باسمه تعالی

استفاده از شبکه‌های حافظه کوتاه مدت بلند برای حل مسئله پیش‌بینی سری‌های زمانی

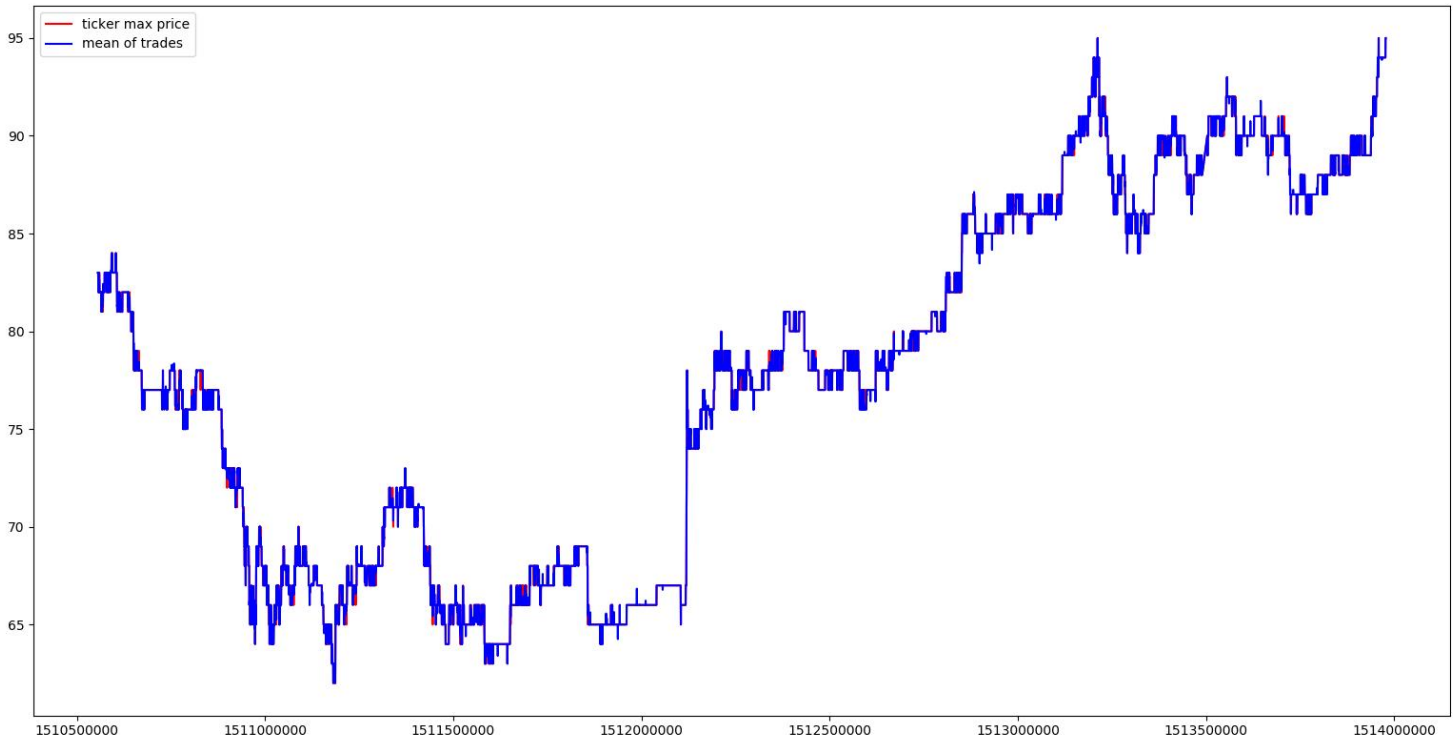
مسئله پیشی‌بینی سری‌های زمانی یکی از مسائل مهم به شمار می‌آید که در زمینه‌هایی از قبیل پیش‌بینی جریان رودخانه، دما، بارش، آب و هوا، میزان استفاده از برق یا انرژی و بورس یا قیمت ارز کاربرد دارد. هدف از این مسئله استفاده از داده‌های موجود و داده‌های پیشین برای پیش‌بینی تغییرات این داده‌ها در آینده است. این کار از طریق یافتن ارتباطاتی و روابطی که داده‌ها در زمان‌های مختلف با یکدیگر دارند انجام می‌شود.

یک سری زمانی می‌تواند ارتباط خطی یا غیر خطی با داده‌های پیشین داشته باشد. ویژگی اصلی سری‌های زمانی همبستگی خودکار بین داده‌های فعلی با داده‌های موجود در زمان‌های گذشته است. تعداد قدم‌هایی که به عقب برمیگردیم تا از داده‌های آن زمان‌ها استفاده کنیم به تاخیرزمانی معروف است. به عنوان مثال اگر داده‌ی لحظه پیشین مهم‌ترین داده برای پیش‌بینی لحظه بعد باشد تاخیر زمانی که استفاده شده یک می‌باشد. ممکن است فقط از تاخیرهای زمانی داده‌ای که میخواهیم پیش‌بینی کنیم استفاده شود یا اینکه از داده‌های دیگری نیز علاوه بر آن استفاده شود. برای مثال برای پیش‌بینی قیمت یک فلز می‌توان فقط از قیمت‌های آن در روزهای گذشته استفاده کرد یا اینکه علاوه بر قیمت از ویژگی‌های دیگری مثل میزان فروش هم بهره جست. نحوه انتخاب داده‌ها به تحلیلی که از هم بستگی بین داده‌ی مورد پیش‌بینی و داده‌های پیشین هست و هم‌چنین سعی و خطا به دست می‌آید. چون هم بستگی خطی را می‌توان مشخص کرد اما همبستگی غیر خطی را خیر.

استفاده از شبکه‌های عصبی برای پیش‌بینی سری‌های زمانی بسیار متداول بوده و در بسیاری از مسائل نتایج خوبی را به دست آورده است. در این میان شبکه‌های حافظه‌ی کوتاه مدت بلند از پتانسیل زیادی در حل مسائلی که بعد زمان در آن‌ها مطرح است دارند. به همین خاطر برای حل این مسئله یک شبکه‌ی حافظه کوتاه مدت بلند استفاده شده است.

بررسی داده‌ها

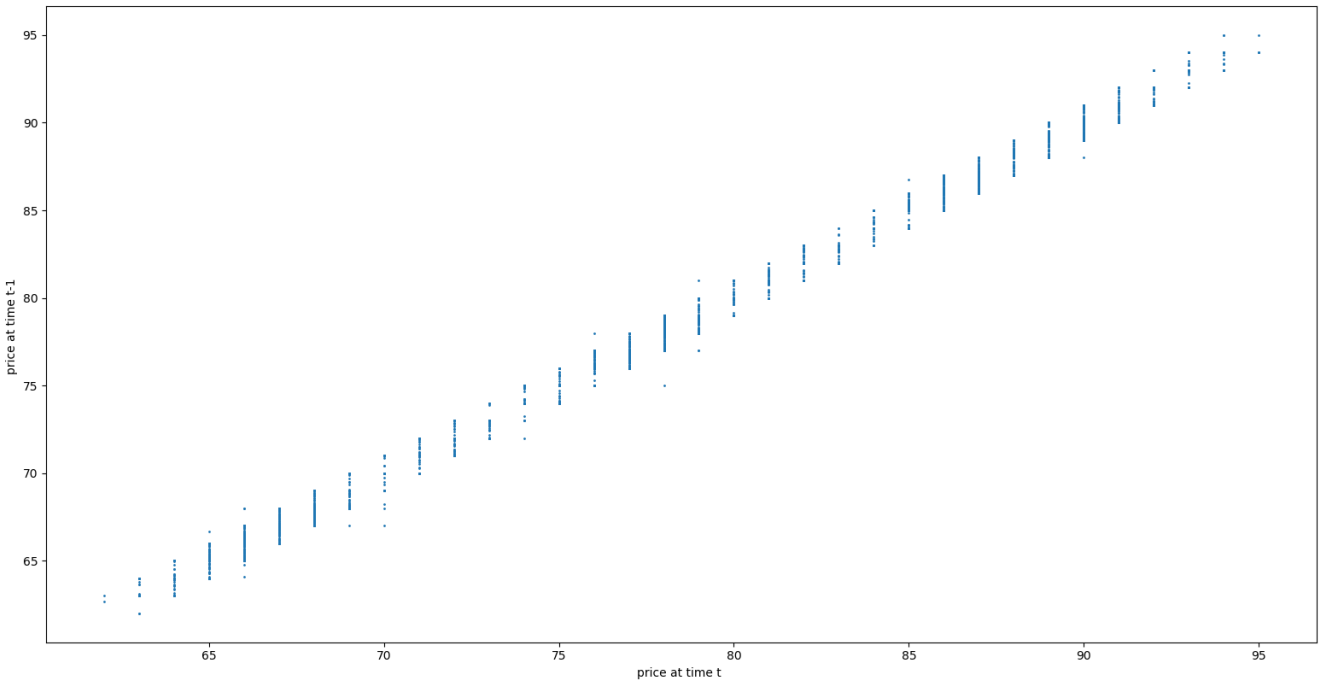
در داده‌هایی که برای هر ارز یا فلز داده شده است دو ویژگی اصلی مشاهده می‌شود یکی قیمت فلز و دومی میزان خرید و فروش. در فایل معاملات مربوط به هر کالا اطلاعات ۱۲۰ معامله صورت گرفته در هر پنج دقیقه موجود می‌باشد. به نظر می‌آید میانگین قیمت‌های این معاملات شکل خوبی از قیمت آن پنج دقیقه را نشان دهد. چرا که واریانس تغییرات قیمت در هر پنج دقیقه بسیار کم است. اما با مقایسه قیمت کالاها در فایل ticker با میانگین قیمت در فایل trade مشخص میشود که تفاوت چندانی بین آن ها وجود ندارد (تصویر ۱).



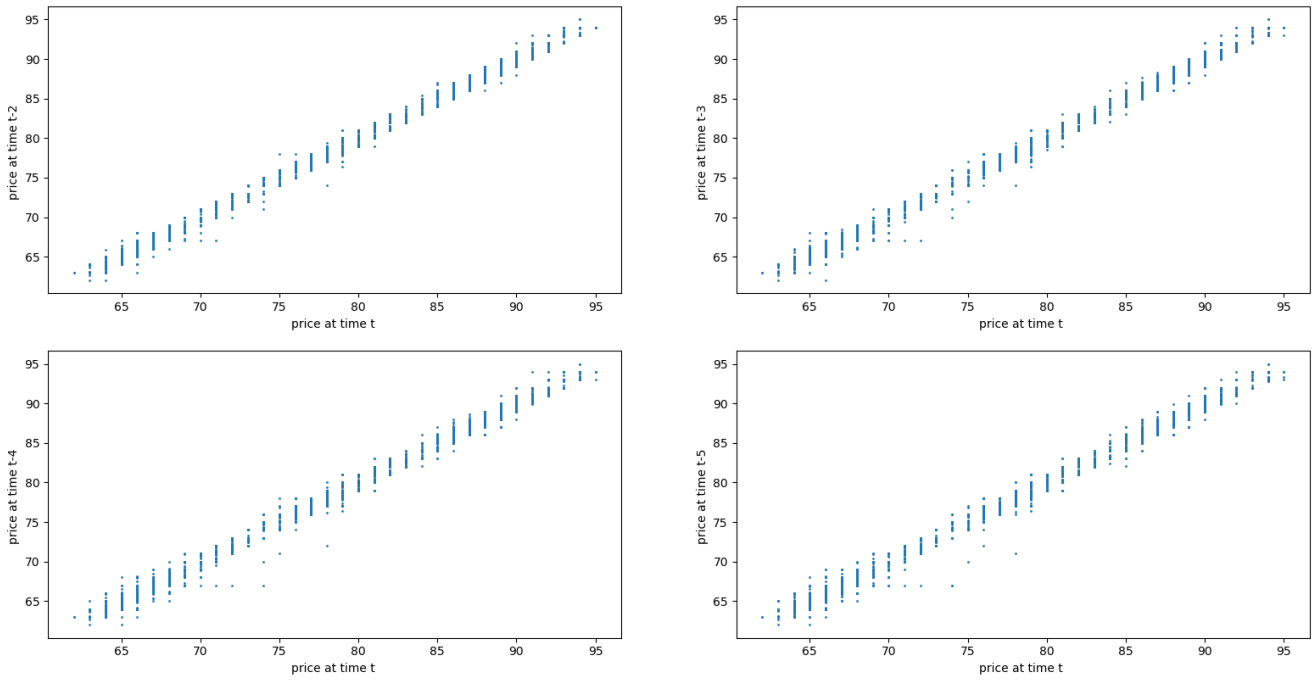
تصویر ۱ مقایسه تغییرات حداکثر قیمت در فایل ticker و میانگین قیمت در فایل trades (داده‌های نویز حذف شده‌اند)

بنابراین میتوان ستون‌های مربوط به قیمت در فایل trades را به عنوان داده زائد در نظر گرفت و از آن‌ها اجتناب کرد. البته این مورد بدون نمودار هم کاملا بدیهی بود و صرفا جهت اثبات موضوع در نمودار آورد شده است.

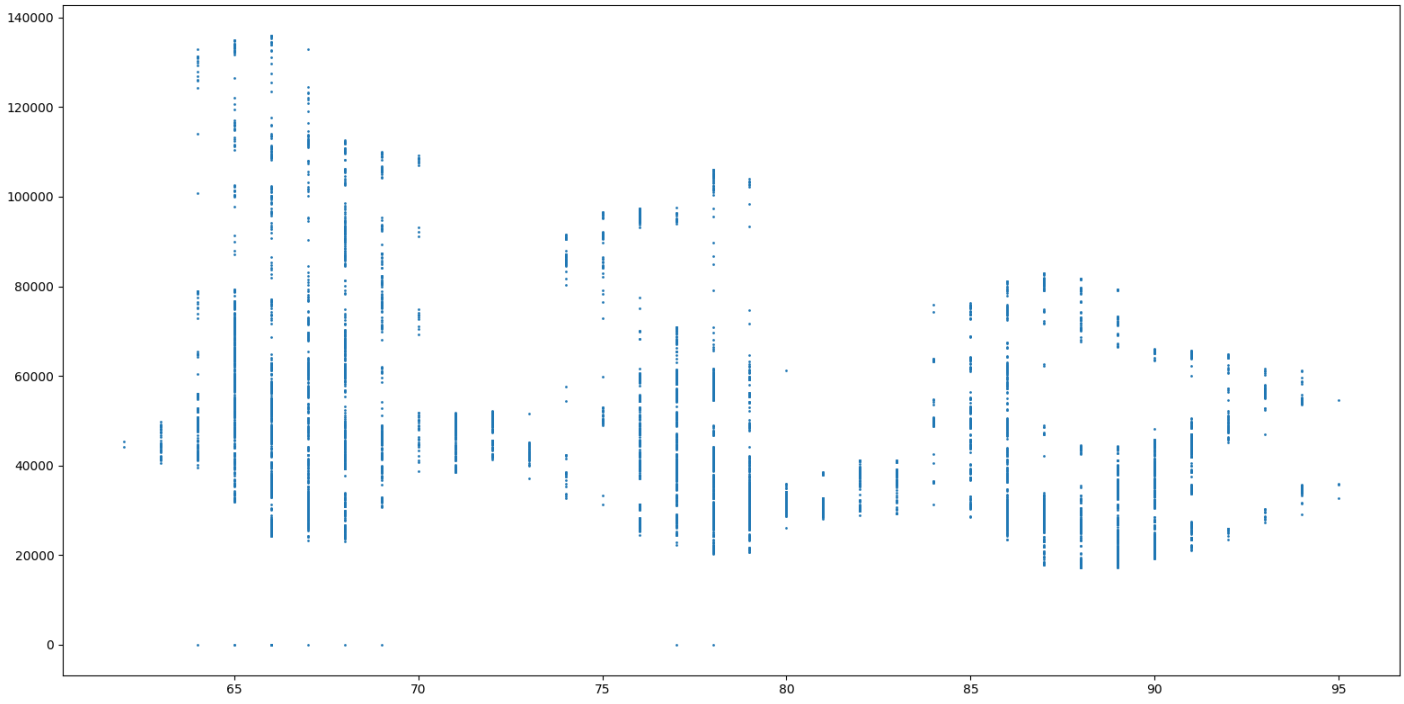
با مقایسه میزان همبستگی قیمت فعلی با قیمت روزهای قبل مشاهده می‌شود که هم بستگی مثبتی بین آن‌ها وجود دارد(تصویر ۲). این حقیقت بیانگر این است که مهمترین ویژگی برای پیش‌بینی قیمت در لحظه قبل می‌باشد. همچنین مقایسه همبستگی قیمت فعلی با قیمت لحظه‌های قبل در تصویر ۳ آمده است.
همانطور که مشاهده می‌شود با لحظه‌های قبل‌تر هم همبستگی خوبی وجود دارد و باید با سعی و خطا در آموزش مدل بهترین تعداد انتخاب شود. میزان هم بستگی قیمت لحظه t با داده‌های دیگر در لحظه t-1 در آمده است.



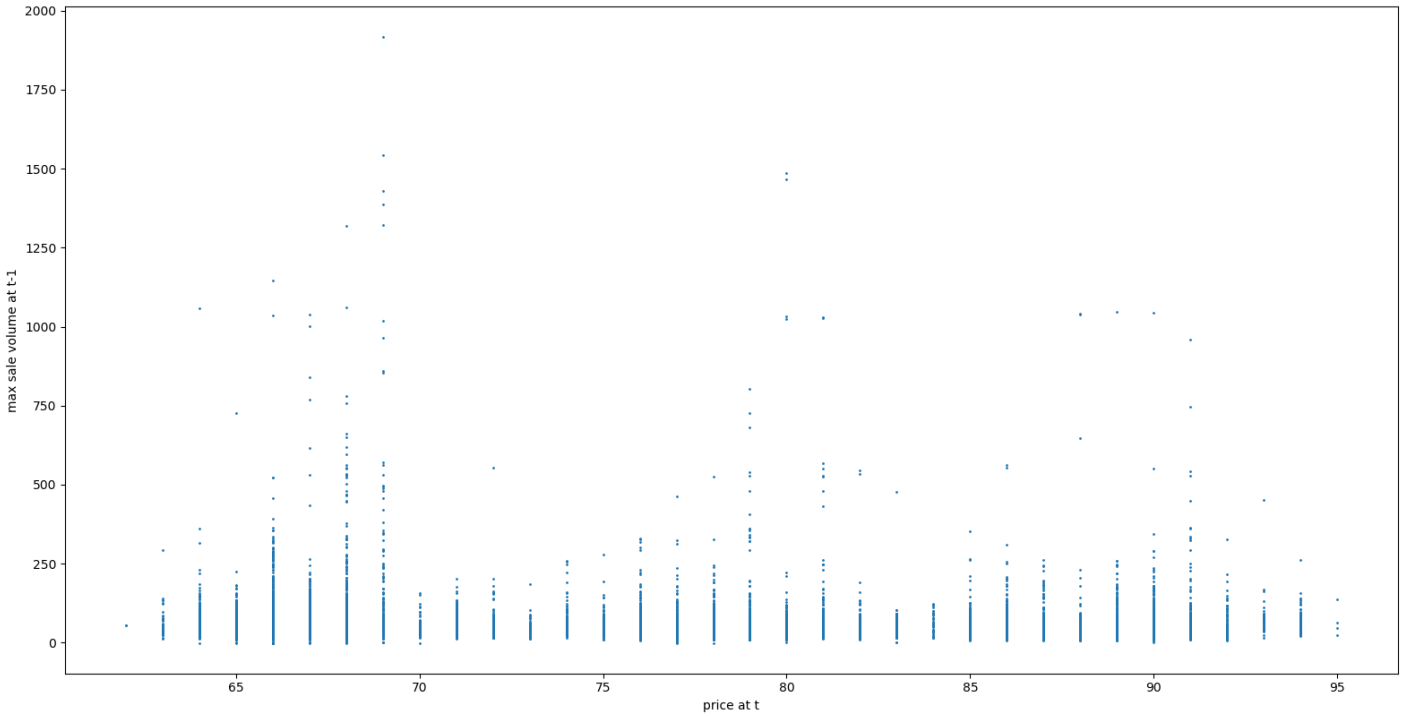
تصویر ۲ بررسی همبستگی قیمت در لحظه t با لحظه t-1



تصویر ۳ بررسی همبستگی قیمت در لحظه t با لحظه‌ها قبل



تصویر ٤ بررسی همبستگی قیمت در لحظه t با حجم معاملات روزانه در لحظه t-1



تصویر ٥ بررسی همبستگی قیمت در لحظه t با حداکثر میزان فروش در لحظه t-1

مقایسه هم بستگی داده‌ها به صورت عددی در جدول ۱ آمده است.

حداکثر حجم خرید در لحظه t-1	حجم معاملات روزانه در لحظه t-1	قیمت در لحظه t-2	قیمت در لحظه t-1	
-۰/۰۹	-۰/۳۵	۰/۹۹	۰/۹۹	قیمت در لحظه t

جدول ۱ بررسی همبستگی قیمت در لحظه t با داده‌های دیگر

همانطور که مشاهده می‌شود حداکثر حجم خرید و حجم معلاملات در لحظه t-1 همبستگی کمی با قیمت دارند. ذکر این نکته هم حائز اهمیت است که با اینکه قیمت‌ها هر پنج دقیقه جمع آوری شده‌اند ولی حجم معاملات روزانه برای ۵ دقیقه‌هایی که در یک روز هستند متفاوت است. به همین دلیل این ویژگی هم به نظر نامناسب می‌آید و درست نمی‌باشد.

با بررسی اطلاعات دیگر نیز به این نتیجه میرسیم که بسیاری از آن‌ها زائد هستند و تنها اطلاعات مفید قیمت‌های روزهای گذشته می‌باشد. هم چنین از فایل book به دلیل توضیحات مبهم استفاده نشد.

البته این تحلیل ارتباط قیمت فعلی با داده‌های قبلی را به صورت خطی بررسی می‌کند. مزیتی شبکه‌های عصبی دارند این است که می‌توانند ارتباطات غیر خطی را هم پیدا کند و باید ویژگی‌های مختلف را در مرحله آموزش به مدل داده شوند تا در صورت وجود ارتباط غیرخطی آن‌ها را پیدا کند.

پیاده سازی

برای پیاده سازی از زبان برنامه نویسی پایتون استفاده شده و مدل شبکه عصبی با استفاده از کتابخانه معروف keras[1] طراحی شده است. مدل طراحی شده یک شبکه LSTM[2] است که ۱۰ سلول LSTM دارد که به یک نود خروجی وصل شده‌اند. تابع فعالیتی که در نود خروجی استفاده شده تابع elu[3] و همچنین تابع خطای که استفاده شد تابع MSE[4] می‌باشد. هم چنین از بهینه‌ساز RMSprop[5] برای آموزش استفاده شد.

[1] Keras.io
[2] Long Short Term Memory
[3] https://arxiv.org/abs/1511.07289
[4] Mean Square Error
[5] http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

معیارهای ارزیابی

برای ارزیابی مدل‌ها از چهار معیار استفاده شد:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(z_t - T_t)^2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|z_t - T_t|$$

$$\mathrm{RMAE} = \frac{\mathrm{MAE}}{\bar{T}}$$

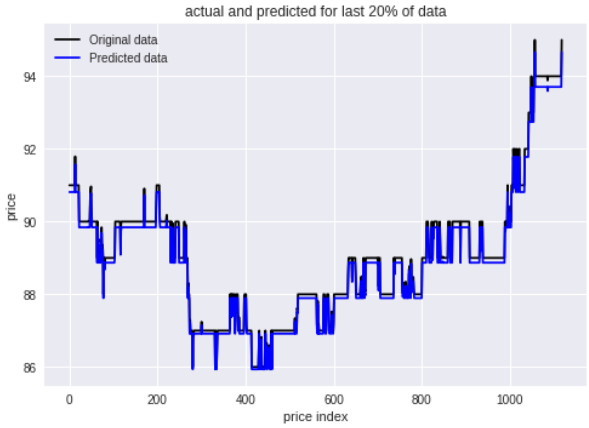
$$\mathrm{PI} = 1 - \frac{\sum_{t=1}^{n}(z_t - T_t)^2}{\sum_{t=1}^{n}(T_t - T_{t-1})^2}.$$

$$R^2 = \frac{n(\sum zT) - (\sum z)(\sum T)}{\sqrt{[n\sum z^2 - (\sum z)^2][n\sum T^2 - (\sum T)^2]}}$$

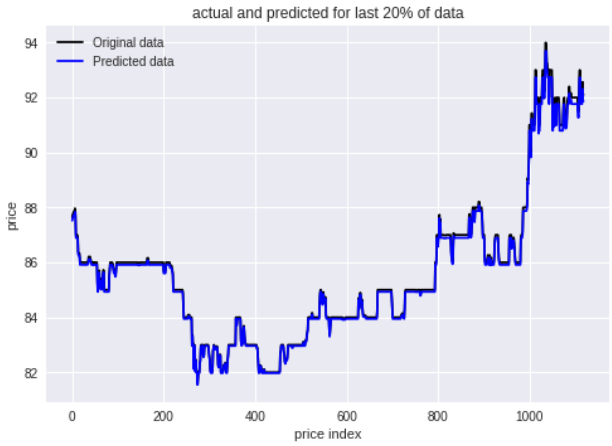
که در آن z مقدار پیش‌بینی شده و T مقدار واقعی می‌باشد. هر میزان RMSE، MAE و RMAE کمتر باشد و PI و R^2 بیشتر باشد مدل بهتر است.

نتایج تجربی

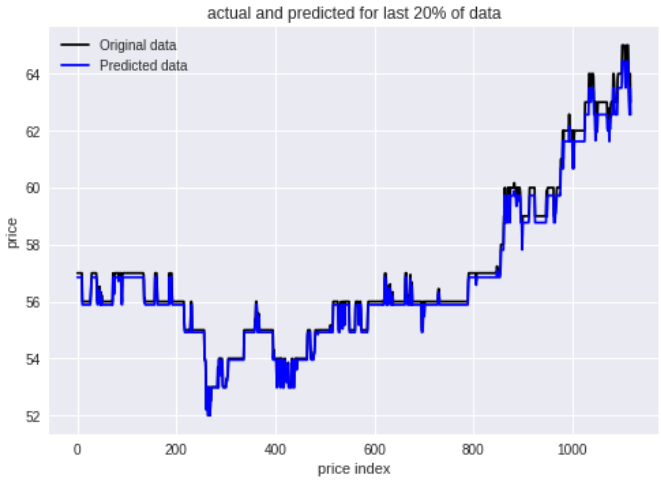
برای هر کالا یک مدل جداگانه آموزش داده شد. ۸۰ درصد اول داده‌ها در فاز آموزش و ۲۰ درصد بقیه برای فاز تست استفاده شد. داده‌ها برای استفاده در شبکه بین صفر و یک مقیاس بندی شدند.البته مقیاس‌دهنده فقط بر روی داده‌های تست آموزش داده شد تا هیچ دیدی نسبت به داده‌های تست موجود نباشد. هر مدل ۱۰۰۰ دور آموزش داده شد. نمودار خطای آموزش و ارزیابی و همچنین نمودار مقدار پیش بینی شده با مقدار واقعی برای هر کالا به ترتیب قابل مشاهده است. هم چنین میزان خطا و امتیازهای معیار مختلف آورده شده است.



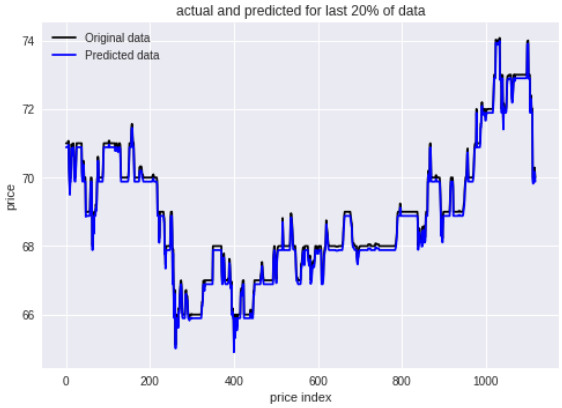
تصویر ٦ مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای A



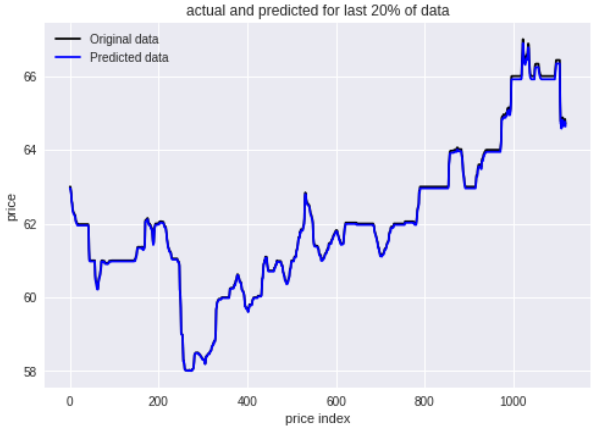
تصویر ۷مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای B



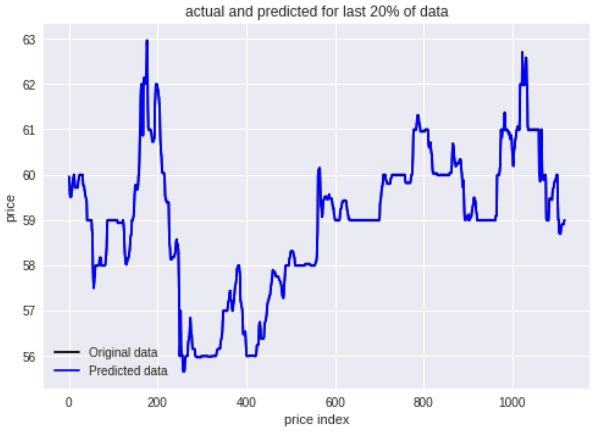
تصویر ۸ مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای C



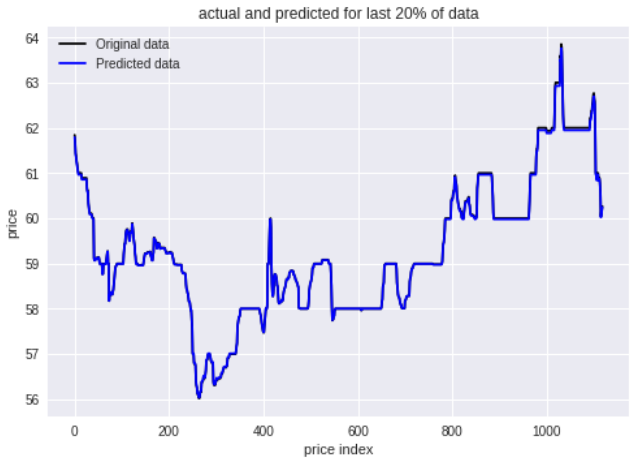
تصویر ۹مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای D



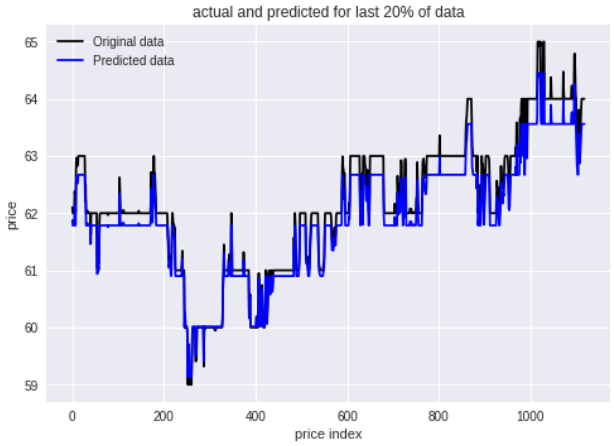
تصویر ۱۰مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای E



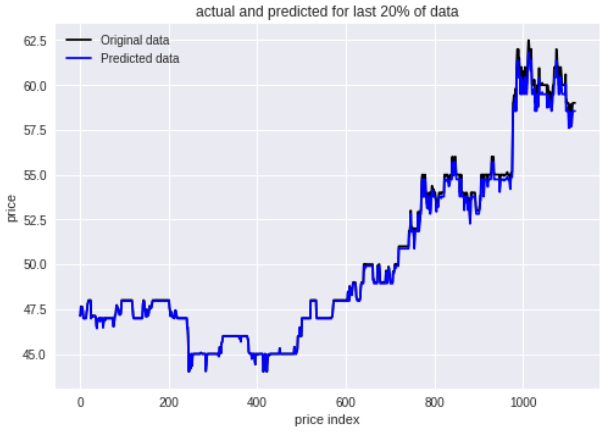
تصویر ۱۱مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای F



تصویر ۱۲مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای G



تصویر ۱۳مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای H



تصویر ۱٤مقایسه نمودار واقعی با نمودار پیش‌بینی برای کالای I

	rmse	mae	rmae	PI	R^2
A	۰/۱۴۶	۰/۱۳۶	۰/۳۶۹	۰/۹۹۴	۰/۷۵۹
B	۰/۱۰۴	۰/۰۸۷	۰/۲۹۵	۰/۹۹۹	۰/۷۶۸
C	۰/۱۹۸	۰/۱۵۷	۰/۳۹۶	۰/۹۹۵	۰/۵۳۵
D	۰/۱۱۵	۰/۱۱۵	۰/۳۳۹	۰/۹۹۶	۰/۷۶۵
E	۰/۰۴۳	۰/۰۳۷	۰/۱۹۳	۱	۰/۷۷۵
F	۰/۰۱۵	۰/۰۱۴	۰/۱۱۹	۱	۰/۹۹۱
G	۰/۰۲۴	۰/۰۱۷	۰/۱۳	۱	۰/۹۵۱
H	۰/۲۶۶	۰/۲۳۴	۰/۴۸۴	۰/۹۵۰	-۰/۷۳۵
I	۰/۲۱	۰/۱۳	۰/۳۷	۰/۹۹۸	۰/۴۹۵

جدول ۲میزان خطا و امتیاز بر اساس معیارهای مختلف برای همه کالاها

توضیحات فایل های اجرایی

فایل اصلی برای اجرا فایل main.py می‌باشد که ابتدا تعداد خط ها را میگیرد سپس در یک خط ورودی‌ها را میگیرد و در خط بعدی مقدار پیش بینی شده برای همه مدل ها را چاپ می‌کند. این کار به تعداد مرتبه تعداد خطوطی که ابتدا دریافت کرده انجام میشود.