Market Research using Machine Learning Techniques for a New Car Company Entering The Space

Author: Gregory Hovesen

ID: 500769246
Section: D10
Supervisor: Dr. Kachana Padmanabhan
November 8th, 2021

Ryerson University

Table of Contents

Abstract

A new automotive company is deciding to enter the market with their new range of products. They are implementing data science tools to understand the current market. The goal of the analysis is to understand the leading factors that go into the price of a vehicle, which in turn, will help with the future price generation of their products to remain competitive, or exceed the current market. As you can assume, many factors go into the pricing of a vehicle. Parameters such as weight, capacity, engine type, and many other factors can be said to influence the final sales price of each unit. To accomplish this analysis, multiple linear regression will be used to create a model. This will give insight regarding each contributing factor. This analysis will be performed on previous market data, however, this can be replicated with a forever changing/expanding dataset to provide more accurate, to-date modeling. This analysis will be implemented through various softwares. These softwares include Scikit, Pandas, using the Python Language along with ggplot using the R Language, to name a few. These will be used for

data-cleaning of the datasets, exploratory analysis to understand the factors of the dataset, regressionional analysis, as well as use of and manipulation of classification algorithms. After the completion of the analysis, ggplot will be used to visualize the data in a more viewer friendly manner to present the analysis' findings. This in turn will help the company in making future decisions.

The main datasets used in this analysis can be found on Kaggle at:

Birla, N. (2020, October 24). *Vehicle dataset*. Kaggle. Retrieved September 27, 2021, from https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho?select=Car%2Bdetails%2Bv3.csv.

Pouyaaskari. (2021, January 15). *Automobile-dataset*. Kaggle. Retrieved September 27, 2021, from https://www.kaggle.com/pouyaaskari/automobile-dataset.

To conclude, this project will be able to determine a price prediction model for future products. What attributes have the strongest relationship with the dependent variable, price? And finally, will be able to determine whether or not the company's products are at a competitive rate compared to industry averages.

Literature Review

As a new car company trying to enter a very competitive market, loads of market research is required to have a successful entry. Relying on market data from various public datasets regarding vehicle attributes, we can determine the main factors that contribute to the pricing of a vehicle. In our datasets, attributes such as weight, dimensions, capacity and engine size, to name a few, are all contributing factors to our model. This literature review examines the data science methods required to follow through with creating a pricing model. Along with researching the methods being used, various articles that reveal more regarding the pricing of a vehicle will be examined to cross reference our results. As this form of market research has been completed before, this literature review will also highlight the similarities and differences between this project and the others.

According to D'Allegro (2021), mileage and condition are the key factors into pricing a used car. This information helps us determine whether or not information in our dataset coincides with real world instances of car pricing. As condition is a very subjective attribute, a more nominal representation of a vehicle's condition can be solely based on mileage as the vehicle's main asset is the engine. Higher mileage, lower selling price is the relation in question. Along with this, D'Allegro (2021) mentions that even though there are many tangible attributes regarding car pricing, some soft attributes may hold a higher weight regarding a minivan versus a sports car. For example, a manual transmission is more desirable in comparison to an automatic transmission for the sports car market than the minivan market. Finally, an attribute such as colour may have greater weight for one buyer compared to the other. This can be crucial to the cars price as more common car colours will sell faster than others which will come with a higher price tag than those that sit.

The way we are able to determine all of the factors that go into the pricing of a vehicle, many data science techniques will be used. Foremost, a multiple linear regression model will be created based off of the data. With that being said, this process has been carried out many times regarding a price prediction model in any industry. An example for this would be in the paper by Y. E. Cakra and B. Distiawan Trisedya (2015) regarding the prediction of stock prices using linear regression models. As mentioned by Y. E. Cakra and B. Distiawan Trisedya (2015), stock pricing models have to be continuously updated to create accurate predictions, which is also the case for car pricing. Although car pricing is not up to the second like stock pricing, everytime a vehicle is sold, the market value changes, requiring us to reconsider the current model. Another valid pull from this article is the fact that this paper is using linear regression based on sentiment analysis. Should our dataset have intangible values such as colour, we will have to consider them in our model as it was stated by Y. E. Cakra and B. Distiawan Trisedya (2015).

As stated in the article by Dubin, R.A. (1998), regression analysis was used to predict housing prices. The statistical technique used for this model was the ordinary least squares. This paper is particularly useful for this project as, like other regression models, many attributes such as house type, location, and number of rooms, to name a few were used; however, the model mentions it is missing a large quantity of information. This being the correlation between the prices of neighbouring houses. This is directly related to our dataset regarding missing market information. Although this article is from 1998, the process of model creation remains the same.

In the article "Support vector regression analysis for price prediction in a car leasing application.", Listiani, M. (2009) mentions the fact that many factors go into the pricing of a vehicle. This meaning there is a large amount of independent variables making up the dependent value, car price. As there are so many variables, the only way we can create an accurate model is through machine learning. Listiani, M. (2009), states that “a good model is required to have a good performance on future or previously unseen data”, which is exactly what we are trying to accomplish throughout this project. Listiani, M. (2009) proceeds to mention two key faults with regression models. Under-fitting, and Over-fitting. These two problems have the same effect on the model as they both render the model less accurate. When a model is overfit, the model is too heavily reliant on the training data and is poor at accurately predicting new values. On the other hand, under-fitting models is the result of the model being too simple. This does not allow for all attributes to be accounted for in the model.

In the article “OLD CAR PRICE PREDICTION WITH MACHINE LEARNING”, Gajera, P., Gondaliya, A., & Kavathiya (2021) compare five different models of car price prediction. The result of this paper is that using the Random Forrest Regression model was the most efficient model at 93.11% success rate. This is particulairly impressive as the comparison is against KNN-Regression, Decision-Tree Regression, XG Boost Regression and finally Linear Regression. As linear regression directly relates to my project, I look forward to creating a more accurate model than Gajera, P., Gondaliya, A., & Kavathiya (2021), with their success rate of 76.46%. As mentioned by D'Allegro (2021), mileage has a huge role in pricing a vehicle, and this was also the case for the article in question. As Gajera, P., Gondaliya, A., & Kavathiya (2021) used a dataset of over 92000 records, their model may be vary intricate, which may help them with their accuracy. On the other hand,

Finally, to understand the implementation of regression models, Sci-Kit Learn is the Python package in question. As mentioned in the article "Scikit-learn: Machine learning in Python", Sci-Kit is a Python package with many machine learning algorithms at its disposal and is directed at the non-specialist user for medium scale problems. Sci-kit takes pride in its interactive nature to make machine learning more accessible for the Python community. The backend processes of this package are Numpy and Scipy. As mentioned, the majority of the package is underwritten python, however there are some C++ libraries that are also used in its implementation. Sci-kit learn also has a website with all of its documentation and real examples to assist with your problems at hand. As this will be used in my analysis, this information is very pertinent and useful.

The datasets in question have similar independent variables such as transmission and engine type. These being categorical values along with other numerical attributes like mileage. We will use the required packages from Sci-Kit Learn, in conjunction with data cleaning techniques to improve our data. In conclusion, as this project has been implemented in the past, the aim of this project is to improve the accuracy of our model in comparison to the others. Along with other papers such as stock price prediction, we can use this new found information to compare and contrast the methods used by others previously.

Below is the link to the project Git Repository:

https://github.com/GHovCIND820/CIND820-Final-Project

Initial Results

The dataset contained many object attributes which had to be handled accordingly. A few attributes such as normalized-losses contained some missing values. This however was denoted from the source using a “?”. This was handled by converting the appropriate cells to NaN values and then <Na> values prior to being assigned the columns average. Other columns such as num-of-cylinders was denoted by the english spelling of the numbers, hence were changed from “three” to 3. To complete this data cleaning, the dataset was split between the object attributes and the integer/float attributes. After the cleaning, the dataset was concatenated and Cleaned_Car_Data.csv was created. Looking at the data through some visualizations and methods, I could find that the dataset’s average car price was just over 13,000 dollars. The upper and lower bounds were 45,400 and 5,000 dollars respectively. We can conclude that majority of car prices are below the dataset’s mean by referring to the mean and median. We can conclude that the dataset has a positively asymmetric distribution. When deciphering the correlation matrix heatmap, we can determine some of the driving factors in the dependent variable that is price. Engine-Size, Curb-Weight and Horsepower are the three attributes with the greatest positive correlation to the price. Whereas Highway-Mpg and City-Mpg have the greatest negative correlation. These findings can help with the optimization of the initial model created. This model took every numerical value into account. This included all the attributes that were converted as well. The initial model resulted in the following accuracy and statistics. Mean

Absolute Error: 2929.7148881358653

Mean Squared Error: 19681481.204782143

Root Mean Squared Error: 4436.381544094483

R Squared Score: 0.6724444782976456

Explained Variance Score: 0.6770069282135331

With this being said based on the mean price of 13207.13 and the Root Mean Squared Error of 4436.38, we can conclude that the model has an accuracy of predicting the cars price roughly two thirds of the time. This will have to be optimized for greater accuracy to render this model valid. This hopefully can be somewhat accomplished by the omittance of irrelevant attributes in the model to reduce clutter. Other manipulations to the dataset such as using dummy variables for the cars “make”, may help improve accuracy.

References:

D'Allegro, J. (2021, September 13). *Just what factors into the value of your used car?* Investopedia. Retrieved October 18, 2021, from https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp.

Dubin, R.A. Predicting House Prices Using Multiple Listings Data. *The Journal of Real Estate Finance and Economics* **17,** 35–59 (1998). https://doi-org.ezproxy.lib.ryerson.ca/10.1023/A:1007751112669

Gajera, P., Gondaliya, A., & Kavathiya, J. OLD CAR PRICE PREDICTION WITH MACHINE LEARNING.

Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application. *Unpublished. https://www. ifis. uni-luebeck. de/~ moeller/publist-sts-pw-andm/source/papers/2009/list09. pdf.*

Nelli, F., & Safari, an O'Reilly Media Company. (2015). *Python data analytics: Data analysis and science using pandas, matplotlib, and the python programming language* (1st ed.). Apress.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*(1S3).

Wright, D. B., & London, K. (2009). *Modern regression techniques using R*. SAGE Publications Ltd https://www-doi-org.ezproxy.lib.ryerson.ca/10.4135/9780857024497

Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, pp. 147-154, doi: 10.1109/ICACSIS.2015.7415179.