# 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 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: 10.0

R2 Score: -0.0864614516599298

* The high negative R2 and MSE indicate as above, poor model performance
* The negative R2 score suggests that the model does not explain any 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 not 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': 'lsqr'}

Best Negative Mean Squared Error: -82750190.87526745

R2 Score on Test Set: 0.6906777638298627

Mean Squared Error on Test Set: 85492245.53149173

* This indicates that the model explains 69% of the variance in the 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: 121907663.0426

RMSE: 11041.1803

R²: 0.5589

* This suggests that R2 score of .5589 or 55.89% 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: 0.8772
* Mean Squared Error (MSE): 33952791.5704
* Root Mean Squared Error (RMSE): 5826.9024
* Feature Importance:
* region 0.013506
* year 0.037516
* manufacturer 0.013576
* model 0.023850
* condition 0.006092
* cylinders 0.010201
* fuel 0.008217
* odometer 0.037026
* title\_status 0.000628
* transmission 0.007703
* drive 0.006621
* size 0.001918
* type 0.010772
* paint\_color 0.007218
* state 0.010168
* price\_category 0.804988
* dtype: float64

A graph with numbers and letters

Description automatically generated

* As the previous chart indicated, the R2 is able to explain 87% 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 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': 2, 'min\_samples\_split': 2, 'n\_estimators': 200}

Best cross-validation accuracy: 0.1970

Test set accuracy with best model: 0.2232

Mean Squared Error (MSE): 128799808.9174

R² score: 0.5340

A graph with blue dots and red line

Description automatically generated

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

**Model Run 6**

Random Forest Regressor with Randomized Search

Best parameters: {'max\_depth': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 3, 'n\_estimators': 178}

Best cross-validation MSE: 31193328.6885

Test set Mean Squared Error (MSE): 33718988.2918

Test set R² score: 0.8780

<ipython-input-304-4e3e8285f3ed>:48: MatplotlibDeprecationWarning: Auto-removal of overlapping axes is deprecated since 3.6 and will be removed two minor releases later; explicitly call ax.remove() as needed.

plt.subplot(1, 2, 2)

A graph with text overlay

Description automatically generated

* 87 % 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': 30, 'min\_samples\_leaf': 1, 'min\_samples\_split': 3, 'n\_estimators': 178}

Best cross-validation MSE: 31193328.6885

Test set Mean Squared Error (MSE): 33718988.2918

Test set R² score: 0.8780

Feature importances:

price\_category: 0.8081

year: 0.0372

odometer: 0.0365

model: 0.0234

manufacturer: 0.0136

region: 0.0131

type: 0.0105

cylinders: 0.0099

state: 0.0099

fuel: 0.0081

transmission: 0.0076

paint\_color: 0.0071

drive: 0.0066

condition: 0.0060

size: 0.0018

title\_status: 0.0007

A graph with numbers and a blue line

Description automatically generated with medium confidence

* 87 % 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: 82751083.8146

Test set Mean Squared Error (MSE): 85454379.0829

Test set R² score: 0.6908

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