بسمه تعالی

داده کاوی – پیش‌بینی قیمت ده ارز و فلز گرانبها

فهرست مطالب

۱- مقدمه و روش‌های مورد استفاده

برای پیش‌بینی قیمت در بازار های مختلف روش‌های متعددی وجود دارد. استفاده از هر یک از این روش‌ها بسته به موقعیت های مختلف و با توجه به داده‌ها متفاوت می‌باشد، به عنوان مثال در بسیاری موارد به اینکه قیمت افزایش پیدا میکند یا کاهش، بسنده میکنیم، که در این موارد می‌توان از روش‌های کلاس بندی دودویی (binary classification) استفاده نمود.

در موقعیت متفاوت دیگری نیاز داریم تا با توجه به داده‌های از قبل موجود قیمت دقیق یک جنس در بازار را پیش‌بینی و بررسی نماییم، که در این موارد روش‌های رگرسیون می‌توانند بسیار مفید باشند.

اما گونه دیگری از مسایل نیز وجود دارند که علاوه بر موارد گفته شده مسیله زمان نیز در آن‌ها مطرح است. این دسته از مسایل که به سری های زمانی معروف هستند به پیش‌بینی قیمت دقیق یک کالا در آینده و با توجه به داده‌های موجود می‌پردازد، ایده اصلی برای پیش‌بینی قیمت در چنین مسایلی استفاده از روش‌های آمار و شبکه عصبی است. به عنوان مثال می‌توان روش‌های Additive model و ARIMA (مخفف Auto-Regressive Integrated Moving Averages) و LSTM (یک روش بر مبنای شبکه‌ها عصبی) را برای حل اینگونه مسایل نام برد.

در این تحقیق من نیز از روش Additive model برای پیش‌بینی استفاده نمودم که در لینک زیر مبنای آن توضیح داده شده است.

https://en.wikipedia.org/wiki/Additive_model

همچنین این روش نیاز به پیش پردازش خاص خود دارد و یا استفاده از ابزار ها و کتابخانه‌ها مختص خود که آن‌ها را در هر بخش مرتبط آن توضیح می‌دهم.

تذکر: در هر مرحله تصاویری از کد پایتون مربوطه برای درک بهتر قرار داده شده است.

۲- توضیح ابزار ها و محیط اجرا

در این تحقیق زبان برنامه نویسی پایتون مورد استفاده قرار گرفته است. همچنین کتابخانه‌های زیر نیز باید بر روی سیستم عامل مربوطه نصب باشد. نام هر کدام و مورد کاربرد آن نیز بیان شده است.

Pandas library: برای دستکاری داده‌ها، خواندن از فایل csv

matplot library:مصور سازی داده‌ها و رسم نمودار

datatime library:تبدیل تاریخ از timestamp

numpy library:برای دستکاری داده‌ها و کار با آرایه های چند بعدی

fbprophet library: برای پیاده‌سازی روش آماری Additive model

scikit-learn library:برای ارزیابی نهایی مدل

تذکر: برای اجرای فایل ارسالی باید این کتابخانه ها نصب باشند و همچنین فایل‌های داده و فایل اجرایی برنامه در یک دایرکتوری قرار داشته باشند.

۳- تحلیل داده‌ها

در این قسمت قصد تشکیل یک دیتاست برای پیش‌بینی قیمت در آینده داریم، بنابراین نیاز است که فایل‌های مختلف را برای کالا ها بررسی نماییم.

برای هر فلز یا ارز گرانبها سه فایل با نام های ticker و book و trades داده شده است. برای هر کدام در زیر بیان نموده‌ام که چه اطلاعاتی برای پیش‌بینی مهم هستند و باید به دیتاست افزوده شود.(همراه با دلیل)

فایل ticker: از این فایل همه ستون‌ها برای پیش‌بینی قیمت مهم می‌باشند به غیر از ستون ۸ و ۹ که مربوط به حداکثر و حداقل قیمت روزانه می‌باشد که این دو ستون از دیتاست حذف شده‌اند و میانگین آن‌ها به عنوان قیمت آن ارز یا فلز گرانبها مورد استفاده قرار گرفته است. این ستون همان ستونی است که در آینده باید پیش‌بینی شود.

در تصویر زیر کد پایتون مربوط به این کار را ملاحضه می نمایید.

```
# reading data
data_set = pd.read_csv( file_name + "_ticker.csv",header=None)
book_file = pd.read_csv( file_name + "_book.csv",header=None)
trades_file = pd.read_csv( file_name + "_trades.csv",header=None)
data_set['price'] = (data_set[7] + data_set[8]) / 2      # target
data_set = data_set.drop([7 , 8], axis=1)
```

فایل book: این فایل نشان دهنده سفارش ها است. یکی از ستون‌های مهم در این فایل حجم سفارش است. به دلیل اینکه این ستون حاوی مقادیر منفی نیز می‌باشد می‌توان چنین استنباط نمود که به ازای مقادیر مثبت حجم سفارش ما آن ارز یا فلز گرانبها را به بازار عرضه نموده ایم.(مانند عرضه دلار بانک مرکزی به بازار) و مقادیر منفی حجم سفارش به معنی باز پس گرفتن آن ارز یا فلز گرانبها از بازار است.

حال موردی که مطرح است مجموع حجم سفارش ها برای یک زمان خاص است. که ممکن است این مجموع حجم برای یک زمان خاص مثبت یا منفی به دست آید، برای هر کدام استنباط زیر مطرح است:

اگر حجم سفارش منفی بدست آید: یعنی ما آن ارز یا فلز گرانبها را بیش از حد به بازار عرضه نموده ایم و مشتری برای آن وجود ندارد. (عرضه زیاد بوده است و باعث کاهش قیمت می شود.)

اگر حجم سفارش ها مثبت بدست آید: یعنی ما آن ارز یا فلز گرانبها را کمتر از تقاضا به بازار عرضه نموده ایم و مشتری برای آن وجود دارد. (تقاضا زیاد بوده است و باعث افزایش قیمت می شود.)

بنابر اطلاعات بالا، باید برای هر زمان خاص حجم سفارش‌ها محاسبه شود و به دیتاست اضافه گردد.

در تصویر زیر کد پایتون مربوط به این کار را ملاحضه می نمایید.

```
# calculate important data from *_book and *_trades file
order_volume = []
for i in range(0,len(book_file)):
    sum_of_orders = 0
    for j in range(3,151,3):
        sum_of_orders = sum_of_orders + book_file.loc[i,j]
    order_volume.append(sum_of_orders)
mapping = dict(enumerate(order_volume))
data_set['order_volume'] = data_set[1].map(mapping)
```

فایل trades: این فایل نیز حاوی معاملات انجام شده است و یکی از ستون هایی که میتواند در قیمت آینده تأثیر گزار باشد، حجم معاملات انجام شده در یک زمان خاص است (به ازای هر سطر از این فایل). بنابراین می توانیم مانند فایل book به ازای هر زمان خاص (هر سطر از این فایل) مجموع حجم معاملات را محاسبه نماییم و به دیتاست اضافه کنیم.

همچنین از این فایل مجموع قیمت معاملات انجام شده برای یک زمان خاص مهم است که آنرا نیز دقیقاً به طریق بالا محاسبه می نماییم و به دیتاست اضافه می‌کنیم. کد آن در زیر آورده شده است.

```
turnover = []
for i in range(0,len(trades_file)):
    sum_of_turnover = 0
    for j in range(3,481,4):
        sum_of_turnover = sum_of_turnover + trades_file.loc[i,j]
    turnover.append(sum_of_turnover)
mapping = dict(enumerate(turnover))
data_set['turnover'] = data_set[1].map(mapping)
data_set['transaction_price'] = data_set[1].map(mapping)
```

بنابر توضیحات بالا دیتاست ما از ستون‌های زیر تشکیل شده است:

ستون ۱: همان زمان ثبت اطلاعات با رزولوشون ۵ دقیقه است.

ستون ۲: همان ستون ۲ فایل ticker است.

ستون ۳: همان ستون۳ فایل ticker است.

ستون ٤: همان ستون ٤ فایل ticker است.

ستون ٥: همان ستون ٥ فایل ticker است.

ستون ٦: همان ستون ٦ فایل ticker است.

ستون ۷: همان ستون ۷ فایل ticker است.

ستون ۸ (price): میانگین ستون‌های ۸ و ۹ فایل ticker است. (این ستون قیمت است و باید در آینده پیش‌بینی شود.)

ستون ۹(order_volume): مجموع حجم سفارش‌های داده شده برای آن زمان. (ستون اول دیتاست)

ستون ۱۰(turnover): مجموع حجم معاملات انجام شده برای آن زمان.

ستون ۱۱(price_transaction): مجموع قیمت معاملات انجام شده برای آن زمان.

در زیر چند سطر اول این دیتاست را در محیط terminal ملاحضه می نمایید.

	0	1	2	3	4	5	6	price	order_volume	turnover	transaction_price
0	1510555672	83	17	83	85	83	40782	84.5	-17	31	31
1	1510555977	83	32	83	49	83	40795	84.5	-17	31	31
2	1510556278	83	39	83	24	83	40807	84.5	-17	31	31
3	1510556577	83	21	83	26	83	40787	84.5	-17	31	31
4	1510556877	82	69	82	39	82	40847	84.5	-15	0	0
5	1510557177	82	73	82	38	82	40993	84.5	-15	0	0
6	1510557477	82	68	82	39	82	41106	84.5	-15	0	0
7	1510557777	83	74	83	38	83	41131	84.5	-17	31	31
8	1510558077	82	23	82	48	82	41181	84.5	-15	0	0
9	1510558377	83	30	83	32	83	41076	84.5	-17	31	31
10	1510558677	83	22	83	51	83	41072	84.5	-17	31	31
11	1510558977	83	31	83	62	83	40932	84.5	-17	31	31
12	1510559278	82	26	82	50	82	40902	84.5	-15	0	0
13	1510559579	83	32	83	64	83	40863	84.5	-17	31	31
14	1510559941	82	56	82	39	82	41035	84.5	-15	0	0
15	1510560177	82	37	82	23	82	41042	84.5	-15	0	0
16	1510560477	82	46	82	35	82	40299	84.5	-15	0	0
17	1510560777	82	37	82	54	82	39696	84.5	-15	0	0
18	1510561077	82	32	82	58	82	39235	84.5	-15	0	0
19	1510561378	82	18	82	82	82	39054	84.5	-15	0	0
20	1510561677	82	69	82	31	82	38413	84.5	-15	0	0
21	1510561977	82	47	82	74	82	38226	84.5	-15	0	0
22	1510562277	82	19	82	37	82	38012	84.5	-15	0	0
23	1510562577	82	34	82	66	82	37916	84.5	-15	0	0
24	1510562877	82	18	82	86	82	37841	84.5	-15	0	0
25	1510563178	82	63	82	34	82	37841	84.5	-15	0	0
26	1510563480	82	22	82	74	82	37723	84.5	-15	0	0
27	1510563777	82	19	82	76	82	37639	84.5	-15	0	0
28	1510564078	82	18	82	70	82	37561	84.5	-15	0	0
29	1510564377	82	14	82	81	82	37444	84.5	-15	0	0
...	...	..	...	..	...	..	...	...	...	...	...
11171	1513968636	94	71	94	24	94	35116	92.0	49	10	10
11172	1513968937	94	60	94	165	94	35280	92.0	49	10	10
11173	1513969237	94	44	94	128	94	34962	92.0	49	10	10
11174	1513969536	94	57	94	122	94	35004	92.0	49	10	10
11175	1513969836	94	32	94	107	94	35035	92.0	49	10	10
11176	1513970137	94	29	94	90	94	35059	92.0	49	10	10
11177	1513970437	94	85	94	171	94	34920	92.0	49	10	10
11178	1513970736	94	79	94	150	94	35019	92.0	49	10	10
11179	1513971036	94	69	94	129	94	35055	92.0	49	10	10
11180	1513971336	94	40	94	159	94	34555	92.0	49	10	10
11181	1513971636	94	39	94	71	94	34594	92.0	49	10	10

٤- پیش پردازش داده‌ها

در بخش پیش پردازش داده‌ها ابتدا نیاز است تا قالب زمان داده‌ها را از timestamp درآوریم و به current datetime تبدیل نماییم.

بحث دوم در مورد outlier ها در داده‌ها می‌باشد. مدل Additive از کتابخانه prophet تنها در صورتی که مقادیر آن‌ها را به np.pd.NaN (به صورت خیلی ساده منظور همان None در برنامه نویسی می‌باشد.) تبدیل کنیم قادر به شناسایی آن‌ها است و مانع از تغیر مقادیر پیش‌بینی در آینده می‌شود. البته در مستندات کتابخانه آمده بود که بهترین راه برای پیش پردازش آن‌ها حذف آن‌ها است، منتها در این کار outlier ها حدف نشده‌اند به دلیل اینکه در بعضی موارد تعداد آن‌ها بسیار کم بود و تأثیری نداشت اما برای بعضی از ارز ها یا فلز های گرانبها تعداد آن‌ها زیاد بود و باعث تغیر در نتیجه می‌شد که می‌توان برای بهبود مدل آن‌ها را حذف نمود.

در زیر کد مربوط به پیش پردازش داده‌ها را مشاهده می نمایید.

```
# preprocessing
# change date format
date_list = []
for i in range (0,len(data_set)):
    date_list.append(datetime.datetime.fromtimestamp(
        int(data_set[0][i])
        ).strftime('%Y-%m-%d %H:%M'))

date_sries = pd.Series( date_list , name='date')
data_set[0] = date_sries


# set NaN value for noisy data in pandas
data_set.loc[(data_set['price'] == 0) & (data_set[0] == -1), 'price'] = None
# split for train and test
test_set = data_set[round(len(data_set)*(0.9))+1:len(data_set)]
data_set = data_set[0:round(len(data_set)*(0.9))]
```

در نهایت پس از اعمال پیش پردازش های بالا، داده‌ها را به دو قسمت داده‌های آموزشی و آزمایشی تقسیم کردم. توجه نمایید که تنها ۱۰ درصد داده‌ها را برای آزمایش قرار دادم و مابقی آن‌ها را برای آموزش مدل گزاشته ام.

۵- پیش‌بینی قیمت با استفاده از روش‌های آماری

در این قسمت به توضیح مدل آماری Additive model که برای پیش‌بینی استفاده شده است می پردازم.

در ابتدا موضوعی که مطرح است فهم این مدل است که چون فرصت مناسبی برای بیان آن در این گزارش نیست و همچنین در استفاده از آن تأثیر چندانی ندارد، از بیان آن خودداری می‌کنم.

موضوع بعدی کتابخانه prophet و توابع آن است که برای استفاده از مدل آماری Additive model مورد استفاده قرار می گیرد.

دو تابع مهم آن fit و predict است. در تابع اول به عنوان آرگومان ورودی یک دیتاست را دریافت می نماید و برای آن یک مدل پیش گویی برای زمان های بعدی ایجاد می نماید. سپس با تابع predict و توسط آرگومان های ورودی آن سری های زمانی آینده را پیش‌بینی می‌کند.

موضوع بعدی از این کتابخانه بازه زمان هایی است که پیش‌بینی می‌کند، این بازه می‌تواند سالانه، ماهانه، روزانه و یا ساعتی باشد. اما قابلیت پیش‌بینی آینده در بازه دقیقه‌ای را ندارد. با توجه به بررسی هایی که در مستندات این کتابخانه داشتم در ورژن های آتی این قابلیت اضافه خواهد شد.

بنابراین در این پروژه تا ۷۲ ساعت بعد از آخرین زمانی که داده شده بود را پیش‌بینی نمودم و به جای اندازه‌گیری ۵۰ دقیقه بعدی با رزولوشن ۵ دقیقه، ۵۰ ساعت بعدی را با رزولوشن یک ساعت به یک ساعت اندازه‌گیری نمودم. البته این مورد در نسخه های بعدی این کتابخانه برطرف خواهد شد. در زیر کد مربوط به این قسمت را مشاهده می نمایید.

```
# use statistic model for forecasting
data_set = data_set.rename(columns={ 0 : 'ds', 'price' : 'y'})
A_prophet = fbprophet.Prophet()
A_prophet.fit(data_set)

A_forecast = A_prophet.make_future_dataframe(freq='H',periods=72)
A_forecast = A_prophet.predict(A_forecast)

A_prophet.plot(A_forecast, xlabel = 'Date', ylabel = 'Price')
plt.title('prediction for ' + file_name);
```

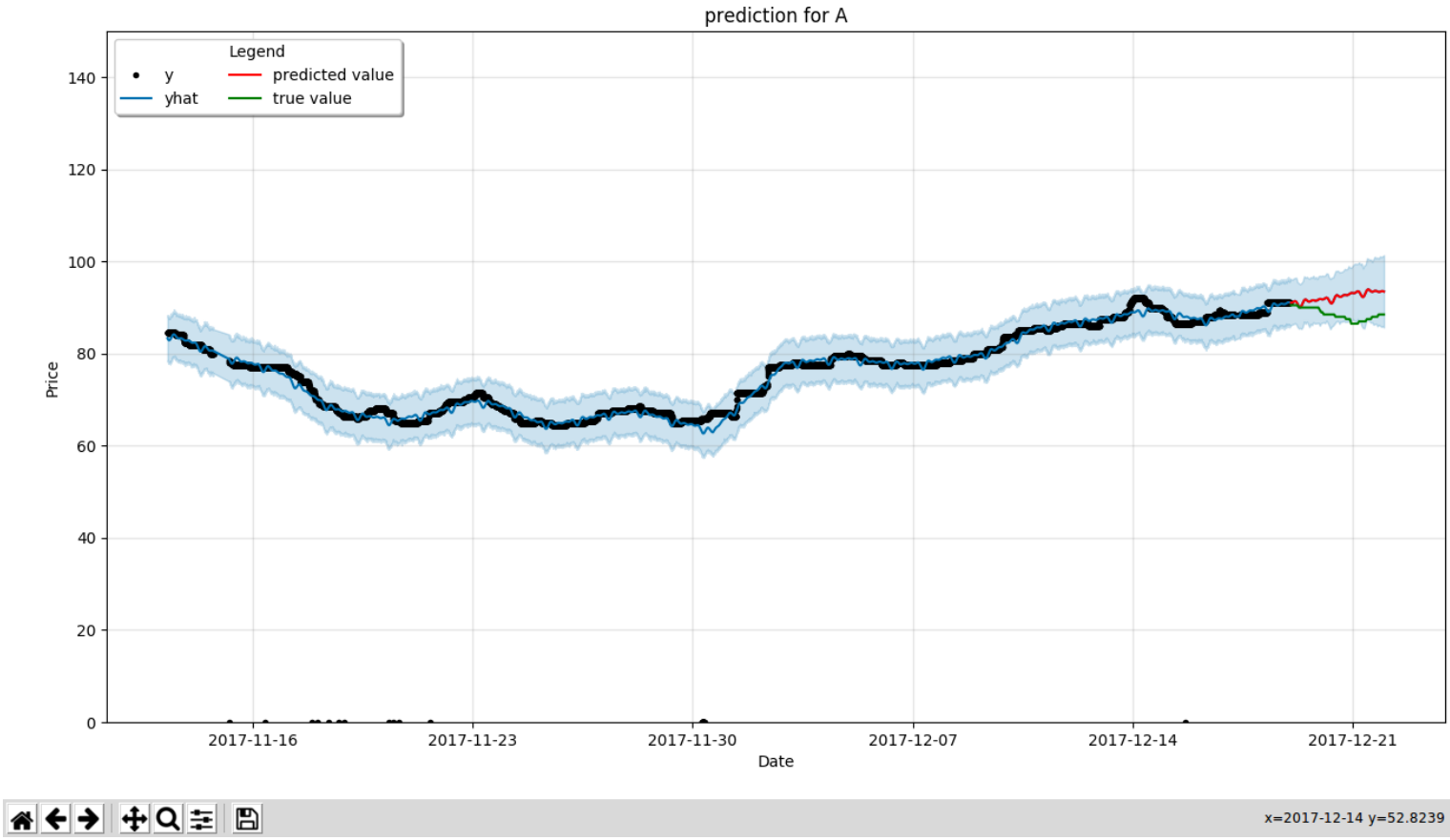
٦- ارزیابی مدل

برای ارزیابی مدل از سنجه خطای میانگین مربعات استفاده شده است که همراه پیش‌بینی قیمت برای هر ازر یا فلز گرانبها محاسبه می‌شود و در خروجی نمایش داده می‌شود.

```
# Evaluate prediction
predicted_df = A_forecast.set_index('ds').join(test_set.set_index(0))
predicted_df = predicted_df.dropna()
predicted_df = predicted_df[["yhat","price"]]
y_pred = predicted_df['yhat'].tolist()
y_true = predicted_df['price'].tolist()
print("=======================================================")
print("mean squar error for " + file_name + ":" , mean_squared_error(y_true, y_pred))
print("=======================================================\n")
input("press any key to present forecasting plot...\nand to continue close the figure.")
```

٧- مصور سازی نتایج

در قسمت مصور سازی نتایج برای هر ارز یا فلز گران بها یک نمودار رسم می‌شود. ستون عمودی مربوط به قیمت است و خط افقی مربوط به زمان (در فرمت year-mounth-day). در زیر تصویر مصور سازی نتایج را برای ارز یا فلز گرانبهای A مشاهده می کنید. هر خط داخل این نمودار نشان دهنده مطلبی است که در زیر عکس به آنها اشاره خواهم کرد.



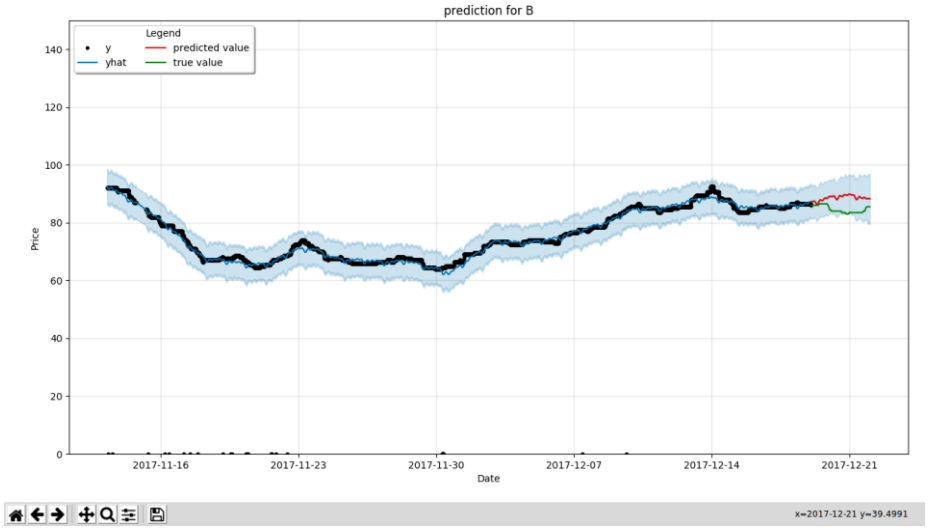
Y: نشان دهنده همان مقادیر واقعی قیمت در آن زمان است که درواقع به عنوان داده‌های آموزشی ما آن‌ها را با الگوریتم داده‌ایم. (رنگ مشکی داخل تصویر)

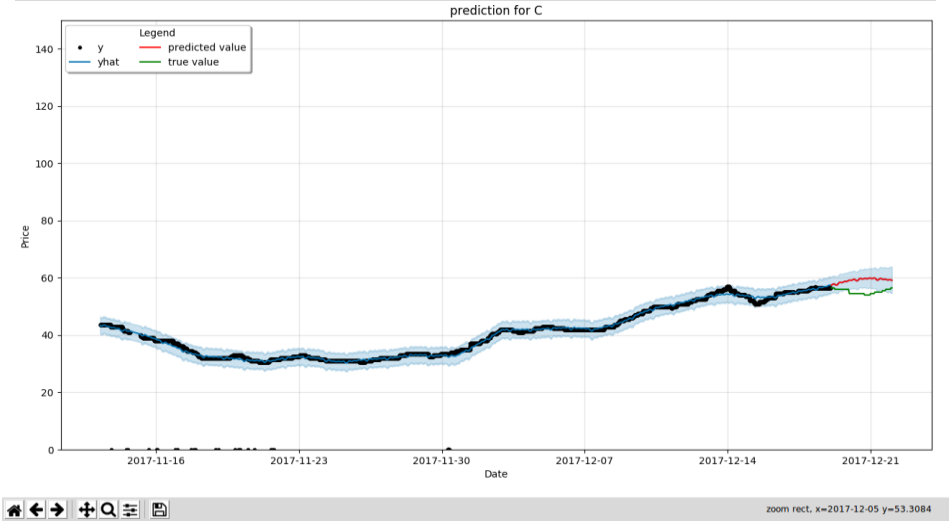
yhat: نشان دهنده همان مقادیر پیش‌بینی شده قیمت در آن زمان است. (رنگ آبی داخل تصویر)

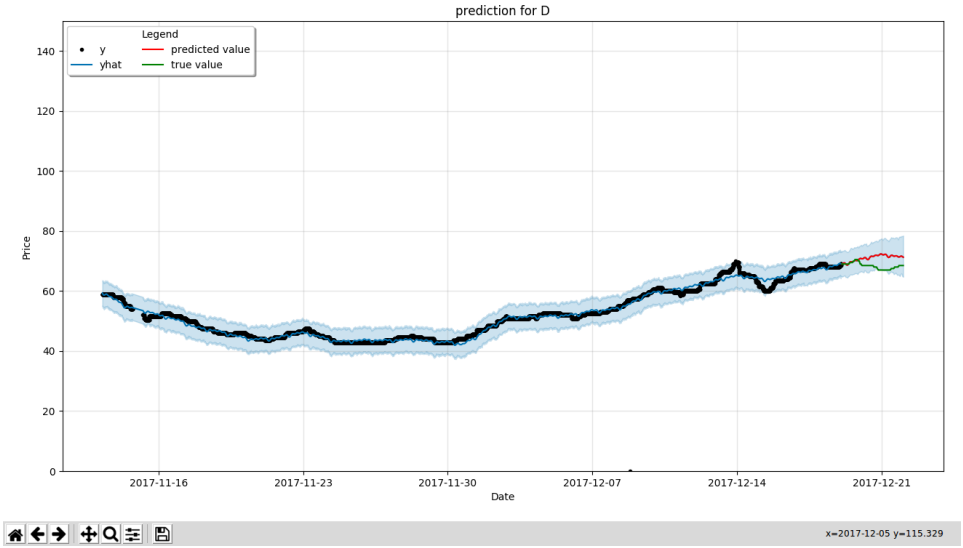
true vlaue: این خط نشان دهنده مقادیر واقعی برای زمان های آینده است، این خط با توجه به داده‌های تست رسم شده است. (رنگ سبز)

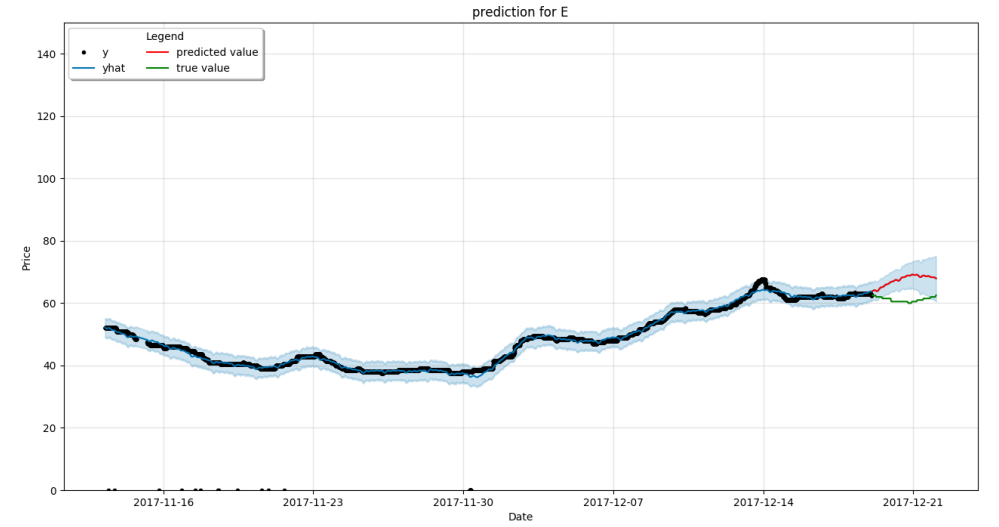
predicted value: این خط نشان دهنده مقادیر پیش‌بینی شده برای زمان های آینده است. (رنگ قرمز)

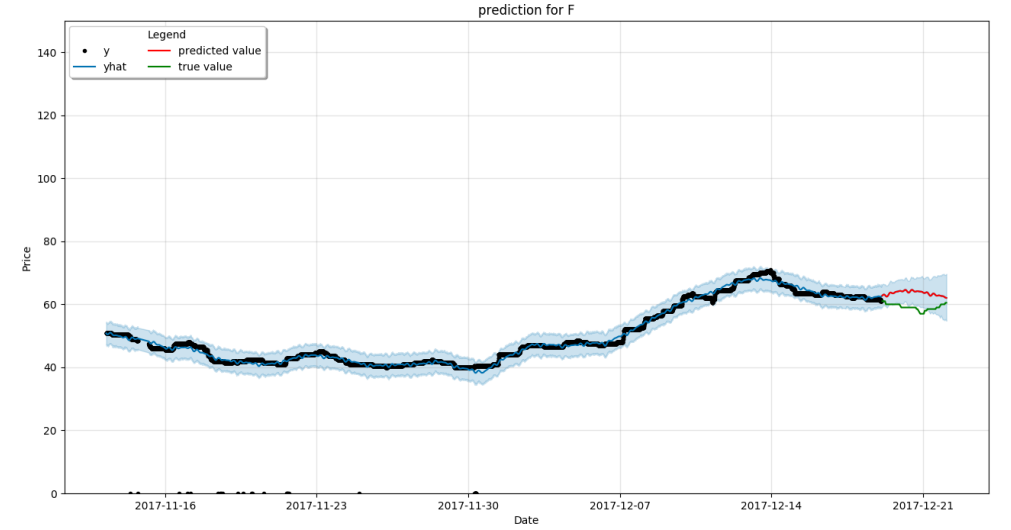
همچنین در زیر نمودار مقادیر پیش‌بینی شده برای کالاهای A تا I محاسبه و رسم شده است.

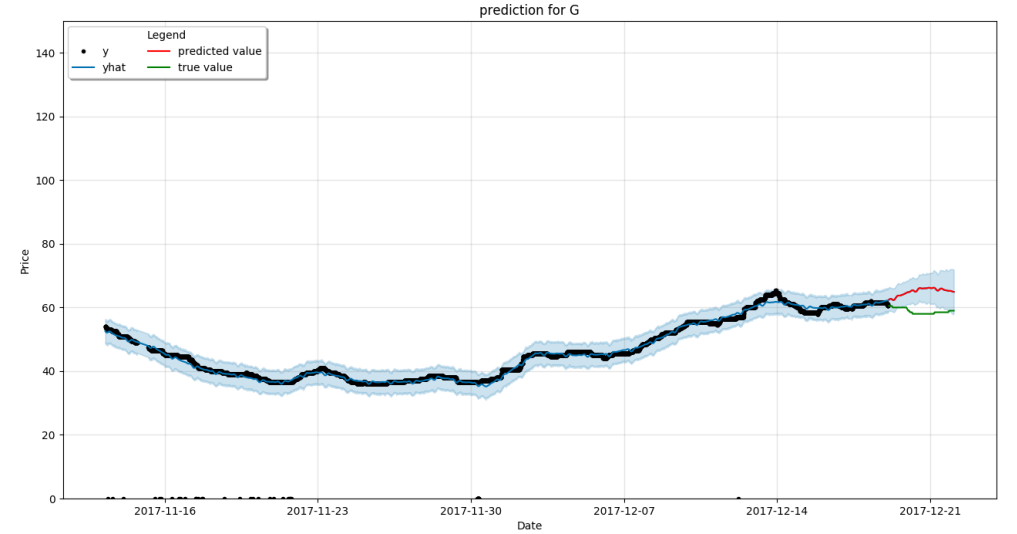

prediction for B
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
x=2017-12-21 y=39.4991

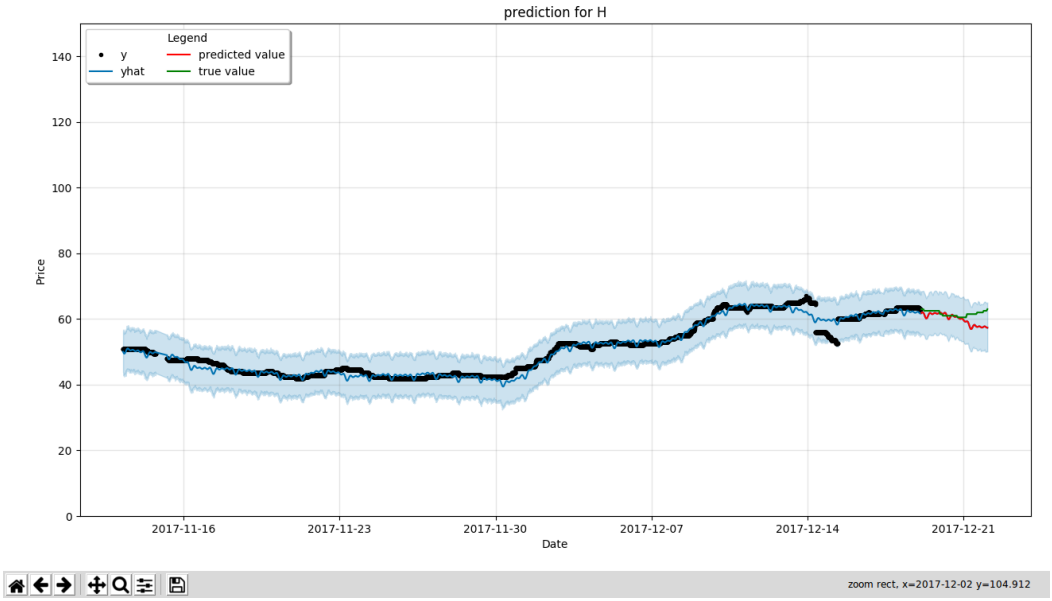

prediction for C
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
zoom rect, x=2017-12-05 y=53.3084

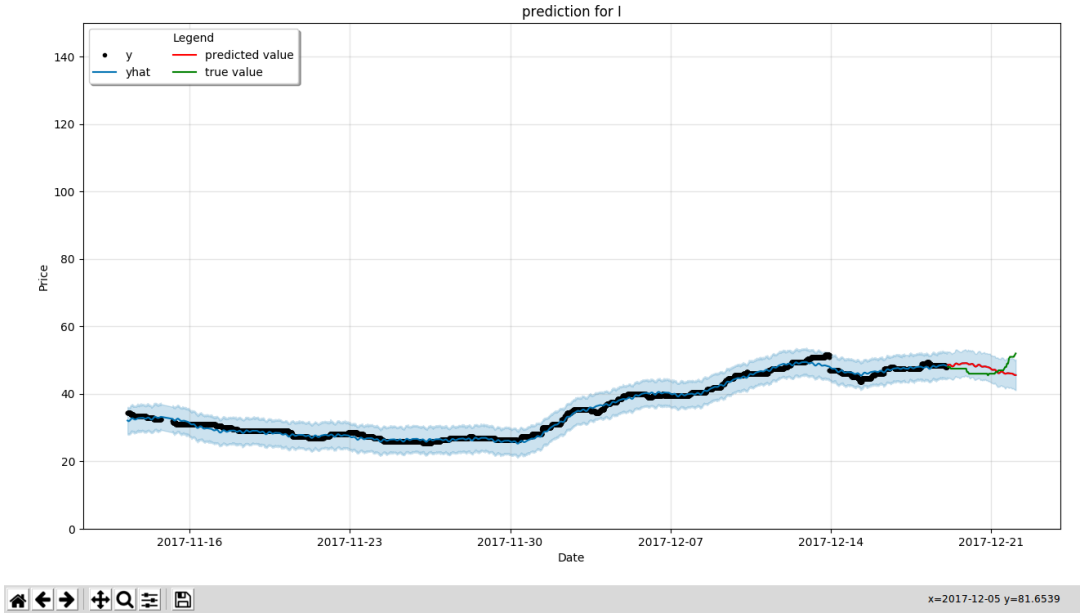

prediction for D
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
x=2017-12-05 y=115.329


prediction for E
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
zoom rect, x=2017-12-21 y=19.1485


prediction for F
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
zoom rect, x=2017-12-12 y=15.0299


prediction for G
Legend
y
yhat
predicted value
true value
Price
Date
2017-11-16
2017-11-23
2017-11-30
2017-12-07
2017-12-14
2017-12-21
x=2017-12-13 y=29.3238





۸- بهبود نتایج و روش‌های دیگر

در این تحقیق سعی شد تا با استفاده از مدل آماری Additive model یک پیش‌بینی نزدیک به واقعیت از آینده داشته باشیم. در صورت نصب پیش نیاز ها و اجرای فایل پایتون ارسالی متوجه می‌شوید که این مدل برای داده‌های بدون outlier خوب عمل میکند و سنجه خطای میانگین مربعات برای آن مقادیر بین ٤ الی ١٠ را محاسبه میکند و در صورتی که داده‌ها دارای outlier باشند سنجه خطای میانگین مربعات حتی تا مقادیر بین ٣٠ الی ٤٠ هم می‌رود.

بنابراین یکی از روش‌های بهبود این مدل حذف outlier ها می‌باشد.

علاوه بر بحث بالا برای بهبود نتایج روش‌های آماری و مبتنی بر شبکه‌های عصبی نیز می‌توانند مفید واقع شوند. ازین روش‌ها می‌توان به روش ARIMA (مخفف Auto-Regressive Integrated Moving Averages) و LSTM (یک روش بر مبنای شبکه‌های عصبی) نام برد.

پایان

تذکر۱: برای اجرای فایل ارسالی باید این کتابخانه ها نصب باشند و همچنین فایل‌های داده و فایل اجرایی برنامه در یک دایرکتوری قرار داشته باشند.

تذکر۲: به دلیل کمبود زمان فرمت خواسته شده برای ورودی و تست برنامه آماده نشد و در عوض از ۱۰ درصد داده‌ها برای تست برنامه استفاده نمودم و نمودار ها و ارزیابی مدل و همه موارد خواسته شده دیگر انجام شد و مدل ایجاد شده کامل است، امیدوارم مورد تأیید باشد.

با تشکر، پیروز باشید.

میلاد چراغی
miladchraghi@gmail.com