Car Popularity Prediction

Introduction

A car company has the data for all the cars that are present in the market. They are planning to introduce some new ones of their own, but first, they want to find out what would be the popularity of the new cars in the market based on each car's attributes.

We will provide you a dataset of cars along with the attributes of each car along with its popularity. Your task is to train a model that can predict the popularity of new cars based on the given attributes.

Dataset

You are given a training dataset, `train.csv`. The file is a comma separated file with useful information for this task:

- `train.csv` contains the information about a car along with its popularity level. Each row provides information on each car. Information such as buying_price, maintenance_cost, number_of_doors, number_of_seats, etc. The definition of each attribute is as follows:

    1. `buying_price`: The buying_price denotes the buying price of the car, and it ranges from [1...4], where buying_price equal to 1 represents the lowest price while buying_price equal to 4 represents the highest price.

    2. `maintenance_cost`: The maintenance_cost denotes the maintenance cost of the car, and it ranges from [1...4], where maintenance_cost equal to 1 represents the lowest cost while maintenance_cost equal to 4 represents the highest cost.

    3. `number_of_doors`: The number_of_doors denotes the number of doors in the car, and it ranges from [2...5], where each value of number_of_doors represents the number of doors in the car.

    4. `number_of_seats`: The number_of_seats denotes the number of seats in the car, and it consists of [2, 4, 5], where each value of number_of_seats represents the number of seats in the car.

    5. `luggage_boot_size`: The luggage_boot_size denotes the luggage boot size, and it ranges from [1...3], where luggage_boot_size equal to 1 represents smallest luggage boot size while luggage_boot_size equal to 3 represents largest luggage boot size.

    6. `safety_rating`: The safety_rating denotes the safety rating of the car, and it ranges from [1...3], where safety_rating equal to 1 represents low safety while safety_rating equal to 3 represents high safety.

    7. `popularity`: The popularity denotes the popularity of the car, and it ranges from [1...4], where popularity equal to 1 represents an unacceptable car, popularity equal to 2 represents an acceptable car, popularity equal to 3 represents a good car, and popularity equal to 4 represents the best car.

We also provide a test set of 100 car along with the above attributes excluding popularity, in `test.csv`. The goal is to predict the popularity of the car based on its attributes.

You can download the zip (MD5 checksum: `35c4588462cdf8f6a455d45a44284b96`) containing the training and test files.

Submission Details

You are required to upload the following three files:

- The output file *prediction.csv* (maximum allowed size is `1MB`) containing the predicted values of the *popularity* attribute for each of the cars given in *test.csv*.
    - The output file will be evaluated against our hidden data and the grader will return a score.

    - If the uploaded file does not contain the same number of rows as the *test.csv* file, the grader will reject it.
    - Each line of the uploaded file must be either of 1, 2, 3 or 4 denoting your model's prediction for that line in the test dataset. If any line contains anything other than a 1, 2, 3, or 4 its validation will fail.

  A valid *prediction.csv* file has the following format:

```
1
3
4
2
2
```

- A *PDF* file (maximum allowed size is `2MB`) providing the findings and justification on the following topics:
    - Write a few lines about training dataset quality and any errors found in the training dataset.
    - Explain the data preprocessing steps.
    - Explain and justify the model you've chosen for the classification system.
- The source code of your approach for this task. Upload a *zip* file (maximum allowed size is `5MB`) with all relevant files to reproduce your results. The submitted file must have a `README` file with a detailed description about how to run the model to predict the popularity and generate the `prediction.csv`. Do not forget to include links to any external libraries or packages you use for the generation of your model.

  There is no limit on execution time, but the code should generate the output file: `prediction.csv`.

Evaluation

For each of the four class labels, $L = 1,\ 2,\ 3,\ 4$, We calculate the F_1 score separately. Let $T_{P,\,L}$ be the number of *true positives*, $F_{P,\,L}$ be the number of *false positives*, $T_{N,\,L}$ be the number of *true negatives* and $F_{N,\,L}$ be the number of *false negatives* for the class label L, then the precision P_L and recall R_L for the class label L are calculated as:

$$P_L = \frac{T_{P,\,L}}{T_{P,\,L} + F_{P,\,L}}$$

$$R_L = \frac{T_{P,\,L}}{T_{P,\,L} + F_{N,\,L}}$$

Let $F_{1,\,L}$ be the F_1 score for the class label L, then:

$$F_{1,\,L} = \frac{2 \times P_L \times R_L}{P_L + R_L}$$

The final score is calculated as:

$$S = 1000.0 \times \frac{F_{1,\,1} \times T_{P,\,1} + F_{1,\,2} \times T_{P,\,2} + F_{1,\,3} \times T_{P,\,3} + F_{1,\,4} \times T_{P,\,4}}{T_{P,\,1} + T_{P,\,2} + T_{P,\,3} + T_{P,\,4}}$$

Note that if the F_1 score for any class label is zero, then the final score will be zero.

Ranking

Prior to the end of the contest, all evaluations will be performed on a randomly selected 50% data points in the *test.csv* file. At the end of the contest, your *last* uploaded file (i.e., the most recently uploaded file) will be used to calculate your final score and position on the leaderboard. Because of this, make sure that your final (very last) submission is the output file with the maximum score.

File Upload