به نام خدا

دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده برق

مبانی سیستم های هوشمند

گزارش پروژه پایانی

لینک کولب کدها

سیده زهرا عربی

۴۰۰۰۷۱۷۳

استاد : آقای دکتر مهدی علیاری

بهمن ۱۴۰۳

فهرست مطالب

عنوان شماره صفحه

چکیده

در این پروژه، پیش‌بینی قیمت سهام با استفاده از مدل‌های مختلف یادگیری ماشین و شبکه‌های عصبی مورد بررسی قرار گرفته است. داده‌های مورد استفاده شامل قیمت‌های تاریخی سهام، شاخص‌های کلان اقتصادی، احساسات بازار و سایر ویژگی‌های مرتبط است که از منابع معتبر جمع‌آوری شده‌اند. پس از انجام فرآیند پیش‌پردازش داده‌ها، مدل‌های LSTM، GRU، CNN-LSTM، Transformer، MLP و سیستم نروفازی (ANFIS) برای پیش‌بینی قیمت سهام مورد آزمایش قرار گرفتند.

بخش ۱ - جمع‌آوری دیتا

هدف این پروژه پیش‌بینی قیمت سهام در بازار آمریکا است. قیمت سهام تحت تاثیر عوامل متعددی قرار دارد که شامل قیمت‌های تاریخی سهام، احساسات بازار، شاخص‌های کلان اقتصادی و داده‌های اختیارات سهام می‌شود .

قیمت‌های تاریخی سهام نشان‌دهنده رفتار گذشته سهم است و اطلاعاتی در مورد روندهای کوتاه‌مدت و بلندمدت آن فراهم می‌کند. این داده‌ها به تحلیلگران کمک می‌کند تا تغییرات قیمت در گذشته را بررسی کرده و پیش‌بینی‌هایی برای آینده انجام دهند .

احساسات بازار از تحلیل اخبار و رویدادهای مالی ناشی می‌شود و می‌تواند به‌طور قابل توجهی بر قیمت سهام تاثیر بگذارد. اخبار مثبت یا منفی می‌توانند باعث تغییرات فوری در قیمت سهام شوند، زیرا سرمایه‌گذاران به سرعت به اخبار واکنش نشان می‌دهند .

شاخص‌های کلان اقتصادی مانند تولید ناخالص داخلی(GDP) ، نرخ بیکاری و شاخص قیمت مصرف‌کننده (CPI) نیز تاثیر زیادی بر بازار دارند، زیرا این شاخص‌ها وضعیت کلی اقتصاد را نشان می‌دهند. به‌طور مثال، رشد اقتصادی و کاهش نرخ بیکاری می‌تواند منجر به افزایش قیمت سهام شود، در حالی که رکود اقتصادی یا تورم بالا می‌تواند برعکس عمل کند .

اختیارات سهام، که شامل معاملات خرید و فروش سهام با قیمت‌های از پیش تعیین‌شده در آینده است، می‌تواند نشانه‌ای از پیش‌بینی‌های فعالان بازار در مورد آینده قیمت‌ها باشد. افزایش حجم معاملات اختیارات خرید (Call) می‌تواند به معنای انتظار برای افزایش قیمت سهام باشد، در حالی که افزایش حجم معاملات اختیارات فروش (Put) نشان‌دهنده نگرانی نسبت به کاهش قیمت سهام است.

به‌طور کلی، تمامی این عوامل به‌صورت هم‌زمان بر قیمت نهایی سهام تاثیر می‌گذارند و درک ارتباط آن‌ها به تحلیل دقیق‌تر روندهای بازار کمک می‌کند.

برای انجام این پیش‌بینی، داده‌های مختلفی جمع‌آوری شده که شامل داده‌های قیمت سهام، احساسات بازار، شاخص‌های کلان اقتصادی، داده‌های اختیارات سهام و شاخص‌های کلی بازار است. این داده‌ها از منابع معتبر مانند Yahoo Finance ، Finnhub API، Alpha Vantage و FRED API جمع‌آوری می‌شوند.

قیمت سهام

Yahoo Financیکی از منابع اصلی برای دریافت داده‌های سهام است. داده‌های قیمت سهام و شاخص‌های بازار از این پلتفرم به‌دست می‌آید.

- **منبع داده‌ها** :از کتابخانه yfinance برای دانلود داده‌های تاریخی سهام از Yahoo Finance استفاده شده است.
- **ویژگی‌ها**:
  - **Close Price**: قیمت پایانی هر روز.
  - **Open Price**: قیمت بازگشایی هر روز.
  - **High Price**: بالاترین قیمت روز.
  - **Low Price**: پایین‌ترین قیمت روز.
  - **Volume**:حجم معاملات سهام در هر روز.
- **تحلیل** :این داده‌ها روند قیمت‌ها و حجم معاملات را در طول زمان نشان می‌دهد.

احساسات بازار

Finnhub یک API برای دسترسی به داده‌های اخبار، تحلیل احساسات بازار و تحلیل‌های مربوط به سهام است.

- **منبع داده‌ها** :داده‌های اخبار بازار از طریق Finnhub API به دست آمده و تحلیل احساسات با استفاده از .TextBlob انجام شده است.
- **ویژگی‌ها**:
  - **Sentiment Score**: نمره احساسات اخبار که می‌تواند مثبت، منفی یا خنثی باشد.
  - **Headlines**: عنوان اخبار که برای تحلیل احساسات استفاده می‌شود.
  - **Daily Sentiment**: میانگین احساسات اخبار برای هر روز.
- **تحلیل** :این ویژگی‌ها به شناسایی تاثیر اخبار و احساسات بازار بر روی قیمت سهام کمک می‌کنند.

شاخص‌ها

FRED یک API متعلق به Federal Reserve Economic Data است که داده‌های اقتصادی کلان مانندGDP ، نرخ بیکاری، و شاخص قیمت مصرف‌کننده (CPI) را فراهم می‌کند.

- **منبع داده‌ها** :از yfinance برای دریافت شاخص‌های بازار و از FRED API برای جمع‌آوری شاخص‌های اقتصادی استفاده شده است.
- **ویژگی‌ها**:
  - **شاخص‌های بازار**:

۱. شاخص **S&P 500** یکی از مهم‌ترین شاخص‌های بورس آمریکاست که شامل ۵۰۰ شرکت بزرگ آمریکایی است. این شاخص نشان‌دهنده وضعیت کلی بازار سهام آمریکا می‌باشد و تاثیر زیادی بر پیش‌بینی قیمت سهام دارد. تغییرات در **S&P 500** می‌تواند به‌عنوان یک سیگنال برای روند کلی بازار سهام در نظر گرفته شود.

۲. **NASDAQ** شاخص بورس است که بیشتر شرکت‌های تکنولوژی را شامل می‌شود. تغییرات این شاخص معمولاً به شدت به بازار فناوری و نوآوری وابسته است.

۳. **Dow Jones** نیز از شاخص‌های بورس است که به‌طور خاص به ۳۰ شرکت بزرگ آمریکایی اشاره دارد و تأثیر زیادی بر روند کلی اقتصاد و بازار سهام دارد.

  - **شاخص‌های اقتصادی**:

۱. **GDP** تولید ناخالص داخلی کشور است که نشان‌دهنده ارزش کل کالاها و خدمات تولید شده در یک کشور در یک دوره زمانی مشخص است. این شاخص نشان‌دهنده وضعیت رشد اقتصادی است.

۲. **Unemployment Rate** نرخ بیکاری و یکی از مهم‌ترین شاخص‌ها برای تحلیل وضعیت بازار کار است. تغییرات در نرخ بیکاری می‌تواند سیگنالی از تغییرات در وضعیت اقتصادی و پیش‌بینی بازار سهام باشد.

۳. **CPI** شاخص قیمت مصرف‌کننده است که نشان‌دهنده نرخ تورم در یک کشور است. این شاخص برای ارزیابی تغییرات قیمت کالاها و خدمات مصرفی در طول زمان استفاده می‌شود.

۴. **Federal Funds Rate** نرخ بهره‌ای است که بانک‌های تجاری باید برای قرض گرفتن پول از بانک مرکزی آمریکا (Federal Reserve) پرداخت کنند. تغییرات در این نرخ تأثیر زیادی بر بازارهای مالی و قیمت سهام دارند.

اختیارات سهام

- **منبع داده‌ها** :استفاده از yfinance برای دریافت داده‌های مربوط به اختیارات سهام.(Options)
- **ویژگی‌ها**:
  - **Call Volume**: حجم معاملات اختیارات خرید.(Call)
  - **Put Volume**: حجم معاملات اختیارات فروش.(Put)
  - **Put/Call Ratio**: نسبت بین حجم معاملات Put و Call.
- **تحلیل** :این ویژگی‌ها به شناسایی روندهای بازار اختیارات و پیش‌بینی جهت احتمالی بازار کمک می‌کنند.

در نهایت تمام ویژگی‌های استخراج شده از منابع مختلف برای ۵ نماد مربوط به شرکت‌های اپل، گوگل، مایکروسافت، آمازون و تسلا درون یک اکسل (هرنماد یک شیت) ذخیره شده است.

در این دیتا اطلاعات ۱۰۰۰ روز متوالی (۶۸۸ روز کاری) جمع آوری و مورد بررسی قرار گرفته است.

بخش ۲ – استراتژی‌های انتخاب شبکه‌های عصبی

برای پیش‌بینی قیمت سهام با استفاده از هوش مصنوعی، چندین روش مختلف وجود دارد که بسته به نوع داده‌ها و اهداف پروژه می‌توانند مفید باشند.

LSTM (Long Short-Term Memory)

LSTM یکی از مدل‌های پیشرفته شبکه‌های عصبی بازگشتی (RNN) است که برای داده‌های زمان‌سری مانند قیمت سهام بسیار مناسب می‌باشد. این مدل قادر است وابستگی‌های بلندمدت در داده‌ها را یاد بگیرد و الگوهای پیچیده‌ای که در طول زمان ایجاد می‌شوند را شبیه‌سازی کند. در داده‌های مالی، که تغییرات قیمت‌ها تحت تاثیر روندهای گذشته است، این مدل می‌تواند اطلاعات گذشته را در حافظه خود نگه دارد و از آن برای پیش‌بینی قیمت‌های آینده استفاده کند. به دلیل توانایی در پردازش داده‌های طولانی و وابسته به زمان، LSTM انتخاب مناسبی برای پیش‌بینی قیمت سهام است، به‌ویژه زمانی که قیمت‌های روزانه و شاخص‌های اقتصادی به‌صورت متوالی مورد بررسی قرار می‌گیرند.

GRU (Gated Recurrent Unit)

GRU نسخه ساده‌تر و سریع‌تر از LSTM است که همچنان قابلیت‌های مشابهی در یادگیری وابستگی‌های بلندمدت دارد. این مدل به‌طور خاص برای حجم داده‌های متوسط و کم، به‌ویژه زمانی که سرعت آموزش مهم است، مناسب است .GRU مصرف حافظه کمتری نسبت به LSTM دارد و به دلیل ساختار ساده‌تر، نسبت به LSTM زمان آموزش کمتری نیاز دارد. اگر حجم داده‌ها زیاد باشد یا محدودیت‌های محاسباتی داشته باشیم، GRU می‌تواند یک انتخاب خوب باشد که کارایی مشابه LSTM را با مصرف منابع کمتر ارائه دهد.

ترکیبCNN-LSTM

مدل ترکیبی CNN-LSTM از دو بخش مختلف تشکیل شده است: بخش CNN برای استخراج ویژگی‌های مهم و الگوهای پیچیده از داده‌های ورودی (مانند تصاویر یا داده‌های چندبعدی) و بخش LSTM برای پیش‌بینی سری زمانی استفاده می‌شود. این ترکیب به‌ویژه برای داده‌هایی که شامل متغیرهای مختلف و ویژگی‌های پیچیده مانند قیمت سهام، حجم معاملات و شاخص‌های اقتصادی است، بسیار مناسب است. بخش CNN قادر است ویژگی‌های مفیدی را از داده‌های ورودی استخراج کند، سپس LSTM این ویژگی‌ها را برای پیش‌بینی وابستگی‌های زمانی و روندهای آینده استفاده می‌کند. این مدل ترکیبی می‌تواند دقت پیش‌بینی را در داده‌های پیچیده و چندبعدی بهبود بخشد.

Transformerبا مکانیزم توجه

مدل Transformerبا مکانیزم توجه (Attention Mechanism) یکی از مدل‌های پیشرفته و موثر برای تحلیل داده‌های زمان‌سری و چندبعدی است. این مدل برخلاف RNN ها که داده‌ها را به‌صورت ترتیبی پردازش می‌کنند، قادر است اطلاعات را به‌طور موازی پردازش کرده و وابستگی‌های بلندمدت و پیچیده میان داده‌ها را شبیه‌سازی کند. در داده‌های مالی که ممکن است شامل متغیرهای مختلف و وابستگی‌های پیچیده باشند، Transformer به دلیل توانایی‌اش در توجه به بخش‌های مختلف داده و پردازش موازی، می‌تواند پیش‌بینی‌های دقیق‌تری ارائه دهد. این مدل به‌ویژه در مواردی که داده‌ها پیچیده هستند و نیاز به مدل‌سازی وابستگی‌های طولانی‌مدت دارند، عملکرد بسیار خوبی دارد.

MLP (Multi-Layer Perceptron)

MLP یک شبکه عصبی پیشرفته است که از چندین لایه متصل به‌هم تشکیل شده و برای شناسایی الگوهای غیرخطی در داده‌ها مناسب است. این مدل به‌ویژه برای پیش‌بینی‌های کوتاه‌مدت، مانند پیش‌بینی قیمت سهام برای روز آینده، مناسب می‌باشد .MLP از ویژگی‌های ورودی (مانند قیمت، حجم معاملات، و شاخص‌های اقتصادی) استفاده کرده و به‌طور غیرخطی بین آن‌ها روابط برقرار می‌کند. این مدل برای تحلیل داده‌هایی که به وابستگی‌های تاریخی بلندمدت نیاز ندارند، مناسب است. با این حال، MLP فاقد حافظه تاریخی است و به همین دلیل نمی‌تواند وابستگی‌های زمانی پیچیده مانند آنچه که در داده‌های مالی معمول است را به‌طور کامل مدل‌سازی کند.

سیستم نروفازی (ترکیب شبکه عصبی و فازی – ANFIS)

مدل سیستم نروفازی ترکیبی از شبکه‌های عصبی و سیستم‌های فازی است که مزایای هر دو روش را در بر دارد. این مدل به ما این امکان را می‌دهد که از قدرت یادگیری شبکه عصبی در کنار تفسیر سیستم فازی استفاده کنیم. در بازار سهام، جایی که داده‌ها به‌طور کامل دقیق نیستند و احتمال وجود عدم قطعیت در تحلیل‌ها وجود دارد، سیستم‌های فازی می‌توانند به‌عنوان یک مدل مفید عمل کنند. این سیستم‌ها با استفاده از قوانین "اگر-آنگاه" می‌توانند دانش کارشناسی را در کنار یادگیری از داده‌ها قرار دهند و نتایج قابل تفسیرتری ارائه دهند.

بخش ۳ - پیش پردازش داده‌ها

چرا باید داده‌ها را قبل از مدل‌سازی پیش‌پردازش کنیم؟

قبل از این‌که مدل یادگیری ماشین را اجرا کنیم، داده‌های خامی که جمع‌آوری می‌شوند، معمولاً دارای مشکلات متعددی هستند که بدون پردازش مناسب، مدل را به نتایج نادرست، ناپایدار و غیرقابل اعتماد می‌رساند. دلایل اصلی نیاز به پیش‌پردازش عبارتند از:

وجود نویز و داده‌های نادرست :داده‌های مالی و اقتصادی ممکن است دارای مقادیر پرت، داده‌های از دست رفته یا اشتباه باشند. این باعث می‌شود که مدل نتواند الگوهای صحیح را شناسایی کند.

عدم مقیاس‌بندی صحیح متغیرها :مقیاس‌های مختلف متغیرها (مثلاً قیمت سهام در مقیاس صدها دلار، ولی نرخ بهره در مقیاس درصد) می‌تواند باعث شود که مدل نسبت به برخی ویژگی‌ها بیش از حد حساس باشد و نسبت به برخی دیگر بی‌تفاوت عمل کند.

وجود همبستگی‌های نادرست و ناهنجاری‌های آماری :برخی از ویژگی‌ها ممکن است همبستگی بالا داشته باشند که باعث افزونگی اطلاعات و کاهش کارایی مدل می‌شود.

وابستگی زمانی در داده‌های سری‌زمانی :داده‌های بازار مالی وابسته به زمان هستند، بنابراین حفظ ترتیب زمانی در پردازش داده‌ها ضروری است .استفاده از تکنیک‌هایی که داده‌ها را تصادفی مرتب کند مثل shuffle می‌تواند باعث خراب شدن دنباله منطقی اطلاعات شود.

لزوم استخراج اطلاعات هدفمند :بسیاری از ویژگی‌ها در شکل خام خود اطلاعات مناسبی برای مدل ارائه نمی‌دهند و نیاز به تبدیل، فیلتر و مهندسی ویژگی دارند.

اجزاء اصلی پیش‌پردازش داده‌ها در مدل‌های پیش‌بینی قیمت سهام

فرآیند پیش‌پردازش معمولاً شامل چندین بخش کلیدی است:

پاک‌سازی داده‌ها (Cleaning)

- بررسی و حذف مقادیر گمشده.(Null)
- حذف یا اصلاح مقادیر پرت با استفاده از آماره‌های مناسب.
- بررسی و حذف داده‌های تکراری که می‌توانند باعث ایجاد وزن نامتعادل در مدل شوند.

تعریف متغیر هدف (Target) و ویژگی‌ها (Features)

- مدل باید یاد بگیرد که بر اساس ویژگی‌های داده‌شده، مقدار متغیر هدف را پیش‌بینی کند.
- اگر متغیر هدف به درستی تعریف نشود، مدل یادگیری مفهوم اشتباهی پیدا می‌کند.
- مشخص کردن ویژگی‌هایی مثل قیمت باز، بسته، بالاترین، پایین‌ترین، حجم معاملات، داده‌های اقتصاد کلان، احساسات بازار و شاخص‌ها و اختیارات سهام.
- متغیر هدف را یک روز جلوتر شیفت (Shift) می‌دهیم، چون می‌خواهیم قیمت بسته شدن فردا را پیش‌بینی کنیم.

مقیاس‌بندی و نرمال‌سازی داده‌ها

- **استانداردسازی (Standardization)** : تبدیل ویژگی‌ها به توزیع نرمال با میانگین صفر و انحراف معیار یک.
- **نرمال‌سازی (Min-Max Scaling)** : تبدیل همه مقادیر به بازه $[0,1]$ تا تغییرات کوچک، تأثیر زیادی روی مدل نداشته باشند.

تبدیل داده‌ها به توالی‌های زمانی(Sequence Creation)

- مدل‌های سری‌زمانی مثل LSTM، GRU و Transformer باید اطلاعات چند روز گذشته را به عنوان ورودی دریافت کنند تا روندها را یاد بگیرند.
- اگر داده‌ها را به شکل توالی‌های متحرک ذخیره نکنیم، مدل هیچ درکی از روند زمانی نخواهد داشت.
- ایجاد دنباله‌ای از روزهای گذشته (مثلاً ۶۰ روز اخیر) برای هر روز جدید.
- خروجی هر توالی، مقدار قیمت بسته شدن روز بعد است.

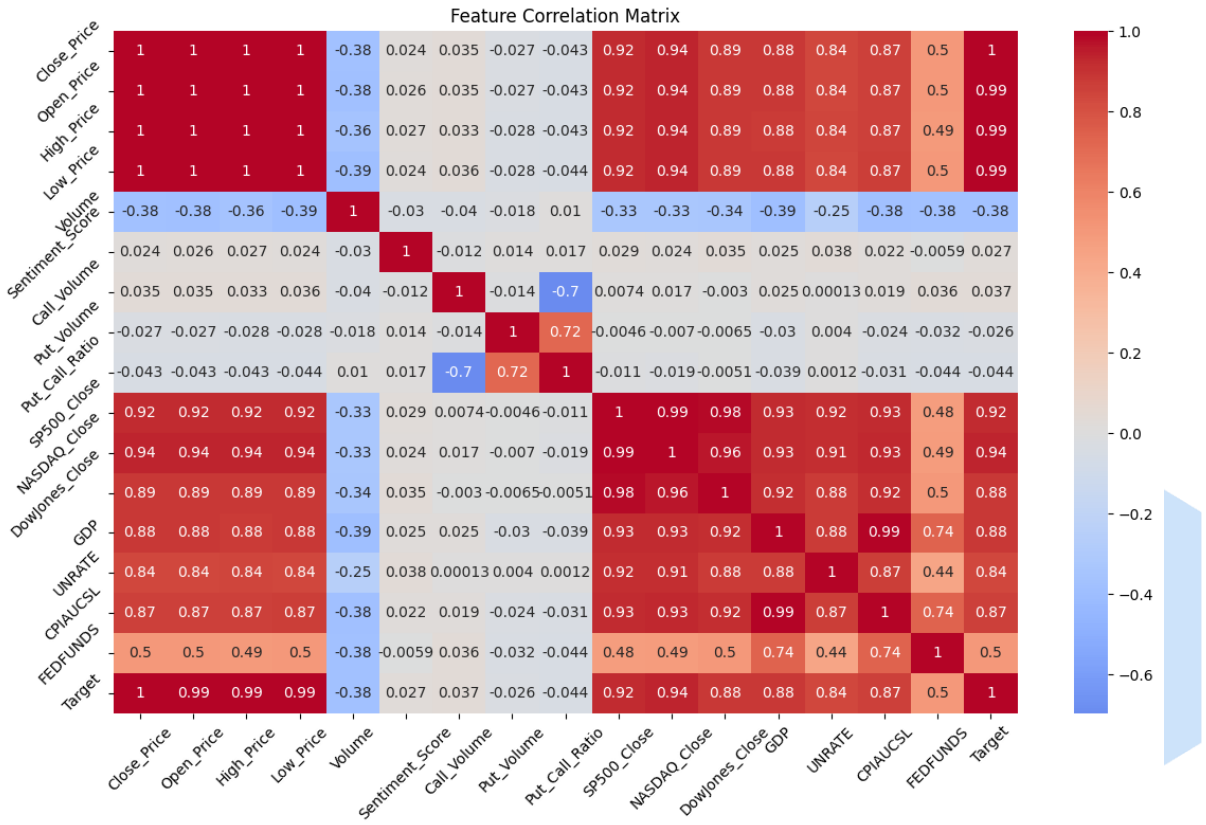
تقسیم داده‌ها به مجموعه‌های آموزشی، اعتبارسنجی و تست

- اگر مدل روی تمام داده‌ها آموزش ببیند، ممکن است فقط آن داده‌ها را حفظ کند و روی داده‌های جدید عملکرد ضعیفی داشته باشد.(Overfitting)
- مجموعه‌های تست و اعتبارسنجی کمک می‌کنند تا مدل روی داده‌های جدید ارزیابی شود.
- ٪۶۰ داده‌ها برای آموزش، ٪۲۰ برای اعتبارسنجی و ٪۲۰ برای تست استفاده می‌شود.
- تقسیم داده‌ها بدون شافل انجام می‌شود تا ترتیب زمانی حفظ شود.

Data shapes:
Training set: (375, 60, 17)
Validation set: (125, 60, 17)
Test set: (125, 60, 17)

تحلیل ماتریس همبستگی و نمودار توزیع ویژگی‌ها

ماتریس همبستگی ویژگی‌ها



این ماتریس همبستگی نشان‌دهنده روابط خطی بین ویژگی‌های مختلف و متغیر هدف در داده‌های مورد استفاده برای پیش‌بینی قیمت سهام است. همبستگی بین دو متغیر عددی بین ۱- تا ۱+ است:

- ۱ + یعنی دو متغیر کاملاً همبسته هستند (با افزایش یکی، دیگری هم افزایش می‌یابد).
- ۱ -یعنی دو متغیر رابطه معکوس دارند (با افزایش یکی، دیگری کاهش می‌یابد).
- ۰ یعنی هیچ رابطه‌ای بین دو متغیر وجود ندارد.

تحلیل همبستگی ویژگی‌ها با متغیر هدف (Target)

متغیر هدف، قیمت بسته شدن روز بعد است. مهم‌ترین عوامل مرتبط با آن:

ویژگی‌های قیمت (Close, Open, High, Low)

- قیمت‌های باز، بسته، بیشینه و کمینه دارای همبستگی بسیار قوی (0.99~) با متغیر هدف هستند. این طبیعی است زیرا قیمت بسته شدن یک روز بسیار نزدیک به قیمت‌های دیگر همان روز است.
- این سطح از همبستگی نشان می‌دهد که مدل باید روی روندهای تاریخی این ویژگی‌ها تمرکز کند.

شاخص‌های بازار (S&P500, NASDAQ, DowJones)

- **S&P500** (۰.۹۲)، **NASDAQ** (۰.۹۴) و **DowJones** (۰.۸۸) با قیمت بسته شدن سهام رابطه قوی دارند.
- این به این معناست که سهام مورد بررسی تحت تأثیر کلی شاخص‌های بازار قرار دارد.

شاخص‌های اقتصادی (GDP, UNRATE, CPI, FEDFUNDS)

- GDP (۰.۸۸)، نرخ بیکاری (۰.۸۴) و CPI (۰.۸۷) تأثیر مثبت روی قیمت سهام دارند.
- نرخ بهره فدرال (**۰.۵**) همبستگی متوسطی دارد، که نشان می‌دهد سیاست‌های پولی بر قیمت سهام تأثیر دارد، اما تأثیر آن به‌اندازه شاخص‌های دیگر نیست.

حجم معاملات و احساسات بازار

- حجم معاملات همبستگی منفی (-۰.۳۸) با قیمت دارد .این نشان می‌دهد که در روزهایی که حجم معاملات بالا است، احتمالاً نوسانات افزایش می‌یابد و ممکن است کاهش قیمت رخ دهد.
- احساسات بازار تقریباً بدون همبستگی (0.02~) **است**. این نشان می‌دهد که احساسات خبری در قالب فعلی اطلاعات زیادی به مدل اضافه نمی‌کند.

تحلیل همبستگی بین ویژگی‌ها

قیمت‌های سهام با هم همبستگی کامل دارند(1.0~)

- این موضوع نشان‌دهنده عدم تنوع اطلاعاتی بین این ویژگی‌ها است.
- از نظر مدل‌سازی، استفاده از همه این ویژگی‌ها به‌صورت همزمان ضروری نیست و ممکن است باعث افزونگی داده‌ها (Multicollinearity) شود.

همبستگی بالای شاخص‌های بازار با یکدیگر(S&P500 ~ NASDAQ ~ DowJones)

- این شاخص‌ها همبستگی شدیدی دارند تا (۰.۹۸ تا ۰.۹۹) که نشان می‌دهد تغییر در یکی از آنها معمولاً در دیگری هم منعکس می‌شود.
- از نظر انتخاب ویژگی‌ها، استفاده از هر سه شاخص ممکن است اطلاعات تکراری به مدل اضافه کند.

حجم معاملات با بیشتر ویژگی‌ها همبستگی منفی دارد (-۰.۳ تا -۰.۴)

- افزایش حجم معمولاً نشانه‌ای از نوسان بالا یا تغییر در احساسات بازار است.
- همبستگی منفی نشان می‌دهد که در شرایطی که حجم زیاد است، کاهش قیمت محتمل‌تر است.

نسبت Put/Call همبستگی نسبتاً قوی با حجم معاملات دارد(0.72)

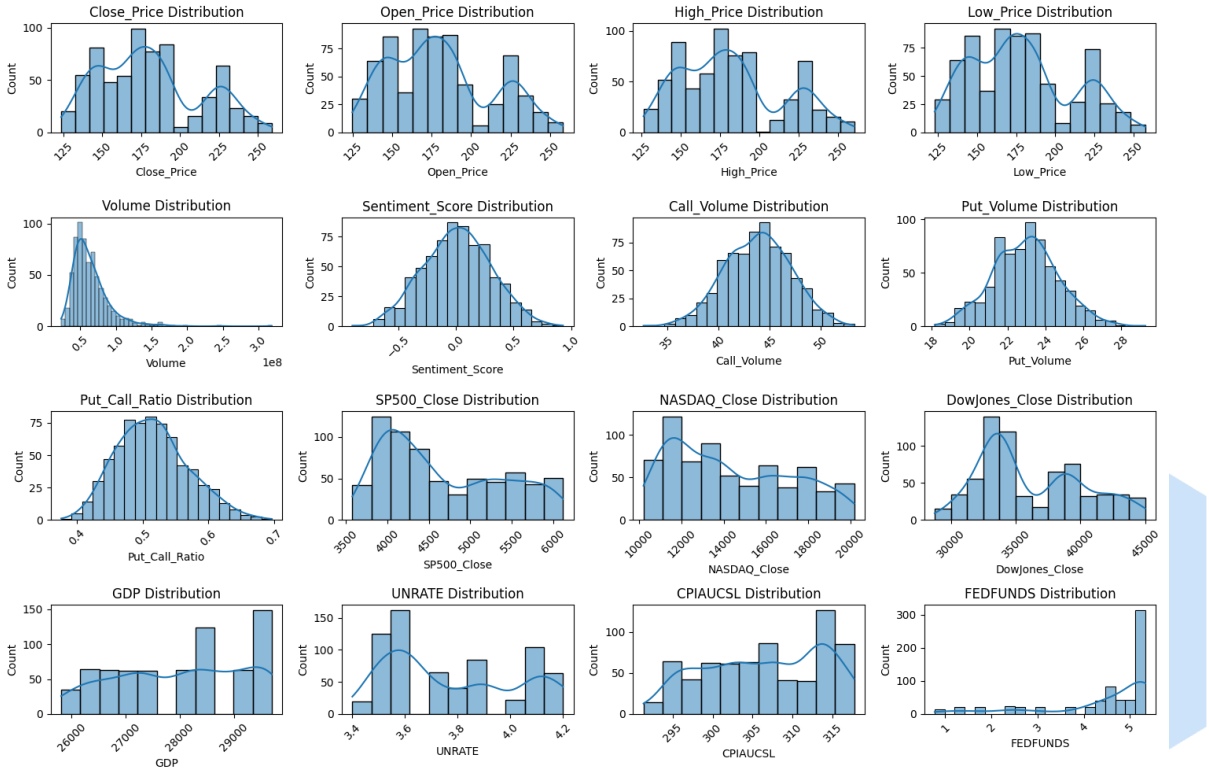
- نسبت Put/Call نشان‌دهنده احساسات سرمایه‌گذاران نسبت به خرید یا فروش سهام است.
- همبستگی ۰.۷۲ نشان می‌دهد که وقتی این نسبت افزایش می‌یابد، احتمالاً حجم کلی معاملات نیز افزایش پیدا می‌کند.

چه نتایجی می‌توان گرفت؟

- ویژگی‌های قیمتی (Close, Open, High, Low) مهم‌ترین متغیرها هستند، اما باید از انتخاب چندتایی آنها اجتناب کرد.
- شاخص‌های کلان اقتصادی و شاخص‌های بازار اطلاعات خوبی ارائه می‌دهند، اما همه آنها ضروری نیستند (مثلاً شاید انتخاب یک شاخص به‌جای سه شاخص بهتر باشد).
- حجم معاملات و نسبت Put/Call می‌توانند نشانه‌های خوبی از نوسانات بازار باشند.

- احساسات بازار (Sentiment Score) در این تحلیل تأثیر خاصی ندارد، اما شاید با روش‌های دیگر پردازش داده‌های متنی، بهبود یابد.

نمودار توزیع ویژگی‌ها



این نمودارها توزیع ویژگی‌های کلیدی در داده‌های مورد استفاده برای مدل‌سازی را نمایش می‌دهند . هر نمودار نشان می‌دهد که مقدار هر ویژگی چگونه در داده‌ها پراکنده شده است . بررسی این توزیع‌ها کمک می‌کند تا متوجه شویم:

- آیا ویژگی‌ها به‌صورت نرمال توزیع شده‌اند یا خیر؟
- آیا ویژگی‌ها دارای داده‌های پرت (Outlier) هستند؟
- آیا نوسانات شدیدی در ویژگی‌ها دیده می‌شود؟
- آیا نیاز **به** تبدیل (Transformation) **یا** مقیاس‌بندی (Scaling) داریم؟

تحلیل ویژگی‌های قیمتی(Close Price, Open Price, High Price, Low Price)

ویژگی‌های قیمت سهام دارای توزیع چند قله‌ای (Multimodal) هستند.

- به جای یک توزیع نرمال یکنواخت، چندین قله در داده‌ها مشاهده می‌شود که احتمالاً نشان‌دهنده دوره‌های مختلف تغییر روند بازار است.
- این نشان می‌دهد که قیمت سهام در چندین بازه قیمتی نوسان داشته و مدل باید روندهای مختلف را یاد بگیرد.
- **راهکار** : می‌توان **از** روش‌های خوشه‌بندی (Clustering) یا فیلتر روندها برای دسته‌بندی این داده‌ها استفاده کرد.

حجم معاملات

توزیع بسیار چوله به راست (Right-Skewed) دارد.

- مقدار حجم معاملات برای اکثر نمونه‌ها در محدوده کم قرار دارد و مقادیر بالا نادر هستند.
- وجود مقادیر پرت (Outliers) در حجم معاملات دیده می‌شود که نشان‌دهنده روزهای خاص با نوسان زیاد است.
- راهکار:
  - استفاده از تبدیل لگاریتمی (Log Transformation) برای متعادل‌سازی داده‌ها.
  - بررسی روزهای پرت برای یافتن دلیل افزایش حجم (مثلاً رویدادهای اقتصادی یا گزارش‌های مالی).

امتیاز احساسات بازار

توزیع نزدیک به نرمال دارد.

- مقدار احساسات از حدود -۱ تا +۱ پراکنده است.
- بیشتر داده‌ها حول مقدار صفر متمرکز شده‌اند که نشان می‌دهد اغلب اخبار خنثی هستند.
- **راهکار:** بررسی نقش این ویژگی در مدل و شاید تغییر روش پردازش متن برای بهبود کیفیت داده‌های احساسات.

حجم معاملات آپشن‌ها

توزیع تقریباً نرمال دارند.

- حجم معاملات قراردادهای خرید (Call) و فروش (Put) دارای توزیع متقارن و نرمال هستند که نشان می‌دهد بیشتر مقادیر حول مقدار میانگین توزیع شده‌اند.
- نسبت Put/Call نیز دارای توزیع تقریباً نرمال است.
- این نشان می‌دهد که استفاده از این ویژگی‌ها در مدل می‌تواند ارزشمند باشد، چون دارای نوسان شدید یا چوله‌گی نیستند.

شاخص‌های بازار

توزیع‌های چند قله‌ای (Multimodal) دارند.

- به دلیل تغییرات در بازار سهام، این شاخص‌ها دارای چندین ناحیه پرتراکم هستند.
- مقادیر پایین شاخص‌ها احتمالاً به دوره‌های نزولی بازار (Bear Market) مربوط هستند، درحالی‌که مقادیر بالاتر مربوط به دوره‌های صعودی (Bull Market) هستند.
- **راهکار:**بررسی این شاخص‌ها با استفاده از تحلیل روند (Trend Analysis) برای جدا کردن فازهای مختلف بازار.

شاخص‌های اقتصادی کلان

توزیع‌های مختلف، از نرمال تا چوله به راست.

- GDP تولید ناخالص داخلی و CPI شاخص قیمت مصرف‌کننده توزیع‌های نسبتاً یکنواخت دارند که نشان می‌دهد این مقادیر در طول زمان رشد تدریجی داشته‌اند.
- نرخ بیکاری دارای یک قله واضح است که نشان می‌دهد بیشتر داده‌ها در یک محدوده خاص قرار دارند.
- نرخ بهره فدرال به شدت چوله به راست است که نشان‌دهنده وجود دوره‌هایی با نرخ بهره بسیار پایین و سپس افزایش شدید است.
- **راهکار:** استفاده از تبدیل‌های آماری برای بهبود پخش مقادیر و بررسی اثر تغییرات نرخ بهره روی بازار سهام.

چه نتایجی می‌توان گرفت؟

- ویژگی‌های قیمت دارای توزیع چند قله‌ای هستند که نیاز به تحلیل روند دارند.
- حجم معاملات و نرخ بهره دارای مقادیر پرت (Outliers) هستند که نیاز به مقیاس‌بندی دارند.
- امتیاز احساسات توزیع نرمال دارد و احتمالاً باید بهتر پردازش شود تا اطلاعات بیشتری ارائه دهد.
- شاخص‌های بازار دارای تغییرات زیاد هستند که می‌توانند نشانه‌های خوبی از روندهای کلی باشند.

بهبود پیش‌پردازش داده‌ها

برای بهبود کیفیت داده‌ها و افزایش دقت پیش‌بینی مدل تغییرات مهمی اعمال شده است.

۱. حذف مقادیر ناموجود (Null) با روش Interpolation : داده‌های از دست رفته بر اساس مقدار قبلی و بعدی تخمین زده می‌شوند.
۲. تبدیل لگاریتمی برای بهبود حجم معاملات (Volume, Call Volume, Put Volume) : برای متعادل‌سازی توزیع داده‌هایی که چوله به راست هستند.
۳. ایجاد ویژگی‌های جدید:
    - توزیع تغییرات قیمت(Price_Change)
    - توزیع دامنه قیمت(Price_Range)
    - توزیع تغییرات حجم(Volume_Change)
    - توزیع میانگین متحرک ۲۰ روزه(MA20)
    - توزیع نوسانات(Volatility)
    - توزیع روزهای هفته(DayOfWeek)
۴. حذف ویژگی‌های با همبستگی بالا:
    - از بین قیمت‌ها فقط Close و High حفظ شده است.
    - از بین شاخص‌ها فقط SP500 حفظ شده است.
    - از بین MA ها فقط MA20 حفظ شده است.
۵. حفظ ویژگی‌های مهم با همبستگی کم :
    - Volume Logو مشتقات آن
    - Put Call Ratio
    - شاخص‌های اقتصادی کلیدی
    - ویژگی‌های تکنیکال جدید
    - Sentiment Scor، DayOfWeek

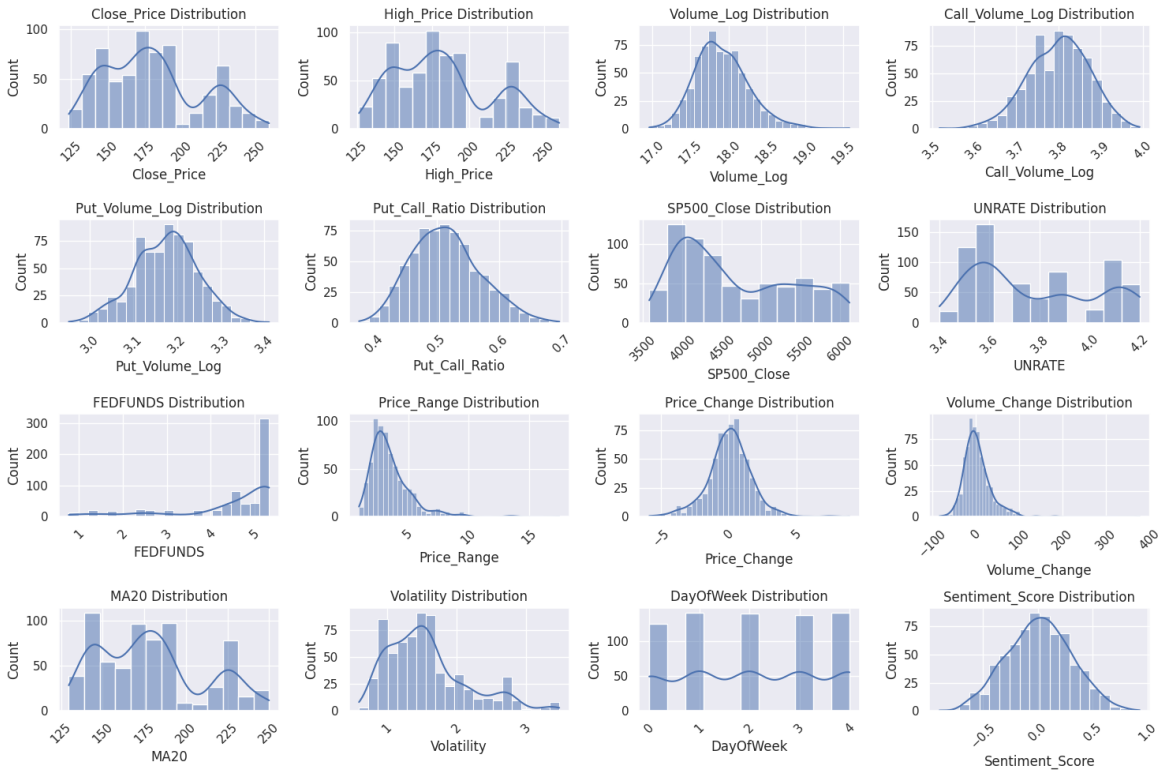
بهبود مقیاس‌بندی داده‌ها

۱. استفاده از RobustScaler به جای StandardScaler : مقاومت بیشتر در برابر داده‌های پرت.
۲. اســـتفاده از MinMaxScaler پس از RobustScaler برای تبدیل داده‌ها به بازه [0,1] پس از حذف اثر مقادیر پرت.

بهبود پردازش داده‌های سری‌زمانی

۱. ایجاد دنباله‌های ۶۰ روزه از داده‌ها : مدل برای پیش‌بینی، ۶۰ روز گذشته را مشاهده می‌کند.
۲. اســـتفاده از ویژگی‌های جدید مانند "Day of Week Encoding" برای اضـــافه کردن اطلاعات مربوط به روز هفته به عنوان یک فاکتور موثر در پیش‌بینی‌ها.

نمودار توزیع ویژگی‌ها پس از بهبود پیش پردازش



۱. توزیع Close Price و High Price همچنان چندقله‌ای است.

- این نشانه وجود فازهای مختلف در روند بازار است و باید در مدل لحاظ شود.

۲. Volume Log، Call Volume Logو Put Volume Log دارای توزیع متعادل‌تر هستند.

- تبدیل لگاریتمی باعث کاهش چوله بودن توزیع شده و این برای مدل مفید است.

۳. Put Call Ratioتوزیع نسبتاً نرمالی دارد.

- این نشان می‌دهد که این ویژگی ممکن است برای مدل مفید باشد.

٤. SP500_Closeدارای دو قله مشخص است.

- احتمالاً این شاخص در دو فاز مختلف بازار قرار دارد (رکود و صعود).

٥. نرخ بیکاری و نرخ بهره فدرال همچنان توزیع خاصی دارند.

- نرخ بیکاری دارای نوسانات دوره‌ای مشخص است.
- نرخ بهره فدرال همچنان چوله به راست است، که طبیعی است.

٦. Price_Change و Volume_Change دارای توزیع تقریباً نرمال هستند.

- این ویژگی‌ها برای تشخیص نوسانات بازار مفید خواهند بود.

۷. Volatilityهمچنان دارای چوله به راست است.

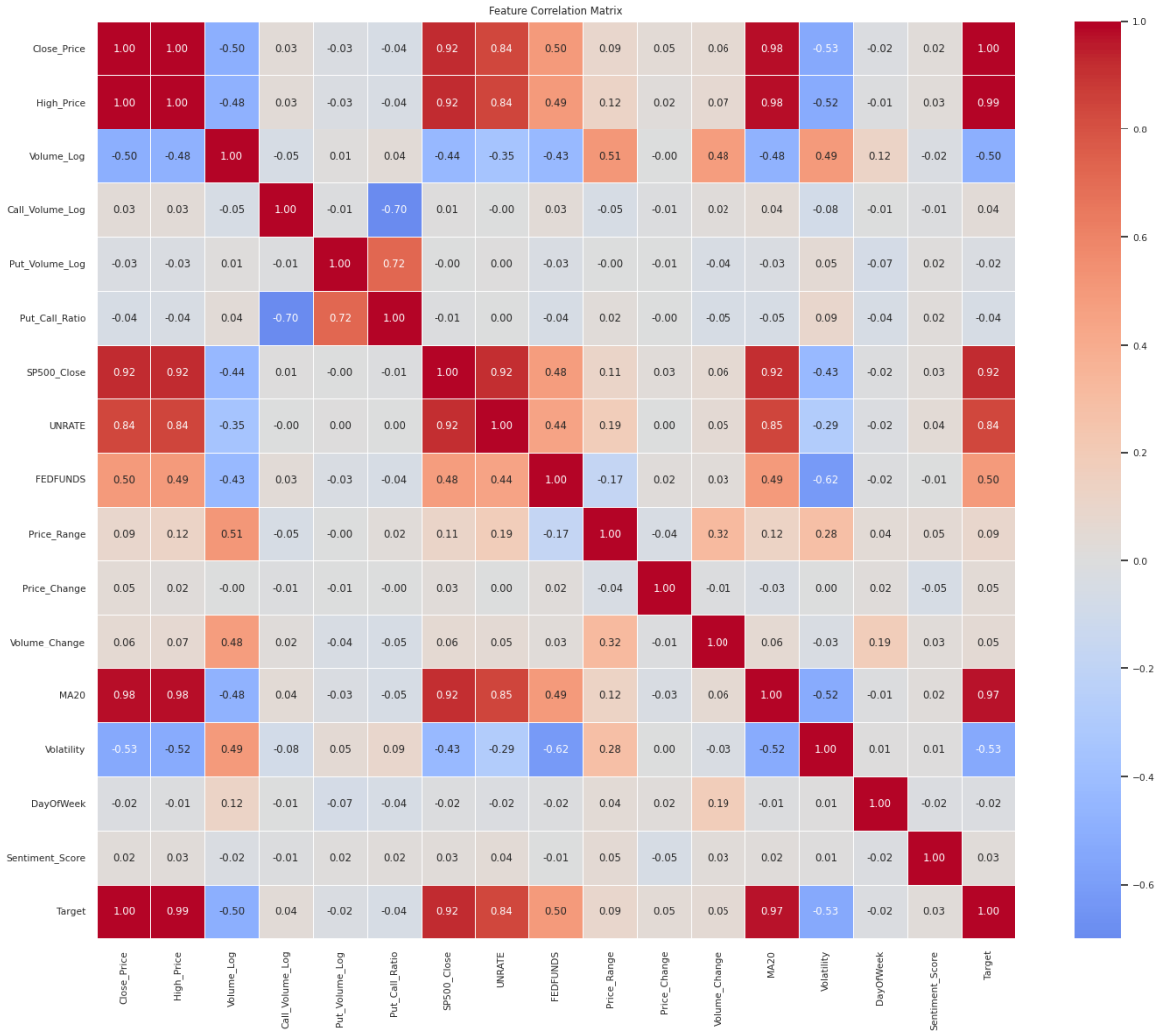
۸. DayOfWeekبه‌صورت توزیع یکنواخت قرار گرفته است.

- یعنی همه روزهای هفته به‌طور یکسان در داده‌ها توزیع شده‌اند.

۹. Sentiment Scoreدارای توزیع تقریباً نرمال است.

- اما همبستگی آن با قیمت سهام همچنان کم است.

ماتریس همبستگی میان ویژگی‌ها پس از بهبود پیش پردازش



۱. همبستگی شدید بین Close Price و High Price . (~1.00)

- این دو متغیر تقریباً معادل هم هستند.

۲. Volume Logو Close Price همبستگی منفی دارد(-0.50)

- نشان‌دهنده این است که افزایش حجم معاملات اغلب با کاهش قیمت همراه است.

۳. Put Call Ratioو Volume Log همبستگی نسبتاً قوی (~۰.۷۰) دارند.

- افزایش قراردادهای Put نسبت به Call معمولاً همراه با افزایش حجم معاملات است.

٤. SP500 Closeهمچنان همبستگی بالا با Close Price دارد. (~0.92)

- یعنی حرکت این شاخص تأثیر زیادی بر قیمت سهام دارد.

٥. نرخ بیکاری و نرخ بهره فدرال همبستگی زیادی با Close Price دارند (~۰.۸۴ و ~۰.۵۰).

- نرخ بیکاری و نرخ بهره همچنان فاکتورهای اقتصادی مهمی در قیمت‌گذاری هستند.

٦. Volatilityدارای همبستگی منفی قوی با Close Price است. (~-0.53)

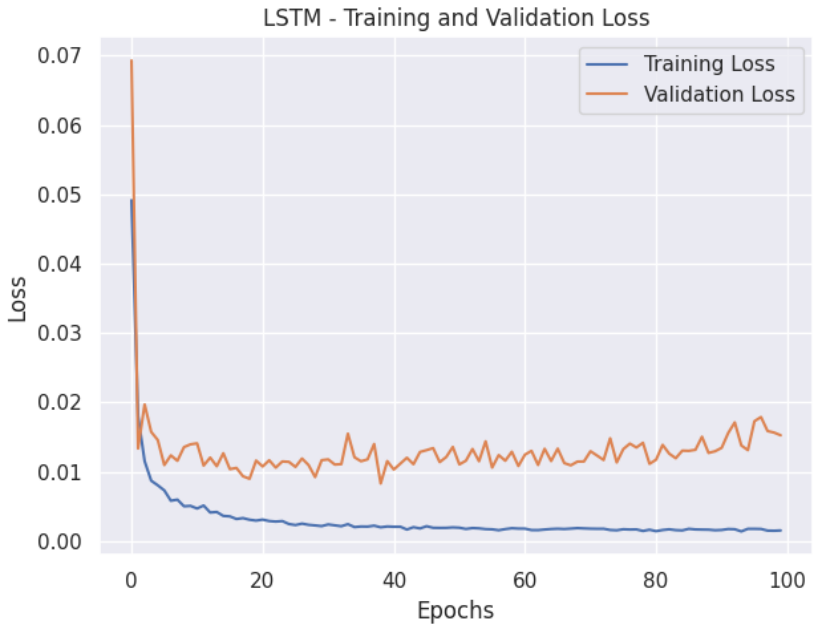
- این نشان می‌دهد که وقتی نوسانات زیاد می‌شود، قیمت‌ها معمولاً افت می‌کنند.

٧. Sentiment Scoreهمچنان همبستگی ضعیفی با Close Price دارد.(~0.02)

بخش ۴ – بررسی عملکرد شبکه‌ها

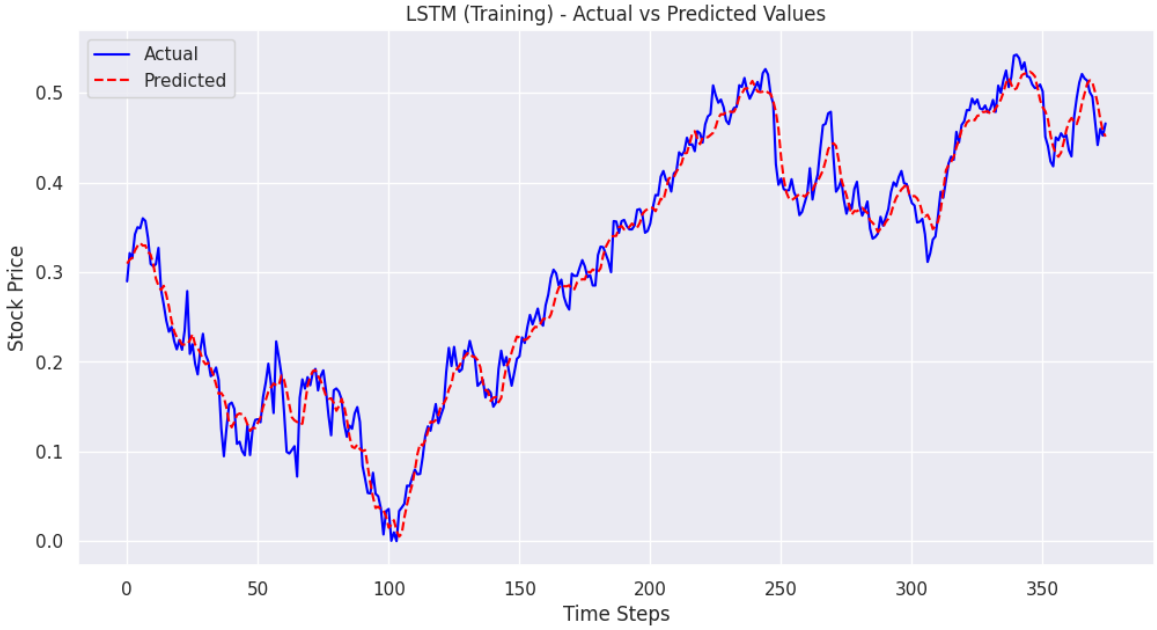
LSTM

دقت مجموعه تست و اعتبارسنجی



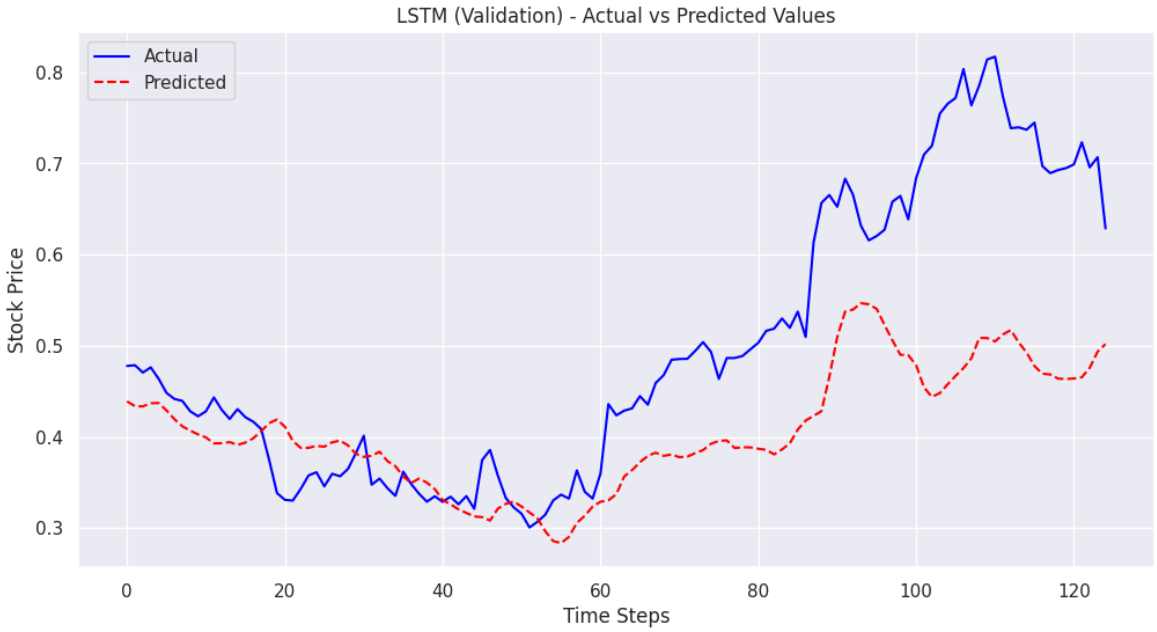
مدل LSTM دچار Overfitting شده است؛ Training Loss کاهش یافته و پایدار شده، اما Validation Loss در سطح بالاتری باقی مانده و نوسان دارد، که نشان‌دهنده ضعف مدل در تعمیم به داده‌های جدید.

دقت مجموعه آموزش



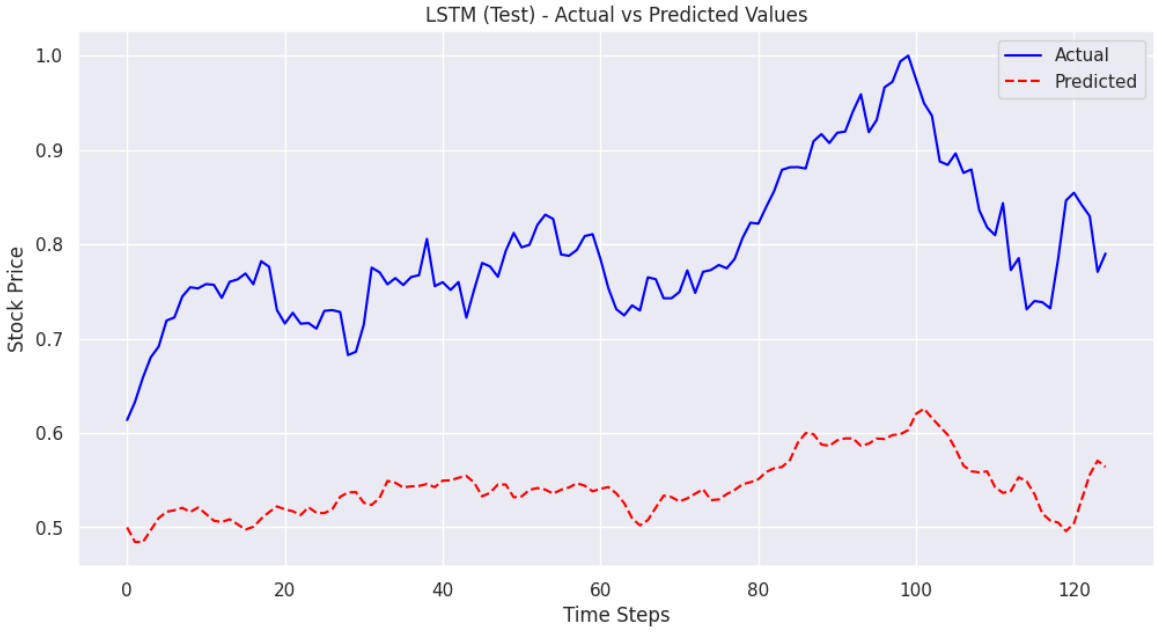
مدل داده‌های آموزشی را بیش‌ازحد حفظ کرده و یادگیری بیش‌ازحد (Overfitting) دارد.

دقت مجموعه اعتبار سنجی



مدل در یادگیری داده‌های جدید مشکل دارد که تأیید دیگری بر **Overfitting** است.

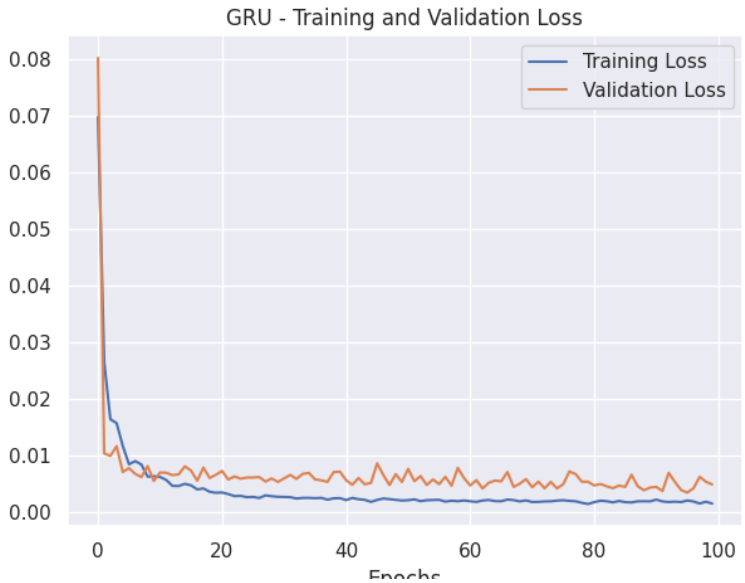
دقت مجموعه تست



مدل کاملاً در پیش‌بینی داده‌های تست شکست خورده است و مقدار پیش‌بینی‌شده هیچ همخوانی‌ای با مقدار واقعی ندارد.

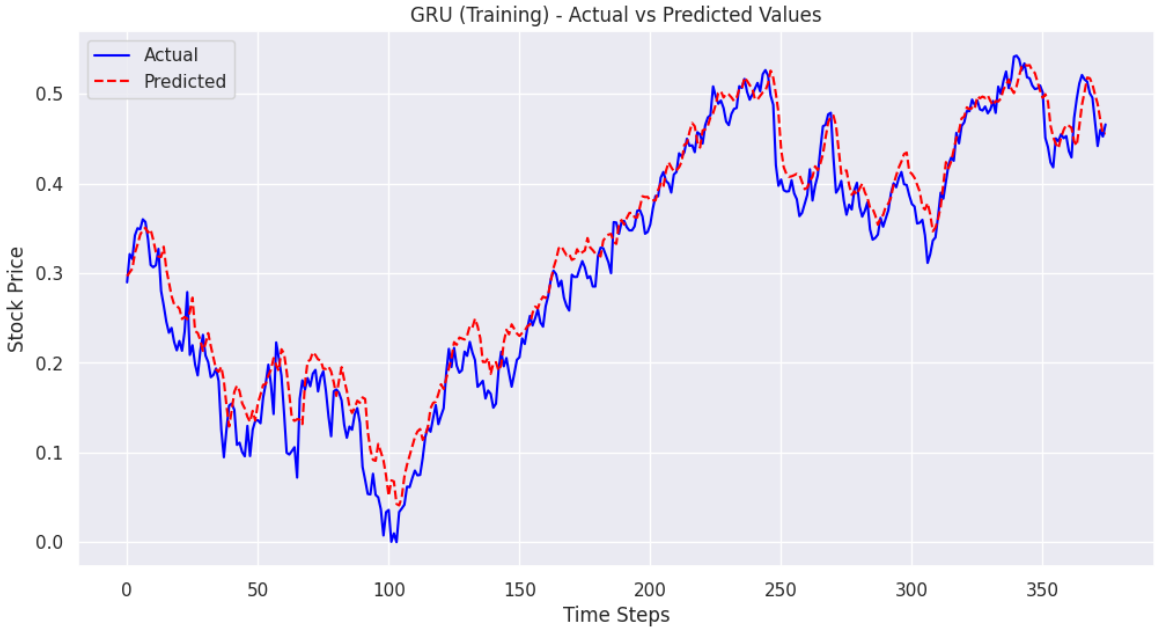
GRU

دقت مجموعه تست و اعتبارسنجی



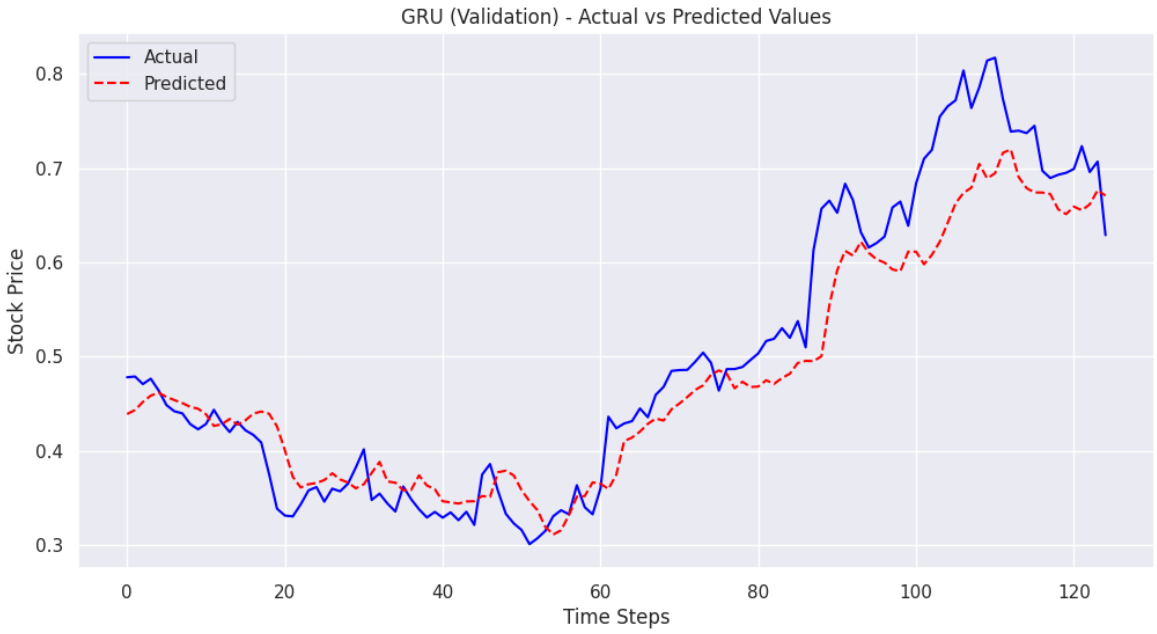
این نشان می‌دهد که GRU نسبت به LSTM در این سناریو تعمیم‌پذیری بهتری دارد.
Overfitting به شدت کاهش یافته است، اما هنوز مقداری وجود دارد.

دقت مجموعه آموزش



مدل روی داده‌های آموزشی عملکرد بسیار خوبی دارد، اما این مقدار کمی پایین‌تر از مدل LSTM است که نشان می‌دهد مدل GRU نسبت به LSTM کمتر پیچیده شده است.

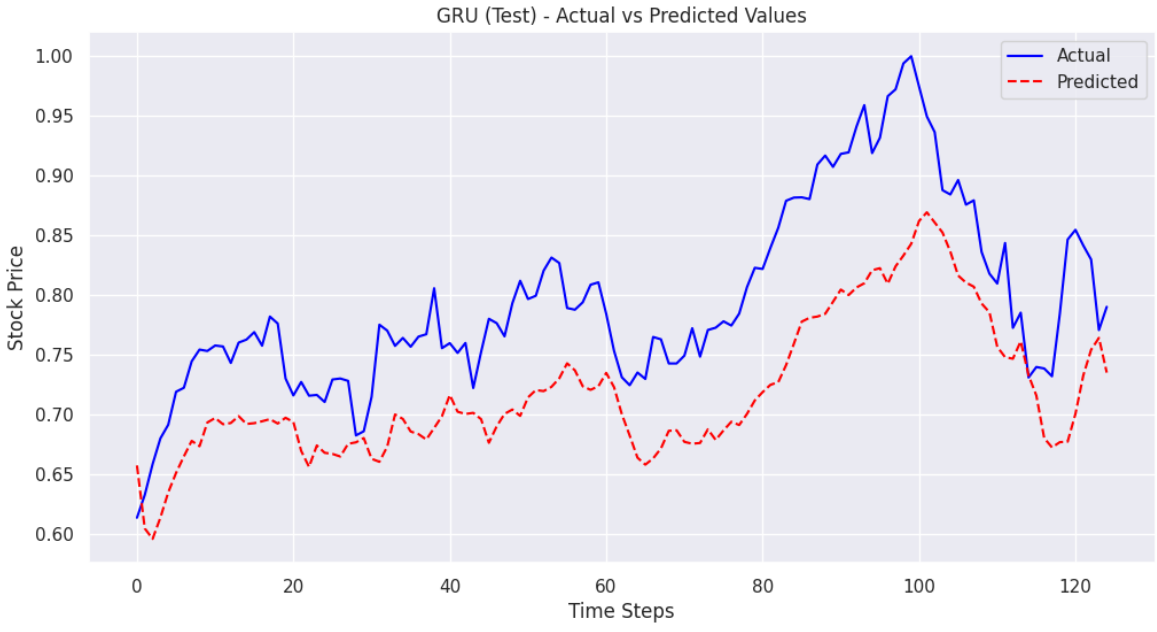
دقت مجموعه اعتبارسنجی



برخلاف LSTM ، مدل GRU توانسته است داده‌های اعتبارسنجی را بهتر پیش‌بینی کند.

Overfittingکاهش یافته و مدل تعمیم‌پذیری بهتری دارد.

دقت مجموعه تست



مدل در پیش‌بینی داده‌های جدید هنوز مشکل دارد، اما بهتر از LSTM است.

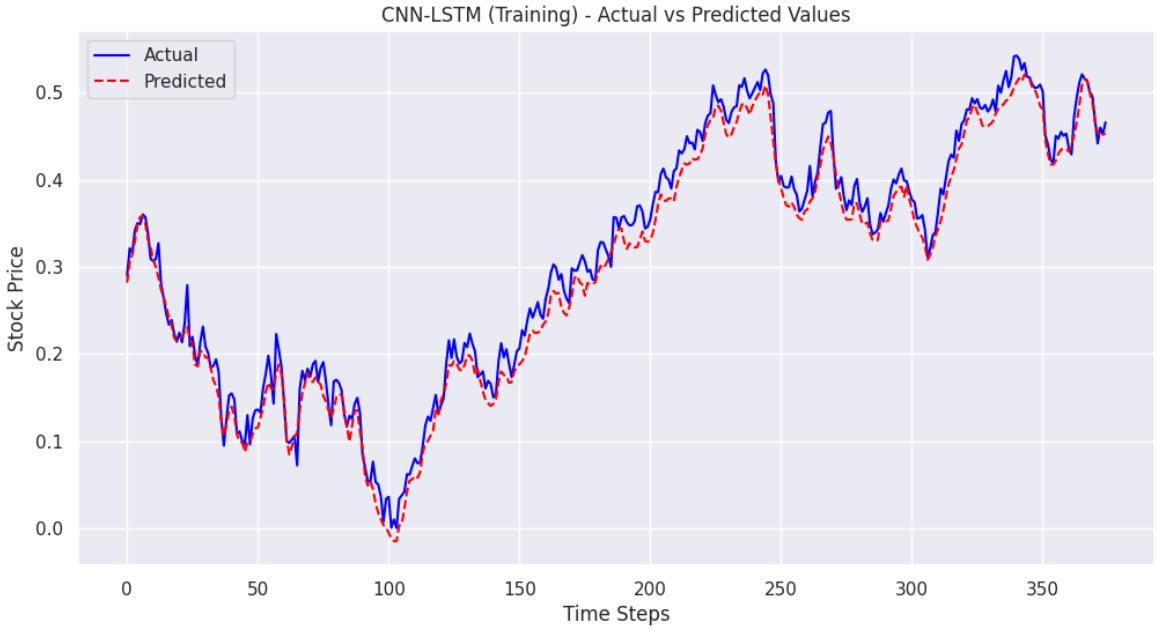
ممکن است داده‌های تست دارای رفتار متفاوتی باشند که مدل در آموزش ندیده است.

CNN-LSTM

دقت مجموعه تست و اعتبارسنجی

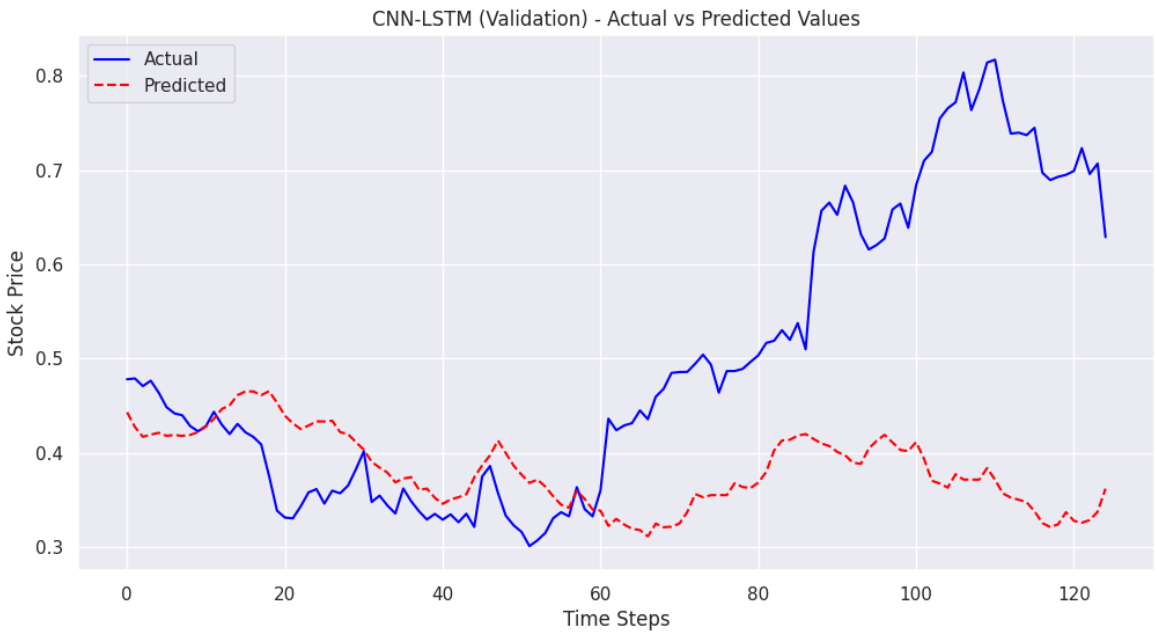
مدل در اوایل آموزش عملکرد خوبی دارد اما در اواخر دچار Overfitting می‌شود.

دقت مجموعه آموزش



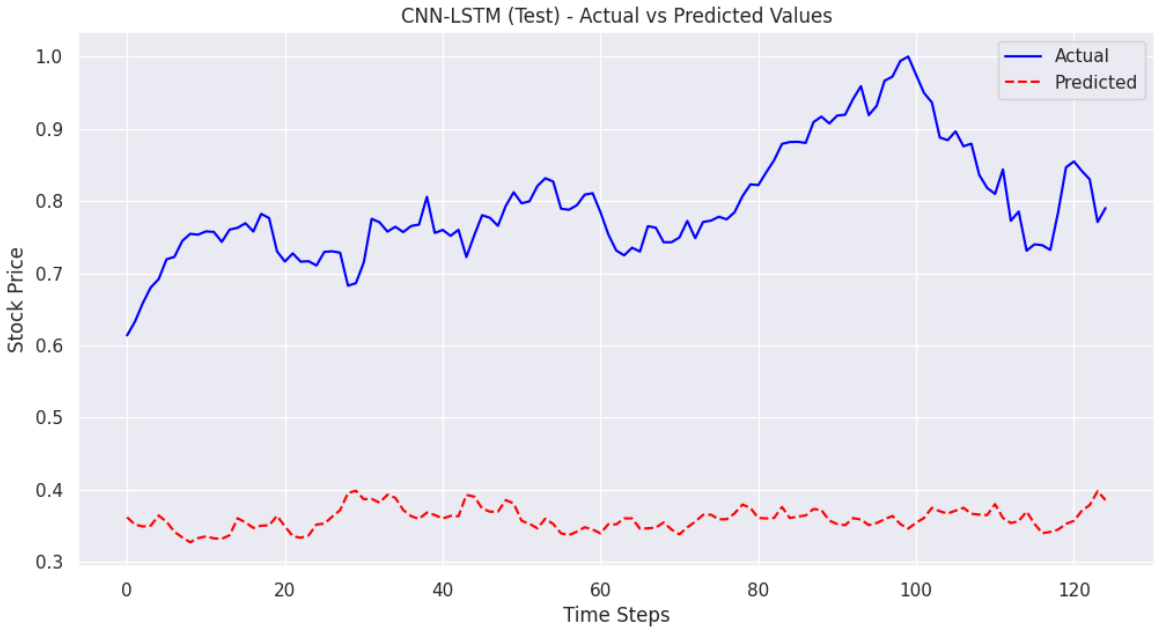
مرحله آموزش عالی عمل کرده است اما این نشانه Overfitting نیز می‌تواند باشد.

دقت مجموعه اعتبارسنجی



Overfitting تایید می‌شود؛ مدل در داده‌های جدید عملکرد ضعیفی دارد.

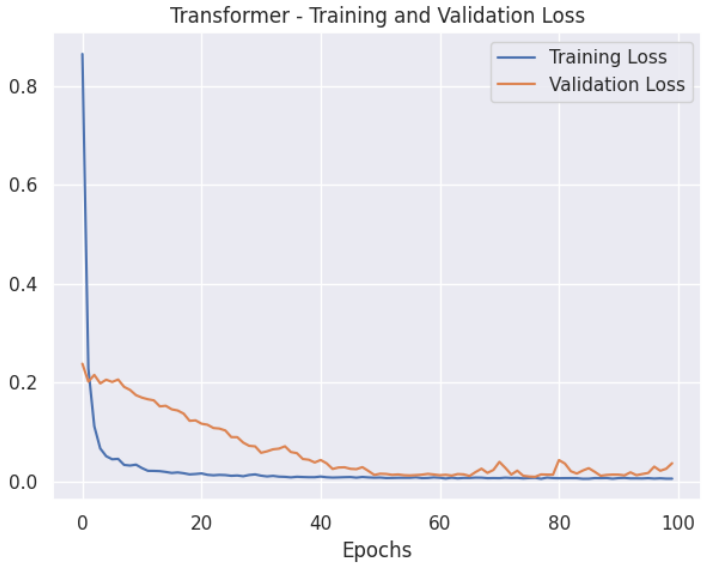
دقت مجموعه تست



مدل ناتوان در تعمیم‌دهی به داده‌های جدید است و در یادگیری ویژگی‌های جدید شکست خورد است.

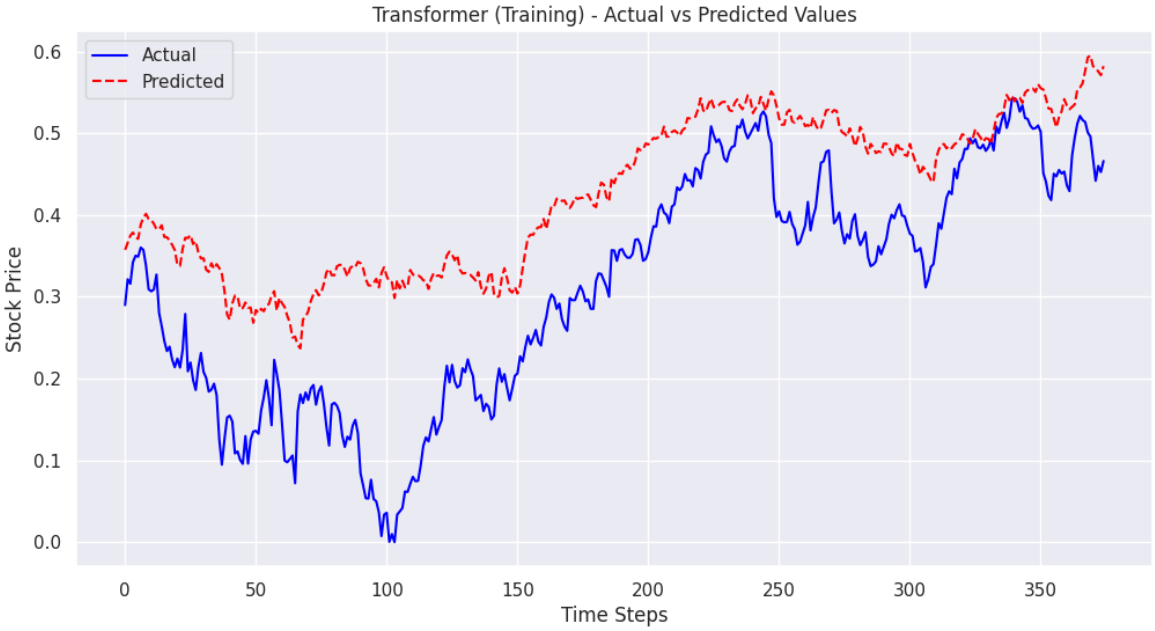
TRANSFORMER

دقت مجموعه تست و اعتبارسنجی



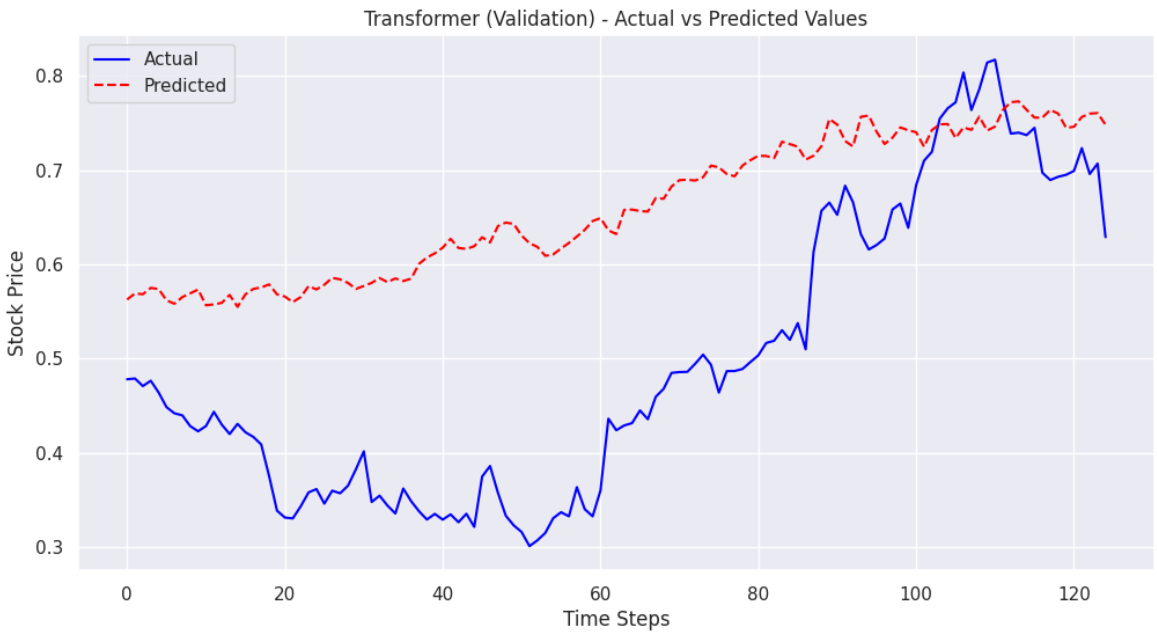
Training Loss کاهش یافته اما Validation Loss پس از کاهش اولیه، نوسان دارد و از روند واقعی منحرف شده‌است، **که نشان‌دهنده** مشکل در تعمیم‌دهی و بیش‌برازش (Overfitting) **است**

دقت مجموعه آموزش



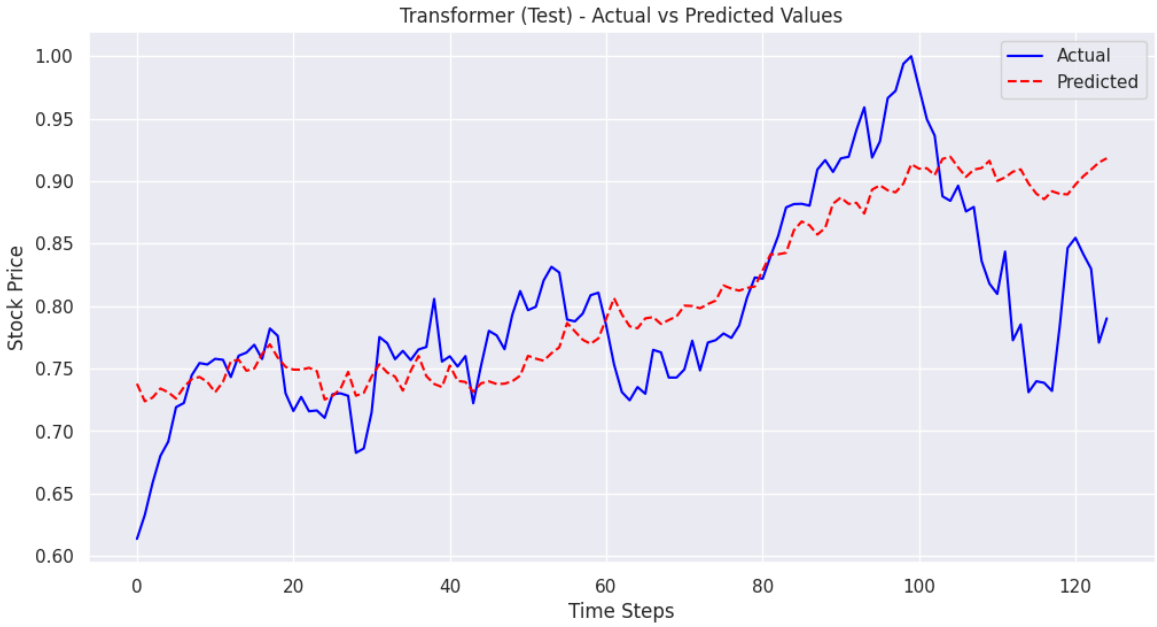
مدل در داده‌های آموزشی تطابق خوبی دارد، اما به دلیل حفظ بیش از حد الگوها، تعمیم‌پذیری به داده‌های جدید پایین است.

دقت مجموعه اعتبارسنجی



پیش‌بینی‌های مدل نسبت به **داده‌های واقعی** بیش از حد صاف و غیر دقیق است، که نشان‌دهنده ضعف مدل در تشخیص نوسانات بازار است

دقت مجموعه تست



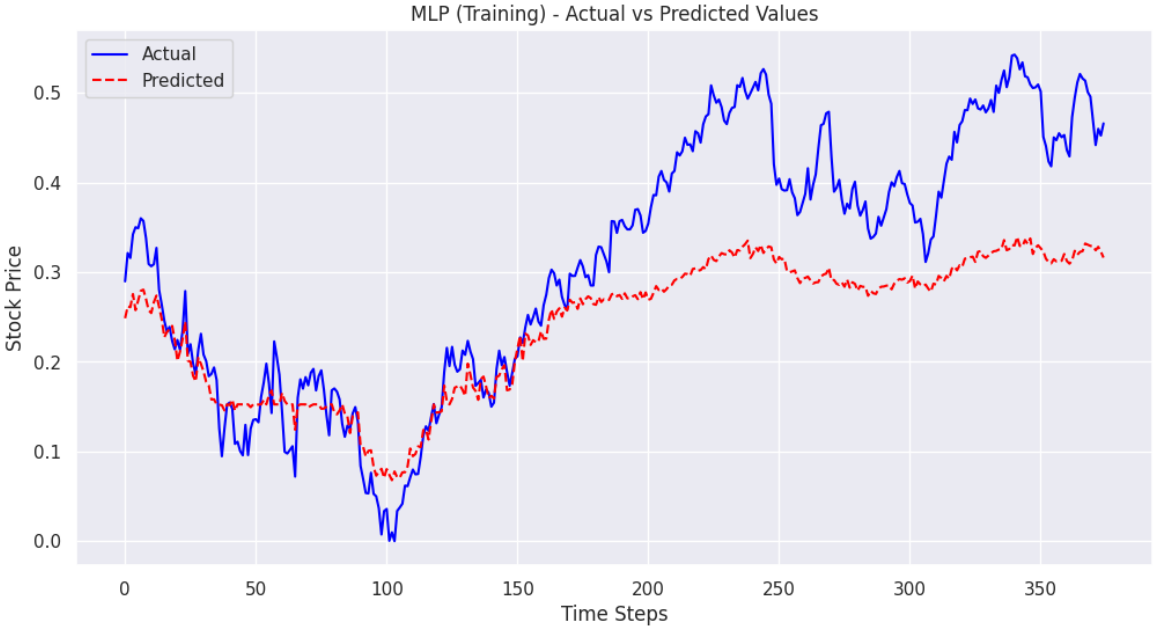
مدل به خوبی روند کلی را شناسایی می‌کند اما پیش‌بینی‌ها همچنان با انحراف از داده‌های واقعی همراه است، که نشان از عدم تعمیم کافی دارد.

MLP

دقت مجموعه تست و اعتبارسنجی

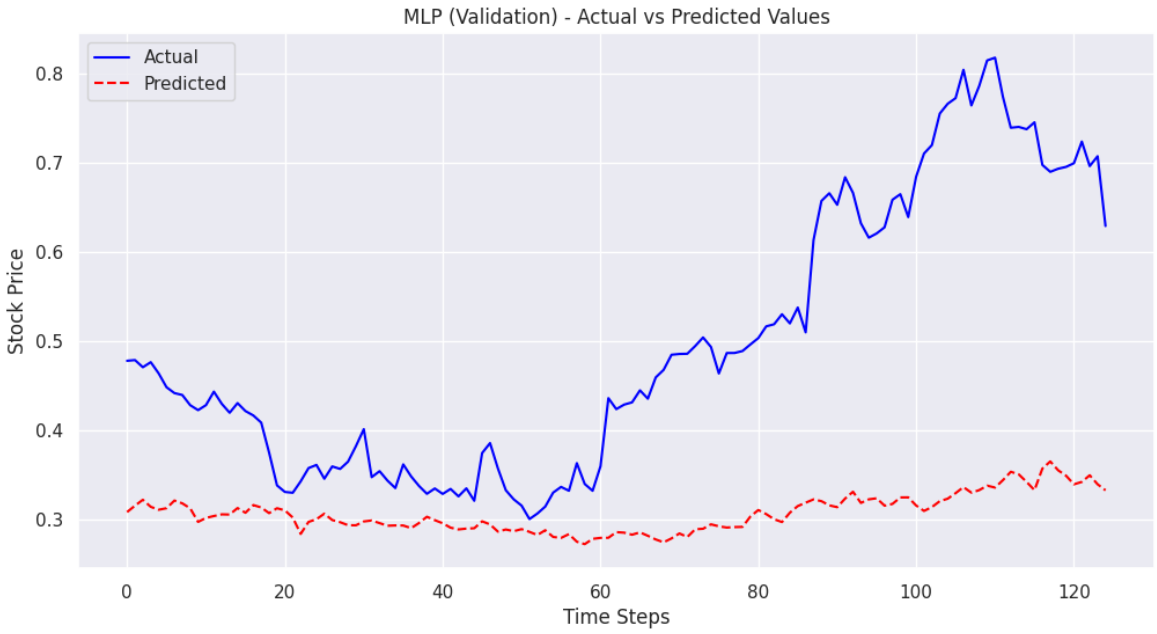
مدل MLP دچار Overfitting شده؛ Training Loss کاهش یافته اما Validation Lossپس از کاهش اولیه، در سطح بالایی نوسان دارد.

دقت مجموعه آموزش



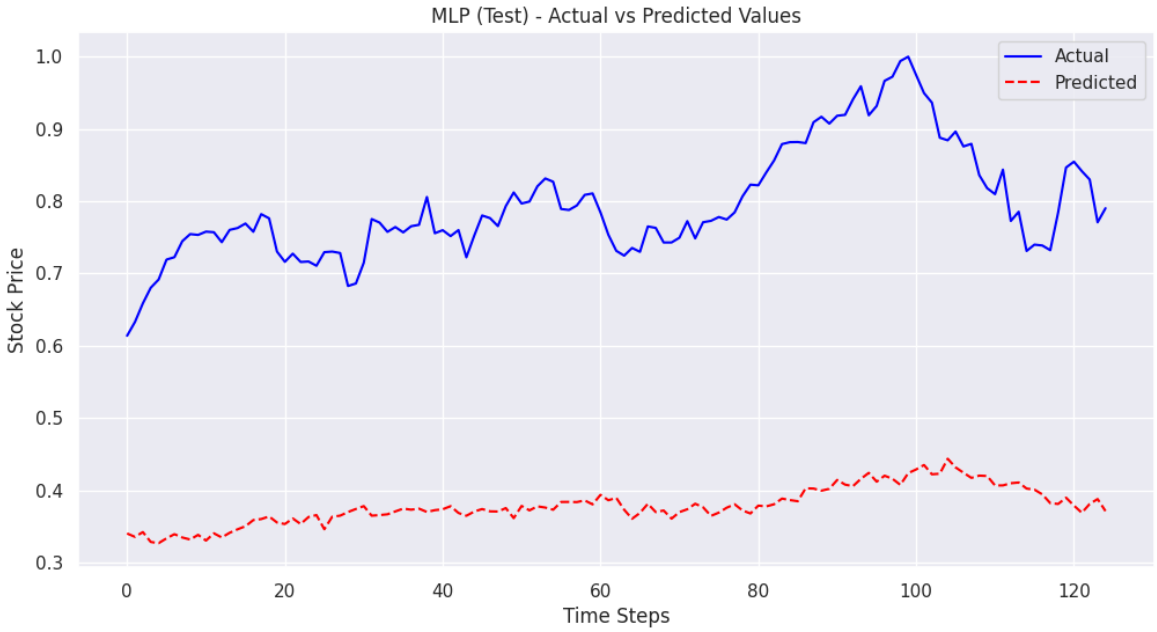
مدل در داده‌های آموزشی پیش‌بینی‌های یکنواخت و نادقیق ارائه می‌دهد که نشان‌دهنده ضعف در یادگیری نوسانات پیچیده بازار است.

دقت مجموعه اعتبارسنجی



مدل توانایی تطبیق با روند واقعی را ندارد و پیش‌بینی‌ها به‌طور مداوم کمتر از مقدار واقعی هستند که نشان از کم‌برازش (Underfitting) دارد.

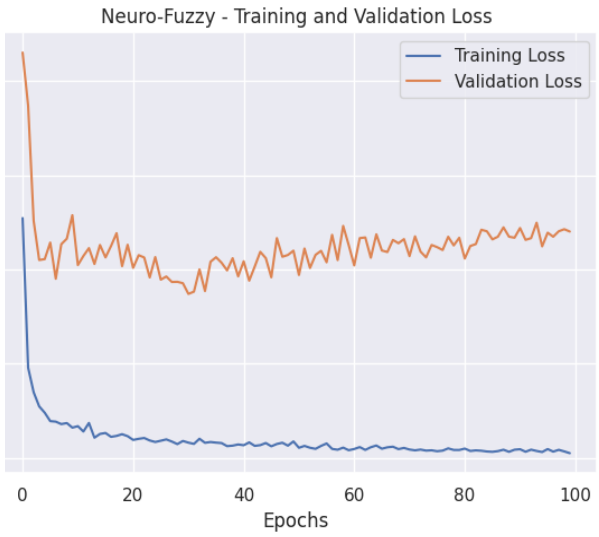
دقت مجموعه تست



در داده‌های تست، مدل دچار خطای سیستماتیک شده و مقدار واقعی را کمتر از حد پیش‌بینی می‌کند، که نشان‌دهنده ضعف در تعمیم و تطبیق با تغییرات واقعی بازار است.

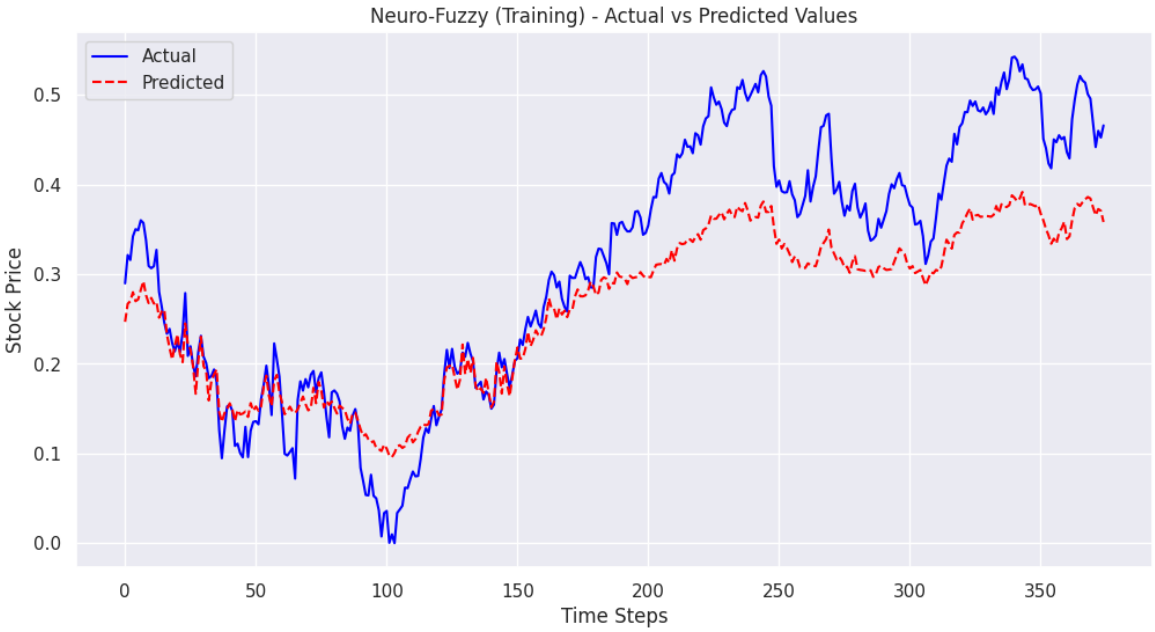
ANFIS

دقت مجموعه تست و اعتبارسنجی



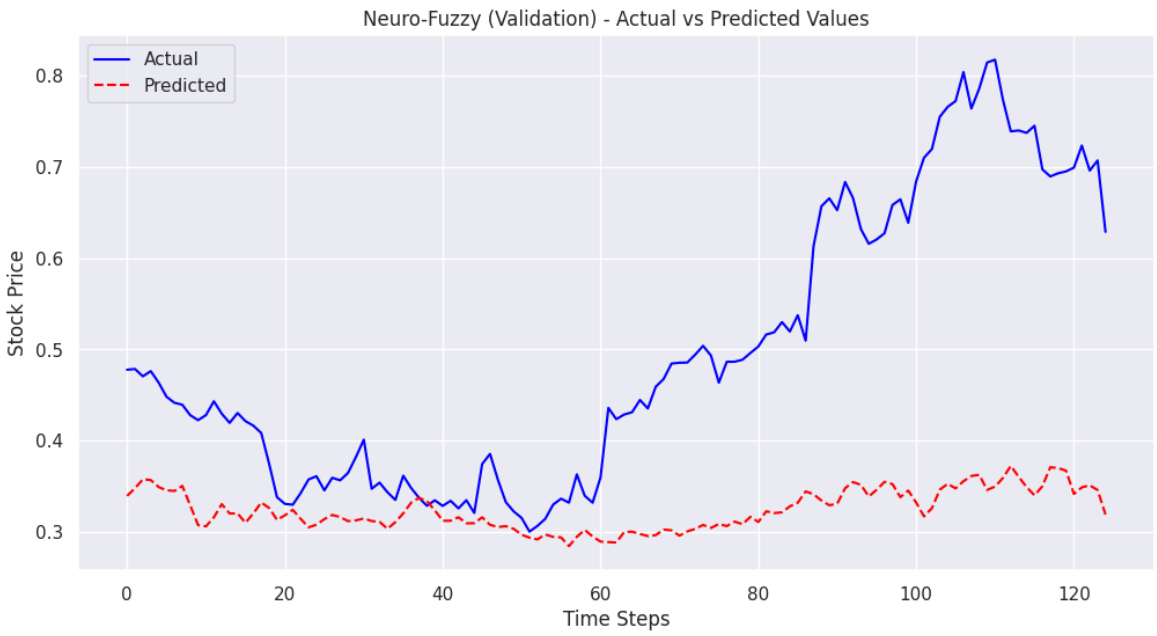
مدل در طول آموزش، کاهش مداومی در خطای آموزش دارد اما خطای اعتبارسنجی نسبتاً نوسانی باقی مانده است که نشان‌دهنده تطبیق بیش از حد (Overfitting) است.

دقت مجموعه آموزش



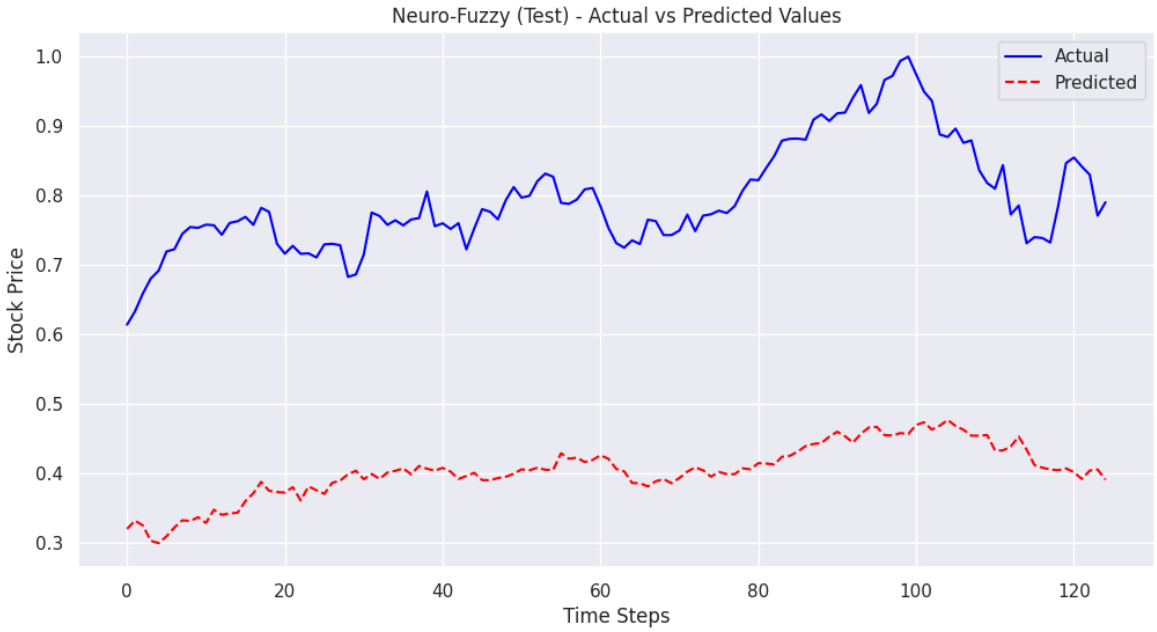
مدل در مجموعه آموزش عملکرد خوبی دارد و پیش‌بینی‌ها تقریباً هم‌راستا با داده‌های واقعی هستند، اما هنوز برخی اختلافات مشاهده می‌شود.

دقت مجموعه اعتبارسنجی



در داده‌های اعتبارسنجی، پیش‌بینی‌ها از داده‌های واقعی عقب هستند و مدل قادر به تطبیق دقیق با نوسانات واقعی قیمت نیست، که نشان‌دهنده ضعف در تعمیم است.

دقت مجموعه تست



مدل در مجموعه تست عملکرد ضعیفی دارد و پیش‌بینی‌ها دقت کمی دارند، به‌طوری که مقادیر پیش‌بینی شده به‌طور قابل توجهی کمتر از مقادیر واقعی هستند، نشان‌دهنده کم‌برازش (Underfitting) است.

مقایسه عملکرد مدل‌ها باهم

LSTM

- **نقاط قوت :**عملکرد خوب روی داده‌های آموزش، اما در داده‌های اعتبارسنجی و تست بهبود لازم دارد.
- **نقاط ضعف :**تفاوت بین خطای آموزش و اعتبارسنجی نشان‌دهنده اورفیتینگ است.

GRU

- **نقاط قوت :**عملکرد بهتر نسبت به LSTM با کاهش پیچیدگی، تطبیق بهتر در داده‌های تست.
- **نقاط ضعف :**هنوز اورفیتینگ مشاهده می‌شود، اما کمتر از LSTM.

CNN-LSTM

- **نقاط قوت :**عملکرد خوب در داده‌های آموزش، اما عدم تعمیم‌پذیری در داده‌های تست.
- **نقاط ضعف :**مدل بیش‌ازحد پیچیده شده و ممکن است برای این نوع پیش‌بینی مناسب نباشد.

TRANSFORMER

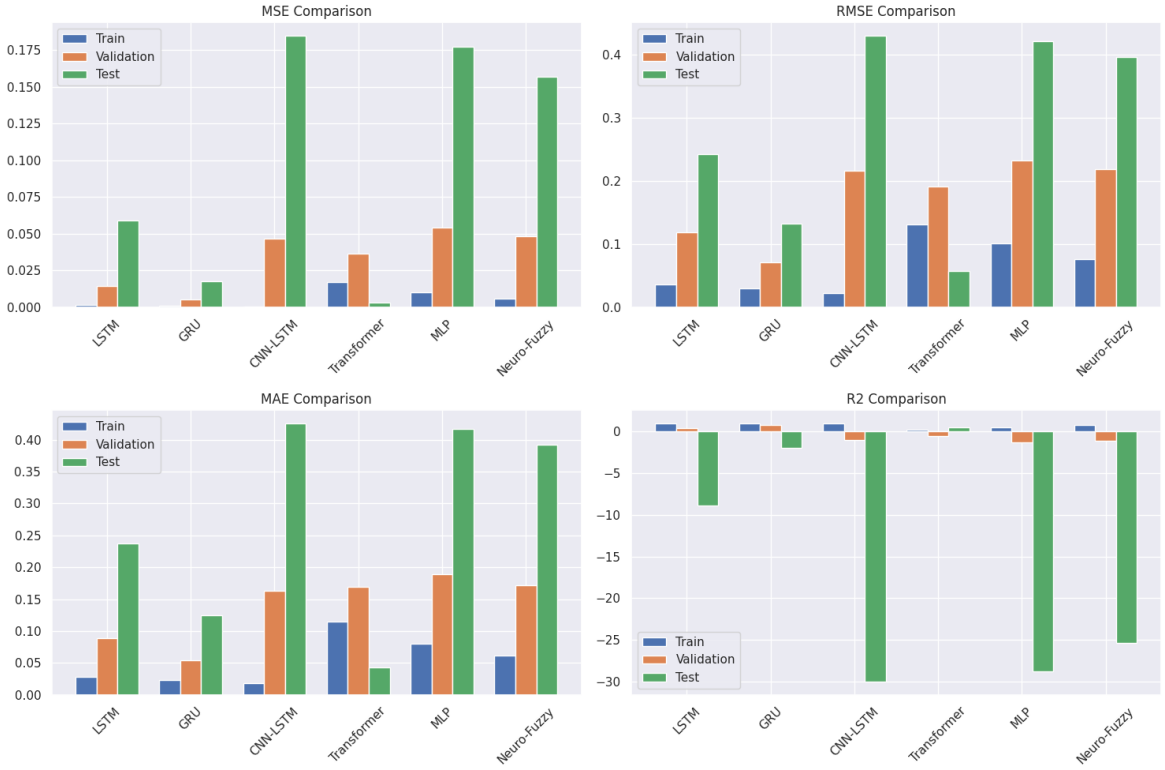
- **نقاط قوت :**عملکرد ضعیف روی داده‌های تست، مشکل بیش‌برازش و تفاوت زیاد بین داده‌های واقعی و پیش‌بینی شده.
- **نقاط ضعف :**این مدل در پیش‌بینی سری‌های زمانی با این داده‌ها عملکرد ضعیفی دارد.

MLP

- **نقاط قوت :**مدل ساده با سرعت اجرا بالا، اما پیش‌بینی دقیق نیست.
- **نقاط ضعف :**عملکرد ضعیف روی داده‌های تست، نشان‌دهنده این است که MLP برای سری‌های زمانی مناسب نیست.

ANFIS

- **نقاط قوت :**تعمیم‌پذیری بهتر ازMLP ، اما هنوز پیش‌بینی با دقت پایین انجام می‌شود.
- **نقاط ضعف :**نوسانات زیاد در اعتبارسنجی نشان‌دهنده عدم پایداری مدل است.



MSE

- GRU کمترین مقدار MSE را دارد، به این معنی که این مدل کمترین میزان خطای کلی را دارد.
- CNN-LSTM، MLP و Neuro-Fuzzy دارای بیشترین مقدار MSE در تست هستند، که نشان‌دهنده عملکرد ضعیف آن‌ها در تعمیم روی داده‌های جدید است.
- Transformerعملکرد متوسطی دارد، اما مقدار MSE در تست نسبتاً بالا است.

RMSE

- GRU مجدداً بهترین عملکرد را دارد، زیرا مقدار RMSE آن در تست پایین‌تر از سایر مدل‌ها است.
- CNN-LSTM و MLP دارای بالاترین مقدار RMSE در تست هستند، به این معنی که پیش‌بینی‌های این مدل‌ها در تست تفاوت زیادی با مقادیر واقعی دارند.
- LSTM مقدار RMSE متوسطی دارد اما هنوز بالاتر از GRU است.

MAE

- GRU کمترین مقدار MAE را دارد، به این معنی که این مدل کمترین میزان خطای مطلق را در پیش‌بینی‌ها دارد.

- CNN-LSTM و MLP دارای بالاترین مقدار MAE در تست هستند که نشان می‌دهد این مدل‌ها انحراف زیادی در پیش‌بینی دارند.
- Transformer نسبت به سایر مدل‌های ضعیف مانند MLP بهتر است، اما همچنان بهبود لازم دارد.

R^2

- GRU بهترین مقدار R^2 را دارد، به این معنی که این مدل بیشترین میزان واریانس را در داده‌های واقعی توضیح می‌دهد.
- CNN-LSTM و MLP مقدار R^2 منفی دارند که نشان می‌دهد این مدل‌ها حتی بدتر از یک پیش‌بینی خطی ساده عمل کرده‌اند.
- Transformer مقدار R^2 بهتری نسبت به MLP و CNN-LSTM دارد، اما همچنان عملکرد آن ضعیف است.

نتیجه‌گیری کلی

بهترین مدل

- **GRU**در تمامی معیارها عملکرد بهتری دارد و خطای کمتری روی داده‌های تست دارد.

بدترین مدل

- **MLP**و **CNN-LSTM**: عملکرد ضعیفی در پیش‌بینی داشتند و در تست مقدار خطای بالایی دارند و R^2 آن‌ها منفی است.

مدل‌های قابل بهبود:

- LSTM را می‌توان با تنظیم هایپرپارامترها (کاهش پیچیدگی مدل و بهبود مقدار Dropout) بهبود داد.
- Transformer عملکرد متوسطی دارد، اما نیاز به افزایش داده‌های آموزشی و تنظیم بهتر دارد.
- Neuro-Fuzzy عملکرد بهتری از MLP دارد، اما هنوز نیاز به بهینه‌سازی پارامترهای فازی دارد.

و من الله التوفيق