Applying Machine Learning

in Estimate Used Car Listing Price

Huy Tran

Springboard Capstone Project

Mentor: Karel Verhoeven

1. **Introduction**
2. **Initial Problem**

Everyone at one point will probably thinking about selling their old car. The number one question is: how much my car worth? To answer this question, many sites such as Kelly Blue Book, CarGurus, eBay Car Sales, etc. can recommend a price range for the vehicle based on the previous sale of the similar vehicle. Machine learning, in this case, plays a very big role in analyzing and recommending the price.

1. **Business Case**

We should know that every site like Kelly Blue Book or CarGurus usually have two versions, one for the regular users, like us who are sometimes doing a little research about the car price, and one for the dealers, who use it for daily business. Now if the businesses have enough data, they can build their own system not only to estimate price for a used car, but also to track how long the car will be in the lot, which model will be the best, what will be the best price, how much discount should they give to get the car out, etc. to maximize the profit.

1. **Constraints**

There are not enough data in this project to actually fix all the problems above. However, even with small data, the result is promising, and could be expanding easily when there are larger and more diversify data.

1. **Working with Data**
2. **Data source**

The data is originally scrapped from Craigslist, an advertisement website for everything, including used car. This data is gotten from Kaggle.

<https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

1. **Data Wrangling**

The data originally has 458,213 rows and 25 columns.

**Graphical user interface, application

Description automatically generated**

In 25 columns of data, there are a few columns that are not useful for analyzing such as id, url, region\_url, VIN, image\_url, description, lat, long. These will be removed to have a more focusing data, which helps the machine learning a lot easier.

**A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence**

State and region columns have different info due to the way Craigslist is set up, and we only need one of them. In this case, state will probably make more sense (and easier to understand) since we would not know all the regions in the US.

**Text

Description automatically generated**

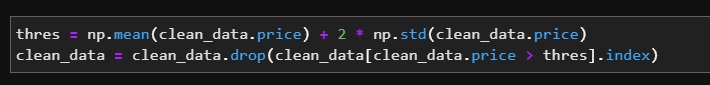
We also will create a ‘age’ column to replace ‘year’ to get a more insightful data, so instead of having a categorical column, we can have a numeric column, which will be is easier for data cleaning and training.

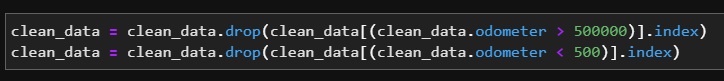
Text

Description automatically generated

During this data wrangling, the data was also cleaned by removing duplicates, removing the outliners of ‘age’ (vehicles that are more than 40 years old), ‘odometer’ (vehicles that have more than 500,000 mileage and less than 500 mileage), and ‘price’ (vehicle that are too expensive or too cheap).

Graphical user interface, text

Description automatically generated



The data was only collected within three months, from 10/2020 to 12/2020, and we do not have any data regarding when the vehicle was sold, so unfortunately, we were not able to use this feature to predict how long the car will be on listing.

Text

Description automatically generated

1. **Exploratory Data Analysis**

EDA provides us a clearer understanding of how the data is. It is very self-explanatory when we go through the plots.

4-year-old vehicles are listed the most. This might be because the regular lease term is three years, so either owner or dealer will try to sell the car when the lease end.

Chart, histogram

Description automatically generated

California has highest number of listing, far above the following such as Florida, Texas, New York. We could explain this as because California is a capitol of high tech, which they tend to make use of online tool such as Craigslist.

Chart, bar chart

Description automatically generated

Ford is the most popular car on Craigslist, follow by Chevrolet, surprisingly. Toyota and Honda are only the third and fourth. Ferrari has highest average price, Tesla, Porsche, Aston-martin are the next ones (and yes they are luxury cars).

Chart

Description automatically generated

Lean title (finance, lease, etc.) has the highest average price while only have very small percentage on the listing. Clean title is listed the most. Parts only vehicle has the least value, of course. Surprisingly, Salvage title vehicles' average price is almost $10000.

Chart, bar chart

Description automatically generated

It is easy to understand that new car will cost more. Interesting thing here is excellent car is priced less than good and like new car, and salvage car cost more than fair condition!

Chart, bar chart, pie chart

Description automatically generated

In general, the more cylinders, the faster the car, and of course the more gas it uses. 4, 6, and 8 cylinders are listed the most, while 12 cylinders car has highest price.

Chart, bar chart, pie chart

Description automatically generated

Gas vehicle dominates the listing as expected. Somehow diesel car has highest average listing price while hybrid car is lowest. One explanation is that cars that use diesel are usually big like van or truck, so the average price is higher. Hybrid car, on the other hand, lose value because of the battery that will need to be changed after about 10 years.

Chart, bar chart

Description automatically generated

Automatic cars are listed the most, not surprising, but I thought manual car usually cost more than automatic? Other types of transmission are CVT and Semi-automatic.

Chart, bar chart

Description automatically generated

No surprise here. 4WD and RWD cars are usually more expensive than FWD. 4WD and FWD is more popular than RWD so they will have more listing.

Chart, bar chart, pie chart

Description automatically generated

Sedan and SUV are listed the most. Somehow minivan has lowest average price. Pickup and truck have highest average price.

Chart, Excel

Description automatically generated

White cars are the most popular and cost the most, following by black and orange.

Chart, bar chart

Description automatically generated

1. **Data Processing**

As we already know, the price will be the target feature for prediction. However, we can see here that it has very long right tail. Therefore, in order to have a better prediction, we will transform this column to log\_price to have a more normal distribution.

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

The model column has many rows with special characters. The easiest way to clean this column is just remove all those rows.

Text

Description automatically generated

Finally, we will check all the missing data and fill them with something.

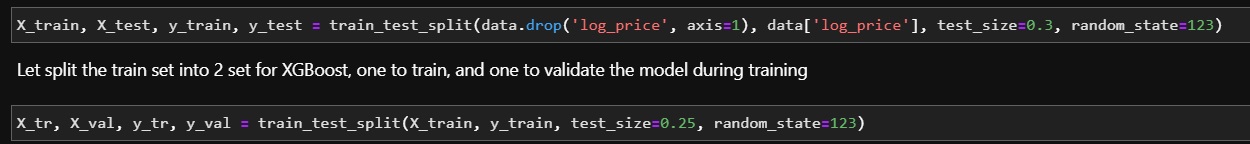
A picture containing text, monitor, screen

Description automatically generated

For numeric columns such as ‘age’ and ‘odometer’, we will use the mean to fill in all the missing data. For categorical columns, we will fill the missing rows with ‘other’.

1. **Modeling**
2. **XGBoost**

As always in machine learning, we split the data to train and test set. For XGBoost, we will split the train set to two set, one to train and one to validate during training.



We then use RandomizeSearchCV to find the best parameters for the model.

Text

Description automatically generated

Finally, after the model was trained, the test set is used to check the performance.

Graphical user interface, application

Description automatically generated

Below is feature importance by XGBoost.

Chart

Description automatically generated with low confidence

1. **Linear Regression**

Using the same train/test set, a pipeline was created to scale the data and get the best k for cross-validation, then fit the data to the test set and check the performance.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

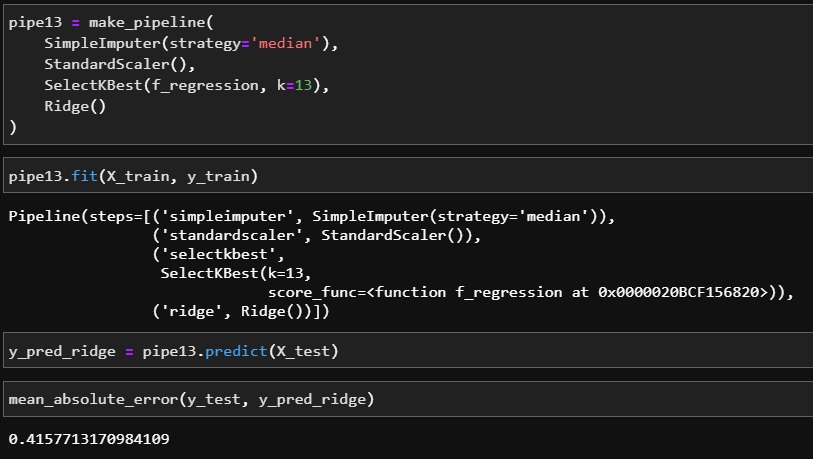
Correlation coefficients

Text

Description automatically generated

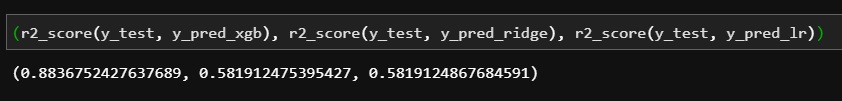
1. **Ridge Regression**

Ridge Regression model is created similarly to Linear Regression. Starting with the pipeline, using cross-validation to get the best model, then fit it to the test set and check performance.



1. **Conclusion**
2. **Model Performance**

As seen above, XGBoost is the model with the least mean absolute error. It is also the model with highest R2 score.



A sample here for the first 1000 rows, XGBoost shows a lot better prediction than Ridge Regression.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

XGBoost also has a lot better residual plot.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

1. **Conclusion**

Car model and odometer are the most important factors that affect the listing price. One surprised is that title status is the least important, even though I thought this should be the most important. I think this is due to the data we have is for listing price, not sold price, so title status probably reduces a lot of view which eventually drive down the price at the end.

In one test data, which is my car that I was trying to sell, a 2013 Toyota Prius C 3, rebuilt title, good condition with about 30000 mileage. The recommended listing price was over $12,000. KBB gives me the price range about $6000 - $11000. This is close, but again, this is only the listing price, and the model did not have that much data as KBB has.

We also need to acknowledge the data and hardware constraint. The model could be more accurate if we use certain coding technique which requires a better computer.