به نام او که تاریخ بلند کبریاست

دانشگاه صنعتی شریف
دانشکده مهندسی صنایع

راهنمای پروژه – فاز اول
درس برنامه‌ریزی حمل‌ونقل

استاد درس: دکتر حسن‌نایبی

پاییز 1403

توضیحات داده‌ها

مجموعه داده‌ای همراه فایل راهنما پروژه داده شده است (**لینک داده‌ها**). این مجموعه داده شامل سوابق سفرهای تاکسی سبز است که توسط ارائه‌دهندگان خدمات فناوری تاکسی (TSPs) ثبت شده‌اند. هر ردیف نمایانگر یک سفر منحصر به فرد در یک تاکسی سبز است. داده‌ها اطلاعات حیاتی مانند زمان‌های شروع و پایان سفر، مکان‌های سوار و پیاده شدن، فاصله طی شده، مقدار کرایه و نوع پرداخت را شامل می‌شود. همچنین مالیات‌ها و هزینه‌های اضافی مختلفی که به هر سفر اعمال می‌شود را ثبت می‌کند. این مجموعه داده برای تحلیل الگوهای استفاده از تاکسی، ساختارهای کرایه، حجم مسافران و دینامیک حمل و نقل در مناطقی که تاکسی‌های سبز فعالیت می‌کنند، بسیار ارزشمند است. توضیحات هر ستون‌های این مجموعه داده به شرح زیر است.

VendorID: یک کد عددی که ارائه‌دهنده خدمات فناوری (TSP) را که مسئول ثبت سفر بوده، مشخص می‌کند.

lpep_pickup_datetime: زمان و تاریخ شروع سفر و زمانی که تاکسی‌متر روشن شده است.

lpep_dropoff_datetime: زمان و تاریخ پایان سفر و زمانی که تاکسی‌متر خاموش شده است.

store_and_fwd_flag: یک علامت متنی که نشان می‌دهد آیا رکورد سفر به دلیل عدم اتصال به سرور، موقتاً در حافظه وسیله نقلیه ذخیره شده است یا خیر؛ "Y" به معنی ذخیره داده و "N" به معنی عدم ذخیره است.

RatecodeID: یک کد عددی که نوع نرخ کرایه استفاده شده برای سفر را نشان می‌دهد (مانند نرخ استاندارد، نرخ مذاکره شده، سفر گروهی و غیره). این ستون در محاسبه کرایه تأثیر دارد.

PULocationID: یک کد عددی که ناحیه تاکسی TLC (کمیسیون تاکسی و لیموزین) را که سفر از آنجا شروع شده است (محل سوار شدن) مشخص می‌کند.

DOLocationID: یک کد عددی که ناحیه تاکسی TLC را که سفر در آنجا پایان یافته است (محل پیاده شدن) نشان می‌دهد.

passenger_count: تعداد مسافران گزارش شده توسط راننده برای سفر در این ستون نشان داده شده است.

trip_distance: فاصله کل سفر، به مایل، همانطور که توسط تاکسی‌متر ثبت شده است.

fare_amount: کرایه‌ای که بر اساس زمان و مسافت سفر توسط تاکسی‌متر محاسبه شده است، که در آن هزینه‌های اضافی مانند انعام یا مالیات در نظر گرفته نشده است.

Extra: این ستون هزینه‌های اضافی و گوناگون مانند هزینه اضافه ۰,۵۰ دلار برای ساعات شلوغی یا ۱ دلار برای ساعات شب را ثبت می‌کند.

mta_tax: مالیاتی که به صورت خودکار بر اساس نرخ متری استفاده شده در سفر اعمال می‌شود.

tip_amount: انعامی که به راننده داده شده است و به صورت خودکار برای انعام‌های کارت اعتباری ثبت می‌شود. انعام‌های نقدی در این ستون گنجانده نشده‌اند.

tolls_amount: مجموع کل مبالغ پرداخت شده برای عوارض جاده‌ای در طول سفر در این ستون آمده است.

ehail_fee: این ستون برای هزینه‌های مربوط به درخواست‌های الکترونیکی تاکسی (ehail) اختصاص دارد.

improvement_surcharge: هزینه اجباری ۰,۳۰ دلاری که از سال ۲۰۱۵ برای بهبود خدمات تاکسی در ابتدای هر سفر اخذ می‌شود.

total_amount: مجموع مبلغی که از مسافر برای سفر اخذ شده است، شامل کرایه، هزینه‌های اضافی، مالیات و عوارض می‌شود. انعام‌های نقدی در این محاسبه گنجانده نشده‌اند.

payment_type: یک کد عددی که نشان می‌دهد مسافر چگونه هزینه سفر را پرداخت کرده است. مثال‌ها شامل پرداخت با کارت اعتباری، نقدی یا بدون هزینه است.

trip_type: یک مقدار عددی که نوع سفر (مانند محلی، فرودگاهی) را نشان می‌دهد. این ستون به دسته‌بندی سفرها برای ساختارهای مختلف کرایه یا خدمات کمک می‌کند.

congestion_surcharge: این ستون هزینه مربوط به ازدحام ترافیکی را ثبت می‌کند که برای مدیریت ترافیک در مناطق خاص و در ساعات اوج تراکم اعمال می‌شود.

خواسته‌های فاز اول پروژه

خواسته اول – پیش‌پردازش داده‌ها

1) ستون‌های جدول زیر نباید در تحلیل‌ها مورد استفاده قرار بگیرند.

RatecodeID	store_and_fwd_flag	improvement_surcharge	ehail_fee
mta_tax	extra	fare_amount	congestion_surcharge

2) اگر در سفری، انعام پرداخت شده است، نحوه پرداخت با کارت اعتباری بوده است. مقادیر خالی را با آن پر کنید.
3) اگر تعداد مسافران از 4 نفر بیشتر است، از ون استفاده شده است و نوع سفر dispatch و نوع پرداخت unknown است. این موارد را در داده‌ها اعمال کنید.
4) سطرهایی که مقدار VendorID آن‌ها خالی است را حذف کنید.
5) سطرهایی که سال‌ها، غیر از سال 2021 است را حذف کنید.

خواسته دوم – توصیف و تفسیر داده‌ها

1) یک نمودار نقشه گرمایی[1] رسم کرده و در یک پاراگراف تحلیل خود را از زمان‌های پیک و نحوه تقسیم‌بندی روز به کمک آن‌ها شرح دهید.
2) با رسم نمودارهایی، روند پرداخت مسافران در گذر زمان، روند در نوع سفر، روند در استفاده از تاکسی سبز در هر یک از بخش‌های شهر[2] و روند استفاده در بازه‌های زمانی روز استخراج کرده و به شکل مختر (یک پاراگراف کوتاه) هر یک را تحلیل کنید.

خواسته سوم – تحلیل دقیق داده‌ها

1) ماتریس همبستگی ستون‌ها مورد استفاده در تحلیل را استخراج کرده و تحلیل همبستگی با هدف پیش‌بینی قیمت نهایی ارائه کنید.
2) با اجرای روش‌های backward, forward, chi-sq test, random forest importance با در نظر داشتن هدف **پیش‌بینی قیمت**، ستون‌های موثرتر برای ساخت مدل رگرسیونی را استخراج کنید. باتوجه به داده‌هایی که در دست است، تحلیل کنید که کدام روش بهتر است.
3) دو مدل پیش‌بینی برای دسته‌بندی پرداخت یا عدم پرداخت انعام بسازید. به این منظور می‌توانید از بین اگوریتم‌های XGBoost, RandomForest، Decision Tree استفاده کنید.
4) دو مدل شامل یک مدل رگرسیونی و یک مدل تقویت‌شده (XGBoost, RandomForest, …) برای پیش‌بینی قیمت نهایی سفر توسعه دهید.

[1] Heat Map
[2] Borough

5) حداقل سه پارامتر مدل‌های گفته شده در خواسته سوم و چهار را با روش‌هایی مانند Taguchi یا Grid Search بهینه کنید و سپس نتایج مدل‌ها به شکل جدا هر یک در یک پاراگراف کوتاه تحلیل کنید. برای بهینه‌یابی پارامترهای این مدل‌ها، نیازی نیست از کل داده‌ها استفاده کنید و استفاده از تنها 5 درصد داده‌ها کافی است.

* توسعه مدل‌های ترکیبی و خلاقانه برای **خواسته‌های سوم و چهارم** شامل نمره امتیازی خواهد شد و لازم است رفرنس ها و مقالاتی که از آن استفاده کرده اید را حتما در پاسخ ذکر کنید.

نکات قابل توجه

- ❖ به جهت توضیح خواسته‌ها و روش‌هایی که در پروژه آمده و در کلاس مطرح نشده‌اند، کلاس حل‌تمرین برگزار خواهد شد.
- ❖ تمامی موارد خواسته شده را با استفاده از کد اجرا کنید و از دست بردن در فایل داده‌ها خودداری کنید.
- ❖ گروه‌های پروژه می‌تواند حداکثر شامل **سه نفر** باشد.
- ❖ برای پاسخ خواسته‌های پروژه، یک فایل گزارش به فرمت ”.pdf“ با رعایت اصول نگارش یک گزارش علمی و یک فایل کد پایتون (فرمت ”.py“ یا ”.ipynb“) با کامنت‌گذاری‌های حداقلی جهت تصحیح دقیق‌تر و راحت‌تر پروژه در سامانه CW پیش از مهلت پروژه بارگذاری کنید.
- ❖ بارگذاری پوشه زیب شده، شامل فایل‌های مذکور، تنها توسط یکی از اعضای گروه کافی است.
- ❖ فرمت ”TP Project – Phase I – [student_number_1, student_number_2, student_number_3]“ برای نام‌گذاری پوشه استفاده شود.