



پردیس علوم
دانشکده‌ی ریاضی، آمار و علوم کامپیوتر

تحلیل داده‌های بازار بورس ایران با استفاده از الگوریتم‌های داده‌کاوی

نگارنده

محمدحسین خوش رفتار

استاد راهنما: دکتر سمانه افتخاری مهابادی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر

تیر ۱۴۰۰

چکیده

هدف این پروژه پیش بینی وضعیت سهم های بازار بورس تهران با استفاده از تاریخچه ی سهم و سایر شاخص های اقتصادی از جمله قیمت دلار ، تورم ، نقدینگی و ... میباشد. با استفاده از روش ها و الگوریتم های داده کاوی سعی کردیم که به کمک تاریخچه ی سهم و شاخص های اقتصادی مهم به مدلی دست پیدا کنیم که آینده ی سهم را پیش بینی کند. در گام نخست دیتای مورد نیاز را جمع آوری کردیم. دیتای سال ۲۰۱۵ تا سال ۲۰۲۰ تمامی سهم ها و شاخص های اقتصادی نام برده شده را با فرمت CSV ذخیره کردیم. در گام بعدی با رسم نمودار های مناسب و تحلیل شاخص ها کنار هم به بررسی اینکه کدام شاخص ها بر قیمت سهم تاثیر گذارند پرداختیم و شاخص های مناسب را انتخاب کردیم. در گام آخر هم با استفاده از دیتای موجود به بررسی و تست مدل های مناسب جهت پیش بینی سری های زمانی با دقت مناسب پرداختیم و در نهایت با استفاده از مدل میانگین متحرک خود همبسته یکپارچه ^۱ یک مدل جهت پیش بینی قیمت سهم ها آموزش دادیم و دقت سنجی کردیم.

^۱ARIMA model

پیشگفتار

مهم ترین مسئله تو بازار سرمایه پیشبینی آینده سهم است. با مشخص شدن وضعیت سهم تو آینده ، استراتژی خرید و فروش به آسانی به دست می آید و می توانیم میزان سود را ماکزیم کنیم. فاکتور های بسیار زیادی بر قیمت سهم تاثیر گذار هستند. مهم ترین فاکتور ها عملکرد مالی شرکت صاحب سهم و تاریخچه ی خود سهم هستند ولی از آنجایی که در کشور وضعیت شاخص های اقتصادی پایدار نیست ، عواملی چون قیمت دلار ، تورم ، نقدینگی و ... تاثیر قابل توجهی روی قیمت سهم ها می گذراند. ازینرو برای پیش بینی قیمت سهم نیازمند مدلی هستیم که علاوه بر توجه به تاریخچه ی سهم یا همان ویژگی های سری های زمانی به تاریخچه ی دیگر شاخص های اقتصادی نیز توجه نماید و تغییرات آن ها را در تصمیم گیری لحاظ کند. یکی از چالش های مهم تو این راه این است که واحد ها و مقیاس های موارد مطرح شده یکسان نمی باشد و این موضوع باعث ایجاد بایاس در مدل و تصمیم گیری خواهد شد که باید با استفاده از نرمال سازی و تبدیل های مناسب این مشکل را برطرف کرد

فهرست مطالب

۱	مقدمه	۱
۲	۱.۱ مفاهیم	۲
۲	۱.۱.۱ قیمت دلار	۲
۲	۲.۱.۱ تورم	۲
۲	۳.۱.۱ نقدینگی	۲
۳	۴.۱.۱ تاریخچه قیمت	۳
۳	۵.۱.۱ قیمت پایانی	۳
۴	۲ مراحل فنی پروژه	۴
۴	۱.۲ جمع آوری و پیش پردازش دیتا	۴
۴	۱.۱.۲ جمع آوری دیتای بورس	۴
۵	۲.۱.۲ جمع آوری دیتای شاخص های اقتصادی	۵
۵	۳.۱.۲ حذف نوسانات مقطعی دیتا	۵
۵	۴.۱.۲ رسیدگی به داده های گم شده	۵
۶	۵.۱.۲ یکسان سازی مقیاس ها	۶
۶	۶.۱.۲ یکسان سازی مولفه ی x دیتا	۶

۶	تحلیل پیشگو ها	۲.۲
۷	بررسی تاثیر قیمت دلار	۱.۲.۲
۱۰	بررسی تاثیر نقدینگی	۲.۲.۲
۱۰	طراحی و پیاده سازی مدل	۳.۲
۱۱	انتخاب مدل پایه	۱.۳.۲
۱۱	Prophet مدل	۲.۳.۲
۱۲	LSTM مدل	۳.۳.۲
۱۳	ARIMA مدل	۴.۳.۲
۱۴	GARCH مدل	۵.۳.۲
۱۵	تحلیل نتایج و جمع بندی	۳
۱۵	نتایج بدست آمده از مدل ARIMA	۱.۳
۱۶	نتایج عملی تست	۱.۱.۳
۱۸	تحلیل و جمع بندی	۲.۱.۳

فصل ۱

مقدمه

پروژه از ۳ فاز اصلی تشکیل شده که به ترتیب جمع آوری و پیش پردازش دیتا، بررسی شهودی و انتخاب پیشگو ها و در نهایت طراحی و پیاده سازی مدل می باشد. تمام مراحل فنی این پروژه با استفاده از زبان برنامه نویسی پایتون^۱ پیاده سازی شده است و تمام مراحل پیاده سازی پروژه تحت گیت^۲ ثبت شده است و میتوانیم مراحل و تاریخچه را به کمک آن ببینیم. به منظور تصمیم بهتر راجع به پیشگو^۳ های اقتصادی مطرح شده لازم داریم به طور شهودی با آن ها آشنا باشیم که در ادامه به طور مختصر توضیحاتی راجع به آن ها فراهم شده است. علاوه بر این نیازمند برای فهم بهتر تحلیل های ارائه شده خوب است که با مفاهیم اولیه ی بازار بورس آشنا باشیم ولی به دلیل گستردگی این مطالب از اشاره به همه ی آن ها در این مقاله خودداری کرده ایم و فقط به موارد ضروری اشاره کرده ایم

^۱Python programming language

^۲Git

^۳Predictor

۱.۱ مفاهیم

۱.۱.۱ قیمت دلار

در این پروژه منظور از قیمت دلار، ارزش یک دلار آمریکا به ریال در تاریخ مشخصی و با توجه به قیمت رسمی اعلام شده از سوی بانک مرکزی می باشد.

۲.۱.۱ تورم^۴

از نظر علم اقتصاد اشاره به افزایش سطح عمومی تولید پول، درآمدهای پولی یا قیمت است. تورم عموماً به معنی افزایش غیرمتناسب سطح عمومی قیمت در نظر گرفته می شود. تورم، روند افزایشی و نامنظم افزایش قیمت ها در اقتصاد است. هر چند بر پایه نظریه های گوناگون، تعریف های متفاوتی از تورم ارائه می شود، اما، تمامی آن ها به روند فزاینده و نامنظم افزایش در قیمت ها اشاره دارند. در این پروژه از آمار رسمی که توسط سازمان آمار اعلام می شود و بر اساس میانگین افزایش قیمت کالا محاسبه می شود استفاده شده است.

۳.۱.۱ نقدینگی^۵

منظور از نقدینگی، میزان سرمایه ی نقدی که دست مردم است می باشد. این سرمایه شامل اعتبار حساب بانکی افراد، پول نقد در دسترسشون، و یا دیگر سرمایه های نقدی مردم می شود. باید توجه داشت که خودرو، مسکن و یا دلار جز سرمایه های نقدی افراد حساب نمی شود و جز آمار نقدینگی محسوب نمی شود چون با نقد شدن مقداری فاصله دارد. در این مقاله از آمار رسمی بانک مرکزی راجع به نقدینگی استفاده شده است.

^۴ Inflation

^۵ Cash

۴.۱.۱ تاریخچه قیمت^۶

منظور از تاریخچه ی قیمت، سری زمانی^۷ قیمت سهم می باشد. یعنی اینکه به ازای هر روز خاص قیمت سهم تو اون روز چقدر بوده است.

۵.۱.۱ قیمت پایانی^۸

منظور از قیمت پایانی یک روز، میانگین وزن دار تمام معاملات آن روز با وزن حجم معاملات انجام شده می باشد

Price history^۶

Time series^۷

Closing price^۸

فصل ۲

مراحل فنی پروژه

۱.۲ جمع آوری و پیش پردازش دیتا

برای هر نوع تحلیل داده کاوی ما نیازمند دیتا هستیم. بدون دیتا هیچ تحلیلی هم امکان پذیر نیست. همچنین این دیتا می بایست یک سری ویژگی ها داشته باشد که تحلیل ما را آسان تر و دقیق تر بکند.

۱.۱.۲ جمع آوری دیتای بورس

برای تحلیل ما نیازمند تاریخچه ی سهم های بازار بورس ، تاریخچه ی قیمت دلار ، تورم و نقدینگی بودیم. برای دسترسی به دیتای تاریخچه ی سهم های بازار بورس راه های فراوانی هست و ما از کتابخانه ی پاتونی `pytse_client` استفاده کردیم که قابلیت دانلود تاریخچه ی تمام سهم های بازار سرمایه را به ریزدانگی روز از سال ۲۰۱۵ را در اختیارمون قرار میدهد. خروجی استفاده از این کتابخانه یک فایل `csv` به ازای هر سهم بازار می باشد که حاوی تاریخچه ی سهم مورد نظر با ریزدانگی روز می باشد. این فایل ها در پوشه ی `tickers_data` از پروژه قرار دارند. به این نکته

توجه شود که قیمت پایانی سهم به عنوان قیمت اون روز سهم در نظر گرفته شده است.

۲.۱.۲ جمع آوری دیتای شاخص های اقتصادی

برای جمع آوری دیتای تاریخچه ی قیمت دلار از منبع شبکه ی اطلاع رسانی قیمت ارز^۱ استفاده شد و قیمت دلار به صورت ماهانه در فایل با فرمت CSV ذخیره گردید. برای سایر شاخص های اقتصادی نام برده شده از منبع مرکز آمار و بانک مرکزی استفاده گردید و دیتای ماهانه ی شاخص های مذکور در فایل با فرمت CSV با اسامی مشخص هر شاخص ذخیره گردید که در ادامه مورد استفاده قرار گیرد.

۳.۱.۲ حذف نوسانات مقطعی دیتا

در گام نخست به منظور ساده تر شدن تحلیل و حذف نوسان های مقطعی به پیشنهاد استاد سطح ریزدانگی داده را از روزانه به ماهانه تغییر دادیم بدین شکل که به ازای هر سهم و یا شاخص اقتصادی به جای اینکه به ازای هر روز قیمت را نگه داریم به ازای هر ماه قیمت ابتدا یا انتهای آن ماه رانگهداری کنیم و بر اساس این دیتا تحلیل رو انجام دهیم.

۴.۱.۲ رسیدگی به داده های گم شده^۲

با توجه به اینکه نماد های بازار بورس بر اثر افشا ها یا سایر موارد ممکن است روز هایی بسته باشند و دیتا اون روز ها موجود نباشد ، ممکن است در تحلیل ها ما را دچار مشکل کنند و این مسئله باید رسیدگی شود. ازین رو از روش آخرین مشاهده انجام شده به جلو^۳ برای پر کردن جای خالی

tgju^۱

Missing data^۲

last observation carried forward^۳

این روز ها استفاده کردیم. بدین شکل که قیمت اولین روزی که سهم بعد از بسته بودن باز شده و تعیین قیمت شده است را به جای روز های بسته بودن سهم قرار می دهیم

۵.۱.۲ یکسان سازی مقیاس ها

با توجه به اینکه مقیاس سهم های بازار سرمایه ، قیمت دلار و سایر شاخص های اقتصادی تفاوت بسیار زیادی دارد و این تفاوت در مقیاس باعث ایجاد خطا در تحلیل ها و همچنین خراب کردن عملکرد مدل می شوند به پیشنهاد استاد به جای استفاده از مقدار متغیر ها از درصد تغییر شاخص ها برای تحلیل استفاده کردیم. درصد تغییر در سهم های بازار بورس به همان معنی بازدهی ماهانه سهم هستند و در سایر شاخص ها هم درصد تغییر شاخص در ماه مورد نظر هستند.

۶.۱.۲ یکسان سازی مولفه ی X دیتا

روشی که ما برای رسیدگی به داده های گم شده انتخاب کردیم باعث میشد که مولفه ی X دیتای ما که از جنس تاریخ بود در بهضی از ماه ها یکسان نباشد بدین شکل که در حالت عادی روز اول ماه هست ولی در حالتی که جای گذاری اتفاق افتاده روز دوم سوم یا جلوتر از ماه باشد و این تفاوت در مرحله ی پیاده سازی مدل برامون مشکلاتی به وجود آورد برای همین به جای نگهداری روز در مولفه ی X دیتامون مستقیما ماه مورد نظر رابه عنوان مولفه ی X در نظر گرفتیم. به طور مثال تاریخ ۲۰۱۹-۰۴-۰۳ به ۲۰۱۹-۰۴-۰۴ تغییر کرد

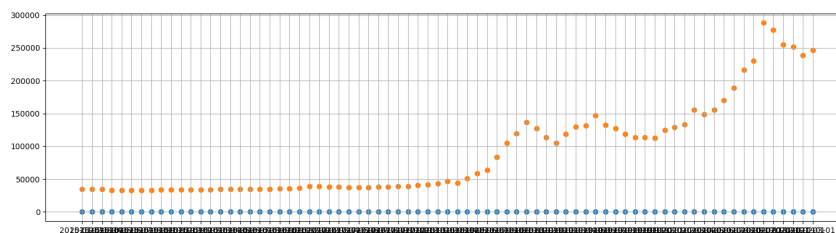
۲.۲ تحلیل پیشگو ها

اکنون که دیتا آماده و تمیز شده است نیازمند این هستیم که پیشگو هایی که انتخاب کرده ایم را بررسی نماییم تا از بین پیشگو های مورد نظر فقط آن هایی که در متغیر هدف ما تاثیر گذار هستند

را به مرحله ی طراحی مدل ببریم. لازم به ذکر است که تاریخچه ی خود سهم با توجه به شناختی که از دیتا داریم پیشگویی اجتناب ناپذیر می باشد و بدون تحلیل به مرحله مدل سازی میبریم. از ابزارهای شهودی آمار جهت تحلیل پیشگوها استفاده کردیم. به منظور رسم نمودارهای مورد نیاز از کتابخانه ی پایتونی matplotlib استفاده کردیم.

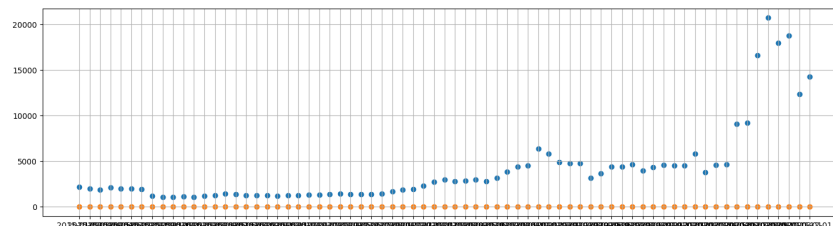
۱.۲.۲ بررسی تاثیر قیمت دلار

از بین سهم های بازار بورس نماد های فولاد، اخبار و خودرو را به نمایندگی از شاخص های بزرگ جهت تحلیل انتخاب میکنیم. ابتدا نمودار اسکتر^۴ قیمت دلار را در کنار سهم های مذکور مورد بررسی می نماییم.

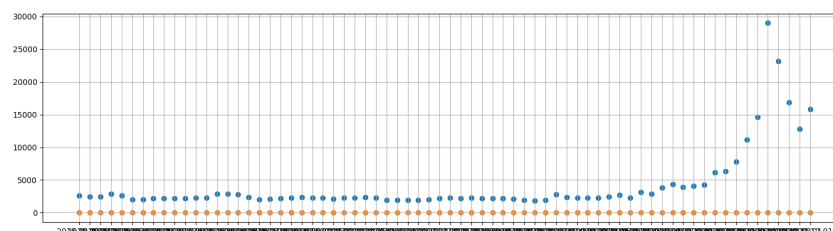


شکل ۱.۲: نمودار قیمت دلار

Scatter plot^۴



شکل ۲.۲: نمودار قیمت سهم فولاد

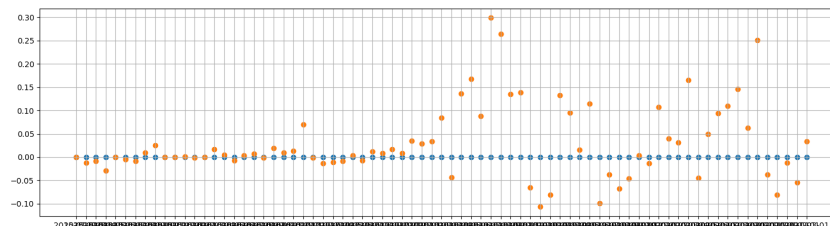


شکل ۳.۲: نمودار قیمت سهم اخبر

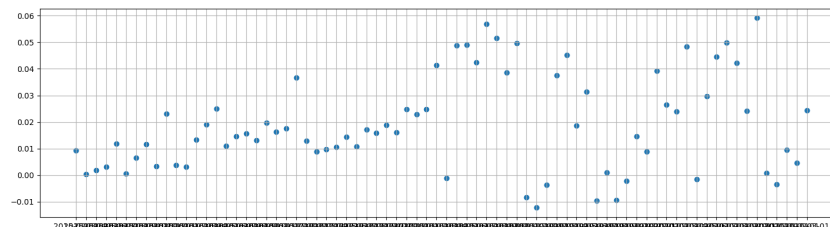
همانطور که مشاهده میشود همبستگی بسیار زیاد بین قیمت دلار و قیمت نماد فولاد و نماد اخبر وجود دارد. با رسم نمودار تغییرات این دو نیز دقیقاً همین به نتیجه خواهیم رسید پس از رسم مجدد آن صرف نظر میکنیم. پس با توجه به تحلیل بالا شواهد کافی برای حضور قیمت دلار در مرحله طراحی مدل وجود دارد.

بررسی تاثیر تورم

طبق تحلیل شناختی از دیتا قابل حدس می باشد که تورم و قیمت دلار به یک معنی هستند و باید یکدیگر را تایید کنند. به منظور اطمینان بیشتر از صحت حدس خود نمودارهای تغییرات دلار و تغییرات تورم را در کنار یکدیگر مورد بررسی قرار می دهیم



شکل ۴.۲: نمودار درصد تغییر قیمت دلار



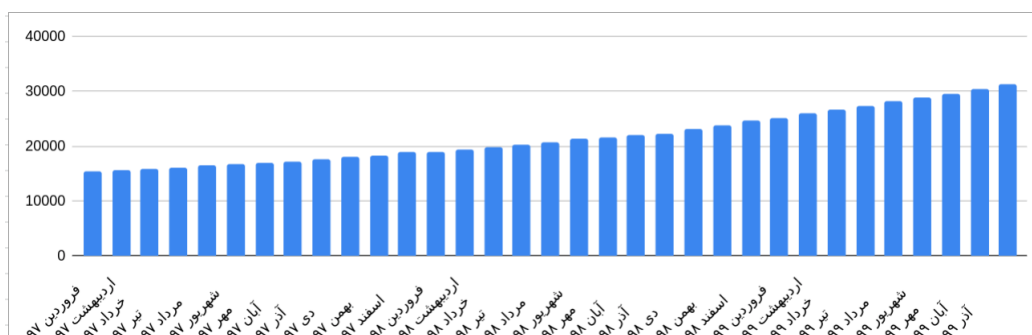
شکل ۵.۲: نمودار درصد تغییر تورم

همانطور که مشاهده میشه تا حد خوبی این دو شاخص با هم همبستگی دارند ولی برای اثبات شهود خود از ضریب همبستگی پیرسون استفاده میکنیم. طبق رابطه زیر r را محاسبه میکنیم

$$r = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

در نهایت با محاسبه r برای دیتای مذکور به عدد نزدیک ۱ میرسیم که نشان دهنده همبستگی مثبت خوبی هست با توجه به همبستگی بالای این دو شاخص میتوان از یکی آن ها در مرحله ی طراحی مدل صرف نظر کرد ولی ممکن است به هر حال یک سری اطلاعات را از دست بدیم چون همبستگی کامل نداشتن این دو شاخص. با توجه به حجم کم دیتا شاخص تورم را نیز در مرحله ی طراحی مدل قرار می دهیم.

۲.۲.۲ بررسی تاثیر نقدینگی



شکل ۶.۲: نمودار نقدینگی ماهانه

با توجه به نمودار نقدینگی ماهانه متوجه میشویم که نقدینگی کاملاً به صورت خطی در حال افزایش است و نوسان خاصی هم ندارد. این به این معنی هست که نقدینگی هیچ اطلاعات زیادی به ما نمیدهد. با ترسیم نمودار تغییرات نقدینگی یک خط موازی با محور x ها دریافت میکنیم و همبستگی پایینی هم بین نقدینگی و قیمت سهم ها و دلار وجود خواهد داشت. پس میتوان از ورود شاخص نقدینگی به مرحله ی مدل سازی صرف نظر کرد ولی با توجه به پیشنهاد استاد و کم بودن حجم دیتا تصمیم بر این شد که نقدینگی هم وارد مرحله مدل سازی شود.

۳.۲ طراحی و پیاده سازی مدل

بعد از تحلیل و انتخاب پیشگوها با توجه به جنس و خصیصه^۵ های دیتا باید به بررسی مدل های پایه و موجود پردازیم و با استفاده از یک مدل پایه مدلی جهت پیشبینی قیمت سهم پیاده سازی کنیم. به طور مشخص دیتا از جنس سری زمانی می باشد پس باید از بین مدل های مخصوص سری زمانی مدلی انتخاب کنیم که خصیصه های سری های زمانی را به خوبی مورد توجه قرار دهد.

^۵Feature

۱.۳.۲ انتخاب مدل پایه

با بررسی اولیه تمام مدل های موجود مناسب برای سری های زمانی مدل های زیر برای بررسی و تست بیشتر انتخاب شدند و مورد بررسی قرار گرفتند.

- Prophet model
- LSTM model
- ARIMA model
- GARCH model

در ادامه توضیحات بیشتری راجع به هر کدام و دلیل رد یا انتخابشون را بیان کردیم

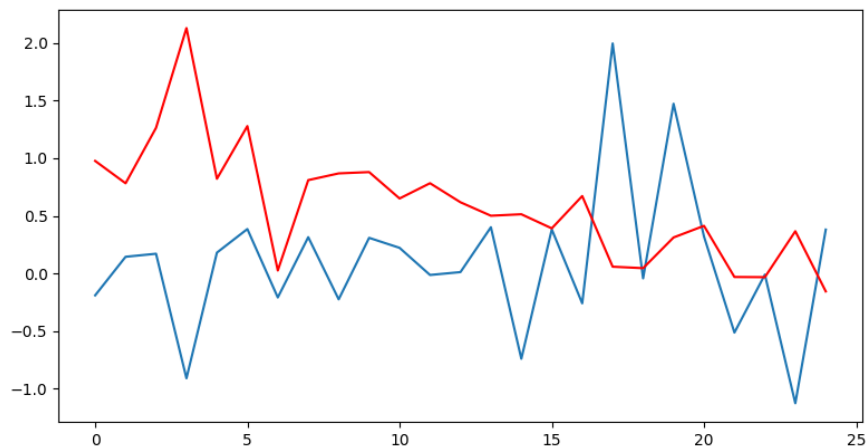
۲.۳.۲ Prophet مدل

این مدل بر پایه ی مدل افزودنی^۶ طراحی شده است که مدلی رگرسیونی ناپارامتری^۷ می باشد. این مدل بر اساس فصلی بودن^۸ دیتا تصمیم میگیرد و به دنبال الگو های تکرار شونده می باشد. بر اساس تست های اولیه ای که با این مدل گرفتیم دقت بسیار پایینی دریافت کردیم که با توجه به جمله ی قبلی خیلی دور از انتظار نبود با توجه به جنس دیتای مورد تحلیل ما.

Additive model^۶

Nonparametric regression^۷

seasonality^۸



شکل ۷.۲: نمودار دقت مدل

همانطور که در شکل مشاهده می کنید، مدل خطای بالایی در پیشبینی داشته است و حتی در تشخیص روند^۹ دچار خطا شده است و الگو رو به پایین پیشبینی کرده است. طبق محاسبه ی میانگین مربعات خطا^{۱۰} به عدد نزدیک ۳ میرسیم که خطای بسیار زیادی میباشد.

۳.۳.۲ مدل LSTM

این مدل بر پایه ی شبکه ی عصبی بازگشتی^{۱۱} طراحی شده است و در مسائل الگویابی کاربرد دارد از آنجایی که این مدل بسیار پیچیده است و مباحث پایه ای استفاده شده در آن خارج مباحث داده کاوی مقدماتی می باشد تلاشی برای تست این مدل به صورت عملی صورت نگرفت و صرفاً در حد تحقیق باقی گذاشته شد.

^۹trend

^{۱۰}mean squared error

^{۱۱}Recurrent neural network

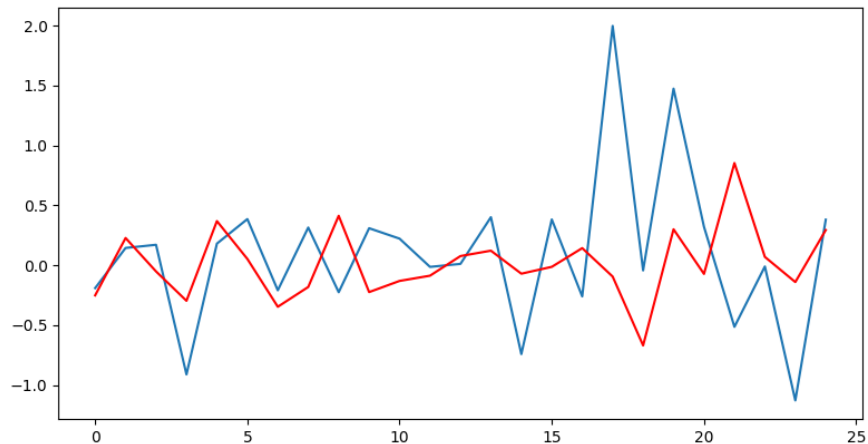
۴.۳.۲ مدل ARIMA

این مدل از ترکیب دو مدل AR یا همان مدل خودهمبسته^{۱۲} و مدل میانگین متحرک^{۱۳} می باشد. مدل خود همبسته نوعی از فرایند تصادفی می باشد و بر اساس اینکه دیتا به صورت خطی به گذشته ی خودش وابسته است کار میکند. مدل ARIMA با استفاده از ترکیب مدل AR و مدل MA ساخته شده است برای پیشینی متغیر های پیچیده تر در سری های زمانی. این مدل دارای ۳ پارامتر می باشد که هر کدام به ترتیب مشخص کننده ی پارامتر مدل خودهمبسته، درجه تفاوت^{۱۴} و درجه ی مدل میانگین متحرک می باشد. راجع به مدل های میانگین متحرک و خودهمبسته با توجه به مربوط بودن به مباحث فرایند تصادفی و سری زمانی خیلی در این مقاله توضیح نمیدهیم و به خواننده واگذار میکنیم. درجه تفاوت تعداد بار هایی است که باید داده با داده ی گذشته ی خود تفریق گردد تا به حالت ایستا برسیم ولی با توجه به نرمال سازی دادمون و استفاده از درصد تغییرات به جای خود قیمت این پارامتر را در مدل برابر ۰ قرار میدهم پس با این اوصاف در واقع ما داریم از مدل ARMA استفاده میکنیم.

^{۱۲}Autoregressive model

^{۱۳}Moving-average model

^{۱۴}Degree of difference



شکل ۸.۲: نمودار دقت مدل رو سهم فولاد

همانطور که در شکل مشاهده میکنید دقت بسیار بیشتری بدست آماده است و مدل علاوه بر تشخیص صحیح روند اختلاف کمی با جواب اصلی دارد. خطا در این حالت حدود ۵ دهم میباشد که به نسبت مدل prophet بسیار کمتر است.

۵.۳.۲ مدل GARCH

با توجه به نتایج خوبی که از مدل قبلی گرفتیم این مدل را کنار گذاشتیم و مورد تست و بررسی قرار ندادیم. از آنجایی که پایه ی این مدل هم مدل خودهمبسته می باشد حدس میزدیم که نتایج بسیار متفاوتی نخواهیم گرفت. در نهایت تصمیم بر این شد که روی مدل ARMA بیشتر وقت گذاشته بشه و با تغییر پارمترها به بهترین دقت برسیم.

فصل ۳

تحلیل نتایج و جمع بندی

۱.۳ نتایج بدست آمده از مدل ARIMA

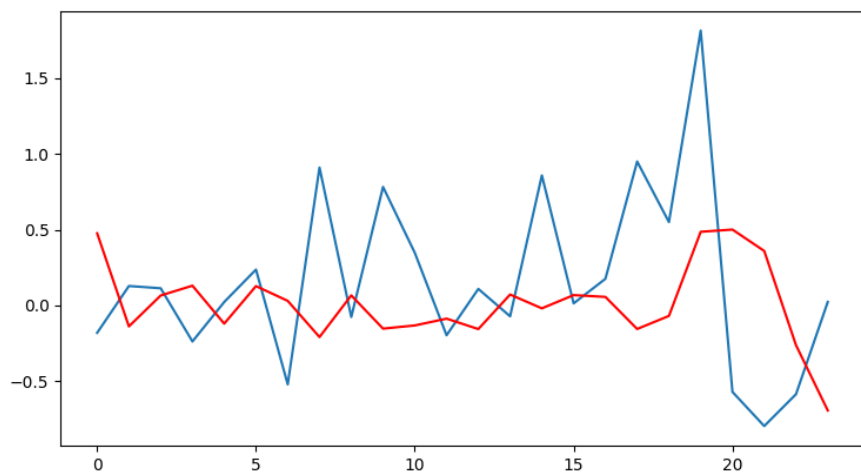
ابتدا نتایج بدست آمده از مدل را مشاهده کنیم و در آخر تحلیل نتایج را ببینیم. دو حالت برای انجام تست رو مدل امکان پذیر بود. در هر دو حالت ابتدا داده به دو قسمت آموزشی و آزمایشی تقسیم میشود که در تست های انجام شده ی ما ۶۶ درصد داده ها از ابتدای سری زمانی به عنوان داده ی آموزشی انتخاب شدند و ما بقی به عنوان داده ی آزمایشی انتخاب شدند. حالت اول برای تست به این روش عمل میکردیم که مدل با توجه به کل داده های آزمایشی بعلاوه ی داده هایی که خودش آن ها را پیش بینی کرده آموزش داده شود و نقطه ی بعدی را پیش بینی کند و به همین ترتیب جلو برود. حالت دوم این است که در هر مرحله عضوی که مدل پیشبینی کرده است را مقدار اصلیشو به عنوان مشاهده به مجموعه داده های آزمایشی اضافه کنیم و مدل مرحله ی بعدی را پیشبینی کند و به همین ترتیب پیش برویم. ما در تمام تست ها از حالت اول استفاده کردیم و در بخش تحلیل بیشتر راجع به تفاوت این دو روش صحبت خواهد شد

۱.۱.۳ نتایج عملی تست

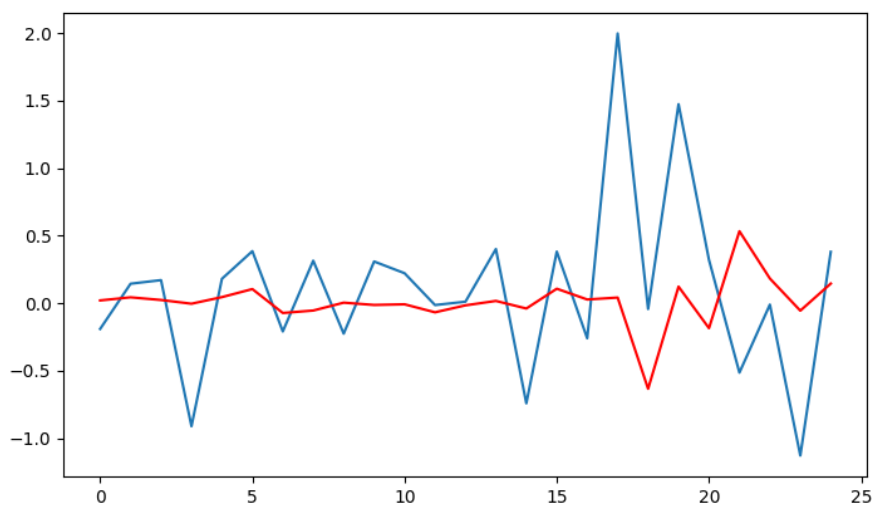
چند نمونه از نتایج به دست آمده توسط مدل طراحی شده در ادامه آمده است.



شکل ۱.۳: نمودار دقت مدل رو سهم خودرو با پارامترهای (۳، ۰، ۴)



شکل ۲.۳: نمودار دقت مدل رو سهم اخبار با پارامترهای $(۳, ۰, ۴)$



شکل ۳.۳: نمودار دقت مدل رو سهم اخبار با پارامترهای $(۲, ۰, ۰)$

۲.۱.۳ تحلیل و جمع بندی

همانطور که مشاهده می شود مدل در شروع کار دقت بهتری دارد و در ادامه رفته رفته از دقت آن کم میشود. علت این اتفاق این است که از روشی برای آزمایش و تست مدل استفاده میکنیم که پیشبینی خود مدل را به عنوان مشاهده جدید به مدل میدهیم در کنار داده های آموزشی و از آنجایی که این پیشبینی خطا دارد رفته رفته خطا به نقاط جلو تر بازشر میشود و خطای نقاط جلوتر بیشتر میشود ولی با توجه به اینکه این روش در دنیای واقعی بیرون عملی تر و ساده تر است نسبت به روش دوم و دید بلند مدت تری ارائه می دهد ترجیح دادیم مدلمون را با توجه به این روش بهینه کنیم. در شکل ۳.۳ همانطور که مشاهده میکنید پارامتر مدل میانگین متحرک را برابر صفر قرار دادیم و مدل صرفا یه مدل ساده ی خودهمبسته شد و دقت نهایی کاهش پیدا کرد.

در نهایت برای رسیدن به بهترین دقت با روش آزمون و خطا سعی کردیم که به بهترین پارامتر ها برای مدل برسیم ولی این روش نقطه تعادلی نداشت برامون و در نهایت مشاهده کردیم که برای هر سهم با توجه به رفتار و نوسانات بهترین دقت در پارامتر های مختلفی کسب میشود و نمیشود قانون کلی براش صادر کرد.

واژه‌نامه فارسی به انگلیسی

ARIMA model.....	مدل میانگین متحرک همبسته یکپارچه
Python programming language	زبان برنامه نویسی پایتون
Git.....	گیت
Predictor	تورم
Asset Distribution	نقدینگی
Financial Risk.....	تاریخچه قیمت
Sharpe Ratio	قیمت پایانی
Correlation.....	داده ی گم شده
Eigenvalue	آخرین مشاهده ی انجام شده به جلو
Eigenvector.....	نموداراسکتر
Missing Value	خصیصه
Cast	مدل افزودنی
Weight	رگرسیون ناپارامتری
Weight	فصلی بودن
Cast	روند

مراجع

- [1] Larose D.T. and Larose C.D. (2014) Discovering knowledge in data: an introduction to data mining (Second edition). John Wiley & Sons.
- [2] Ruey S. Tsay, “Analysis of Financial Time Series, 3rd Edition”, WILEY, 2010.
- [3] Introduction to Stochastic Processes. Lothar Breuer

Abstract

The purpose of this project is to predict the future of Tehran Stock using the history of stocks and other economic indicators such as dollar price, inflation, liquidity and etc. we tried to implement a model using data mining algorithms that can predict stock future using stock's and economic indicators history. In the first step, we collected the required data. we saved all the stocks and mentioned economic indicators from 2015 to 2020 in csv format. next step, we examined which indicators affect the stock price by drawing appropriate graphs and analyzing the indicators together then we selected the appropriate indicators. In the last step, we trained an ARIMA model to predict stock future and we tested our model.



College of Science

School of Mathematics, Statistics, and Computer Science

Analyse Iran stock exchange using data mining algorithms

Mohammad Hossein Khoshraftar

Supervisor: Dr. Samaneh Eftekhari Mahabadi

A thesis submitted to Graduate Studies Office
in partial fulfillment of the requirements for the degree of

B.Sc.in

Computer Science

July , 2021