چالش مرحله اول

پیش‌بینی تقاضای خرید بلیط هواپیما

تخصیص بهینه و متناسب با نیاز منابع و تجهیزات در صنایع و تجارت‌های مختلف بسیار حائز اهمیت است. داشتن یک پیش‌بینی دقیق و درست نسبت به نیازهای پیش‌رو می‌تواند یک گام بسیار مهم برای تخصیص بهینه‌ی منابع باشد. در همین راستا، امسال در دومین دوره مسابقات ملی داده‌کاوی امیرکبیر قصد داریم تا بتوانیم با استفاده از ابزارها و راه‌حل‌های نوین داده‌کاوی و یادگیری ماشین، تعداد درخواست مسافران برای خرید بلیط هواپیما بین شهرهای مختلف را در روزهای مختلف سال پیش‌بینی کنیم.

سناریو

داده‌های مربوط به حدود ۲ میلیون خرید بلیط هواپیما مربوط به خطوط مختلف هوایی موجود است. تاریخ خرید بلیط مربوط به سال‌های ۱۳۹۵ و ۱۳۹۶ می‌باشد. هر سطر داده شامل اطلاعاتی از جمله «تاریخ و زمان خرید بلیط»، «تاریخ و زمان پرواز»، «قیمت بلیط»، «شماره فرودگاه مبدا»، «شماره فرودگاه مقصد» و «شماره ایرلاین» می‌باشد.

با استفاده از این داده‌ها می‌توان تعداد خرید بلیط در هر روز را برای هر زوج «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» محاسبه کرد. در مرحله‌ی اول مسابقه، هدف پیش‌بینی تعداد خرید بلیط برای زوج‌های مختلفی از «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» در روزهای آتی است.

از شرکت کنندگان انتظار می‌رود با استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین مدل مناسبی بر اساس داده‌های ۲ سال گذشته طراحی کنند و از آن مدل برای پیش‌بینی دقیق متغیر هدف در روزهای مشخصی در سال ۱۳۹۷ استفاده کنند. فایل آزمون شامل حدود ۱۲۰۰۰ سطر است که شرکت‌کنندگان باید با توجه به «تاریخ خرید بلیط»، «شماره فرودگاه مبدا» و «شماره فرودگاه مقصد» مربوطه، پیش‌بینی خود را از تعداد خرید بلیط در آن تاریخ و بین شهر مبدا و مقصد ارائه کنند.

داده‌ها

داده‌های ارائه شده مربوط به خریدهای بلیط از سایت علی‌بابا است که به همراه صورت مسئله در سایت بارگذاری شده است. در ادامه چند نکته در مورد فایل‌های ضمیمه شده آورده شده است:

۱. هر سطر از داده‌ها نشانگر یک درخواست خرید بلیط است.
۲. خط اول هر فایل داده حاوی نام ویژگی‌ها است.
۳. در تمام سطرها، علامت "," ستون‌ها را از هم جدا می‌کند.
۴. علامت "." نشانگر جدا کننده اعشار است.

فایل "AUT DMC 2018 - Features.pdf" موجود در "task.zip" حاوی لیستی از تمام ویژگی‌ها و توضیحات مربوط به آن ویژگی‌ها است.

دقت کنید که امکان وجود داده‌ی پرت وجود دارد و تیم‌ها بنا به صلاحدید خود می‌توانند هر پردازشی روی آن‌ها انجام دهند.

بارگذاری فایل پیش‌بینی

تیم‌ها در هر روز می‌توانند حداکثر **۵** فایل پیش‌بینی بارگذاری کرده و امتیاز خود را در کنار امتیاز تیم‌های دیگر مشاهده کنند. امکان بارگذاری نتایج تا **۱۴ بهمن** وجود خواهد داشت.

نام ستون	توضیحات	محدوده‌ی مقادیر
Log_Date	تاریخ درخواست خرید	YYYY/MM/DD
FROM	مبدا مسیر	اعداد صحیح مثبت
TO	مقصد مسیر	اعداد صحیح مثبت
Sales	تعداد درخواست خرید	اعداد صحیح مثبت

فایل نهایی باید دقیقا با فرمت بالا باشد. مثال زیر یک بارگذاری صحیح را نشان می‌دهد:

Log_Date,FROM,TO,Sales

1397/01/01,3,30,10

1397/01/01,3,49,10

1397/01/01,3,66,10

...

برای فایل نهایی از فرمت csv استفاده شود.

معیار ارزیابی

برای هر ترکیب "Log_Date, FROM, TO" که در فایل "test.csv" داده شده است، مقدار "Sales" باید پیش‌بینی شود. ارزیابی با معیار MAPE (Mean Absolute Percentage Error) انجام خواهد پذیرفت.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right|$$

A_i و P_i نشاندهنده‌ی مقادیر واقعی و پیش‌بینی شده هستند.

گزارش و کد

هر تیم باید اقدام به تهیه‌ی گزارشی در مورد روش حل مسئله و مدل انتخابی و همچنین دید بیزینسی مسئله کند. بررسی دید بیزینسی از اهمیت بیشتری برخوردار است. همچنین کد استفاده شده برای تولید نتایج نیز باید بارگذاری شود تا امکان بررسی صحت کار و جلوگیری از تقلب وجود داشته باشد. امکان بارگذاری گزارش و کد تا **۱۵ بهمن** وجود خواهد داشت.

ارزیابی نهایی

امتیاز نهایی تیم‌ها با در نظر گرفتن ضریب ۸۰٪ برای معیار ارزیابی MAPE و ۲۰٪ برای گزارش نهایی محاسبه خواهد شد. برای این منظور MAPEهای پرت حذف شده و تیم‌ها با توجه به MAPE کسب شده از ۰ تا ۸۰ نمره خواهند گرفت که با نمره‌ی گزارش جمع شده و نمره‌ی نهایی به دست خواهد آمد.