



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 سهم‌های 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 داده‌های مورد نیاز را جمع آوری کردیم. داده‌های سال ۲۰۱۵ تا سال ۲۰۲۰ از تمامی سهم ها و شاخص های اقتصادی نام برده شده را با فرمت CSV ذخیره کردیم. در گام بعدی با رسم نمودارهای 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 داده‌های 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 <sup>۱</sup> یک مدل جهت پیش بینی قیمت سهم ها آموزش دادیم و دقت سنجی کردیم.

---

<sup>۱</sup>ARIMA model

## پیشگفتار

مهم ترین مسئله در بازار سرمایه پیش‌بینی آینده سهم است. با مشخص شدن وضعیت سهم در آینده، استراتژی خرید و فروش به آسانی به دست می‌آید و می‌توانیم میزان سود را ماکزیم کنیم. فاکتورهای 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 زیادی روی قیمت سهم ها می گذرانند. ازین رو 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 این راه یکسان نبودن واحد و مقیاس در شاخص های اقتصادی مذکور می باشد که باعث ایجاد خطا در مدل و تصمیم گیری خواهد شد که باید با استفاده از نرمال سازی و تبدیل های مناسب این مشکل را برطرف کرد. در زمینه ی پیش بینی قیمت سهم مقالات زیادی وجود دارد ولی اکثرا به شاخص های اقتصادی و تاثیرات آن ها پرداخته اند و تمرکز را روی تحلیل سری زمانی خود سهم و الگوی افزایش و کاهش قیمت قرار داده اند. به طور مثال در سال ۲۰۱۵ آقای اگاندی در مقاله ای [۳] به کمک مدل های داده کاوی از جمله مدل ARIMA به تحلیل داده های بورسی پرداخته است. در این پروژه سعی شده است که تاثیرات شاخص های اقتصادی اعمال شود.

## 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
۳	۶.۱.۱ آخرین قیمت	۳
۴	۷.۱.۱ پیشگو	۴
۴	۸.۱.۱ متغیر هدف	۴
۴	۹.۱.۱ مدل	۴
۴	۱۰.۱.۱ پیش‌بینی	۴
۵	۱۱.۱.۱ خطای پیش‌بینی	۵
۵	۱۲.۱.۱ داده‌های گم شده	۵
۵	۱۳.۱.۱ بازه زمانی	۵

۶	۲	مراحل فنی پروژه
۶	۱.۲	جمع آوری و پیش پردازش داده ها
۶	۱.۱.۲	جمع آوری داده بورس
۷	۲.۱.۲	جمع آوری داده های شاخص های اقتصادی
۷	۳.۱.۲	حذف نوسانات مقطعی داده
۷	۴.۱.۲	رسیدگی به داده های گم شده
۸	۵.۱.۲	یکسان سازی مقیاس ها
۸	۶.۱.۲	یکسان سازی مولفه ی $x$ داده
۸	۲.۲	تحلیل پیشگو ها
۹	۱.۲.۲	بررسی تاثیر قیمت دلار
۱۳	۲.۲.۲	بررسی تاثیر نقدینگی
۱۳	۳.۲	انتخاب مدل پایه
۱۴	۱.۳.۲	مدل Prophet
۱۵	۲.۳.۲	مدل LSTM
۱۶	۳.۳.۲	مدل ARIMA
۱۷	۴.۲	پیاده سازی مدل به کمک ARIMA
۱۷	۱.۴.۲	تعیین پارامتر ها
۲۱	۲.۴.۲	اثردهی پیشگو ها
۲۳	۳	تحلیل نتایج و جمع بندی
۲۳	۱.۳	نتایج بدست آمده از مدل ARIMA
۲۴	۱.۱.۳	نتایج عملی تست
۲۵	۲.۱.۳	تحلیل و جمع بندی

# فصل ۱

## مقدمه

پروژه از ۳ فاز اصلی تشکیل شده که به ترتیب جمع آوری و پیش پردازش داده، 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<sup>۱</sup> پیاده سازی شده است و تمام مراحل پیاده سازی پروژه تحت گیت<sup>۲</sup> ثبت شده است و میتوانیم مراحل و تاریخچه را به کمک آن ببینیم. به منظور تصمیم بهتر راجع به پیشگو<sup>۳</sup> های اقتصادی مطرح شده نیاز است که 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 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 خودداری کرده ایم و فقط به موارد ضروری اشاره شده است.

---

<sup>1</sup>Python programming language

<sup>2</sup>Git

<sup>3</sup>Predictor

## 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

تورم<sup>۴</sup> نرخ تغییر سطح عمومی قیمت ها می باشد و بر اساس میانگین تغییرات قیمت مجموعه ای از کالاهای مشخصی محاسبه می شود. شاخص تورم به صورت ماهانه توسط بانک مرکزی و همچنین سازمان آمار محاسبه و اعلام می شود. در این پروژه از آمار رسمی که توسط سازمان آمار اعلام می شود استفاده شده است.

### ۳.۱.۱ نقدینگی

نقدینگی<sup>۵</sup> مجموع پول و شبه پولی است که در اقتصاد کشور جریان دارد. پول، طبق تعریف شامل اسکناس و مسکوک در دست اشخاص به علاوه سپرده های دیداری است و شبه پول نیز شامل سپرده های غیردیداری است. نقدینگی در اقتصاد ضریبی از پایه پولی است که توسط بانک مرکزی خلق شده است.

عوامل موثر بر حجم نقدینگی در اقتصاد عمدتاً شامل سه قلم «خالص دارایی های خارجی سیستم بانکی»، «خالص بدهی بخش دولتی به سیستم بانکی» و «بدهی بخش غیردولتی به سیستم بانکی» است. مجموع اقلام فوق به صورت پول و شبه پول در اقتصاد به گردش درمی آید.

---

<sup>4</sup>Inflation

<sup>5</sup>Cash

#### ۴.۱.۱ تاریخچه قیمت

منظور از تاریخچه ی قیمت، سری زمانی<sup>۶</sup> قیمت سهم می باشد. یعنی اینکه به ازای هر روز خاص قیمت سهم در اون روز چقدر بوده است. معمولاً در بازار بورس تهران قیمت پایانی سهم به عنوان قیمت سهم در یک روز خاص در نظر گرفته می شود ولی به طور کلی در بقیه بازار ها معمولاً از آخرین قیمت استفاده می شود.

#### ۵.۱.۱ قیمت پایانی

منظور از قیمت پایانی<sup>۷</sup> یک روز، میانگین وزن دار تمام معاملات آن روز با وزن حجم معاملات انجام شده می باشد. در بازار بورس تهران با توجه به اینکه محدودیت قیمت برای سهم ها وجود دارد قیمت پایانی اهمیت پیدا می کند زیرا تعیین کننده ی محدوده نوسانی قیمت روز معاملاتی آینده می باشد.

محدوده نوسانی قیمت هر سهم در هر روز معاملاتی معمولاً بازه ی ۵ درصد کمتر تا ۵ درصد بیشتر از قیمت پایانی روز معاملاتی گذشته است.

#### ۶.۱.۱ آخرین قیمت

منظور از آخرین قیمت هر سهم، قیمت آخرین معامله ی انجام شده ی سهم در روز معاملاتی مشخصی می باشد. این قیمت مستقل از حجم معامله انجام شده می باشد.

---

<sup>6</sup>Time series

<sup>7</sup>Closing price



### ۷.۱.۱ پیشگو

پیشگو متغیری می‌باشد که جهت پیش‌بینی متغیری دیگر که به عنوان متغیر هدف شناخته می‌شود استفاده می‌شود. متغیر پیشگو معادل متغیر مستقل در علم ریاضیات می‌باشد. برای پیش‌بینی متغیر هدف ممکن است چندین پیشگو وجود داشته باشد. پیشگو ها می‌توانند کاملاً مستقل از یکدیگر یا وابسته به هم باشند.

### ۸.۱.۱ متغیر هدف

متغیر هدف متغیری است که با استفاده از پیشگوها پیش‌بینی می‌شود. متغیر هدف معادل متغیر وابسته در علم ریاضیات می‌باشد. از پیشگوهایی که متغیر هدف به آن ها وابسته است جهت پیش‌بینی استفاده می‌شود و مابقی کنار گذاشته می‌شوند.

### ۹.۱.۱ مدل

در این پروژه، منظور از مدل، مدل آماری می‌باشد. مدل آماری نوعی از مدل ریاضی می‌باشد که دارای مجموعه ای از فرض‌های آماری می‌باشد و روابط بین متغیر های تصادفی را فرمول بندی می‌کند. مدل‌های آماری کاربرد های فراوانی دارند که در این پروژه برای پیش‌بینی مورد استفاده قرار می‌گیرند.

### ۱۰.۱.۱ پیش‌بینی

در این مطالعه، منظور از پیش‌بینی بدست آوردن مقادیر متغیرهای هدف به کمک متغیر های پیشگو می‌باشد.

### ۱۱.۱.۱ خطای پیش‌بینی

منظور از خطای پیش‌بینی اختلاف مقدار واقعی متغیر هدف با مقدار پیش‌بینی شده می‌باشد در وضعیت مشخصی می‌باشد. این مقدار می‌تواند صفر باشد و به معنای این است که خطایی وجود نداشته است.

### ۱۲.۱.۱ داده‌های گم شده

منظور از داده‌های گم شده<sup>۸</sup> داده‌هایی می‌باشد که برا اثر خطای نمونه‌گیری یا سیستمی از دست رفته باشند و قابل دست‌یابی نباشند. به طور مثال برای داده‌های سهم‌های بورسی ممکن است قیمت بعضی روزها به دلایلی موجود نباشد که به آن داده‌ی گم شده می‌گویند و نیازمند اقدامات خاص خود می‌باشد.

### ۱۳.۱.۱ بازه زمانی

در این پروژه منظور از بازه زمانی<sup>۹</sup>، فاصله‌ی نقاط محور  $x$  در سری زمانی مورد بحث می‌باشد. به طور مثال در داده‌های بورس اگر قیمت روزانه‌ی سهم‌ها را در نظر بگیریم بازه زمانی مورد بررسی روز می‌شود و اگر قیمت ماهانه سهم‌ها را در نظر بگیریم بازه زمانی مورد بررسی ماه می‌شود.

---

<sup>۸</sup>Missing data

<sup>۹</sup>Time interval

## فصل ۲

### مراحل فنی پروژه

#### ۱.۲ جمع آوری و پیش پردازش داده ها

برای هر نوع تحلیل داده کاوی ما نیازمند داده هستیم. بدون داده هیچ تحلیلی هم امکان پذیر نیست. همچنین این داده می بایست یک سری ویژگی ها داشته باشد که تحلیل را آسان تر و دقیق تر کند ازینرو قبل از تحلیل داده و یا مدل سازی آن نیازمند اقداماتی هستیم که در ادامه مورد بحث قرار گرفته اند.

##### ۱.۱.۲ جمع آوری داده بورس

برای تحلیل ما نیازمند تاریخچه ی سهم های بازار بورس ، تاریخچه ی قیمت دلار ، تورم و نقدینگی بودیم. برای دسترسی به داده ی تاریخچه ی سهم های بازار بورس راه های فراوانی هست و ما از کتابخانه ای در زبان پایتون به نام `pytse_client` استفاده کردیم که قابلیت دانلود تاریخچه ی تمام سهم های بازار سرمایه را به دقت روز از سال ۲۰۱۵ را در اختیارمان 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 دقت 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 آن روز سهم در نظر گرفته شده است.

## ۲.۱.۲ جمع آوری داده‌های شاخص های اقتصادی

برای جمع آوری داده‌های تاریخچه ی قیمت دلار از منبع شبکه ی اطلاع رسانی قیمت ارز<sup>۱</sup> 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 داده ی 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 داده

در گام نخست به منظور ساده تر شدن تحلیل و حذف نوسان های مقطعی، سطح دقت یا بازه ی زمانی 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 را ذخیره کنیم.

## ۴.۱.۲ رسیدگی به داده های گم شده

با توجه به اینکه نماد های بازار بورس بر اثر افشا ها یا سایر موارد ممکن است روز هایی بسته باشند و داده در آن روز ها موجود نباشد ، ممکن است در تحلیل ها ما را دچار مشکل کنند و این مسئله باید رسیدگی شود. ازین رو از روش جاینبی آخرین مشاهده<sup>۲</sup> برای پر کردن جای خالی این روز ها

---

<sup>1</sup>tgju

<sup>2</sup>last observation carried forward

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 کاهش عملکرد مدل می شوند، 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 $x$ داده

روشی که ما برای رسیدگی به داده های گم شده انتخاب کردیم باعث میشد که مولفه ی  $x$  داده که از جنس تاریخ بود در بعضی 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 مشکلاتی به وجود آورد برای همین به جای نگهداری روز در مولفه ی  $x$  داده مستقیماً ماه مورد نظر را به عنوان مولفه ی  $x$  در نظر گرفتیم. به طور مثال تاریخ ۰۴-۲۰۱۹ به ۰۳-۰۴-۲۰۱۹ تغییر کرد

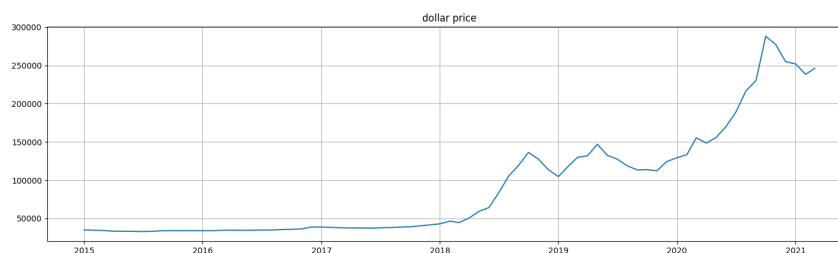
## ۲.۲ تحلیل پیشگو ها

اکنون که داده ها 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 روی 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 داده داریم پیشگویی اجتناب ناپذیر می باشد و بدون تحلیل به مرحله مدل سازی میبریم. از ابزار های ترسیمی آمار جهت تحلیل پیشگو ها استفاده کردیم. به منظور رسم نمودار های مورد نیاز از کتابخانه ی پایتونی matplotlib استفاده کردیم.

## ۱.۲.۲ بررسی تاثیر قیمت دلار

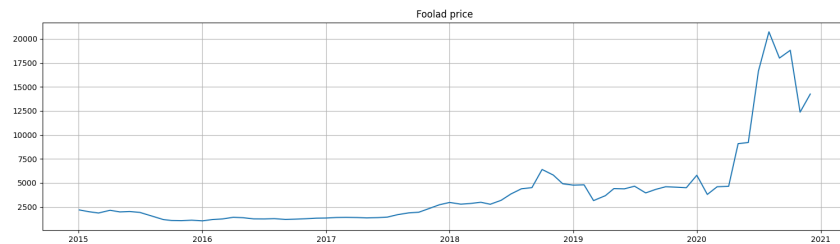
از بین سهم های بازار بورس نماد های فولاد، اخبر و خودرو را به نمایندگی از شاخص های بزرگ جهت تحلیل انتخاب میکنیم. ابتدا نمودار سری زمانی<sup>۳</sup> قیمت دلار را در کنار سهم های مذکور مورد بررسی می نماییم. در شکل ۱.۲ نمودار سری زمانی قیمت دلار به صورت ماهانه ترسیم شده است و در شکل های ۲.۲ و ۳.۲ نمودار های سری زمانی سهم های فولاد و اخبر ترسیم شده است.



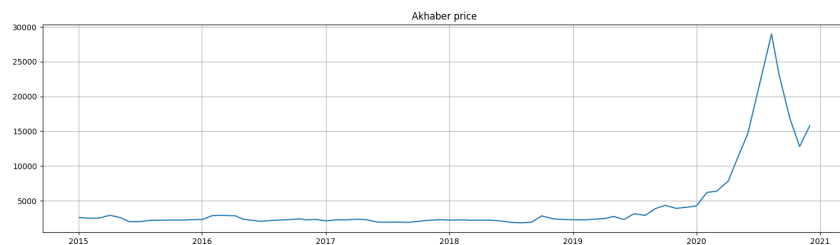
شکل ۱.۲: نمودار قیمت دلار

---

<sup>3</sup>Time series plot



شکل ۲.۲: نمودار قیمت سهم فولاد



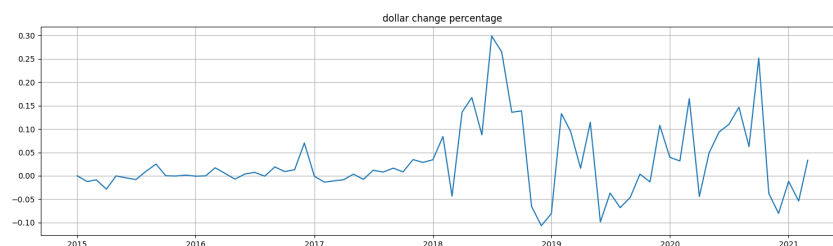
شکل ۳.۲: نمودار قیمت سهم اخابر

همانطور که مشاهده میشود همبستگی بسیار زیاد بین قیمت دلار و قیمت نماد فولاد و نماد اخابر وجود دارد. با رسم نمودار تغییرات این دو نیز دقیقاً به همین 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. در ادامه برای تایید شهود ارائه شده شواهد کافی ارائه خواهد شد.

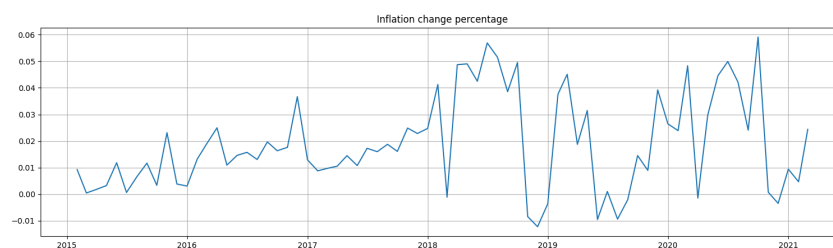
#### بررسی تاثیر تورم

طبق تحلیل شناختی از داده، 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. در شکل ۴.۲ نمودار بازدهی دلار که

به معنای درصد تغییرات قیمت دلار نسبت به ماه قبل می باشد ترسیم شده است و در ادامه در شکل ۵.۲ نمودار بازدهی تورم که به معنای درصد تغییرات تورم ماهانه می باشد ترسیم شده است.



شکل ۴.۲: نمودار درصد تغییر قیمت دلار



شکل ۵.۲: نمودار درصد تغییر تورم

همانطور که مشاهده می شود تا حد خوبی این دو شاخص با هم همبستگی دارند ولی برای اثبات شهود خود از ضریب همبستگی پیرسون استفاده میکنیم. جزئیات بیشتر راجع به همبستگی و ضریب همبستگی پیرسون در کتاب داده کاوی لارس [۱] ذکر شده است. 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 داده‌ی مذکور به عدد ۸۲ صدم با پی مقدار بسیار پایین می‌رسیم که نشان دهنده همبستگی مثبت خوبی هست. با توجه به همبستگی بالای این دو شاخص میتوان از یکی آن



	foolad	akhaber	dollar	inflation	cash
dollar	0.787	0.775	1	0.820	0.013
inflation	0.680	0.710	0.820	1	0.025
cash	0.023	0.033	0.013	0.025	1

جدول ۱.۲: همبستگی پیرسون

ها در مرحله ی طراحی مدل صرف نظر کرد ولی ممکن است به هر حال یک سری اطلاعات را از دست بدهیم چون همبستگی کامل نداشتن این دو شاخص. با توجه به حجم کم داده، شاخص تورم را نیز در مرحله ی طراحی مدل قرار می دهیم.

برای تایید شهود های ارائه شده در این بخش در جدول ۱.۲ مقادیر همبستگی شاخص های مختلف گردآوری شده است.

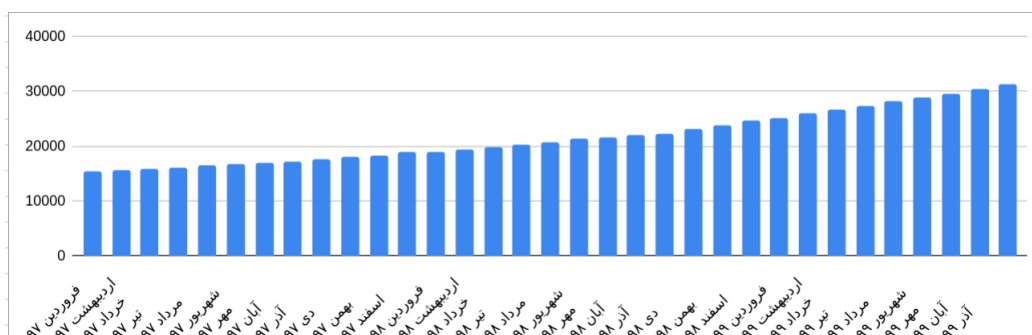
از آنجایی که تعداد اعضا ثابت و بزرگ می باشد مقادیر پی-مقدار<sup>۴</sup> ها برای عدد های همبستگی نزدیک ۱ به شدت به صفر نزدیک می باشد و برای عدد های نزدیک به صفر عددی به شدت نزدیک به ۱ می باشد در نتیجه معنا داری اعداد بالا بدیهی می باشد.

در ادامه با بررسی نمودار نقدینگی علت عدد پایین همبستگی هم متوجه می شویم همچنین با توجه به توضیحات قسمت قبل مبنی بر اینکه دلار و تورم باید یکدیگر را تایید نمایند با مشاهده عدد همبستگی نزدیک ۸ دهم، این جمله تایید میشود.

---

<sup>4</sup>p-value

## 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 ارزشمندی 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 کم بودن حجم داده تصمیم بر این شد که نقدینگی هم وارد مرحله مدل سازی شود که مطمئن شویم اطلاعاتی از دست نمی رود و تحلیل بیشتر این شاخص را به مدل می سپاریم.

## ۳.۲ انتخاب مدل پایه

بعد از تحلیل و انتخاب پیشگو ها با توجه به جنس و خصیصه <sup>۵</sup> های داده 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 داده از جنس سری زمانی می باشد پس باید از بین مدل های مخصوص سری

<sup>۵</sup>Feature

زمانی مدلی انتخاب کنیم که خصیصه های سری های زمانی را به خوبی مورد توجه قرار دهد. 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

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

## ۱.۳.۲ مدل Prophet

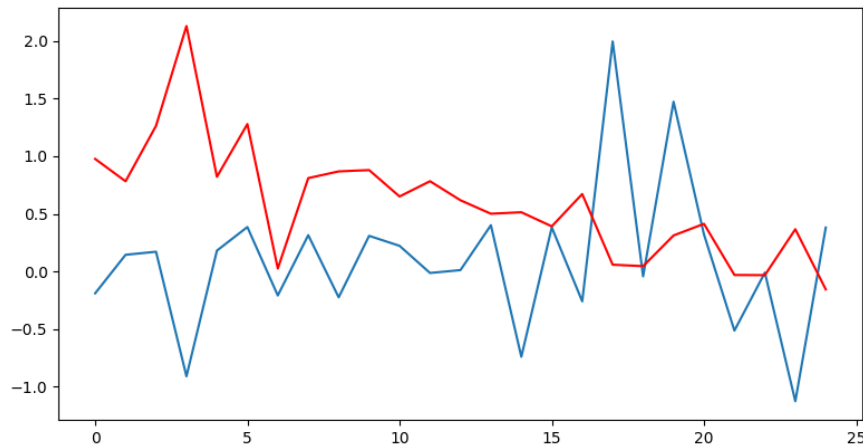
این مدل بر پایه ی مدل جمعی<sup>۶</sup> طراحی شده است که مدلی رگرسیونی ناپارامتری<sup>۷</sup> می باشد. این مدل بر اساس فصلی بودن<sup>۸</sup> داده تصمیم میگیرد و به دنبال الگو های تکرار شونده می باشد. بر اساس تست های اولیه ای که با این مدل گرفتیم دقت بسیار پایینی دریافت کردیم که با توجه به جمله ی قبلی خیلی دور از انتظار نبود.

---

<sup>۶</sup>Additive model

<sup>۷</sup>Nonparametric regression

<sup>۸</sup>seasonality



شکل ۷.۲: نمودار مقایسه پیش بینی بازدهی سهم فولاد و بازدهی واقعی آن توسط مدل prophet

در شکل ۷.۲ خط قرمز رنگ بازدهی پیش بینی شده توسط مدل prophet و خط آبی رنگ بازدهی واقعی آن می باشد. همانطور که مشاهده می کنید، مدل خطای بالایی در پیش بینی داشته است و حتی در تشخیص روند<sup>۹</sup> دچار خطا شده است و الگو رو به پایین پیش بینی کرده است. طبق محاسبه ی میانگین مربعات خطا<sup>۱۰</sup> به عدد نزدیک ۳ میرسیم که خطای بسیار زیادی میباشد.

## ۲.۳.۲ مدل LSTM

این مدل بر پایه ی شبکه ی عصبی بازگشتی<sup>۱۱</sup> طراحی شده است و در مسائل الگویابی کاربرد. جزئیات بیشتر راجع به این مدل را میتوانید در گزارش دکتر جیسون براونی [۴] مطالعه فرمایید. از آنجایی که این مدل بسیار پیچیده است و مباحث پایه ای استفاده شده در آن خارج از مباحث داده

<sup>9</sup>trend

<sup>10</sup>mean squared error

<sup>11</sup>Recurrent neural network

کاوی مقدماتی می باشد، تلاشی برای تست این مدل به صورت عملی صورت نگرفت و صرفاً در حد تحقیق باقی گذاشته شد.

## ۳.۳.۲ مدل ARIMA

این مدل از ترکیب دو مدل AR یا همان مدل خودبازنگر<sup>۱۲</sup> و MA یا همان مدل میانگین متحرک<sup>۱۳</sup> ساخته شده است. مدل خودبازنگر نوعی از فرایند تصادفی می باشد و بر اساس اینکه داده ها به صورت خطی به گذشته ی خودش وابسته است و مدل MA بر اساس ارتباط بین مشاهده فعلی و خطای ایجاد شده بین پیش بینی مدل و مشاهدات قبلی کار میکند.

مدل ARIMA با استفاده از ترکیب مدل AR و مدل MA برای پیشبینی متغیرهای پیچیده تر در سری های زمانی ساخته شده است. مدل ARIMA مدلی شناخته شده و قدیمی می باشد. برای آشنایی بیشتر با مفاهیم تئوری این مدل مقاله ی پاول نیبولد<sup>[۵]</sup> که در سال ۱۹۸۳ به چاپ رسیده است و همچنین کتاب تحلیل سری زمانی تیزی<sup>[۲]</sup> حاوی مطالب مفیدی می باشد. این مدل دارای ۳ پارامتر می باشد که هر کدام به ترتیب مشخص کننده ی پارامتر مدل خودهمبسته، درجه تفاضل<sup>۱۴</sup> و درجه ی مدل میانگین متحرک می باشد. درجه تفاضل تعداد دفعاتی است که باید داده با داده ی گذشته ی خود تفریق گردد تا سری زمانی ایستا<sup>۱۵</sup> شود. راجع به پارامترهای این مدل و نحوه ی محاسبه ی هرکدام در بخش بعد بیشتر توضیح داده شده است.

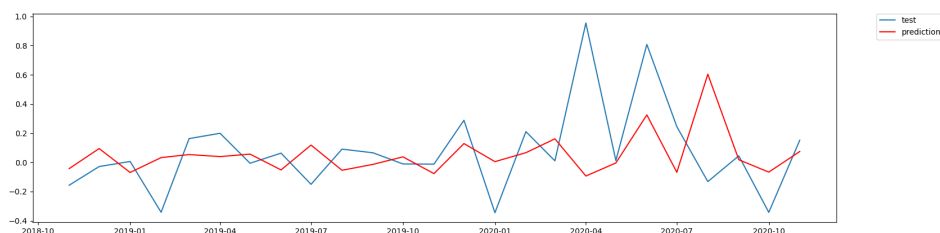
---

<sup>12</sup>Autoregressive model

<sup>13</sup>Moving-average model

<sup>14</sup>Degree of difference

<sup>15</sup>Stationary



شکل ۸.۲: نمودار دقت مدل رو سهم فولاد

همانطور که در شکل ۸.۲ 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 ۲۴ صدم می باشد که به نسبت مدل prophet بسیار کمتر است. در این نمودار خط قرمز رنگ پیش بینی مدل و خط آبی رنگ بازدهی سهم فولاد در ۲ سال ۲۰۲۰ و ۲۰۱۹ می باشد. در بخش بعدی به طور کامل راجع به روند پیاده سازی مدل به کمک ARIMA صحبت خواهد شد.

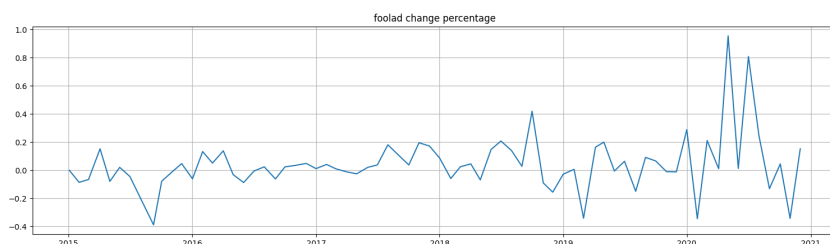
## ۴.۲ پیاده سازی مدل به کمک ARIMA

بعد از نتایج مثبت اولیه که با مدل ARIMA کسب شد، این مدل به عنوان مدل پایه در نظر گرفته شد.

### ۱.۴.۲ تعیین پارامترها

همانطور که گفته شد مدل ARIMA دارای ۳ پارامتر اصلی به ترتیب درجه ی مدل AR ، درجه تفاضل و درجه ی مدل MA می باشد. در گام نخست درجه ی تفاضل مدل را تعیین میکنیم. برا تعیین درجه ی تفاضل ابتدا وضعیت ایستایی سری زمانی را بررسی میکنیم. اگر سری زمانی ایستا باشد درجه تفاضل صفر در نظر گرفته می شود. برای بررسی وضعیت ایستایی سری زمانی ابتدا باید

نمودار سری زمانی را مورد بررسی قرار داد.



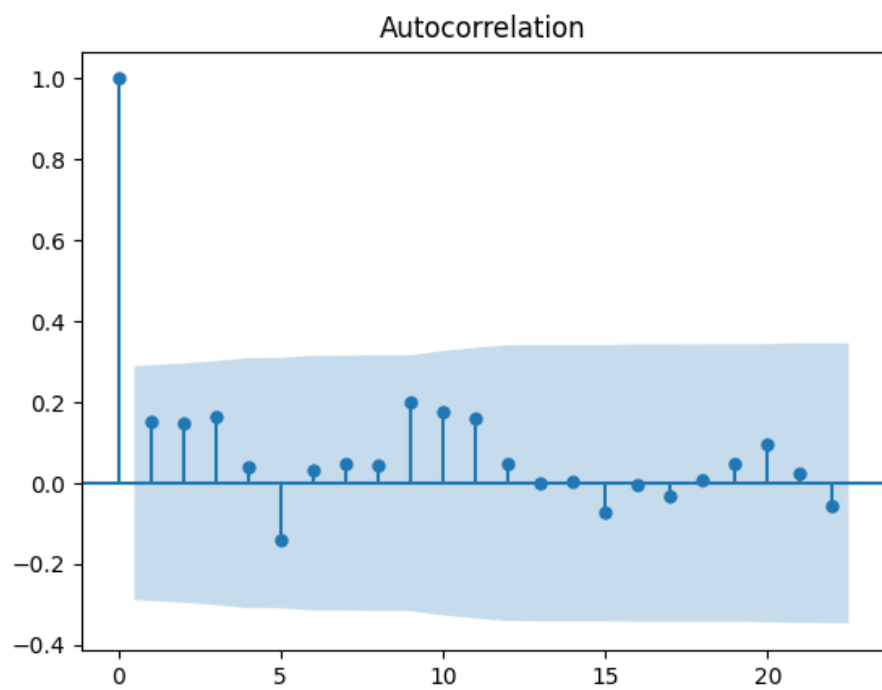
شکل ۹.۲: نمودار نرمال شده ی بازدهی نماد فولاد

با توجه به شکل ۹.۲، سری زمانی به طور شهودی ایستا می باشد و نمیتوان قاطعانه ایستایی سری زمانی را رد کرد به همین دلیل از آزمون دیکی فاولر<sup>۱۶</sup> جهت بررسی ایستایی مدل استفاده میکنیم. فرض صفر در این آزمون غیر ایستا بودن سری زمانی می باشد و فرض یک ایستایی سری زمانی می باشد. با استفاده از تابع پایتونی adfuller آزمون فرض بر روی سری زمانی مذکور اجرا شد و پی-مقدار<sup>۱۷</sup> حدود ۱۳ صدم بدست آمد که با توجه به اینکه بزرگ تر از ۵ صدم می باشد شواهد کافی برای رد فرض صفر وجود ندارد پس در نتیجه سری زمانی ایستا نمی باشد. در گام بعدی باید ایستایی سری زمانی با درجه تفاضل ۱ مورد بررسی قرار گیرد. ازین رو با استفاده از تابع diff سری زمانی جدیدی که حاصل اختلاف هر عضو با عضو قبلی میباشد را تشکیل میدهیم و همین آزمون را مجددا روی آن اجرا میکنیم. در این آزمون پی-مقدار عددی نزدیک صفر می شود و از آنجایی که کمتر از ۵ درصد می باشد میتوانیم به این نتیجه برسیم که سری زمانی بدست آمده ایستا می باشد و میتوانیم درجه تفاضل را برای مدل ۱ در نظر بگیریم.

برای تعیین پارامترهای مدل های MA و AR از نمودارهای ACF و نمودارهای PACF استفاده شده است و به کمک توابع پایتونی plot\_acf و plot\_pacf نمودارهای مذکور ترسیم شده اند.

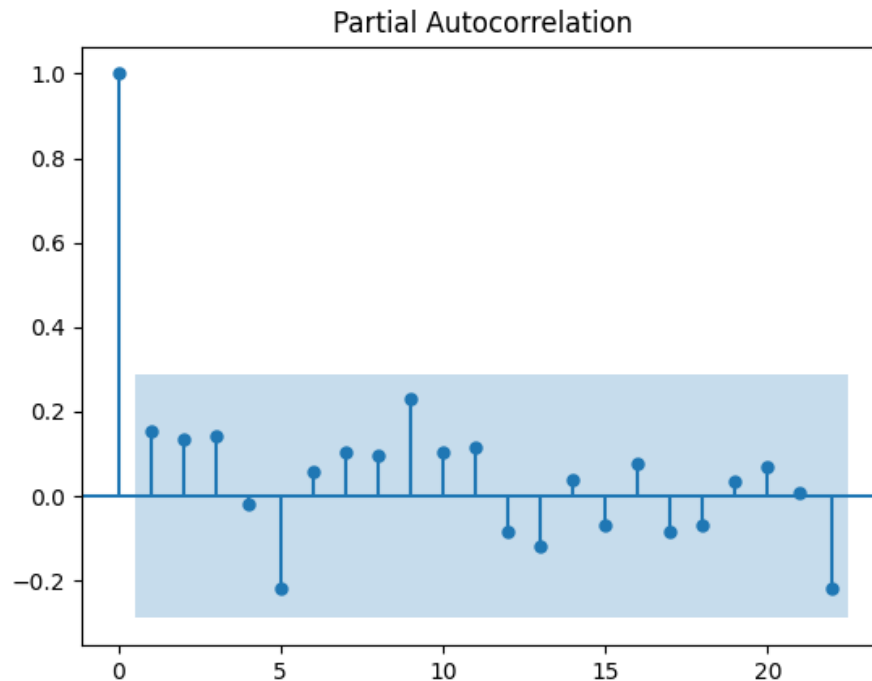
<sup>16</sup>Dickey-Fuller test

<sup>17</sup>p-value



شکل ۱۰.۲: نمودار ACF سری زمانی نماد فولاد





شکل ۱۱.۲: نمودار PACF سری زمانی نماد فولاد

از نمودار PACF جهت تعیین پارامتر مدل AR استفاده شده است همانطور که در نمودار PACF مشخص است از اولین لگ <sup>۱۸</sup> به بعد شاهد برش <sup>۱۹</sup> هستیم و همه ی نقاط در ناحیه ی آبی رنگ که ناحیه معناداری <sup>۲۰</sup> نام دارد قرار دارند با توجه به این موضوع نتیجه میگیریم که مدل روی این سری زمانی اصلاً درجه‌ی مدل AR ندارد.

برای تعیین پارامتر مدل MA از نمودار ACF استفاده می کنیم همانطور که در نمودار ACF مشخص است از اولین تاخیر شاهد برش هستیم و در ناحیه معناداری قرار میگیریم پس نتیجه می‌شود

<sup>18</sup>Lag

<sup>19</sup>Cut off

<sup>20</sup>Significance area

که مدل برای این سری زمانی درجه ی MA هم ندارد. با تکرار تمامی فرایندهای بالا برای سهم های دیگر مثل اخبار و خودرو نتیجه مشابهی بدست می آید. با توجه به اینکه نه پارامتر AR داریم نه پارامتر MA نتیجه میگیریم که این دسته سری زمانی ها همگی نویز سفید<sup>۲۱</sup> هستند و نمیتوانیم با استفاده از مدل AR و مدل MA پیش بینی خیلی دقیقی ازین سری زمانی ها داشته باشیم و مدل ARIMA هم به طبع صرفاً یک مدل تفاضل گیر ساده خواهد شد و تعیین پارامترهای بهینه برای مدل ARIMA کار ساده ای نخواهد بود. در بخش بعدی سعی میشود با اثر دهی بقیه پیشگو ها و بررسی حالت های مختلف برای پارامتر مدل ARIMA به بهترین دقت ممکن برای پیشبینی برسیم.

## ۲.۴.۲ اثردهی پیشگو ها

همانطور که گفته شده است هدف، پیشبینی قیمت سهم به کمک تاریخچه ی خود سهم و پیشگوهای اقتصادی ذکر شده می باشد. سری زمانی هر سهم حاوی اطلاعات تاریخچه ی سهم می باشد که با وارد کردن سری زمانی به مدل، این اطلاعات در دسترس مدل قرار میگیرند ولی برای شاخص های اقتصادی نیاز می باشد که اطلاعات این پیشگو ها به نحوی در اختیار مدل قرار بگیرند. مدل ARIMA یک سری زمانی اضافه تحت عنوان متغیر بیرونی<sup>۲۲</sup> در حین آموزش و پیشبینی دریافت میکند و به عنوان پیشگو اثر گذاری متغیر دریافتی را با سری زمانی اصلی که دریافت کرده بررسی میکند و در پیشبینی سری زمانی اعمال میکند. در پیاده سازی پایتون، این متغیر را با عنوان exog دریافت میکند که معادل آن در زبان R ، xreg می باشد.

در پیاده سازی پایتون ARIMA، مدل فقط یک سری زمانی تحت عنوان متغیر بیرونی دریافت میکند و در صورت وجود چندین پیشگو رسیدگی به این موضوع به عهده ی استفاده کننده می باشد. برای حل این چالش در این پروژه از ترکیب خطی بازگشت پذیر سری های زمانی پیشگو های اقتصادی استفاده شده است. بدین شکل که از ترکیب خطی سری زمانی نرمال شده ی قیمت دلار

<sup>21</sup>White noise

<sup>22</sup>Exogenous variables

، شاخص تورم و نقدینگی یک سری زمانی جدیدی بدست آمده است و به مدل داده شده است. از آنجایی که با نرمال سازی انجام شده واحد و مقیاس همه ی پیشگو های اقتصادی یکسان شده است این نوع ترکیب خطی از نظر علمی صحیح می باشد و کافی است ضرایب هر شاخص به طور صحیحی انتخاب شود. با توجه به جدول ۱.۲ و خروجی اولیه مدل ترکیب خطی زیر به عنوان متغیر بیرونی به مدل تحویل داده شد.

$$0.5 * dollar + 0.5 * inflation - 0 * cash\_values$$

در فصل بعدی نتایج بدست آمده ازین مراحل مورد بررسی قرار خواهد گرفت.

## فصل ۳

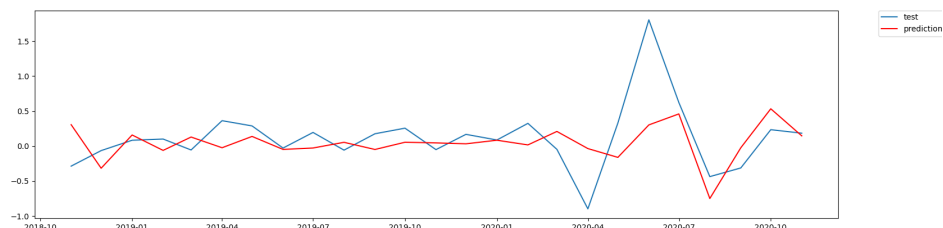
# تحلیل نتایج و جمع بندی

### ۱.۳ نتایج بدست آمده از مدل ARIMA

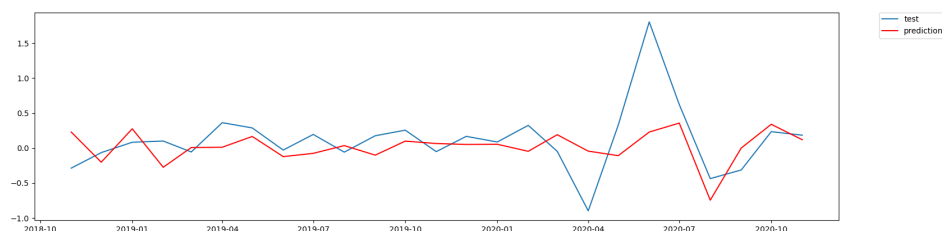
ابتدا نتایج بدست آمده از مدل را مشاهده کنیم و در آخر تحلیل نتایج را ببینیم. دو حالت برای انجام تست روی مدل امکان پذیر بود. در هر دو حالت ابتدا داده به دو قسمت آموزشی و آزمایشی تقسیم میشود که در تست های انجام شده ی ما ۶۶ درصد داده ها از ابتدای سری زمانی به عنوان داده ی آموزشی انتخاب شدند و ما بقی به عنوان 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 مقدار واقعی عضوی که مدل پیشبینی کرده است را به عنوان مشاهده به مجموعه داده های آموزشی 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

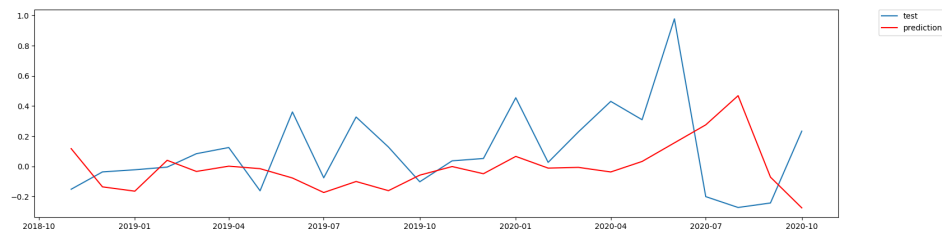
چند نمونه از نتایج به دست آمده توسط مدل طراحی شده در ادامه آمده است.



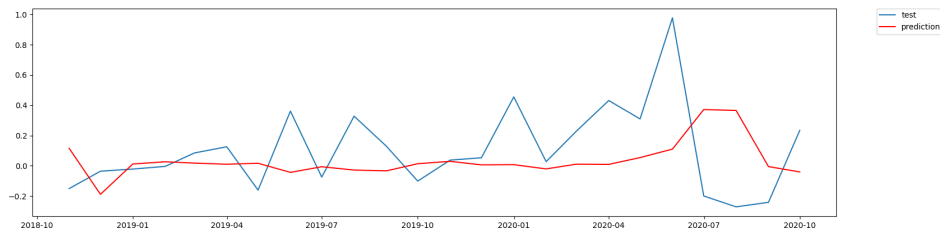
شکل ۱.۳: نمودار دقت مدل روی سهم خودرو با پارامترهای (۲، ۱، ۲)



شکل ۲.۳: نمودار دقت مدل روی سهم خودرو با پارامترهای (۳، ۰، ۴)



شکل ۳.۳: نمودار دقت مدل روی سهم اخبار با پارامترهای (۳، ۰، ۴)



شکل ۴.۳: 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 می دهیم و از آنجایی که این پیشبینی خطا دارد رفته رفته خطا به نقاط جلو تر باز نشر 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 مدلمان را با توجه به این روش بهینه کنیم. در شکل ۴.۳ همانطور که مشاهده میکنید پارامتر مدل میانگین متحرک را برابر صفر قرار دادیم و مدل صرفاً به مدل ساده ی خودهمبسته شد و دقت نهایی کاهش پیدا کرد.

در گام اول تست، بدون در نظر گرفتن نمودارهای ACF و PACF و صرفاً با آزمون و خطا سعی شد که به بهترین دقت برسیم، سپس به کمک نمودارهای نام برده شده سعی کردیم پارامترهای بهینه برای مدل ARIMA را بدست آوریم که با شکست مواجه شدیم در نتیجه مجدداً به گام اول بازگشتیم و سعی کردیم به کمک آزمون و خطا و یا استفاده از تابع `auto_arima` که با بررسی حالت های مختلف سعی میکند به بهترین پارامتر ممکن برسد، به بهترین دقت برسیم. چند نمونه از نتایج بدست آمده در جدول ۱.۳ آمده است.

	foolad	akhaber	khodro
(0,1,0)	0.325	0.330	0.477
(0,1,2)	0.263	0.258	0.374
(1,1,0)	0.267	0.304	0.534
(3,1,3)	0.241	0.266	0.343
(1,1,1)	0.272	0.261	0.445
(1,1,2)	0.272	0.268	0.356
(2,1,2)	0.264	0.272	0.346
(2,1,0)	0.255	0.316	0.407

جدول ۱.۳ : MSE

همانطور که مشاهده می‌شود برای سهم‌های مختلف پارامترهای مختلفی بهترین دقت را ارائه می‌دهند و قانون کلی برای تمامی سهم‌ها نمی‌توان ارائه کرد. نکته‌ای که در آخر لازم به ذکر می‌باشد این است که تاکنون مدل‌های بسیاری برای پیش‌بینی قیمت سهم‌ها ارائه شده است ولی هیچ‌کدام به گونه‌ای دارای دقت نبوده‌اند که برای سرمایه‌گذاری امن پیشنهاد شوند. عوامل بسیار زیادی بر قیمت سهم تاثیرگذار هستند که از چشم مدل‌ها مخفی می‌مانند، از جمله عملکرد شرکت‌ها که توسط کامپیوتر خیلی قابل پیش‌بینی نمی‌باشد.

## واژه‌نامه فارسی به انگلیسی

ARIMA model.....	مدل میانگین متحرک همبسته یکپارچه
Python programming language .....	زبان برنامه نویسی پایتون
Git.....	گیت
Predictor.....	پیشگو
Inflation.....	تورم
Cash .....	نقدینگی
Price history .....	تاریخچه قیمت
Time series .....	سری زمانی
Closing price.....	قیمت پایانی
Missing data .....	داده ی گم شده
Last observation carried forward .....	جاینبهی آخرین مشاهده
Scatter plot.....	نموداراسکتر
Feature.....	خصیصه
Additive model.....	مدل جمعی
Nonparametric regression .....	رگرسیون ناپارامتری
Seasonality .....	فصلی بودن



Trend .....	روند
Mean squared error .....	میانگین مربعات خطا
Recurrent neural network .....	شبکه عصبی بازگشتی
Autoregressive model .....	مدل خود همبسته
Moving-average model .....	مدل میانگین متحرک
Degree of difference .....	درجه تفاضل
Dickey-Fuller test .....	آزمون دیکی فاولر
P-value .....	پی-مقدار
Lag .....	تاخیر
Cut off .....	برش

## 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] Time Series Data Analysis For Stock Market Prediction Using Data Mining Techniques With R, by Mahantesh Angadi
- [4] <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>, by Jason Brownlee
- [5] ARIMA model building and the time series analysis approach to forecasting, by Paul Newbold, 1983

## **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

# Analysis of 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