DESAUTELS Faculty of Management Faculté de gestion

McGill

McGill University

Desautels Faculty of Management

MGSC 661 – Data Analytics and AI for Business

Fall 2024

Final Project

Every Man Can Talk About Cars Like A Pro:

Gauging How Much Your Friends' Cars Are by Just Chitchatting

12/6/2024

Table of Contents

Chapter 1: Introduction

In grown men's world, cars have remained one of the timeless topics among a group of friends and workplace or even small talks that can take place nearly everywhere. Men treasure their wheels, and they sure like to talk about them. Sometimes, especially men, converse on the topic because they are th enthusiast, who spend all their downtime looking at all things about cars; sometimes it is because they recently have purchased ones or looking for one. But for a lot of times, people like to throw what they drive into a conversation to signal their social status. Regardless, car talks are omnipresent thus being able to hold up such conversation is essential to social skills.

1.1. Who and How This Project Can Benefit

The project aims to impart knowledge of how much a car might be worth to anyone who catch themselves in a conversation where people talk about cars by referring to the specifications as to how much horsepower it has, what type of engine there is, and so forth. Regardless of the intent behind the conversation, bringing up the tag price seems like a faux pas for most occasions. Therefore, being able to gauge the price without bluntly poking the question helps people in the situation to follow up with the topic, where they add in information about other cars within the similar price range.

The project is to build a classification model that considers things that people typically mention regarding their car, and render a prediction that will tell which price tier the car in question may lie within, i.e. luxury, premium, mainstream or economy. Additionally, the prediction might bestow insights into what characteristics are most likely to be the dominant determinant of the price tag that would be helpful for consumer when they shop for their rides.

Chapter 2: Exploratory Data Analysis

The dataset comprises 205 observations, with 26 distinct features, including both numerical and categorical variables. The primary dependent variable is ***price***, which simply is the price of an vehicle. Note that we do not know for certain what type of market the car prices in the dataset are nor the denominating currency, however the dataset consists of data collected from an automotive yearbook from 1985 published in the United States. Therefore, the prices are assumed to be the MSRP for new cars in USD in 1985's dollar value. The mean car price in the dataset is $13,207, with the range of price from $5,118 to $45,400. The median price of $10,295 indicated a right-skewed distribution. The histogram and boxplot for the distribution of prices are shown in Figure 1 and 2 respectively in Appendix 1.

To explore the relationships between the features and the target variable, i.e. car price , we employed scatter plots for numerical variables and box plots for categorical variables. Scatter plots allowed us to visually

assess the potential linear and nonlinear relationships between continuous predictors can the target variable. Such continuous predictors in a broader term include the curb weight, dimensions of vehicles, dimensions and performance of engines, and fuel efficiency. Similarly, box plots were used to examine the distribution of price across various levels of categorical variables, such as car maker, powertrain type, e.g. 4WD, FWD or RWD and body style, e.g. sedan, convertible, etc., number of doors, engine placement location, fuel and aspiration type. These visualizations, shown in **Error! Reference source not found.**, provided key insights into the correlations and potential predictive power of each feature.

2.1. Numerical Variables

The numerical variables included in EDA process are grouped into 2 types, the first of which is the dimension of the vehicle itself and the second is the dimension and performance metrics of the engine. The first type of variables include: 1) curb.weight, i.e. the weight without passenger or cargo, 2) length, i.e. the full length of the vehicle, 3) wheel.base, i.e. the distance between the front and rear axis, and 4) width, i.e. the width of the vehicle. The second type of the variables include: 1) engine.size, i.e. the engine displacement in air volume, 2) horsepower, 3) peak.rpm, i.e. the engine's RPM measured at the peak horsepower, and 4) highway.mpg, i.e. the fuel efficiency measured when vehicle operating on highway, and.

Under the first type of numerical variables, we can see that in general, heavier, longer and wider vehicles come with a more hefty price tag. Clear upward trend can be observed from the scatterplots in Figure 1, 2 and 3 of Appendix 2. Counterintuitively, the longer wheel base does not seem to lead to higher prices as observed in the relationships amongst other dimension and the prices (shown in Figure 4 of Appendix 2).

Under the second type of variables which primary describe the characteristics of engines, we can see clear upward trends amongst engine size, horsepower and peak RPM, entailing better performance cars come with higher price (shown in Figure 5, 6 and 7 of Appendix 2). In contrast, there is a downward pattern in fuel efficiency, indicating more fuel-efficient car are cheaper (shown in Figure 8 of Appendix 2). These observation from the scatterplots conform to common sense knowledge of cars, where premium cars typically have higher-performance engines, and economic choices of cars have better fuel efficiency catering to price-sensitive consumers.

2.2. Categorical Variables

The EDA process include categorical variables of 1) drive.wheels, i.e. the driving wheels of the vehicle, e.g. 4WD, RWD or FWD, 2) engine.location, i.e., engine placement, e.g. in the front or back of the car, 3) num.cylinders, i.e., the number of cylinders in the engine, 4) aspiration, i.e. how the engine intakes air for combustion, e.g. standard or turbo-charge, 5) fuel.type, e.g. diesel or gasoline, 6) body.style, e.g. sedan,

convertible, hatchback, etc., 7) num.of.doors, i.e. the number of doors, and 8) make, i.e. the make of the vehicle.

Once again, the observations from the categorical variables' relationships with price conforms to our common understanding about cars. Features like rear-wheel drive, rear and turbo-charged engine, and convertible top are more commonly seen in premium cars, especially from the European car makers. As already implied by the engine size, more cylinders in the engine gets higher price. The relationship between car makes and the prices clearly shows that the commonly known luxury car brands have higher price, such as BMW, Mercedez, Jaguar, Volvo and Porche. (See Figure 9, 10, 11, 12, 13, 14, 15 and 16)

Chapter 3: Model Selection

To develop a classification model which can predict the tier of vehicle price, using the predictors that describe the vehicle, Random Forest was chosen primarily due to the following reasons. First, since we have a mixture of data types and some seem to behave in a non-linear relationship with the target variable, Random Forest is more adept in capturing complex pattern and robust to outliers in data than Logistic Regression due to its focus on splits rather than absolute values. Secondly, we have predictors that are interconnected by nature and together they would influence the vehicle prices, e.g. how number of cylinder and engine size together influence price, Random Forest inherently captures interactions between those interactions without requiring manual specification of interaction terms. Thirdly, Random Forest comes with built-in feature selection method, conducive to the understanding of which car characteristics are most predictive of the price tier. And lastly, Random Forest are less prone to overfitting owing to its bagging approach, compared to boosting where there is a larger tendency to overfit and may require careful tuning.

In the steps of preparing, training and extracting the results from Random Forest model, we will do the following:

1. **Data Cleaning and Preprocessing**
    i. Handled missing values, in which we replace the data shown as "?" in string format with NA.
    ii. Converted variables that are numeric in nature whereas present as string type were con into integer, such as price, bore for engine's bore diameter, stroke for engine's stroke length, horsepower, peak.rpm, num.of.cylinders, and num.of.doors.
    iii. Factoring data: Converted the data in either categorical columns or the data where there are only few discrete unique values, such as num.of.cylinders and num.of.doors, into factors.
2. **Feature Engineering**

i. Divide the car prices into 4 tiers that denote luxury, premium, mainstream and economy as follows: (shown in Figure 1.)
   a. **Economy:** prices from the 1st to 30th percentile
   b. **Mainstream:** car prices from the 31st to 70th percentile
   c. **Premium:** car prices from the 71st to 90th percentile
   d. **Luxury:** car prices from the 91st to 100th percentile

Figure 1.

Distribution of Car Price Tiers

	Price Tier	Frequency
1	Economy	61
2	Mainstream	80
3	Premium	40
4	Luxury	20

   e. Dropped the original price variable in numeric form to not leak the price information to the model while predicting its price tier.

3. Model Training

Though random forest is considered relatively robust to noisy data, multicollinearity and overfitting, the parameters still require careful setting to prevent it from plainly memorizing the training data and failing to generalize the pattern. Thus it remains crucial to balance the trade-off between accuracy in prediction and overfitting, so that the spread between the accuracy score across training data and testing data are reasonable while achieving as high prediction accuracy as possible.

For all the iterations of the Random Forest, the model is trained on 70% of the data and tested on the remaining 30%. In the first iteration of the Random Forest, the parameters have 500 trees, and use the default setting of the square root of the number of features to be considered at each split. Whereas the rest of the parameters were left with default values. Under this setting, 83.33% of accuracy in the testing data and 99.28% in the training data were attained, with the spread of 16%. A 5-fold cross validation ensued shows an average of 82.08% in accuracy on testing data. Here the overfitting issues was suspected given the discrepancy between accuracy on training and testing datasets may have room to improve.

Next step is to experiment with different parameters to narrow down the spread between accuracy on testing and training dataset. Varying setting of ***ntree*** and ***mtry*** both show little impact on narrowing the spread. However by adjusting ***maxnodes,*** which specifies the maximum number of leaf nodes each tree in the Random Forest can have. Without specifying the value of ***maxnodes*** in the model, the tree will grows trees

without an cap on the number of terminal nodes, consequently creating fully grown trees that may lead to overfitting. By decreasing ***maxnodes***, it can limit growth into overly complex trees and prevent overfitting. See shown in Table 1 for the summary of different ***maxnodes*** settings and its impact on the model's performance, with maximum nodes of the branches are limited to 20, we can bring down the spread in accuracy without sacrificing the prediction accuracy on the testing data. And the cross-validation accuracy with 5-fold converges to the average accuracy of 77.78%. As the result, we will use such to extract important features and insights in the next chapter.

Table 1.

Accuracy/*maxnodes*	**Unspecified**	**5**	**10**	**20**
Test Accuracy	83.33%	70.37%	77.78%	83.33%
Train Accuracy	99.28	78.42%	89.93%	94.24%
Spread	15.95%	8.05%	12.15%	10.91%

Chapter 4: Result Analysis and Insights

4.1. Confusion Matrix

The following Table 2. shows the confusion matrix for the prediction on testing data. The first column denotes the actual price tier in the testing data and the top row denotes the predicted categories. We see the model does most poorly in predicting the premium tier.

Table 2.

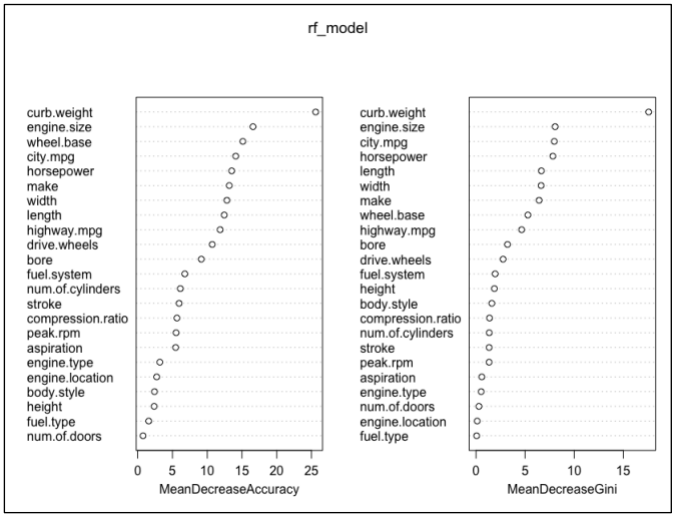
Actual/Prediction	Economy	Mainstream	Premium	Luxury
Economy	13	2	0	0
Mainstream	0	18	2	0
Premium	0	4	9	0
Luxury	0	0	1	5

4.2. Feature Importance

Based on the Random forest model with the tuning of the parameter, *maxnodes*, we obtained the important features shown in Figure 2. On the left panel, the importance of a given feature is measured by the decrease in accuracy when the feature's values are randomly shuffled so as to break the association with the target; on the right panel, a feature's importance is measured by its contribution to reduction in impurity across all splits in all trees.

From both measurements, curb_weight, engine_size, city_mpg and horsepower are common features shown in the top 5 important features for both measures respectively. This presents us the insight that the weight, engine size in terms of air displacement, fuel economy when driven in the city and the power performance are the critical factors in determining its price tier. The top 5 important features conform to the commonsense understanding of cars in regard to how the specifications relate to their price. Bigger and heavier cars are typically associated with the roomier space the car has, and that leads to higher price, possibly due to more safety features, larger engine size, etc. Similarly with the horsepower and fuel efficiency, they are frequently used as the metrics when people make purchase decisions or value cars.

Figure 2.



Chapter 5: Conclusion

We started the project by doing EDA to spot potential relationships among the independent and target variables, where we saw some obvious trends between the independent and target variable pairs. Given that we have more than 20 variables and multicollinearity are likely to be present, Random Forest was chosen to be the model for classification which borrows strength from being less prone to overfitting and less sensitive to outliers. Additionally, Random Forest requires less fine-tuning on parameters, compared to Boosting methods. Thus, we can save a tremendous amount of time in data-preprocessing and tuning the model. The accuracy of the model is 83.3% on testing data, which gives us a good classification result considering the context and objective of the project, that is to equips individuals with the ability to infer the price tier of cars from key specifications, enhancing their ability to participate in car-related discussions without directly mentioning prices.

References

[1] Demuro, D. (2024). Car Specs Guide: Everything You Need to Know. Autotrader.

[2] Maddali, S. (2022). Predicting Car Prices Using Machine Learning and Data Science. ODSCJournal.

Appendices

Appendix 1

Figure 1

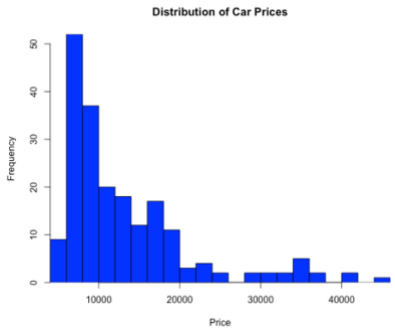


Figure 2

Appendix 2

Figure 1

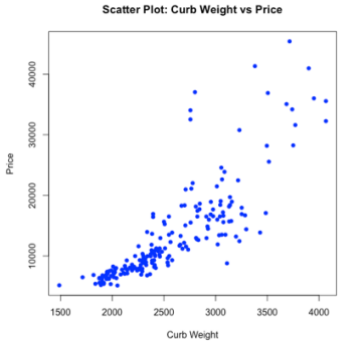


Figure 2

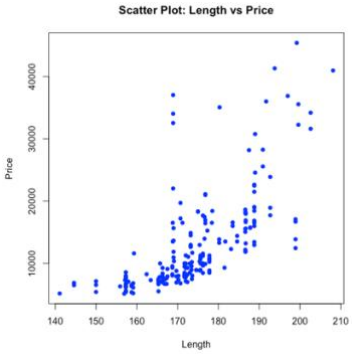


Figure 3

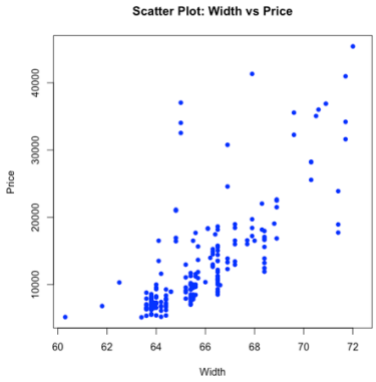


Figure 4

Figure 5

Figure 6

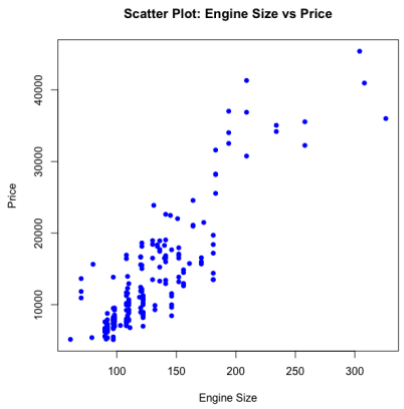


Figure 7

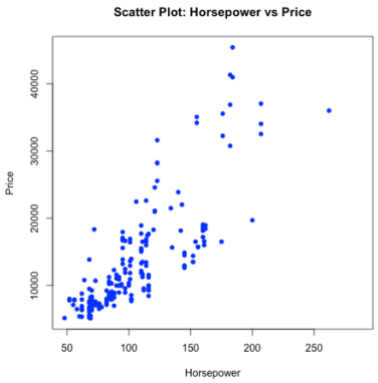


Figure 8

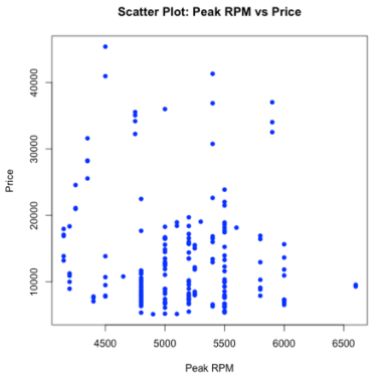


Figure 9

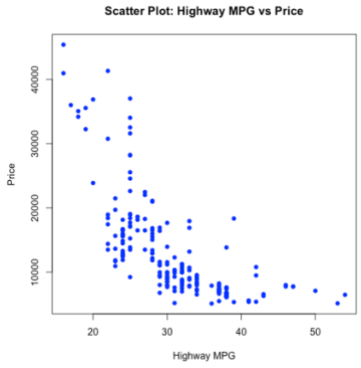


Figure 10

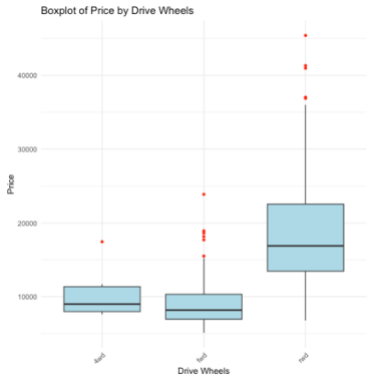


Figure 11

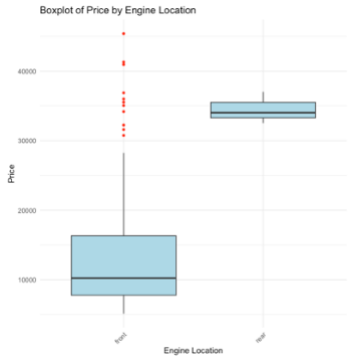
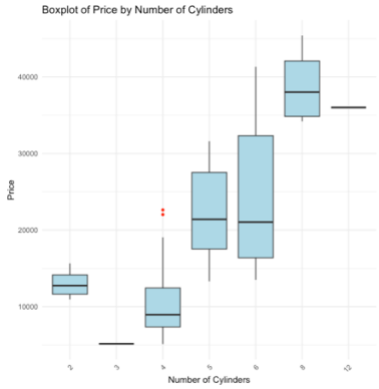


Figure 12



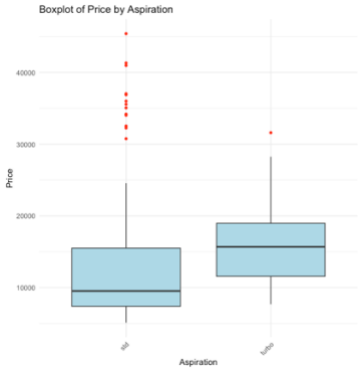


Figure 13

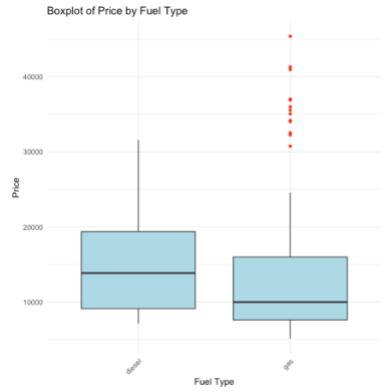


Figure 14

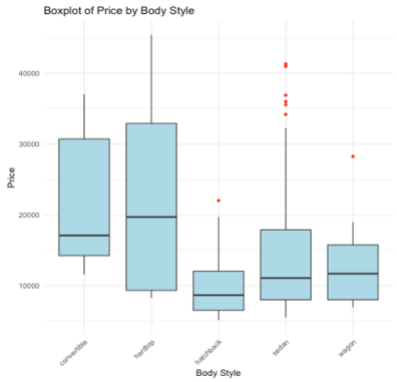


Figure 15

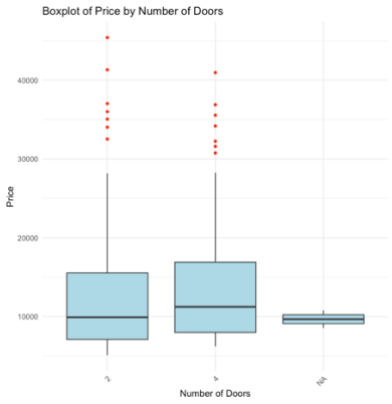


Figure 16