Final Project: Predicting New Bike Shares

Due date: July 6th

In this final project, you will conduct exploratory analysis on a bike sharing dataset to understand the effect of time, weather and temperature on the number of new bike shares. Based on the analysis, you will construct a prediction model to estimate the number of new bike shares. You can work on this project either *individually or in groups of two.*

Dataset Description

The data is provided to you in NYU Brightspace. It contains 17414 records in total, ranging from 2015-01-04 to 2017-01-03. In each record, we have the following columns:

"timestamp" - timestamp field for grouping the data
"cnt" - the count of new bike shares
"t1" - real temperature in Celsius
"t2" - "feels like" temperature in Celsius
"hum" - humidity in percentage
"windspeed" - wind speed in km/h
"weathercode" - category of the weather (1-clear; 2-scattered clouds; 3-broken clouds; 4-cloudy; 7-light rain; 10-rain with thunderstorm; 26-snowfall; 94-freezing fog)
"isholiday" – binary variable - 1 holiday / 0 non holiday
"isweekend" - binary variable - 1 if the day is weekend / 0 non weekend
"season" – category of the season (0-spring ; 1-summer; 2-fall; 3-winter)

You are expected to split the dataset into the training set (data records between 2015-01-04 and 2016-09-30) and the test set (data records between 2016-10-01 and 2017-01-03)

Your goal is to learn a prediction model using the data records in the training set to predict the number of new bike shares (the “cnt” column) in the test set.

Deliverables:

1. Project Report (in the form of Jupyter Notebook or PDF)

Please include your **codes and descriptions** in the report. You are expected to **clearly explain** the design of your prediction model, and **justify the selection** of your model and related hyperparameters. In addition, you are expected to include **all prediction or analytical results** in the report, such as your failed attempt or inferior models that you have tried. Besides, you need to explain why these models do not work well, and what efforts you have taken to improve the performance.

To help you better prepare the project report, we have provided a set of suggested questions in the following for you to explore and answer in the report. You are not limited to these questions and any additional effort will be welcome.

Exploratory Analysis

1. Which feature is the most associated with the number of bike shares?
2. What bike sharing patterns do you observe in different time of day, day of week, month or holiday?
3. What is the relation between weather situations and the number of bike shares? And what is your explanations?

Predictive Model Construction

1. Which model is the most suitable for this bike sharing prediction task? And why?
2. What is the learning objective?
3. How would you select the optimal hyperparameter values (if applicable)?

Performance Analysis

1. What are the appropriate evaluation metrics?
2. How to further improve the prediction performance?
3. What are the business insights that you have obtained through the analysis?
4. How would you explain your findings to the manager of the bike sharing company?

Your term project report will be judged based on the **business insights** you have obtained through exploratory analysis, **the efforts you have made** to try different types of prediction models, your **analysis on the pros & cons** of each prediction model, the **clarity** of your prediction model design (including parameter details and justifications), and the final **prediction performance** of the new bike shares.

2. Project Representation (in the form of PPT)

Besides the project report, you are expected to make a **10-minute presentation** to the class on July 6th. You are expected to explain the **key insights** you have obtained from your analysis, the final **prediction model** that you have selected, and the **prediction performance**. You are also expected to take questions from your colleagues.

Your project presentation will be judged based on the **clarity** of the business insights, the **coherence** of the presentation, the **cohesiveness** of the takeaways, and the **ability to handle questions** from the audience.