Department of Computer Engineering

INF508: Machine Learning
Final Project Report

Used Car Price Estimation

Cem Yeniçeri
15411007

Galatasaray University
Fall 2017

Contents

1. INTRODUCTION

Recommendation plays a significant role in many fields such as financial systems, social network applications, e-commerce web sites and insurance systems. It can also be used for price estimation in real estate agency systems and used car trade systems. Automobile industry is getting more competitive each day and brands present new models every year. This fast changing causes more second-hand cars placed on market. As a result of this increase, importance of second-hand car price evaluation grows day by day.

Many researches showed that machine learning provides smart solutions to these kinds of problems. In this project, a supervised learning method has been applied to estimate second hand car price. To obtain a large dataset, kaggle database is used. There are three main steps on development which are data analyzing, data preparation and neural network application. Data mining processes, which are cleaning and feature selection have been applied to raw dataset to have inputs for artificial neural network algorithm. According to the results, my solution to the problem has been considerably successful, and is able to produce price estimations for whom wants to sell or buy a second car.

2. TOOLS & ENVIRONMENT

This section provides information about the tools, development environment used.

The project was developed on the Windows 7 Professional operating system. The computer used is Intel Core i7-5600U CPU with 2.60 GHz processor and 16.00 GB RAM on the device.

R Studio was used for analyzing, preprocessing and predicting, plotting and testing. R is a free software environment for statistical computing and graphics.

3. DATA COLLECTION

Data is obtained from public dataset provider kaggle. Kaggle fetched over 189349 used cars data from Ebay. The content of data is in German. They are handled at preprocessing phase of project.

The format of the data file is CSV and fields are as follows:

dateCrawled : When this ad was first crawled, all field-values are taken from this date

name : Name of the car

seller : Private or dealer

offerType : Offered or Looked for

price : The price on the ad to sell the car

abtest : Test or control

vehicleType : Body type of car

yearOfRegistration : At which year the car was first registered

gearbox : Manuel or automatic

powerPS : Engine Power of the car

model : Model of cars

kilometer : How many kilometers the car has driven

monthOfRegistration : At which month the car was first registered

fuelType : Gasoline or Diesel

brand : Brand of the car

notRepairedDamage : If the car has a damage which is not repaired yet

dateCreated : The date for which the ad ebay was created

nrOfPictures : Number of pictures in the ad

postalCode : Postal code of who owns ad

lastSeenOnline : When the crawler saw this ad last online

4. EVALUATION

Data needs to preprocess before get into the neural network process. Raw data has empty and noisy instances. In data preparation process, first of all, some of fields are removed from dataset, because they do not affect price of cars or same data provided by other columns. Then, incomplete and empty and noisy data are deleted.

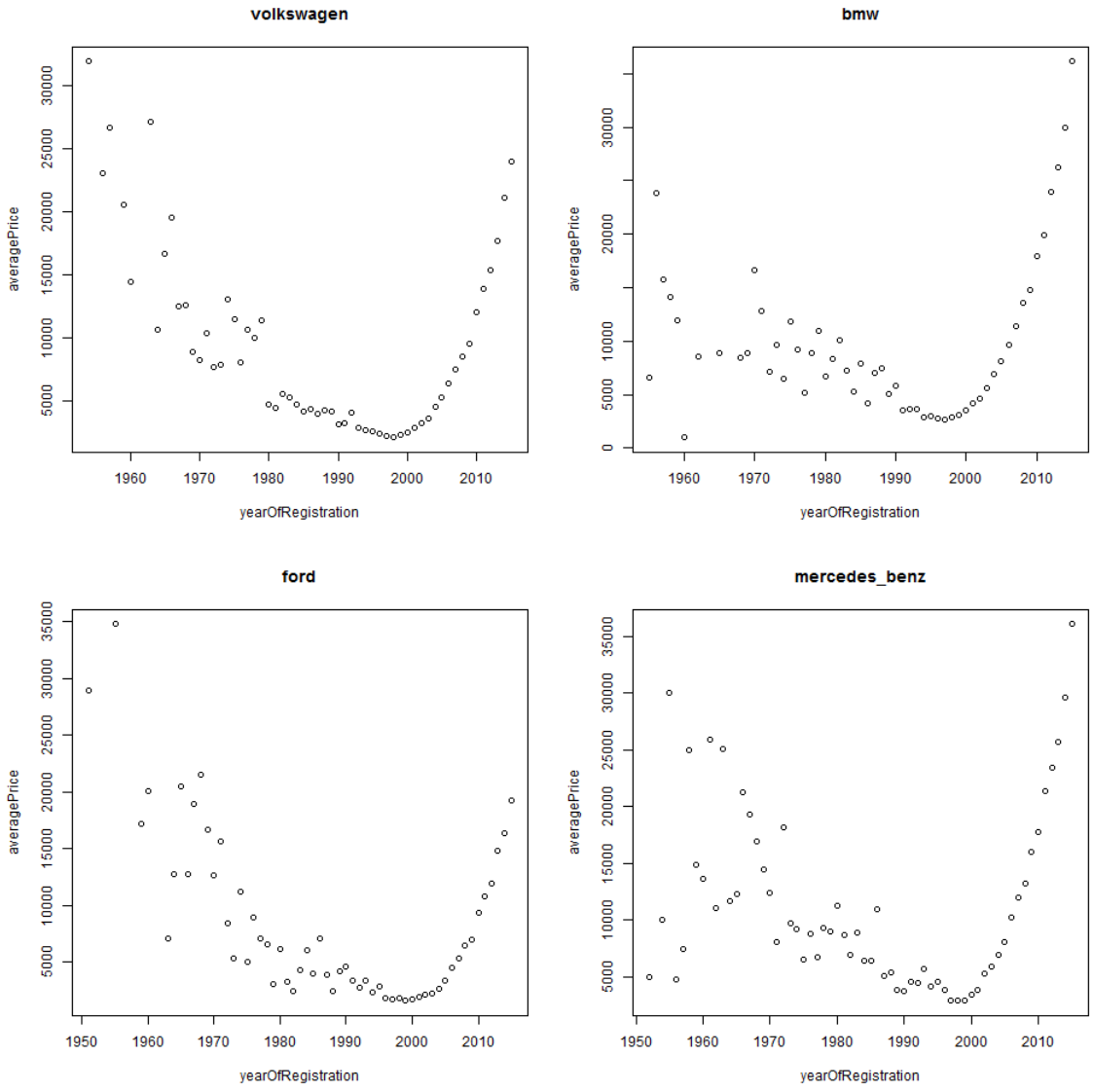
4.1 Preprocessing

Feature selection preocess has been done manually in this project. Relying on domain knowledge on buying and selling second-hand car, postal code and last seen can be ignored directly. They have been removed. Number of pictures could affect possibility of selling a product but it does not increase or decrease price of a product. Therefore, it has been also removed. Year of registration is important feature for price of a car, but month is detail for them. Therby, month information has been removed. If date crawled column is analyzed, all rows had been collected in two months. Therefore, this atrribute also has been removed. Name has been also removed, because name column is just detailed description about model and brand columns. Similar information could be obtained from these two columns. Abtest is known as testing method that holds control and test values, which are not related for predicting price value. Vehicle Type such as cabrio, coupe, hatchback etc. is normally important for classify cars. In this dataset, almost most of cars has unique vehicle type already. For example, if a user selects a golf as brand, hatchback vehicle type is already selected automatically.Golfs have not any other vehicle types. If this dataset is used for another aim, vehicle type could be useful but in our case not. Therefore, it has been removed. Normally, dateCreated is important for price prediction in this domain

because price can increase or decrease due to lots of reasons in timeline. When datecreated field is analyzed, all advertisements have been created between 2015 and 2016 years. Therefore, it is assumed that price of car does not change drastically. For simplicity, this field has been removed with regard to above explanation.

In second phase of preprocessing, attributes are investigated a bit more. When seller type is analyzed, it is realized that two options can be found which are “gewerblich” and “privat”. Those words are german as you guess. Gewerblich means dealer and privat means “this attribute becomes insignificant for price estimation and has been removed. Another attribute offerType has two variants as “Angebot” and “Gesuch”. “Angebot”s mean who create advertisements for selling their car. On the contrary, “gesuch”s mean who create advertisement for buying a car. In remaining dataset after cleaning mentioned above just eight “gesuch” raws exist and they are not meaningfull for our bussiness model. Then, offer type has been removed. Cars have price which is lower than 1000 euro and higher than 60000 euro, have been excluded because more expensive cars(ultra lux) cause so much errors in model. Also, cars has price is lower than 1000 euro, are very cheap and should be out of our scope. They cause again outliers in our model. “sonstige_autos” label has been given to cars which of brand is not determined by crawler. “sonstige” means other in english. The brands are “sonstige_autos”, which are also removed from dataset to pbtain meaningfull dataset. When engine power of cars are explored, it is realized that some power ps values that makes no sense, exist like over than 700 ps. They also should be out of scope, so they have been excluded.

Year of registration contains so many noisy data like 9999 and 1000. Since data has been crawled one year ago(2016) and our aim is to predict second-hand car price, minimum age for a used car should be one year to get better result. Therefore, upper limit has been selected 2015 for year registration. In addition to upper limit, lower limit should be determined. Plots of price distributions by year helped to determine lower limit. For Volkswagen, BMW, Ford and Mercedes, plots have been drawn as below. According to graphics, after 1995 in year, dataset makes sense. After that time, increasing price can be shown easily. Before this time, probably, some noisy data and vintage car instances exist. Therefore, intances with before 1995 in year have been excluded.



After removing irrelevant year instances, when price data are investigated, some noisy instances are found as 111111,22222,3333. All they have been excluded by using regex. After all, clean dataset is obtained for artificial neural network.

4.2 ANN Process

"nnet" library of R is used on ANN process. After preprocessing, dataset has 8 features with price value and 128031 instances. Brand, model, fuel type, gearbox and notRepairedDamage properties are key attributes for prediction. However, they are nominal values. By the way, just numeric values can be used in ann process. Therefore, model can not contain these attributes directly. Eventhough this seems a problem, if they are used as filter while creating model, this can provide better models for ann. For example, if a user wants to learn price of a golf, ,it is expected to him giving some information about vehicle such as model, fuel type, gear box and notRepairedDamage status. Lets assume that this user wants to buy a volkswagen golf that consumes gasoline, is used manually and can have damage, training data are generated along these filter information. Then, all remaining columns which

are price, yearOfRegistration, kilometers and powerPs, have been scaled as 0 to 1. To fit our model, linOut property of nnet library has been selected as true. Using nnet for a regression problem setting lineout to TRUE is required. One hidden layer has been choosen with 100 units. Maximum number of iteration is determined as 1000. Other settings remained as default. The best results have been obtained with size of 100 and maxit of 1000. Obtained prediction values are scaled after ann process has been done, so rescaled is needed to compare with original value.

Results have been calculated as maximum, minimum and mean absolute percentage error and maximum, minimum and mean absolute error.

The unit of price is euro(€). The results are coming from the Volkswagen Golf example.

	Maximum	Minimum	Mean
Absolute Error	2359.835	1.247607	409.0194
Absolute Percentage Error	138.8138	0.007343185	19.40027

A few instances predicted from nnet and original price values as an example:

Original Price	Predicted Price
8000	7915.4215
9500	9176.9923
2500	2596.2216
1700	1373.5832
2599	3202.4303
1300	1168.3217
3600	4114.8668

5. CONCLUSION & FUTURE WORK

The aim of this project was to predict the price of used cars by applying artificial neural network. Although there are many car websites in Turkey but crawling data is very painful activity because of so many blocking, etc. Therefore, a ready dataset which is downloaded from kaggle.com. The raw dataset has twenty attributes but some of them have been excluded. As a consequence, three attributes with price values have been used. Size of dataset was 128031. The kilometers, engine power and year of registered have been used to predict the price of used cars using nnet library of R.

The system predicts price with about 19.4% error. This system can be useful for car sellers and car buyers who need to assess the value of their cars.

In the future, algortihm can be tried with more features. Algorithms parameters can be changed and even other machine learning algorithms can used. If real up to date used cars data can be used by accessing popular car web sites, results will be more realistic.