Pricing Cars

Tyler Engalla

2024-12-09

Pricing Cars

To predict the price of a car based on a set of the car's features, the supervised learning method I chose was a Linear Regression model. This method was chosen as a straight forward means of getting to a predicted price based on 15 features: trim, subTrim, condition, isOneOwner, mileage, year, color, displacement, fuel, state, region, soundSystem, wheelType, wheelSize, featureCount.

The first step was to explore the data set. There was some pre-processing that needed to be done such as checking for missing values and converting the categorical variables into factors.

Once the data was prepped, the next step was to split the data into training and testing data sets so that we can see how the model performs against unseen data.

Then we run the linear model. This allows us to identify significant predictors and also refine the model as we go by excluding features based on contextual hunches and statistical significance.

```
## [1] 0
```

```
##
## Call:
## lm(formula = price ~ ., data = train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64604  -5115   -876   3521 272037
##
## Coefficients: (15 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.300e+06  1.103e+05 -75.227  < 2e-16 ***
## X                          -6.174e-03  4.637e-03  -1.332 0.183032
## trim350                     3.113e+04  1.362e+04   2.286 0.022283 *
## trim400                    -1.123e+04  1.475e+04  -0.761 0.446572
## trim420                     5.257e+04  1.356e+04   3.877 0.000106 ***
## trim430                     2.922e+04  1.353e+04   2.160 0.030807 *
## trim450                    -3.723e+04  1.619e+04  -2.300 0.021441 *
## trim500                     3.178e+04  1.353e+04   2.348 0.018859 *
## trim55 AMG                  3.088e+04  1.355e+04   2.279 0.022656 *
## trim550                    -6.477e+03  1.081e+04  -0.599 0.549216
## trim600                     3.791e+03  1.082e+04   0.350 0.726164
## trim63 AMG                  4.347e+04  1.084e+04   4.012 6.05e-05 ***
## trim65 AMG                  5.463e+03  1.088e+04   0.502 0.615730
## trimunsp                    2.834e+04  1.086e+04   2.609 0.009083 **
## subTrimunsp                        NA         NA      NA       NA    
## conditionNew                3.536e+04  2.975e+02 118.865  < 2e-16 ***
## conditionUsed              -4.390e+03  2.597e+02 -16.906  < 2e-16 ***
## isOneOwnert                -5.323e+02  2.228e+02  -2.389 0.016916 *  
## mileage                    -1.304e-01  3.255e-03 -40.057  < 2e-16 ***
## year                        4.143e+03  5.459e+01  75.887  < 2e-16 ***
## colorBlack                 -5.406e+02  8.369e+02  -0.646 0.518319    
## colorBlue                  -7.402e+02  8.820e+02  -0.839 0.401356    
## colorBronze                 3.706e+03  3.893e+03   0.952 0.341091    
## colorBrown                  2.922e+02  1.676e+03   0.174 0.861579    
## colorGold                   1.087e+03  1.129e+03   0.963 0.335555    
## colorGray                  -1.487e+03  8.730e+02  -1.703 0.088563 .  
## colorGreen                 -1.557e+02  1.238e+03  -0.126 0.899924    
## colorPurple                 5.627e+03  4.156e+03   1.354 0.175803    
## colorRed                   -2.283e+02  1.063e+03  -0.215 0.829988    
## colorSilver                -1.429e+03  8.422e+02  -1.697 0.089706 .  
## colorTurquoise             -2.535e+03  4.882e+03  -0.519 0.603632    
## colorunsp                  -2.314e+02  8.876e+02  -0.261 0.794316    
## colorWhite                  7.910e+02  8.505e+02   0.930 0.352359    
## colorYellow                -7.170e+03  7.640e+03  -0.938 0.348000    
## displacement3.2 L           5.113e+04  1.355e+04   3.773 0.000161 ***
## displacement3.5 L           3.700e+04  5.540e+03   6.679 2.46e-11 ***
## displacement3.7 L          -1.325e+04  6.016e+03  -2.202 0.027652 *  
## displacement4.2 L                  NA         NA      NA       NA    
## displacement4.3 L                  NA         NA      NA       NA    
## displacement4.6 L           3.485e+04  8.258e+03   4.221 2.44e-05 ***
## displacement4.7 L           5.677e+04  8.270e+03   6.864 6.87e-12 ***
## displacement5.0 L                  NA         NA      NA       NA    
## displacement5.4 L                  NA         NA      NA       NA    
## displacement5.5 L           3.097e+04  8.259e+03   3.749 0.000178 ***
## displacement5.8 L           3.006e+04  8.381e+03   3.586 0.000336 ***
## displacement6.0 L           7.450e+04  8.301e+03   8.974  < 2e-16 ***
## displacement6.3 L          -7.516e+03  8.296e+03  -0.906 0.364958    
## displacement8.0 L           5.857e+04  1.357e+04   4.316 1.60e-05 ***
## displacementunsp            3.960e+04  8.228e+03   4.813 1.49e-06 ***
## fuelGasoline                6.828e+03  5.855e+03   1.166 0.243537    
## fuelHybrid                         NA         NA      NA       NA    
## fuelunsp                    1.786e+04  5.805e+03   3.076 0.002099 ** 
## stateAL                     6.203e+03  7.626e+03   0.813 0.416007    
## stateAR                     6.978e+03  7.754e+03   0.900 0.368186    
## stateAZ                     7.504e+03  7.622e+03   0.984 0.324892    
## stateCA                     6.534e+03  7.600e+03   0.860 0.389903    
## stateCO                     7.121e+03  7.620e+03   0.934 0.350071    
## stateCT                     5.539e+03  7.624e+03   0.727 0.467479    
## stateDC                    -8.395e+03  9.804e+03  -0.856 0.391841    
## stateDE                     6.531e+03  7.699e+03   0.848 0.396345    
## stateFL                     5.994e+03  7.600e+03   0.789 0.430308    
## stateGA                     5.863e+03  7.605e+03   0.771 0.440694    
## stateHI                     6.167e+03  7.685e+03   0.802 0.422291    
## stateIA                     6.731e+03  7.832e+03   0.859 0.390076    
## stateID                     1.136e+04  7.969e+03   1.425 0.154117    
## stateIL                     6.839e+03  7.605e+03   0.899 0.368521    
## stateIN                     6.057e+03  7.645e+03   0.792 0.428186    
## stateKS                     7.725e+03  7.707e+03   1.002 0.316219    
## stateKY                     9.203e+03  7.644e+03   1.204 0.228638    
## stateLA                     8.009e+03  7.660e+03   1.046 0.295757    
## stateMA                     6.514e+03  7.609e+03   0.856 0.391984    
## stateMD                     6.674e+03  7.612e+03   0.877 0.380596    
## stateME                     3.533e+03  7.988e+03   0.442 0.658318    
## stateMI                     6.571e+03  7.639e+03   0.860 0.389745    
## stateMN                     7.597e+03  7.638e+03   0.995 0.319944    
## stateMO                     7.621e+03  7.624e+03   1.000 0.317530    
## stateMS                     9.155e+03  7.665e+03   1.194 0.232363    
## stateMT                     9.674e+03  8.402e+03   1.151 0.249613    
## stateNC                     7.222e+03  7.609e+03   0.949 0.342570    
## stateND                     1.049e+04  8.992e+03   1.167 0.243298    
## stateNE                     1.005e+04  8.133e+03   1.236 0.216646    
## stateNH                     7.857e+03  7.665e+03   1.025 0.305378    
## stateNJ                     6.147e+03  7.603e+03   0.809 0.418773    
## stateNM                     7.498e+03  7.849e+03   0.955 0.339411    
## stateNV                     8.106e+03  7.622e+03   1.064 0.287521    
## stateNY                     5.427e+03  7.602e+03   0.714 0.475302    
## stateOH                     5.655e+03  7.613e+03   0.743 0.457587    
## stateOK                     7.895e+03  7.649e+03   1.032 0.302010    
## stateON                     8.347e+03  9.313e+03   0.896 0.370124    
## stateOR                     7.881e+03  7.661e+03   1.029 0.303584    
## statePA                     6.620e+03  7.610e+03   0.870 0.384406    
## stateRI                     5.073e+03  7.773e+03   0.653 0.513966    
## stateSC                     7.132e+03  7.634e+03   0.934 0.350185    
## stateSD                     2.489e+04  1.075e+04   2.316 0.020543 *  
## stateTN                     5.347e+03  7.620e+03   0.702 0.482847    
## stateTX                     7.653e+03  7.601e+03   1.007 0.314056    
## stateunsp                  -5.830e+04  1.324e+04  -4.403 1.07e-05 ***
## stateUT                     8.214e+03  7.676e+03   1.070 0.284570    
## stateVA                     6.235e+03  7.607e+03   0.820 0.412467    
## stateWA                     8.294e+03  7.627e+03   1.088 0.276802    
## stateWI                     7.181e+03  7.661e+03   0.937 0.348619    
## stateWV                     8.512e+03  7.819e+03   1.089 0.276345    
## stateWY                    -4.696e+02  1.316e+04  -0.036 0.971529    
## regionESC                          NA         NA      NA       NA    
## regionMid                          NA         NA      NA       NA    
## regionMtn                          NA         NA      NA       NA    
## regionNew                          NA         NA      NA       NA    
## regionPac                          NA         NA      NA       NA    
## regionSoA                          NA         NA      NA       NA    
## regionunsp                         NA         NA      NA       NA    
## regionWNC                          NA         NA      NA       NA    
## regionWSC                          NA         NA      NA       NA    
## soundSystemBang Olufsen    -8.254e+02  7.661e+03  -0.108 0.914208    
## soundSystemBose            -6.211e+03  7.618e+03  -0.815 0.414958    
## soundSystemBoston Acoustic -5.166e+03  1.321e+04  -0.391 0.695657    
## soundSystemHarman Kardon   -6.473e+03  7.609e+03  -0.851 0.394993    
## soundSystemPremium         -4.626e+03  7.607e+03  -0.608 0.543118    
## soundSystemunsp            -4.176e+03  7.607e+03  -0.549 0.582987    
## wheelTypeChrome            -8.771e+02  1.345e+03  -0.652 0.514477    
## wheelTypePremium            3.001e+02  6.066e+02   0.495 0.620798    
## wheelTypeSteel              1.865e+04  1.683e+03  11.081  < 2e-16 ***
## wheelTypeunsp               4.459e+02  1.758e+02   2.537 0.011179 *  
## wheelSize17               -1.005e+04  1.548e+03  -6.487 8.92e-11 ***
## wheelSize18               -6.285e+03  1.258e+03  -4.997 5.87e-07 ***
## wheelSize19               -6.100e+03  1.273e+03  -4.793 1.65e-06 ***
## wheelSize20               -1.277e+03  1.299e+03  -0.983 0.325442
## wheelSize21               -5.340e+03  7.699e+03  -0.694 0.487918
## wheelSize22               -1.380e+03  2.553e+03  -0.540 0.588874
## wheelSizeunsp             -5.577e+03  1.221e+03  -4.568 4.95e-06 ***
## featureCount               1.872e-01  3.300e+00   0.057 0.954772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10740 on 23459 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.9428
## F-statistic:  3408 on 114 and 23459 DF,  p-value: < 2.2e-16
```

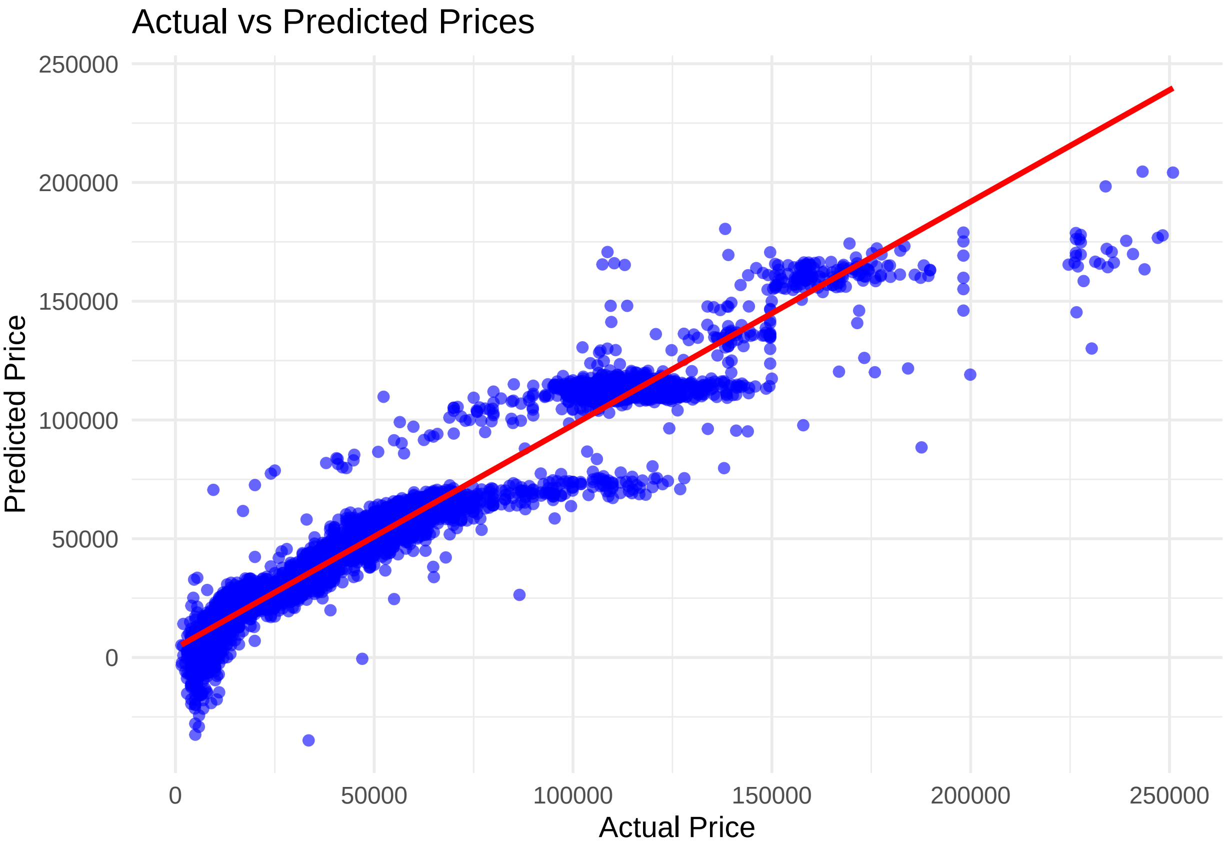
```
## [1] "Linear RMSE: 10559.2843411631"
```

The first model was run with the full feature set and gave an Adjusted R-squared: 0.9428 and a RMSE of 10559. Meaning our model was able to account for ~95% of the variability involved in estimating the price, while on average the price that was predicted differed from the actual by roughly ~$10k.

Vairables with p-values less than .05 such as certain trim packages (trim420, trim63 AMG), whether the car is new or not, the mileage, year, and many specific displacement levels significantly impact price.

For example, having a trim420 package is associated with an increase of ~52,570 in price (holding other variables constant) and mileage, as expected, decreases prices by ~.13 with each mile.

Non-significant predictors that do not impact price included sound system, most states the vehicle is from, and color.



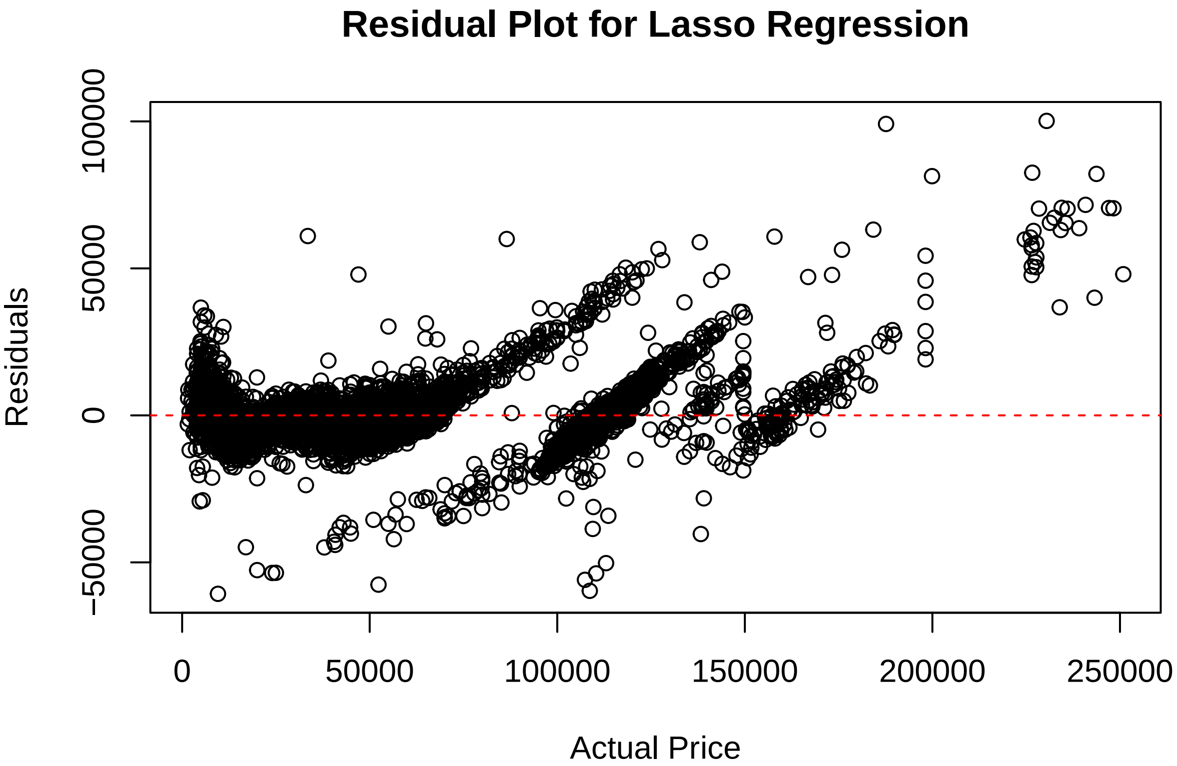
But being off by $10k on average in the car market isn't quite the accuracy we'd want when trying to accurately predict a car's price.

Trying to improve on this model, I took more of an intuitive approach to see if we can vastly simplify our model and improve our RMSE by only taking variables that were statistically significant and what I perceived as generally the most important factors in buying a car - condition, mileage, and year of the vehicle.

This attempt at refining and simplifying our model reduced our Adjusted R-squared to 0.8726 and provided a RMSE of 15764. So we accounted for less of the variability and are now on average off by ~$16k. With an attempt to greatly simplify we have made the model worse at predicting the price.

The next iterative approach was to try and regularize the data using Lasso Regression. This method allows us to eliminate some of the features that are irrelevant. This produced a RMSE of 10565. However, this is still slightly higher than our original linear model. Looking at the plot of the residuals of the Lasso Regression, one of the reasons for this may be due to some non-linearity in the data. Additionally, we can notice the spread of residuals increase as the actual price increases (showing heteroscedasticity), meaning our model is not great at predicting expensive car prices.

```
## [1] "Lasso RMSE: 10565.9658579047"
```



In an alternative attempt to get a better predicting model of price, the next approach taken was with a non-linear model using a Random Forest.This will combine multiple decision trees in an effort to improve accuracy and account for outliers we're seeing in the linear model.

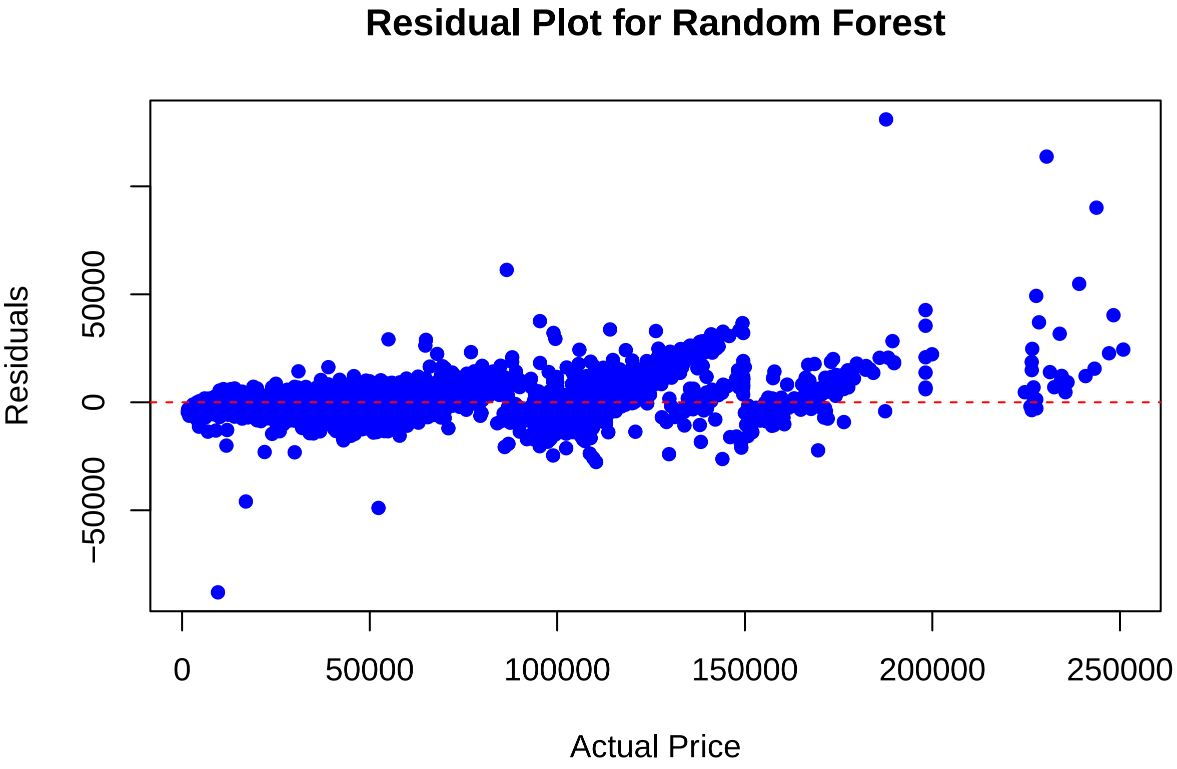
The random forests provides us with a feature importance score that lets us know that trim is the most important feature, followed by state (however from our linear model we see that only state being unspecified or SD plays any statistical significance), year, displacement, and finally mileage in determining a car's price. Intuitively, these features impacting price would all generally make sense.

The RMSE for the random forest was 6855 and when calculating the multiple-r squared value to compare it to our previous models, we get .9765. Meaning, with this model we can account for ~98% of the variability in price based on these 15 features and on average get that price right within $6900 (much better than our

$10k earlier from the linear model). Looking at the residual plot, we notice a more uniform spread, but the model still struggles with accurately predicting the higher priced vehicles. But overall, the random forest provided the best performance at predicting a car's price.

```
##                %IncMSE IncNodePurity
## X             1.373148  1.919395e+11
## trim         33.435783  3.843004e+12
## subTrim       8.524344  3.712593e+09
## condition    17.963513  6.963353e+12
## isOneOwner    9.031726  6.163395e+10
## mileage      25.273963  1.382298e+13
## year         32.161796  1.602483e+13
## color        19.980809  1.402999e+11
## displacement 31.028877  4.737586e+12
## fuel          8.758645  8.842881e+10
## state        33.044382  2.650199e+11
## region       21.741609  8.725120e+10
## soundSystem  13.063892  2.423409e+11
## wheelType    10.583246  2.837630e+10
## wheelSize    13.374145  3.766422e+11
## featureCount 20.612614  3.546533e+11
```

```
## [1] "Random Forest RMSE: 6855.24430003652"
```



```
## [1] "Random Forest Calculated R-squared: 0.9765"
```