((به نام خدا))

گزارش مسابقه پیش‌بینی مصرف اینترنت کاربران همراه اول

اعضای گروه:

زهرا دهقانیان، جواد ظهرابی

سعید سراوانی، علیرضا عموزاد

زمستان ۱۴۰۰

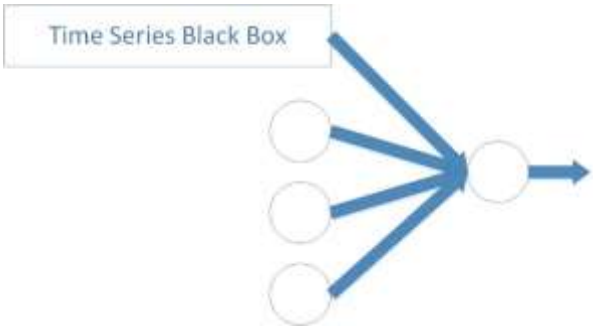
۱- روند پیش‌برد پروژه

در ابتدا پیاده‌سازی این پروژه به بررسی دقیق‌تر ساختار داده پرداختیم، این دادگان شامل دو نوع ویژگی Nominal و Categorical می‌باشد. در ادامه دسته‌بندی این ویژگی‌ها می‌پردازیم:

- **Nominal**: day, subscriber_age, months_of_subscription, subscriber_total_expenses, nonpackage_voice_expenses, package_voice_noncash_expenses, call_in_network_duration, call_off_netwrok_duration, nonpackage_call_in_network_expenses, nonpackage_call_off_network_expenses, total_call_duration, #inter_operator_calls, xyz_score, hxr_score, data_cash_expenses, nonpackage_data_expenses, package_data_noncash_expenses, subscriber_data_expenses, subscriber_nondata_expenses, #activated_monthly_data_packages, #activated_short_term_data_packages, #activated_type_one_data_packages, #activated_type_two_data_packages, #activated_type_three_data_packages, data_usage_volume

- **Categorical**: subscriber_ecid, subscriber_gender, registration_province, most_used_province, is_usage_nonzero, is_voice_expenses_nonzero, is_voice_usage_nonzero, is_data_expenses_nonzero, is_data_usage_nonzero

در دسته Nominal نیز دادگان به صورت سری زمانی و یا غیر سری زمانی هستند. مدل پیشنهادی برای آموزش مقدار هدف با توجه به نوع دادگان به صورت زیر می‌باشد:



در بخش time series black Box دادگان سری زمانی را به عنوان ورودی می‌دهیم و خروجی این بخش را به کمک یک یا چند لایه Dense با ویژگی‌های Categorical ترکیب می‌کنیم. برای پیشنهاد مدل به روز و بهینه ابتدا مجموعه داده موجود را با دیتاست‌های شاخص موجود مقایسه کردیم، شبیه‌ترین دیتاستی که توانستیم

بدان دسترسی داشته باشیم، دیتاست ETTh (پیش‌بینی مصرف برق) بود. در ادامه بهترین مدل‌هایی که با اختلاف نتایج قابل ملاحظه‌ای داشتند را انتخاب و آزمایش کردیم.

برای بخش سری زمانی شبکه عصبی، چهار مدل مختلف را آزمایش می‌کنیم که در ادامه به بررسی جزئیات هر یک از این مدل‌ها خواهیم پرداخت:

۱-۱- مدل SCINet

این مدل بر روی دادگان ETTh در حالت Multivariate به خطای MSE ۰.۴۹۷ و در حالت Univariate به خطای MSE ۰.۰۷۵ می‌رسد که رتبه اول را در میان مدل‌های موجود بدست آورده است.

در اینجا نیز این مدل را در هر دو حالت Multivariate و Univariate بررسی می‌کنیم.

	SCINet
univariate	2.83
multivariate	3.14

همانطور که مشاهده می‌شود، برخلاف انتظار ما خطای مدل در حالت Univariate کمتر بوده و ویژگی‌های اضافه عملا کمکی به بهبودی تخمین نمی‌کند.

۱-۲- مدل Informer

این مدل یکی مدل‌های بسیار خوب بر روی دیتاست مورد ماست و برترین مقاله AAAI 2021 می‌باشد، در این مدل نیز به نتیجه فوق در خصوص نتیجه بهتر برای حالت Univariate می‌رسیم. نتایج عملکرد این مدل به صورت زیر می‌باشد:

	Informer
univariate	3.25
multivariate	3.85

۱-۳- مدل LSTM

این مدل در ابتدا برای بررسی بیشتر و explore دادگان به کار گرفته شد و همانطور که انتظار می‌رفت بیشترین خطا برای این مدل به مقدار ۴.۳ بدست آمد.

۱-۴- مدل ARIMA

این مدل یکی از مدل‌های قدیمی و پرقدرت حوزه سری زمانی می‌باشد که در ابتدای بررسی‌ها کمتر مورد توجه این تیم بود اما با دستیابی به بهترین نتایج عملا قدرت خود را نشان داده و به مدل نهایی برای این مسابقه تبدیل شد. در ابتدا یک مدل برای همه افراد موجود در دیتاست آموزش داده شد اما با هدف شخصی سازی و حفظ بیشتر ویژگی‌های فردی مدل به ۹۵ زیر مدل شکسته شد، بدین ترتیب که برای هر فرد مقادیر p، q و r بهینه پیدا شد و به کمک order بدست آمده با کمترین خطا، تخمین‌های نهایی محاسبه شد. مقادیر خطا برای این مدل به صورت زیر می‌باشد.

	ARIMA
general	2.035
personal	1.76

❊❊ تمامی کد مدل‌های مختلف به همراه نتایج بر روی آدرس گیت‌هاب زیر موجود می‌باشد:

https://github.com/zahraDehghanian97/Data-Usage-Prediction