Predicting the Price of a Used Car with Linear Modeling

Maxwell Lee

Viterbi School of Engineering, University of Southern California,

Los Angeles, California 90089, USA

(Dated: January 19, 2022)

Abstract

In this report, we will detail the findings of simple models used to predict the price of a used car based on a set of features. While there are concerns regarding the linearity of the data and thus the appropriateness of this approach, we are ultimately able to predict price using a Ridge Regression model with some success. We are also able to determine the importance of the features with regards to price. The most important feature is the mileage of the car. The car will lose about $824 of value for every 10,000 miles it drives. Additionally, the age of a car is an important feature. On average, a car will lose about $4,000 of value per year in the first 5 years of its life, while losing about $450 per year after that. This works about to about a 5% loss in value annually. Finally, we analyzed the importance of changes we can make to the car to improve its value. The condition of the car is also important, as a car that is like new will be worth about $9,000 more than a fair car. The modifications F1 and F3 are both changes that significantly raise the value of the car, F4 might have some impact, and F2 is very unlikely to have any impact.

I. INTRODUCTION

For this task, we have been asked to predict the price of a used car given a set of features. Due to state-law, we are limited to simple linear methods for our prediction. This will have the added benefit of interpretability for our model. Our data set is given as a csv file with 9 features: year, manufacturer, conditions, cylinders in the engine, fuel type, mileage, transmission, car type, and paint color. We would like to understand which, if any of these, play an important role in predicting the price of the car. Additionally, we are given four modifications we could make to the car and would like to understand if they impact the price of the car and thus are worth doing.

We will be using four different methods to predict price: OLS, Ridge Regression, Lasso Regression, and Elastic Net. OLS is the most traditional approach to linear regression, while the remaining three are rooted in the principles of OLS with methods of feature selection built into the fitting of the model. The best of these models will be chosen for further analysis and interpretation.

II. DATA EXPLORATION

To explore the data, it was read into a Jupyter Notebook using a Pandas DataFrame. We then examined the variable type for each of the features. Price, year, mileage, F1, F2, and F3 are treated as continuous variables, while the remaining are all treated as categorical variables. The cylinders of an engine could be treated as a continuous variable, however the choice was made to treat cylinders are a categorical variable as it was not clear that price would have a linear relationship with price. Box plots of each feature were created to detect reasonable ranges of each feature. These decisions will be discussed in greater length in Section III Data Preprocessing.

The first important discovery in data exploration is the skewness of price data. The distribution is shown below in Figure 1 This is to be expected; the majority of cars will fit into a narrow distribution of value, while a small number of luxury cars will skew the distribution upwards. This isn't inherently a problem, but may pose some challenges to linear modeling. An assumption of linear regression is the randomness of residuals, which is difficult to achieve in a skewed distribution. In an attempt to remedy this, a log transform is

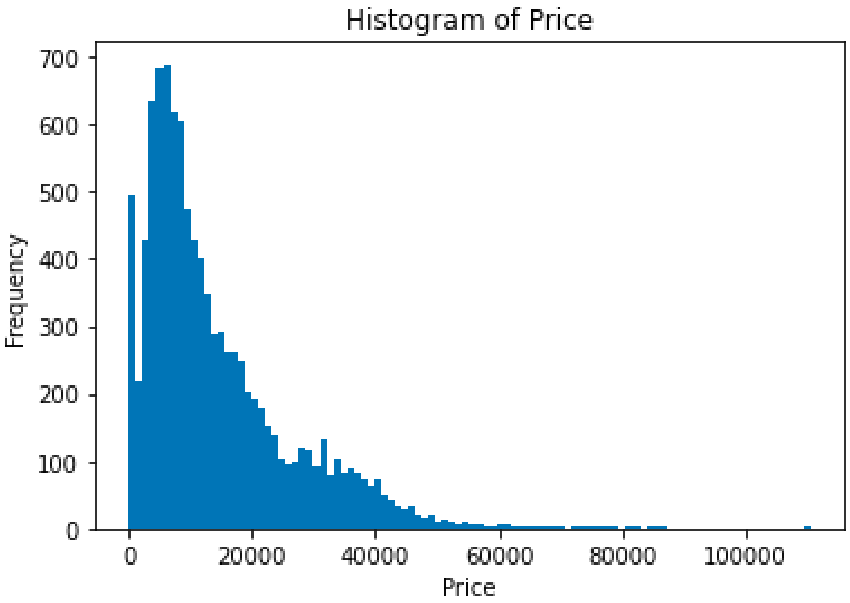


FIG. 1. Histogram of price displays significant skewness in price data

taken of price after an initial linear regression attempt displayed highly correlated residuals.

Additionally, some important limitations to the dataset became evident in the categorical variables. While the continuous variables generally had wide spanning ranges, some categorical features were heavily populated with only a subset or one of the potentially categories. The manufacturers were almost entirely Ford, with a few cars made by Subaru, almost all transmissions were automatic as opposed to manual and there were only gas cars in the dataset. As such, this may not be representative of the entire car market, so the specific model generated may not be best suited beyond the constraints of this dataset. However, the methodology should remain sound for any similar dataset.

III. DATA PREPROCESSING

The first step in Data Preprocessing was to detect outliers in the data. This was done by examining the boxplots, as well as applying a measure of common sense for reasonable values. A $400,000 truck is the only point removed on the basis of high price. There are a decent number of cars with a value of $0. This isn't totally unreasonable, as some cars are so old or so worn that they cannot be resold. However, these points are ultimately removed, as they proved a great challenge to the linear nature of our problem. Different methods of machine learning may be more appropriate if detecting a worthless car is important to the task.

The outlier cutoff for odometer is put at 500,000 miles. This cutoff is tricky. It's extremely rare for a car to have more than 500,000 miles, which is why this is a conservative estimate. The boxplots would indicate any mileage over about 275,000 would be an outlier, however there is still a not insignificant portion of the data with greater than that mileage. Because these values are within the realm of possibility, they are kept. Values outside the range are discarded, however imputation could be done on these values in further work, as they are likely to be input errors. Additionally, some rows were missing values for the mileage, and were also discarded.

The next step was to use One Hot Encoding for categorical variables. This was done using the Pandas *get_dummies*() function. Importantly, we drop one column from each of these encodings. Failure to do so would result in perfect colinearity amongst the category.

Then, feature selection for the OLS model was done using the statsmodels library [1]. VIF was used to remove features with high colinearity. Year and F2 both had very high VIF values, so they were removed from consideration for the OLS model.

The next process was feature scaling. First, we rescaled the odometer to be in the units of 1000 miles, for interpretability. Fitting OLS with all features to price immediately revealed that the skewness of price mentioned in Section II was an issue, as the residuals were strongly correlated. This fit is displayed in Figure 2. To attempt to remedy this, the dependent variable of price was transformed with $log(price)$. The hope is that this will unskew the price data and result in a more linear fit. This did not entirely fix the problem, but does look much more linear in nature. The fit to $log(price)$ is shown in Figure 3. Interestingly, the R^2 value decreased somewhat from .488 to .339. However, based on the plot of the results, the plot with $log(price)$ looks far more linear, and knowing the price data is skewed this approach seems better.

IV. MODEL SELECTION

We will be comparing the results of OLS, Ridge Regression, Lasso Regression, and Elastic Net on our data. Each of these models will be built using the sklearn library [2] in Python and fit to the dependent variable $log(price)$. Portions of this code was inspired by Jason Brownlee [3]. We start by randomly splitting the data into training and testing partitions

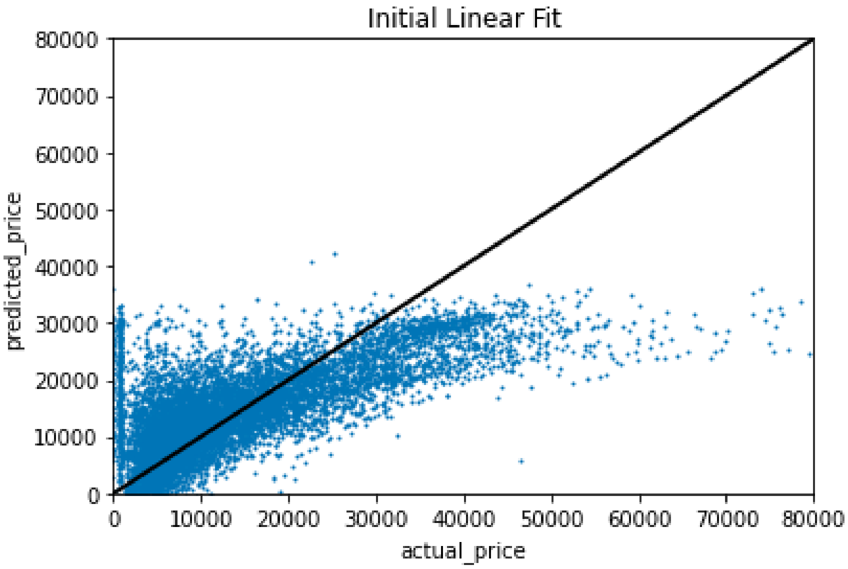


FIG. 2. Initial fit to price reveals a poor fit with clear correlation in the residuals

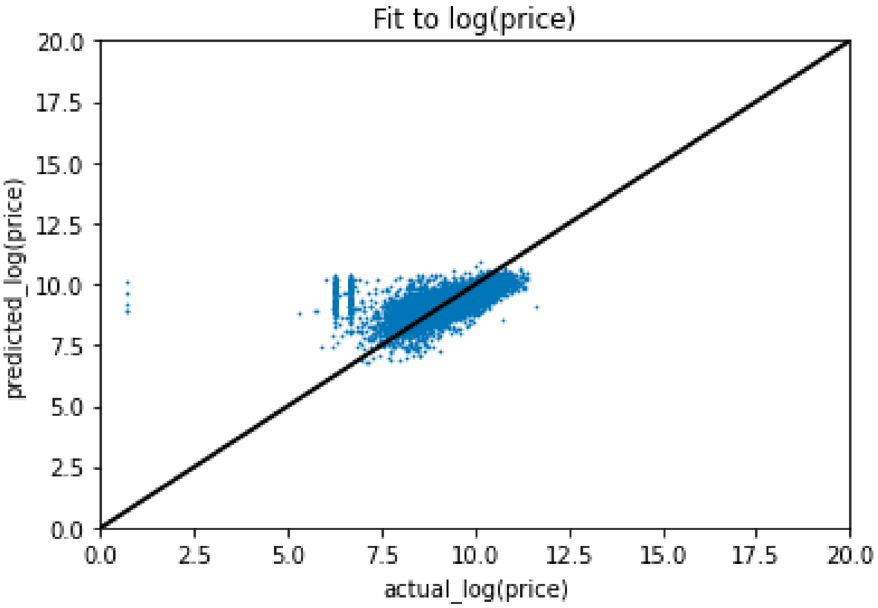


FIG. 3. The model fit to log(price) looks more linear in nature

with 80% training data. With our training data, we then will use a K-fold cross validator with 5 folds, looking to optimize the average of the root mean square error for each of the 5 folds.

From our previous knowledge of the VIF values of year and F2, these features are dropped for the OLS model. However, as the remaining models have feature selection built into the implementation, we will allow the features to be included. For the non-OLS models, we also perform a grid search for the optimal parameters for that model. The average root mean square error across the folds along with the found optimal parameters are provided the Table I below.

Model	Parameters	RMSE
OLS	N/A	0.779
Ridge	α=2.9	0.747
Lasso	α=0	0.747
Elastic Net	α=0 L1=0	0.747

TABLE I. Parameters and performance of each model

Here we can see that Ridge, Lasso, and Elastic Net achieved the same RMSE, all of which are better than OLS. But the optimal parameters of Lasso and Elastic Net are very interesting. The α value of 0 indicates that Lasso was run as a simple OLS model. The better performance than our OLS, which excluded year and F2 due to high VIF, indicates that these features should be included. Indeed, refitting OLS with all features generates the same RMSE, displayed in Table II. This makes the optimal parameter for Ridge puzzling, as it was able to find some level of feature reduction that resulted in the same performance, perhaps by sheer luck. Regardless, it is clear that the more advanced algorithms are not giving us any greater performance. When performance is the same, simple is better, so we will declare OLS with all features the best model.

Model	Parameters	RMSE
OLS (all features)	N/A	0.747
Ridge	α=2.9	0.747
Lasso	α=0	0.747
Elastic Net	α=0 L1=0	0.747

TABLE II. Parameters and performance of each model

V. MODEL EVALUATION

Having declared OLS with all features the best model, we now refit that model with all of the training data using the statsmodel library, as it has superior feature interpretation. We use this model to predict the testing data we set aside. The plots of predicted vs actual as well as a residual plot is shown in Figures 4 and 5. The model does not look dramatically

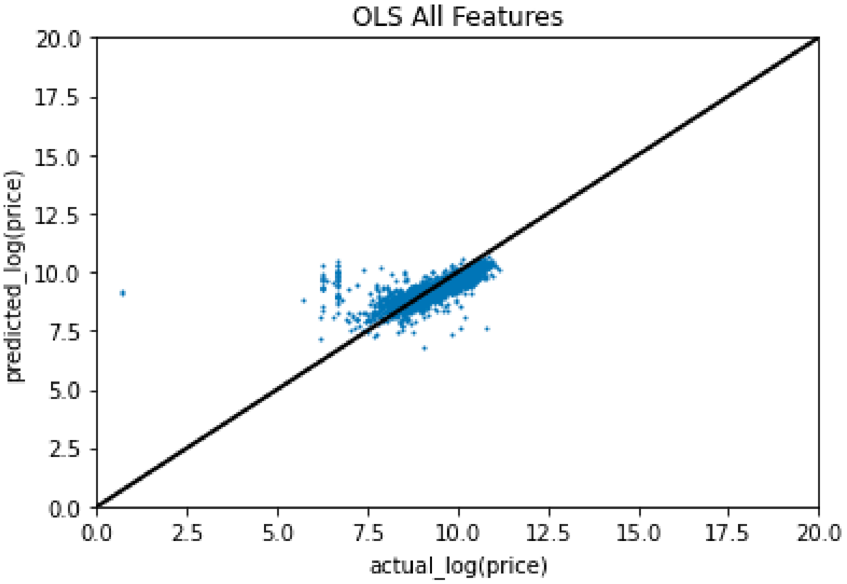


FIG. 4. Final fit to testing data with OLS on all features

different from the model we excluded year and F2 from, but does have a decent R^2 value of .4 and a similar RMSE to the cross validation RMSE at .743.

The residuals are still problematic, as they are clearly negatively correlated. This problem may simply not be well suited to a simple linear approach. However, this model performs well for the vast majority of the data. Any tradeoff to a more complex model will come at the expense of simplicity and interpretability, a cost that may not be worth the gain in performance for a small subset of the overall data.

VI. FEATURE IMPORTANCE

As OLS was selected as the best model, we have the great benefit of easily interpretable feature importance. For the statistically inclined, the results summary provided by the statsmodels is provided in Figure 6. For everyone else, we can distill some of these findings into plain English. Virtually all information we have is important to knowing the price of the car, except for F2 and to a lesser extent if the car is red. The intercepts are harder to interpret as we have log transformed the dependent variable. If we set each coefficient as x and take $(e^x - 1) * 100$, we can get the percentage change in price due to each feature. Table III displays the percentage change as a result some of the most important features.

Our percentage change measure is not useful for understanding the impact of our modi-

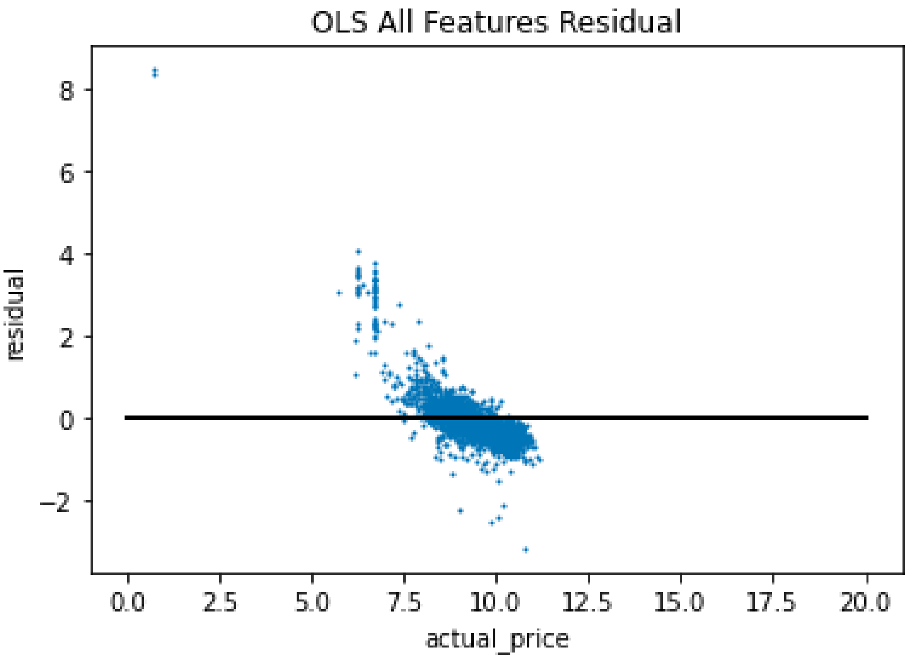
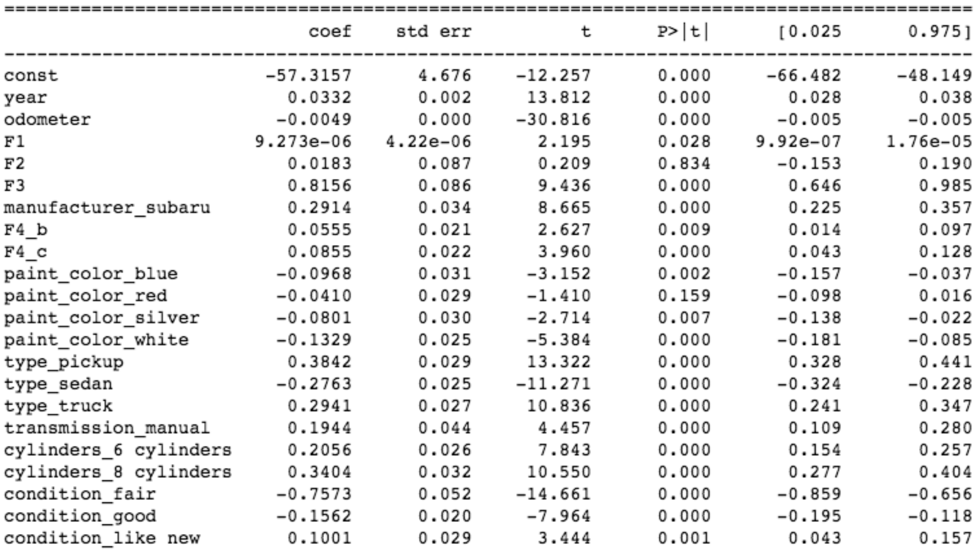


FIG. 5. Residuals for OLS still problematically correlated

fications, so we will simply rely on the p-value. F1 and F3 are continuous variables, which makes their interpretation straightforward. They are positive and have a very small p-value, so we can confidently say F1 and F3 raise the price of the car. F2 has a large p-value, so we can say F2 likely does not impact the price of the car, although not with the confidence with which we made the claim of F1 and F3. F4 is a bit harder to interpret as it is a categorical variable. Our output is in comparison to the category we dropped in one hot encoding, so we can confidently say that and F4 category of B or C is significantly higher than the F4 category of A. C has a higher coefficient than B, but whether that difference is statistically significant is diffiult to determine.

Feature	Unit	Percentage Change in Price
Year	years	3%
Mileage	1000 miles	-.5%
Condition fair	w.r.t. excellent	33%
Truck	w.r.t. SUV	34%

TABLE III. Percentage change in price due to important features



```
=====================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 -57.3157      4.676    -12.257      0.000     -66.482     -48.149
year                    0.0332      0.002     13.812      0.000       0.028       0.038
odometer               -0.0049      0.000    -30.816      0.000      -0.005      -0.005
F1                   9.273e-06   4.22e-06      2.195      0.028    9.92e-07    1.76e-05
F2                      0.0183      0.087      0.209      0.834      -0.153       0.190
F3                      0.8156      0.086      9.436      0.000       0.646       0.985
manufacturer_subaru     0.2914      0.034      8.665      0.000       0.225       0.357
F4_b                    0.0555      0.021      2.627      0.009       0.014       0.097
F4_c                    0.0855      0.022      3.960      0.000       0.043       0.128
paint_color_blue       -0.0968      0.031     -3.152      0.002      -0.157      -0.037
paint_color_red        -0.0410      0.029     -1.410      0.159      -0.098       0.016
paint_color_silver     -0.0801      0.030     -2.714      0.007      -0.138      -0.022
paint_color_white      -0.1329      0.025     -5.384      0.000      -0.181      -0.085
type_pickup             0.3842      0.029     13.322      0.000       0.328       0.441
type_sedan             -0.2763      0.025    -11.271      0.000      -0.324      -0.228
type_truck              0.2941      0.027     10.836      0.000       0.241       0.347
transmission_manual     0.1944      0.044      4.457      0.000       0.109       0.280
cylinders_6 cylinders   0.2056      0.026      7.843      0.000       0.154       0.257
cylinders_8 cylinders   0.3404      0.032     10.550      0.000       0.277       0.404
condition_fair         -0.7573      0.052    -14.661      0.000      -0.859      -0.656
condition_good         -0.1562      0.020     -7.964      0.000      -0.195      -0.118
condition_like new      0.1001      0.029      3.444      0.001       0.043       0.157
```

FIG. 6. The rather dense summary of OLS regression

VII. INTERPRETATION

While a good deal of the interpretation of results is done in Section VI, it is important to spend some time discussing the limitations of our findings. First, the assumptions we made regarding outliers may hurt our model's performance given a new data set. The model is particularly sensitive to the exclusion of cars deemed to have no value. We removed these cars as our model was terrible at fitting to them, so it was better to remove the points and allow the model to fit more directly to the more relevant points. But cars with no value may be important to other tasks, and so this should be kept in mind for the future.

As mentioned many times, the underlying assumption in OLS is the residuals are not correlated. This assumption is clearly violated and our attempts to remedy this do make some progress, but not nearly enough. This ultimately isn't a huge problem for the pure performance of the model, as we can point to relatively successful R^2 and RMSE values for proof. However, this will hurt any statistical tests done to interpret the data. Additionally, we detected high colinearity among some features, but did not remove features as doing so hurt the RMSE performance. This too will have an impact on interpreting the statistical tests done on the intercepts.

Ultimately, this approach is fine as a simple way to understand some major factors that drive car price. Our results align with many of our prior assumptions, such that age of the car and mileage play an important factor into the price of the car. But, drilling further down and extracting more specific meaning should be done with caution.

VIII. CONCLUSIONS

With the given data, we are able to create a simple linear model that is successful at predicting the price of cars. However, there are some underpinning statistical assumptions that are broken by the data, which would require further work to remedy, if even possible. Overall, the results align with our prior assumptions that the age of the car, mileage, and other features that common sense would say impact the price of the car. Importantly for us, we are able to detect which modifications we can make that will impact the price of the car. The modifications F1 and F3 are the two modifications we can most confidently say impact the price of the car.

[1] S. Seabold and J. Perktold, statsmodels: Econometric and statistical modeling with python (2010).

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research **12**, 2825 (2011).

[3] J. Brownlee, How to develop ridge regression models in python (2020).