REGENERATION

ΠΑΠΑΣΤΡΑΤΟΣ
ΕΤΑΙΡΙΑ ΤΗΣ PHILIP MORRIS INTERNATIONAL

Code.Hub

ReGeneration Academy on Data Science
(powered by Papastratos)

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

Group Project

September/October 2021

Overview

This document describes the scope of the project that will be undertaken by the different teams. The project is about building a model that predicts the **city-cycle fuel consumption in miles per gallon** of a car. Your team is asked to explore the given data, process them as you see fit, present them (before and after preprocessing) and build a machine learning model using scikit-learn python library.

Contents

Introduction

This group project aims to encouraging students to apply the knowledge and experience learned in the class towards a real-life business intelligence system. You are employed as a data scientist in a company that is involved in car rentals company consulting and it cares about city-cycle fuel consumption in mpg You are tasked with building a Proof of Concept (POC) of a service that will predict the mpg value given 8 different attributes. This service will be marketed by your company to various car rental companies to be used when they want to renew their fleet.

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.

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.
- Built, 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-D05 in the following sections. You will collect all deliverables and submit them as your project portfolio work. Interim submissions should be done during the development process (see below).

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 by Airbnb 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

Visualization (before preprocessing)

Present the initial data (before any processing) using Power-BI.

D01: a Power BI file containing your data visualization before preprocessing **(interim submission is due to 03/10/2021).**

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

D02: a python notebook containing any Exploration Data Analysis (EDA) you performed **(interim submission is due to 10/10/2021).**

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.

D03: a notebook showing the preprocessing steps as you applied them **(interim submission is due to 10/10/2021).**

Visualization (after preprocessing)

Present your data after preprocessing using Power-BI.

D04: a Power BI file containing your data visualization after preprocessing **(interim submission is due to 12/10/2021).**

Modelling - Predictions

Here you must build a model that accurately predicts the price of a given listing. Use the appropriate metrics for evaluating your results any way you see fit!

D05: a notebook containing your model, showing the training phase and the evaluation of the model **(there is no interim submission for this part).**

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.

D06: production-level code

Project submission

Final Submission and Presentation: 22/10/2021

- You must push your work (interims and final submission) 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 (interims and final submission) 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. This file should also be pushed on Github and stored on the Files area of your team on MS Teams