ISE 291 Term Project

Data Science on JML business data [1]

Due date: May 12, 2022.

A local car software company, JML inc., is looking for data scientists to help them understand the possible patterns that will affect the car resale prices. Currently the company plans to build an application (app) for the used cars that are sold/purchased across the country. The company is interested in estimating the sale price of cars for the app users. Such an app will help the company to attract investor for potential projects. The company has provided the relevant data that they have collected over the years. Following table presents an overview of the collected data:

Table 1: Data Description

Fields	Description
Manufacturer	Cars's manufacturer.
Type	Type of the car. The car could be sedan, crossover or SUV.
Cylinders	Total number of cylinders in the car's engine.
Model_Year	The year of the car was manufactured.
Model_Price	Sale Price of the car in 'Model-Year' in SAR.
Sold_Year	The year in which the car was sold.
Sale_Price	Sale Price of the car in 'Sold-Year' in SAR.
KM_Driven	Total KiloMeters the car is driven before the sale in km units.
KPL	Estimated kilometers per liter at the time of sale in kpl units.
Power	Four wheel drive (4WD), or two wheel drive (2WD).
Class	Standard or Premium features available on the car.
Location	City in which the car was sold, 'Riyadh', 'Dammam', 'Jeddah', 'Dhahran', 'Makkah', 'Madinah'.
Accident_History	Total number of major accidents reported on car's history.
Owner_History	Total number of owners of the car.
Color	The color of the car.
Seller_Type	Whether the seller is an individual or a company.
Service_History	It either 0 or 1, where 1 indicates the service history from the car's authorized dealer is available.

Aim. The aim of this project is to explore the data, and find possible patterns/relationships in the data. The key variable of interest to JML inc. is Sale_Price. Assume that the cars that depreciate by 10% or less are high return cars, and those that depreciate by 20% or more are low return cars. The depreciation is defined as:

$$\frac{[Model\_Price] - [Sale\_Price]}{([Sold\_Year] - [Model\_Year])[Model\_Price]} * 100$$

Data. The data related to the project is provided in three different files, named in the following format: Group_SS_XX_A, Group_SS_XX_B and Group_SS_XX_C files, where SS is your section number and XX is your group number. In addition to that, Table 1 presents the meta data related to the given data.

Expectations. At the end of this project, you are expected to provide JML with answers to the following questions. Support your answers with corresponding/appropriate data science methods and visualizations (wherever applicable).

[1] Case study for ISE 291 Term 212 students. Developed by: Dr. Mujahid Syed.

For the following task use Group_XX_A file:

Task-1: Prepare the data given in Group_SS_XX_A file, i.e., handle the missing values, remove outliers, and fix inconsistencies. You can pick any set of methods, but clearly justify your approach.

For the following task use Group_SS_XX_B file:

Task-2: Draw the pair-wise plots between all the input variables and the output variable (Sale_Price).

Task-3: Identify top and bottom three numerical variables that are strongly related to the output variable (Sale_Price)? Use the relevant analysis approach.

Task-4: Show if the input variables contain the information to separate low and high return cars? Use plots to justify.

Task-5: What are the common patterns for the low return cars? Use plots to justify.

Task-6: What are the common patterns for the high return cars? Use plots to justify.

For the following task use Group_SS_XX_B and Group_SS_XX_C files:

Task-7: From the input and output columns given in Group_SS_XX_B file; identify how the input variables together are related to the output. Assume that all the input variables are relevant to output variable (Sale_Price).

Task-8: It was observed that some of the input columns are correlated, and this may make the above analysis unreliable. Redo Task-(7), with the consideration of correlation issue between input variables.

Task-9: It was observed that some of the input columns may not be relevant to the output variable, and this may make the analysis unreliable. Redo Task-(7), with the consideration of possible unrelated input variables.

Task-10: Predict the estimated Sale_Price values given in Group_SS_XX_C file. Consider all the numerical and categorical variables for the analysis. If you skip any column, then provide strong justification. Also, justify your transformation and modification of the columns for the analysis.