Stat 220 Lab 4

Background

With the increasing focus on sustainability and public health, bike-sharing services have gained immense popularity. A local bike-sharing company named CycleHub aims to better understand the factors that influence the demand for their bikes. Accurate estimation and prediction can help them allocate resources more efficiently, thereby optimizing operations and maximizing customer satisfaction.

Scenario

CycleHub collects data on the number of bikes rented per hour, along with several associated variables like temperature, humidity, wind speed, and whether it is a holiday or not. This data is publicly available at `https://richardson.byu.edu/220/bike_sharing_data.csv`.

CycleHub is particularly interested in the following:

- Identifying the variables that significantly influence the demand for bikes.
- Estimating the impact of these significant variables.
- Creating predictive models that can forecast future demand accurately.

Your analysis will help to start the process of fulfilling these tasks.

Data Dictionary

Below is a description of each variable in the dataset:

Variable	**Type**	**Description**
bikes_rented	Continuous	Number of bikes rented during the hour.
temperature	Continuous	Temperature in degrees Fahrenheit during the hour.
humidity	Continuous	Humidity percentage during the hour (0-100%).
wind_speed	Continuous	Wind speed in miles per hour during the hour.
is_holiday	Categorical (0/1)	Indicator of whether the day is a holiday (1) or not (0).

The data contains both continuous and categorical variables that will be used to explore and model the number of bikes rented.

Your Role

As students of statistics and data science, you are tasked to help CycleHub analyze this data using simple and multiple regression as well as regression trees. Your findings will aid in decision-making processes and strategic planning.

Instructions

1. Load the provided dataset from `https://richardson.byu.edu/220/bike_sharing_data.csv`
2. Conduct exploratory data analysis to get a sense of the data.
3. Perform simple linear regression:
   (a) Choose one variable that you believe would most influence bike demand.
   (b) Estimate the parameters using Ordinary Least Squares (OLS).
   (c) Interpret the model (see below)
4. Perform multiple linear regression:
   (a) Add more variables to your model from the simple linear regression.
   (b) Estimate the new parameters.
   (c) Interpret the model (see below)
5. Implement a regression tree model:
   (a) Build a regression tree with the variables you found significant.
   (b) Interpret the tree. (see below)
6. Combined Prediction Exercise:
   (a) Use the following week's weather forecast to make daily predictions of the number of bikes rented. Note that all temperatures are in °F, humidity in %, and wind speed in mph. Additionally, indicate whether the day is a holiday (1) or not (0):

Day	**Temperature**	**Humidity**	**Wind Speed**	**Is Holiday**
Monday	72	45	10	0
Tuesday	68	50	8	0
Wednesday	75	40	12	0
Thursday	70	60	6	0
Friday	77	35	15	0
Saturday	80	30	20	0
Sunday	65	55	5	1

   (b) Use your models (simple linear regression, multiple linear regression, and regression tree) to make predictions for each day.
   (c) Create a visualization that displays the predicted number of bikes rented per day for each model. (see below)
   (d) Compare the predictions across your models and discuss any differences or similarities.
7. Submit a detailed report containing all your analyses, code, visualizations, and conclusions. Additionally, provide recommendations for CycleHub based on your findings, which they can implement to improve their service.

Interpreting the Models

Linear Regression Models

When interpreting the coefficients of the linear regression models, focus on explaining the following:

- **Coefficients**: Explain how each coefficient represents the expected change in the response variable (number of bikes rented) for a one-unit change in the predictor variable, holding other variables constant.

- **Significance**: Discuss the statistical significance of each variable and its practical implications for CycleHub.

Regression Tree Model

For the regression tree, describe the structure and decision-making process of the model:

- **Splits**: Discuss how each split is made based on a variable and threshold that best reduces prediction error.

- **Terminal Nodes**: Explain the interpretation of terminal nodes as predictions for the response variable.

Provide a clear explanation for CycleHub on how the model makes predictions, ensuring they understand the impact of each predictor and how different conditions (like temperature, humidity, or holiday status) influence the predicted demand.

Visualization of Predictions

After generating predictions for the upcoming week, create a plot that displays the predicted number of bikes rented per day. Make sure to:

- Clearly label the x-axis as the days of the week and the y-axis as the predicted number of bikes rented.

- Use different line types or colors to distinguish between predictions from the simple linear regression, multiple linear regression, and regression tree models.

- Add a legend to identify each model's predictions.