# What determines the price of a car?

Sudi Narasimhan

For the purposes of this exercise different models were attempted including

* Ridge
* RandomForest
* Linear

Evaluation Criteria used MSE and R2 as the best determinants of how well the model was performing.

Ridge + Polynomial **Model Run 1**

* Attempting to do a GridSearch, which is a technique for finding the optimal parameters given a set of parameters or choosing the best combination of hyperparameters.
* Attempting to run a Polynomial with Ridge results in the following results:
* Polynomial transformation can capture non-linear relationships to support to fit more complex functions.

Fitting 5 folds for each of 5 candidates, totalling 25 fits

Best alpha: 1.0

R2 Score: 0.7879546886922705

Mean Squared Error: 48805749.65502113

* The MSE indicate as above, poor model performance
* The high R2 score suggests that the model does explain 79% of the variance
* The MSE or Mean Squared Error (MSE) which measures the average squared difference between the predicted values and the actual values indicates that the model is able to accurately capture the underlying data and indicating poor model performance.

Standardize the data with StandardScaler and Ridge **Model Run 2**

Fitting 5 folds for each of 7 candidates, totalling 35 fits

Best Parameters: {'regressor\_\_alpha': 10.0, 'regressor\_\_solver': 'sag'}

Best Negative Mean Squared Error: -71013997.0541757

R2 Score on Test Set: 0.6955929446520976

Mean Squared Error on Test Set: 70064338.81940931

* This indicates that the model explains 69~70% of the variance in the test data.
* MSE which measures the average squared difference between the predicted and actual values on the test data is still substantial which indicates questionable model performance.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* Standardizing the data with the Random Forest Classifier **Model Run 3** resulted in:

MSE: 46560022.2610

RMSE: 6823.4905

R²: 0.7977

* This suggests that R2 score of .7977 or 79~80 % of the variance is explained by the model.
* High MSE score here indicates the average squared difference between the predicted and actual values. Naturally, a lower MSE indicates a better model performance in terms of prediction accuracy.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Model Run 4**

R² Score: 0.9029

Mean Squared Error (MSE): 22358984.1553

Root Mean Squared Error (RMSE): 4728.5288

Feature Importance:

region 0.013497

year 0.041504

manufacturer 0.014132

model 0.030752

condition 0.005752

cylinders 0.012368

fuel 0.007928

odometer 0.046792

title\_status 0.000992

transmission 0.006050

drive 0.008859

size 0.001608

type 0.011136

paint\_color 0.008631

state 0.006224

price\_category 0.783776

dtype: float64

A graph with a bar graph

Description automatically generated

* As the previous chart indicated, the R2 is able to explain 90% of the variance in the data. The MSE as the average squared error between the actual and predicted values remain pretty substantial.
* What the model identified is the most important features of the data include the price\_category, the year, odometer readings, which influence or dictate the price of a car. This is intuitively accurate since the number of miles and the year of a car have historically been factors in the price of a car. However, what seems surprising is other criteria like vehicle model have not weighed as substantially as expected.

**Model Run 5**

Best parameters: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}

Best cross-validation accuracy: 0.2434

Test set accuracy with best model: 0.2444

Mean Squared Error (MSE): 45939390.4828

R² score: 0.8004

A graph with blue dots and red line

Description automatically generated

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Model Run 6**

Random Forest Regressor with Randomized Search

Best parameters: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 233}

Best cross-validation MSE: 30852538.4438

Test set Mean Squared Error (MSE): 22152791.2064

Test set R² score: 0.9038

A graph with text overlay

Description automatically generated

* 90 % of the variance can be explained by the model
* Once again, price category is very correlated to the price. Beyond that the year, and odometer, and model are the most important features and factors associated with the price of a vehicle.

**Model Run 7**

Best parameters: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 233}

Best cross-validation MSE: 30852538.4438

Test set Mean Squared Error (MSE): 22152791.2064

Test set R² score: 0.9038

Feature importances:

price\_category: 0.7842

odometer: 0.0468

year: 0.0415

model: 0.0308

manufacturer: 0.0141

region: 0.0137

cylinders: 0.0122

type: 0.0112

drive: 0.0086

paint\_color: 0.0085

fuel: 0.0078

state: 0.0063

transmission: 0.0060

condition: 0.0056

size: 0.0016

title\_status: 0.0010

A graph with numbers and a bar

Description automatically generated with medium confidence

* 90% of the variance can be explained by the model
* Once again, price category is very correlated to the price. Beyond that the year, and odometer, and model are the most important features and factors associated with the price of a vehicle.

**Model Run 8**

With Linear Regression

Best parameters: {'regressor\_\_fit\_intercept': True, 'regressor\_\_positive': False}

Best cross-validation MSE: 71014858.2577

Test set Mean Squared Error (MSE): 70050610.7626

Test set R² score: 0.6957

69% of the variance can be explained by the model.