MCS 7103: Machine Learning

Assignment 1: Exploratory Data Analysis Process

Name: Okabo Jasper

Registration No: 2024/HD05/21943U

Student No:2400721943

Introduction

A number of people have experienced difficulties at a time when they have to look up for a new car to buy. But then the journey begins with a lot of frauds, negotiating deals, researching the local areas both off and online. This hassle usually results into exploitation of both the buyers and the car manufacturers by the middle men who, in most cases determines the prices of cars basing on their best interest.

To tackle this challenge, the machine learning model trained using the car prices data will definitely reduce exploitation of prospective buyers/new car owners by the middlemen by predicting the car prices basing on the car's year of manufacture, model and the gasoline usage.

The Linear Regression Machine Learning model was used to compute the linear relationship between dependant variable and one or more independent features by fitting a linear equation to observed data

Procedures

This project was done in google Collaboratory platform using step by step approach up to the final point of performing prediction.

i. Import the necessary python Libraries:

These are the preprogramed modules that does different functions;

- Pandas: Used for loading the data frame
- Matplotlib: This was used for visualizing features of the data.
- Seaborn: It helps in determining the correlation between features using heatmap.

Data Pre-processing

This involved categorizing the features depending on their datatype (int, float, object) and then calculate the number of them.

Data Wrangling

This involved things such as gathering, collecting, and transforming the raw data into another format for better understanding, decision-making and analysis in less time.

Exploratory Data Analysis (EDA)

A deep analysis was done so as to discover different patterns and spot anomalies on the data before making inferences. All the variables were examined and an heatmap was made using the seaborn library that was earlier imported.

Data Cleaning

Having understood the structure of the dataset, it was improved by removing the incorrect and irrelevant data. As in the dataset, there were some columns that were not important and irrelevant for the model training. So, those columns dropped before the model could be trained.

There were 2 approaches to dealing with empty/null values; 1. By deleting the columns/rows that were not important. 2. Filling the empty slots with mean/mode/0/NA/etc.

One Hot Encoding

This was meant to convert categorical data into binary vectors to map the values to integer values and by using OneHotEncoder, I converted object data into integers.

Splitting Dataset into Training and Testing

X and Y splitting (i.e. Y is the Car Price column and the rest of the other columns are X)

Model and Accuracy

To train the model to determine the continuous values, I used the linear regression model to predict the final output-dependent value based on the given independent features.

Like, here we have to predict car prices depending on features like car model, Year of Manufacture and fuel type.