Assignment: 12.1 – Term Project

Tae-Hyun Sakong

DSC 530-T303-2243-1

03/02/2024

**Summary**

**Research Question:**

What are good predictors of used-car sales? Specifically, how does the vehicle's year, KM\_Driven, mileage, engine size, or max power predict whether or not a vehicle is sold?

**Aim:**

The aim of this term project was to identify good predictors of used-car sales in India, focusing specifically on how the vehicle's year, kilometers driven, mileage, engine size, and maximum power influence the likelihood of a car being sold.

**Outcome of the Exploratory Data Analysis:**

The exploratory data analysis (EDA) revealed several key insights into the used-car dataset. Histograms for year, km driven, mileage, engine, and max power showcased the distribution of these variables. The year of the vehicles tended to be more recent, with a peak around 2015 to 2017, suggesting a market preference for newer used cars. KM driven and max power showed a right-skewed distribution, indicating that most cars have lower kilometers driven and power, with a few exceptions having very high values. The analysis of outliers using the Interquartile Range (IQR) method led to the capping of extreme values to reduce their impact while retaining all data points. This step was essential for preparing the data for further analysis, such as correlation tests and modeling.

**What Was Missed During the Analysis:**

While the analysis covered several critical aspects, there were potential areas that might have been missed. For instance, the impact of categorical variables such as fuel type, seller type, transmission, and ownership status on the likelihood of a car being sold was not explicitly analyzed. These variables could provide additional insights into buyer preferences and how they influence sales outcomes. Moreover, interaction effects between variables, such as the combined impact of year and mileage or engine size and max power on sales, were not explored. These interactions could reveal more complex relationships that single variables alone cannot capture.

**Additional Variables for Analysis:**

Incorporating variables related to the car's condition, such as maintenance history, accident history, or any modifications made to the vehicle, could have enriched the analysis. Additionally, macroeconomic factors, such as economic growth rates, fuel prices, or policy changes affecting the automotive sector, might also influence used-car sales but were not considered in this analysis.

**Assumptions Made During the Analysis:**

One assumption made during the analysis was that the dataset accurately represents the broader used-car market in India. This assumption could be incorrect if the dataset is biased or not comprehensive. Another assumption was that the linear relationships explored through correlation tests and linear modeling adequately capture the relationships between variables and sales outcomes. In reality, these relationships might be non-linear or influenced by unobserved factors.

**Challenges Faced and Areas of Confusion**

One of the significant challenges faced during the analysis was dealing with class imbalance in the dataset, particularly in predicting whether a car is sold. The logistic regression model tended to predict the majority class (not sold), highlighting the difficulty of modeling imbalanced datasets. Additionally, determining the best way to handle outliers and deciding whether to remove or cap them required judgment calls that could significantly impact the analysis outcomes. Lastly, interpreting the results of the logistic regression model, especially understanding the practical significance of the model's coefficients and their implications for predicting sales, was complex and required careful consideration.