Homework 1 Report

Jae Young Kim

Applied Data Science, University of Southern California,
Los Angeles, California 90089, USA

(Dated: February 11, 2021)

Abstract

As interest in used cars of the people has increased, it has become important to our company to predict the appropriate price of the used car. Through this report, I explored used car data set to find which variable is significant to the price of the used cars and fit a predictive model which predicts the price of the used cars base on proper features obtained through exploration. To fit an easily interpretable model, linear regression model is chosen. Among many regression models, Simple linear regression scored the lowest RMSE score. Simple linear regression scored RMSE score 8605.66.

I. INTRODUCTION

To predict the price of the used cars, I explored used car data set and trained a prediction model. While exploring data, I tried to find interesting trend of the data from the scatter plots and box plots. I used scatter plot to see the distribution of numerical features such as year, odometer, F1, F2 and F3. When it was not easy to find obvious trend, I applied log transformation to the features. To prove it with statistics, I made single linear regression model and interpreted the statistics. For categorical variables such as manufacturer, condition, cylinders, fuel, transmission, type, I drew box plot to check whether there is obvious difference between the categories. Through the exploration, I found there is strong positive relation between year and price and strong negative relation between odometer and price. I could measure the relation with correlation matrix. However, there were some features which seems not that informative to determine the price. In case of F4, the boxplot and result of ANOVA both showed that there is no difference in price when the category of F4 changes.

After exploration, preprocessing step was implemented. I dropped rows with null values and removed some outliers with too high values and applied one-hot encoding to categorical variables such as manufacturer, transmission, type and paint_color. However, I did not apply one-hot encoding to condition and cylinders since it seems to be there is order.

For the model selection, as easily interpretable models were recommended, I tried linear models such as simple liner regression, Ridge Regression, Lasso and Elastic Net. I used simple linear regression model as baseline and tried other methods. To do hyperparameter search, I tried thousands of models with different parameters and calculated RMSE score with validation set. After the hyperparameters were chosen, I calculated the final score after training the models with train set and validation set and tested it with the test set.

II. DATA EXPLORATION

To explore the data set, I calculated correlations between features, plotted box plot and scatter plot. Especially, I focused on the correlation with the target variable, price. There were some numerical features and categorical features. Year, odometer, F1, F2 and F3 are numerical features and Manufacturer, condition, cylinders, fuel, transmission, type,

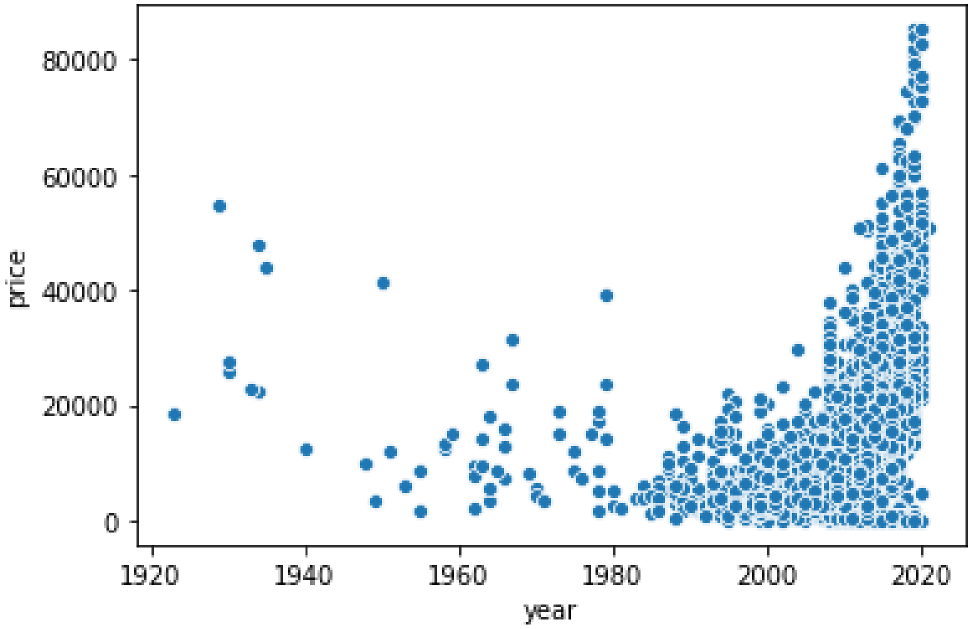
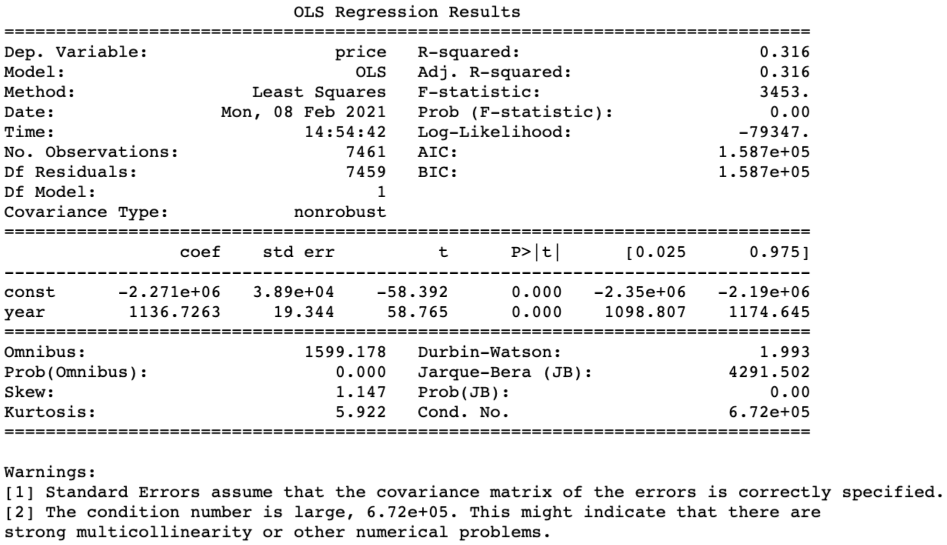


FIG. 1. Scatter Plot of year and price

paint_color and F4 are categorical features.

I could find interesting trend from the scatter plot of year versus price. From the figure 1, most cars in the data set were manufactured after 1980. About 99% of the cars were manufactured after 1980. The cars manufactured before 1980 seems to have relatively random price. However, generally, holding all other factors constant, the price of a car increased 341.14$ as the age of the car increases one year. This appears to be due to the increasing rarity of cars. On the other hand, the cars manufactured after 1980 has shown opposite trend. Generally, the newer the car, the higher the price. Holding all other factors constant, the price of the car decreased 1136.73$ as the age of the car increases one year. The correlation between 'year' variable and price is 0.40 and it is high score compared to other correlations of variables with price. As the correlation of 'year' variable was high, the age of a used car is one of the most important factors determining the price of a used car. Additionally, other numerical feature odometer showed interesting trend. After removing some outliers with too high odometer, it showed quite strong negative linear relation to price. Holding all other factors constant, the price of a car decreased 105.3$ as the odometer of the used car increases 1000miles. It is easy to find the trend when you watch the scatter plot between price and odometer in Fig3. After removing some outliers, the correlation between 'odometer' variable and price is -0.52 and it was the highest absolute correlation with price.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.316
Method:                 Least Squares   F-statistic:                     3453.
Date:                Mon, 08 Feb 2021   Prob (F-statistic):               0.00
Time:                        14:54:42   Log-Likelihood:                -79347.
No. Observations:                7461   AIC:                         1.587e+05
Df Residuals:                    7459   BIC:                         1.587e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2.271e+06   3.89e+04    -58.392      0.000   -2.35e+06   -2.19e+06
year        1136.7263     19.344     58.765      0.000    1098.807    1174.645
==============================================================================
Omnibus:                     1599.178   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4291.502
Skew:                           1.147   Prob(JB):                         0.00
Kurtosis:                       5.922   Cond. No.                     6.72e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.72e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

FIG. 2. Linear regression result of Year vs Price for the cars manufactured after 1980

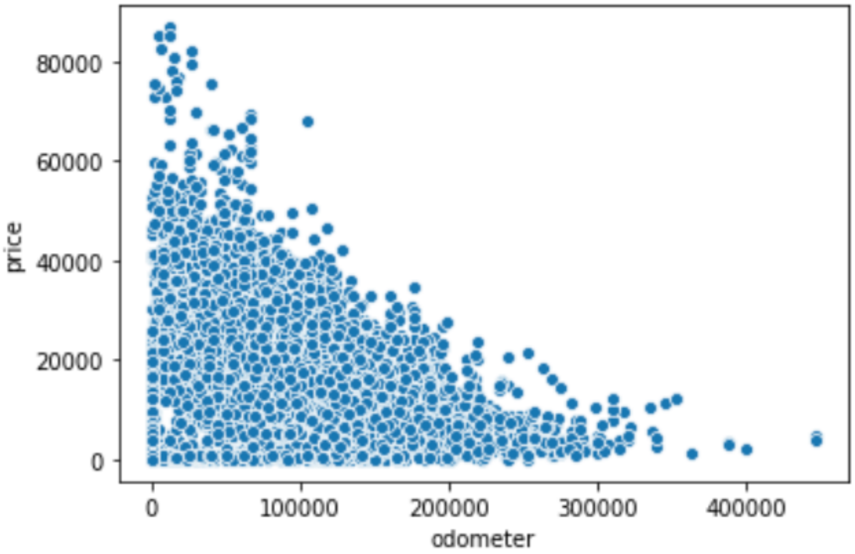


FIG. 3. Scatter Plot of Odometer and price

In case of categorical variables, type of the car and condition of the car seems to be an important feature determining the price. For condition, it seems that there is an order for the groups. ‘Like new’ car scored the highest price and then ‘excellent’, ‘good’ cars. ‘fair’ cars scored the lowest price. Below numbers are the mean values of the price calculated based on each group. This trend can be checked from Fig4

Like new : 19777.62$ Excellent : 15158.72$ Good : 12987.89$ Fair : 3559.74$

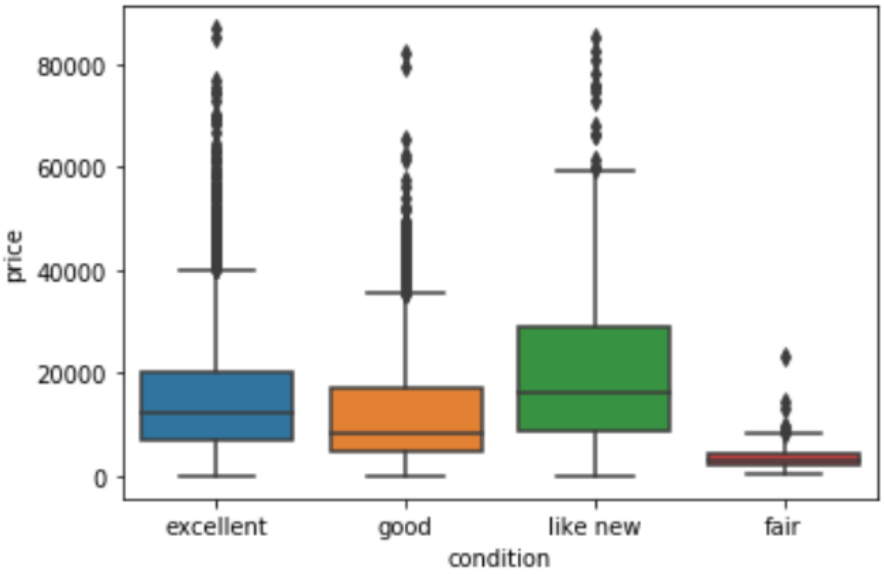


FIG. 4. Box Plot of condition and price

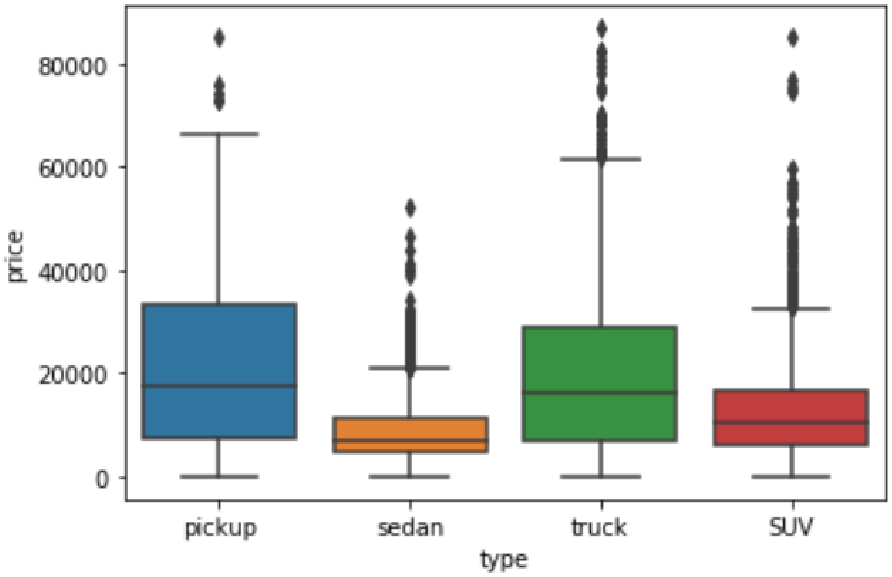


FIG. 5. Box Plot of type and price

Type of the car also seems to have obvious difference between groups. ‘Pickup’ and ‘truck‘ scored high price and ‘sedan’ showed the smallest price. Below numbers are the mean values of the price calculated based on each group. This trend can be checked from Fig5

pickup : 20240.99$ sedan : 8822.15$ truck : 19007.48$ SUV : 12425.29$

To sum up, year, odometer, condition and type seems to be the most important features from the data set.

For the special modifications, as F1, F2, F3 are continuous variable and F4 is categorical variable, I used linear regression for F1,F2,F3 and ANOVA for F4 to access the impact. To access the impact with the same scale, min max scaler was adapted to F1,F2,F3 values. The result of linear regression between price and F1 showed p-value 0. It means that there is a linear relation between F1 and price when other features were fixed. Similarly, p value of F2 is 0 and it means that F2 has a linear relation with price. For F3, it also show p value 0. The coefficients of F1, F2, F3 were 21650, 38890, 10330. Therefore, F2 has the highest impact on price and F3 had the lowest impact on price. For F4, to test the impact of the variable, I used ANOVA. However, the p value was 0.86 and it is too big to say it has impact on price. To summarize, F1 and F2 seems to have some relation with price. F3 seems to have some relation but the impact was relatively low. Finally, F4 has no price difference between the subgroups.

III. DATA PREPROCESSING

Before data preprocessing, I dropped the rows with null values and I split data into three groups, train set, validation set and test set. The ratio was 8:1:1. To sustain data as much as possible, I removed the outliers with way too extreme values such as rows with price over 100,000$ and rows with odometer over 500,000 miles. I tried to fill null values with -1 or mean values. However, as the result was worse, I just dropped the rows with null values.

For categorical features, manufacturer, transmission, paint_color and F4, I used one hot encoding to express categorical variable with numerical values. However, for categorical features such as condition and cylinders I did not use one hot encoding since the order of the groups has meaning. For example, I found there is an order of ‘like new’ > ‘excellent’ > ‘good’ > ‘fair’. Therefore, I assigned 0 for ‘fair’, 1 for ‘good’, 2 for ‘excellent’ and 3 for ‘like new’. For cylinders, I just used the number of cylinder as numerical variable since the number of cylinders had numerical meaning.

For ‘year’, I subtracted 1922, the minimum value of ‘year’, from the values since the scale of ‘year’ variable was too big.

As the values of ‘odometer’ were too high, I created a dummy variable ‘odometer_log’ which is a variable applied log(x+1) function to ‘odometer’. Then I could find there is a

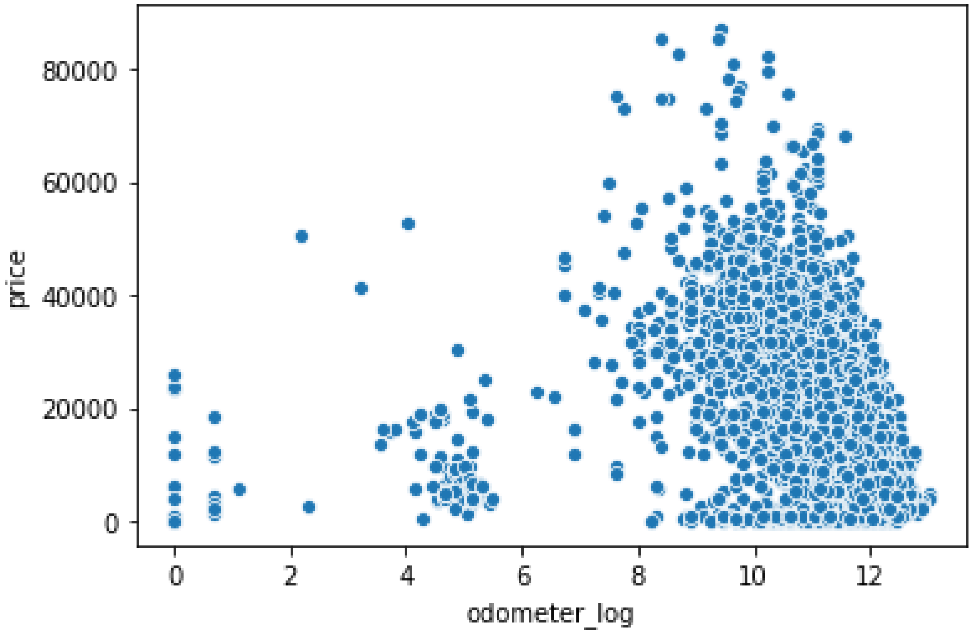


FIG. 6. Scatter plot of log transformed odometer

trend difference in the rows with 'odometer_log ' <7 and rows with 'odometer_log ' ⩾7. Therefore, I created one more dummy variable named 'odometer_log_dummy' which shows whether 'odometer_log' is lower than 7 or not.

As there was only one value for the column 'fuel', I dropped the column 'fuel'.

In addition, I created a correlation matrix of predictor variables and found that variable 'year' and 'F2' are highly correlated to each other. Therefore, I dropped 'F2' variable from the predictor variables. Figure 7 is the heat map of the correlation of the variables. Through this heat map, 'F2_log' showed high correlation with F2 and it is also dropped.

IV. MODEL SELECTION AND EVALUATION

There are many models which can be applied to this data set. However, I tried linear models to make the model simple and easily interpretable. I used simple linear regression, Ridge regression, Lasso and Elastic Net. To measure the performance of the models, root mean square error is used.

Before training data, the data is scaled with min max scaler. It is used because it will let the absolute value of the parameters to be a measure of feature importance.

To get the best hyperparameter of Ridge regression, Lasso and Elastic Net, many models

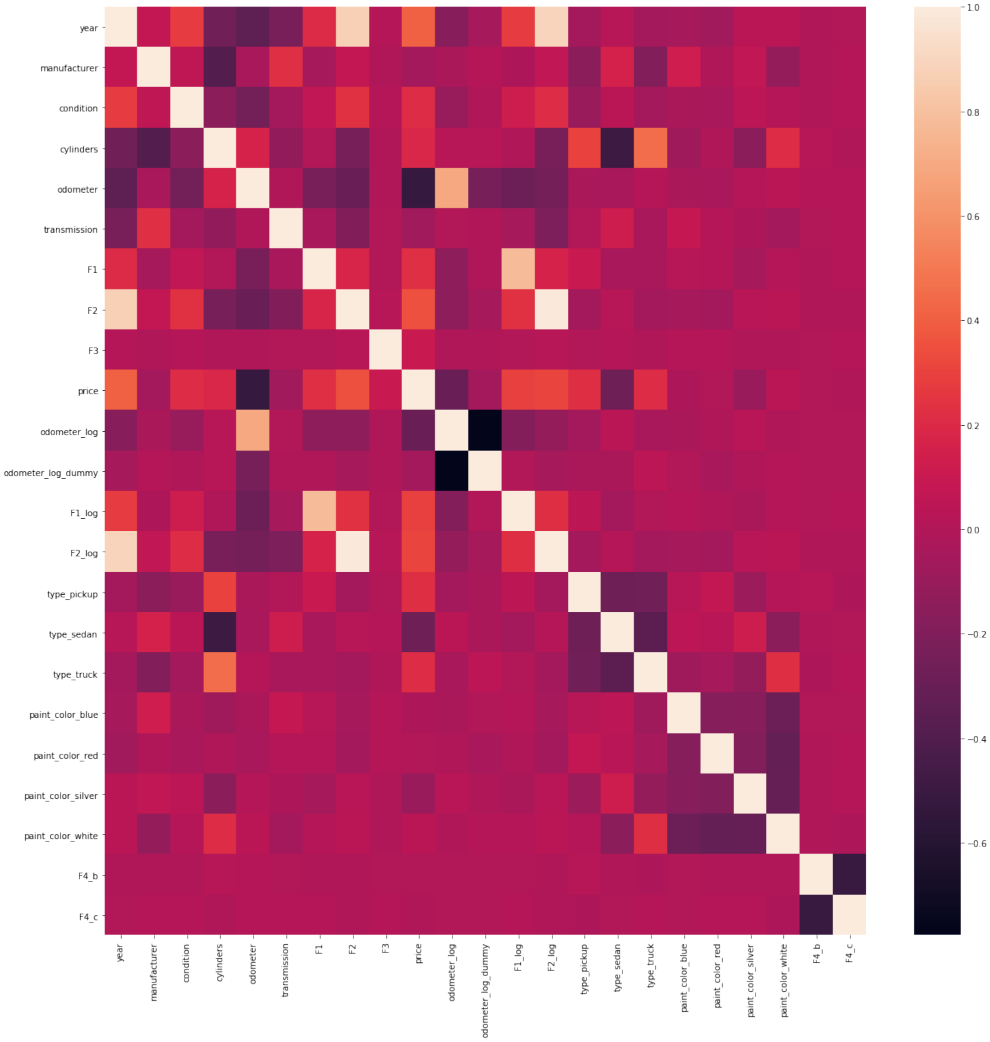


FIG. 7. Heat map of the correlation of the features

with different hyperparameters were tried. 100 models with alpha from 0 to 1 with 0.01 interval for Ridge regression and Lasso and 1000 models with different alpha and l1_ratio were tried for Elastic Net.(0 to 1 with 0.01 interval for alpha and 0 to 1 with 0.1 interval for l1_ratio)

Finally, the hyperparameter chosen for Ridge regression and Lasso were each 0.54 and 0. It means that L2 regularization had no effect on the linear regression model. For Elastic net, best alpha was 0.13 and best l1_ratio was 0.9. Below table shows the result of the models in terms of RMSE score.

Model	RMSE for validation set	RMSE for test set
Linear Regression	9230.38	8605.66
Ridge Regression	8800.70	8699.54
Lasso	9230.38	8605.66
Elastic Net	8953.83	8803.89

The result showed that even though Ridge regression showed the best result on the validation set, the baseline linear regression model showed better result on the test data. Finally, Linear regression model was the best model. Linear regression model achieved the lowest RMSE score, 8605.66.

V. FEATURE IMPORTANCE AND INTERPRETATION

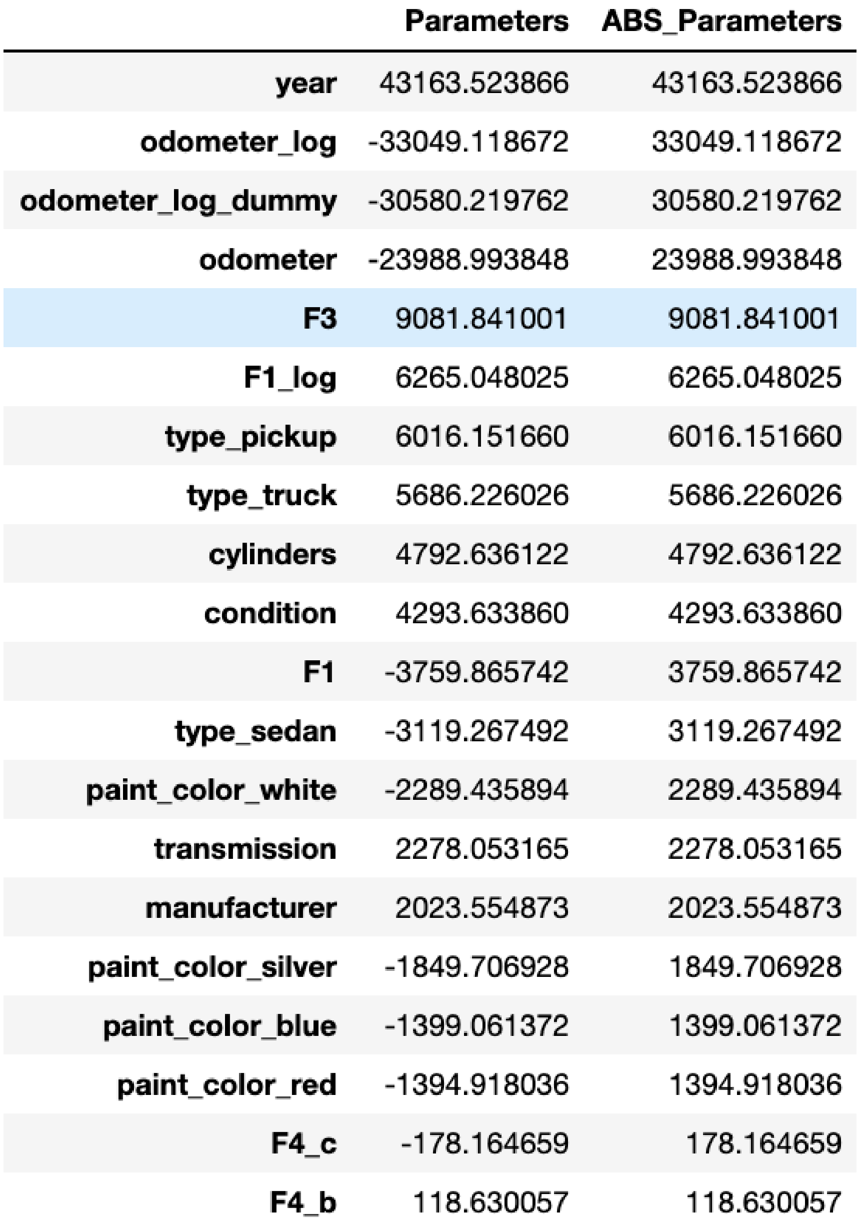
The absolute value of the parameters of the features are a measure of the feature importance since the data is already scaled. The ranked absolute value of parameters are shown in the Fig8.

From figure 8, year showed the highest absolute parameter value and odometer, dummy feature of log transformed odometer(whether log(odometer+1) is larger than 7 or not), log transformed odometer and F3 followed behind. This result can be interpreted as year and odometer has the highest impact on the price.

VI. CONCLUSIONS

To sum up, year and odometer were the features having the biggest impact on determining price of the used cars. There was obvious trend that the new cars are relatively expensive and the price of the cars with high odometer is cheap. Besides, there were some trends such as cars having many cylinders such as pickup and truck are expensive and good condition cars are expensive. For the special modifications, F1,F2,F3 all showed linear relation with price. F2 showed the strongest impact on price and F3 showed the smallest impact on price. However, F2 showed too strong correlation with year variable. F4 showed almost no impact on price.

Simple linear regression is selected as a final model. It scored the lowest RMSE score,



	Parameters	ABS_Parameters
year	43163.523866	43163.523866
odometer_log	-33049.118672	33049.118672
odometer_log_dummy	-30580.219762	30580.219762
odometer	-23988.993848	23988.993848
F3	9081.841001	9081.841001
F1_log	6265.048025	6265.048025
type_pickup	6016.151660	6016.151660
type_truck	5686.226026	5686.226026
cylinders	4792.636122	4792.636122
condition	4293.633860	4293.633860
F1	-3759.865742	3759.865742
type_sedan	-3119.267492	3119.267492
paint_color_white	-2289.435894	2289.435894
transmission	2278.053165	2278.053165
manufacturer	2023.554873	2023.554873
paint_color_silver	-1849.706928	1849.706928
paint_color_blue	-1399.061372	1399.061372
paint_color_red	-1394.918036	1394.918036
F4_c	-178.164659	178.164659
F4_b	118.630057	118.630057

FIG. 8. Parameters of Ridge Regression model

8605.66. Through the feature importance, I could recheck year and odometer are the most influential factor in determining the price of the used cars.

DATA AVAILABILITY

Data is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps1-jeayoung114

CODE AVAILABILITY

Code is available at
https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps1-jeayoung114