Predicting Car Mileage and Make using Linear and Logistic Regression

Jeffrey Roeder, Jeppe Hallgren, and Kasper Møller

April 1, 2024

Abstract

Is it possible to predict the outcome of a car's miles per gallon based on various factors such as the number of cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name? Is it possible to predict a car's make based on the same factors? This paper explores these questions using linear regression and logistic regression models.

1 Introduction

2 Research Question

3 Methods

We decided to follow **goldstein2017deconstructing** and follow the six steps of Data Science Deconstucted. We also needed to consider which data wrangling to perform **endel2015data**, **langer2023python**.

During our analysis of the data set, we discovered several outliers, which led us to investigate and identify additional errors in the data. Among other things, we found four cars where the horsepower values were missing, and we searched for relevant information to fill in these missing values. Additionally, we encountered typos in the 'car name' variable, which we decided to split into 'make' and 'model'. To clean the dataset, we ran a dynamic process where we checked and fixed typos in all 'make' values. We also reviewed the dataset to identify and correct any null values, empty strings, and special characters that had crept into the dataset. These steps were essential to ensure that our data was of high quality and reliable to train and evaluate our models. Another important step was to input missing data for the 'horsepower' variable. Since it was only 6 entires that were missing, we decided to manually search for the missing values. We found the missing values by searching for the car model and year on the internet. The dataset also included different naming for the same car make, such as Chevy and Chevrolet, VW and Volkswagen, Mercedes and Mercedes Benz. We decided to merge these into one category to avoid having the same car make represented by different names.

4 Analysis

5 Findings

6 Conclusion