Can I Sell My Car?*

Keiwun Turner†
Department of Physics and Astronomy, Texas Tech University

(Dated: May 10, 2025)

During the wake of the COVID Pandemic, car sell prices have risen tremendously. While this is not an uncommon phenomenon, car prices fluctuate based on many things, the raising prices have also affected the used car market. In this paper, I plan on using data to predict the sale price of a used car based on key factors such as model, mileage, gas type and others.

I. INTRODUCTION

In the age of information and data, understanding how to manipulate and analyze data becomes paramount. In the area of Statistics and Probability, there are many ways to understand data and numbers, however it is now incumbent upon modern researchers to figure out how to do this with large amounts of data in record time. A large part of High Energy Physics is to sift through the scattering events at CERN and determine which events to throw away and which to investigate further. As a concrete example of the scale of data is being produced, CERN data centers would process about a petabyte worth of data per day$citeCERNhere$. To this end, techniques like machine learning have become a necessity for the field. This is not to say that HEP is the only area that will benefit from technique advancement or that HEP is the driving force for innovation in this area of mathematics and computer programming, but rather to serve as an example of modern science. In the area of Astronomy or Astrophysics where data on stars and exoplanets as well as blackholes are studied, much data is produced and evaluated as well. In the biological sciences, bioinformatics has become a large field, doing much of the same discussed earlier.

A. Data Structure

The data used for this project came from a website called Kaggle. The data set was titled "Vehicle Dataset". This data reports car sales and list relevant parameters about the sale namely, car make and model, the year, gas type, transmission type, mileage and whether the seller is an individual or a company, and the amount for which the car was sold. This data type cannot tell if a car will be sold, but rather how much the car sold for. As such, the author will look for strongly correlated parameters as a way to predict a selling price. To do this, a correlation matrix was constructed to see the strength of correlation and anti-correlation between different variables. Now seems like a good time to express the currency choice of the data set. The currency of choice in the data in called the "LAC" which stands for LaCucina. Currently, the LAC's conversion to USD is as follows 1 LAC is equivalent to 0.0185 USD. As the author did not produce the data, the author cannot give a guess as to why this currency was used for the unit of currency. Now, an astute reader could guess that there are strong correlations between certain parameters, for example, as mileage goes up it is expected that the car would sell at a lower price, or if the car was of a brand that is said to always breakdown, the selling price would also go down, but if the car was made in a recent year, then one would expect for the price to be higher. A simple correlation matrix can be seen below.

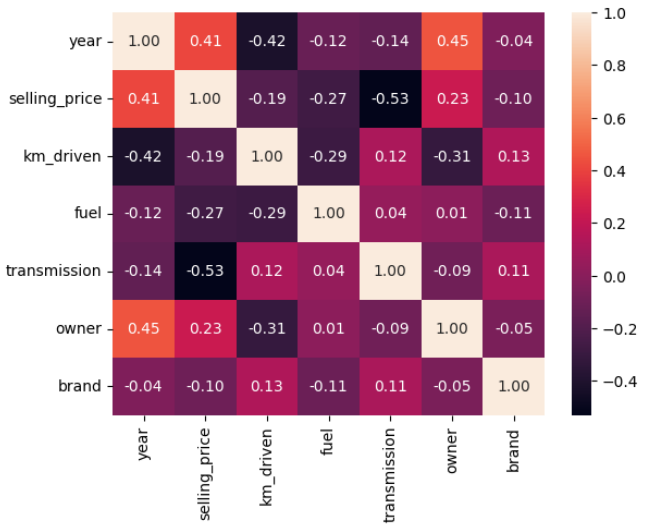

FIG. 1. Correlation map of selling parameters

Correlation matrices are symmetric matrices with 1's on the diagonal to show that a variable always perfectly correlates with itself. These matrices, in a small number of variables can seem tedious, however for larger number of variables, it can be a way to quickly see redundancy in data and to see strong, weak or anti-correlation. One could also use these matrices as ways to check for the validity of other methods. If there are a large number of correlations, then linear regression methods may not yield reliable results. However, the author has not done this, instead the author's intention was to perform a san-

* A footnote to the article title
† Also at Physics Department, Texas Tech University.

ity check. As stated above, one could expect correlations, and this matrix shows just that.

Now, seeing the correlations can give insight, but looking at other graphs and representations of the data can give even more insight. For example, it is strange to correlate a brand with a fuel. So one may be lead to ask if there are other methods by which the author can show correlations that will be useful in predicting the selling price of the vehicle, and the author gives a resounding yes! Take for example the amount of cars sold of a single brand, if a car belongs to a company that is known to make reliable cars then it stands to reason, that particular car can sell for a higher price. This representation can be seen below.

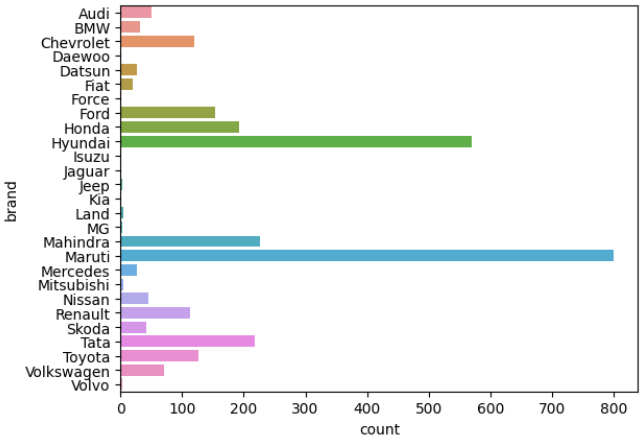

FIG. 2. Graph showing the relative selling volume of brand vs number of sells

It is apparent that there are favored companies and that there are companies that do not perform well at all and this gives valuable insight. Maruti is the highest selling car by volume, whereas something like Kia is hardly ever bought and this information is vital to know.

Another graph that can show information is the price at which the car sold vs the year of the car, this relationship can be seen below.

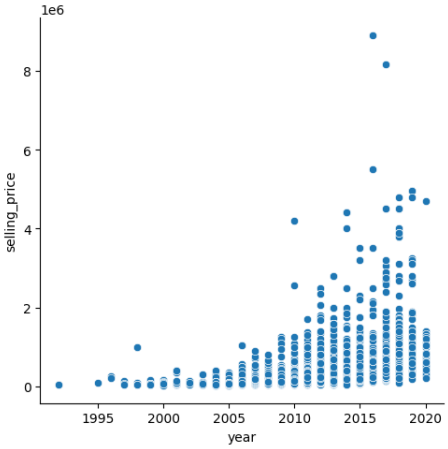

FIG. 3. Plot of price of car vs the year of the car

II. MACHINE LEARNING

In this project the use of Machine Learning is necessary, as a way to predict the price of a car being sold on the market. To do this unsupervised and supervised methods were used, namely, KMeans Clustering and Random Forrest Classifier. The choice for unsupervised was to allow the machine to recognize patterns on its own, without human intervention, while the Random Forrest Classifier was used to predict a value in the future.

A. KMeans Clustering

The word clustering in the title gives a good idea as to what this method does. The goal is for the code to group data points into clusters based on similarities seen in the data, this is what is meant by seeing trends. This is done by creating a cluster center, which is just the median or mean of the data points. The algorithm works iteratively and places data points into clusters. It is a efficient and simple method, which is why it is a very common clustering method. In this project, 5 clusters were made, starting with the random state being 42, so as to provide reproducibility.

The Random Forrest Classifier is a meta estimator, meaning it used other estimators as an input. This method was used as a way to improve the estimation accuracy as many different variables were used. The data was split into test and training data.

Next was to determine the importance of each parameter. Not every metric is as important as another, to that end the relative importance, as determined form the algorithm, goes as:

For the purposes of this project, having the cutoff being at 3 seems reasonable. This means any parameter that is at 3 or above will be classified as an important variable,

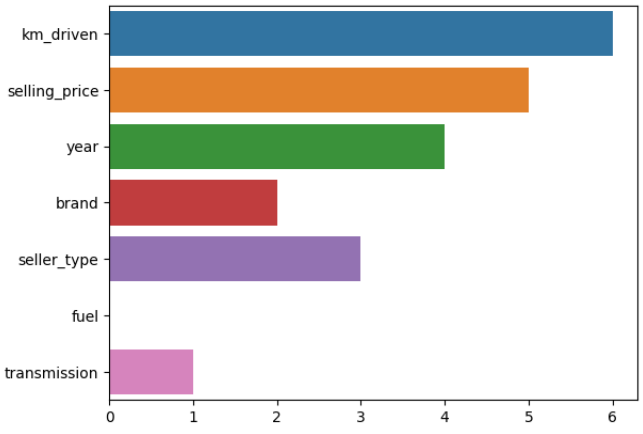

FIG. 4. Relative importance of each varible

while the others are deemed as unimportant. To that end, it can be seen that variables like fuel type plays no role in the decision, in contrast, the km driven parameter seems to be the parameter with the most sway, even more so than the selling price.

Armed with this knowledge, and all the pieces an estimator can be constructed. The test will be a 2014 Hyundai with an automatic transmission, gas, sold from an individual for 650000 with a km of 100000. The algorithm gives an approximation of the chance of selling the car and for a recommended price, which can be seen below.

```
Your car has a 0.44% chance of being sold!!
We recommend You to reduce the price of Your Car
The Average price of cars likes yours is 627647
```

FIG. 5. Prediction given by the algorithm

III. CONCLUSION

The author set out in this project to make an algorithm to predict the selling price of a car in the European market based on a few variables. By using methods like machine learning and correlation maps to express relationships between the variables. In the end it was shown that the algorithm was able to give a percent chance of a vehicle with particular parameters to be sold. The algorithm also was also able to give an estimate of how much the car's selling price should be to optimize the chance that the car gets bought. In future works the author would like to add factors such as brand sentiment, the amount of crashes a car has been in and possibly the geographical location of the sell. For the brand sentiment, the general public considers Toyota to make reliable cars and therefore will have a good brand sentiment, however the general public does not hold such a high opinion of say Ford. If a car has been in an abnormally high number of accidents, then there is a higher chance that car will have maintenance issues. As far as location goes, if one lives in an area that experiences high amounts of snow yearly, then vehicles with all wheel drive(AWD) might out perform others. Though these are parameters that the average buyer thinks of, the amount of variables will make it hard to collect adequate data and then the algorithm needed to give accurate results may take a while to compute.

Thanks is given to the Kaggle website for being a free source of many different types of data.