ALY 6020:

PREDCTIVE ANALYTICS

Mid-Week 2: Linear Regression Analysis for Car Prices

Submitted To:
Prof. Chinthaka Pathum Dinesh Herath Gedara, Faculty Lecturer

Submitted By:
Abhilash Dikshit

Academic Term: Winter 2024
Northeastern University, Vancouver, BC,Canada
Master of Professional Studies in Analytics

January 20, 2024

Title: Linear Regression Analysis for Car Prices

I. Abstract

This report presents a linear regression analysis aimed at predicting the prices of cars using a dataset and data dictionary. The analysis involves feature selection, model fitting, and interpretation of results. The objective is to identify the most significant variables impacting car prices and evaluate the accuracy of the predictive model.

II. Introduction

The dataset under consideration contains information on various car attributes such as engine specifications, fuel type, and performance metrics. The analysis utilizes the linear regression model to establish relationships between these features and the target variable, which is the price of the cars.

III. Methodology

Data Loading and Preprocessing

The dataset was loaded and examined for any missing or inconsistent values. Non-numeric columns were dropped or encoded appropriately for compatibility with the linear regression model. The dataset was split into training and testing sets to assess the model's performance.

Linear Regression Model

A linear regression model was fitted to the training data using the Ordinary Least Squares (OLS) method. The model's summary provides insights into the coefficients, p-values, and R-squared values.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.851
Method:                 Least Squares   F-statistic:                     63.28
Date:                Sat, 20 Jan 2024   Prob (F-statistic):           1.71e-56
Time:                        18:43:13   Log-Likelihood:                -1536.5
No. Observations:                 164   AIC:                             3105.
Df Residuals:                     148   BIC:                             3155.
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.507e+04   1.57e+04     -3.512      0.001   -8.61e+04   -2.41e+04
x1           -12.7237      4.510     -2.821      0.005     -21.635      -3.812
x2           227.8530    262.367      0.868      0.387    -290.616     746.322
x3           116.5145    114.011      1.022      0.308    -108.786     341.815
x4           -54.0264     58.018     -0.931      0.353    -168.677      60.624
x5           478.0259    248.126      1.927      0.056     -12.302     968.354
x6           253.1410    142.636      1.775      0.078     -28.724     535.006
x7             0.6704      1.798      0.373      0.710      -2.883       4.224
x8           112.9261     15.403      7.332      0.000      82.488     143.364
x9           510.8291   1254.610      0.407      0.684   -1968.433    2990.091
x10        -3267.6119    800.405     -4.082      0.000   -4849.310   -1685.914
x11          359.6388     91.484      3.931      0.000     178.856     540.422
x12           28.8914     16.728      1.727      0.086      -4.166      61.949
x13            2.3040      0.696      3.312      0.001       0.929       3.679
x14         -390.4061    194.712     -2.005      0.047    -775.181      -5.631
x15          209.7428    163.331      1.284      0.201    -113.019     532.505
==============================================================================
Omnibus:                        7.254   Durbin-Watson:                   1.854
Prob(Omnibus):                  0.027   Jarque-Bera (JB):               12.853
Skew:                           0.054   Prob(JB):                      0.00162
Kurtosis:                       4.367   Cond. No.                     3.88e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.88e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

IV. Results and Discussion

Model Summary

The linear regression model resulted in the following key statistics:

- R-squared: 0.865 (Adjusted R-squared: 0.851)
- F-statistic: 63.28
- P-value (Prob F-statistic): 1.71e-56

These values indicate that the model explains a significant proportion of the variance in car prices.

Significant Variables

Three notable methods—forward selection, backward elimination, and stepwise selection—were employed to identify the most significant variables.

Forward Selection

```
[16]: def forward_selection(X, y):
    features = list(X.columns)
    selected_features = []
    remaining_features = features.copy()

    while remaining_features:
        p_values = []
        for feature in remaining_features:
            model = sm.OLS(y, sm.add_constant(X[selected_features + [feature]])).fit()
            p_values.append((feature, model.pvalues[feature]))

        best_feature, min_p_value = min(p_values, key=lambda x: x[1])

        if min_p_value < 0.05:  # Adjust the significance level as needed
            selected_features.append(best_feature)
            remaining_features.remove(best_feature)
        else:
            break

    return selected_features

# Perform Forward selection on your data
selected_features_forward = forward_selection(X_train_numeric, y_train)

# Print the selected features
print("Selected Features (Forward Selection):", selected_features_forward)
```

```
Selected Features (Forward Selection): ['enginesize', 'horsepower', 'carwidth', 'stroke', 'car_ID', 'compressionratio', '
peakrpm', 'citympg', 'carheight']
```

Stepwise:

```
Add 1 feature, "enginesize", P-value: 0.0000
Add 1 feature, "horsepower", P-value: 0.0000
Add 1 feature, "carwidth", P-value: 0.0000
Add 1 feature, "stroke", P-value: 0.0033
Add 1 feature, "car_ID", P-value: 0.0056
Add 1 feature, "compressionratio", P-value: 0.0028
Add 1 feature, "peakrpm", P-value: 0.0070
Add 1 feature, "citympg", P-value: 0.0035
Remove 1 feature, "horsepower", P-value: 0.0988
Selected Features (Stepwise Selection): ['enginesize', 'carwidth', 'stroke', 'car_ID', 'compressionratio', 'peakrpm', 'ci
tympg']
```

Interpretation of Coefficients

The coefficients associated with each feature provide insights into their impact on car prices. For instance, the coefficient for 'enginesize' suggests that a one-unit increase in engine size results in a decrease of $12,723.7 in car prices.

V. Accuracy Assessment

The model's accuracy was evaluated based on the R-squared value, which indicates the proportion of variance in car prices explained by the model. With an R-squared of 0.865, the model demonstrates a high level of accuracy.

VI. Conclusion

The linear regression analysis revealed that 'enginesize,' 'carwidth,' 'stroke,' 'car_ID,' 'compressionratio,' 'peakrpm,' and 'citympg' are significant variables influencing car prices. Of these, 'enginesize' had the greatest positive influence, as indicated by its high coefficient.

The model, with an R-squared of 0.865, is deemed accurate in predicting car prices. However, it is essential to consider potential limitations, such as multicollinearity, as indicated by the large condition number.

VII. Recommendations

The findings from this analysis can be valuable for car manufacturers and sellers to understand the factors contributing to car prices. Further refinements and validations may enhance the model's robustness.