UNIVERSITY OF
MARYLAND

ROBERT H. SMITH
SCHOOL OF BUSINESS

POLAND CARS FOR SALE

TREVOR DRAKE, STEFFANY FLORES, KEHAN LI,
JASON TAYLOR, AND RUBY TUCKER

Introduction

The following report details a comprehensive analysis of a dataset documenting cars available for purchase in Poland. It was created by web scraping over 200,000 car offers from one of Poland's largest car advertisement sites. The dataset contains 208,304 observations of 25 variables. Some variables included are price, condition, production_year, mileage_km, power_hp, and features. The objective of this analysis was to identify key variables that exert the greatest influence on car pricing and understand how they contribute to price variation.

Initial Modeling Efforts and Results

Comprehensive Model

The comprehensive linear regression model initially included all predictor variables and yielded an R^2 value of 0.81 *(Figure 1)*, indicating that 81% of the variance in car prices could be explained by the model.

Simplified Model

After narrowing down to a simpler model, it was found that a few variables were strong predictors of price. Our model narrowed down the most influential predictors to `Power_HP` and `Production_Year` as predictor variables. After running the model only considering these 2 variables, the R^2 dropped to 0.70 *(Figure 1)*, signifying that only 11% of the variance explanation was lost. The results of the final simplified model are seen in Appendix A of this report.

Technical Interpretation

- Power_HP (Horsepower): Higher horsepower is associated with significantly higher car prices. With a standardized coefficient of `54,050.62`, this variable plays a dominant role in predicting the price.
- Production_Year: Newer cars are also strongly correlated with higher prices, as evidenced by a standardized coefficient of `24,373.04`.

Multicollinearity Check

- After checking for multicollinearity, we found that `Power_HP` and `Production_Year` exhibited moderate to high correlations with many other predictor variables, underscoring their influence.

Pivoting from the Initial Approach
Our initial model was one of inference which was aimed to inform management on which variables were most influential towards price. The model revealed that `Power_HP` and `Production_Year` alone accounted for a significant portion of price variance. Using our intuition, we began to realize that this finding was common knowledge, and we needed to dig deeper to find more useful, actionable insights. However, with our findings we still have some implications that can be useful to know.

Despite these findings aligning with intuition, the insights offer practical value in the following way:

Targeted Sales Strategy

- ***Educating Sales Teams***: Sales teams can leverage this information to understand the primary drivers of a car's worth. Training should emphasize the role of horsepower and production year in determining price.
- ***Price Justification***: When negotiating with customers, salespeople can justify higher prices by pointing to a car's production year and horsepower as significant value drivers.
- By illustrating how important these factors are in determining a car's worth, salespeople will be able to play an integral role in maximizing profit margins.

Although our findings reiterate common-sense information, integrating our price model findings will improve sales training, allowing salesmen at car reselling services and dealerships to make more informed decisions. This improvement in negotiation ability will enhance financial performance through margin optimization.

Following our initial discoveries, it became evident that the variables with the most significant impact on car prices, like horsepower and production year, are inherently intuitive. Consequently, we recognized the need to delve deeper into the dataset to unearth novel insights worthy of recommendation. In response to this realization, we constructed ten binary variables corresponding to ten distinct features that we hypothesized could influence pricing. A value of 1 was assigned if the car exhibited the desired feature. The ten features selected were: rear camera, lane assistant, alloy wheels, LED lights, bluetooth, heated rear seats, automatic AC, GPS, twilight sensors and blind spot sensors (BSS).

Analysis

Random Forest Model
First, we employed a random forest model with price as the dependent variable and all ten features as the predictor variables. We set the seed to 123 and the number of trees to 50. This model explained 24.35% of the variability in car prices (*Figure 2)*. While this is significantly lower than our original model, it offers greater utility as it provides insights into the relative importance of each of the ten features. Looking at the plot for variable importance, we found that the most important feature was the rear camera. In fact, if the rear camera had not been included in the model, the accuracy would have decreased by over 35%. The subsequent nine variables, ranked in descending order of importance, were as follows: lane assistant, alloy wheels, LED lights, bluetooth, heated rear seats, automatic AC, GPS, twilight sensors, and BSS (*Figure 2)*.

We then ran another random forest model removing the three least important variables, twilight sensor, GPS, and bluetooth to reduce overfitting and better understand the variation in price that the top seven variables contribute to. This model explained 23.48% of the variance, showing that the removal of the three least important features didn't have a significant effect on the variance explained by the model. Again, the rear camera was the most impactful feature, followed by alloy wheels, LED lights, lane assistant, BSS, heated rear seats, and automatic AC.

Forward-stepwise Regression Model
Given our understanding of the relative importance of each feature, we then ran two regression models to quantify the precise impact of each feature on car prices. We set price as the dependent variable and car features, including rear camera, heated seats, GPS, Bluetooth, BSS, automatic AC, alloy wheels, twilight sensor, lane assistant, and LED lights as independent variables. To find out which features should be included in the regression model for predicting the car price, we ran the forward-stepwise regression Model.

At the beginning, with 0 variables, the CP value is at its peak around 29787.731. As more variables are added, the CP value decreases, improving overall model performance. By step 11, the model includes 10 variables with the lowest Cp of 11.00, suggesting that 10 variables should be selected to gain a good balance of complexity and performance *(Figure 6*) for the regression model. This also means that all features should be included in the linear model.

Full model search

As in the forward-stepwise regression model, we chose the same dependent and independent variables to run the model. The result also suggests that the CP value is the lowest (11) if all 10 variables are included *(Figure 5).*

Linear regression model

Based on the previous two regression models, we include all features to predict the car price. The model explains approximately 24.07% of the variation in car prices, and the model has a p-value of 2.2e-16, meaning the model is significant *(Figure 7)*. Below are the findings based on the output:

Positive Influences

Features such as rear cameras, heated seats, GPS, Bluetooth, blind spot sensors, automatic air conditioning, lane assistants, and LED lights can significantly increase car prices. Specifically, adding a rear camera can increase the car price by an average of $39,518. Similarly, installing heated seats adds around $25,854, while adding a lane assistant can increase prices by $29,341. GPS systems raise the price by about $4,691, Bluetooth functions can boost the price by roughly $3,124, BSS can increase the price by $23,828, automatic air conditioning can increase by $8,006, and finally, LED lights can add $22,355.

Negative Influences

Alloy wheels and twilight sensors negatively influence car prices, with alloy wheels significantly decreasing the car's price by about $13,418, while twilight sensors will reduce the price by $3,566.

Recommendations

Focus on High-Impact Features

Features including lane assistants, rear cameras, BSS, heated seats, and LED lights can significantly increase car prices. Car resellers should prioritize these features to enhance market competitiveness and profitability.

It is also recommended that the sales team be trained to acknowledge the specific benefits and functionality of features like lane assistants, rear cameras, BSS, heated seats, and LED lights. Understanding how these features improve safety, comfort, and the overall driving experience can help the team communicate effectively with potential customers.

Segmentation

The sales team should also customize offers based on different customer segments. Offer virtual or in-person demonstrations specifically tailored to highlight the features each customer segment values. For example, individuals who prioritize safety and are interested in technology focus on demonstrating advanced safety technologies like lane assistants, BSS, and rear cameras.

Similarly, for customers who value comfort, the sales team can highlight the luxury and convenience features of the vehicles. Offering personalized demonstrations emphasizing comfort-enhancing features, such as heated seats and LED lights, can enhance the driving experience, especially in colder climates or during nighttime.

For features that harm car prices, such as alloy wheels and twilight sensors, the sales team should explore options to enhance the appeal of these features or consider strategies to decrease their costs, which could help boost profit.

Marketing and Promotions

The marketing team can develop offers focusing on upgrading to premium models with these features. This could include limited-time discounts on the cars including these features or give attractive financing options that make the upgrades more affordable. For example, develop limited-time promotional offers that significantly discount cars featuring add-ons like lane assistants, rear cameras, and heated seats. These limited-time discounts encourage customers to purchase while the offer lasts.

Limitations

The dataset used in our analysis is geographically limited, consisting solely of data from cars sold in Poland. This limitation may affect the generalizability of our findings to car resellers in the U.S. market, where the environment and consumer preferences can be very different. Also, the price variable in our dataset demonstrates a left-skewed distribution, likely due to a lack of luxury cars in the region, which may skew average pricing perceptions downward. To enhance the validity and applicability of our conclusions, future studies should examine car data from the U.S., which features a broader variation in pricing.

Appendix:

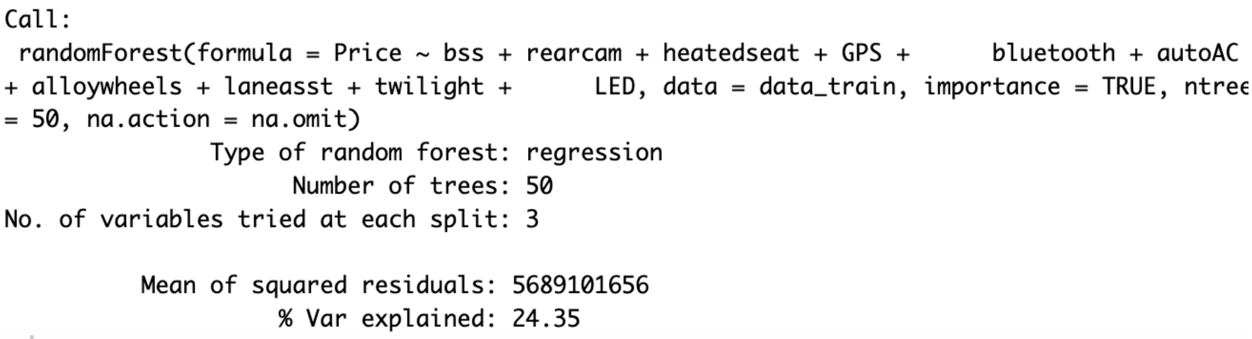
Figure 1 : Results of the linear regression model including Power_HP and Production_year as predictor variables of ‘Price’.

```
 OLS Regression Results (Simplified Model)
Dep. Variable:                 Price   R-squared:                       0.703
Model:                           OLS   Adj. R-squared:                  0.703
Method:                Least Squares   F-statistic:                 1.503e+04
Date:               Thu, 09 May 2024   Prob (F-statistic):               0.00
Time:                       15:10:21   Log-Likelihood:            -1.5323e+05
No. Observations:              12714   AIC:                         3.065e+05
Df Residuals:                  12711   BIC:                         3.065e+05
Df Model:                          2
Covariance Type:           nonrobust
==============================================================================
=================
                    coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const           7.323e+04    368.013    198.994      0.000    7.25e+04     7.4e+04
Production_year  2.437e+04    376.477     64.740      0.000    2.36e+04    2.51e+04
Power_HP        5.405e+04    376.477    143.570      0.000    5.33e+04    5.48e+04
==============================================================================
============
Omnibus:                    13036.025   Durbin-Watson:                   1.690
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2804026.851
Skew:                           4.629   Prob(JB):                         0.00
Kurtosis:                      75.162   Cond. No.                         1.24
==============================================================================
============
```

Most Influential Predictors (Standardized Coefficients):

```
Power_HP           54050.621666
Production_year    24373.045179
```

Figure 2: Random Forest Model with all 10 Variables



```
Call:
 randomForest(formula = Price ~ bss + rearcam + heatedseat + GPS +      bluetooth + autoAC
+ alloywheels + laneasst + twilight +      LED, data = data_train, importance = TRUE, ntree
= 50, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 3

          Mean of squared residuals: 5689101656
                    % Var explained: 24.35
```

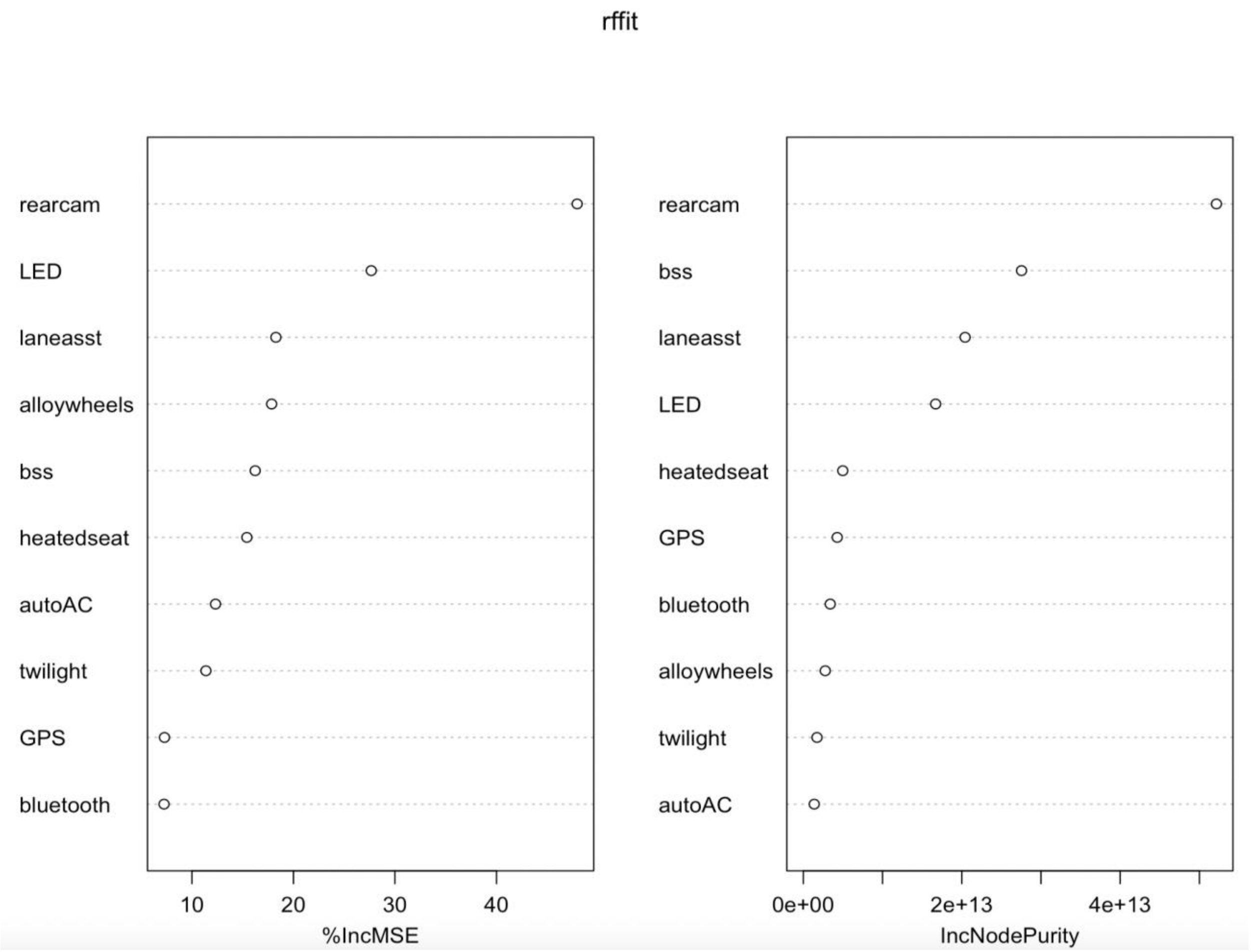
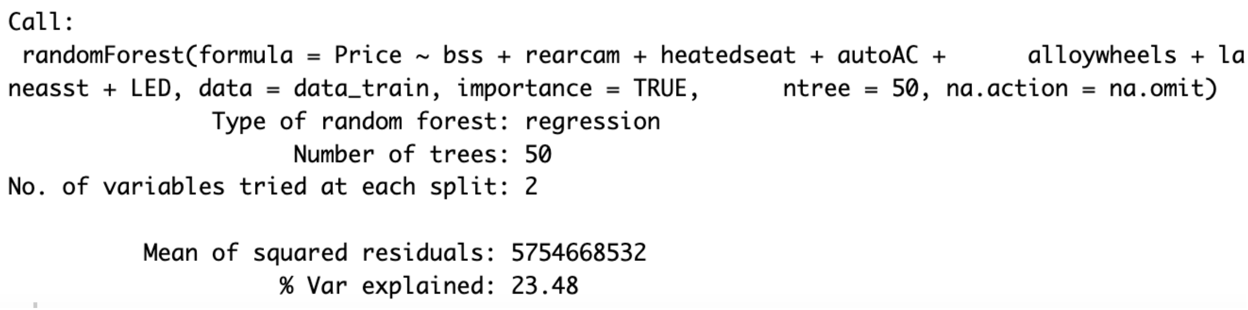


Figure 2 also features the variable importance plot for this model showing the relative importance of each feature.

Figure 3: Random Forest Model with 7 Variables



```
Call:
 randomForest(formula = Price ~ bss + rearcam + heatedseat + autoAC +      alloywheels + la
neasst + LED, data = data_train, importance = TRUE,      ntree = 50, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 2

          Mean of squared residuals: 5754668532
                    % Var explained: 23.48
```

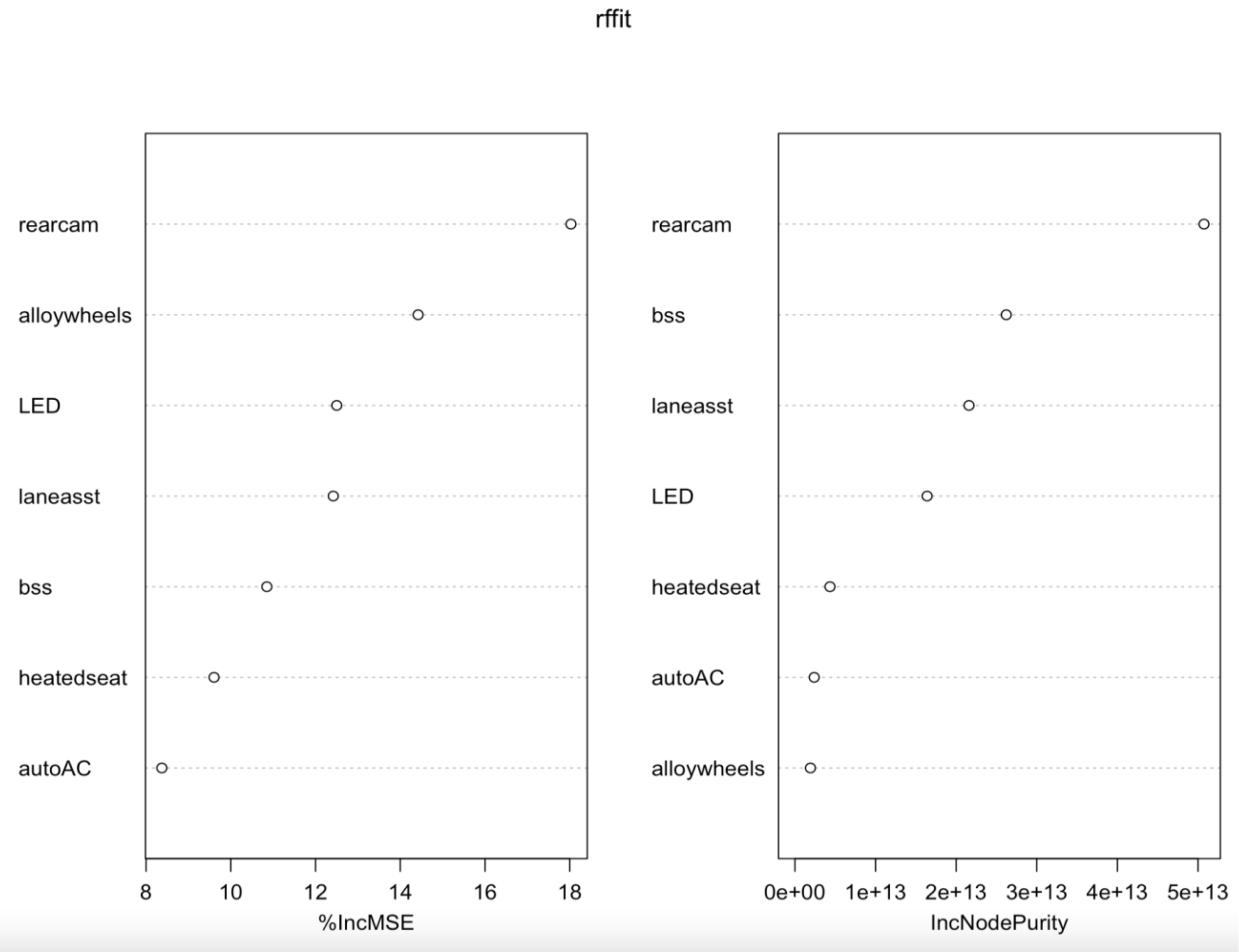
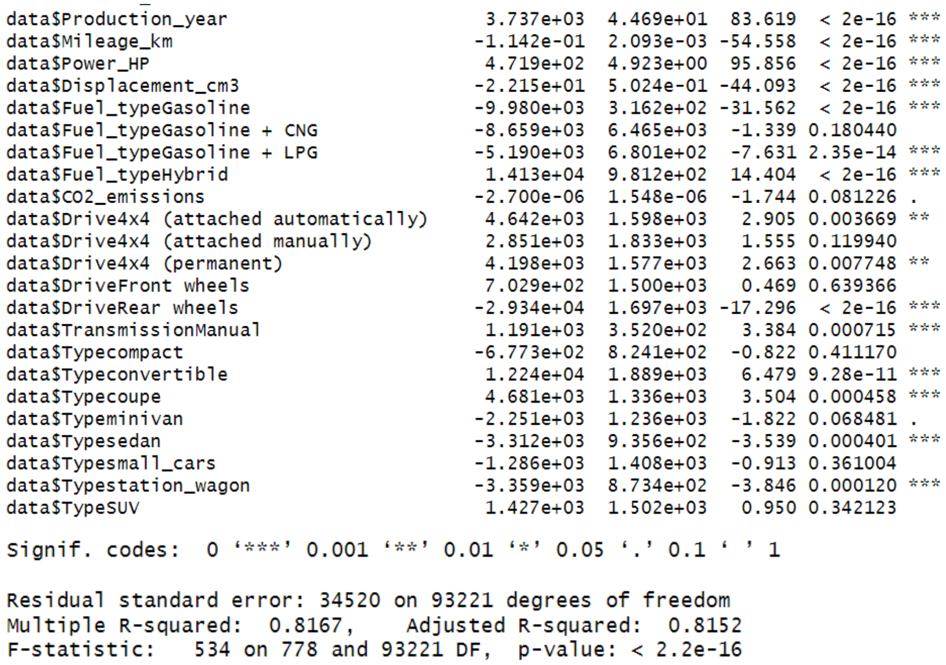


Figure 3 also features the variable importance plot for this model showing the relative importance of each feature.

Figure 4: Comprehensive Model



```
data$Production_year                    3.737e+03  4.469e+01  83.619  < 2e-16 ***
data$Mileage_km                        -1.142e-01  2.093e-03 -54.558  < 2e-16 ***
data$Power_HP                           4.719e+02  4.923e+00  95.856  < 2e-16 ***
data$Displacement_cm3                  -2.215e+01  5.024e-01 -44.093  < 2e-16 ***
data$Fuel_typeGasoline                 -9.980e+03  3.162e+02 -31.562  < 2e-16 ***
data$Fuel_typeGasoline + CNG           -8.659e+03  6.465e+03  -1.339 0.180440    
data$Fuel_typeGasoline + LPG           -5.190e+03  6.801e+02  -7.631 2.35e-14 ***
data$Fuel_typeHybrid                    1.413e+04  9.812e+02  14.404  < 2e-16 ***
data$CO2_emissions                     -2.700e-06  1.548e-06  -1.744 0.081226 .  
data$Drive4x4 (attached automatically)  4.642e+03  1.598e+03   2.905 0.003669 ** 
data$Drive4x4 (attached manually)       2.851e+03  1.833e+03   1.555 0.119940    
data$Drive4x4 (permanent)               4.198e+03  1.577e+03   2.663 0.007748 ** 
data$DriveFront wheels                  7.029e+02  1.500e+03   0.469 0.639366    
data$DriveRear wheels                  -2.934e+04  1.697e+03 -17.296  < 2e-16 ***
data$TransmissionManual                 1.191e+03  3.520e+02   3.384 0.000715 ***
data$Typecompact                       -6.773e+02  8.241e+02  -0.822 0.411170    
data$Typeconvertible                    1.224e+04  1.889e+03   6.479 9.28e-11 ***
data$Typecoupe                          4.681e+03  1.336e+03   3.504 0.000458 ***
data$Typeminivan                       -2.251e+03  1.236e+03  -1.822 0.068481 .  
data$Typesedan                         -3.312e+03  9.356e+02  -3.539 0.000401 ***
data$Typesmall_cars                    -1.286e+03  1.408e+03  -0.913 0.361004    
data$Typestation_wagon                 -3.359e+03  8.734e+02  -3.846 0.000120 ***
data$TypeSUV                            1.427e+03  1.502e+03   0.950 0.342123    

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34520 on 93221 degrees of freedom
Multiple R-squared:  0.8167,	Adjusted R-squared:  0.8152 
F-statistic:   534 on 778 and 93221 DF,  p-value: < 2.2e-16
```

Figure 5: Full Model Search

```
> out <- leaps(x, y, method="Cp", nbest=1)
> print(out)
$which
      1     2     3     4     5     6     7     8     9     A
1  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
2  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
3  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
4  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
5  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE
6  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
7  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
8  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
9  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
10 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           
"5"          
 [7] "6"           "7"           "8"           "9"           "A"          

$size
 [1]  2  3  4  5  6  7  8  9 10 11

$Cp
```

```
 [1] 9186.22895 4066.27469 2122.87366 1180.95677  615.35141  323.94524
117.11683   55.28695
 [9]   35.58218   11.00000
```

Figure 6: Forward-Stepwise Regression

```
> #Forward-stepwise regression
> library(lars)
Loaded lars 1.3

> res = lars(x, y, type="stepwise")
> print(summary(res))
LARS/Forward Stepwise
Call: lars(x = x, y = y, type = "stepwise")
   Df        Rss        Cp
0   1 6.0611e+14 29787.731
1   2 5.0522e+14  9186.229
2   3 4.8014e+14  4066.275
3   4 4.7062e+14  2122.874
4   5 4.6681e+14  1346.947
5   6 4.6322e+14   615.351
6   7 4.6178e+14   323.945
7   8 4.6076e+14   117.117
8   9 4.6045e+14    55.287
9  10 4.6034e+14    35.582
10 11 4.6021e+14    11.000
```

Figure 7: Linear Regression Model

```
> summary(model)

Call:
lm(formula = Price ~ rearcam + heatedseat + GPS + bluetooth + 
    bss + autoAC + alloyWheels + twilightSensor + laneAssistant + 
    ledLights, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-169897  -24966  -10747    5353 2311064 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     34244.7      499.7  68.536  < 2e-16 ***
rearcam         39518.4      683.2  57.840  < 2e-16 ***
heatedseat      25854.5     1083.2  23.868  < 2e-16 ***
GPS              4690.6      609.8   7.692 1.46e-14 ***
bluetooth        3123.8      605.9   5.156 2.53e-07 ***
bss             23828.3      883.0  26.987  < 2e-16 ***
```

```
autoAC           8006.2      601.0  13.321  < 2e-16 ***
alloyWheels    -13418.1      624.8 -21.477  < 2e-16 ***
twilightSensor  -3566.1      618.1  -5.769 8.00e-09 ***
laneAssistant   29341.6      830.2  35.344  < 2e-16 ***
ledLights       22355.0      610.0  36.648  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69970 on 93989 degrees of freedom
Multiple R-squared:  0.2407,  Adjusted R-squared:  0.2406
F-statistic:  2980 on 10 and 93989 DF,  p-value: < 2.2e-16
```