Practical Exam - Electric Moped Reviews

Instructions

- Use Python or R to perform the tasks required
- Write your solutions in the workspace provided from your certification page
- Include all of the visualizations you create to complete the tasks
- Visualizations must be visible in the published version of the workspace. Links to external visualizations will not be accepted.
- You do not need to include code unless the question says you must

Background

EMO is a manufacturer of electric motorcycles.

EMO launched its first electric motorcycle in India in 2019.

The product team has been asking website users to rate the motorcycles.

Ratings from owners help the product team to improve the quality of the motorcycles.

Ratings from non-owners help the product team add new features. They hope the new features will increase the number of new customers.

The product team wants to extend the survey. But, they want to be sure they can predict whether the ratings came from owners or non-owners.

Data

The dataset contains rating information about mopeds collected by the product team.

The dataset can be downloaded from here.

Column Name	Criteria
owned	Nominal. Whether the reviewer owns the moped (1) or not (0). Missing values should be removed.
make_model	Nominal. The make and model of the bike, one of six possible values (Nielah-Eyden, Nielah-Keetra, Lunna-Keetra, Hoang-Keetra, Lunna-Eyden, Hoang-Eyden). Replace missing values with “unknown”.
review_month	Nominal. The month the review was given in English short format (Jan, Feb, Mar, Apr etc.). Replace missing values with “unknown”
web_browser	Nominal. Web browser used by the user leaving the review, one of Chrome, IE, Firefox, Safari, Android, Opera Replace missing values with “unknown”.
reviewer_age	Discrete. Age of the user leaving the review. Integer values from 16. Replace missing values with the average age.
primary_use	Nominal. The main reason the user reports that they use the bike for. One of Commuting or Leisure Replace missing values with “unknown”.
value_for_money	Discrete. Rating given by the user on value for money of the bike. Rating from 1 to 10. Replace missing values with 0.
overall_rating	Continuous. Total rating score after combining multiple rating scores. Continuous values from 0 to 25 are possible. Replace missing values with the average rating.

Tasks

Submit your answers directly in the workspace provided.

1. For every column in the data:
    a. State whether the values match the description given in the table above.
    b. State the number of missing values in the column
    c. Describe what you did to make values match the description if they did not match.
2. Create a visualization that shows how many reviews were from owners and how many were not owners. Use the visualization to:
    a. State which category of the variable owned has the most number of observations
    b. Explain whether the observations are balanced across categories of the variable owned
3. Describe the distribution of the overall rating across the possible values. Your answer must include a visualization that shows the distribution.
4. Describe the relationship between ownership and overall rating. Your answer must include a visualization to demonstrate the relationship.
5. The business wants to predict whether a review came from an owner or not using the data provided. State the type of machine learning problem that this is (regression/classification/clustering).
6. Fit a baseline model to predict whether a review came from an owner or not using the data provided. You must include your code.
7. Fit a comparison model to predict whether a review came from an owner or not using the data provided. You must include your code.
8. Explain why you chose the two models used in parts 6 and 7.
9. Compare the performance of the two models used in parts 6 and 7, using any method suitable for the type of model. You must include your code.
10. Explain which model performs better and why.