ORIE 4741 Project Used Car Pricing Midterm Report

Yue Han, Zixin Huang, Yuqi Wang

1. Avoid over-fitting and under-fitting

To avoid over-fitting and under-fitting, we plan to apply appropriate linear and non-linear models, choose a suitable number of features and evaluate by comparing training and testing errors. Considering that our data set is very large, we feel confident that we will be able to overcome this challenge.

2. Visualization

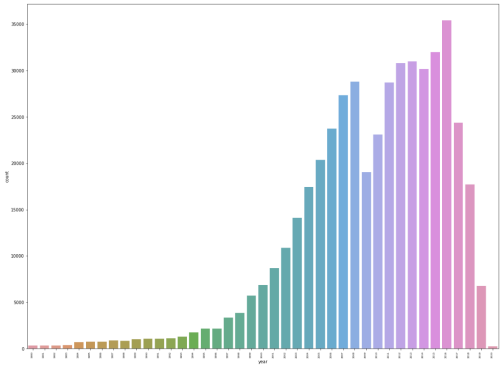

Figure 1. Count group by state

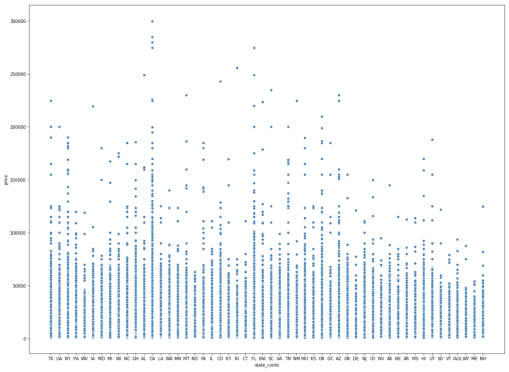

Figure 2. Price of cars group by state

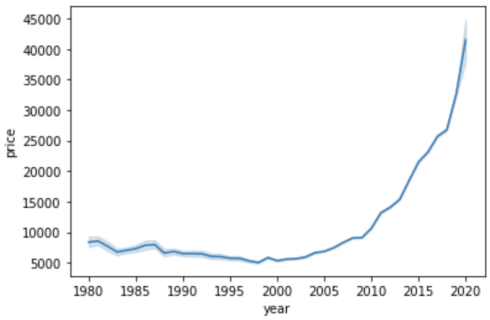

Figure 3. Year vs Price

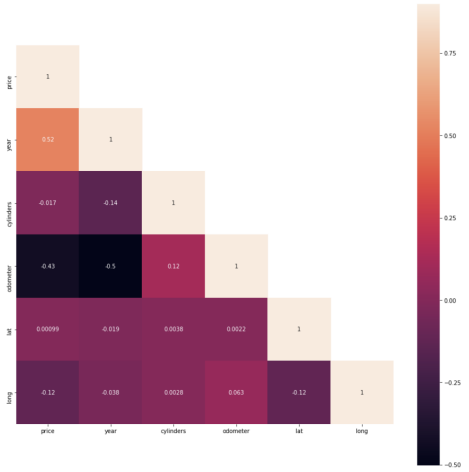

Figure 4. Correlation between features

3. Features and examples

3.1. statistics

Total number of features: 22

Total number of example: 550313

Feature	Type	Missing	% Missing	mean	std	min	max
url	object	0	0%	-	-	-	-
city	object	0	0%	-	-	-	-
city_url	object	0	0%	-	-	-	-
price	int64	0	0%	9.40e+04	1.24e+07	0.00e+00	3.79e+09
year	float64	1487	0.27%	2008	9.61	1900	2020
manufacturer	object	26915	4.89%	-	-	-	-
make	object	9677	1.76%	-	-	-	-
condition	object	250074	45.44%	-	-	-	-
cylinders	object	218997	39.79%	-	-	-	-
fuel	object	4741	0.86%	-	-	-	-
odometer	float64	110800	20.13%	1.04e+05	2.13e+05	0.00e+00	1.02e+08
title_status	object	4024	0.73%	-	-	-	-
transmission	object	4055	0.74%	-	-	-	-
VIN	object	239238	43.47%	-	-	-	-
drive	object	165838	30.14%	-	-	-	-
size	object	366256	66.55%	-	-	-	-
type	object	159179	28.93%	-	-	-	-
paint_color	object	180021	32.71%	-	-	-	-
image_url	object	26	0.005%	-	-	-	-
desc	object	30	0.005%	-	-	-	-
lat	float64	11790	2.14%	38.45	5.82	-81.88	84.15
long	float64	11790	2.14%	-93.85	17.61	-178.15	122.37

descriptive statistics

3.2. Missing and corrupted data

- **useless features**: The features url, VIN, image_url and desc are unique to each car. The city_url does not contribute to this project. Also, we assume that paint_color is not a significant factor for pricing. Thus, we also dropped these features.
- **NaN value**: We would like to drop the features that has more than 50% of NaN value. Size has the most NaN value, so we dropped it.
- **cylinders**: The cylinder feature is of object type. We transformed it to integer.
- **price**: By checking the price data, we observed some outliers such as 123,456,789. Based on common sense, we decided to drop the cars with price greater than 300,000.
- **odometer**: Similar to price, odometer also has some outliers. We dropped the examples with odometer greater than 1,000,000.
- **year**: We observed that most cars are produced after 1980. So we dropped the data of cars produced before 1980.
- **state**: We believe it will be sufficient to use the state as the indicator of location, so we extract a new column named "state" from the "city" column of the data set, which stands for the state which the city is located in. With examination we identify 180,000 missing values, which counts for almost 40% of the data. To deal with this, we notice the state

can be inferred with the latitude and longitude of the location. We use the USA state latitude and longitude data set, and for each sample with missing state information, we find the state whose latitude and longitude are closest to those of the sample and use it as the inferred state. As a result, the number of missing values is reduced to 2,520, and we discard these samples.

3.3. Feature Expansion: Household Income Data

We believe the differences between the price levels of different states play an important role in car pricing, and household income is a major driver of the differences, so we decide to incorporate this data into our data set. Specifically, we use the latest median household income data for each state released by United States Census Bureau, and we map that to each sample with the state information processed before.

3.4. Encoding

There is a number of categorical data in our data set, so we need to encode them before further analysis. We apply one-hot encoding to columns "condition", "cylinders", "title_status", "type" and "drive", and for columns "transmission" and "fuel" which primarily contain only two levels, we use a single integer to encode them.

4. Preliminary analysis

To begin with the analysis, we try to fit the data set with a linear model. Specifically, we use Ridge Regression and apply grid search for the optimal regularization parameter. We first select features that we believe are most powerful, and then scale these features to guarantee the effectiveness of the model. We measure the prediction error with mean relative error, which is defined as the following:

$$mean\ relative\ error = \frac{1}{n}\sum_{i=1}^{n}(\frac{|price_i^{prediction} - price_i|}{price}).$$

We split the samples into training and testing sets, build model for different regularization parameters and choose the best parameter base on the mean relative error of the test set.

As a conclusion, the optimal regularization parameter is 39,200. The training error (mean relative error) is 0.4895 and the testing error is 0.4945. Given that both training and testing error are high, we can see the Ridge Regression model under-fits and more complex models are needed.

5. Future work

- Build non-linear models to address the under-fitting problem, such as decision trees.
- Build sub-models for different categories of cars, such as grouping by the way of transmission or by the fuel used. Sub-models will better capture the differences between each category and make more accurate predictions.
- Choose the best models and combine them into one ultimate model with the stacking or blending method (which can further improve predictive power).
- Estimate out-of-sample error by testing the ultimate model on the out-of-sample data set.