Used Car Price Prediction

CISC 372 – Report

Clare McMullen
20121629
18cim@queensu.ca
April 19, 2021

CONTENTS

ABSTRACT

The objective of this study is to create a machine learning model to predict used car prices in order to aid buyers and sellers in navigating the used car market, and to determine the driving factors behind these predictions. This experiment was achieved by using the Craigslist Used Car Dataset from Kaggle and the Sci-kit learn function RandomForestRegressor to build a model. The main result found was that this dataset and this method provided a very accurate price prediction, and that size, year and manufacturer were the top three most important characteristics in predicting a vehicle's price. Overall, we can conclude that this method is a successful way of predicting prices for used car data, and that these factors are important in determining the price of a used vehicle.

1 INTRODUCTION

Every year millions of used cars are bought and sold in the US and Canada. Buyers and sellers utilize mediums such as Craigslist, Kijiji and Autotrader to list and discover vehicles for sale in their area. For example, let’s say you are looking to buy a used car. You have a specific model and year that you are searching for and you discover an advertisement for the perfect car. After viewing the asking price, you are unsure if this is a reasonable price. Going to the trouble of researching multiple other prices of similar vehicles is tiresome and tedious. Conversely, I propose a method where one could simply input basic information about a specific vehicle into a program that would then generate a reasonable asking price based on previous sales. This would allow buyers to determine if an asking price is above average, below average, or fair. It would also allow sellers to easily determine a reasonable asking price at which to list the vehicle.

Many have attempted to use machine learning to predict used car prices in the past, however I hope to contribute addition information on the subject. My goal is to identify the most important factors in my price prediction model so that buyers and sellers may better understand the reasoning behind the current value of their vehicles. Some challenges that arose during the experiment came due to the abundance of data from such a broad spread of car sales. However, this will be addressed in detail in the Proposed Methods section.

2 PROBLEM STATEMENT

The problem is to create a machine learning model that can accurately predict numeric car prices using specific variables such as year and model. After successfully finding a model that achieves accurate results on training and testing sets, the feature importance can be extracted. The data being utilized in this problem will be from the *Craigslist Used Car Dataset*. This set includes data from used car advertisements on Craigslist from across the United States and can be found on the Kaggle website here. The data set includes variables such as manufacturer, model, year, odometer, size, type, paint_color, cylinders, etc. The column 'price' will be used as the target variable. Since this is a numeric target, we will be using a regressor model.

As work began on this experiment, it was quickly realized that the scope was too large. Prices across the country vary incredibly and therefore finding a model that could accurately apply to every state was no plausible. Therefore, I decided to narrow the scope to include only a few states near the Canadian border, so that the model may be easily comparable to Canadian used car data as well. These states included New York, Vermont, Michigan and New Hampshire. Additionally, many similar models use the region variable as a high predicter of price. However, my goal was to identify features of the car that most significantly impact the price prediction, so this variable was omitted from all calculations.

3 Proposed Solution

The experiment began by first cleaning up the data file. I dropped any null rows and manipulated the posting date variable from a string to the number of days since posting. I then extracted the data we required to make the model, by taking only rows which had the column variable 'state' as New York, Michigan, Vermont or New Hampshire. I also restricted the variable 'year' to be anything above 2009. This restriction was implemented because as cars got quite old (such as 1960s or 1970s), the price began to increase again, which negatively impacted the effectiveness of the model. Next, I chose the variables to drop from features in the model. Some unimportant variables such as VIN, url and description were dropped, as well as region as explained above. Finally, the target variable was set to be price, and the dataset was split using the train test split function. I chose

20% to be held back for testing, while the remaining 80% would be used to train the model.

I decided to use a pipeline set up to create the model. First, the data needed to be preprocessed. Therefore, I created pipelines to process the numeric data and the categorical data separately. The numeric columns were checked for missing values and replaced by the mean of the column. They were then standardized, subtracting the mean and scaling to unit variance. The categorical columns were also checked for missing values and then were OneHotEncoded. This creates binary columns for each categorical label in the column, thus allowing us to use it in the model. In the main pipeline I chose the type of regressor and selected the RandomForestRegressor from scikit-learn.

Grid Search was used to find the ideal parameter tuning for this model. I began by testing with 50-100 n_estimators, however the model was underfitting, as the training score was quite low. Therefore, I increased the range of n_estimators to 100-200, which resulting in a 98% training accuracy and a testing accuracy of 87%.

I then attempted to find the feature importance. Since the categorical features had been OneHotEncoded, I first needed to associate the feature importance of the model with their variable names. Next I used the eli5 function to display each feature with its importance.[1] However, I wanted to determine the full importance of each column variable, not each column variable label. Therefore, I decided to use the permutation importance method. This function determines feature importance by calculating the increase in the method's prediction error after permuting the feature.[2] After finding the feature importances, I simply plotted them in a bar chart using the pyplot function.

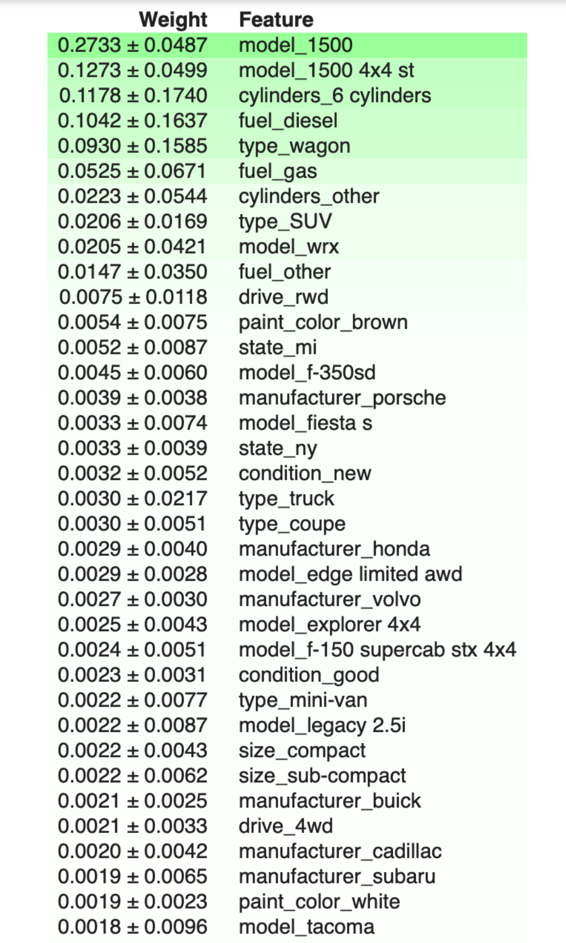
[1] https://towardsdatascience.com/extracting-feature-importances-from-scikit-learn-pipelines-18c79b4ae09a

[2] https://christophm.github.io/interpretable-ml-book/feature-importance.html

4 EXPERIMENT RESULTS

Training Accuracy	Testing Accuracy
98%	87%

Table 1: The training and testing results of the model.



Weight	Feature
0.2733 ± 0.0487	model_1500
0.1273 ± 0.0499	model_1500 4x4 st
0.1178 ± 0.1740	cylinders_6 cylinders
0.1042 ± 0.1637	fuel_diesel
0.0930 ± 0.1585	type_wagon
0.0525 ± 0.0671	fuel_gas
0.0223 ± 0.0544	cylinders_other
0.0206 ± 0.0169	type_SUV
0.0205 ± 0.0421	model_wrx
0.0147 ± 0.0350	fuel_other
0.0075 ± 0.0118	drive_rwd
0.0054 ± 0.0075	paint_color_brown
0.0052 ± 0.0087	state_mi
0.0045 ± 0.0060	model_f-350sd
0.0039 ± 0.0038	manufacturer_porsche
0.0033 ± 0.0074	model_fiesta s
0.0033 ± 0.0039	state_ny
0.0032 ± 0.0052	condition_new
0.0030 ± 0.0217	type_truck
0.0030 ± 0.0051	type_coupe
0.0029 ± 0.0040	manufacturer_honda
0.0029 ± 0.0028	model_edge limited awd
0.0027 ± 0.0030	manufacturer_volvo
0.0025 ± 0.0043	model_explorer 4x4
0.0024 ± 0.0051	model_f-150 supercab stx 4x4
0.0023 ± 0.0031	condition_good
0.0022 ± 0.0077	type_mini-van
0.0022 ± 0.0087	model_legacy 2.5i
0.0022 ± 0.0043	size_compact
0.0022 ± 0.0062	size_sub-compact
0.0021 ± 0.0025	manufacturer_buick
0.0021 ± 0.0033	drive_4wd
0.0020 ± 0.0042	manufacturer_cadillac
0.0019 ± 0.0065	manufacturer_subaru
0.0019 ± 0.0023	paint_color_white
0.0018 ± 0.0096	model_tacoma

Figure 1: The top features of the model ordered by importance. Green represents the most important.

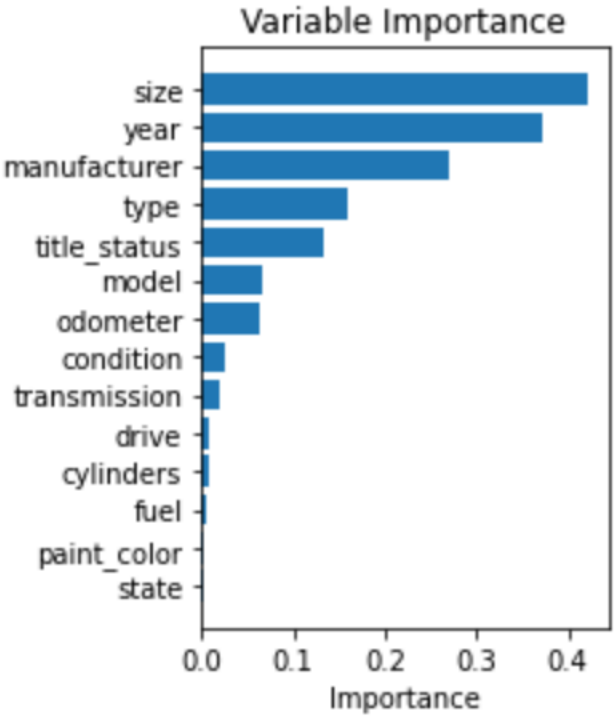


Figure 2: Feature importance of the model using the permutation importance method.

Price	**Prediction**	**Good Price?**
$22888	$18881.27	$4006.73 overpriced
$9990	$10482.87	$492.87 underpriced
$18593	$22919.87	$4326.87 underpriced

Table 2: Example of a use case. Predictions of local cars for sale. From the prediction, one can tell if a vehicle is overpriced, underpriced or fair.

From the results above we can see that both the training accuracy and the testing accuracy were quite good for this model. This was a major improvement from earlier scores where the area had not been specified. These results were approximately 40% training accuracy and 0.1% testing accuracy. After narrowing the scope, the scores improved drastically. Finally, after tuning the parameters, the best scores were arrived upon which are displayed above.

Above we can also see some of the most important factors of the model. Interestingly, the model 1500 was one of the most important factors. However, since there are hundreds of factors with minimal importance, it is more beneficial to inspect figure 2, which displays the permutation importances. From this we can see that size and year are the two most important factors. This makes a great deal of sense. In general, larger cars will be much more expensive than smaller cars. Another interesting revelation is that year is much more significant than odometer. This means that mileage is not as important in determining price as one might have thought. Manufacturer is also of high importance which also makes sense, as different companies have different qualities and therefore prices of cars. Additionally, paint colour has a very low significance in determining the price, which is somewhat expected.

I have also provided a sample use case for the model. In table 2 we can see the model's predictions for a few cars found at autotrader.ca. I have multiplied each prediction by the current exchange rate to compare the prices in Canadian dollars. A buyer can use this number to determine if a certain listing is worth the current asking price.

5 Conclusions

There are many important takeaways from this project. First, it has been demonstrated that resolution plays a major role in the success of a model. If the scope of a dataset is too large, and the data is too different, this severely impacts the model's accuracy. Once a more appropriate area was chosen, the model was very effective in learning the patterns. Clearly, resolution is an important factor that must be considered when investigating a dataset.

Another important take away is the feature importances of the model. It is valuable for sellers to know that odometer(mileage) is not as important as many believe. It is also interesting to know that year plays a major role in determining price, even more so than the manufacturer does. All of these facts are valuable to anyone buying or selling a car who is hoping to better understand the current market and why certain vehicles are priced in such a way.

6 References

1. Christophhm.github.io. *Permutation Feature Importance*. https://christophm.github.io/interpretable-ml-book/feature-importance.html

2. Reese, Austin. *Used Car Data Vehicle Listings from Craigslist*. https://www.kaggle.com/austinreese/craigslist-carstrucks-data

3. Rebecca Vickery. *Extracting Feature Importance from Sci-kit Learn Pipelines*. https://towardsdatascience.com/extracting-feature-importances-from-scikit-learn-pipelines-18c79b4ae09a

4. Sci-kit Learn. *Permutation Importance vs Random Forest Feature Importance*. https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html

5. Autotrader Canada. https://www.autotrader.ca/cars/subaru/outback/2018/?gclid=Cj0KCQjwse-DBhC7ARIsAI8YcWLaENUcKwObiZwNbo4UiF4MzBr4IXBbCzFg_y6i3i5URLOZIW1Lg6gaArG4EALw_wcB