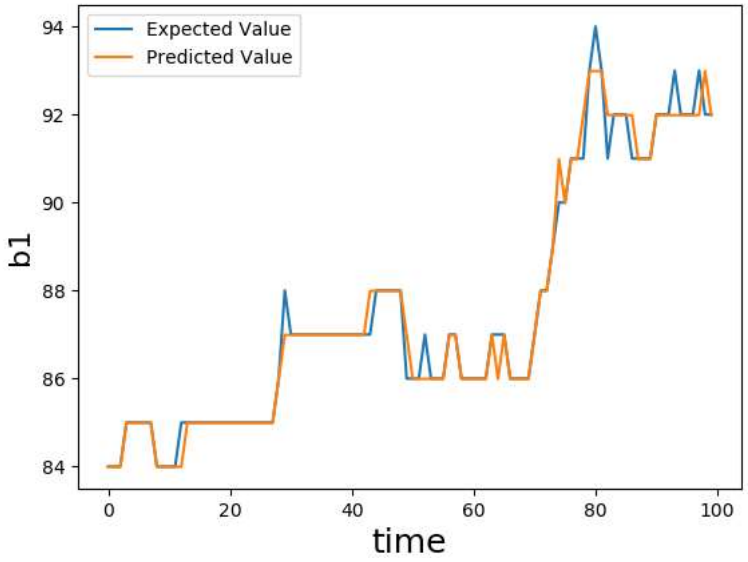
مدل خطی :

از آن جایی که پس از تضویر سازی داده ها (رسم مقدار قیمت در هر زمان بر اساس زمان قبلی) نمودار ظاهری خطی پیدا کرد مدل خطی را امتحان کردیم و نتایج خوبی گرفتیم.

تصویر ۱۰۰ تا از داده های تست فایل A_ticker **برای مدل خطی :**



تصویر ۱۰۰ تا از داده های تست فایل B_ticker **برای مدل خطی :**



شکل ها برای همه داده ها کشیده شده و بررسی شده ولی برا کوتاه کردن گزارش شکل دو جنس اول را در گزارش قرار داده ایم .

میانگین خطای RMSE **و واریانس بعد از ۱۰ بار اجرای مدل** LINEAR **روی فایل** a_ticker **تا** i_ticker **:**

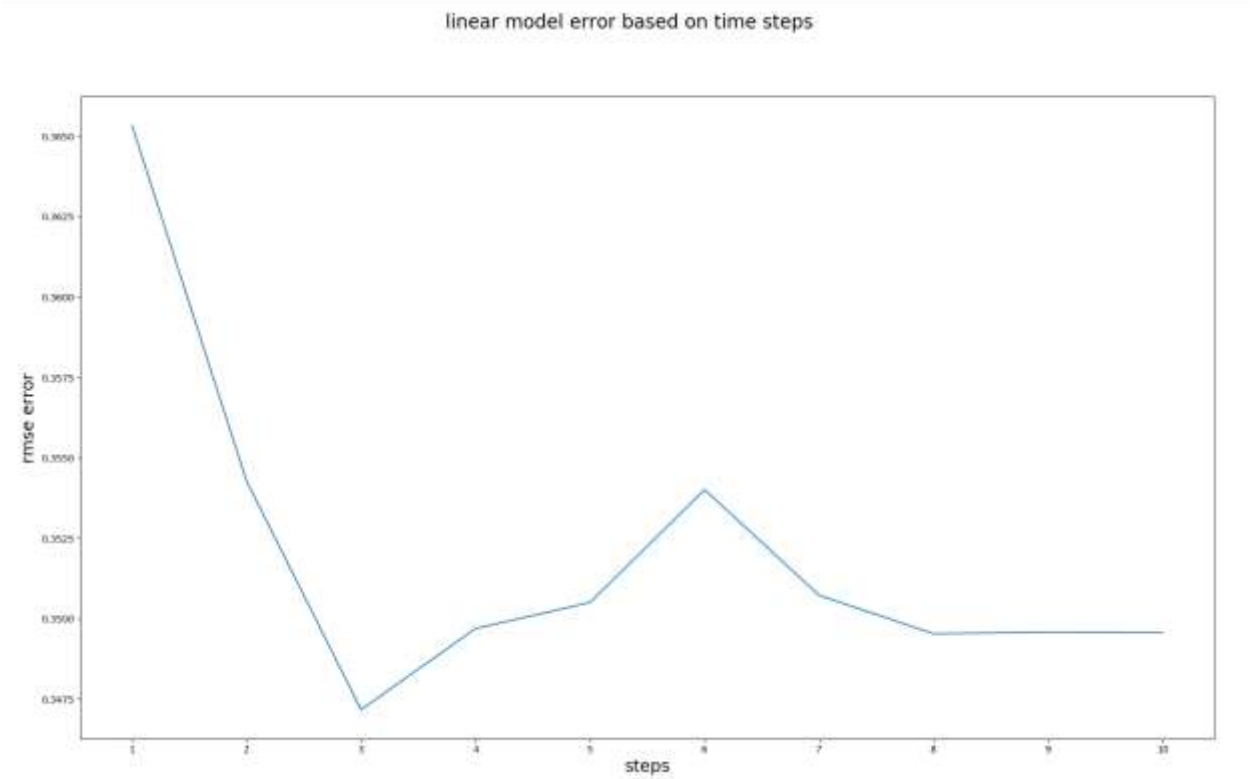
I_TICKER	H_TICKER	G_TICKER	F_TICKER	E_TICKER	D_TICKER	C_TICKER	B_TICKER	A_TICKER	
0.386	0.377	0.387	0.430	0.368	0.401	0.304	0.436	0.370	RMSE
0.0369	0.0170	0.0212	0.0232	0.0167	0.0033	0.0001	0.0019	0.0000	VAR

بخش پنجم :

استفاده از چند گام زمانی قبلتر برای تخمین :

همانطور که در بالاتر هم گفته شد برای تخمین از مقدار گام قبلی استفاده کردیم در این بخش از چند گام زمانی قبلی برای تخمین نیز استفاده کردیم تا تفاوت خطاها را بررسی کنیم .

- **نمودار خطای مدل خطی بر حسب تعداد گام در نظر گرفته شده:**



این نمودار میانگین خطای مدل خطی به ازای مقدار گام های در نظر گرفته شده را در ۱۰ اجرای متوالی نشان می دهد. همان طور که مشاهده می شود نقطه کمینه این نمودار در ۳ گام رخ داده از همین رو برای مدل خطی بهینه این است که تا۳ گام قبل را در نظر بگیریم.

البته برای بررسی اینکه این تفاوت معنی دار است یا نه با استفاده از T test این مسئله را بررسی می کنیم:

تعداد گام زمانی	میانگین	واریانس	تعداد دفعات اجرا
1	0.3653	0.0002	10
3	0.3472	0.0002	10

به کمک ابزار موجود در این سابت تست ttest انجام شد و نتیجه این بود که این تفاوت معنی دار است.

https://www.graphpad.com/quickcalcs/ttest1/?Format=SEM

Review your data:

Group	Group One	Group Two
Mean	0.347200	0.365300
SD	0.000632	0.000632
SEM	0.000200	0.000200
N	10	10

Unpaired *t* test results

P value and statistical significance:

The two-tailed P value is less than 0.0001

By conventional criteria, this difference is considered to be extremely statistically significant.

Confidence interval:

The mean of Group One minus Group Two equals -0.018100

95% confidence interval of this difference: From -0.018694 to -0.017506

Intermediate values used in calculations:

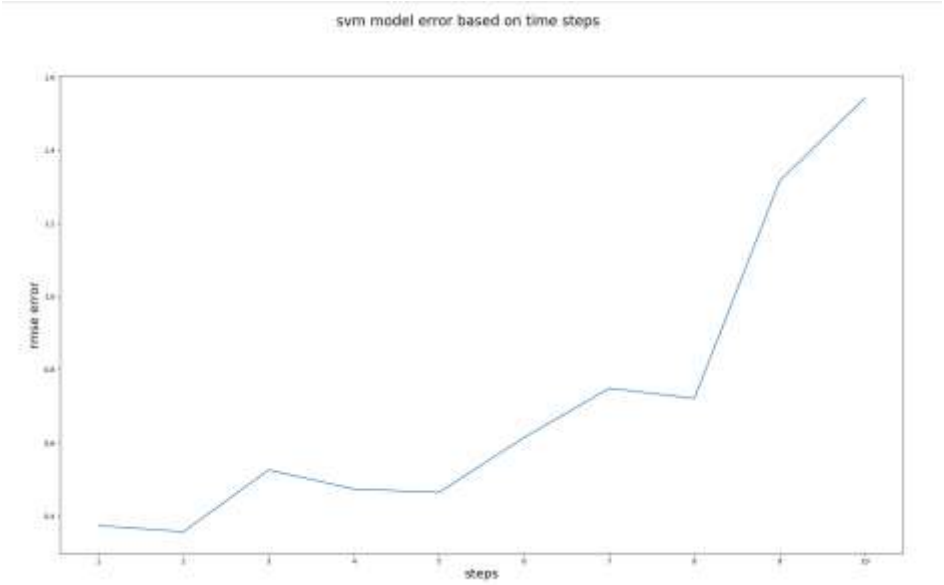
t = 63.9932

df = 18

standard error of difference = 0.000

برای همین در مدل خطی تا ۳ گام قبل را در نظر خواهیم گرفت.

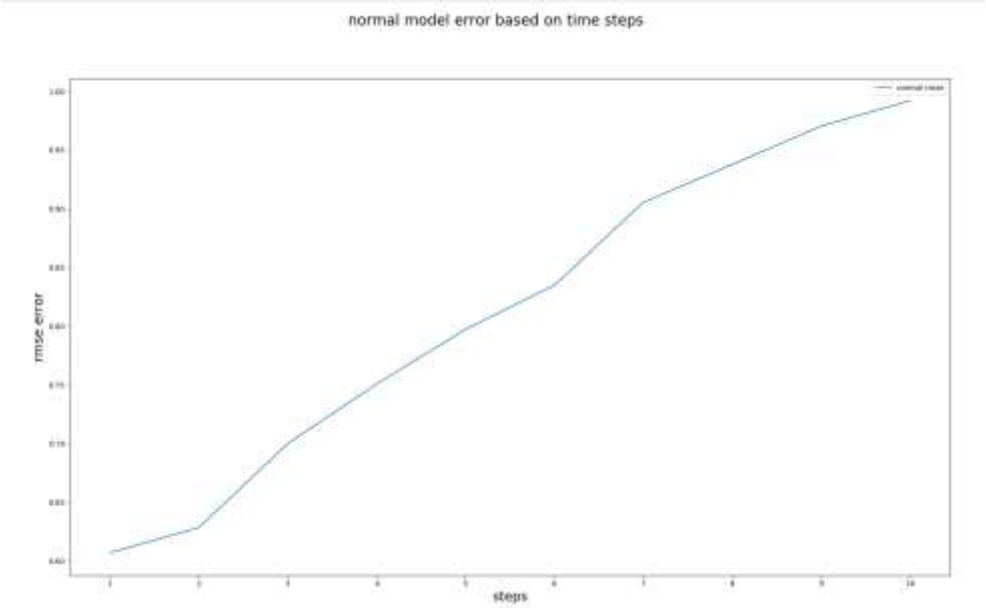
- **نمودار خطای مدل svm بر حسب تعداد گام در نظر گرفته شده:**



این نمودار نیز همانند نمودار مدل قبل کشیده شده. برای این مدل مشاهده می شود که با افزایش تعداد گام خطا نیز افزایش می یابد.

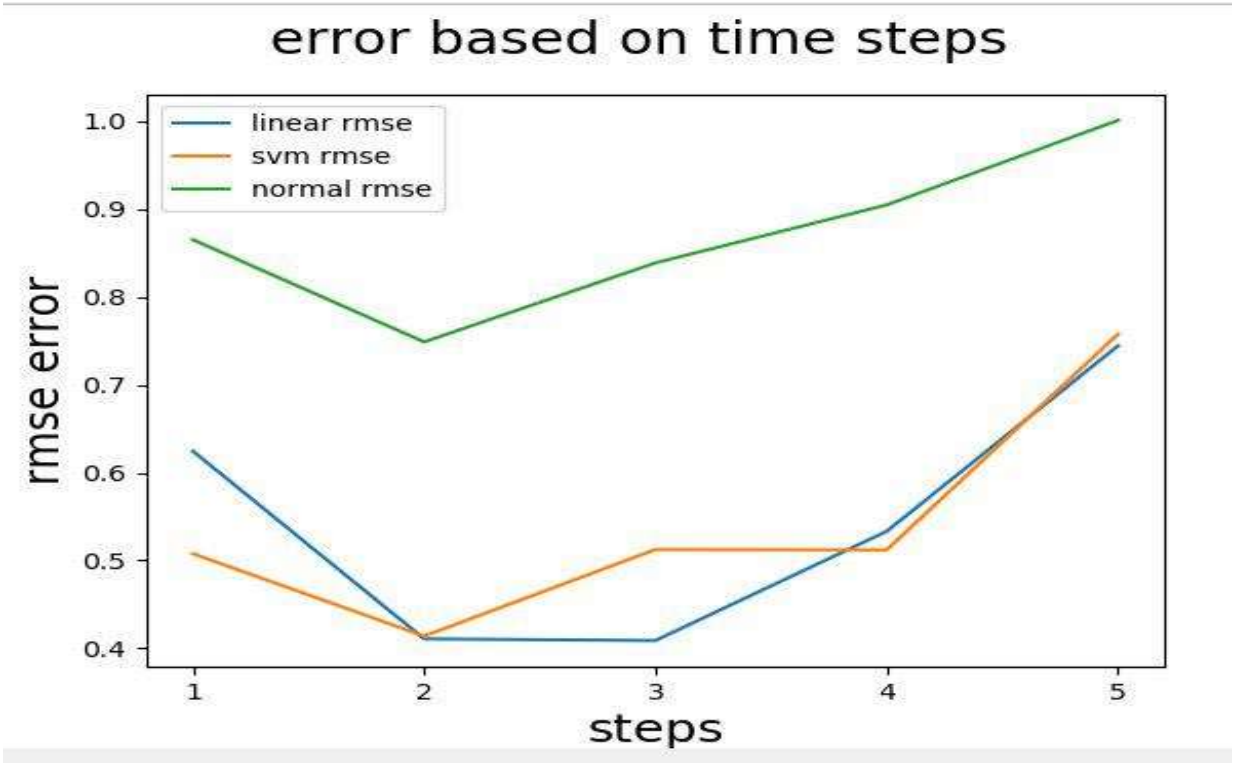
- **نمودار خطای مدل نرمال بر حسب تعداد گام در نظر گرفته شده:**

برای مدل نرمال هم این ازمایش انجام شد. به این معنی که مقدار قیمت در هر گام برابر با میانگین k گام قبلی قرار می گرفت (که k تعداد گام است که مورد ازمایش قرار گرفته است).



برای این مدل نیز مشاهده می شود که با افزایش تعدا گام خطا افزایش می یابد.

همان برای داده B_ticker:



به طور مشابه می توان میزان خطا بر حسب تعداد گام برای مدل های مختلف را روی همه ارز ها امتحان کرد و بهترین تعداد گام برای هر مدل را پیدا کرد.

بخش ششم :

بخش نهایی :

نتیجه گیری:

برای تخمین زدن قیمت طلا کارهای مختلیف انجام شد و ۳ مدل خطی، svm و نرمال امتحان شد. همین طور این موضوع که در نظر گرفتن چند گام قبل خطا را کمتر می کند نیز بررسی شد. تاثیر هموار سازی داده ها در خطا نیز بررسی شد و مشاهده شد که اثر گذار نیست.

مدل پیش بینی قیمت به صورت کلی به حالت زیر در امد (برای هر کدام از ارز ها به صورت جداگانه):

1. **در نظر گرفتن هر ۳ مدل خطی، svm و نرمال و یافتن بهترین تعداد گام برای هر کدام از این مدل ها به صورت اتومات**
2. **اجرای هر کدام از مدل ها با بهترین تعداد گام محاسبه شده مربوط به خودشان از مرحله قبل و محاسبه خطا**
3. **انتخاب مدل با خطای کمتر**
4. **پیش بینی قیمت بر اساس ان مدل و تعداد گام مربوطه**

با توجه به نکات بالا مقادیر پیش بینی شده برای کالا ها از A تا I برای ۱۰ گام بعدی قیمت تخمین زده شده است :

کالای A :

خطای تست : 0.355531722749

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 94.11330435
2. 94.22748863
3. 94.72328569
4. 94.20705679
5. 94.3318566
6. 94.56895323
7. 94.2750855
8. 94.37669578
9. 94.4847405
10. 94.32096423

کالای B :

خطای تست : 0.385677212873

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 91.99583115
2. 91.9957573
3. 91.99557975
4. 91.99536606
5. 91.99461913
6. 91.99159703
7. 91.99141735
8. 91.99108489
9. 91.99063723
10. 91.98954211

کالای C :

خطای تست : 0.307129149305

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 63.95202193
2. 63.90414063
3. 63.85635591
4. 63.80866757
5. 63.76107542
6. 63.71357926
7. 63.6661789
8. 63.61887415
9. 63.57166482
10. 63.52455071

کالای D :

خطای تست : 0.366080843065

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 69.98619524
2. 69.98566894
3. 69.98571668
4. 69.98523875
5. 69.98268256
6. 69.9721605
7. 69.97136999
8. 69.97125073
9. 69.97005822
10. 69.96615221

کالای E :

خطای تست : 0.29702542833

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 64.973988
2. 64.97352647
3. 64.9731298
4. 64.97284244
5. 64.97060446
6. 64.9478935
7. 64.9470188
8. 64.94627104
9. 64.94556487
10 64.94164542

کالای F :

خطای تست : 0.316292574675

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 58.98449908
2. 58.98435278
3. 58.98431369
4. 58.98375823
5. 58.98275289
6. 58.96899093
7. 58.96870652
8. 58.96856409
9. 58.96751267
10. 58.96571788

کالای G :

خطای تست : 0.285297247099

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 60.97667288
2. 60.95338363
3. 60.93013217
4. 60.90691846
5. 60.88374243
6. 60.86060401
7. 60.83750315
8. 60.81443978
9. 60.79141385
10. 60.7684253

کالایH :

خطای تست : 0.325150961861

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 63.98086664
2. 63.97880619
3. 63.97761769
4. 63.96130945
5. 63.95776197
6. 63.95555746
7. 63.9414384
8. 63.93682591
9. 63.93376364
10 63.92133778

کالایI :

خطای تست : 0.330926023627

مقادیر پیش بینی شده به ترتیب برای ۱۰ گام زمانی بعدی :

1. 58.93692796
2. 58.8740303
3. 58.81130655
4. 58.74875623
5. 58.68637886
6. 58.62417395
7. 58.56214104
8. 58.50027964
9. 58.43858929
10. 58.3770695