Report on the analysis of the mtcars data set using logistic regression to estimate the probability of manual transmission

"Milica Djurkovic"

2023-08-27

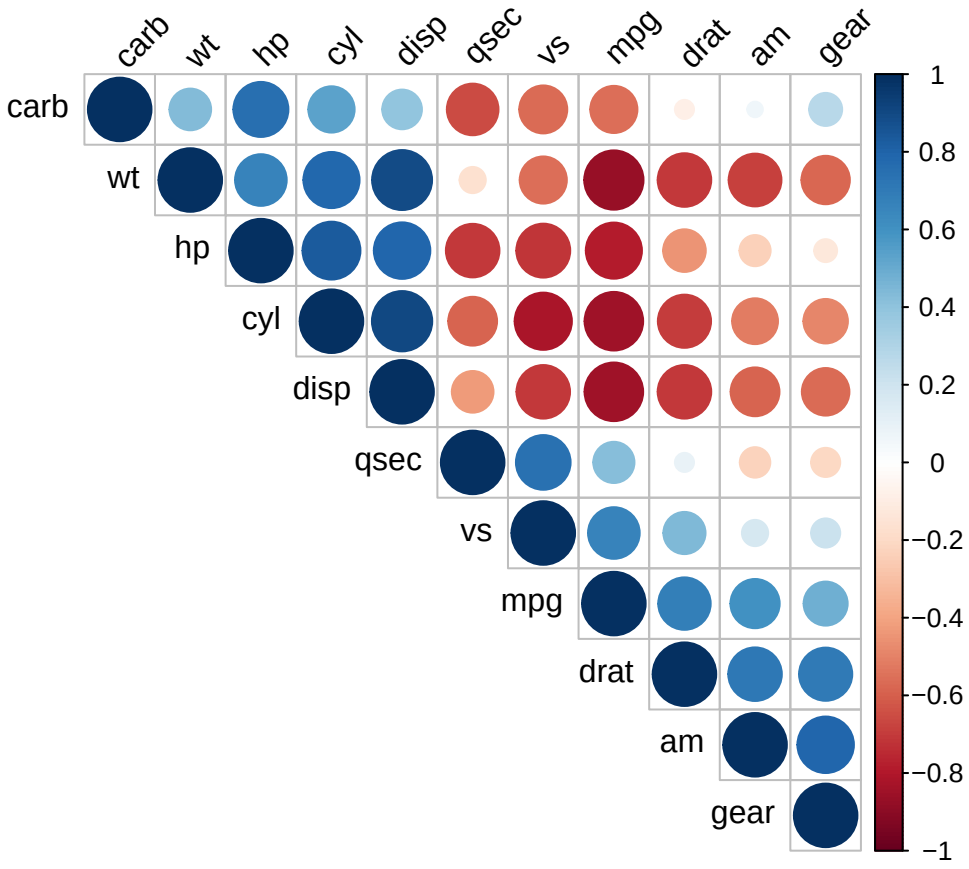
Problem Statement

This report analyzes the "mtcar" dataset to estimate the probability that a car's transmission is manual or automatic, using a logistic regression technique. The goal of the research was to develop a model that can classify the type of transmission based on available car characteristics.

The dataset "mtcar" contains information about various car specifications. The target variable for our analysis is the transmission column, "am", which represents the type of transmission (1 for manual and 0 for automatic). We aim to predict the transmission type based on rear axle ratio, "drat", and a number of forward gears "gear".

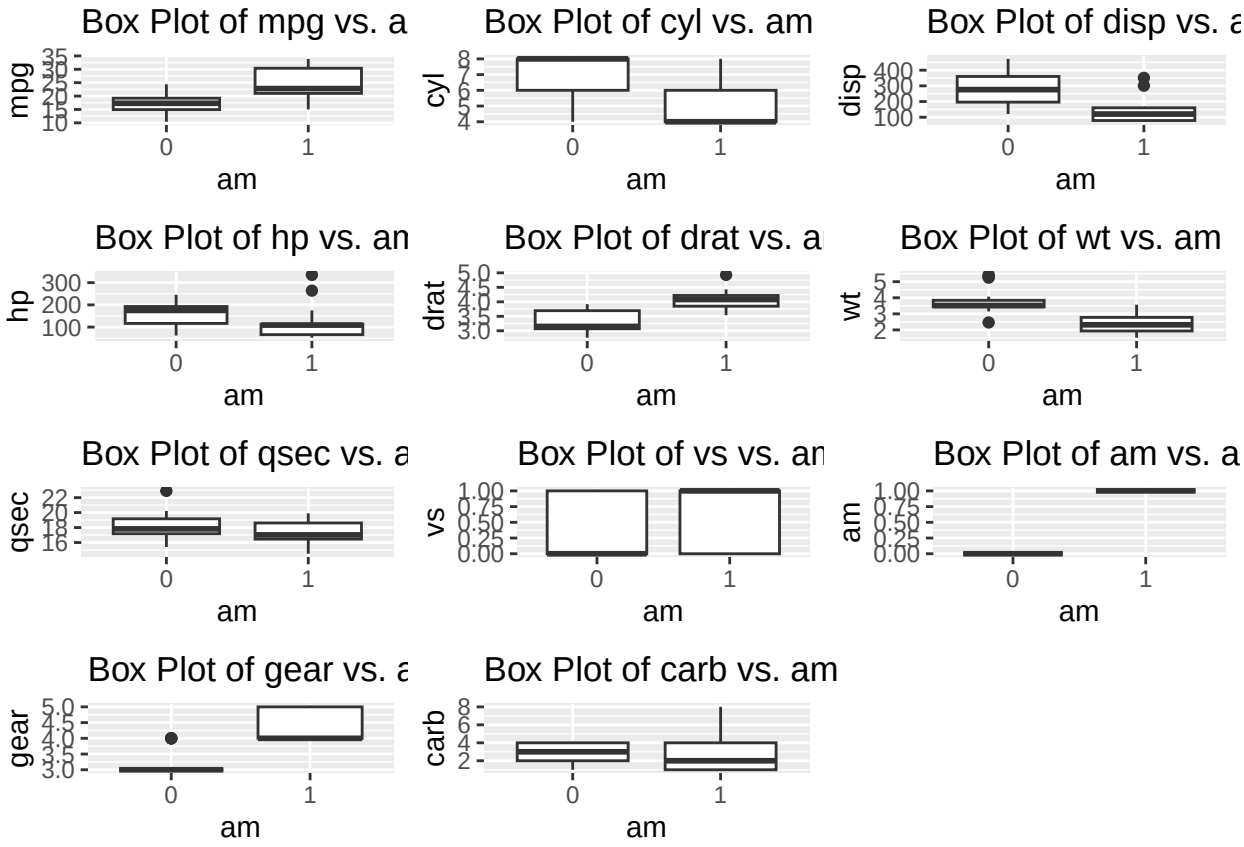
Based on the analysis of the correlation table, and later applying the VIF technique for assessing multicollinearity, we notice that the variables number of cylinders - 'cyl', displacement - 'disp', 1/4 mile time - 'qsec' and a number of carburetors - 'carb', show a pronounced mutual correlation. High absolute correlations between these variables indicate the potential presence of multicollinearity between them.

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```



Visualization of Box Plot

Box Plot: 'am' vs. Numerical Variable Below is the box plot showing the distribution of 'am' variable in relation to a numerical variable.



Logistic Regression

The logistic regression analysis was performed to explore the relationship between the transmission type ("am") and the rear axle ratio, "drat", and a number of forward gears "gear" variables. The model indicates that both weight and horsepower play a significant role in predicting the type of transmission in automobiles.

Based on the analysis of the model applied to the mtcars_data dataset using a binomial model, the following conclusions can be drawn:

VIF Values

VIF Values: The examination of Variance Inflation Factor (VIF) values did not indicate a significant issue of multicollinearity among the independent variables. All VIF values are close to 1, indicating low mutual correlation between predictors.

Model Summary

A binomial model with the formula am ~ drat + gear was employed. The analysis of residuals suggests that the model has relatively low deviance, indicating a good fit to the data.

Coefficients

The computed coefficients were transformed using the exponential function for better interpretability. Interpreting these coefficients provides insights into the probability of change in the dependent variable relative

to changes in the independent variables.

Predictions and Metrics

Based on the applied model, predicted values were generated using a threshold of 0.5. Accuracy, hit rate, and false alarm rate metrics were calculated. The model exhibited a high accuracy of 87.5%, a high hit rate of 92.31%, and a low false alarm rate of 15.79%.

The model appears to fit the mtcars_data well, as indicated by the low deviance, good accuracy metrics, and high prediction success rates.

```
## [1] "VIF values: 1.00000053710826" "VIF values: 1.00000053710826"
##
## Call:
## glm(formula = am ~ drat + gear, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3283   0.0000   0.0000   0.0007   1.4042
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -137.798  16793.940  -0.008    0.993
## drat           12.382      8.831   1.402    0.161
## gear           22.402   4198.470   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 10.535  on 29  degrees of freedom
## AIC: 16.535
##
## Number of Fisher Scoring iterations: 20
##
##  (Intercept)         drat         gear
## 1.429189e-60 2.385525e+05 5.357542e+09
## $accuracy
## [1] 0.875
##
## $hit_rate
## [1] 0.9230769
##
## $false_alarm_rate
## [1] 0.1578947
```

Evaluation of "am" Classification Using Different Thresholds

An evaluation was conducted to assess the performance of the classification model for predicting the "am" variable based on the rear axle ratio, "drat", and a number of forward gears "gear" variables. Various decision thresholds were employed to determine how the model performs under different conditions. The evaluation included the following metrics:

Hit Rate (Recall): Represents the proportion of actual positive cases (manual transmission) that were correctly identified by the model.

False Alarm Rate: Indicates the proportion of actual negative cases (automatic transmission) that were incorrectly classified as positive cases.

Accuracy: Refers to the proportion of all cases that were correctly classified by the model.

Analysis of ROC Curve

The provided code generates a Receiver Operating Characteristic (ROC) curve and calculates corresponding metrics using a range of decision thresholds. The ROC curve illustrates the trade-off between true positive rate (hit rate) and false positive rate (false alarm rate) as the decision threshold changes.

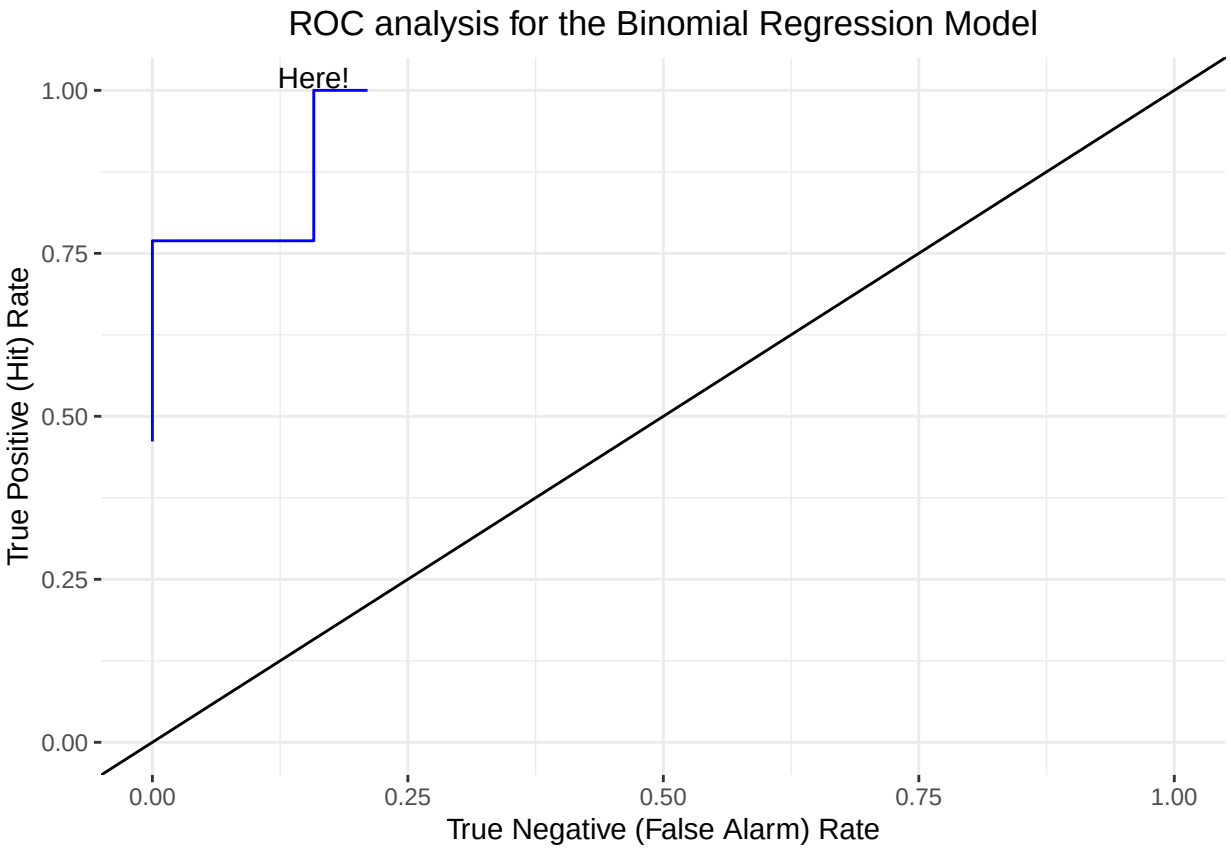
The ROC frame displays hit rates, false alarm rates, accuracies, and the decision thresholds used. The highest difference between hit rate and false alarm rate occurs at a threshold of 0.08, resulting in a value of 0.8421053. This indicates a point on the ROC curve that balances well between high hit rate and low false alarm rate.

```
##   hit_rate false_alarm_rate accuracy  dec      diff label
## 1        1        0.2105263    0.875 0.01 0.7894737
## 2        1        0.2105263    0.875 0.02 0.7894737
## 3        1        0.2105263    0.875 0.03 0.7894737
## 4        1        0.2105263    0.875 0.04 0.7894737
## 5        1        0.2105263    0.875 0.05 0.7894737
## 6        1        0.2105263    0.875 0.06 0.7894737
```

The ROC curve plot

In our analysis, we've identified a point on the ROC curve that stands out as the optimal operating point. At this threshold, the hit rate is impressively high at 0.923, indicating the model's ability to correctly classify positive instances. Simultaneously, the false alarm rate remains relatively low at 0.158, implying the model's effectiveness in minimizing false positives.

The key metrics further reinforce the model's commendable performance. The overall accuracy of 0.875 demonstrates that the model's predictions align well with the actual outcomes. The high hit rate showcases the model's capability to capture a substantial portion of positive instances, while the low false alarm rate signifies its proficiency in avoiding false positives.



Conclusion

Based on our analysis of the linear model, we can conclude the linear regression model exhibited strong predictive capabilities, as demonstrated by its accuracy and hit rate metrics. The comprehensive evaluation, including the analysis of coefficients, prediction metrics, and ROC curve, collectively indicates that the model is well-suited for capturing patterns and making predictions within the given dataset.