Car Sales Prediction (Problem Set 1)

Ananya Sharma

Department of Computer Science,
University of Southern California, Los Angeles, California 90089, USA

(Dated: February 4, 2021)

Abstract

In this report we explore the datasets given, divide them into training, validation and test datasets. Trends amongst the data is observed. A regression model is chosen depending upon its residual sum of squares. Before proceeding, data is cleaned.

I. INTRODUCTION AND DATA EXPLORATION

We have 2 datasets. One for training and validating our model and the other to test the tightness of our model. The datasets show the price, year, manufacturer etc of the cars.

From the data we see that the price of the cars over the years more or less remains the same.

```
[13]: sns.regplot(x = "year", y="price", data=df, fit_reg = False, scatter_kws={"alpha": 0.2})

[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7fcb4cf567f0>
```

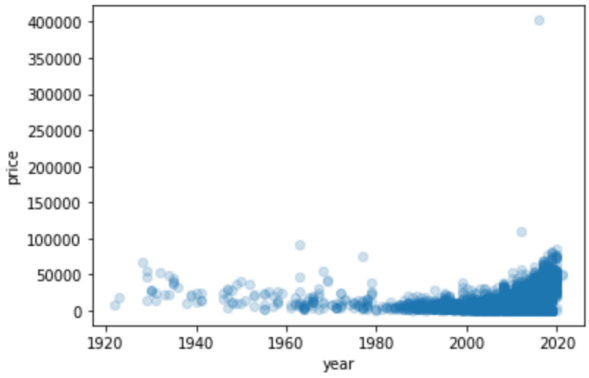
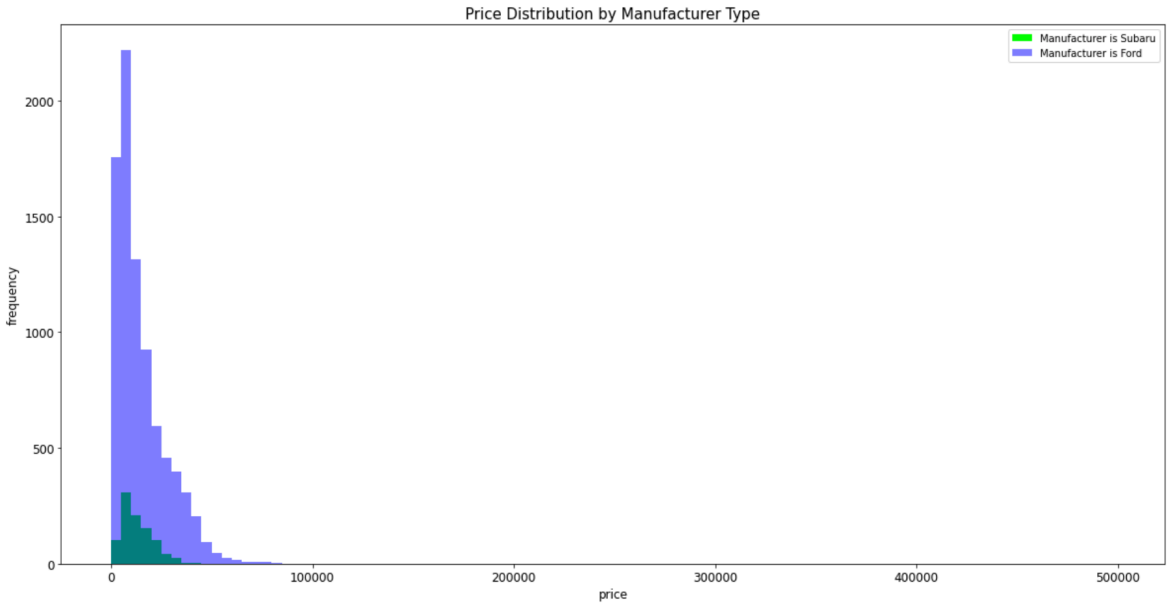


FIG 1 : Shows that the price of cars more or less remains the same over the years. However, there are a few anomalies

Further, the amount of cars manufactured by Ford are way more in comparison to Subaru.



Ss2

FIG 2: It shows that Ford is more popular.

Another pattern seen is that the price distribution is uneven and generally stays under the $15k bracket.

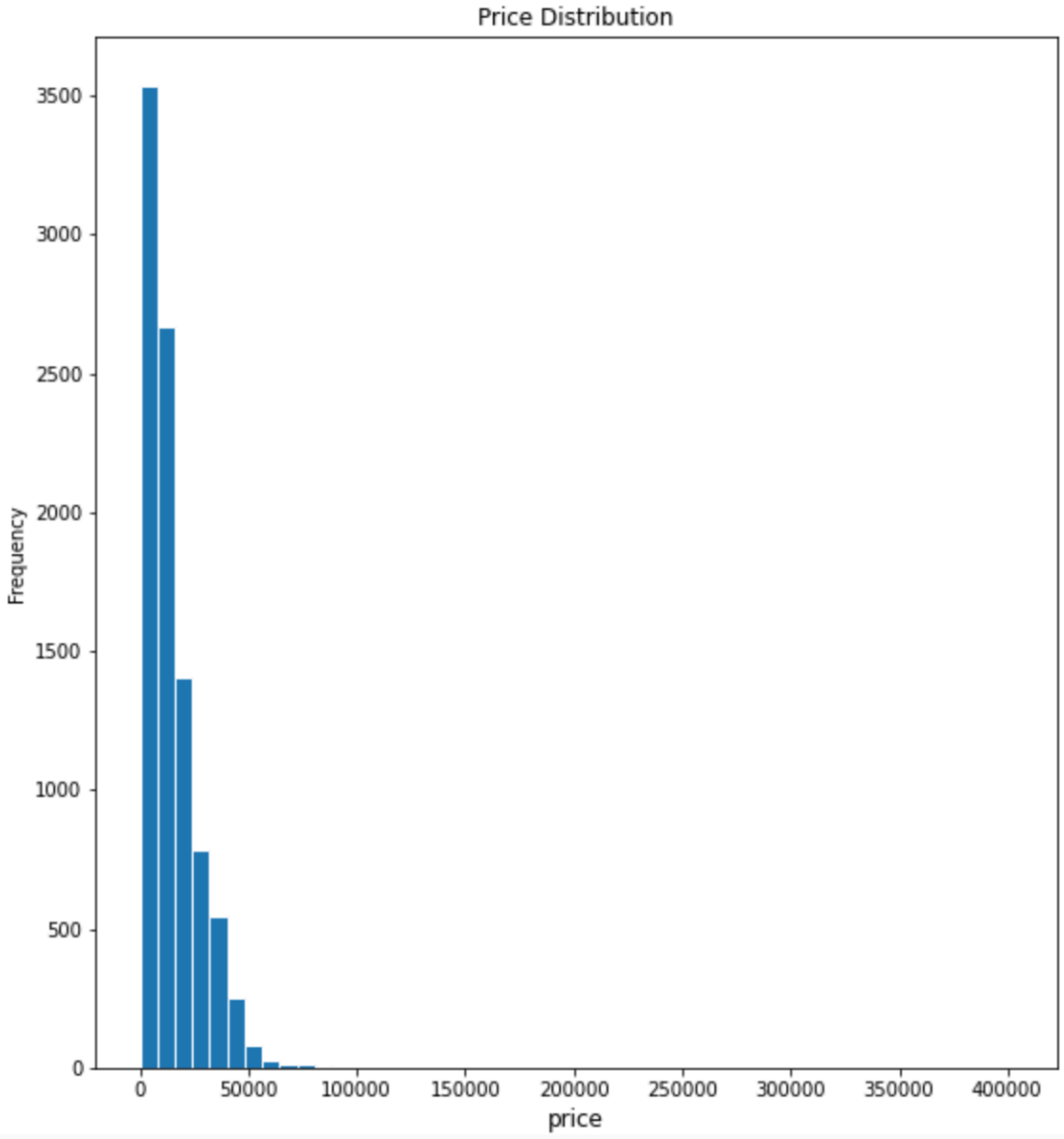


FIG 3 : Shows the frequent cost of a car.

II. DATA PREPROCESSING

The first thing to be done is cleaning data.

1. There were values with empty /missing odometer readings (0). Removed them. Ss1 before and see 2 after

```
[2]: df.shape

[2]: (9997, 14)
```

FIG 4 : With empty values, number of rows and columns

```
[17]: df = df.dropna(how='any',axis=0)

[18]: print(df)
```

```
        price    year manufacturer  condition    cylinders fuel  odometer  \
0     18219.0  2008.0         ford  excellent  8 cylinders  gas   86238.0
1       800.0  2008.0         ford  excellent  6 cylinders  gas  170953.0
2     23660.0  2016.0         ford       good  8 cylinders  gas  119026.0
3      5335.0  2009.0         ford  excellent  4 cylinders  gas   69000.0
4      1597.0  1999.0         ford       good  6 cylinders  gas   59130.0
...       ...     ...          ...        ...          ...  ...       ...
9991   7589.0  2008.0         ford  excellent  6 cylinders  gas   96800.0
9992  18924.0  2017.0         ford       good  4 cylinders  gas  122612.0
9994  26269.0  2017.0         ford  excellent  6 cylinders  gas   52541.0
9995   6149.0  2013.0         ford       good  4 cylinders  gas  197000.0
9996   9831.0  2015.0         ford  excellent  4 cylinders  gas  139000.0

     transmission    type paint_color    F1        F2        F3 F4   Age
0       automatic  pickup       black  5823  2.193844 -0.031986  b  12.0
1       automatic     SUV         red  2024  2.133691  0.097985  b  12.0
2       automatic   truck       white   294  2.160859  0.046984  c   4.0
3          manual   sedan        blue  3544  2.114929 -0.110121  c  11.0
4       automatic   sedan        blue  1329  1.829625 -0.060615  c  21.0
...           ...     ...         ...   ...       ...       ... ..   ...
9991    automatic     SUV       white   329  2.123854 -0.016047  b  12.0
9992    automatic   sedan      silver  3588  2.202934  0.212334  b   3.0
9994    automatic     SUV       white  1499  2.392569  0.094751  c   3.0
9995    automatic     SUV       black   180  2.269796  0.129762  a   7.0
9996    automatic     SUV       white   386  2.407066  0.311618  c   5.0

[9331 rows x 15 columns]
```

FIG 5 : After data cleaning

2. Handling outliers. There were a few outliers in the data which were removed. Outliers are observed by making scatterplots, histograms, etc.

3. Scaling of data was done so as to ensure all columns have equal weightage and none is heavier than the other, hence improving numerical stability of the model.

IV. MODEL SELECTION

A model consists of

- X Train — Training data of independent variables, also known as features
- X Test — Test data of independent variables
- Y Train — Training data of dependent variable
- Y Test — Test data of dependent variable

For example, we are forecasting the price of cars based on their usage, then the car price is represented as Y (dependent variable) and the usage is X (independent variables or features). Training data of X is then known as X Train which you we use to train the model.

Data set is divided into three parts:

1. Training Set
2. Validation Set
3. Test Set

Train the model on the training set (60% of the data), then perform model selection (tuning parameters) on validation set (20% of the data) and once ready, test the model on the test set (20% of the data).

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line Ridge Regression is a technique used when the data suffers has too many from multicollinearity (independent variables are highly correlated). In

multicollinearity, even though the least squares estimates (OLS) are unbiased, their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.
The correlation matrix :

```
[6]: corrMatrix = df.corr()
     sns.heatmap(corrMatrix, annot=True)
     plt.show()
```

FIG 6 : The correlation matrix

Lasso matrix sometimes struggles with a kind of data and can hence cause a slight variance in the model.

The Lasso regression technique can also cause a small bias in the model if the prediction is too dependent on a variable.

V. MODEL EVALUATION

The final performance for regression is calculated by checking its mean squared error. residual sum of squares resulting from comparing the predictions

Since the MSE is based on squared residuals, it is on the scale of the squared outcomes. Thus, the root of the MSE, which is on the scale of the outcome, is often used to report model fit:

A disadvantage of the mean-squared error is that it is not very interpretable because MSEs vary depending on the task and thus cannot be compared across different tasks.

VIII. CONCLUSIONS

We did a linear regression for analysis to predict the car sales based on their usage. Some other factors that were noted were the consistency in car sales and more popular manufacturer.

DATA AVAILABILITY

Data is available at . . . ‘https://www.kaggle.com/c/usc-dsci552-32415d-spring2021/data'

CODE AVAILABILITY

Code is available at ‘https://github.com/usc-dsci552-32415D-spring2021/problem-set-01-AnanyaSharma25/blob/main/ProblemSet1'

ACKNOWLEDGMENTS

[1]https://www.kaggle.com/mediasittich/linear-regression-for-car-price-prediction

[2] https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe/notebook

[3] ‘Hands-On Machine Learning with Scikit-Learn and TensorFlow’, 2nd edition