CSIS 3290

Fundamentals of Machine Learning

Mini Project 01

Almas A

Kateryna Timonina

Scott L.

Ricardo S.

References

Jupyter, Python and Markdown

- Python documentation: https://docs.python.org/3/
- Markdown Guide: https://www.markdownguide.org
- Jupyter Notebook: https://jupyter.org/documentation

Libraries

- Pandas documentation: https://pandas.pydata.org/
- Seaborn: https://seaborn.pydata.org/index.html
- Stats Models: https://www.statsmodels.org/stable/index.html

Regression Model

- Linear Regression in Python (by Mirko Stojiljkovic, 2020): https://realpython.com/linear-regression-in-python/
- A step-by-step guide to Simple and Multiple Linear Regression in Python (by Nikhil Adithyan, 2020): https://medium.com/codex/step-by-step-guide-to-simple-and-multiple-linear-regression-in-python-867ac9a30298

Dataset Analysis

This project is to analyze car pricing and create linear regression models to predict the pricing of certain used cars. In order to complete this analysis we utilized Python with the Jupyter Notebook and the following libraries: Math, Pandas, Numpy, Matplotlib.pyplot, Seaborn, Statsmodels.api. We also imported the linear regression library from Scikit-Learn packs, Model Selection library from SkLearn package, and Metrics library from Scikit-Learn package.

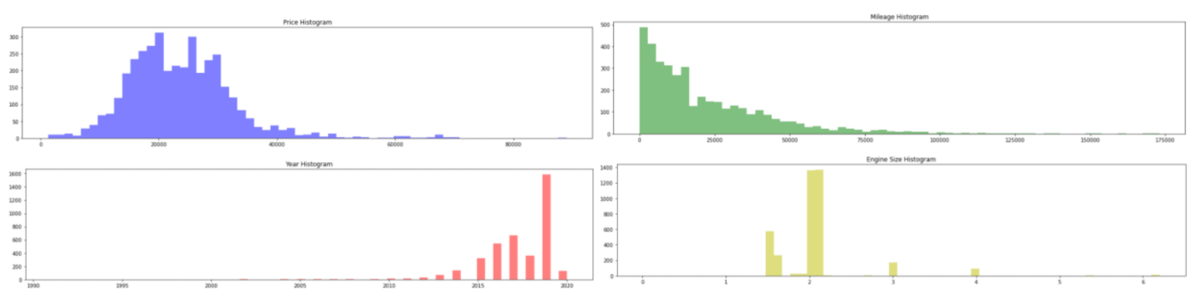
We started by reading in the cleaned data set into the Data Frame with 3899 data objects with 12 attributes each.

```
Data Frame Head:
    Year    Price  Mileage  Engine_size  Automatic  Manual  Other_Transmission  \
0  2020  30495.0   1200.0          2.0          1       0                   0
1  2020  29989.0   1000.0          1.5          1       0                   0
2  2020  37899.0    500.0          2.0          1       0                   0
3  2019  30399.0   5000.0          2.0          1       0                   0
4  2019  29899.0   4500.0          2.0          1       0                   0

   Semi-Auto  Diesel  Hybrid  Other_Fuel_Type  Petrol
0          0       1       0                0       0
1          0       0       0                0       1
2          0       1       0                0       0
3          0       1       0                0       0
4          0       1       0                0       0
```

EDA

In this phase of the project we determined the attributes Year, Price, Mileage, and Engine_size are numeric values representing interval data. In this case, Year, Price, and Mileage are considered whole values. Engine_size is a ratio value. The other attributes are categorical values that were vectorized using a boolean representation to indicate the respective Transmission and Fuel Type of each car. We then removed price outliers, year outliers, and mileage outliers. The following histograms portray the price, year, mileage and engine size of the data.



Feature Observation and Hypothesis

Based on a few observations, it is possible to predict the relationship between the resale price and the attributes of the car. The following are our predictions:

- Price x Year: Newer cars will have higher resale prices as they will have more modern features already installed in the vehicles.
- Price x Milage: Cars with lower mileage will have higher resale prices as there is less wear and tear on the vehicle.
- Price x Engine Size: Cars with larger engines are more expensive to begin with and will have a higher resale price.

Simple Linear Regression Report

We applied a variety of regression models on the data the following is a selection of the models used.

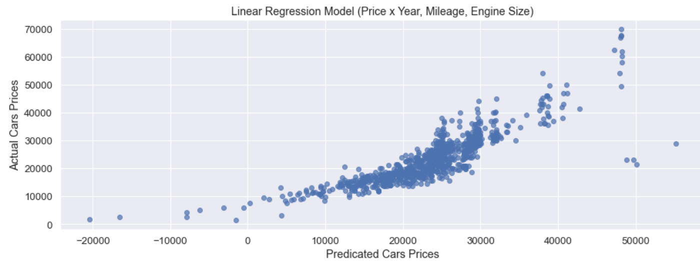
Ordinary Least Squares Regression for attributes Year, Mileage, Engine Size:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.764
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     4207.
Date:                Wed, 03 Feb 2021   Prob (F-statistic):               0.00
Time:                        13:57:25   Log-Likelihood:                -38102.
No. Observations:                3893   AIC:                         7.621e+04
Df Residuals:                    3889   BIC:                         7.624e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.259e+06   1.04e+05    -40.832      0.000   -4.46e+06   -4.05e+06
x1          2115.3193     51.639     40.964      0.000    2014.077    2216.562
x2            -0.1382      0.005    -28.152      0.000      -0.148      -0.129
x3          8895.7046    144.105     61.731      0.000    8613.175    9178.234
==============================================================================
Omnibus:                      716.931   Durbin-Watson:                   1.626
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6226.995
Skew:                           0.624   Prob(JB):                         0.00
Kurtosis:                       9.069   Cond. No.                     4.73e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.73e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Simple Linear Regression Model using Year, Mileage, and Engine Size:

Our dependent variable is price and our independent variables are year, mileage, and engine size. We then split the dataset into training data and test data, the X train shape is (2919, 3), Y train shap (2919,), X test shape (974, 3), and the Y test shape is (974,). Coefficient of Determination (R^2) in the Linear Regression Training set is: 0.7728755111094647. Coefficient of Determination (R^2) in the Linear Regression Test set is: 0.7361555786000258. The RMSE of the Linear Regression Model is: 4431.70134292943




```
The Linear Regression Coefficients per Feature are:
       Feature           Coefficients
0         Year     2117.7476357137134
1      Mileage   -0.13719501318849436
2  Engine_size      9166.468747554634
```

Linear Regression with Lasso/Ridge Report

For our linear regression with lasso and ridge report, our dependent variable was price, and the independent variables were all other attributes. When applying ridge regression, our alpha values were: [0.01, 0.1, 1, 100, 150, 160, 180, 200]. Our results from the ridge regression with alpha of 150 were:

	Model	R-square Train Set	R-square Test Set	RMSE
0	Simple Linear Regression (Price x Year, Mileage, Engine Size)	0.772876	0.736156	4431.701343
1	Simple Linear Regression (Price x Year, Engine Size)	0.726094	0.684534	4845.885023
2	Ridge Regression - alpha = 150 (Price x All Features)	0.787007	0.753078	4287.226549

When applying lasso regression, our alpha values were [0.01, 0.1, 1, 10, 50, 75, 100, 150] and our results using alpha 75 were:

	Model	R-square Train Set	R-square Test Set	RMSE
0	Simple Linear Regression (Price x Year, Mileage, Engine Size)	0.772876	0.736156	4431.701343
1	Simple Linear Regression (Price x Year, Engine Size)	0.726094	0.684534	4845.885023
2	Ridge Regression - alpha = 150 (Price x All Features)	0.787007	0.753078	4287.226549
3	Lasso Regression - alpha = 75 (Price x All Features)	0.787979	0.752654	4290.908678

Based on the values of R-Square and RMSE computed before, the performance obtained applying the Ridge Regression Model using alpha equal to 150 was the best one comparing Ridge and Lasso Models.

Polynomial Regression Report

Polynomial Regression will be performed based on the best linear regression model so far, using 3 features/independent variables: Year, Mileage, and Engine_size, and price being our dependent variable. Our polynomial degrees were set to: [1, 2, 3, 4, 5]. Based on our results, the best performance applying polynomial regression was using 4 for the degree and are as follows:

	Model	R-square Train Set	R-square Test Set	RMSE
0	Simple Linear Regression (Price x Year, Mileage, Engine Size)	0.772876	0.736156	4431.701343
1	Simple Linear Regression (Price x Year, Engine Size)	0.726094	0.684534	4845.885023
2	Ridge Regression - alpha = 150 (Price x All Features)	0.787007	0.753078	4287.226549
3	Lasso Regression - alpha = 75 (Price x All Features)	0.787979	0.752654	4290.908678
4	Simple Polynomial Regression - degree = 4 (Price x Year, Mileage, and Engine Size)	0.898262	0.885361	2921.211998

Summary Table

The comparison table shows that the Polynomial Regression using 4 degrees has the lowest RMSE and highest R-Square computed using the Test data set. Based on that, it is possible to conclude that the **Polynomial Regression using 4 degrees** has the best performance among all the models tested.

	Model	R-square Train Set	R-square Test Set	RMSE
0	Simple Linear Regression (Price x Year, Mileage, Engine Size)	0.772876	0.736156	4431.701343
1	Simple Linear Regression (Price x Year, Engine Size)	0.726094	0.684534	4845.885023
2	Ridge Regression - alpha = 150 (Price x All Features)	0.787007	0.753078	4287.226549
3	Lasso Regression - alpha = 75 (Price x All Features)	0.787979	0.752654	4290.908678
4	Simple Polynomial Regression - degree = 4 (Price x Year, Mileage, and Engine Size)	0.898262	0.885361	2921.211998

In addition, based on RMSE and Test Set R-squared, the Polynomial Regression model showed much better results than the Linear Regression, Ridge Regression, and Lasso Regression models. This statement attests to the feasibility of using the Polynomial Features approach in this case.

Please refer to our attached Jupyter notebook for more plots and charts of the data.