Car Price Determination Report

Jared Barnes
October 2022

Business Problem

Car manufacturers must determine Manufacturer's Suggested Retail Price (MSRP) for newly made cars. MSRP contains subjective components that consider market value and manufacturing profit on top of manufacturing, distribution, and sales costs. A calculated determination of market value for existing cars, based on car features would greater inform the MSRP, allowing for optimal profit without overpricing.

Car dealerships, specifically used car dealerships, lose thousands of dollars every year buying cars that sit unsold on their lots. Profit margins are limited based on car popularity and market value, depending on the type of car. Not only is it difficult to sell every car the dealership has, but even when the car is sold, there is a limited profit the dealership can make. If a dealership could understand what car features led to highest car value and highest car popularity, they could significantly reduce the number of cars sitting unsold on their lot. This would save them thousands. Further, if the dealership had access to a calculated market value for each car, they could dial in on maximum profit and reduce overspending on used cars.
See the following link for more info: Do Car Dealers Really Lose Money? How To Prevent It | Carketa App (getcarketa.com)

Car buyers often spend too much for cars whose values are significantly less than expected. A shiny exterior or a persuasive salesman can deceive buyers into thinking a car is worth more than it is. It would be immensely valuable if a buyer could quickly determine true market value for a car so they can save the most and get the most for their money.

Dataset

The car data are in one data set, scraped from Edmunds.com and Twitter. Each row is a different car and each column is a different feature of that car. Some features include make, model, year, engine, and other properties of the car used to predict its price.

Anticipated Data Science Approach

My approach aims to accomplish two things:
(1.) Identify the top 3 to 5 car features that have the greatest influence on market value. These features can be used by dealerships and manufacturers in purchasing and/or building the most valuable and popular selling cars.
(2.) Build a machine learning (ML) predictive model to make accurate, reproducible predictions of price, given car features, so as to dial in on the true market value. This model and its information can be used by dealers, manufacturers, and buyers to maximize profit or savings.

I will complete steps in the data science method including data cleaning and wrangling, EDA, feature analysis, feature manipulation, train/test splitting, scaling, ML, and prediction and evaluation. I plan to test and compare the following 3 ML algorithms:

- Linear Regression
- Random Forest Regression
- XGBoost Regression

Data Cleaning:

All columns of the data set were of the expected data types. There were, however, numerous null values in 5 out of 16 columns. I calculated the percentage of the columns that were null, and none were above 0.58 % null, except for one column: “Market Category”. I replaced the nulls from that column with the string “No Category” and removed the other null rows completely. Overall, I only removed 101 rows, leaving a remaining 11,812 rows of data. The data were clean after dropping null values.

EDA:

1.) Histograms of all numeric features showed the expected normal distributions, except for Popularity and Number of Doors. I expected this because there is nothing to suggest that popularity as a quantified value should follow a normal distribution and number of doors on a car is overwhelmingly either 2 doors or 4 doors (a few 3 door cars existed in the data).

2.) I calculated a correlation heatmap to determine the most correlated features. Engine Horsepower and Engine Cylinders had the highest correlations with MSRP so I looked into those further, noting that the Bugatti Veyron 16.4 had the highest MSRP, highest horsepower, and greatest number of cylinders.

3.) I plotted the mean and median MSRP values by year, looking for any large trend in price over time. Because of large outliers, I chose to follow the median MSRP per year (**Figure 1**), seeing that there was a large spike in car prices in the year 2000. After that, MSRP has gone up steadily with the exception of minor drops along the way.

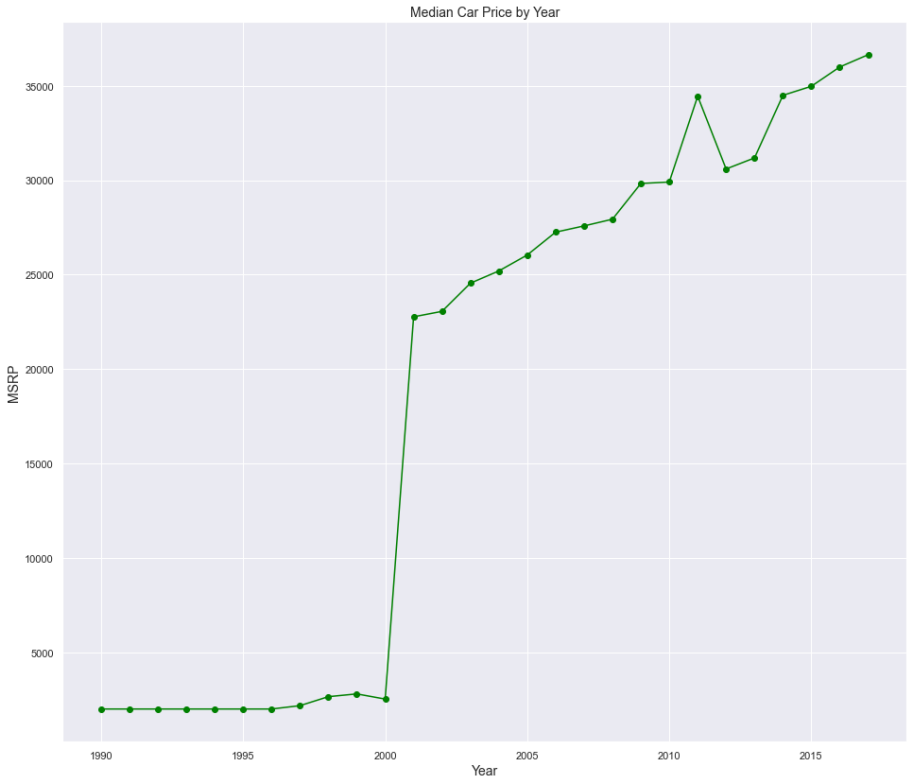

Figure 1: A plot of median MSRP as a function of time. A massive increase in MSRP occurred in the year 2000. Afterward, a steady increase in car value has continued with the exception of minor drops in MSRP along the way.

Converting Categorical Data to Numeric Data:

Since the goal was to create a regression model, I needed to convert categorical data to numeric data. That meant that every category option was given its own column with either a 1 or a 0 in each entry as a Yes or No for the category. After doing this, I had a total of 1059 columns.

Train Test Splitting and Scaling:

Splitting: The data set was split into training and testing data - 80% training data, 20% testing data. All features except for MSRP were assigned as the X variable, while MSRP was assigned as the Y variable.

Scaling: The X variable was scaled using sklearn's MinMaxScaler() function. Both training and testing data were scaled for the X variable features. All numeric values were thus scaled between 0 and 1.

Model Algorithm Comparison:

I fit 3 models to the data to determine which works best at predicting the target variable using three different algorithms. The following 3 algorithms were applied and compared to one another:

- Linear Regression
- Random Forest Regression
- XGBoost Regression

To compare the three algorithm models, I calculated the following scores for each model and both training and testing:

- R^2
- Mean absolute error
- Mean squared error
- Root mean squared error

Of the three algorithms, *the model for the XGBoost Regressor algorithm performed the best (0.986 R^2 test score)*, slightly better than the Random Forest Regression (0.970). Both XGBoost and Random Forest outperformed Linear Regression (-3.300855e+19).

XGBoost Hyperparameter Tuning:

I moved forward with the XGBoost model, doing hyperparameter tuning. However, it was worth noting that the R^2 scores were so close to 1 that the model was most likely overfitting. A 5-fold cross-validation was used necessary to identify the best parameters that did not yield overfitting.

I used GridSearchCV to tune the model and achieved an **R^2 value of 0.910**, with the following parameters:

{'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 180}

Model Effectiveness:

After tuning, the model achieved an **accuracy score of 0.980**. Clearly, the model was predicting the test data very accurately, leading me to move forward in completing the second goal of the project: determine top car features in importance to MSRP.

Calculating Feature Importances:

I made predictions and identified features importances using the "model.get_booster().get_score(importance_type='weight')" method on the tuned XGBoost Regressor model. The following were the five most important features influencing MSRP in descending order (see **Figure 2** and **Figure3** below):

1.) Engine HP
2.) Year
3.) highway MPG
4.) Popularity
5.) city mpg

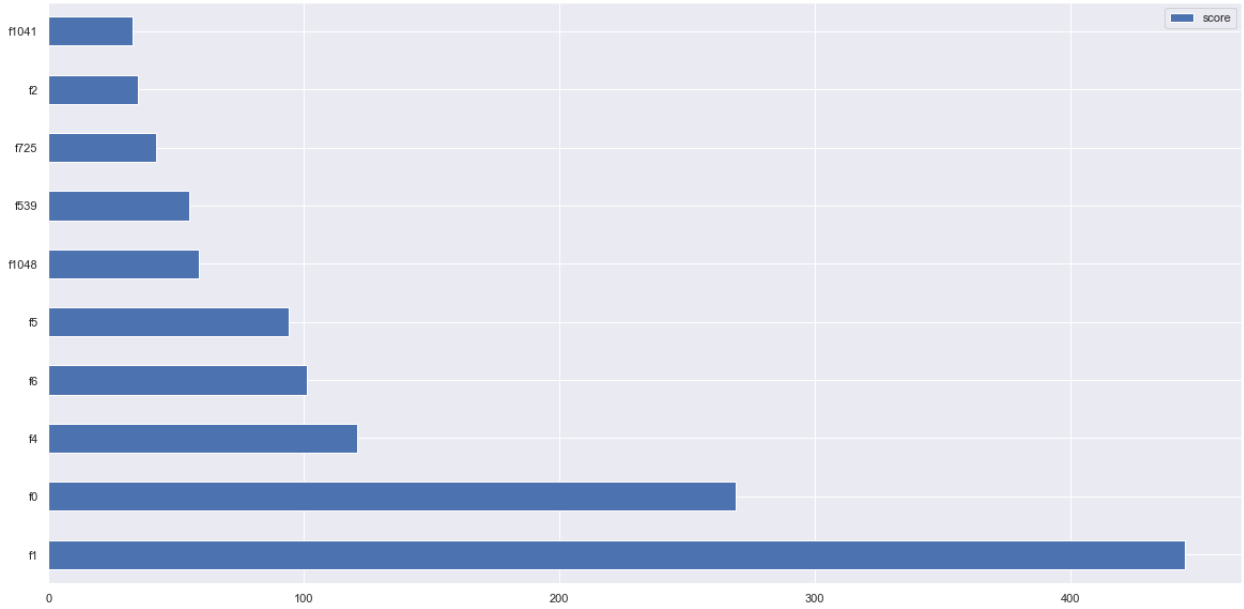

Figure 2: A barplot of features and their respective weights of importance on MSRP. See Figure 3 as the legend for features (i.e. f1 = "Engine HP", f0 = "Year", f4 = "highway MPG", f6 = "Popularity", and f5 = "city mpg")**.**

	Year	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg	Popularity	Make_Alfa Romeo	Make_Aston Martin	Make_Audi
0	2011	335.0	6.0	2.0	26	19	3916	0	0	0
1	2011	300.0	6.0	2.0	28	19	3916	0	0	0
2	2011	300.0	6.0	2.0	28	20	3916	0	0	0
3	2011	230.0	6.0	2.0	28	18	3916	0	0	0
4	2011	230.0	6.0	2.0	28	18	3916	0	0	0

5 rows × 1058 columns

Figure 3: A snip of the car dataframe showing the column names that correspond to the feature names shown in Figure 2 (i.e. f1 = “Engine HP”, f0 = “Year”, f4 = “highway MPG”, f6 = “Popularity”, and f5 = “city mpg”).

Recommendations:

If a car dealership were to use this model, they could infer that the 5 main aspects of cars shown above are the most important to focus on when buying or selling vehicles. Further, they could identify the cars that have the best value for those qualities if they are buying, and try to sell them for higher price knowing that the market values those features.

A car buyer could use this model to calculate the exact value of the car they are looking at. They could also decide to look at cars based on the important features, choosing to compromise feature value with cost if needed.

Moving forward, I would determine the following in order to develop the project even more:

- Obtain more data: I would attempt to obtain more data columns/car features. Specifically, I would like to see a feature that describes the wear and tear levels on a car. Is it lightly driven/used or greatly driven/used? What are the conditions of each car? Further, it would be interesting to see if color or design play a part in MSRP / market value.
- Determine cost to produce features: I would determine what it cost to make each feature, given the car make and model, in order to produce a cost-effectiveness metric. This metric could be used to show which cars have the good important features that were cheaply made, maximizing profit.
- Specifics of important features: I would identify specific attributes/quantities of the 3 most important features that lead to highest MSRP, so as to provide a more detailed recommendation for car buyers and sellers looking to maximize value for car features.