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

CIND820 - Big Data Analytics Project

Ryerson University, Chang School of Continuing Education

Data Analytics, Big Data, and Predictive Analytics Certification 2021

Author: Gregory Hovesen

gregory.hovesen@ryerson.ca

ID: 500769246

Section: D10

Supervisor: TBA

September 27th, 2021

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. We are also able to use classification models such as decision trees to understand whether or not a proposed price for a product is above or below the current market price. 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.