Problem Set 1

Ran Tao

Master of Science in Applied Data Science,

University of Southern California, Los Angeles, California 90089, USA

(Dated: February 4, 2021)

Abstract

In this problem set, I evaluated three models and found, in this case, linear regression is the most suitable model. In addition, I found there are two main factors that affect the value of a car, namely, year and odometer of a car. When it comes to the impact of some special modifications, I found F1 and F2 do influence the price whereas F3 and F4 have a very little impact.

I. INTRODUCTION

Firstly, my CEO asked me to look at the data set and find the main factors that affect the value of a car. Additionally, he/she also asked me to assess the impact of some special modifications (denoted as F1, F2, F3 and F4 in my data set) on the price.

Secondly, my Technical Manager asked me to propose a predictive model and evaluate how accurate the model is.

Lastly, a senior developer wanted to deploy my model to production.

And I finished these tasks separately.

II. DATA EXPLORATION

This data set includes 9997 cars' information, including price, year, manufacturer...It also contains four special modifications. There are six numerical information, see Fig.1, as well as eight categorical information about cars, see Fig.2.

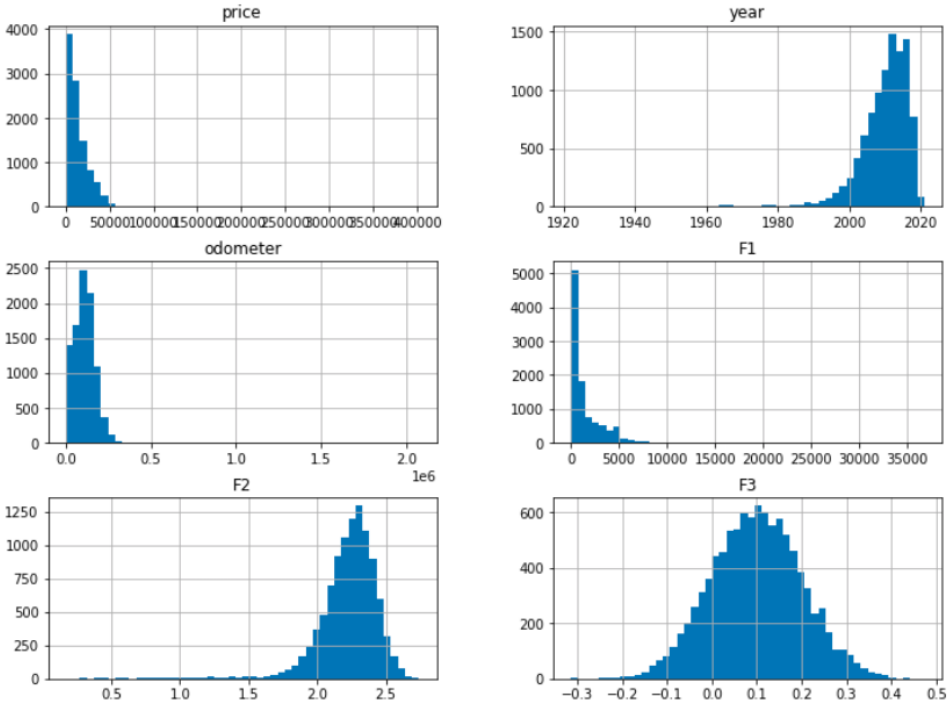

FIG. 1. Distribution of each numerical data

column	value	counts
F4	c	3246
	a	3313
	b	3438
condition	fair	370
	like new	1181
	good	3567
	excellent	4879
cylinders	4 cylinders	3127
	8 cylinders	3233
	6 cylinders	3637
fuel	gas	9997
manufacturer	subaru	989
	ford	9008
paint_color	blue	1372
	red	1580
	silver	1604
	black	2052
	white	3389
transmission	manual	557
	automatic	9440
type	pickup	1737
	truck	2569
	sedan	2626
	SUV	3065

FIG. 2. Value and counts of each categorical data

III. DATA PREPROCESSING

Firstly, I split data into training, validation and test sets. Then I applied a certain way of transformation to these three sets. The steps of this transformation are as follow:

1. I dropped missing values because they are just a very small part of the data set.

2. I selected 'price' as the dependent variable and other variables as independent variables.

3. I did one-hot encoding to categorical variables.

IV. MODEL SELECTION

I used three simple models, namely, Linear regression, Ridge regression, LASSO. Then I computed the mean squared error as a way of evaluating these models. The mean squared errors of these three models are $9134, $9175, and $9278. Therefore, the linear regression is the best model.

V. MODEL EVALUATION

I combined data in my training set and validation set. Then I used these data to fit the linear regression model and applied the fitted model to my data in the testing set. The mean squared error is $8944.

VI. FEATURE IMPORTANCE

I computed the correlations between price and other variables. I found there are two factors that are important for the price, namely, year and odometer. When it comes to the impact of some special modifications, I find F1 and F2 have a great impact on the price. Then I got the coefficient of each variable in the linear regression model, and from this, I can understand specifically how each independent variable affects the dependent variable.

VII. INTERPRETATION

In my linear model, the coefficient of year and odometer is $492 and -$0.0446, which means a car made one year earlier will be $492 more expensive. And when a car's odometer increase 10,000 meters, it price will decrease $446. The coefficient of F2 is $890, which means every unit of this modification will increase the price of a car by $890.

VIII. CONCLUSIONS

In conclusion, Linear regression is the most suitable model for this problem. When it comes to the price of a car, there are several factors at play. Firstly, a car made one year earlier will be $492 more expensive. Secondly, when a car's odometer increase 10,000 meters,

it price will decrease $446. Lastly, every unit of F2 modification will increase the price of a car by $890.

DATA AVAILABILITY

Data is available at https://github.com/usc-dsci552-32415D-spring2021/problem-set-01-rantao-usc/datasets/cars/used_car_dataset.csv

CODE AVAILABILITY

Code is available at https://github.com/usc-dsci552-32415D-spring2021/problem-set-01-rantao-usc/Cars.ipynb