دانشگاه صنعتی شریف

دانشکده مهندسی صنایع

پروژه درس برنامه‌ریزی حمل‌ونقل

فاز یک

پاییز 1402

فاز 1

تجزیه و تحلیل داده‌ها و پیش‌بینی سری زمانی

نکات پروژه

- مهلت ارسال فاز اول، 1 آذرماه خواهد بود.
- پروژه به صورت گروهی بوده و تعداد اعضا مجاز برای هر گروه حداکثر 3 نفر است.
- گزارش پروژه با فرمت صحیح گزارش‌نویسی و کد(ها) پروژه، خروجی‌هایی هستند که باید در سامانه CW بارگذاری شوند.
- بارگذاری فایل زیپ پروژه تنها توسط **یک نفر** از اعضای گروه کافی است.
- فرمت صحیح نام‌گذاری فایل زیپ پروژه به صورت [student numbers] Phase I است.

توضیحات داده‌های پروژه

داده‌های قرار گرفته مربوط به یک مرکز اجاره دوچرخه شهر سئول[1] هستند. با رشد روزافزون تقاضا اجاره دوچرخه، فراهم کردن عرضه پایدار خودرو برای اجاره تبدیل به چالشی برای این بنگاه شده است. برای این بنگاه بسیار پر اهمیت است که در راستای حداقل کردن زمان انتظار مشتریان تعداد مناسب دوچرخه را تدارک ببیند؛ زیرا این موضوع زمان انتظار را کاهش می‌دهد. بنابرین پیش‌بینی تقاضای دوچرخه می‌تواند در راستای رسیدن به این هدف راهگشا باشد. مجموعه داده‌هایی که در اختیار قرار گرفته است، مربوط به تعداد دوچرخه‌های اجاره گرفته شده در هر ساعت در هر روز است. این مجموعه داده حاوی ویژگی‌های زیر است:

ویژگی	توضیح ویژگی
Date	تاریخ
Rented Bike Count	تعداد دوچرخه های اجاره شده
Hour	ساعت اجاره
Temperature(°C)	دما (سنتیگراد)
Humidity (%)	رطوبت هوا(درصد)
Wind speed (m/s)	سرعت وزش باد(متر بر ثانیه)
Visibility (10m)	اندازه فاصله ای که در آن جسم یا نور به وضوح در واحدهای 10 متری قابل تشخیص است
Dew point temperature(°C)	دمای ثبت شده در ابتدای روز بر حسب سلسیوس
Solar Radiation (MJ/m2)	شدت نور خورشید
Rainfall(mm)	میزان بارندگی دریافتی بر حسب میلی متر
Snowfall (cm)	میزان بارش برف بر حسب سانتی متر
Seasons	فصل
Holiday	آیا آن روز تعطیل است؟
Functioning Day	اینکه آیا سرویس اجاره در دسترس است (Yes–ساعات کارکردی) یا خیر (ساعت های غیرکارکردی -NO)

[1] SeoulBikeData

سوالات زیر با شرح کامل پاسخ دهید.

1. توزیع تقاضای اجاره دوچرخه به تفکیک در **ساعات مختلف** روز، **فصل‌های مختلف** سال و **روزهای مختلف** هفته چگونه است؟ چه تحلیلی برای شکل این توزیع‌ها می‌توانید ارائه دهید؟
2. نمودار Boxplot برای تقاضای اجاره دوچرخه در روزهای تعطیل و غیر تعطیل رسم کنید، آیا تفاوت معنا دار آماری بین این دو مقدار وجود دارد؟ به منظور فهمیدن این مسئله ابتدا از نرمال بودن داده‌ها با کمک آزمون Anderson-Darling اطمینان حاصل کنید و سپس مبتی بر خروجی بد ست آمده، تست آماری منا سب را اجرا و نتیجه را تحلیل کنید. (در صورت نرمال بودن داده‌ها، می‌توانید از Two sample t-test و در صورت غیرنرمال بودن از Kruskal-Wallis استفاده کنید. لازم به ذکر است که تمامی این آزمون‌ها را می‌توانید با کمک کتابخانه SciPy در پایتون اجرا کنید.)
3. با استفاده از تحلیل همبستگی و الگوریتم Lasso متغیرهایی را که بهترین نحو می‌توانند رفتار تقاضا را پیش‌بینی کنند، انتخاب کنید.
4. برای پیش‌بینی متغیر هدف می‌توان از روش‌های آماری، یادگیری ماشین استفاده کرد. برای پیش‌بینی با کمک روش‌های آماری از مدل ARIMA، با روش‌های یاگیری ماشین از مدل رگرسیون خطی Regression و مدل جنگل تصادفی Random Forest استفاده کنید. با کمک تحلیل‌های خواسته سوم، بهترین متغیرها (feature) را با استفاده از تحلیل‌های همبستگی و Lasso بدست آورده‌اید. حال در اجرای مدل Regression از همان featureهای منتخب استفاده کنید ولی در اجرای الگوریتم Random Forest از تمام featureها استفاده کنید. برای سنجش عملکرد مدل‌ها می‌توانید از شاخص $R^2 - adj$ و میانگین مربع خطا (MSE) استفاده کنید. توضیح دهید که از چه روش‌های برای پیشگیری از Overfit شدن مدل استفاده می‌کنید؟
5. در ادامه و پس از پاسخ دادن به سوالات بالا، بهترین مدل ممکن (بالاترین R^2 که بیش‌برازش نشده باشد) را برای پیش‌بینی این سری زمانی برای **سه روز پس از آخرین تاریخ داده‌ها** انجام دهید.