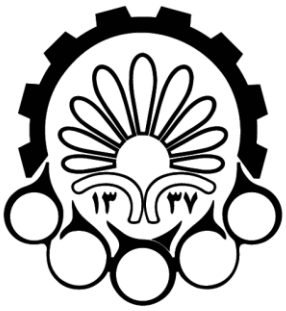


دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

چالش پیش‌بینی میزان مصرف اینترنت مشترکین همراه اول

نام تیم:

Order of the Phoenix

اعضای تیم:

سالار حسینی شمعچی

محمد حسین‌زاده

احمد احمدی

اسفند ۱۴۰۰ - فروردین ۱۴۰۱

چکیده:

با توجه به افزایش اهمیت اینترنت در دهه‌های اخیر، مدیریت منابع و سرویس دهی اینترنت برای اپراتورهای تامین کننده‌ی اینترنت به مسئله‌ی بسیار مهمی تبدیل شده است. یکی از پیش‌نیازهای مدیریت درست در سرویس دهی اینترنتی، نگرشی درست نسبت به میزان مصرف اینترنت کاربران در آینده می‌باشد. با توجه به این مسئله، هدف از این چالش پیش‌بینی میزان مصرف اینترنت کاربران اپراتور همراه اول در طول یک هفته‌ی آینده، با توجه به تاریخچه‌ی مصرف هر کاربر می‌باشد، که در دو مرحله برای دو هفته‌ی متوالی بررسی می‌شود. روش اتخاذ شده توسط تیم ما یک روش مختص پیش پردازش داده‌ها و ساخت سری‌های زمانی است که با شیفت دادن مصرف اینترنت هر کاربر در هر روز به روز بعدی به دست آمده و نتایج بهتری نسبت به روش‌های دیگری که امتحان شده دست داده است. علاوه بر این به عنوان خود الگوریتم پیش‌بینی نیز یک شبکه‌ی بازگشتی LSTM طراحی شده است، که با توجه به نتایج آن بر روی داده‌های تست، عملکرد مناسبی نشان داده است.

فهرست

۱- مقدمه

در دهه‌های اخیر، اینترنت به بخش مهمی از زندگی افراد تبدیل شده است. کاربردهای اینترنت در برقراری ارتباط، مکان یابی، دسترسی و نگهداری اطلاعات، خرید و فروش، آموزش، سرگرمی و هزاران کاربرد دیگر و علاوه بر آن، رشد سریع اینترنت اشیا (IoT)[1] اینترنت را به عضوی جدا ناپذیر و حتی حیاتی در زندگی مدرن تبدیل کرده است.

در دنیایی که اینترنت چنان نقش مهمی را ایفا میکند، تامین کننده‌های خدمات اینترنتی (ISPs)[2] با چالش‌های مهمی در زمینه‌ی تامین اینترنت مواجه هستند. یکی از راهکارهای اصلی این شرکت‌ها در مواجهه با چالش‌های سرویس دهی اینترنتی، پیش‌بینی مصرف اینترنت کاربران است. برای مثال در موارد امنیتی (مانند تشخیص هویت)، مدیریت منابع سرویس دهی، مدیریت اختصاص پهنای باند، پیش‌بینی تراکم مصرف اینترنت، تبلیغات هدف گذاری شده، مدیریت و توسعه‌ی فضاهای ابری، کاهش مصرف انرژی، افزایش کیفیت سرویس دهی (QoS)[3] و تخصیص بسته‌های پیشنهادی برای کاربران، این راهکار می‌تواند مفید باشد [3][2][1]. علاوه بر این، پیش‌بینی مصرف اینترنت کاربران می‌تواند در برخی مسائل پیچیده‌تر نیز کارساز باشد. یکی از مسائل مهم برای تامین کننده‌های مدرن خدمات اینترنتی، طبیعت الاستیک تقاضای اینترنت در طول روز است، که ناشی از حرکت مداوم مصرف کننده‌ها درون شهرها است. به این معنی که تامین کننده‌ها مجبور به تغییر مداوم پهنای باند هستند. عدم رسیدگی به این مسئله می‌تواند منجر به اختلال در سرویس دهی و نارضایتی کاربران شود. پیش‌بینی میزان مصرف اینترنت یکی از راه حل‌های اصلی برای این مسئله می‌باشد [4].

شرکت ارتباطات سیار ایران (همراه اول) به عنوان بزرگ‌ترین اپراتور خاورمیانه که به عنوان بزرگ‌ترین اپراتور خاورمیانه در نظر دارد با توجه موارد ذکر شده، با طرح چالش پیش‌بینی میزان مصرف اینترنت کاربران خود، به افزایش بهره‌وری، ارائه‌ی خدمات متنوع و بهینه و ایجاد تجربه‌ی کاربری عالی در سطوح کاربری مختلف بپردازد.

در این چالش به پیش‌بینی مصرف اینترنت گروه خاصی از مشترکین منتخب همراه اول پرداخته شده است. در واقع هدف این چالش طراحی و توسعه‌ی مدلی هوشمند و منعطف بر پایه‌ی علم هوش مصنوعی روز می‌باشد. در این مسابقه داده‌های اینترنت مصرفی روزانه‌ی کاربران به مدت سه ماه متوالی از سیم کارت‌های مختلف گمنام شده در اختیار شرکت کنندگان قرار گرفته است. هدف نهایی، طراحی مدلی است که با استفاده از ۶۹ روز (دو ماه و یک هفته) داده‌ی دریافتی، به بررسی رفتار داده‌ها پرداخته و در قالب سه مرحله، به پیش‌بینی میزان مصرف

[1] Internet of Things
[2] Internet Service Providers
[3] Quality of Service

اینترنت مشترکین در دو هفته‌ی انتهایی اقدام نمایند. میزان مصرف در هر یک از دو هفته‌ی نهایی در یک مرحله‌ی مجزا پیش‌بینی شده است. در انتهای هر مرحله، مقدار واقعی مصرف اینترنت آن مرحله‌ی مشترکین در اختیار شرکت کنندگان قرار گرفته است، تا در مراحل بعدی از آن استفاده شود. مقدار مصرف اینترنت مرحله‌ی اول به عنوان داده‌های تست مدل برای آموزش مرحله‌ی دوم مورد استفاده قرار گرفته است.

در پیش‌بینی میزان مصرف اینترنت، مسئله‌ی مهمی که باید به آن پرداخته شود، پیش‌بینی پذیری است. پیش‌بینی پذیری به معنی امکان پیش‌بینی دقیق نسبت به بازه‌ی زمانی مطلوب است. از طرفی بازه‌ی زمانی بزرگی لازم است تا فرصت کافی برای اَعمال کنترلی و جبران تاخیر سیستم وجود داشته باشد. از طرف دیگر، افزایش بازه‌ی زمانی منجر به افزایش خطای مدل شده و عملکرد کنترلی را مختل می‌کند. لذا باید به مصالحه‌ای در راستای انتخاب بازه‌ی زمانی پیش‌بینی و دقت مطلوب رسید [5]. در این چالش، بازه‌ی زمانی یک هفته‌ای برای رسیدن به حداقل خطای ممکن در پیش‌بینی میزان مصرف اینترنت در نظر گرفته شده است، اما با توجه به کاربرد مورد نیاز در پیش‌بینی مصرف اینترنت، بهتر است بازه‌ی زمانی مطلوب انتخاب شده و از بروز خطای ناخواسته جلوگیری شود.

این گزارش در ادامه از این بخش‌ها تشکیل شده است؛ کارهای پیشین در بخش دوم مورد بررسی قرار گرفته. بخش سوم مربوط به روش شناسی می‌باشد. در بخش چهارم نتایج کسب شده مورد تحلیل قرار گرفته‌اند. نهایتا در بخش پنجم نتیجه گیری نهایی انجام شده و منابع مورد استفاده در بخش ششم آورده شده‌اند.

۲- کارهای پیشین

آزمایش انجام شده در [1] اطلاعات مصرف اینترنت ۶۶ دانشجو را با استفاده از داده‌های Cisco NetFlow به صورت ناشناس جمع‌آوری کرده، با استفاده از روش میانگین متحرک[4]، میزان مصرف آینده‌شان را پیش‌بینی میکند. با توجه به بررسی‌های این تحقیق، مشاهدات یک سری زمانی را می‌توان به سه بخش تقسیم کرد؛ روند بلند مدت، روند فصلی و تغییرات ناگهانی که بخشی از طبیعت اغتشاشی و تصادفی بودن سری زمانی را نشان می‌دهد.

محققین در [4] برای مقابله با مسئله‌ی تقاضای الاستیک اینترنت، یک مدل پیش‌بینی تطبیقی برای تخصیص منابع اینترنتی و برنامه ریزی استراتژیک برای زیربنای اینترنتی طراحی کرده‌اند. مدل طراحی شده با توجه به

[4] Moving Average

رفتار شبکه، می‌تواند یکی از دو نوع ARIMA[5] و یا NNAR[6] باشد.

بر اساس یافته‌های تحقیق [6] مدل‌های ARMA[7] و ARIMA در تشخیص تاثیر بلند مدت داده‌ها عملکرد ضعیفی دارند. همچنین در این تحقیق روش‌هایی همچون شبکه‌های عصبی MLP[8]، الگوریتم SVR[9] و مدل‌های فازی نیز مورد بررسی قرار گرفته و نهایتا برای در نظر گرفتن رابطه‌ی بلند مدت داده‌ها، روش GPR[10] مورد استفاده قرار گرفته است.

در تحقیق [7] روش شبکه‌ی عصبی بازگشتی از نوع GRU[11] مورد استفاده قرار گرفته است. بر اساس این تحقیق، شبکه‌های عصبی بازگشتی رفتار مناسبی در پیش‌بینی مصرف اینترنت نشان می‌دهند. هم‌چنین در این تحقیق نشان داده شده که روش‌های LSTM[12] و GRU عملکرد نسبتا یکسانی در این مسئله دارند.

در تحقیق [3] چندین روش مرسوم پیش‌بینی سری زمانی با روش خوشه بندی ادغام شده‌اند. روش‌های خطی و غیرخطی به کار برده، WES[13]، HTES[14]، ARMA، مدل ترکیبی موجک‌ها[15] و WES و نهایتا مدل NNAR هستند.

در تحقیق [8] نیز برای مقابله با مشکل روندهای متفاوت سری زمانی در مقیاس‌های متفاوت زمانی و جلوگیری از گسترش خطا در پیش‌بینی چند گام زمانی، از پیش‌بینی در مقیاس‌های زمانی متفاوت استفاده شده و برای این هدف از روش GPR بهره گیری شده است.

در تحقیق [5] برای بهبود عملکرد مدل ARIMA در تشخیص رابطه‌های بلند مدت، این مدل با روش GARCH[16] ادغام شده است.

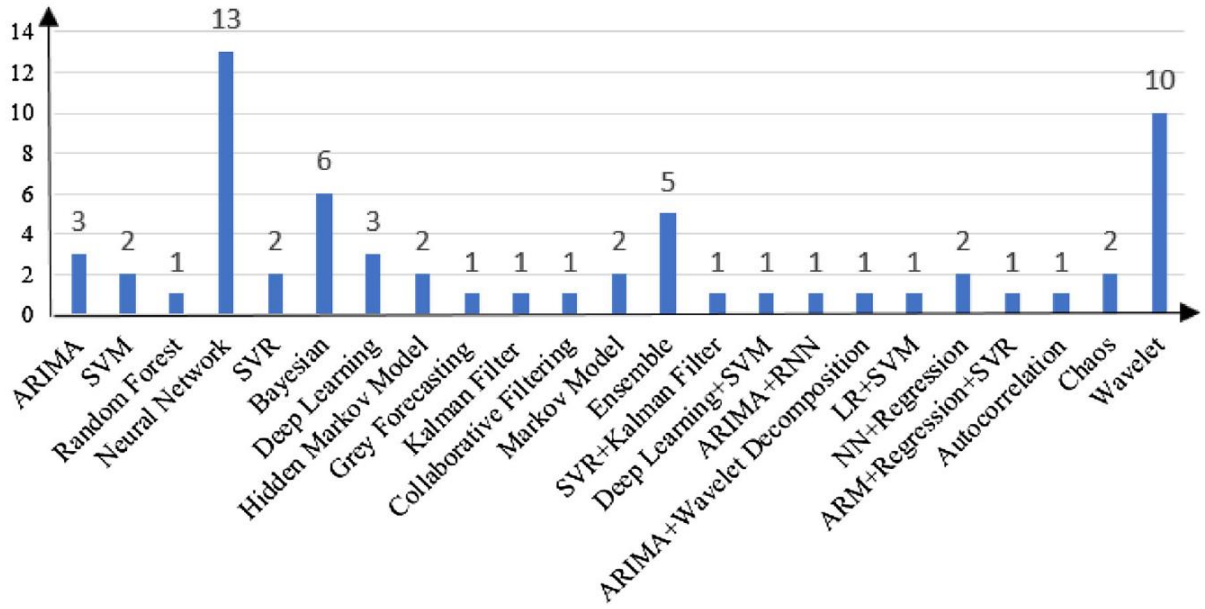
در تحقیق [9] برای کاهش هزینه از روش LSTM برای کاهش هزینه در استفاده از منابع محاسباتی در مصرف اینترنت دستگاه‌های حاشیه[17] استفاده شده است. این تحقیق نشان می‌دهد که LSTM عملکرد بهتری نسبت به

---

[5] Auto-Regressive Integrated Moving Average
[6] Neural Network Auto-Regressive
[7] Auto-Regressive Moving Average
[8] Multi-Layer Perceptron
[9] Support Vector Regression
[10] Gaussian Process Regression
[11] Gated Recurrent Unit
[12] Long Short Term Memory
[13] Weighted Exponential Smoothing
[14] Holt-Trend Exponential Smoothing
[15] Wavelets
[16] Generalized Autoregressive Conditional Heteroscedasticity
[17] Edge Devices

روش‌های AUCROP[18]، XGBoost و Auto-sklearn دارد.

نهایتا در شکل ۱ که برگرفته از تحقیق [2] است، میزان استفاده از روش‌های مختلف در پیش‌بینی بار مصرفی در سری‌های زمانی مورد بررسی قرار گرفته است. همان طور که مشخص است، شبکه‌های عصبی بیشترین استفاده را دارند. علاوه بر این مدل‌های یادگیری عمیق و مدل‌های ترکیبی شبکه عصبی با دیگر روش‌ها نیز به آمار استفاده از شبکه‌های عصبی می‌افزایند.



شکل ۱: میزان استفاده از روش‌های مختلف در پیش‌بینی بار مصرفی در سری زمانی [۲]

۳- روش شناسی

۳-۱- چالش‌های روش شناسی

با توجه به افزایش روز به روز پیچیدگی عملکرد اینترنت، در پیش‌بینی میزان مصرف اینترنت، چالش‌هایی وجود دارند که باید هر یک از آن‌ها در انتخاب روش مورد استفاده در نظر گرفته شوند. چالش‌های پیش آمده از این قرار هستند:

تطبیق پذیری: مدل پیش‌بینی باید قادر به تطبیق با تغییر رفتار سری زمانی و دینامیک رفتار آن باشد [2].

[18] Application and User Context Resource Predictor

فعال بودن: مدل ساخته شده باید در تشخیص تغییرات ناگهانی فعال باشد، تا بتواند در بازه‌ی زمانی مطلوب، این تغییرات را تشخیص داده و برای مدیریت منابع اینترنتی فرصت کافی ایجاد کند [2].

تاریخچه‌ی داده‌ها: مدل مناسب باید از تاریخچه‌ی داده‌ها نهایت بهره‌وری را داشته باشد. این مدل باید وابستگی‌های بلند مدت و کوتاه مدت سری زمانی را تشخیص داده و با استخراج رابطه‌ی صحیح بین داده‌های گذشته، آینده را به درستی پیش‌بینی کند [2].

پیچیدگی: مدل ساخته شده نباید پیچیدگی زیادی داشته و وقت و انرژی زیادی برای پیش‌بینی مصرف کند. در عین حال مدل باید به حد کافی پیچیده باشد تا بتواند پیچیدگی رابطه‌ی سری زمانی را استخراج کند [2].

ریز و درشتی داده‌ها: در ساخت مدل ابتدا باید در نظر گرفته شود که از چه ویژگی‌هایی لازم است استفاده شود. سپس لازم است طول بازه‌های نمونه برداری برای ویژگی‌ها تعیین شود. چرا که نمونه برداری درشت باعث عدم توجه به دینامیک رفتار سری می‌شود، از طرفی نمونه برداری ریز نیز باعث افزایش هزینه‌ی جمع آوری داده و هزینه‌ی محاسباتی می‌شود [2].

طول بازه‌ی زمانی: انتخاب طول بازه‌ی زمانی چالش مهمی است. برای انتخاب درست بهتر است از بازه‌ای استفاده شود که به درست‌ترین شکل رفتار سیستم را نشان می‌دهد [2].

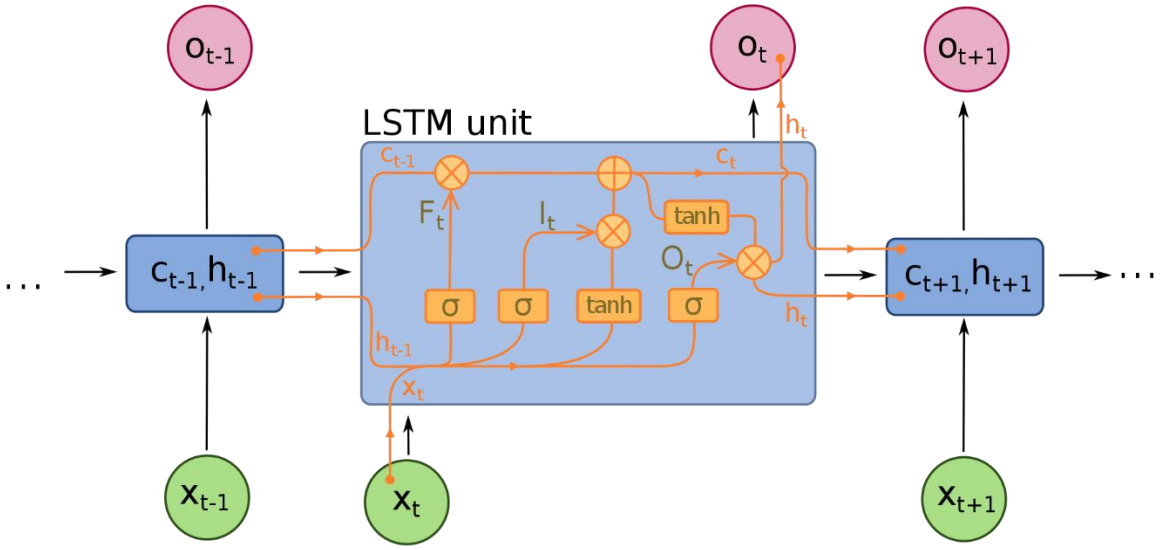
دقت عمل: یک مدل پیش‌بینی باید دقت عمل کافی داشته باشد، تا از عواقب ناخواسته‌ی پیش‌بینی اشتباه و متعاقبا تصمیم گیری‌های اشتباه اجتناب کند. مشخصا یک تصمیم گیری اشتباه منجر به بدتر شدن عملکرد کلی شبکه می‌شود [6].

۳-۲- انتخاب روش

با توجه به چالش‌های ذکر شده، برای رسیدن به تطبیق پذیری، سرعت و دقت عمل مناسب، همچنین برای در نظر گرفتن طبیعت غیرخطی داده‌ها، امکان انتخاب طول بازه‌های متغیر و از همه مهم‌تر برای در نظر گرفتن وابستگی‌های بلند مدت و کوتاه مدت داده‌ها، در این چالش بر استفاده از شبکه عصبی بازگشتی از نوع LSTM تصمیم گرفته شده است.

شبکه‌ی LSTM با ساختار زنجیره‌ای خود، قادر به ایجاد ارتباط بین داده‌ها در بازه‌های طولانی و نگاشت الگوهای رفتاری به پیش‌بینی مصرف است. شبکه‌ی LSTM یک مدل سری زمانی غیرخطی است که قادر به حفظ اطلاعات در حافظه‌ی بلند مدت است. شبکه‌ی LSTM برگرفته از ساختار شبکه‌های عصبی بازگشتی ساده است، اما برتری LSTM نسبت به این شبکه‌ها در دو وجهه‌ی مهم است. وجهه‌ی اول حل مشکل محو گرادیان است که

منجر به عدم یادگیری شبکه در بازه‌های بلند زمانی می‌شود. وجهه‌ی دوم نیز حل مشکل وابستگی‌های بلند مدت است، که با بهره گیری از حافظه‌ی این ساختار، این مشکل حل می‌شود [9]. شکل ۲ نمایانگر ساختار یک بلوک از زنجیره‌ی بلوک‌های LSTM است.



شکل ۲: ساختار یک بلوک از زنجیره‌ی بلوک‌های LSTM [10]

دو روش مرسوم در پیش‌بینی چند گام زمانی در سری‌های زمانی، روش تکراری[۱۹] و روش مستقیم[۲۰] هستند. در روش تکراری، پیش‌بینی یک گام بعدی به صورت مکرر انجام شده و برای گام‌های جلوتر از پیش‌بینی‌های قبلی به عنوان ورودی استفاده می‌شود. در این روش فقط از یک مدل پیش‌بینی برای یک گام بعدی استفاده می‌شود. ایراد اصلی این روش، انباشتگی خطا با افزایش تعداد گام زمانی پیش‌بینی است. اما در روش مستقیم N مدل مختلف برای پیش‌بینی N گام بعدی استفاده می‌شود، یعنی هر یک از مدل‌ها صرفا برای یک گام زمانی مشخص در n گام بعدی (1<n<N) آموزش دیده و استفاده می‌شود. یکی از ایرادهای این روش، پیچیدگی زیاد آن به دلیل آموزش و استفاده از N مدل مختلف است. ایراد دیگر، احتمال عدم آموزش و ایجاد خطا در هر مدل به صورت مجزا است، به این معنی که ممکن است در یک دور از آموزش یکی از مدل‌ها به درستی عمل نکند و در یک دور دیگر، یک مدل دیگر عملکرد درستی نداشته باشد. اما ایراد اصلی این روش این است که صرفا محدود به پیش‌بینی N گام زمانی است، اگرچه می‌توان از این روش برای پیش‌بینی تعداد کم‌تری گام زمانی استفاده کرد، امکان استفاده برای پیش‌بینی تعداد گام زمانی بیشتر وجود ندارد [8]. لذا با توجه به مزایا و معایب دو روش ذکر شده،

[19] Iterative Method
[20] Direct Method

ما در این چالش از روش تکراری استفاده کرده‌ایم.

۳-۳- ساخت شبکه‌ی LSTM

در هر دو مرحله‌ی این چالش از یک شبکه‌ی نسبتا ساده‌ی LSTM در زیرمجموعه‌ی Keras از کتابخانه‌ی TensorFlow با دو لایه‌ی میانی که هر یک ۱۶ واحد دارند استفاده شد، که دلیل استفاده از این ساختار ساده نیز جلوگیری از مشکل overfit شدن می‌باشد. از آنجا که توابع فعال‌ساز لایه‌ی LSTM مخصوص سری‌های مرتب بهینه سازی شده‌اند، دخالتی در تغییر این توابع صورت نگرفته است. در لایه‌ی آخر نیز یک ساختار ساده‌ی dense با یک نورون و یک تابع فعال‌ساز خطی قرار گرفته تا حجم اینترنت مصرفی مربوط به روز آخر سری را پیش‌بینی کند. لایه‌ی اول شبکه نیز سه بعد دارد که دو بعد اول آن متغیر بوده و بعد سوم آن ۳۲ می‌باشد، که تعداد ویژگی‌های ورودی در هر گام زمانی است. دلیل متغیر بودن ابعاد اول و دوم نیز این است که بتوان به شبکه تعداد داده‌های مختلف با طول بازه‌ی مختلف داد.

برای آموزش شبکه، تابع هزینه‌ی انتخاب شده میانگین مربع خطاها[۲۱] است که با استفاده از تابع بهینه‌ساز Adam[22] به مقدار کمینه می‌رسد. سپس داده‌ها به صورت دسته‌های[۲۳] ۵۱۲ تایی در مرحله‌ی اول چالش با تکرار ۳۰۰۰ مرتبه و در مرحله‌ی دوم با تکرار ۲۰۰۰ مرتبه به شبکه داده شده و با تنظیم وزن‌ها روند آموزش پیش می‌رود. دلیل کاهش تعداد مراحل آموزش (ایپاک‌ها)، صرفا این بود که شبکه قبل از ۲۰۰۰ ایپاک می‌تواند به حالت اشباع در آموزش برسد و نیازی به آموزش بیشتر نیست. در طول آموزش نیز هزینه‌ی ناشی از داده‌های صحه گذاری مورد بررسی قرار می‌گیرند. تمامی این پارامترها نیز با آزمون و خطای بسیار بهینه سازی شده‌اند.

برای جلوگیری بیشتر از مشکل overfit شدن و هدایت روند آموزش در راستای مطلوب، در مرحله‌ی اول چالش از ترفندهایی همچون تنظیم کننده‌ی وزن‌های شبکه[۲۴]، کاهش نرخ یادگیری کسینوسی[۲۵] و ذخیره‌ی بهترین وزن‌ها در طول آموزش استفاده شده است، که هر یک تاثیر مثبت خود را در مراحل آزمون و خطا نشان داده‌اند. اما بعدا در مرحله‌ی دوم چالش به دلیل پیش آمدن مشکل underfitting، تنظیم کننده‌ی وزن‌های شبکه حذف شدند.

نهایتا برای تعیین وزن‌های اولیه‌ی شبکه در آموزش و هم‌چنین نوع بُرزنی داده‌ها، از seed گذاری استفاده شده تا حالت تصادفی در اجراهای مکرر کد، نتایج یکسانی را بدهد. سپس با سعی و خطای بسیار، مقداری از

[21] Mean Squared Error
[22] Adaptive Moment Estimation
[23] Batches
[24] Kernel Regularizer
[25] Cosine Learning Rate Decay

seed که بهترین نتایج را داده انتخاب می‌شود.

۳-۴- پیش پردازش داده‌ها

برای مرحله‌ی نهایی چالش که در آن فایل‌های هفته‌های قبل نیز برای استفاده در دسترس هستند، ابتدا درون یک حلقه داده‌های مربوط به هر شخص از فایل آموزش و فایل هفته‌ی اول جدا شده و جای خالی داده‌های گم شده با استفاده از میانگین گیری از ویژگی و شخص مربوط به آن داده پر می‌شود. سپس داده‌های ورودی مربوط به هر شخص به صورت سری‌هایی از داده‌های چند روز متوالی انتخاب می‌شود. این داده‌ها شامل تمام ویژگی‌های داده شده مربوط به هر روز بوده، اما حجم اینترنت مصرفی مربوط به روز قبل استفاده می‌شود. دلیل این کار این است که حجم اینترنت مصرفی روز آخر هر سری به عنوان خروجی آن سری در نظر گرفته شده، لذا این ویژگی برای روزهای قبل نیز لازم است یک روز به جلو شیفت پیدا کند. برای درک بهتر این روند پیش پردازش، در جدول ۱ نمونه‌ی شماتیکی از پیش پردازش داده‌های ۶ روز مربوط به یک شخص آورده شده است. ویژگی ۳۲م در این جدول نشان دهنده‌ی میزان مصرف اینترنت است، که همان طور که مشخص است این ویژگی یک روز به جلوتر شیفت پیدا کرده، که عملا در هر سری، باعث حذف روز اول می‌شود. میزان مصرف اینترنت روز آخر هر سری نیز (که در جدول ۱ باید به صورت D6F32 نشان داده شود) به عنوان خروجی آن سری در نظر گرفته می‌شود. تعداد روزهای هر سری نیز از ۵ الی ۷۵ روز می‌تواند متغیر باشد. به این معنی که اگر طول یک سری n باشد، داده‌های مربوط به یک شخص از روز اول تا روز nام را در بر می‌گیرد ($5<n<75$). نهایتا ویژگی‌های مربوط به روزهای پیشین در سری‌های کوتاه‌تر از ۷۵ روز همگی به عنوان صفر در نظر گرفته شده‌اند، تا تمامی سری‌های آموزشی طول یکسان داشته باشند.

جدول ۱: نمونه‌ی شماتیک پیش پردازش داده‌های ۶ روز مربوط به یک شخص در مرحله‌ی نهایی چالش. روز اول در این سری حذف شده و خروجی این سری نیز D6F32 می‌باشد.

	Feature1	Feature2	...	Feature31	Feature32
Day1	D1F1	D1F2	...	D1F31	-
Day2	D2F1	D2F2	...	D2F31	D1F32
Day3	D3F1	D3F2	...	D3F31	D2F32
Day4	D4F1	D4F2	...	D4F31	D3F32
Day5	D5F1	D5F2	...	D5F31	D4F32
Day6	D6F1	D6F2	...	D6F31	D5F32

اما در مرحله‌ی نیمه نهایی چالش روند متفاوتی اتخاذ شده بود که به دلیل عملکرد نامطلوب آن به روند ذکر

شده تغییر پیدا کرد. در مرحله‌ی نیمه نهایی داده‌های مربوط به ویژگی ۳۲م که همان میزان مصرف اینترنت است، شیفت پیدا نکردند و ویژگی هر روز در روز خود قرار گرفت. اما ویژگی مربوط به روز آخر هر سری صفر در نظر گرفته شد. شماتیک این نوع از پیش پردازش در جدول ۲ قابل مشاهده است. همان طور که در این جدول مشخص است، در این نوع پیش پردازش، هیچ یک از گام‌های زمانی حذف نمی‌شوند، اما در نظر گرفتن مقدار صفر برای روز آخر سری، یک راه حل اشتباه است که ممکن است باعث گمراهی شبکه شود. در این نوع از پیش پردازش نیز، خروجی این سری D6F32 می‌باشد.

جدول ۲: نمونه‌ی شماتیک پیش پردازش داده‌های ۶ روز مربوط به یک شخص در مرحله‌ی نیمه نهایی چالش. روز اول در این سری حذف شده و خروجی این سری نیز D6F32 می‌باشد.

	Feature1	Feature2	...	Feature31	Feature32
Day1	D1F1	D1F2	...	D1F31	D1F32
Day2	D2F1	D2F2	...	D2F31	D2F32
Day3	D3F1	D3F2	...	D3F31	D3F32
Day4	D4F1	D4F2	...	D4F31	D4F32
Day5	D5F1	D5F2	...	D5F31	D5F32
Day6	D6F1	D6F2	...	D6F31	0

در مرحله‌ی نهایی چالش، ابتدا داده‌های مربوط به فایل آموزش و فایل هفته‌ی اول پس از گذراندن مراحل پیش پردازش و آماده سازی سری‌ها به دو دسته‌ی آموزش و صحه گذاری تقسیم شده‌اند تا معیار خوبی برای سنجش عملکرد شبکه داشته باشیم. داده‌های مربوط به فایل هفته‌ی دوم نیز با اضافه شدن به داده‌های آموزش و داده‌های هفته‌ی اول، به عنوان داده‌های تست مورد استفاده قرار می‌گیرند. تست شبکه نیز با روش پیش‌بینی تکراری چند گام زمانی انجام می‌شود، به این معنی که در گام‌های زمانی جلوتر، از خروجی شبکه در گام‌های قبلی استفاده می‌شود. این کار به ما اجازه می‌دهد تا مقاومت و عمومیت پذیری شبکه را به درستی بسنجیم. این روش از تست شبکه در مرحله‌ی نیمه نهایی چالش استفاده نشده بود. همچنین لازم به ذکر است، همه‌ی این داده‌ها بر اساس مقادیر کمینه و بیشینه‌ی داده‌های آموزش نرمالیزه شده و این مقادیر اکسترمم نیز برای استفاده‌های آتی ذخیره می‌شوند.

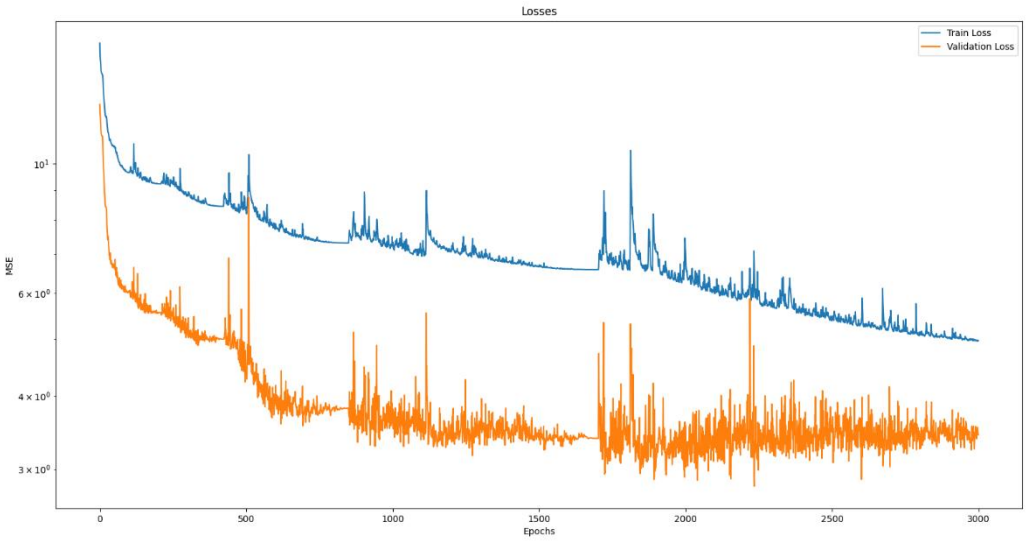
۴- تحلیل نتایج

۴-۱- مرحله‌ی نیمه نهایی چالش

پس از اتمام یادگیری ابتدا برای بررسی عملکرد شبکه، آن را با استفاده از داده‌های صحه گذاری و آموزشی

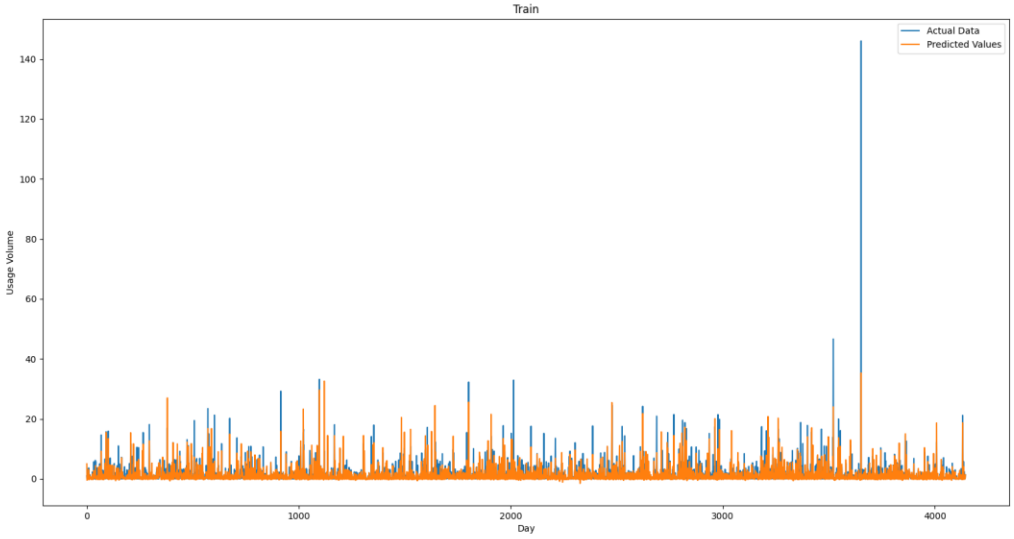
ارزیابی کرده، امتیاز R2 آن را محاسبه می‌کنیم، که برای داده‌های آموزش و صحه گذاری به ترتیب ۰.۵۷۷۳ و ۰.۷۶۵۷ به دست می‌آید. علاوه بر آن نمودار تغییرات تابع هزینه برای داده‌های آموزش و صحه گذاری را نیز برای ارزیابی روند آموزش رسم می‌کنیم که در شکل ۳ قابل مشاهده است. نهایتا مقادیر پیش‌بینی شده توسط شبکه برای داده‌های آموزش و صحه گذاری را در برابر مقادیر واقعی رسم می‌کنیم تا عملکرد شبکه به صورت بصری مشخص شود. این نمودارها نیز در شکل ۴ و شکل ۵ قابل مشاهده هستند.

پس از ارزیابی عملکرد شبکه با روش‌های ذکر شده، داده‌های هفته‌ی دوم به عنوان بخشی از ورودی به شبکه داده شده و حجم اینترنت مصرفی کاربران به عنوان خروجی شبکه گرفته شده و ذخیره می‌شود. این نتایج در فایل week2_results.csv قابل دسترسی هستند. تیم ما در این مرحله بر اساس معیار 1-SMAPE[26] امتیاز ۰.۴۲۳۲ را کسب نمود.

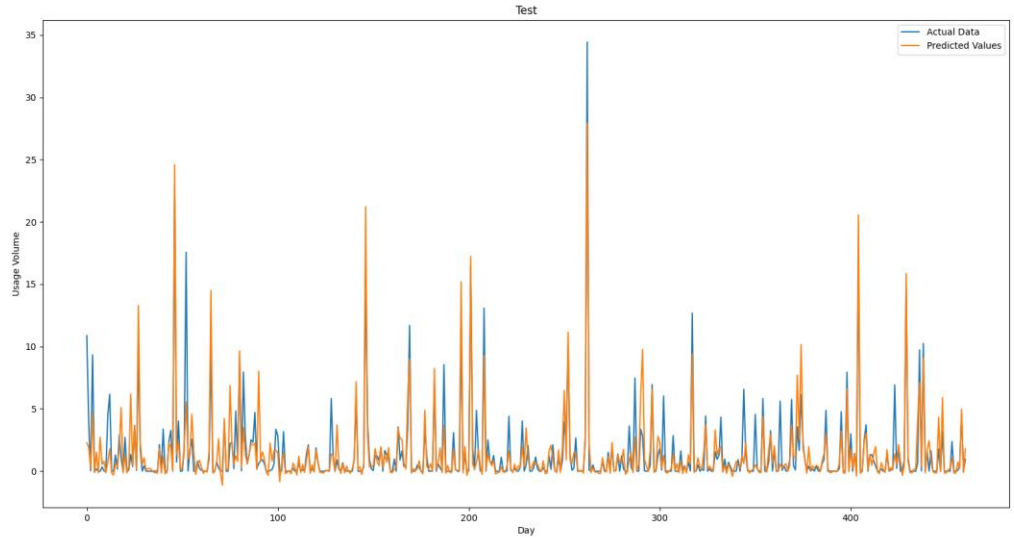


شکل ۳: روند تابع هزینه برای داده‌های آموزش و صحه گذاری در مرحله‌ی نیمه نهایی چالش

[26] Symmetric Mean Absolute Percentage Error



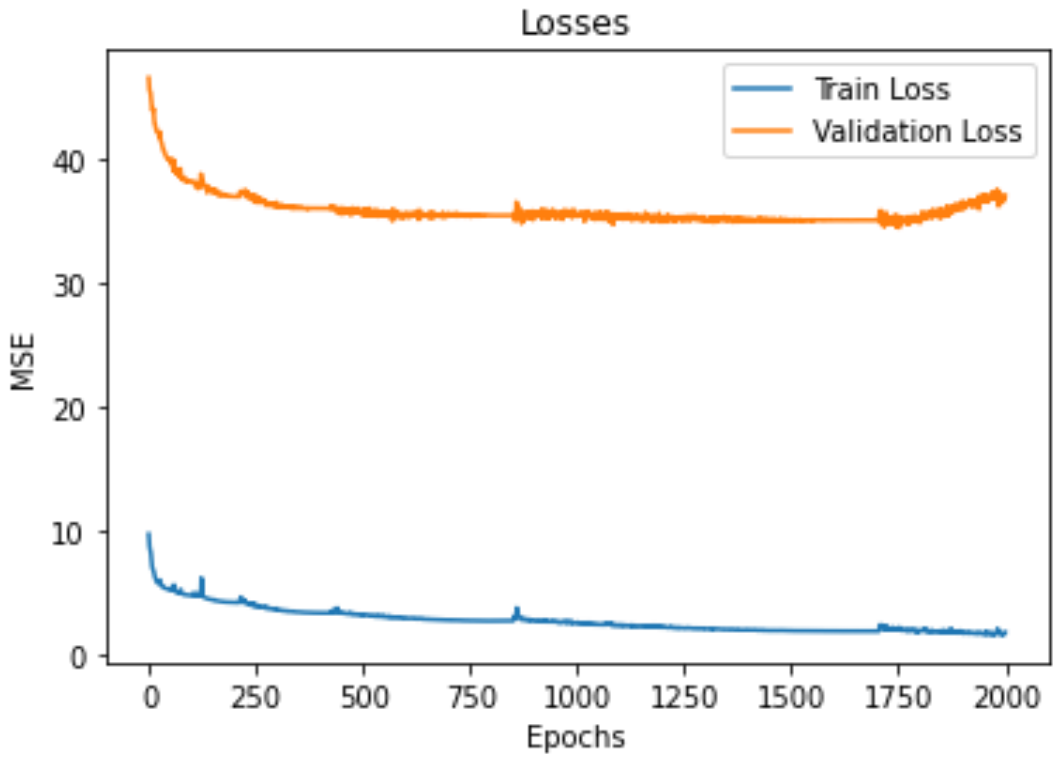
شکل ۴: – مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های آموزش مرحله‌ی نیمه نهایی چالش



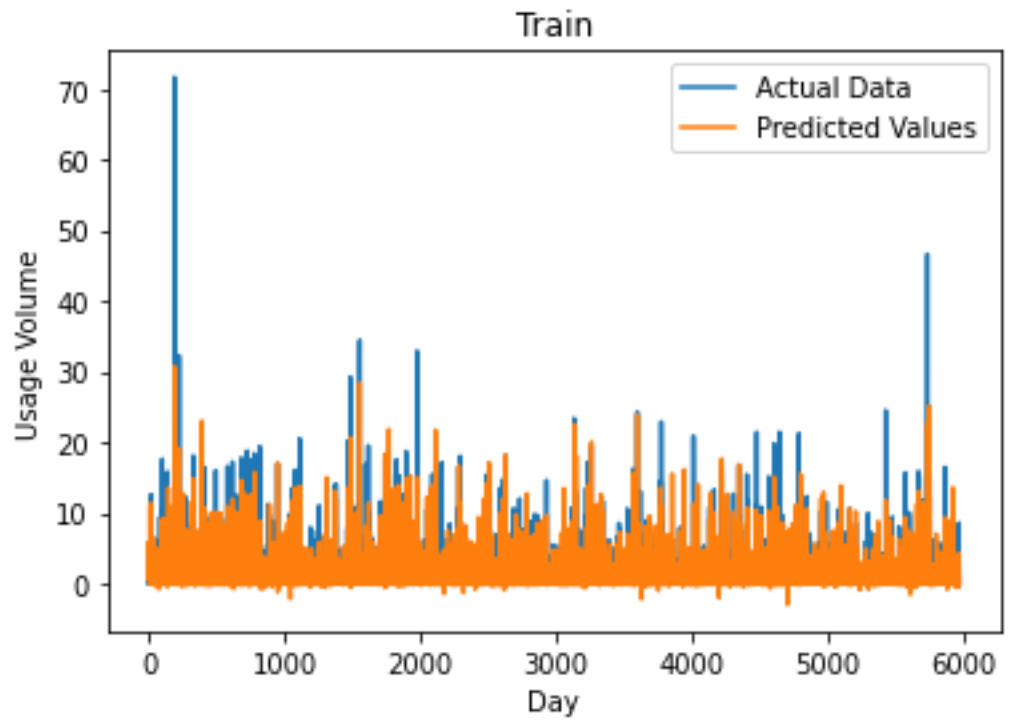
شکل ۵: – مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری مرحله‌ی نیمه نهایی چالش

۴-۲- مرحله‌ی نهایی چالش

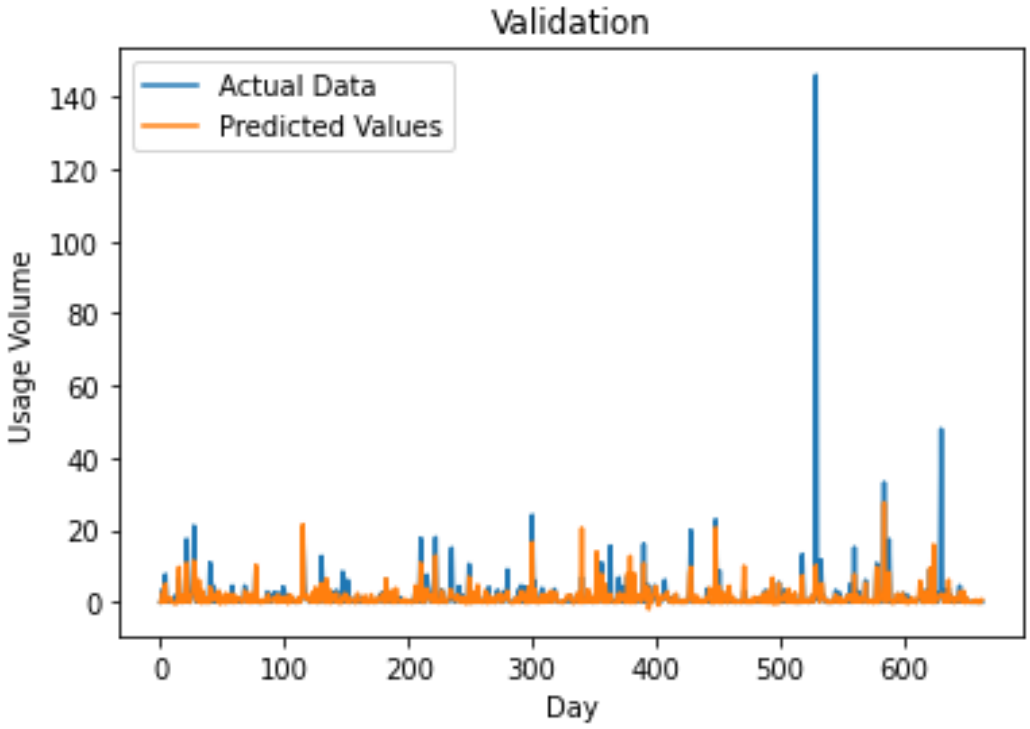
پس از اتمام یادگیری ابتدا برای بررسی عملکرد شبکه، آن را با استفاده از داده‌های صحه گذاری و آموزشی ارزیابی کرده، امتیاز R2 آن را محاسبه می‌کنیم، که برای داده‌های آموزش و صحه گذاری به ترتیب ۰.۷۴۱۶ و ۰.۲۴۷۹ به دست می‌آید. کم شدن امتیاز داده‌های صحه گذاری به دلیل داده‌های موجود در این مجموعه می‌باشد، که یک داده‌ی بسیار بزرگ باعث آن شده است. سپس برای بررسی مقاومت و عمومیت پذیری مدل، داده‌های هفته‌ی دوم به عنوان داده‌های تست بررسی می‌شوند، که امتیاز R2 آن نیز ۰.۵۷۱۲ به دست می‌آید. علاوه بر آن نمودار تغییرات تابع هزینه برای داده‌های آموزش و صحه گذاری را نیز برای ارزیابی روند آموزش رسم می‌کنیم که در شکل ۶ قابل مشاهده است. نهایتا مقادیر پیش‌بینی شده توسط شبکه برای داده‌های آموزش، صحه گذاری و تست را در برابر مقادیر واقعی رسم می‌کنیم تا عملکرد شبکه به صورت بصری مشخص شود. این نمودارها نیز به ترتیب در شکل ۷، شکل ۸ و شکل ۹ قابل مشاهده هستند.



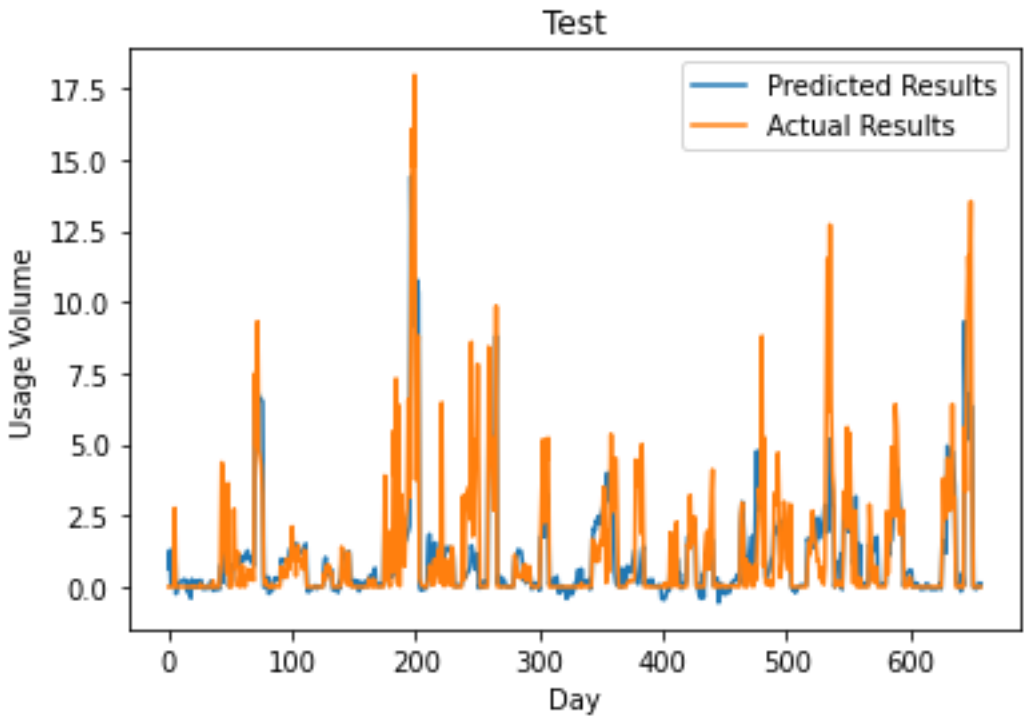
شکل ۶: روند تابع هزینه برای داده‌های آموزش و صحه گذاری مرحله‌ی نهایی چالش



شکل ۷: مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های آموزش مرحله‌ی نهایی چالش



شکل ۸: مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری مرحله‌ی نهایی چالش



شکل ۹: مقایسه‌ی داده‌های پیش بینی شده و واقعی برای داده‌های صحه گذاری مرحله‌ی نهایی چالش

پس از ارزیابی عملکرد شبکه با روش‌های ذکر شده، داده‌های هفته‌ی سوم به عنوان بخشی از ورودی به شبکه داده شده و حجم اینترنت مصرفی کاربران به عنوان خروجی شبکه گرفته شده و ذخیره می‌شود. این نتایج در فایل week3_results.csv قابل دسترسی هستند. تیم ما در این مرحله بر اساس معیار 1-SMAPE امتیاز ۰.۳۵۷۷ را کسب نمود.

۵- نتیجه گیری نهایی

با توجه به امتیازهای R2 کسب شده در مرحله‌ی نهایی، خصوصا برای داده‌های تست که خطای انباشته شده نیز بر روی آن تاثیر گذاشته است و همچنان امتیاز R2 آن بالای ۰.۵ بوده، می‌توان نتیجه گیری کرد که مدل طراحی شده و روش پیش پردازش و آماده سازی داده‌ها به خوبی انجام شده و نتایج مطلوبی ارائه داده است. همچنین در شکل ۹ نیز می‌توان مشاهده کرد که پیش‌بینی‌های شبکه در داده‌های تست، عملکرد خوبی نشان داده است. علاوه بر این باید در نظر گرفت که ویژگی‌های داده شده در داده‌های این چالش، نواقص زیادی دارند و با افزایش تعداد این ویژگی‌ها می‌توان به نتایج بهتری نیز دست یافت. برای مثال از عوامل تاثیرگذار در مصرف اینترنت کاربران، تاریخ، فصل، مناسبت‌های تاریخی و همچنین روزهای هفته هستند، که در این داده‌ها موجود نیستند. لذا می‌توان نتیجه گرفت که روش پیش پردازش داده‌ها که در این تحقیق ارائه شد و همچنین ساختار

شبکه‌ی عصبی بازگشتی LSTM با استفاده از روش پیش‌بینی مکرر در پیش‌بینی چند گام زمانی آینده، گزینه‌ی مناسبی برای پیش‌بینی میزان مصرف اینترنت کاربران با توجه به تاریخچه‌ی مصرف آنها می‌باشد.

۶- منابع

[1] S. Sarmadi, M. Li, and S. Chellappan, “A Statistical Framework to Forecast Duration and Volume of Internet Usage Based on Pervasive Monitoring of NetFlow Logs,” in *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, May 2018, pp. 480–487, doi: 10.1109/AINA.2018.00077.

[2] M. Masdari and A. Khoshnevis, “A survey and classification of the workload forecasting methods in cloud computing,” *Cluster Comput.*, vol. 23, no. 4, pp. 2399–2424, Dec. 2020, doi: 10.1007/s10586-019-03010-3.

[3] T. H. H. Aldhyani and M. R. Joshi, “Integration of time series models with soft clustering to enhance network traffic forecasting,” in *2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Sep. 2016, pp. 212–214, doi: 10.1109/ICRCICN.2016.7813658.

[4] D. H. L. Oliveira, F. M. V. Filho, T. P. de Araujo, J. Celestino, and R. L. Gomes, “Adaptive Model for Network Resources Prediction in Modern Internet Service Providers,” in *2020 IEEE Symposium on Computers and Communications (ISCC)*, Jul. 2020, pp. 1–6, doi: 10.1109/ISCC50000.2020.9219550.

[5] Bo Zhou, Dan He, and Zhili Sun, “Traffic predictability based on ARIMA/GARCH model,” in *2006 2nd Conference on Next Generation Internet Design and Engineering, 2006. NGI ’06.*, pp. 200–207, doi: 10.1109/NGI.2006.1678242.

[6] A. Bayati, V. Asghari, K. Nguyen, and M. Cheriet, “Gaussian Process Regression Based Traffic Modeling and Prediction in High-Speed Networks,” in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–7, doi: 10.1109/GLOCOM.2016.7841857.

[7] J. Violos, S. Tsanakas, T. Theodoropoulos, A. Leivadeas, K. Tserpes, and T. Varvarigou, “Hypertuming GRU Neural Networks for Edge Resource Usage Prediction,” in *2021 IEEE Symposium on Computers and Communications (ISCC)*, Sep. 2021, pp. 1–8, doi: 10.1109/ISCC53001.2021.9631548.

[8] A. Bayati, K. Khoa Nguyen, and M. Cheriet, “Multiple-Step-Ahead Traffic Prediction in High-Speed Networks,” *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2447–2450, Dec. 2018, doi: 10.1109/LCOMM.2018.2875747.

[9] J. Violos, E. Psomakelis, D. Danopoulos, S. Tsanakas, and T. Varvarigou, “Using LSTM Neural Networks as Resource Utilization Predictors: The Case of Training Deep Learning Models on the Edge,” 2020, pp. 67–74.

[10] Wikipedia user: Ixany, “Long short-term memory unit,” *Wikipedia*, 2017. https://en.wikipedia.org/wiki/Recurrent_neural_network#/media/File:Long_Short-Term_Memory.svg.