Code.Hub

WE LEAD

Predicting the City-Cycle Fuel Consumption in Miles per Gallon of a Car

A Classification Problem

December 2024

WE LEAD

Code.Hub

Data Science & Business Intelligence Bootcamp

Predicting the City-Cycle Fuel Consumption in Miles per Gallon of a Car

A Classification Problem

Group Project

December 2024

Contents

Introduction....1

Important Note....2

Project Scope and Deliverables....2

Overview of Project Work....2

Data Description....3

Detailed Objectives....3

- Exploratory Data Analysis....3
- Preprocessing....3
- Modelling - Predictions....3
- Your Code....4
- Project submission....4

Introduction

This group project is designed to help students apply the knowledge and skills gained in class to a real-world business intelligence system. In this scenario, you are a data scientist working for a consulting firm specializing in car rentals. Your company is particularly focused on **city-cycle fuel consumption measured in miles per gallon (mpg)**. Your task is to develop a Proof of Concept (POC) for a service that predicts mpg based on eight different attributes. This service will be marketed to car rental companies to assist them in making informed decisions when updating their fleets.

You are provided with the Auto MPG Dataset having the following characteristics

Data Set Characteristics:	Multivariate	**Number of Instances:**	398
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8
Associated Tasks:	Regression	**Missing Values?**	Yes

Your project focusses exclusively on creating the model for predicting the mpg value. The steps you should follow regarding the data flow are up to you. Your data needs to be well documented and organized so that it can be used in production.

Important Note

The dataset initially approaches the problem as a regression task. However, your task now is to reframe it as a classification problem. To achieve this, divide the data into categories based on percentiles (e.g., low, medium, and high fuel efficiency).

Project Scope and Deliverables

Several subtasks can be spawned from the objective of predicting the mpg value. Such tasks are:

- Explore the given data. See what they describe and gather valuable insights about their properties.
- Preprocess the data so that they can be used for predicting the mpg value.
- Build, train, evaluate a model using the scikit-learn library.

Your project deliverables which will support the implementation of these objectives are identified as deliverables D01-D03 in the following sections. You will collect all deliverables and submit them as your project portfolio work.

WE LEAD Code.Hub

Overview of Project Work

For running this project, you are advised to frequently meet as a team and discuss and agree on your implementation plan and actions. This means that you must end up with a clear understanding of

- the roles and responsibilities of the team members
- the project requirements
- the data requirements
- the way you will run your project
- the tools you will use for the technical work
- the tools you will need for the running of your team
- the deliverables of your work

Data Description

The dataset is provided in a file called **mpg.data.csv** and contains the following information

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Note: The dataset is provided on a Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license, so it is free to use in this project.

Detailed Objectives

Exploratory Data Analysis

Data exploration always helps you to better understand the data. Do not forget graphs helps a lot to gain insights from your data

D01: a python notebook containing any Exploration Data Analysis (EDA) you performed

Preprocessing

In this step you must bring the dataset in a format understandable by most machine learning algorithms. Some steps you might want to consider:

- Handling missing values in the dataset.
- Encoding categorical features.
- Scaling the features.
- Cleaning erroneous values.
- Handling outliers.
- Feature selection/extraction.

Note: Not all these steps are mandatory. You should do what you think better suits your needs.

D02: a notebook showing the preprocessing steps as you applied them

Model – Evaluate – Predict – Fine Tune

Here you must build a model that accurately predicts the mpg category. Try different learners, and parameters. Handle your data appropriately (e.g., split, cross validate, work with hyperparameters, etc) Use the appropriate metrics for evaluating your results any way you see fit!

D03: several notebooks containing your different approaches for training, evaluating and predicting phases of your work. Do not forget to include comments on the evaluation metrics you used.

Your Code

A well-written, documented, and organized code should be submitted.

- Remove all nonessential code.
- Refactor your code into functions.
- Write documentation and type hints for each function.
- Write a readme file explaining how the code is organized, its dependencies and how it should be run.

WE LEAD

Code.Hub

Project submission

Final Submission and Presentation: 24 February

- You must push your work on the main branch of your team's Github repository. The material should be the one you will present during your viva.
- You must also store your work in the Files area of your team on MS Teams.
- Present your work using any tool you feel comfortable with. Details will be discussed during the contact sessions. The presentation file(s) should also be pushed on Github and stored on the Files area of your team on MS Teams