The Analysis of Statistics Models of Predicting Prices of Used Cars

Shankai Liao, Xing Xin, Yiwen Wang

1. Abstract

Building the price prediction model for used cars is one heated topic in economic studies. The price of used cars predicted by certain models is an important reference for vehicles customers. For this report, we chose three typical enterprises, Audi, Bayerische Motoren Werke AG(BMW), and Ford and we also chose variables including model, mileage, and mpg which predict the prices of used cars. We used Linear Regression models, Lasso models, Random Forest, and Boosting to predict the price based on variables of the dataset. By comparing metrics such as root mean square error (RMSE), we concluded that the Random Forest is the best model and there were several variables including model, mpg, and year which customers should consider when they buy used cars and identify different car prices.

2. Introduction

New car sale prices are set by the manufacturer, so their prices are consistent with their actual market value. However, prices in the used car market are set by the dealer, which is a different story. Despite an unprecedented period of growth for the new car market in the United Kingdom, the used car market has consistently been more valuable with a much higher number of vehicles sold. Recently, used-car prices appear to be stuck in high gear, despite slowing consumer demand.

We studied the used car market in the UK to predict the price based on the dataset containing information about three typical car brands Audi, BMW, and Ford. The data comes from Kaggle, a subsidiary of Google LLC, which is an online community of data scientists and machine learning practitioners. Firstly, we visualized the data by giving it visual context through maps or graphs after collecting the data of three enterprises. Then we used the Linear Regression model, Lasso model, Random Forest model, and Boosting model to fit the data to predict the price based on certain variables. Finally, we evaluated the difference by comparing the rmse and we got the conclusion of the four models' comparison and important variables in this process.

3. Dataset and Data Visualization

3.1 Data Introduction

The data describes used car listings, which have been separated into files corresponding to each car manufacturer. We aimed to predict how many old cars should be sold with different features containing information on the price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size.

Thoroughly, the unit of the price is the Great Britain Pound. The car transmission comes in two types: manual and automatic. The sign of the mileage can indicate if the car has been driven a lot but will also refer to if the odometer gives an accurate reading. The fuel type is divided into diesel and petrol which are produced from mineral oil, but the precise refining methods vary. Diesel is in principle easier to refine than gasoline. Per liter, diesel contains more energy than petrol and the vehicle's engine combustion process is

more efficient, adding up to higher fuel efficiency and lower carbon dioxide emissions when using diesel. The car's mpg – which stands for "miles per gallon" – denotes the number of miles it travels on one gallon of gas. The engine size is referred to as 'engine capacity' or 'engine displacement' and is the measurement of the total volume of the cylinders in the engine. The bigger the engine size, the more space there is for air and fuel inside it. As a larger engine is usually able to burn more fuel and produce more power, a car with a larger, more powerful engine is likely to be able to accelerate faster and tow heavier loads than a car with a smaller engine can manage.

3.2 Data Visualization

The data visualization part, firstly, according to Figure 3.2.1, it indicates the expected relationship between price and the mileage for Audi and BMW. For Audi cars, the general trend: as the vehicles have higher mileage, the price will be lower. It meets the common things of used cars. In addition, we also found that the engine size is not a distinctive factor of the price and in this case, we can say that the mileage is an explainable factor for the price of the vehicles. It has a similar trend in BMW and we can find that the price is very low when the mileage is very large clearly.

Next, we discuss the relationship between price and the mpg of Audi. According to Figure 3.2.2, the price is higher for the vehicles whose mpg is lower than Audi cars. It's obvious to find out that the relationship between the mpg and the price is negatively related. What's interesting to find the difference from the previous graph is that the year factor didn't influence a lot on the mpg. For very early registered vehicles, they may have very low mpg. From those 2 graphs, we can conclude that the mileage and mpg factor is 2 explainable factors for the price of used cars.

Figure 3.2.1 Data Visualization: price and mileage

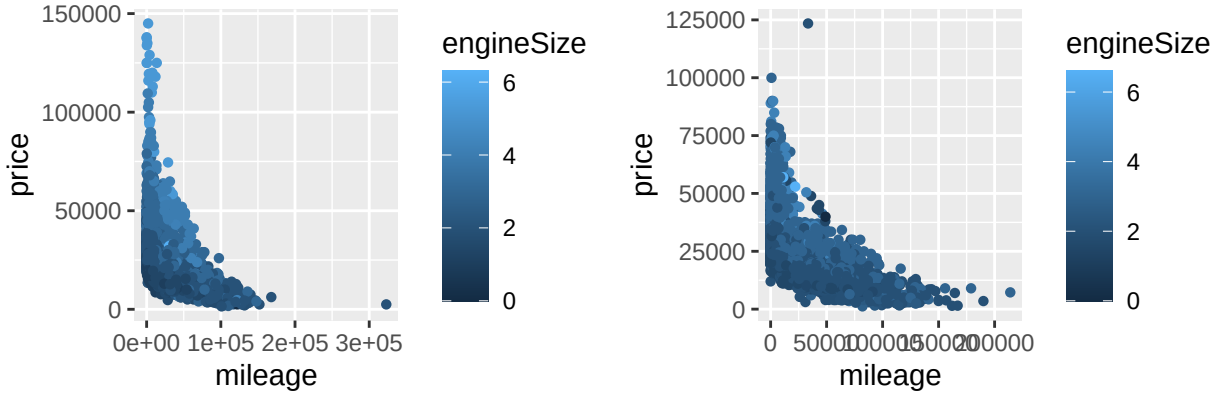
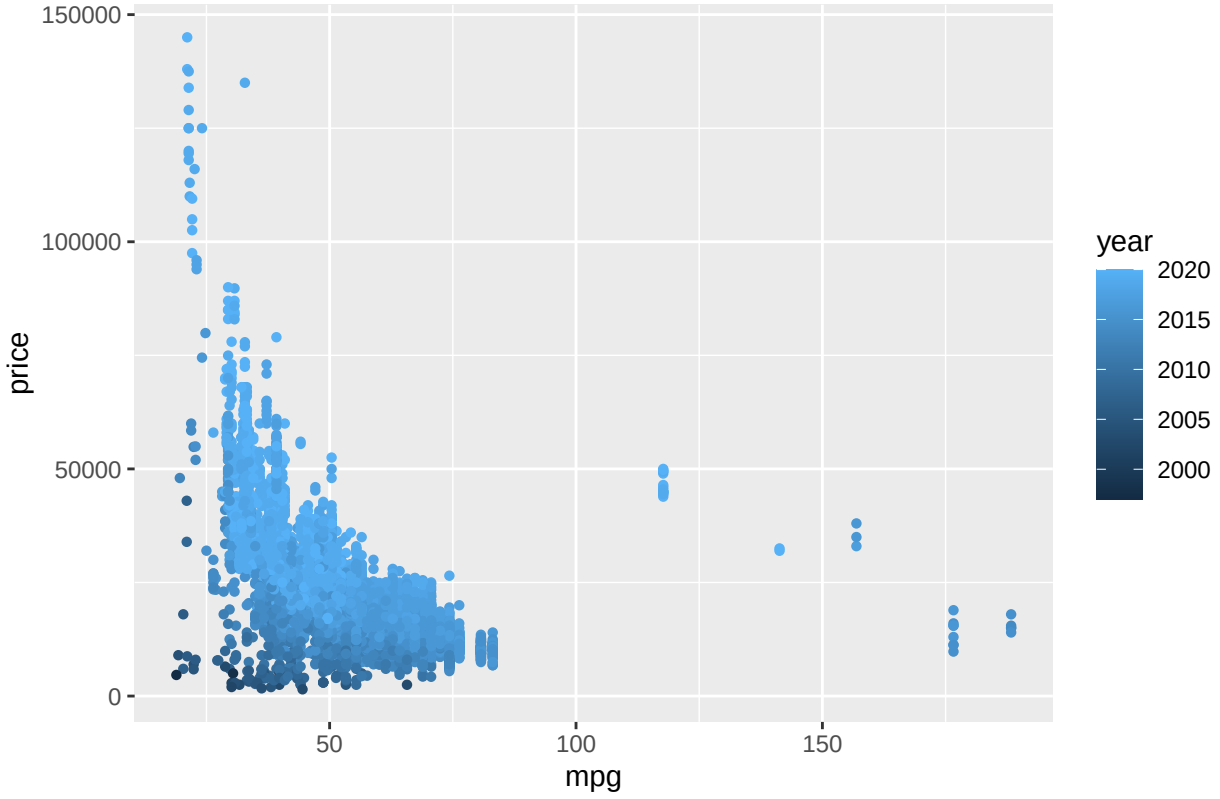


Figure 3.2.2 Data Visualization: price and model



4. Methods

For our methods, we decided to split the data for 90% of training and 10% of testing and we used four models to predict used car prices of Audi, BMW, and Ford. The first is the Linear Regression model which includes variable interactions. The second is the Lasso model which introduces the best lambda for the variables part. Another two models are tree models: Random Forest and Boosting. These two models can capture the nonlinear variables in this process. Finally, we compared the rmse of four models in the Results part.

4.1 Linear Regression Models

Firstly, we conduct the Linear Regression to explore the relationship between the dependent variable of price and the independent variables of features of cars even though not all factors in the data set greatly impact the price change. We examine the coefficients which tell us how much to adjust the trajectory of our predicted price. According to Table 4.1, for example, if we look at different generations of the Audi A series, the latest generation of the model has a more positive influence on the price. Specifically, the coefficients of the models A2-A7 are from around 1.9-4.7 but the coefficient of the model A8 reaches 8, which means the model A8 could increase the additional value of about 8 units in the price, holding all other variables unchanged.

In addition, according to Table 4.1, other coefficients like mileage, mpg, tax, and petrol are negative. This can be explained that if one unit increases in this variable, the price of used cars will decrease to certain units. Like the increase in mileage can predict that the price of used cars will decrease in this process. This is one strength of Linear Regression and can help people identify the cause and effect between price and certain variables. However, we should consider the interaction between certain variables in the Linear Regression

Table 1: **Table 4.1 : The Coefficient of Linear Regression on Audi Price **

	.
(Intercept)	-3.590364e+06
model A2	1.890489e+04
model A3	1.309216e+03
model A4	1.603261e+03
model A5	3.059020e+03
model A6	3.554403e+03
model A7	4.686925e+03
model A8	7.992615e+03
model Q2	1.459974e+03
model Q3	2.937488e+03
model Q5	6.698870e+03
model Q7	1.501591e+04
model Q8	2.501949e+04
model R8	5.789815e+04
model RS3	9.852357e+03
model RS4	2.174984e+04
model RS5	1.912359e+04
model RS6	2.727105e+04
model RS7	1.922542e+04
model S3	5.095630e+03
model S4	9.515788e+03
model S5	2.076478e+03
model S8	9.824044e+03
model SQ5	1.007753e+04
model SQ7	1.932750e+04
model TT	3.403050e+03
year	1.795955e+03
transmissionManual	-1.485948e+03
transmissionSemi-Auto	1.416251e+02
mileage	-8.081020e-02
fuelTypeHybrid	3.373368e+04
fuelTypePetrol	-9.075075e+02
tax	-3.004312e+01
mpg	-2.925126e+02
engineSize	4.526459e+03

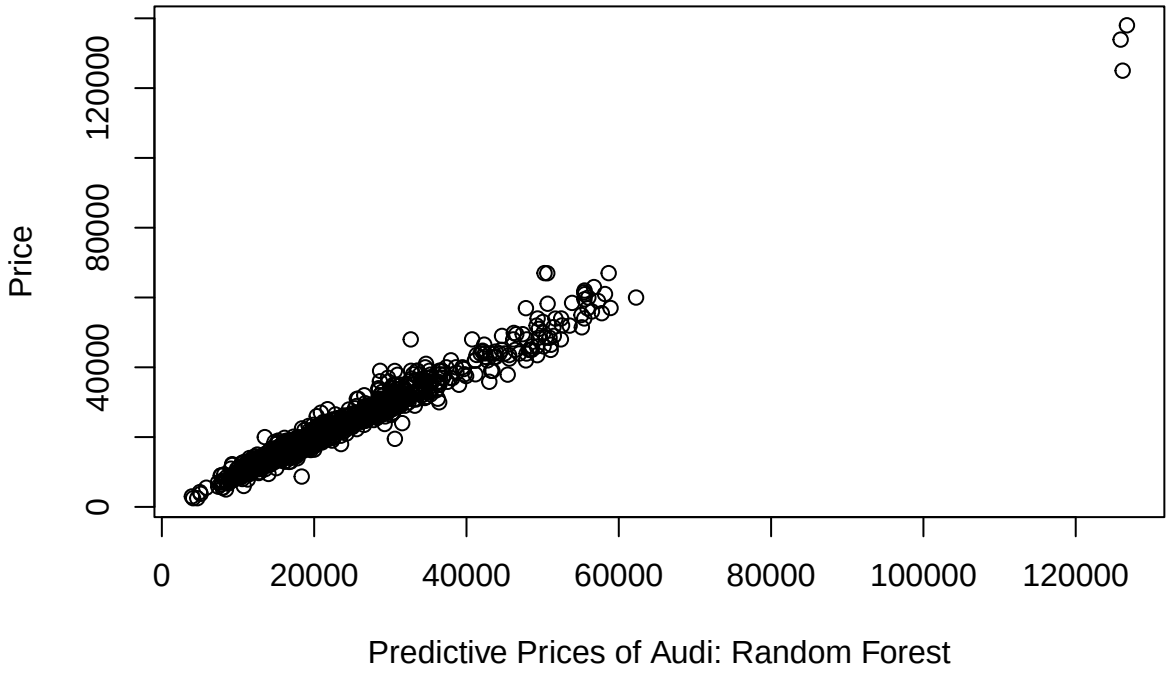
model like two variables year and mileage. If the registration year is early, the mileage could be large. We might not summarize the relationship from the coefficient of linear regression. Linear Regression model may not predict the price effectively but can identify the cause and effect of variables to the predicted price.

4.2 Lasso Models

For the second statistics model, we used the Lasso model which introduces the lambda in this model. The function of the lambda is to push variables near to zero and the Lasso can help us improve the performance and make sure of some important variables during the prediction of prices. We used the Lasso model in the process of our analysis of three brands: Audi, BMW, Ford. During the train set, it selects the best lambda when it became the smallest value of rmse in this process. In the test set, we used this lambda to predict the price of these eight variables and got the rmse of this prediction part. During this process, the Lasso can make us understand the importance of variables and the interactions between variables.

4.3 Random Forest

Figure 4.3: Comparison between Predictive Prices and Prices



The third method we used is Random Forest from tree models. It involves the process of building many trees on the bootstrapped training data. Then Random Forest could reduce the tree correlation and increase the randomness of the tree building process. So, this method could have a good result for predicting the price in our topics. We used Random Forest in three types of cars and predicted the price of three types of used cars. In addition, we find that Random Forest can capture the nonlinear characteristics between several variables which two previous models can't do. According to Figure 4.3, we can find the accurate relationship between predicted prices and registration prices. The Random Forest could predict the used cars' prices in an effective way from this figure.

Table 2: **Table 5.1 : The Comparison of RMSE of Four Predictive Models of Audi, BMW, and Ford **

	V1
Audi_rmse_lm	4152.143
Audi_rmse_la	5779.222
Audi_rmse_rf	2285.014
Audi_rmse_bo	2530.220
BMW_rmse_lm	3897.619
BMW_rmse_la	5532.243
BMW_rmse_rf	2286.634
BMW_rmse_bo	2320.995
Ford_rmse_lm	1887.308
Ford_rmse_la	2461.473
Ford_rmse_rf	1163.883
Ford_rmse_bo	1244.973

4.4 Boosting

Last but not least, the Boosting model can improve the prediction accuracy and deal with issues of bias variances of other models. At the beginning of this model, it fits the data with a single tree and could make some mistakes. This model could fit the new tree which sequentially fixes the last tree's mistake. The new fit is the sum of trees and this Boosting model can improve the performance by building trees that decrease mistakes of previous trees. So we used the Boosting model to predict the price of used cars of Audi, BMW, and Ford. The parameters we chose are the number of trees: 500, interaction depth: 4, and shrinkage: 0.05. The table Relative Influence of BMW in the Appendix illustrates the variables' relative influence of BMW. "Model", "mileage", and "mpg" have higher values and are very important in the boosting method. It has a similar result in Audi and Ford.

5. Results

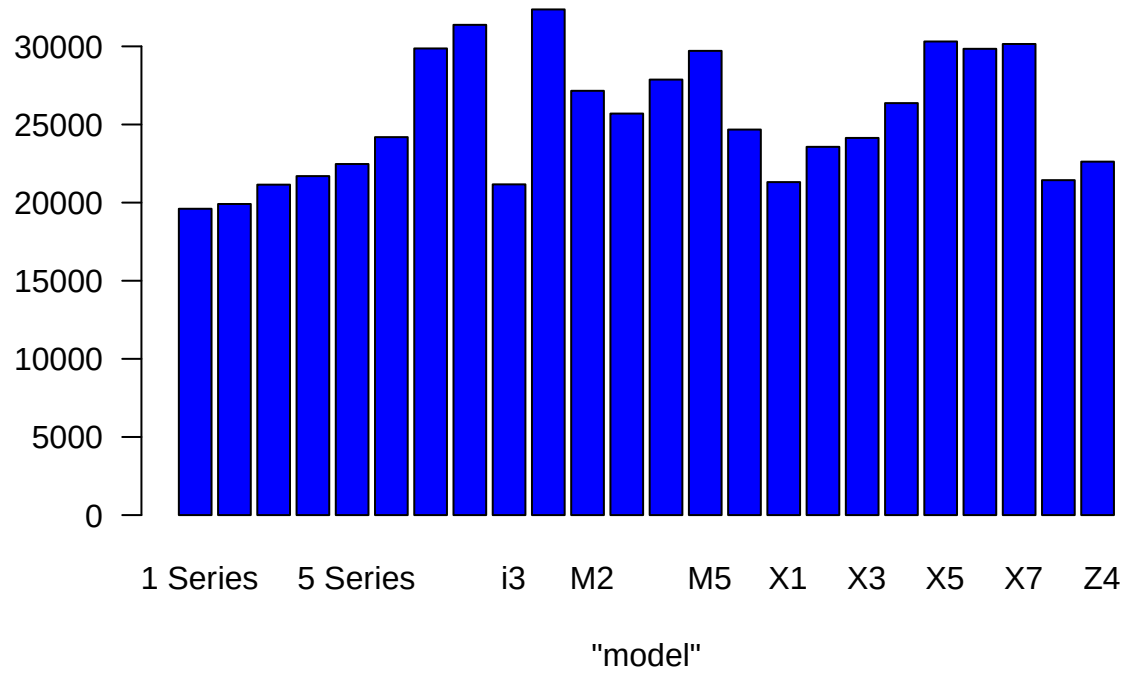
5.1 The Comprasion of RMSE

After comparing the result of rmse of four predictive models of three cars, the Random Forest's rmse is the smallest and this can provide an accurate prediction about the prices of used cars. The Boosting's rmse is similar to the Random Forest due to some common characteristics. In addition, as compared to Audi and BMW, rmse of four models of Ford used cars are so small and four models can predict the price of used cars on Ford more accurately.

5.2 Three important variables and Partial Effects

From the previous performance of Random Forest and the Relative Influence of boosting, we think there are three important variables of predicting the price of used cars. We did three partial dependence plots related to Random Forest of three types of cars. According to Figure 5.2.1, different models of BMW cars could bring different results of price in this process and According to Figure 5.2.2 of Audi cars in the Appendix, it illustrates that when mpg is lower, it becomes one important variable to predict the price of used cars and the price is very high at that time. The third figure 5.2.3 of Ford cars in the Appendix explains one fact that if the car's registration year is early, it has little influence to predict the price and the big engine size can predict the higher price of used cars.

Figure 5.2.1: Partial Dependence Plot on 'model' of BMW



6. Conclusion

According to the RMSE Comparison of Results, tree-based models: Random Forest and Boosting perform better than both Lasso and the Linear Model. It is also clear to see that Random Forest performs on average slightly better than the Boosting method. So we can get the conclusion that the Random Forest is the best model of the four models. It can predict the price of used cars accurately within many variables. Another reason is probably that the data has a non-linear trend and extrapolation outside the training data is not important. This model trains multiple decision trees each tree will draw the random sample, so this prevents overfitting and with a large dataset, the random forest seems to perform well. And Random Forest can handle some variables even if some variables are not important. For example, The Table of Relative Influence in the Appendix illustrates that the value of the tax is very low.

In addition to the best model, we also found many important variables including mpg, model, year, and mileage in the Results. Consumers can pick up the model they favor and we advise that consumers should consider other important variables like the mileage and the mpg of certain cars. This can let them know the range of prices and buy used cars within a reasonable choice. For this project, we use a lot of time visualizing data and doing the prediction process. To improve the prediction or model training, we can apply the adjustment in the model hyperparameter or select only the most relevant attributes for our model.

Appendix

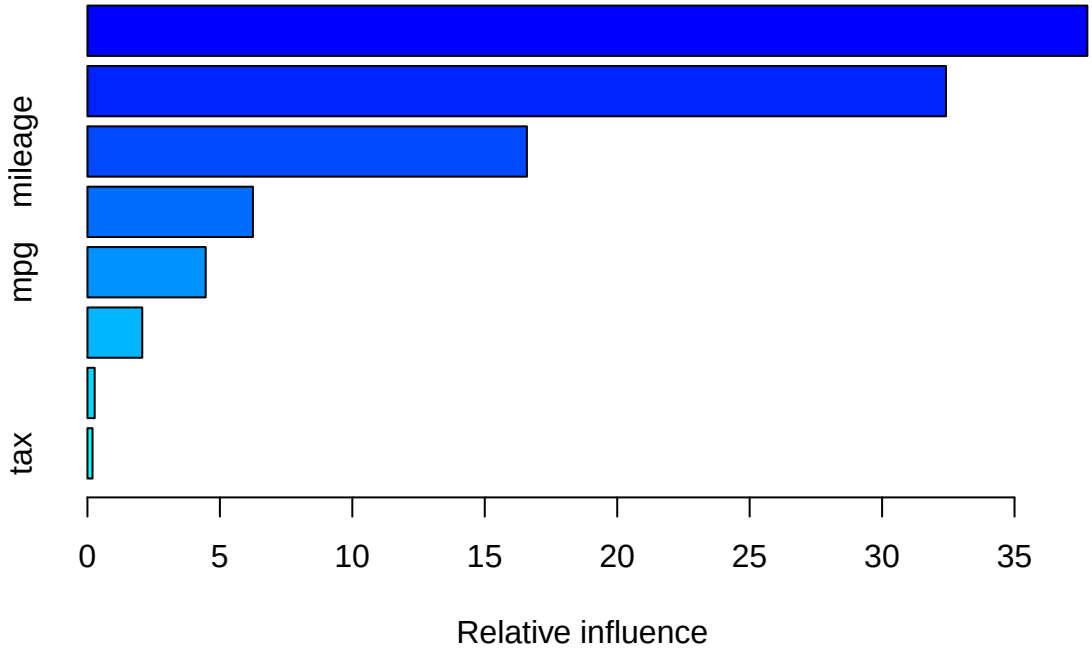


Figure 5.2.2: Partial Dependence Plot on 'mpg' of Audi

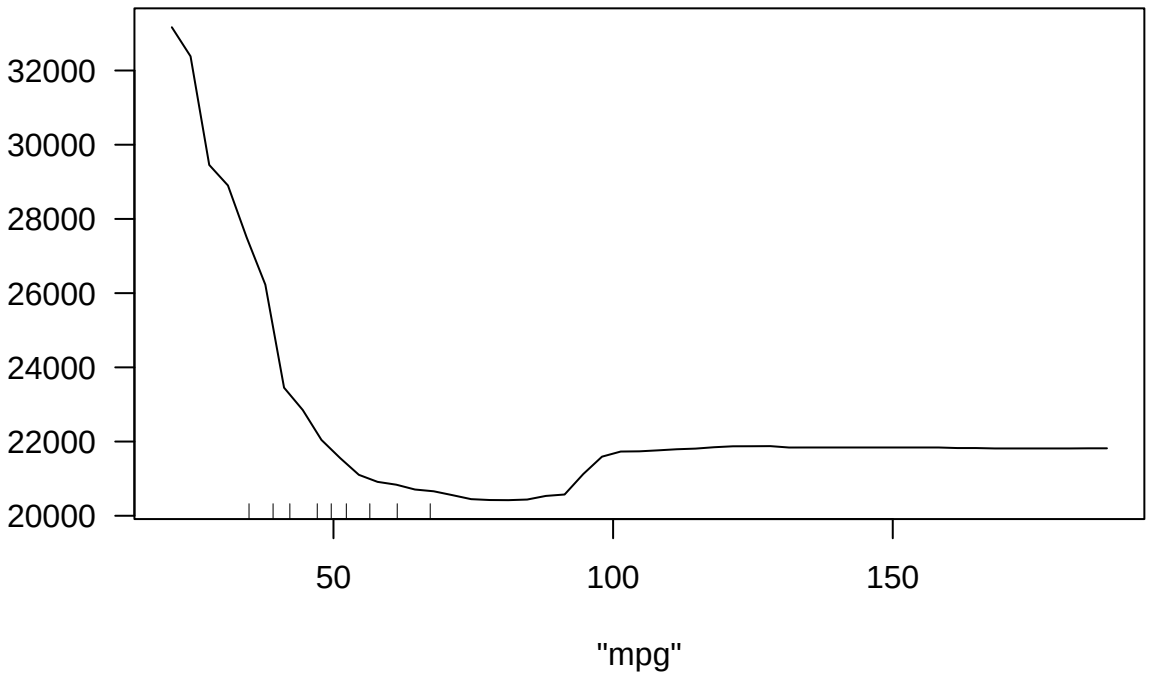


Figure 5.2.3: Partial Dependence Plot on 'year' of Ford