Project Introduction

Your Project for the Week

Goals:

- Use Capital Bike Share Kaggle data set to predict hourly demand for bicycle rentals based on time and weather, e.g.

> "Given the forecasted weather conditions, how many bicycles can we expect to be rented out (city-wide) this Saturday at 2pm?"

- Submit your predictions to Kaggle.

Notes:

- We're predicting a continuous variable, not a class!
- We're predicting the `count` column. It is a sum of `casual` and `registered` columns. The latter two should not be in our features!

Data:

`train.csv` : Your dataset for this weeks project. This is the dataset you are training and evaluating your model on. *It contains the first 19 days of every month.*

`test.csv` : The dataset Kaggle uses to evaluate your model. *It contains days 20 to end of month.*

Evaluation:

- The metric you are optimizing for this week is RMSLE, Root Mean Squared Log Error.

    - Unlike last week, *smaller value is better*!
    - Penalizes under-estimates more than over-estimates
- These are the scores you can expect from the model you build:
    - EASY: `~1.1`
    - MEDIUM: `<0.9`
    - MEDIUM-HARD: `<0.7`
    - HARD: `<0.5`
    - VERY HARD: `<0.45`

Machine Learning Workflow

1. Define business goal:
    - Create a model that predicts bike sharing demand; optimize for RMSLE.
2. Get data:
    - From Kaggle, or `week_03/data` folder on Github.
3. Train-test-split:
    - You already know this!
4. Explore data:
    - Find out what all the variables in your dataset mean
    - Extract time features from the DatetimeIndex
    - Inspect descriptive statistics of your features and the target variable
    - Visualize the relationships between single features and the target variable
    - Look at correlations
    - Think about how to use `ColumnTransformer` and `Pipeline` with your features
5. Train model:
    - This afternoon we'll look at linear regression

6-...

- Later this week