Vehicle Analysis

Glorija

2024-07-02

Introduction

This is a presentation of my findings while investigating the dataset car_prices

The dataset has been cleaned and filtered:

- Removed irrelevant columns such as: trim, vin, body, state, interior, body, seller, and sell date;
- Cleared non-sensical data, missing values, and NAs;
- Made unique makes and colors uniform;

Data Summary

```
summary(data)
```

```
##       year          make              odometer         color
##  Min.   :1984   Length:511711      Min.   :     1   Length:511711
##  1st Qu.:2008   Class :character   1st Qu.: 28430   Class :character
##  Median :2012   Mode  :character   Median : 51838   Mode  :character
##  Mean   :2010                      Mean   : 67566
##  3rd Qu.:2013                      3rd Qu.: 97762
##  Max.   :2015                      Max.   :999999
##       mmr          sellingprice    transmission         condition
##  Min.   :    25   Min.   :     1   Length:511711      Min.   : 1.00
##  1st Qu.:  7300   1st Qu.:  7000   Class :character   1st Qu.:23.00
##  Median : 12300   Median : 12200   Mode  :character   Median :34.00
##  Mean   : 13806   Mean   : 13651                      Mean   :30.57
##  3rd Qu.: 18300   3rd Qu.: 18200                      3rd Qu.:41.00
##  Max.   :182000   Max.   :230000                      Max.   :49.00
```

```
head(data)
```

```
##   year   make odometer color   mmr sellingprice transmission condition
## 1 2015    kia    16639 white 20500        21500    automatic         5
## 2 2015    kia     9393 white 20800        21500    automatic         5
## 3 2014    bmw     1331  gray 31900        30000    automatic        45
## 4 2015  volvo    14282 white 27500        27750    automatic        41
## 5 2014    bmw     2641  gray 66000        67000    automatic        43
## 6 2015 nissan     5554  gray 15350        10900    automatic         1
```

```r
str(data)
```

```
## 'data.frame':    511711 obs. of  8 variables:
##  $ year        : int  2015 2015 2014 2015 2014 2015 2014 2014 2014 2014 ...
##  $ make        : chr  "kia" "kia" "bmw" "volvo" ...
##  $ odometer    : int  16639 9393 1331 14282 2641 5554 14943 28617 9557 4809 ...
##  $ color       : chr  "white" "white" "gray" "white" ...
##  $ mmr         : int  20500 20800 31900 27500 66000 15350 69000 11900 32100 26300 ...
##  $ sellingprice: int  21500 21500 30000 27750 67000 10900 65000 9800 32250 17500 ...
##  $ transmission: chr  "automatic" "automatic" "automatic" "automatic" ...
##  $ condition   : int  5 5 45 41 43 1 34 2 42 3 ...
```

EDA

- I used a correlation matrix to check numerical predictor relationship with the target and Linear Regression to assess the F and p values, as well as ANOVA test for categorical variables

```
## ### Correlation Matrix:
```

```
##                    year        mmr   odometer  condition sellingprice
## year          1.0000000  0.5930293 -0.7732631  0.3324428    0.5826934
## mmr           0.5930293  1.0000000 -0.5839388  0.2777864    0.9836835
## odometer     -0.7732631 -0.5839388  1.0000000 -0.3121395   -0.5783930
## condition     0.3324428  0.2777864 -0.3121395  1.0000000    0.3190660
## sellingprice  0.5826934  0.9836835 -0.5783930  0.3190660    1.0000000
```

```
## ### ANOVA:
```

```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## make              54 9.590e+12 1.776e+11  2466.7 <2e-16 ***
## color             18 1.106e+12 6.145e+10   853.5 <2e-16 ***
## transmission       2 1.245e+11 6.226e+10   864.8 <2e-16 ***
## Residuals     511636 3.683e+13 7.199e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## ###Linear Regression:
```

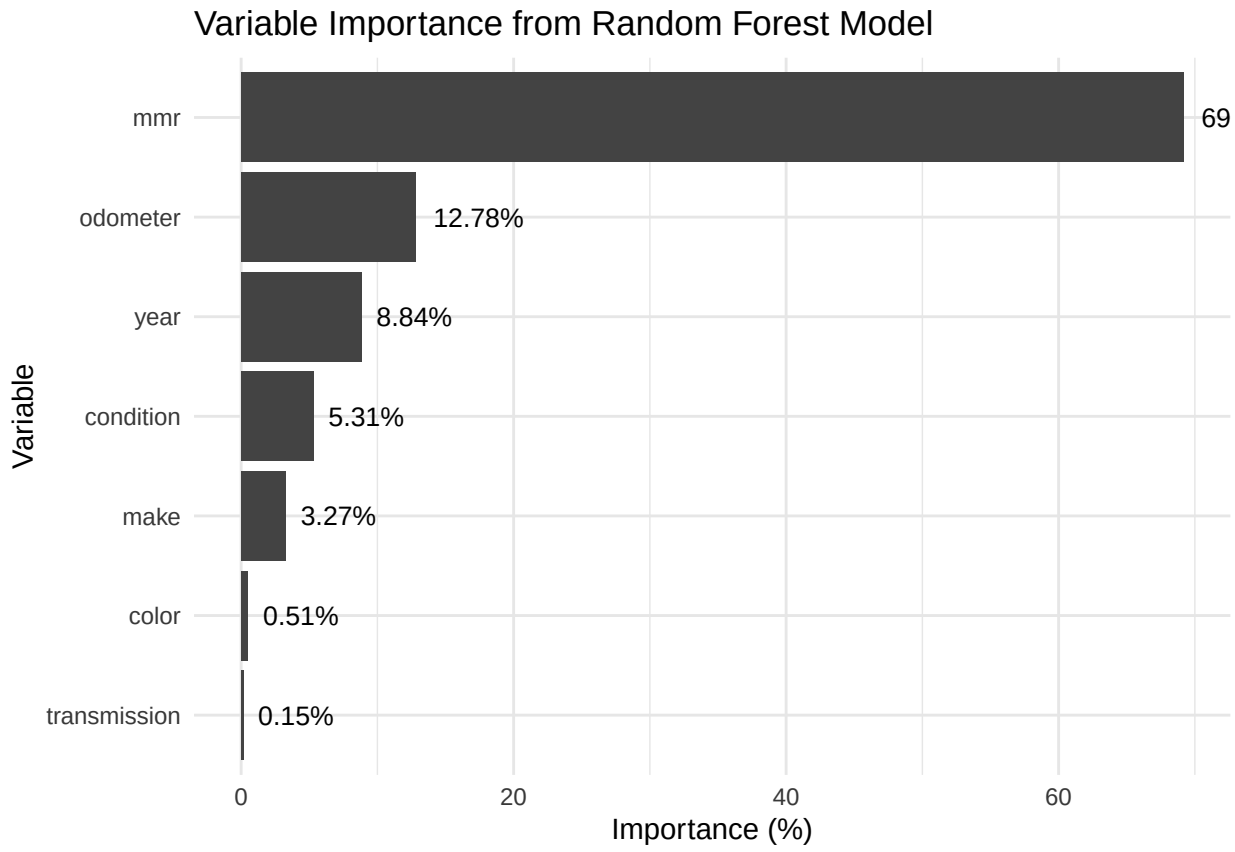
```
## ### Linear Regression:
```

```
##
## Call:
## lm(formula = sellingprice ~ year + mmr + odometer + condition,
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -86652   -668     24    760 207165
##
## Coefficients:
```

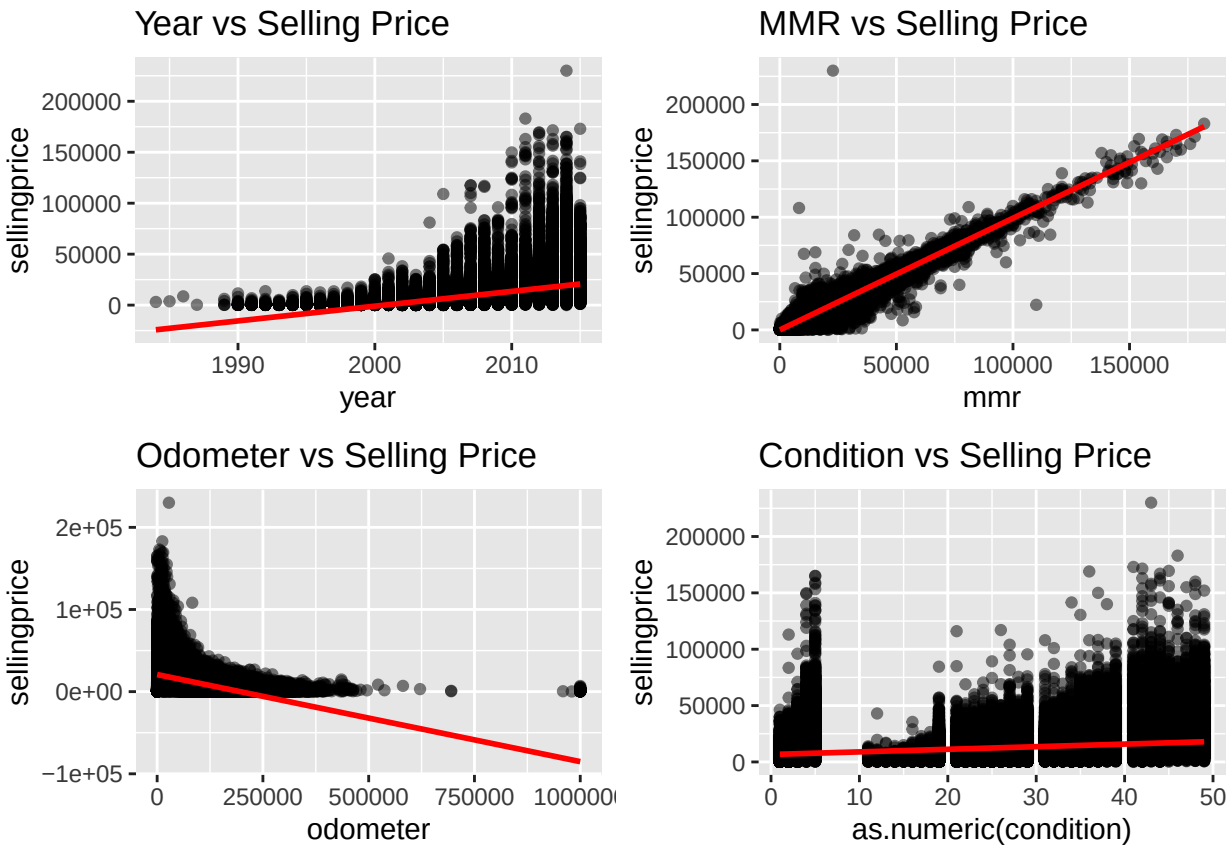
```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.195e+04  2.001e+03   45.96   <2e-16 ***
## year        -4.625e+01  9.946e-01  -46.49   <2e-16 ***
## mmr          9.843e-01  3.136e-04 3139.12   <2e-16 ***
## odometer    -1.146e-03  7.229e-05  -15.86   <2e-16 ***
## condition    3.750e+01  1.865e-01  201.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1670 on 511706 degrees of freedom
## Multiple R-squared:  0.9701, Adjusted R-squared:  0.9701
## F-statistic: 4.144e+06 on 4 and 511706 DF,  p-value: < 2.2e-16
```

Random Forest Model

- Correlation matrix and ANOVA didn't return desirable output so I used Random Forest for its robustness and ability to handle complex interactions between variables without requiring much preprocessing (for data with numerical and categorical variables such as this one)
- Random Forest accounts for interaction between variables rather just individual relationships between target and predictors, unlike in filter methods
- Random Forests are robust to overfitting, especially when dealing with large datasets, because they average the predictions of multiple trees, reducing variance
- Random Forest can model both linear and non-linear relationships by building multiple decision trees that split the data based on different predictor values. Each tree captures different aspects of the data, including linear and non-linear patterns
- I also displayed the importance of each predictor in predicting the target variable, and used Random Forest model as a variable selection method




```
## [1] "Random Forest R-squared: 0.991813039543864"

## [1] "Random Forest RMSE: 882.863389011979"

## [1] "Random Forest MAE: 536.206118764424"

## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```




```
##      year       mmr  odometer condition
##  2.731625  1.653832  2.659237  1.142431
```

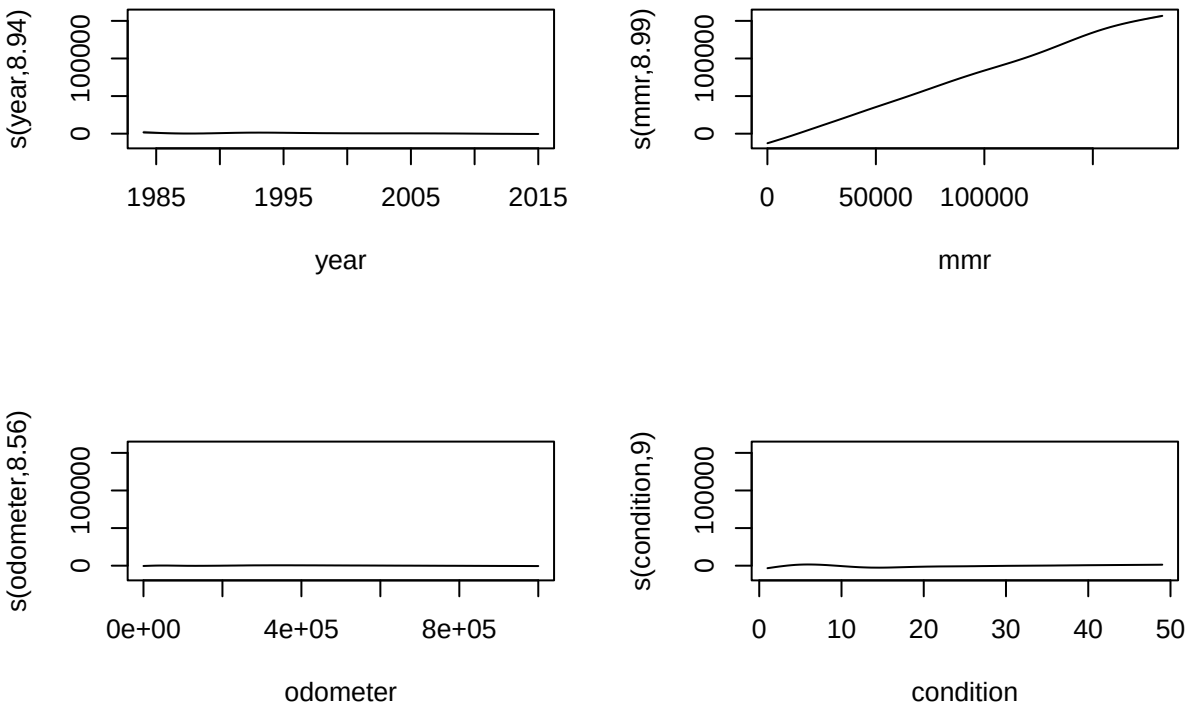
- I used Scatter plots to visualize the relationship between numerical features (year, mmr, odometer, condition) and the selling price -I added a linear regression line to show the trend between the variables

GAM

- Since 3 of the 4 predictors had non-linear relationships with the target, I chose GAM model for its flexibility in capturing the non-linear relationships
- GAMs extend linear models by allowing non-linear functions of predictors. They use smooth functions (splines) to model the relationship between each predictor and the target variable, accommodating non-linearity
- GAMs fit each predictor's effect separately, which can be advantageous when different predictors have different types of non-linear relationships with the target variable

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sellingprice ~ s(year) + s(mmr) + s(odometer) + s(condition)
##
## Parametric coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13650.571      2.166    6303   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df         F p-value
## s(year)      8.943  8.998     477.0  <2e-16 ***
## s(mmr)       8.989  9.000 1155913.2  <2e-16 ***
## s(odometer)  8.562  8.942     386.7  <2e-16 ***
## s(condition) 9.000  9.000   14006.2  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.974   Deviance explained = 97.4%
## GCV = 2.4001e+06  Scale est. = 2.3999e+06  n = 511711
```




```
## [1] "GAM R-squared: 0.974231613527891"

## [1] "GAM RMSE: 1549.11954467324"

## [1] "GAM MAE: 966.604446985693"
```

Random Forest Cross-Validation

- I used k-Fold Cross-Validation to assess the model's performance by dividing the data into k subsets, training on k-1 subsets, and testing on the remaining subset
- Example usage with different k values (k_values <- c(5, 10, 15)), has proven that 10-Fold Cross-Validation provided a good balance between bias and variance in the performance estimate

```
## [1] "Average RMSE from k-fold cross-validation (RF): 1548.74155361922"
```

```
## [1] "Average R-squared from k-fold cross-validation (RF): 0.974209254408973"
```

```
## [1] "Average MAE from k-fold cross-validation (RF): 971.891805997841"
```

GAM Cross-Validation

```
## [1] "Average RMSE from k-fold cross-validation (GAM): 1547.79567585896"
```

```
## [1] "Average R-squared from k-fold cross-validation (GAM): 0.974230090775839"
```

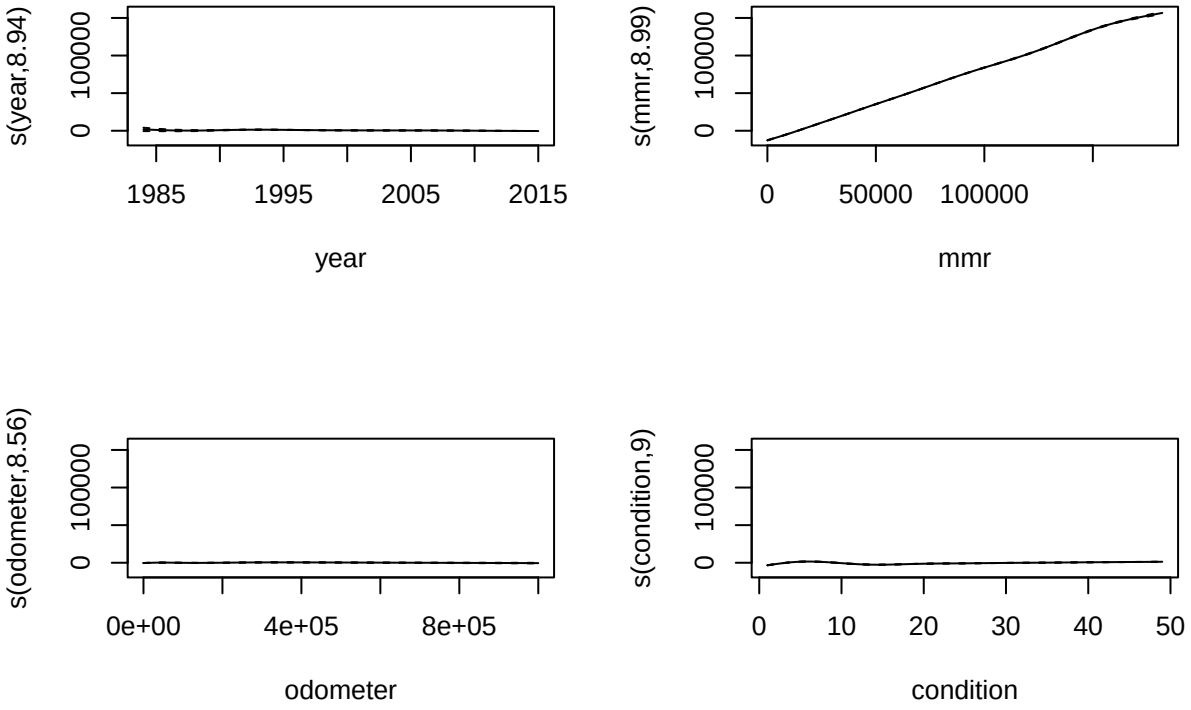
```
## [1] "Average MAE from k-fold cross-validation (GAM): 966.732598882381"
```

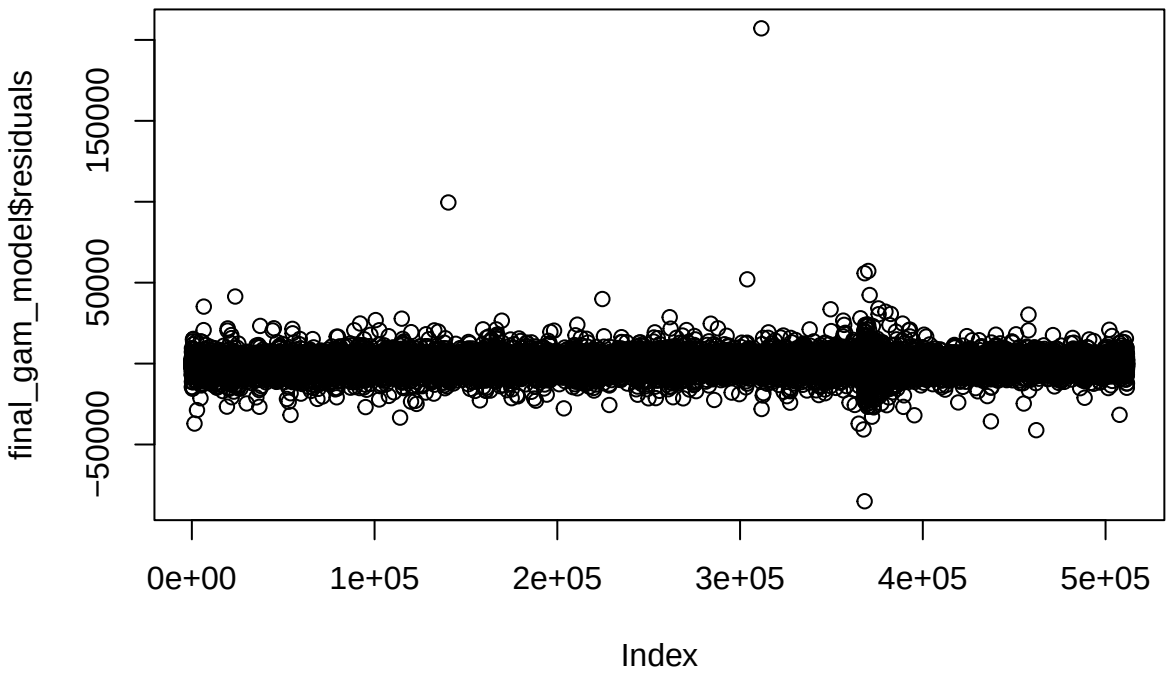
Conlusion

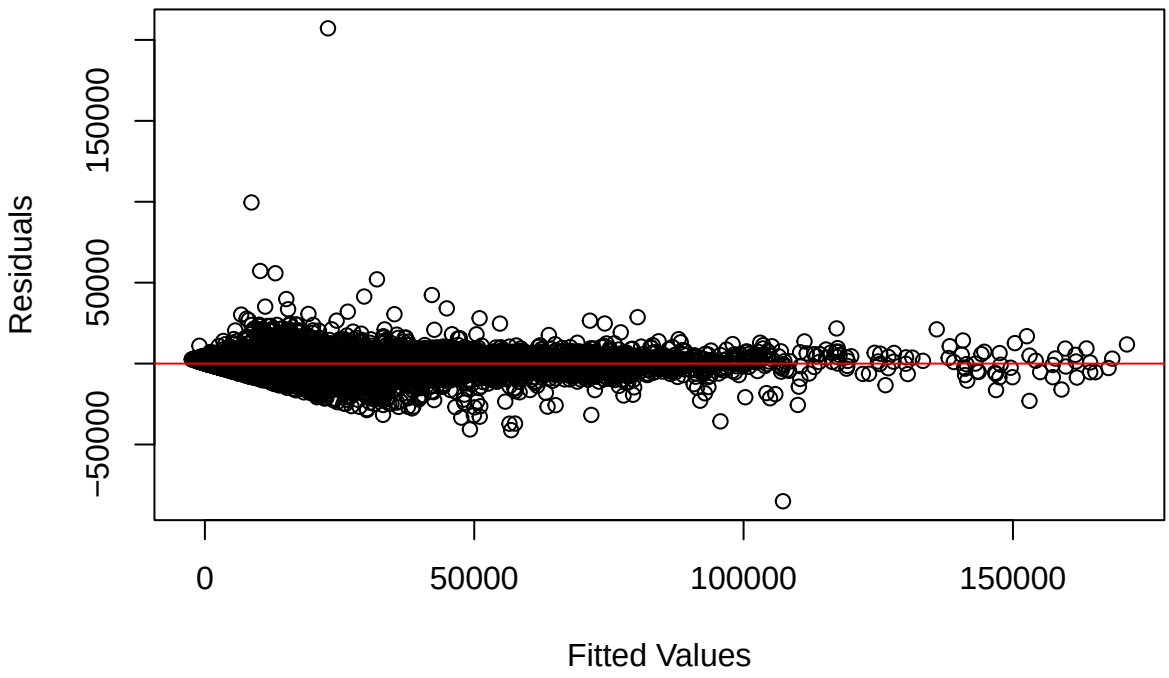
```
##
## Model Selection Summary:

## 1. Both GAM and Random Forest models show
##     comparable performance in predicting vehicle selling prices.

## 2. The GAM model has a slightly lower RMSE ( 1547.796 )
##     compared to the Random Forest model ( 1548.742 ).

## 3. The R-squared for GAM ( 0.9742301 ) is also
##     slightly higher than for Random Forest ( 0.9742093 ).

## 4. The MAE for GAM ( 966.7326 ) is
##     comparable to that for Random Forest ( 971.8918 ).

## 5. Given the small difference in RMSE and R-squared, and
##     considering factors like interpretability and computational
##     efficiency, the GAM model is preferred for its simplicity
##     and better performance.
```

Model Fitting

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## sellingprice ~ s(year) + s(mmr) + s(odometer) + s(condition)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13650.571      2.166    6303   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df         F p-value
## s(year)      8.943  8.998     477.0  <2e-16 ***
## s(mmr)       8.989  9.000 1155913.2  <2e-16 ***
## s(odometer)  8.562  8.942     386.7  <2e-16 ***
## s(condition) 9.000  9.000   14006.2  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.974   Deviance explained = 97.4%
## GCV = 2.4001e+06  Scale est. = 2.3999e+06  n = 511711
```




final_gam_model$residuals
−50000
50000
150000
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Index




```
## [1] "RMSE on the full dataset: 1549.12"
```