Factors that affect prices of used cars using Machine Learning techniques

Saurabh Jain

Department of Computer Science, University of Southern California,

Los Angeles, California 90089, USA

(Dated: February 11, 2021)

Abstract

Machine Learning can be explained as a method that interpret results from exploring a data set, proposing and fitting a predictive model and evaluating the model's performance. This assignment focuses on the data set that describes the condition of various cars and their current prices. Cars experience depreciation over the years by the wear and tear from its use and various other factors. The machine learning model prepared in this assignment proves the same using a data set and creates a 56.36% accuracy from a linear regression model. Each year, the price of a car decreases by $1036, keeping the other factors in model constant. One of the interesting points to be considered is that the newer the car, the better priced it is. The correlation between "year" and "F2" (special modification) is close to 1, which implies that "F2" decreases as the age of the car increases. Linear regression, LASSO, Ridge Regression, Elastic Net methods have been used for the purpose of this assignment. Linear Regression has been the preferred model based on the model score but there has not been a huge difference in the score. The maximum number of cars are from the year 2013 and the maximum amount of cars are priced in the range of $4000-$8000 in the uncleaned data set.

I. INTRODUCTION

The primary task of this assignment was to create a Machine Learning algorithm to predict the price of a used car. A few concepts learned in class such has Linear Regression, Ridge Regression, LASSO were used to compute the score of the model, thus finding the most accurate model.[1] The principal variables included - price, "year", odometer, manufacturer, condition, cylinders, fuel, transmission, type, paint_color, special modification variables such as "F1", "F2", "F3", "F4" with price being the outcome variable for the models.

The primary steps included data cleaning, data processing, removing outliers, calculating correlations between variables, generating regression models, calculating predictions, and finally, choosing the most efficient model. Data was then visualized to effectively assess the data and the results. Root Mean Squared Error, Mean Squared Error, and Mean Absolute Error was calculated for the most effective model. One of the major steps involved performing linear regression models with price as the outcome variable and one of the other variables, keeping every other variable constant.

II. DATA EXPLORATION

The data set consisted of 14 major variables with price being the outcome variable. pandas and numpy libraries were used to read the data and perform one-hot encoding to differentiate the categorical variables into numerical ones. Out of the fourteen variables, eight were categorical variables. The following list consists of the values of different categorical variables:

1. manufacturer: ford, subaru
2. condition: excellent, fair, good, like new
3. cylinders: 4 cylinders, 6 cylinders, 8 cylinders
4. fuel: gas
5. transmission: automatic, manual
6. type: SUV, pickup, sedan, truck

[1] Libraries such as sklearn, scikit, statsmodels, etc. were used which will be referenced later in the report

FIG. 1. Heat-map of correlation of non-categorical variables.

7. paint_color: black, blue, red, silver,white

8. "F4": a, b, c

III. DATA PREPROCESSING

One hot encoding is an excellent tool to convert categorical variables to numerical variables. Therefore, one of pandas libraries' methods, pd.get_dummies was used to convert. Correlation was then computed of between the non-categorical variables and a heat-map was generated for the same as displayed in Fig. 1. Some of the interesting points noticed were the correlation between "F2" and "year" being 0.9, which is close to 1. Price, the outcome variable had the highest correlation with the manufacturing year of the car, with next highest being with "F2". VIF is Variance Inflation Factor and when some features are highly correlated, VIF can be used as a technique to detect multicollinearity. Therefore VIF was calculated to detect multi-collinearity.

$$VIF = \frac{1}{1 - R^2}$$

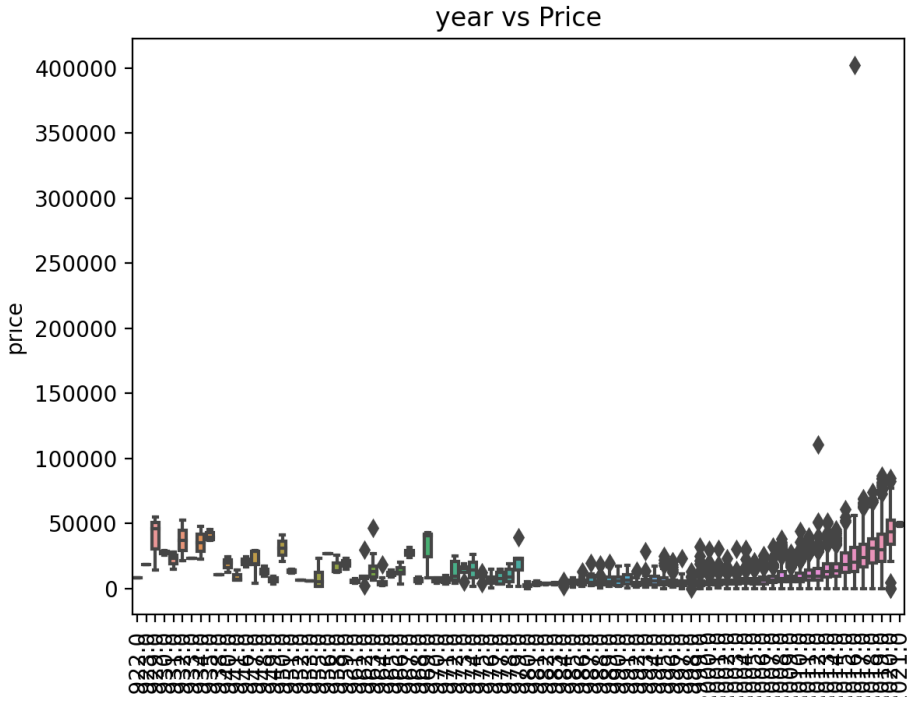

FIG. 2. Box-Plot before treating the outliers.

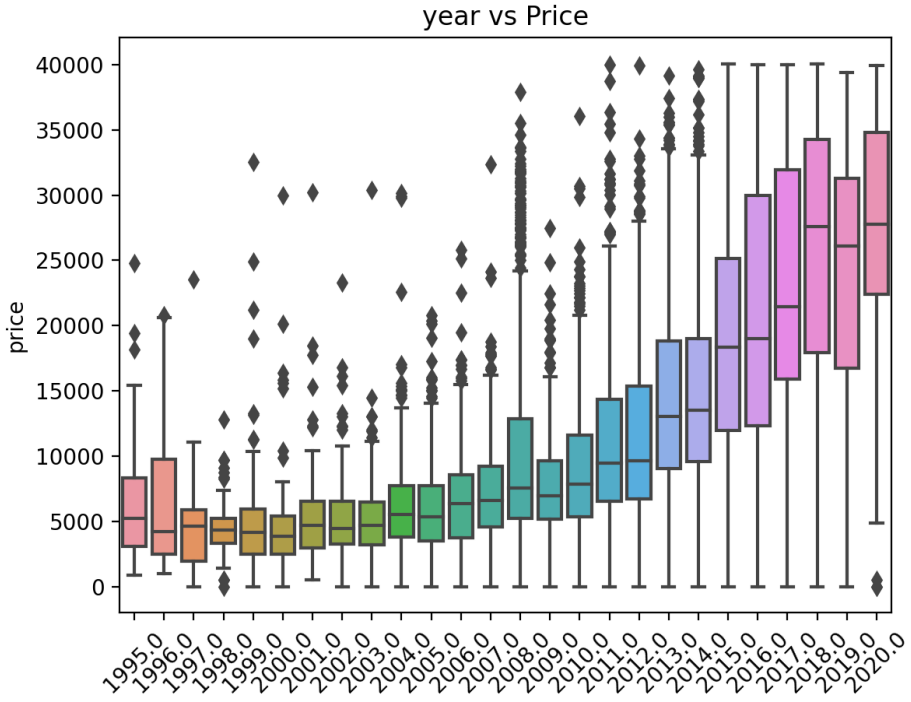

FIG. 3. Box-Plot after treating the outliers.

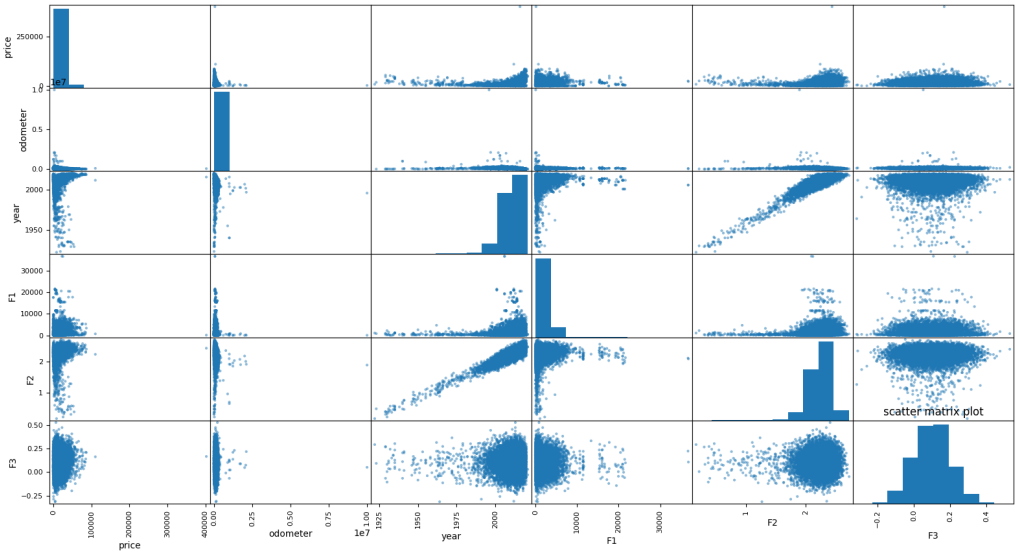

FIG. 4. Scatter-matrix plot of variables before processing and cleaning.

VIF for "year" and "F2" was calculated to be significantly higher than the others. Therefore, box plot was generated for year versus price. Fig. 2 clearly shows that there existed outliers in the data with respect to year. These outliers were removed using the interquartile rule. Fig. 3 shows the modified box plot without the outliers. "odometer" column consisted of no value tabs which were replaced by the mean of that column's values so that the accuracy of the model is maintained.

A scatter matrix plot was generated to visualize the data better and to confirm that it is best to remove outliers from "price" (being one of the outcome variables), "F2" and "year".

IV. MODEL SELECTION

Several methods of models were used namely - Linear Regression using statsmodels library, Linear Regression using sklearn library, LASSO using sklearn library, LASSO Cross-validation using sklearn library, Ridge Regression using sklearn library, Ridge Regression with Cross-Validation using sklearn library, Elastic Net. In each case, the data was split into training set and test set with test_size being 30% of the entire data set. X_train, X_test, y_train, y_test are the common notations used in the code and in the paper.

- **Linear Regression using statsmodels library:** After the data was cleaned and

the outliers were removed using the inter-quartile rile, OLS methods were used. The training data was used to train the model which led to the summary being printed. The test data was used to predict the results, creating an array. The summary of the results were:

$$R^2 = 0.585$$

$$adjustedR^2 = 0.584$$

.

- **Linear Regression using sklearn library:** sklearn library's Linear Regression method was used to compute the score directly of the test data after data-cleaning and removing of outliers. The score of sklearn library's linear regression model was calculated to be 0.567
- **LASSO using sklearn library:** sklearn.linear_model library's Lasso method was used to compute the score directly of the test data after data-cleaning and removing of outliers. The score of this model was calculated to be 0.567
- **LASSO Cross-validation using sklearn library:** sklearn.linear_model library's LassoCV method was used to compute the score directly of the test data after data-cleaning and removing of outliers. The score of this model was calculated to be 0.015
- **Ridge Regression using sklearn library:** sklearn.linear_model library's Ridge method was used to compute the score directly of the test data after data-cleaning and removing of outliers. The score of this model was calculated to be 0.567
- **Ridge Regression with Cross-Validation using sklearn library:** sklearn.linear_model library's RidgeCV method was used to compute the score directly of the test data after data-cleaning and removing of outliers.
- **Elastic net:** The score for elastic net was computed to be 0.568

V. MODEL EVALUATION

Seven different models were used, mainly being Linear regression, Ridge regression and LASSO. Linear regression had the best accuracy. Each model was a result of either sklearn

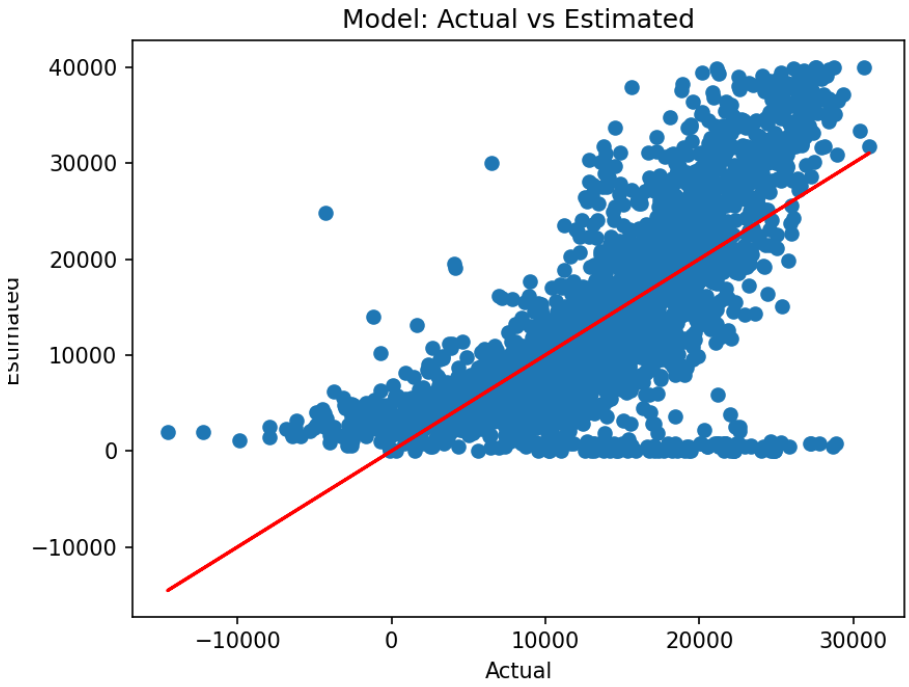

FIG. 5. statsmodels actual vs. predicted plot.

or statsmodels library. These libraries have either model scores or summary which includes values of R^2, therefore providing us with the accuracy of the models.

The linear regression (statsmodels) was plotted as shown in Fig. 5 using predicted y values and the y values of the test set. This can be comprehended as the actual and the predicted values of the test set.

VI. FEATURE IMPORTANCE

Feature Selection technique was used to find out the best 5 variables that will make the model as accurate as possible. SelectKBest and chi2 methods from sklearn.feature_selection library was used to choose 5 variables that are more important. The following were the VIF values computed to detect which variables were used to remove outliers. VIF is a measure of multi-collinearity therefore, price, year and F2 seemed to be choice of variables that had their outliers removed.

1. price: 2.693445

2. year: 115.107018

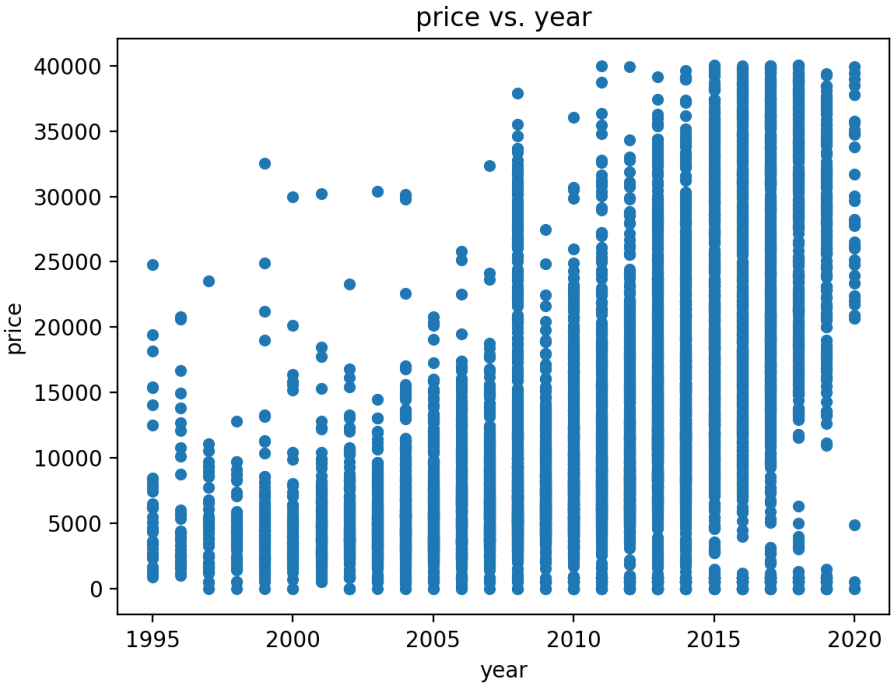

FIG. 6. price vs year

3. odometer: 1.953788

4. F1: 1.605894

5. F2: 117.652081

6. F3: 2.009886

A set of linear regression models were made to compute the single variable effects on price. The best score was computed from price and year with the score being 0.36. Therefore, it can be said that price is hugely dependent on the age of the car.

VII. INTERPRETATION

The major finding from the model is that most of the models have a similar score for this particular data set. The Root Mean Squared Error for each of the model is in the range 6000-6500. It can also be interpreted from Figure 6 that the newer the car, the more expensive it is. Figure 7 shows us that "year" and "F2" are closely related and as the car gets older, the value of F2 decreases.

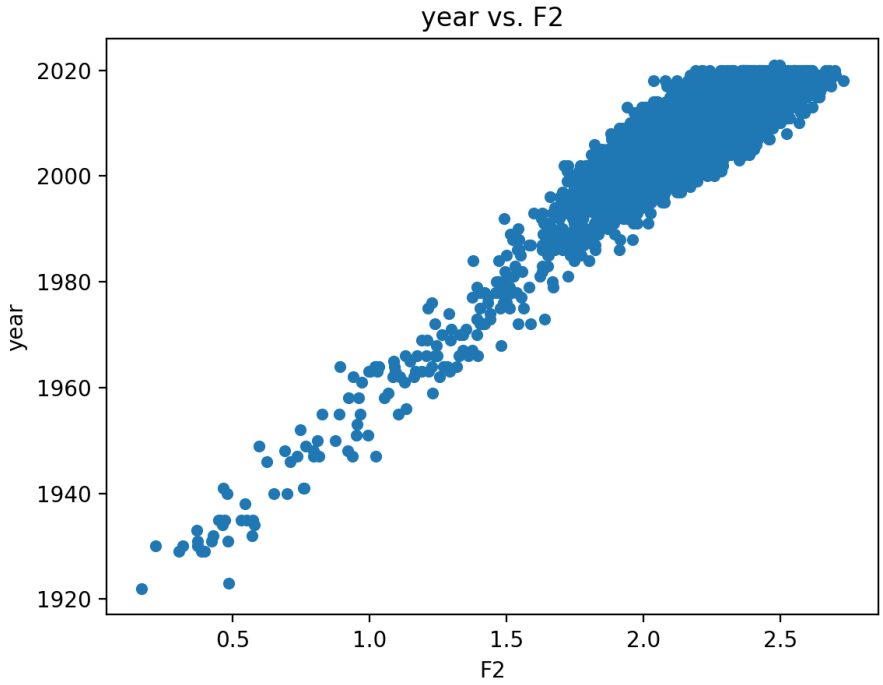

FIG. 7. year vs F2

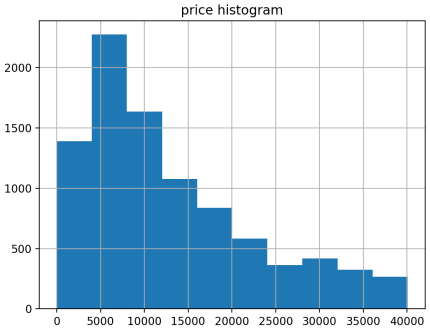

FIG. 8. Price

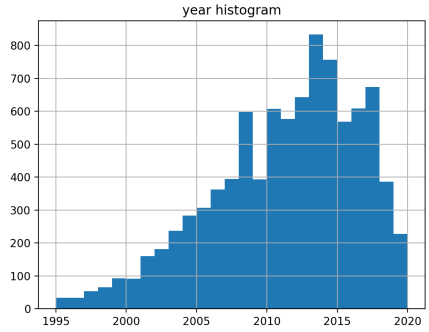

FIG. 9. Year

Figure 8, Figure 9, Figure 10, Figure 11 displays the histogram of variables - "price", "year", "F2", "F3".

Figure 10 and Figure 11 are like bell curves, which depict that most of cars have F2 and F3 values in the mid-value ranges.

Some of the assumptions are as follows:

1. **Linearity:** This assumes that there is a linear relationship between the predictors and the response variable. This also assumes that the predictors are additive.

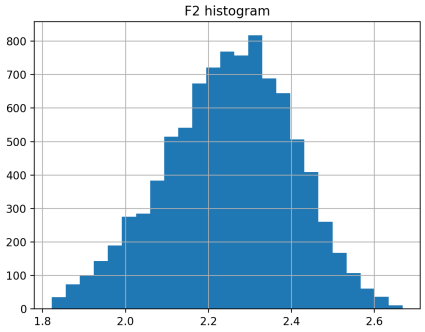

FIG. 10. F2

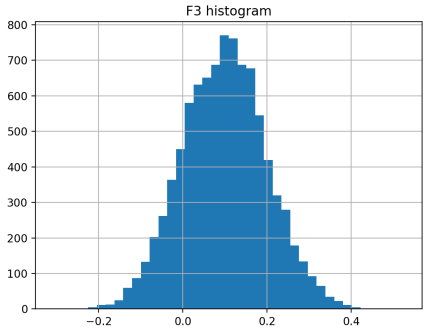

FIG. 11. F3

2. **Normality of the Error Terms:** This assumes that the error terms of the model are normally distributed. Linear regressions other than Ordinary Least Squares (OLS) may also assume normality of the predictors or the label, but that is not the case here.

3. **Outliers:** We are assuming that the outliers removed are bad values but there could be a chance that those values are actually like this. Some good values can be deleted in the process of removing outliers.

VIII. CONCLUSIONS

A Linear Regression model, a LASSO model, an Elastic Net model, and a Ridge regression model with most of them with similar model scores are important aspects of Machine Learning. Models with accuracy of approximately 56% is an excellent representation of the training data which is being tested on a separate part of a data set. We can conclude that the price of the car is highly dependent on the age of the car. There are a few older cars which are as expensive as the newer ones (see Fig. 6) maybe because they are antique and have their prices increased because of less wear and tear. We can also say that with each year, the price of the car decreases by $416, approximately. It can also be concluded that year and F2 are quite related and as the age of the car increases, F2 seems to be decreasing. The root mean squared error is well under the threshold of 10,000, which makes most of these model a good fit for the data.

From the single variable linear regression with price, we can perform some statistical analysis with respect to F1, F2, F3, F4:

1. F1: this model is 4.2% effective. The coefficient is 0.9353 which means that the price does not increase by a lot when F1 increases.

2. F2: this model is 18.4% effective. The coefficient is 2728 which suggests that the price is affected by a lot when F2 changes.

3. F3: this model is 0.4% effective. The coefficient is 5992 which suggests that the price does change with change in F3 but this model is not quite efficient therefore we cannot trust this value.

4. F4: this model is not at all effective, possibly because this only contains categorical variable.

DATA AVAILABILITY

Data is available at blackboard names "used_car_dataset.csv"

CODE AVAILABILITY

Code is available at `https://github.com/USC-DSCI-552-Spring2021/dsci552-spring2021-32416d-ps1-sjain681`

- elastic_net.py: Elastic net code
- feature_selection.py: Feature Selection code
- lasso_lassoCV.py: Lasso and Lasso Cross Validation code
- linear_regression_OLS.py: Linear Regression statsmodels code
- linear_regression_sklearn.py: Linear Regression sklearn code
- ridge_regression.py: Ridge Regression code
- single-1.py: Each variable with price regression sklearn
- single-2.py: Each variable with price regression statsmodels
- visualizations.py: Visualizations

ACKNOWLEDGMENTS

I would like to thank professor Kristina Lerman, professor Keith Burghardt, the TAs and the graders of the course DSCI 552 for all their efforts towards this assignment.

Appendix A: References

1. Jeff Macaluso. "Testing Linear Regression Assumptions in Python." Jeff Macaluso, 27 May 2018, jeffmacaluso.github.io/post/LinearRegressionAssumptions/.

2. "Sklearn.linear_model.LinearRegression." Scikit, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

3. "Statsmodels.regression.linear_model.OLS." Statsmodels, www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html.