1. Introduction

In this project, we aim to get familiar with simple machine learning methods to estimate the price of a car. This project is defined in three phases. In the first phase, we will review and analyze the data. In the next phase, we will get familiar with preprocessing methods, then we'll extract features from the text columns. Finally, we'll build a simple linear regression model from scratch to predict prices.

2. Dataset Description

This dataset contains information on various cars including technical specifications, retail data, and price. The data provides a range of vehicle attributes that can be used to build machine learning models to predict the final price of a car based on its characteristics. The dataset offers an opportunity to practice core skills like data exploration, preprocessing, feature engineering, and modeling for a regression task.

3. Phase One - Explore Data

The first step in any machine learning project is to observe, explore and examine the data and the relationship between the features. For this purpose, do the following steps:

3.1. Check the overall structure of the data using the info() and describe() methods of the Pandas library.

3.2. Find and display the count and percentage of missing data for each column.

3.3. Plot the price column against each feature in subplots. Based on examining these plots, which feature do you think would be best for predicting the price using a linear regression model and why?

3.4. Plot the correlation diagram. Which features are more correlated with the price? Plot the count of unique values for each of the selected features.

3.4. Further analyze, explore, and visualize the relationships between features and the price column using scatter plots and hexbin charts.

4. Phase Two - Preprocessing

Data preprocessing is the most important step in any machine learning project. In this phase, the raw input data must be converted into a set of processable features. For this purpose, do the following steps:

4.1. Which feature has the most missing data?

4.2. There are many methods to solve the problem of missing data. Explore these methods, then choose one to apply to and mention the reason for your choice. (Use fillna() method of the Pandas library)

4.3. Why is Standardization or Normalization being applied to numerical features? Which method do you choose? Why? (You can use sklearn.preprocessing.MinMaxScaler())

4.4. There are many ways to make our model work with categorical features. Explain two of them and state which method you used. (You can use sk.preprocessing.LabelEncoder())

4.5. Sometimes it is necessary to remove unnecessary columns from a dataframe during data preprocessing. Based on exploring the data, explain whether we should drop any columns. What is the rationale for dropping or keeping each column?

4.6. In machine learning, it is essential to split the available data into separate training and test sets in order to properly develop and evaluate models. What would be an appropriate ratio for this train-test split?

5. Phase Three - Model Training and Evaluation

5.1. A Jupyter notebook file has been provided. Complete the specified sections in the notebook to preprocess the data and implement a linear regression model to predict car prices.

5.2. After selecting relevant features from the training data, generate predictions on the test set using the linear regression model. Choose an appropriate evaluation metric to quantify the accuracy and effectiveness of the model's predicted values compared to the actual test labels. Read about different evaluation methods for regression models and explain two appropriate options for this task in your report.

5.3. Implement the two evaluation methods explained in section 5.2 in code cells in the notebook. Use them to evaluate the linear regression model's predictions on the test set. Compare the results of the two metrics. Discuss which one provides a better assessment of the model's performance for predicting car prices and why.

Notes:

- It's recommended to use Jupyter Notebooks for these exercises, mostly because it displays DataFrames clearly.
- Feel free to ask any questions on Telegram/Skype!

- Please upload your responses in this Google form.

Good Luck!