CSEN911: Data Mining
Winter 2022
Author: Hazem Nabil
Course Instructor: Dr. Ayman Alserafi
TAs: Ghada Mansour, Tameem Alghazaly & Nadeen Osama

German University in Cairo

Linear Regression Mini-Project

Due: 6 November 2022, 11:59PM

Introduction & Motivation

According to Statista, the used car market continues to outsell the new car market by more than 2:1. Around 41 million used cars were sold in the US in 2019 compared to the 17 million new vehicles sold in the same year. With about 280 million vehicles in the US with an average age of 11.9 years, and in some European, Asian and African countries this number goes up to 17. According to Grand View Research, the global used car market is estimated to be worth USD 1,332.2 billion in 2019, as such even car dealerships have begun treating used car sales as a primary activity rather than a side by-product of new car sales. Consequently, it is of interest to used car sellers to know how the car's make, type, condition and other characteristics influences the value of the car on the market.

Mini-Project Description & Requirements

You will be provided a dataset of used car auction sales based on real sales made in the US in 2015. You are required to apply multiple linear regression model to predict the car price based on several of its characteristics and comment on your model's performance for this use case. In addition, you are required to answer the following queries and provide an appropriate visualization/plot for each:

1. What are the most popular car brands? (mention at least 3)
2. Draw a bar plot showing the selling price of each of the transmission categories. Do automatic cars, on average (mean), sell for a higher price than manual cars? (Use the following link for reference: `https://seaborn.pydata.org/generated/seaborn.barplot.html`)
3. Draw a box plot showing the distribution of the selling prices of each car make. Which car makes have the widest distribution of selling price? Which are the most expensive? Name at least two for each.
4. How correlated is the odometer reading or mileage with the car's sale price? (Hint: plot a line of best fit.)
5. Likewise, how correlated is the car's condition with its sale price? (Hint: plot a line of best fit.)
6. Draw a bar plot showing the selling price of each of the body categories. Are there differences in the average (mean of) car price due to the car's body type? If so, rank their order and state the most expensive and least expensive car body type.

Data Challenges

You are expected to perform several necessary data pre-processing and cleaning steps. Namely:

- Data cleaning for numerical columns loaded in as object (wrong type)
- Resolving inconsistencies in categorical columns
- Reduce the 'body' categories into a few main types, namely: Sedan, Coupe, Convertible, Hatchback, SUV, Minivan, Wagon, and Pickup Truck. Note: pickup trucks are indicated by their different cab types in the 'body' column.
- Selecting suitable columns for the modelling step
- Applying an appropriate categorical encoding method for the modelling step

Info: You will be given a starter notebook with an example query and its solution. Write your text answers in markdown cells below the code as shown in the example.

Evaluation Criteria

The maximum grade for this submission is 22 and your Jupyter notebook submission will be graded on the following:

- applying appropriate data pre-processing and cleaning methods (data cleaning, column selection, data transformation, etc.) [8]
- answering each query correctly and displaying an appropriate visualization for each query [6]
- training the chosen model to the dataset [4]
- evaluating the performance of the model and commenting on its performance (an appropriate performance metric should be used and justified) [4]

Notice: If you make changes to the dataset, you are required to display the effect of your changes using an appropriate method. You are also required to comment your code, briefly explaining each step you are doing (necessary in data preprocessing, cleaning and feature engineering) You are not allowed to loop over the dataset rows/cells (unnecessary and inefficient); use built-in functions instead.

Submission

You may work on the mini-project individually or in pairs **(cross-tutorial pairs are NOT allowed)**. To get started, register using the following GitHub classroom link `https://classroom.github.com/a/-HwOmd_r` as a team of two or as an individual. You are required to use your name(s) + tutorial number as the team name and the format should be similar to the following example: "Mohamed Ahmed T01 - Ahmed Mohamed T01". Substitute the names/tutorial with your own. A private repository will then be given to you to work on your mini-project assignment along with additional instructions regarding the submission.

Once the team is created, your partner (if in a group of two) can use the link above to join the created team. While only one member is required to join per group, joining allows both members to access and collaborate on the assignment. However, unless you understand git workflow, work together on one notebook instance instead of working in parallel to avoid merge conflicts and breaking changes.

Warning: Your Jupyter notebook submission will be re-run after submission. Submitting a notebook that errors mid-way through execution will result in losing the grade of all subsequent cells! The order of the notebook's code cells and the order in which you execute them matters. Please make sure your notebook runs and outputs the expected results by restarting the kernel and running all the cells before finalizing your submission.

Bonus

Since linear regression is sensitive to outliers and their presence can severely degrade the model's performance, remove the outliers and train a new linear regression model with the outliers removed. Evaluate and compare the new model's performance to your previous model that was trained with outliers present in its training data.

Note: Code related to this bonus task should be performed at the end of your mini-project notebook, not in-between the mini-project's requirements. Indicate the start of your bonus code, either with a comment at the start of the code cell, or add a new section/heading titled 'Bonus' for clarity (similar to the existing sections provided).