Stevens Institute of Technology

**A U.S. Marketing Strategy Report for Geely Auto**

Team member:

Zhao, Xin

Wang, Ruixi

Luo, Guirong

Ruan, Tianhao

Course: BIA 650

Professor: Edward A. Stohr

Date: December 17, 2020

Table of Contents

[**1.Introduction** 3](#_Toc59120102)

[**2.Data description** 4](#_Toc59120103)

[**3.Exploratory Data Analysis (EDA)** 4](#_Toc59120104)

[**4.Multiple Linear Regression** 8](#_Toc59120105)

[**5.Nonlinear Programming Analysis** 10](#_Toc59120106)

[**6.Clustering Model Analysis** 10](#_Toc59120107)

[**7.Conclusion** 14](#_Toc59120108)

# **1.Introduction**

Geely Auto, is a Chinese automobile company which is privately held and multinational. The group was established in 1986 and entered the automotive industry in 1997 with its Geely Auto brand. It sells passenger vehicles under the Geely Auto, Lotus, Lynk & Co, Proton, and Volvo brands as well as commercial only vehicles under the London EV Company and Yuan Cheng Auto brands. The group sold over 1.5 million cars in 2018. Geely’s goal is globalization. They want to occupy markets all over the world. Geely acquired the Swedish passenger car maker Volvo Cars from Ford in 2010, it completed the acquisition of British taxi maker The London Electric Vehicle Company in 2013 and acquired a majority stake in British sports carmaker Lotus Cars in 2017.

The United States is known as the country on wheels and has a mature and stable automobile market. Hence, Geely aspires to enter the US market. They prepare to set up their manufacturing unit there and produce cars locally to give competition to their US and European counterparts. Since the American market could be very different from the Chinese market that Geely is familiar with, they want to understand the factors on which the pricing of cars depends as soon as possible. They have contracted an automobile consulting company to make business strategies for them.

Assume we are in this consulting company and we have gathered a large data set of different types of cars across the American market. Our final goal is to help Geely with their design of cars and marketing strategies. We have four steps to go. First, we will make some exploratory data analysis on the dataset and find the relationships between variables. Second, based on the exploratory data analysis, we will use multiple linear regression to model the price with the independent variables. Then we use profit margin to estimate the cost of each vehicle and also make regression on cost. Third, we use the regression model of price and cost, accompany with the demand function, we make an NLP model to maximize the profit. Finally, we cluster the vehicles to analyze the characteristics for each cluster. After doing all the four steps, we are sure to provide guiding advice on vehicle production and sales.

# **2.Data description**

Before doing analysis, we need to do some data preprocessing, including data cleaning and one hot encoding for categorical variables. We dropped the missing values and unnecessary columns, and we got 13 variables finally.

**Manufacturer**: the categorical variable that supply the cars to the market, such as Acura, Audi, BMW etc.

**Sales in thousands**: the sales of each model in one year.

**Vehicle type**: the categorical variable indicates the vehicle type, ‘car’ or ‘passenger’.

**Price in thousands**: the market price of each model.

**Engine size**: the size of an engine.

**Horsepower**: the horsepower of an engine.

**Wheelbase**: the distance between the front and rear axles of a vehicle.

**Width**: the width of a vehicle.

**Length**: the length of a vehicle.

**Curb weight**: the weight of an automobile without occupants or baggage.

**Fuel capacity**: the capacity of a vehicle that could hold fuel.

**Fuel efficiency**: the capacity of an engine, especially that of a vehicle, to obtain energy from fuel.

**Power performance factor**: the factor that could indicate the power performance of an engine.

The Manufacturer and Vehicle type variables are categorical variables, and the value of each variable does not have numerical relationship. Hence, we use pd.getdummies() method in Python to encode the variables to deal with the problem that the classifier cannot handle attribute data. To a certain extent, it also plays a role on expending variables. After we finish the preprocessing, we could make some deeper analysis on the dataset.

# **3.Exploratory Data Analysis (EDA)**

Since the business objective is to help Geely, one of the most influential Chinese company with a huge capital, enter American car market, our preliminary goals are not only maximizing the total profit of American market for short-term cost control but also obtaining a considerable market share for long-term run. Hence our team choose to use non-linear programming to optimize the profit for the goal 1. To run the nonlinear programming model, we need to use linear regression for attaining the relation functions of price model and cost model. Therefore, we do some data visualization to prepare the qualified variables for price model and cost model and to dig out existing cars in which price and cost intervals are most favorable. If the result from linear programming for pricing is directly in that interval, it is win-win situation which means Geely optimizes the short-term profit and occupies a certain market share meanwhile.

After obtaining the initial data-set of different types of cars across the America market and identify our preliminary objectives, then the first step of goal 1 is to designate the predictor variables and dependent variable for both price model and cost model. We do the same procedure to cost model so as to price model thus here we only illustrate how to process predictor variables and dependent variable for price model. Firstly, to simplex the model, we disregard categorical variables manufacturer and model. We treat the price (unit thousands) as dependent variable undoubtedly, and regard all other variables except price, sales as possible independent variables. Secondly, visualizing the pairwise correlations between possible independent variables to see whether there are multicollinearity and then eliminate multicollinearity (Figure2.1). It’s quite obviously there are multicollinearity between initial possible independent variables since many correlation scores are above 0.8, which means the two variables has high correlations. Because Power performance factor is high correlated with both Horsepower and Engine size, we keep Power performance factor as predictor variable in price model only (Figure2.3 &Figure2.4). We save width but move wheelbase for same reason (Figure2.5). The predictor variable which is insignificant correlated to dependent variable price must also be dropped. Therefore, we may need to drop fuel\_inefficiency. (Figure2.6) Lastly we put all the variables kept into the price model and exclude predictor variable fuel\_inefficiency since it has p-value over than 0.05(Table2.1). We next exclude fuel\_capacity too for same reason when run the new model. For price model, we have five predictor variables at end and. the pairwise correlations scores between predictor variables after adjustment are all lower than 0.8 which means they are weak correlated with each other. It is seeming fine.(Figure2.2)

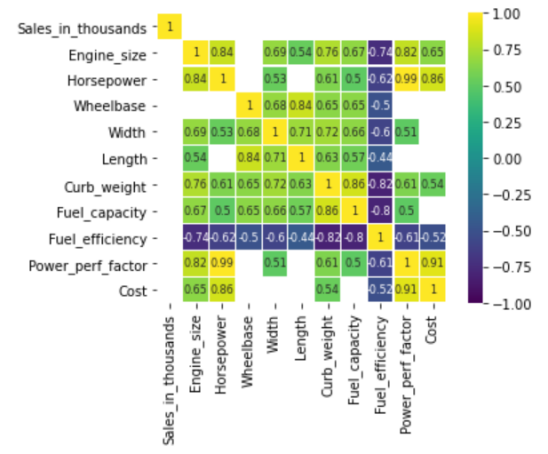
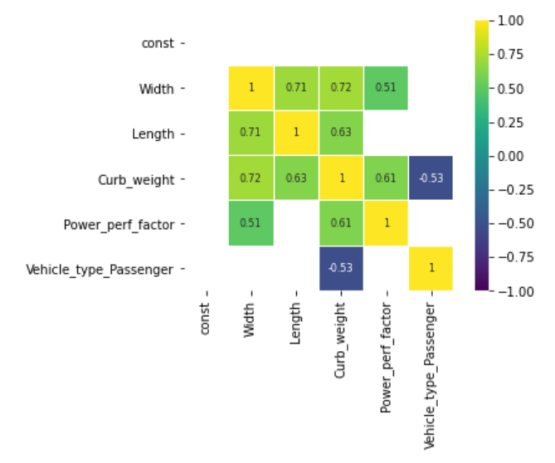


Figure 2.2: pairwise correlation after

Figure 2.1: pairwise correlations before

Figure 2.3:High correlations between Power performance factor and Horsepower

Figure 2.4: Correlations between Power performance factor and Engine\_size

Figure 2.5:High correlations between Width and Wheelbase

Figure 2.6:Insignificant correlations between initial predictor variable(Fuel efficiency) and dependent variable(Price).

Table 2.1: First\_trained Linear Regression Pirce Model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **Std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| constant | 35.3179 | 13.680 | 2.582 | 0.011 | 8.279 | 62.357 |
| Width | -0.6192 | 0.210 | -2.945 | 0.004 | -1.035 | -0.204 |
| Length | -0.2598 | 0.050 | -5.618 | 0.000 | -0.359 | -0.160 |
| Curb\_weight | 7.0447 | 1.765 | 3.992 | 0.000 | 3.557 | 10.532 |
| Fuel\_capacity | 0.5112 | 0.248 | 2.063 | 0.041 | 0.022 | 1.001 |
| Fuel\_efficiency | 0.4078 | 0.213 | 1.919 | **0.057** | -0.012 | 0.828 |
| Power\_perf\_factor | 0.4998 | 0.028 | 17.567 | 0.000 | 0.444 | 0.556 |
| Type\_Passenger | 4.8861 | 1.644 | 2.972 | 0.003 | 1.637 | 8.136 |

After designating the variables of qualified data for price model and cost model, its time conclude the market share of distinct price & cost intervals from exploratory data analysis on probability distribution of price and cost. The graph(Figure 2.7) shows that the cars with relative low price between $10,000 to 35,000 occupies most share market. But for lower price cars, the cost is over than price which means those kinds of models are suffering a marginal loss. It’s kind of weird but reasonable since some companies do so to extend its influence and maintain market share(influence). So directly giving up the extremely low-end market is inadvisable choice for Geely. If the optimal solution for profit shows the pricing targets only high-end market, Geely must do a trade-off to produce some lower prices models for market share. But if the pricing in optimal solution is in the price interval of most share market, it’s outstanding and the pricing doesn’t need to make any change. The only thing we need to focus on next is the existing cars’ collocations and characteristics of corresponding prices, which we will use clustering to explain.

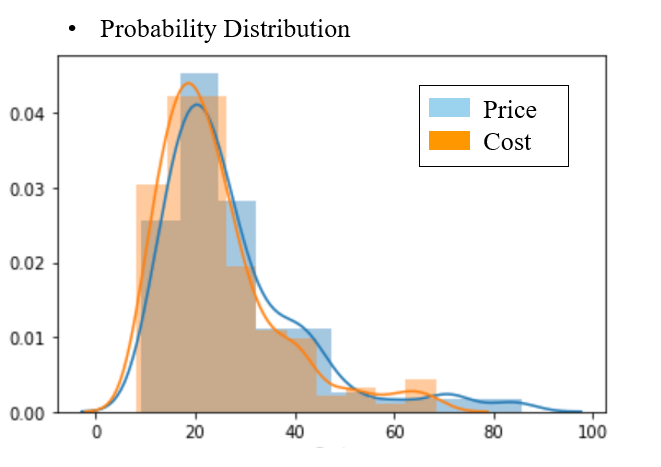


Figure 2.7: Probability distributions of Price and Cost

# **4.Multiple Linear Regression**

The objective of the multiple linear regression model is to best predict pricing of U.S. auto market with as many as possible independent variables. Accordingly, we can use the multiple linear regression model as input of nonlinear programming to learn what characteristics of vehicles that U.S. market prefer, and what price should Geely Auto set to maximize its potential profit in U.S. market.

From exploratory data analysis, we find out that the pricing multiple linear regression should include type\_passenger, width, length, curb\_weight, and power performance factor as independent variables. The following is the regression result:



The overall model performance can be evaluated by the R square factor. The R square of the pricing model is 86.16%, which means that 86.16% of price can be explained by the linear regression model. The p-value of each variable is below 0.05, which means that all independent variables are statistically significant. Furthermore, we removed other variables that are highly correlated with each other as mentioned in the exploratory data analysis, our team believes that this multiple linear regression model is the best model we can obtained from the dataset to predict the pricing of U.S. auto market.

The model demonstrates that the most crucial variable to predict price is the curb weight because the curb weight variable has the highest coefficient. In addition, the model also illustrates that the power performance factor is perfectly correlated with predict price. The following two graphs show the correlations between demonstrated variables and target variable price.

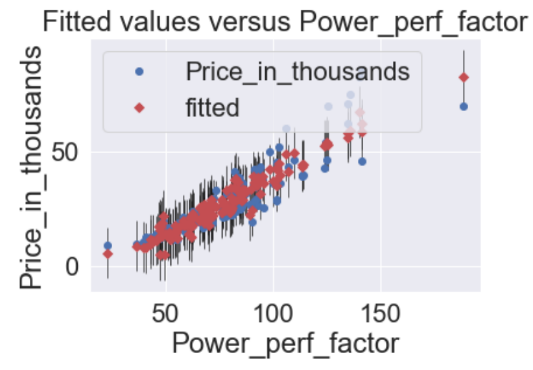
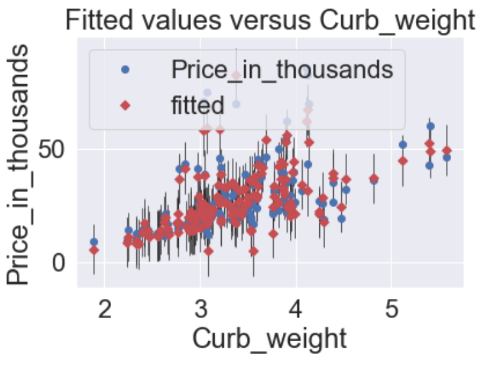
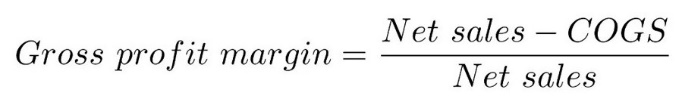


Figure 3.1 curb\_weight VS price Fugure 3.2 power\_perf VS price

In order to create a best nonlinear programming model to predict maximizing profit, our team decides to take advantage of profit margin for each manufacturer to predict cost. The data of profit margin for each manufacturer is from macrotrends financial website. The basic formula is the following:



We calculate the cost of goods sold = (1- Gross profit margin) \* net sales by transforming the formula. As we know that net sales = price \* sales. Thus, we can find the cost of goods sold per unit by (1- Gross profit margin) \* price. Then, we can fit our predict cost into the multiple linear regression model, we get the following result:



The R square of the cost model is 87.76%. The p-value of most of the variables is below 0.05 except for width. The p-value of width is 0.093. However, for the coherence of our two regression models, our team decides to include width in the cost model for better performance of nonlinear programming.

Eventually, we simply use price to predict sales as demand formula for later use in our nonlinear programming analysis:

Demand = 93.1 – 1.447 Price

# **5.Nonlinear Programming Analysis**

The objective of nonlinear programming analysis is to find out what type of vehicles that U.S. auto market prefer, and how much price Geely Auto set to maximize its potential profit in U.S. market.

The nonlinear programming (NLP) takes the regression result of price, cost, and demand as input. Hence, the formulas of price, cost, and sales are the intercept plus the product of coefficient of variables and variables. The profit is price minus cost times sales. The logic upper bound is the maximum value of each variable in the dataset, and the logic lower bound is the minimum value of each variable in the dataset. Accordingly, our model uses GRG nonlinear solver to maximize profit by changing variable cells, subject variables to the logic upper bound and logic lower bound constraint.

The NLP result suggests Geely Auto to set price at 23.706 thousand dollars to get 360.84 profit in million dollars. In addition, our model also indicates that in order to maximize profit, Geely Auto should produce vehicles with passenger type, 62.6 width, 224.5 length, 5.572 curb weight, and 39.76 power performance factor. Based on the result, our team believes that our model implies to produce as long as possible in length and as heavy as possible vehicles to attract customers, and to produce narrow in width and relatively low power performance factor vehicles to save cost to obtain cheap price competitive advantage.

Therefore, our team recommends Geely auto to produce long and heavy SUVs or trucks to attract customers, and to produce narrow and low power performance factor SUVs or trucks to save cost to set price at 23.706 thousand dollars per unit. The projected profit is 360.84 million dollars.

# **6.Clustering Model Analysis**

The objective of the clustering analysis is to understand U.S. auto market and to come up with potential competitors for Geely Auto. The clustering analysis wind up with analyzing the potential profitable market for our recommendation that draws from NLP analysis.

The cluster analysis benefits Geely marketers to identify car market segments and its main competitors from each class. An intuitive sense of what leads to the forming of the homogeneous groups of cars is straightforwardly observing after the use of the Evolutionary Solver in Excel.

The goal is to group these 153 samples(Figure 4.1) into four clusters of cars type that are characteristic similar, where similarity is defined in terms of the 11 aspects including engine size, horsepower, wheelbase, width, length, curb weight, fuel capacity, fuel efficiency, power performance factor, sales, and price. The target is to minimize the sum of the distances from each car to its cluster center by means of the Evolutionary Solver assigning all these cars, and the Evolutionary Solver uses genetic algorithms to find the optimal or good solution.

表格, Excel

描述已自动生成

Figure 4.1: Data for partial selected cars

To avoid the situation where some columns of variables are more spread out than other columns and achieve unit-free(Figure 4.2), standardization, where data are subtracting the attribute’s mean and dividing the difference by the attribute’s standard deviation, is of necessity. After standardizing the values, the next step is building a cell table(Figure 4.4) to measure the distance from the car to each cluster center with the usual Euclidean distance formula.

Distance = 表格

描述已自动生成

Figure 4.2: Summary data for attributes of cars

Then the sum, which is over all the eleven attributes, is calculated for minimum via the Evolutionary Solver. The changing variable cells(Figure 4.3) are the car index, and the only constraints here refer to that car index has to be an integer and should not beyond the range of all the 153 samples.

表格

描述已自动生成

Figure 4.3: Decision variables and objective cell

表格

描述已自动生成

Figure 4.4: Other calculations for cluster analysis

图片包含 散点图

描述已自动生成Additionally, the python allows visualizing the cluster centers(Figure 4.5) into a scatter plot graph, a way to behold the distance between these cars and their centroid.

Figure 4.5: A scatter plot graph of cluster centers

Nevertheless, in most real-life cases, the optimal number of clusters is hard to decide, and the experiment trying here is to identify if this particular point, like four, could provide sufficient information to investigate. The interpretation for the solution, which found LS400, ES300 of Lexus, Bonneville, and Sunfire of Pontiac as the best in minimizing the distance, is explained in the following.

Cluster 1, most of the car manufacturer is Cadillac taking 17%(Figure 4.6), Mercedes-Benz, the same percentage, Porsche, taking 11%, have larger engine size, higher horsepower, longer wheelbase, heavier curb weight, greater fuel capacity, and better power performance factor. Hence it’s taken for granted that these types of cars are the most expensive across all the clusters, namely higher selling prices.

Cluster 2, typical manufacturers like Mitsubishi having 11%, Lexus with 9%, and Volvo, same with 9%, has an average in all levels of function.

Cluster 3, with top manufacturer Dodge occupying 19%, Ford holding 14%, and Buick with 11%, emphasizes a long wheelbase, big car size in width and length, which attracts customers illustrated on the NLP model.

Cluster 4 contains most of the cars of 63, and includes representative manufacturers like Volkswagen engaging 10%, Toyota taking 10%, and Saturn with 8%, outstanding in fuel efficiency. Moreover, sale behaves well, and the price is cheaper.

Moreover, the average performance(Figure 4.7) of each center has been attached. Cluster 1 has an average of 4.772 engine size, 291.889 horsepower, 109.311 wheelbase, 4.140 curb weight, 22.217 fuel capacity, 126.279 power performance factor, and 56.501 price in thousand. Cluster 2 stands in the median. Cluster 3 has 115.462 wheelbase, 74.357 width, and 201.562 length. Cluster 4 overall has 27.063 fuel efficiency, 56.325 sales, and 18.199 prices in thousand.

表格

描述已自动生成

Figure 4.6: Clusters in the solver solution

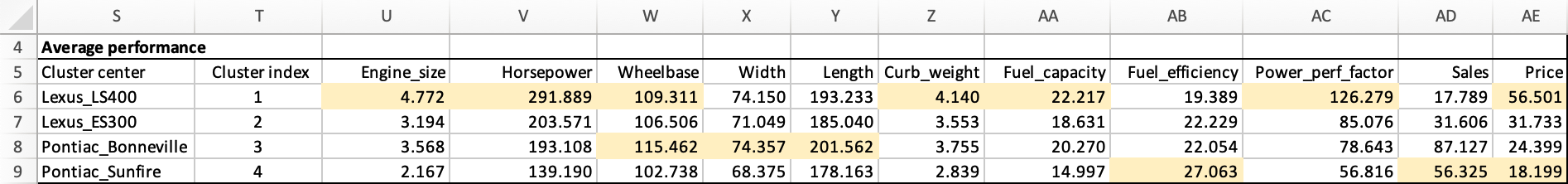


Figure 4.7: Average performance of cluster center without standardization

# **7.Conclusion**

By conducting the research on finding which factors influence customers most in buying the cars and the optimal price, the project facilitates Geely Auto to grasp the opportunity to design a most accepted car and expand the market share through the localization strategy.

In the beginning, the exploratory data analysis process contributes to discovering the correlations between variables and how their relationships behave. The scatter plots demonstrate the high correlations among initial predictor variables as in power performance factor with horsepower, with engine size, and width with wheelbase. Insignificant correlations could be noticed in price with fuel efficiency. The pairwise correlation between variables alongside helps in removing high correlation predicted variables.

After that, the multiple linear regression and non-linear programming aid in achieving an optimal price and advising an ideal design figure conveyed in the results of changing variables. The NLP analysis concludes that customers prefer longer (224.5 in length), heavier curb weight (5.572). In addition, Geely Auto can take advantage of low power performance factor (39.76) cars like sport utility vehicle type to create a cost saving competitive advantage among other manufacturers. More specifically, the suggested prices range at 23 thousand dollars.

Furthermore, the clustering categorizes analogous cars in terms of function. The model indicates that most of the cars that sell well confirmed the similar features proposed in the NLP model. Precisely, the emphasis on width, length, fuel efficiency, and wheelbase.

Therefore, with the step-by-step investigation and combining all of this noteworthy analysis information, the conclusion is that, for Geely auto company, it’s reasonable to capture such market potential to plan or produce this kind of car. To make a breakthrough and have a competitive edge, it shall concentrate on characters like long, wide, and long-wheelbase in car dimensions, low power performance factor and great fuel capacity in car performance, and the most important, cheap, and economical in prices. In addition, for this class of cars, the possible competitors of the manufacturers are Toyota, Volkswagen, Dodge, and Ford, which the Geely company may pay more attention to its latest move.

All in all, our project provides a perspective in making decisions and offers some marketing strategies and business insights for the company to consider with in-depth analysis.