Ebay Used Cars Data Analysis Report

Dayi Fang

June 2, 2017

Introduction:

This project focused on analyzing a real world used cars data from Ebay to predict used cars' prices based on several continuous or categorical factors. Through exploratory data analysis, regression models, and hot map methods, we not only form good understandings of the data trend but also be able to predict and develop several valuable insights. The major methodologies used in this project are general linear model, cross validation, variable selection, random effect regression and gradient boosting algorithm.

Data:

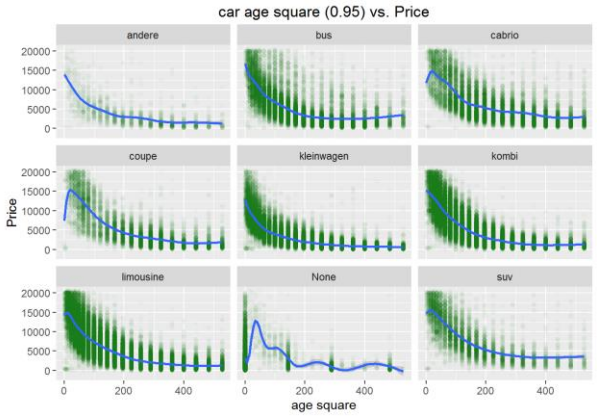
Unlike prepared data, real world data set always contains a great number of missing values and outliers. This Ebay used car dataset is found form Kaggle.com with over 370,000 pieces of used cars information (price, driving distance, car age, power, vehicle type, model…).

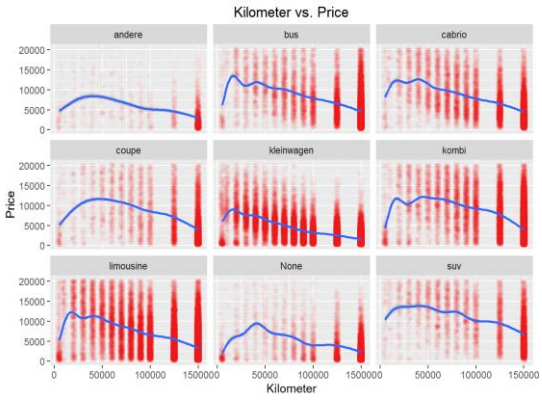
We first clean the data by removing missing values for variables that have a small number of missing values for continuous variables and rename missing values to "unknow" or "none" for categorical variables. Then we create two new date variables from the original dataset to present used cars' age and number of days until they were sold. For the age variable, we add a square transformation of it – age2 to provide better fit for our model. We also remove extreme outliers according to 95% confident interval, and exclude unreasonable/incorrect data points such as the car is registered in the 13^{th} month.

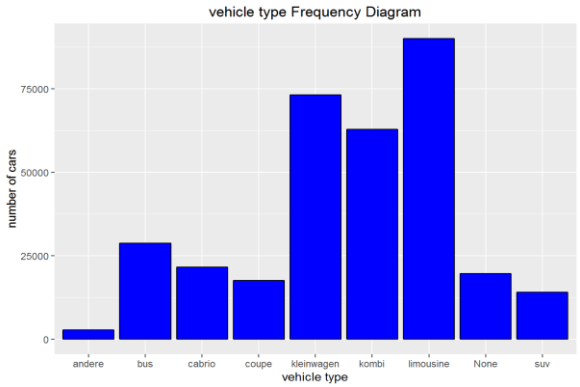
Before performing EDA and modeling, we normalize three continuous variables – price, powerPS, and kilometer to get better idea about the relative scale between them.

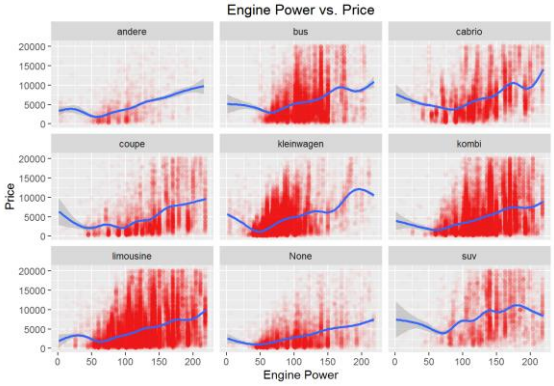
Exploratory Data Analysis:

Before working on the data, we have done some EDA to help us understand the dataset better. We created several comprehensive graphs to present the frequency of important categorical variables such as vehicleType and gearbox, the range of the continuous variables, and the relationships between them. By performing EDA, we have found several interesting facts that affect used cars' prices. Some EDA graphs are presented below:



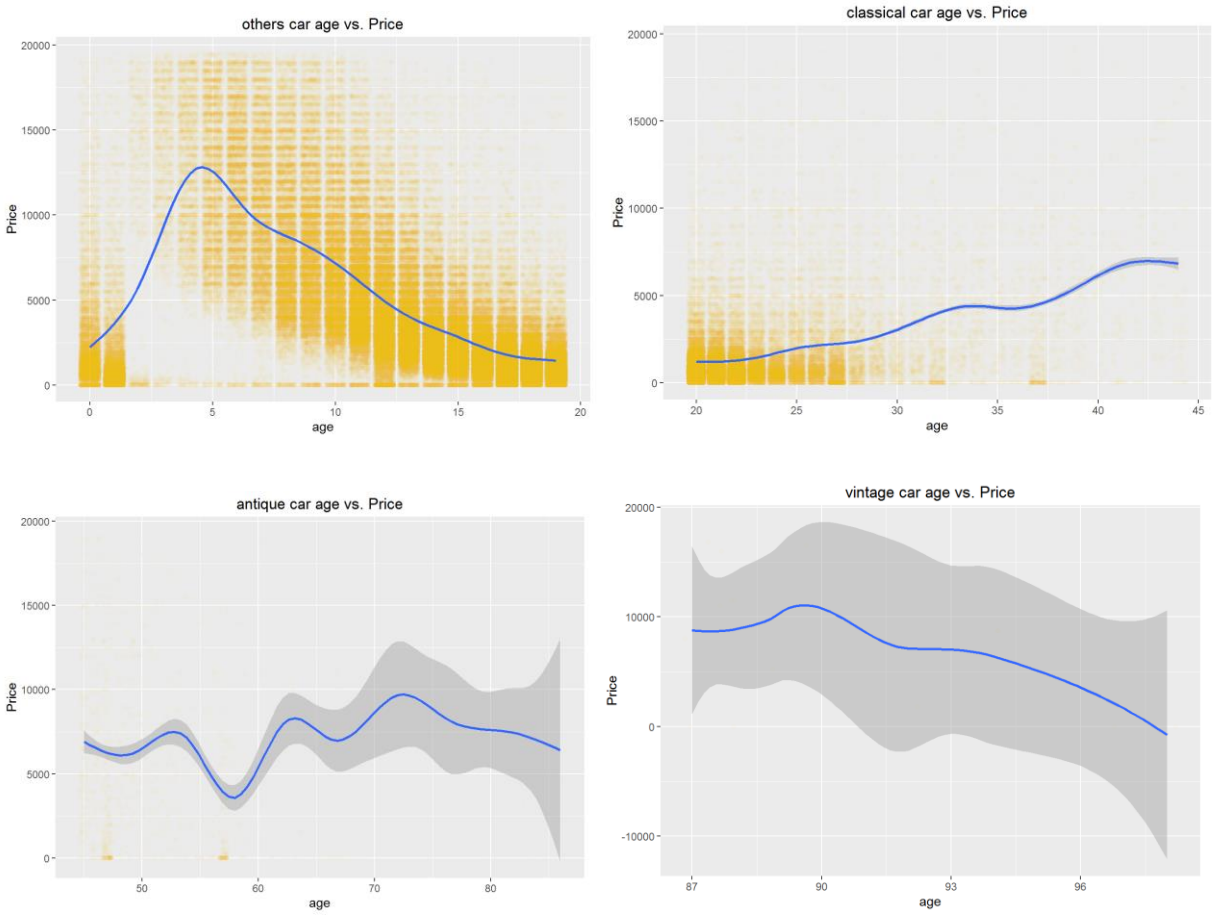






1.) vintage cars' prices

From the Ebay used car data, we can find there are several antique and vintage cars on sale. Do they have similar price patterns as the regular cars? We can find out by plotting 4 different car groups: vintage cars (registered between 1919 and 1930), antique cars (registered between 1930 and 1972), classical cars (registered between 1972 and 1997), and regular cars (registered between 1997 and 2017). We plot these cars' ages versus their prices and get an idea about the relationship change.



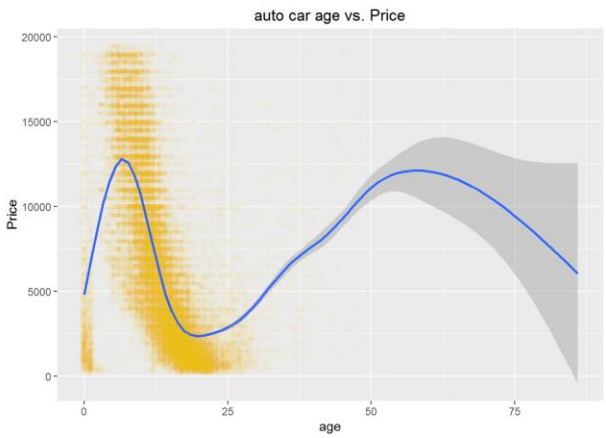
These graphs tell us that different aged cars' groups have different patterns with prices. However, vintage cars and antique cars have much higher starting prices compare to regular cars, and classical cars have their prices increased as their ages go up.

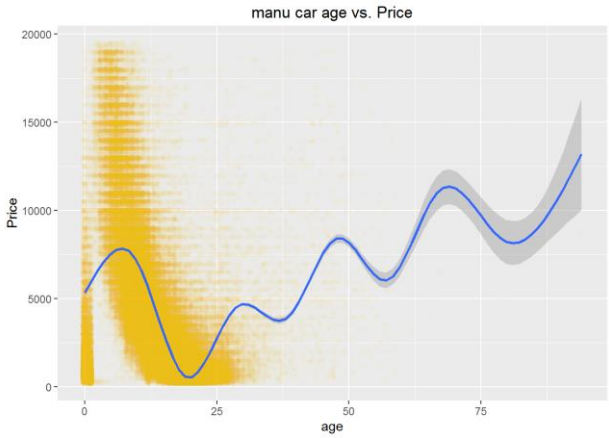
2.) Automatic car versus manual car

What is the difference between the price patterns of automatic cars and manual cars? It is interesting to find out that manual cars generally drop their prices at a slow rate than auto cars.

The following two graphs provide sufficient evidences that manu cars are dropping their prices at a lower rate than auto cars: the slope of auto car's price vs age (5-25 years old) are steeper than the manu cars. This fact shows that the absolute value of age's coefficient for auto cars is great than manu cars and proves that manu cars drop their prices at a slower rate.

Another interesting finding is that manu cars' prices are increasing at a higher rate than auto cars in general.





After EDA, we have found that linear regression can be a good modeling choice for this dataset. However, due to the various price patterns for different age groups, random forest regression and gradient boosting model are also effective for our project.

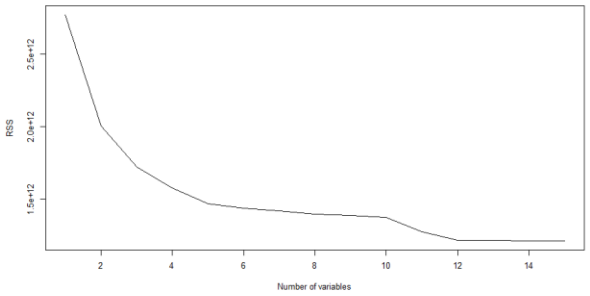
Regression models and cross validation

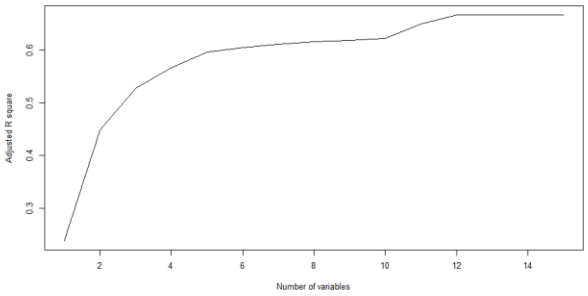
Before running the model, we first set up dummies for our three categorical variables: vehicleType, gearbox, and fuelType, then we split our data into two groups by 20%/80%, one is training set, the other is testing set. We preform several modeling algorithms on the training set and use the testing set to compare their performances. Besides setting regular train/test sets, we also use a 10-fold cross validation to further select our model.

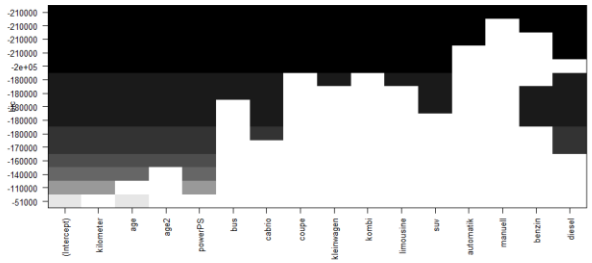
Linear model:

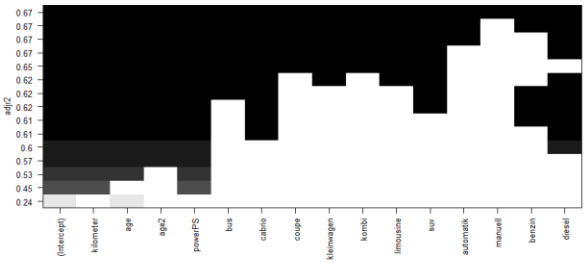
We perform best-subset and forward variable selection for our liner model improvements. Both methods indicate that a liner model with 15 predictors (powerPS, kilometer, age, ages, 7 different vehicle types, 2 different gearbox types, and 2 different fuel types). This model has an adjusted R square value of 66.7%, which is good in our case since we only use linear model to predict our response variable – price.

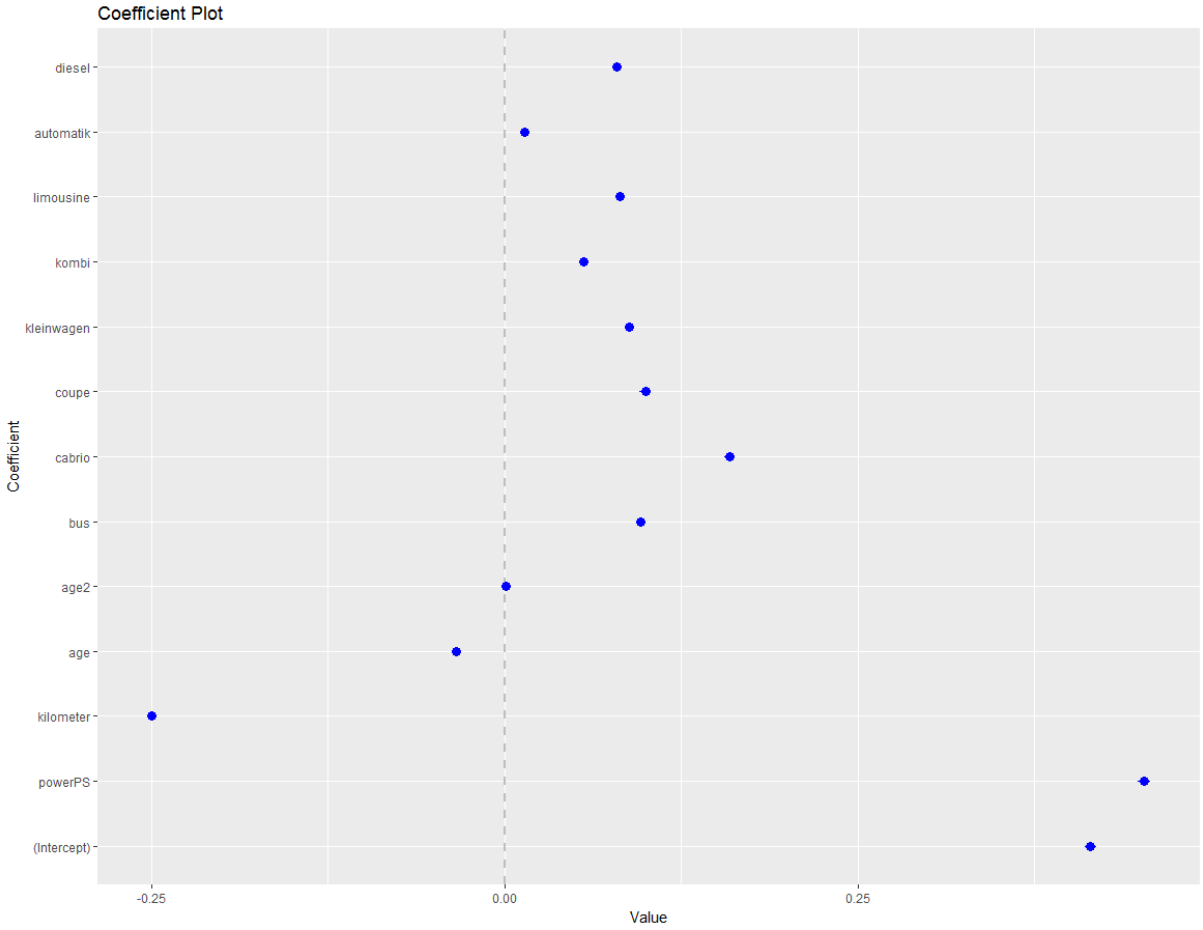
The following graphs show that as the number of variables of the linear regression increases, the r square value also increases, and the BIC value and residual sum of square value decrease. For the coefficients plot, it shows the importance of our variables in this linear model.


RSS
2.5e+12
2.0e+12
1.5e+12
2
4
6
8
10
12
14
Number of variables

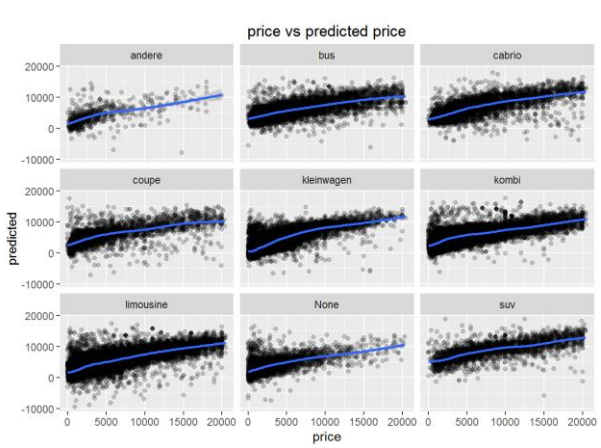

Adjusted R square
0.6
0.5
0.4
0.3
2
4
6
8
10
12
14
Number of variables

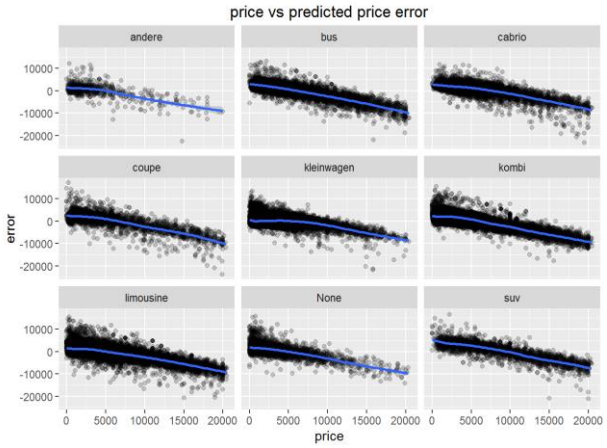

-210000
-210000
-210000
-210000
-2e+05
-180000
-180000
-180000
-180000
-180000
-170000
-160000
-140000
-110000
-51000
(Intercept)
kilometer
age
age2
powerPS
bus
cabrio
coupe
kleinwagen
kombi
limousine
suv
automatik
manuell
benzin
diesel


adjr2
0.67
0.67
0.67
0.67
0.65
0.62
0.62
0.62
0.61
0.61
0.6
0.57
0.53
0.45
0.24
(Intercept)
kilometer
age
age2
powerPS
bus
cabrio
coupe
kleinwagen
kombi
limousine
suv
automatik
manuell
benzin
diesel


Coefficient Plot
diesel
automatik
limousine
kombi
kleinwagen
coupe
cabrio
bus
age2
age
kilometer
powerPS
(Intercept)
Coefficient
-0.25
0.00
0.25
Value

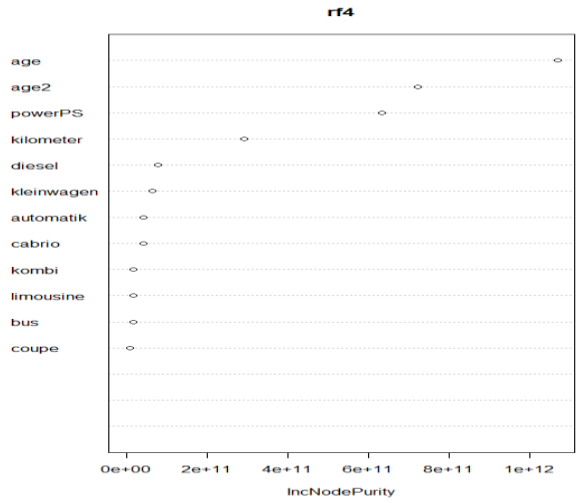
The following two graphs show our predicted price versus actual prices and predicted error versus prices. They inform us that our model results have similar patterns as the actual data, but always underestimate the prices when the actual price is high. We believe this error is caused by several factors such as different effects of the cars' ages for various groups, limitations of the linear model method, or the effects of the outliers.

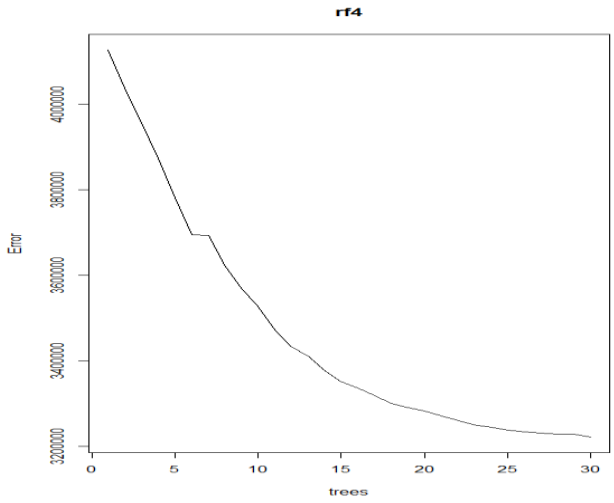




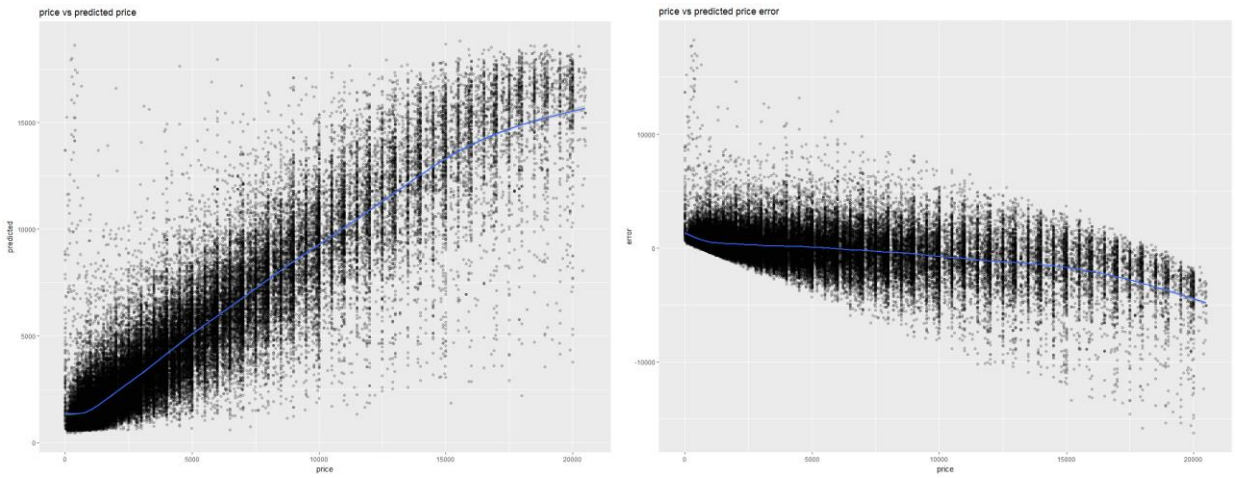
Random Forest:

Since our linear model provides 66.7% R square rate, which is not so bad but still can be improved. We take another look at the random forest regression for capturing the up and down patterns of the cars' ages. Due to the huge computational power requirement by the random forest model, we simplify our data further by removing missing values of the categorical variables, dropping unnecessary predictors, and setting smaller number of trees. Our random forest regression model provides a good improvement by increasing r squared value from 66.7% to around 81%.





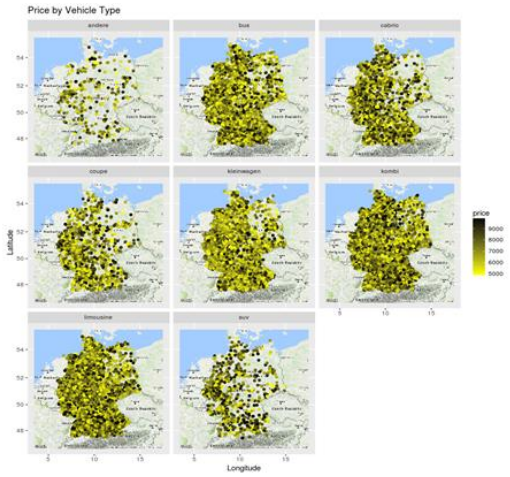
The above graph shows the variable importance and the improvement of random forest model as the number of trees increases. The error line seems to converge as the number of trees goes up, so that we can conclude 30 trees is an effective number for our model.



The predicted price versus actual prices and predicted error versus prices graphs also look better compare to our linear model. Although this is still correlation between error term and the price, the magnitude of the predicted errors is decreased in a significant amount.

Future developments and considerations:

Considering geographic location could be a deciding factor affecting used cars' prices, a hot map (German, since the data is collected there) is generated.



From the graph, it is clear to find that different regions have different price preferences. Given the area zip code, a random effect model can be conducted for future improvements.