Case Study pt.1

Martin S. Rivera Ritchie

29 April 2021

Coursera Capstone

Peer Graded Assignment

1. **A description of the problem and a discussion of the background.**

Used cars sales comprise an important chunk of sales every year. According to Statista in 2019 an estimated 40.8% of sales made in the US were used cars, in comparison to the 17% comprised by new car sales (graph 1).

Graphical user interface

Description automatically generated with medium confidence

(Graph from Statista depicting number of used car sales vs new car sales)

Having this in mind, the business opportunity that I will investigate in the following part is how should a car dealership balance their inventory between new and used cars in order to improve their sales of used cars based on year, make, model, millage, price and state, and condition new vs used.

1. **A description of the data and how it will be used to solve the problem.**

The data that I will evaluate is one that I acquired from *Kaggle* called “*US Cas Dataset”* which in turn was scrapped from *auctionexport.com*. This dataset contains 2,498 entries of which their classification is as follows:

**a) Variable Descriptions**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Type** | **Description** |
| **Price** | Float | Price at which vehicle was sold. |
| **Brand** | String | Make of the vehicle sold. |
| **Model** | String | Model of the vehicle sold. |
| **Year** | Float | Year of model pertaining to the vehicle. |
| **Title\_status** | String | Whether, or not, the vehicle is identified as salvaged by the insurance; binary. |
| **Millage** | Float | Millage as read in the odometer of the vehicle. |
| **Color** | String | Paint of the vehicle. |
| **VIN** | String | Vehicle Identification Number. |
| **Lot** | Integer | Lot pertaining to the manufactured batch that the vehicle belongs to. |
| **State** | String | State in which transaction took place. |
| **Country** | String | Country in which the transaction took place. |
| **Condition** | String | Time that it took to sell the vehicle at auction. |

**b) Cleaning the data**

The dataset as it was downloaded is in a .csv file, let’s look at the first five rows to get an idea of what it looks like in its raw form:

Table

Description automatically generated with low confidence

(Image demonstrating a portion of the dataset to be used from Kaggle)

Looking at the dataset in its raw from there are a few steps that must be taken to clean the data before it can begin to be studied. These steps include eliminating some of the variables, standardizing the data in an existing column, and must importantly the addition of a column which will need to be randomized for the purposes of this exercise which is the characteristic “New vs. Used car”. The changes to the dataset are described in the following table:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Action** | **Description of Change** |
| **Model** | Standardize | The variable contains both numbers interpreted as integers and words interpreted as string values.  All values in the variable must be transformed into string values. |
| **Title\_status** | Eliminate | Variable is removed from the dataset. |
| **VIN** | Eliminate | Variable is removed from the dataset. |
| **Lot** | Eliminate | Variable is removed from the dataset. |
| **Condition** | Eliminate | Variable is removed from the dataset. |
| **Country** | Eliminate | Variable is removed from the dataset. |
| **New/Used** | Append | For the purpose of this exercise, where the range in the *Year* of the cars is from 1973 – 2020, all cars from 2019 – 2020 will be classified as *new*, which comprises 37% of the cars in the dataset. |

Now that the data has been uploaded to a Jupyter Notebook and the mentioned changes have been made we take a look at sample the dataset to confirm the changes.

Table

Description automatically generated

With the check complete, the dataset is ready for exploration, which continues on the following week of the capstone course.

**Sources**

Graph of used vs new car sales in the US used.

https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/

Research of the market share of the used cars market in the US.

Mckinsey & Company, https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/used-cars-new-platforms-accelerating-sales-in-a-digitally-disrupted-market