**Summary Table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NO** | **Task** | **No.of Features** | **Model** | **Accuracy(R2)** |
| **1** | **As is dataset modelling** | **19 vs 1** | **Linear regression** | **1%** |
| **2** | **As is dataset experiment** | **19 vs 19** | **Linear regression** | **1%** |
| **3** | **Removing insignificant features** | **12 vs 1**  **12 vs 12** | **Feature Analysis**  **Correlation matrix**  **P-value coefficient** |  |
| **4** | **Imputing with AmeliaView tool and then modelling** | **12 vs 1**  **12 vs 12** | **Linear regression** | **2%** |
| **5** | **Removing low significant values after imputing** | **12 vs 1**  **12 vs 12** | **Linear regression** | **2 to 6% with individual feature** |
| **6** | **Min-Max normalization** | **12 vs 1** | **Linear regression** | **5%, after removing some insignificant vars its 6%** |
| **7** | **Z-Score normalization** | **12 vs 1** | **Linear regression** | **6%** |
| **8** | **Quantile normalization** | **12 vs 1** | **Linear regression** | **11%** |
| **9** | **Quantile + Z-Score normalization** | **12 vs 1** | **Linear regression** | **11%** |
| **10** | **Quantile + min\_max** | **12 vs 1** | **Linear regression** | **8%** |
| **11** | **Min\_max + Quantile** | **12 vs 1** | **Linear regression** | **48%** |
| **12** | **Square root Normalization** | **5 vs 1** | **Linear regression** | **28%** |
| **13** | **Square root + Quantile** | **5 vs 1** | **Linear regression** | **58%** |
| **14** | **Regularization - Ridge regression** | **12 vs 1** | **Linear regression** | **6%** |
| **15** | **Regularization - Lasso regression** | **12 vs 1** | **Linear regression** | **6%** |
| **16** | **Ensembling- GBBoost** |  |  |  |

**Design & Implementation of Descriptive Analytics, Predictive Analytics, Prescriptive Analytics On Used cars sales data set:**

**Usecase: Predicting price of used cars & sell-ability of used cars & sales forecast**

Details of data of used cars sales:

No.of features: 20

No.of labels: 1

No.of records: 3,71,530

**Descriptive analytics:** a. plotting price vs individual attributes

b. plotting price vs combination of attributes

c. combination of attributes vs dates

d. newly created target label which will be formulated by available attributes

**Predictive analytics:**

Regression model to be be employed with available attributes: {price,abtest,vehicleType,yearOfRegistration,gearbox,powerPS,model,kilometer,

monthOfRegistration,fuelType,brand,notRepairedDamage,

dateCreated,nrOfPictures,lastSeen}

**Prescriptive analytics:**

Best sell-ability

Prediction of price of given model

Sales forecast

**Dataset source:** [**https://www.kaggle.com/orgesleka/used-cars-database**](https://www.kaggle.com/orgesleka/used-cars-database)

**RStudio is chosen for analytics purpose.**

Running following commands would give some idea of what is the data etc. about.

1. Will read the csv data file to origdata:

origdata <- read.csv("D:\\projects\\machineLearning\\PredictiveAnalytics\\autos.csv", quote="")

1. class(origdata): class of data object
2. dim(origdata): Dimension of data
3. names(origdata): column names
4. str(origdata): Preview of data with helpful details
5. glimpse(origdata): better version of str() from deplyr
6. head(origdata): View top data set
7. tail(origdata). View bottom data set
8. hist(): Histogram of single variable
9. plot(): plot of two variables
10. summary(origdata): summary of data

summary(origdata)

dateCrawled name

3/5/2016 14:25 : 68 Ford\_Fiesta : 657

3/5/2016 14:26 : 62 BMW\_318i : 627

3/5/2016 15:48 : 58 Opel\_Corsa : 624

3/5/2016 17:49 : 58 Volkswagen\_Golf\_1.4: 605

3/16/2016 18:49: 55 BMW\_316i : 523

3/20/2016 11:50: 55 BMW\_320i : 492

(Other) :371460 (Other) :368288

seller offerType price

: 1 : 1 0 : 10785

Angebot : 1 4000 : 1 500 : 5674

gewerblich: 3 Angebot:371802 1500 : 5397

privat :371811 Gesuch : 12 1000 : 4654

1200 : 4597

2500 : 4443

(Other):336266

abtest vehicleType yearOfRegistration

: 1 limousine :95961 2000 : 24574

control:179088 kleinwagen:80095 1999 : 22782

kombi : 1 kombi :67625 2005 : 22330

test :192726 :37900 2006 : 20249

bus :30219 2001 : 20231

cabrio :22913 2003 : 19885

(Other) :37103 (Other):241765

gearbox powerPS model

: 20224 0 : 40859 golf : 30085

102 : 1 75 : 24056 andere : 26420

automatik: 77165 60 : 15920 3er : 20581

manuell :274426 150 : 15448 : 20499

140 : 13596 polo : 13105

101 : 13325 corsa : 12584

(Other):248612 (Other):248542

kilometer monthOfRegistration fuelType

Min. : 10 0 : 37706 benzin :224030

1st Qu.:125000 3 : 36191 diesel :107824

Median :150000 6 : 33197 : 33417

Mean :125619 4 : 30943 lpg : 5382

3rd Qu.:150000 5 : 30649 cng : 571

Max. :150000 7 : 28981 hybrid : 279

NA's :2 (Other):174149 (Other): 313

brand notRepairedDamage

volkswagen : 79694 : 72124

bmw : 40301 3/9/2016 0:00: 1

opel : 40166 ja : 36308

mercedes\_benz: 35342 nein :263383

audi : 32897

ford : 25592

(Other) :117824

dateCreated nrOfPictures postalCode

4/3/2016 0:00 : 14450 Min. : 0.00 10115 : 828

4/4/2016 0:00 : 14022 1st Qu.: 0.00 65428 : 638

3/20/2016 0:00: 13547 Median : 0.00 66333 : 349

3/12/2016 0:00: 13379 Mean : 0.06 38518 : 326

3/21/2016 0:00: 13305 3rd Qu.: 0.00 32257 : 324

3/14/2016 0:00: 13088 Max. :22159.00 44145 : 323

(Other) :290025 NA's :2 (Other):369028

lastSeen

4/7/2016 6:45: 708

4/7/2016 7:16: 700

4/7/2016 6:16: 692

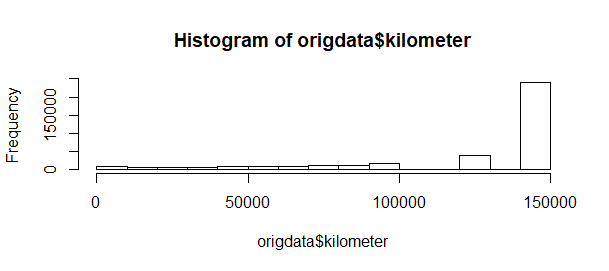
4/6/2016 9:17: 680

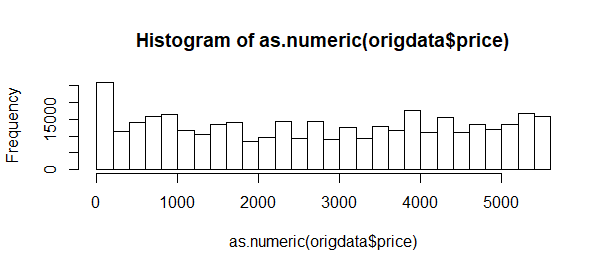
4/6/2016 4:45: 679

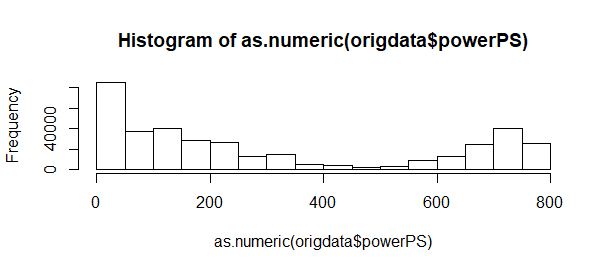
4/6/2016 2:45: 675

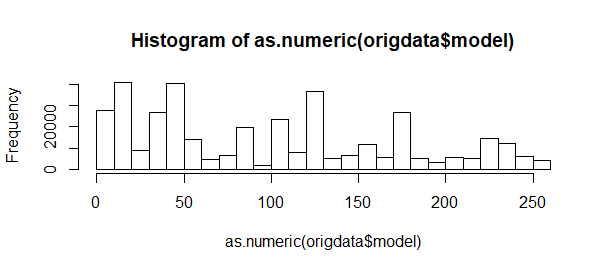
(Other) :367682

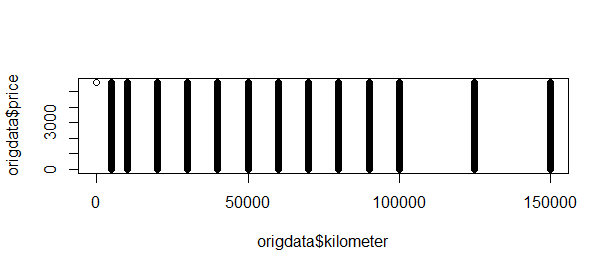
hist(origdata$kilometer):



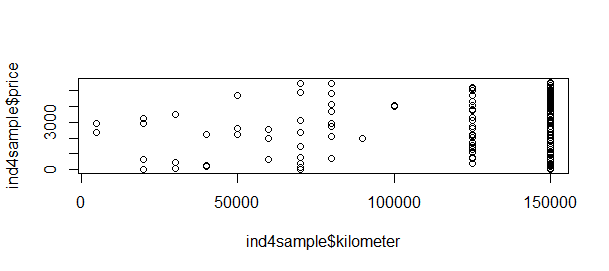


****

****

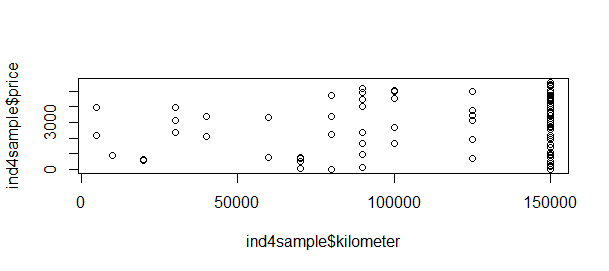
****

**For 192 sample:**

****

**With 100 sample...**

**Seems Kilometers vs Price has no direct influence..**

****

**Then going for correlation coefficient:**

cor(ind4sample$kilometer, as.numeric(ind4sample$price))

[1] 0.1131283

> cor(sdata$kilometer, as.numeric(sdata$price))

[1] 0.03749912

**Which again proves, there’s no impact of kilometer feature on price.**

**Then may go for VIF: Variance Inflation factor**

**Sample Model Script**

**head(autos)**

**#Split dataset into "training" (60%) and "validation" (40%)**

**ind <- sample(2, nrow(autos), replace=TRUE, prob = c(0.6,0.4))**

**#training data**

**tdata <- autos[ind==1,]**

**#validating data**

**vdata <- autos[ind==2,]**

**#multiple linear regression model**

**results <- lm(price~kilometer+powerPS, tdata)**

**summary(results)**

**results$coefficients**

**coef(results)**

**#prediction**

**pred <- predict(results, vdata)**

**head(pred)**

**head(vdata)**

**> cor(as.numeric(ind4sample$powerPS), as.numeric(ind4sample$price))**

**[1] -0.1294626**

**Which is a weak negative linear relationship**

**cor(as.numeric(ind4sample$gearbox), as.numeric(ind4sample$price))**

**[1] 0.1736545**

**Weak positive linear relationship**

**In cor(ind4sample$nrOfPictures, as.numeric(ind4sample$price)) :**

**the standard deviation is zero**

**cor(as.numeric(ind4sample$abtest), as.numeric(ind4sample$price))**

**[1] -0.1559833**

**> cor(as.numeric(ind4sample$gearbox), as.numeric(ind4sample$price))**

**[1] 0.1736545**

**> cor(as.numeric(ind4sample$abtest), as.numeric(ind4sample$price))**

**[1] -0.1559833**

**> cor(as.numeric(ind4sample$fuelType), as.numeric(ind4sample$price))**

**[1] 0.07871191**

**> cor(as.numeric(ind4sample$vehicleType), as.numeric(ind4sample$price))**

**[1] -0.02953545**

**> cor(as.numeric(ind4sample$name), as.numeric(ind4sample$price))**

**[1] 0.01437569**

**> cor(as.numeric(ind4sample$seller), as.numeric(ind4sample$price))**

**[1] NA**

**Warning message:**

**In cor(as.numeric(ind4sample$seller), as.numeric(ind4sample$price)) :**

**the standard deviation is zero**

**> cor(as.numeric(ind4sample$offerType), as.numeric(ind4sample$price))**

**[1] NA**

**Warning message:**

**In cor(as.numeric(ind4sample$offerType), as.numeric(ind4sample$price)) :**

**the standard deviation is zero**

**> cor(as.numeric(ind4sample$abtest), as.numeric(ind4sample$price))**

**[1] -0.1559833**

**> cor(as.numeric(ind4sample$yearOfRegistration), as.numeric(ind4sample$price))**

**[1] -0.1273234**

**> cor(as.numeric(ind4sample$vehicleType), as.numeric(ind4sample$price))**

**[1] -0.02953545**

**> cor(as.numeric(ind4sample$yearOfRegistration), as.numeric(ind4sample$price))**

**[1] -0.1273234**

**> cor(as.numeric(ind4sample$gearbox), as.numeric(ind4sample$price))**

**[1] 0.1736545**

**> cor(as.numeric(ind4sample$powerPS), as.numeric(ind4sample$price))**

**[1] -0.1294626**

**> cor(as.numeric(ind4sample$model), as.numeric(ind4sample$price))**

**[1] 0.02348359**

**> cor(as.numeric(ind4sample$kilometer), as.numeric(ind4sample$price))**

**[1] 0.1131283**

**> cor(as.numeric(ind4sample$monthOfRegistration), as.numeric(ind4sample$price))**

**[1] -0.01819059**

**> cor(as.numeric(ind4sample$fuelType), as.numeric(ind4sample$price))**

**[1] 0.07871191**

**> cor(as.numeric(ind4sample$brand), as.numeric(ind4sample$price))**

**[1] -0.02423699**

**> cor(as.numeric(ind4sample$notRepairedDamage), as.numeric(ind4sample$price))**

**[1] -0.07833866**

**>**

**After plotting data as is following things are observed:**

1. **Price:** Valid data is between 100$ to $70000
2. **yearOfRegistration:** Valid data is greater than 1900 & less than 2018
3. PowerPS: 0 to 2000

**Dimension reduction table**

**19 X 1 table required**

**20 X 20 table required**

**F - Statistics**

**T - Statistics**

**Pearson R**

**VIF**

**Tasks:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Used Cars Sales Dataset - Analytics** |  |  |  |  | Glossary |
| **Tasks** | **status** | **Issues** | **expected completion date** |  | DA-Descriptive Analytics / Statistical an |
| Plotting Price vs Kilometers | Done |  |  |  | PDA-Predictive Analytics |
| Plotting Price vs Brand & Type |  |  |  |  | PSA-Prescriptive Analytics |
| Plotting Price vs Year | Done |  |  |  |  |
| Plotting Price vs fuel |  |  |  |  |  |
| Plotting price vs Damages |  |  |  |  |  |
| Finding combination of attributes vs price |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Experimenting with combination of two or more attributes |  |  |  |  |  |
| Finding potential ranges of individual attributes |  |  |  |  |  |
| Finding potential ranges of combination of attributes |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Best sell-ability |  |  |  |  |  |
| Prediction of price of given model |  |  |  |  |  |
| Sales forecast |  |  |  |  |  |

**Imputing missing values in each column with R**

**DATE: 1/17/2018**

* **have a variable in dataset that does not vary.** 
  + **Removing ‘nrOfPictures’ from the dataset file**
  + **Removed ‘seller’**
  + **Removed ‘offerType’**
* **Repaired or Not column values are replaced with booleans (1 or 0)**
* **Imputing boolean values to the repaired or not column**
* **Gearbox column having boolean values (manual or automatic), mapped manual = 0, automatic = 1**
* **Fuel type (benzin = 0, diesel = 1, lpg = 2, cng = 3, hybrid = 4, elektro = 5)**
* **abtest(test=0, control=1)**
* **Name column removed as it doesnt have any significance**
* **Model names are more than 2000+ out of 50k sample, so removed**
* **Dates related data need to be computed, so for time being not considering**

**{dara crawled, date created, last seen}**

**Vehicle type:**

**1 - andere**

**2 - suv**

**3 - coupe**

**4 - cabrio**

**5 - bus**

**6 - kombi**

**7 - kleinwagen**

**8 - limousine**

**Gearbox:**

**Manuell - 0**

**Automatik - 1**

**Model :**

**lada - 1**

**Lancia - 2**

**Rower - 63**

**Daewoo - 66**

**Saab - 69**

**Trabant - 87**

**Jaguar - 89**

**Daihatsu - 99**

**Subaru - 104**

**Land\_rover - 106**

**Jeep - 107**

**Chrysler - 187**

**Dacia - 119**

**Cheverolet - 241**

**Alpha\_romeo - 302**

**Porsche - 306**

**Suzuki - 317**

**Kia - 349**

**Honda - 373**

**Mitsubishi - 403**

**Volvo - 455**

**Mini - 460**

**Hyundai - 471**

**Sonstige\_autos - 544**

**Toyota - 620**

**Nissan - 686**

**Citroen - 741**

**Mazda - 786**

**Smart - 702**

**Skoda - 813**

**Seat - 995**

**Fiat - 1289**

**Peugeot - 1508**

**Renault - 2470**

**Ford - 3457**

**Audi - 4405**

**Mercedez\_benz - 4721**

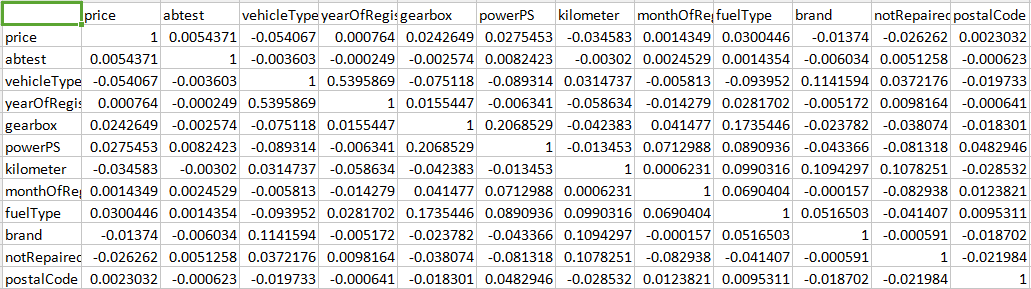
**Opel - 5409**

**Bmw - 5463**

**Volkswagen - 10510**

* **AmeliaView tool is being used for imputation of missing values**

**After converting 12 features to complete numeric values, correlation matrix is plotted**

****

**DATE: 1/18/2018**

**Original Dataset “as is” prediction summary()**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.400e+05 1.189e+05 3.700 0.000216 \*\*\*

abtest 5.331e+02 6.813e+02 0.782 0.433997

vehicleType -1.045e+03 1.955e+02 -5.344 9.18e-08 \*\*\*

yearOfRegistration -2.057e+02 5.921e+01 -3.474 0.000514 \*\*\*

gearbox 2.600e+03 8.492e+02 3.062 0.002201 \*\*

powerPS 2.457e+01 3.207e+00 7.660 1.91e-14 \*\*\*

kilometer -1.248e-01 9.486e-03 -13.153 < 2e-16 \*\*\*

monthOfRegistration 1.063e+02 9.798e+01 1.084 0.278161

fuelType0 3.425e+02 1.952e+03 0.176 0.860680

fuelType1 4.042e+03 2.033e+03 1.988 0.046837 \*

fuelType2 -1.074e+03 3.404e+03 -0.316 0.752356

fuelTypecng 1.490e+03 8.798e+03 0.169 0.865491

fuelTypeelektro -7.126e+03 2.092e+04 -0.341 0.733329

fuelTypehybrid 1.753e+03 1.046e+04 0.168 0.866843

brand -1.434e+02 4.936e+01 -2.906 0.003667 \*\*

notRepairedDamage -3.226e+03 1.084e+03 -2.977 0.002917 \*\*

---

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

Residual standard error: 58830 on 29864 degrees of freedom

(10132 observations deleted due to missingness)

Multiple R-squared: 0.01301, Adjusted R-squared: 0.01251

F-statistic: 26.24 on 15 and 29864 DF, p-value: < 2.2e-16

**After selecting starred features:**

Call:

lm(formula = price ~ vehicleType + yearOfRegistration + gearbox +

powerPS + kilometer + monthOfRegistration + fuelType + brand +

notRepairedDamage, data = tdata)

Residuals:

Min 1Q Median 3Q Max

-124156 -6604 -648 4036 12270282

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.646e+04 1.331e+04 -3.492 0.000480 \*\*\*

vehicleType -3.952e+03 3.356e+02 -11.775 < 2e-16 \*\*\*

yearOfRegistration 4.381e+01 6.807e+00 6.436 1.24e-10 \*\*\*

gearbox 2.240e+03 1.465e+03 1.529 0.126351

powerPS 1.082e+01 4.307e+00 2.513 0.011973 \*

kilometer -7.710e-02 1.487e-02 -5.185 2.17e-07 \*\*\*

monthOfRegistration -1.190e+02 1.582e+02 -0.752 0.452107

fuelType 4.973e+03 1.118e+03 4.448 8.68e-06 \*\*\*

brand -5.628e+01 8.429e+01 -0.668 0.504325

notRepairedDamage -6.889e+03 1.801e+03 -3.824 0.000132 \*\*\*

---

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

Residual standard error: 117000 on 40113 degrees of freedom

Multiple R-squared: 0.00671, Adjusted R-squared: 0.006488

F-statistic: 30.11 on 9 and 40113 DF, p-value: < 2.2e-16

> results$coefficients

(Intercept) vehicleType yearOfRegistration gearbox

-4.646349e+04 -3.951844e+03 4.380761e+01 2.240088e+03

powerPS kilometer monthOfRegistration fuelType

1.082309e+01 -7.710227e-02 -1.189792e+02 4.973287e+03

brand notRepairedDamage

-5.627890e+01 -6.888853e+03

>

> coef(results)

(Intercept) vehicleType yearOfRegistration gearbox

-4.646349e+04 -3.951844e+03 4.380761e+01 2.240088e+03

powerPS kilometer monthOfRegistration fuelType

1.082309e+01 -7.710227e-02 -1.189792e+02 4.973287e+03

brand notRepairedDamage

-5.627890e+01 -6.888853e+03

>

> #prediction

> pred <- predict(results, vdata)

>

> head(pred)

9 13 17 19 21 46

16018.683 7153.916 -10519.865 13895.402 20984.850 1753.634

> head(vdata$price)

[1] 14500 999 300 7550 10400 590

**After removing few more features R-Squared value increaed to 0.0066 which is equivalent to 6% accuracy**

lm(formula = price ~ vehicleType + yearOfRegistration + powerPS +

kilometer + fuelType + notRepairedDamage, data = tdata)

Residuals:

Min 1Q Median 3Q Max

-132745 -6689 -678 4166 12269933

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.984e+04 1.281e+04 -3.891 1.00e-04 \*\*\*

vehicleType -4.018e+03 3.319e+02 -12.107 < 2e-16 \*\*\*

yearOfRegistration 4.470e+01 6.779e+00 6.594 4.33e-11 \*\*\*

powerPS 1.188e+01 4.234e+00 2.807 0.005007 \*\*

kilometer -7.926e-02 1.478e-02 -5.361 8.32e-08 \*\*\*

fuelType 5.148e+03 1.100e+03 4.681 2.86e-06 \*\*\*

notRepairedDamage -6.788e+03 1.796e+03 -3.779 0.000158 \*\*\*

---

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

Residual standard error: 117000 on 40116 degrees of freedom

**Multiple R-squared: 0.006628**, Adjusted R-squared: 0.006479

F-statistic: 44.61 on 6 and 40116 DF, p-value: < 2.2e-16

the relationships between variables are ‘not linear’ or ‘weak linear’,

In this case one may follow three different ways: (i) try to linearize the relationship by transforming the data, (ii) fit polynomial or complex spline models to the data or (iii) fit non-linear functions to the data.

DATE: 1/19/2018

Non-linear , no

**Call:**

**lm(formula = price ~ yearOfRegistration, data = tdata)**

Residuals:

Min 1Q Median 3Q Max

-14234 -5997 -4197 53 12338539

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.370e+03 1.240e+04 0.433 0.665

yearOfRegistration 8.865e-01 6.178e+00 0.143 0.886

Residual standard error: 117500 on 40019 degrees of freedom

Multiple R-squared: 5.145e-07, Adjusted R-squared: -2.447e-05

F-statistic: 0.02059 on 1 and 40019 DF, p-value: 0.8859

**Individual parameters R-Square finding**

Price vs vehicleType:

Residual standard error: 117200 on 40059 degrees of freedom

Multiple R-squared: 0.003505, Adjusted R-squared: 0.00348

F-statistic: 140.9 on 1 and 40059 DF, p-value: < 2.2e-16

Price vs yearofRegistration

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 3.649e-07, Adjusted R-squared: -2.46e-05

F-statistic: 0.01462 on 1 and 40059 DF, p-value: 0.9038

Price vs gearbox

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.0005093, Adjusted R-squared: 0.0004843

F-statistic: 20.41 on 1 and 40059 DF, p-value: 6.266e-06

Price vs powerps

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.0007908, Adjusted R-squared: 0.0007659

F-statistic: 31.7 on 1 and 40059 DF, p-value: 1.807e-08

Price vs kilometer

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.001012, Adjusted R-squared: 0.0009866

F-statistic: 40.56 on 1 and 40059 DF, p-value: 1.924e-10

Price vs fueltype

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.0009527, Adjusted R-squared: 0.0009277

F-statistic: 38.2 on 1 and 40059 DF, p-value: 6.449e-10

Price vs brand

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.0001902, Adjusted R-squared: 0.0001652

F-statistic: 7.62 on 1 and 40059 DF, p-value: 0.005775

Price vs notRepairedDamage

Residual standard error: 117400 on 40059 degrees of freedom

Multiple R-squared: 0.0007005, Adjusted R-squared: 0.0006756

F-statistic: 28.08 on 1 and 40059 DF, p-value: 1.17e-07

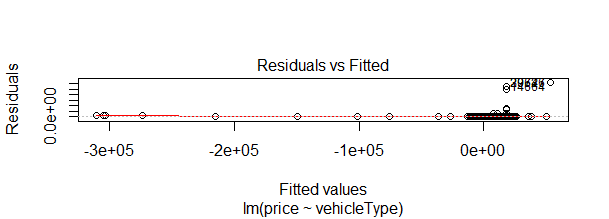
Price vs model

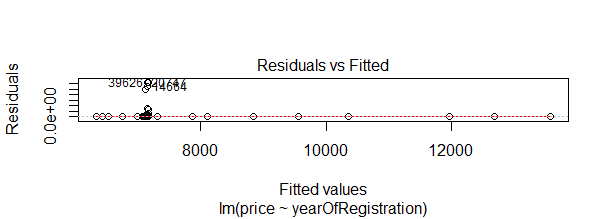
Residual standard error: 117400 on 40059 degrees of freedom

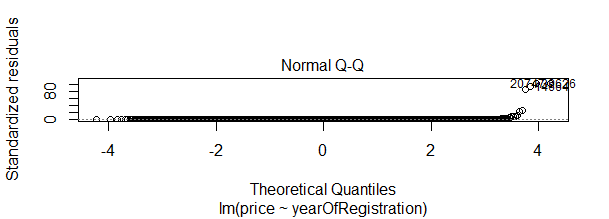
Multiple R-squared: 0.0007005, Adjusted R-squared: 0.0006756

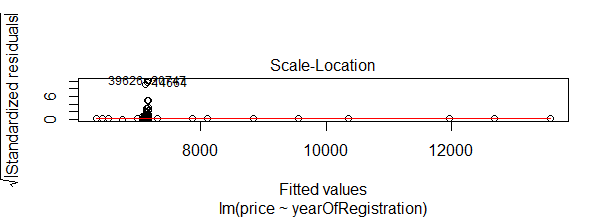
F-statistic: 28.08 on 1 and 40059 DF, p-value: 1.17e-07

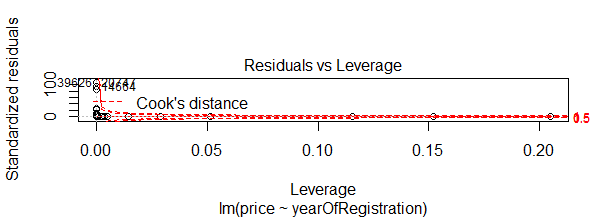
DATE: 1/22/2018











Cook’s distance is useful for identifying outliers in the X values (observations for predictor variables). It also shows the influence of each observation on the fitted response values. An observation with Cook’s distance larger than three times the mean Cook’s distance might be an outlier.

Quartiles

If a data set of values is arranged in ascending order of magnitude, then:

The median is the middle value of the data set.

The lower quartile (Q1) is the median of the lower half of the data set.

The upper quartile (Q3) is the median of the upper half of the data set.

The interquartile range (IQR) = Q3-Q1 is the spread of the middle 50% of the data values. So:

The interquartile range is a more useful measure of spread than the range as it describes the middle 50% of the data values.

Removal of outliers has no impact on the outcome

DATE: 1/23/2018

Min-max normalization

(X - min(X))/(max(X) - min(X))

Quantile Normalization:

lm(formula = price ~ vehicleType + yearOfRegistration + gearbox +

powerPS + kilometer + monthOfRegistration + fuelType + brand +

notRepairedDamage, data = tdata1)

Residuals:

Min 1Q Median 3Q Max

-104504 -1746 831 2451 1039233

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.235e+04 2.808e+02 43.989 < 2e-16 \*\*\*

vehicleType -1.200e-01 4.790e-03 -25.047 < 2e-16 \*\*\*

yearOfRegistration 3.270e-01 8.731e-03 37.457 < 2e-16 \*\*\*

gearbox 3.314e-02 4.121e-03 8.041 9.16e-16 \*\*\*

powerPS 1.034e-01 4.741e-03 21.806 < 2e-16 \*\*\*

kilometer -2.579e-01 1.223e-02 -21.089 < 2e-16 \*\*\*

monthOfRegistration 1.372e-01 7.953e-03 17.257 < 2e-16 \*\*\*

fuelType 8.414e-02 6.437e-03 13.071 < 2e-16 \*\*\*

brand -3.664e-02 9.212e-03 -3.977 6.99e-05 \*\*\*

notRepairedDamage -7.115e-02 4.361e-03 -16.315 < 2e-16 \*\*\*

---

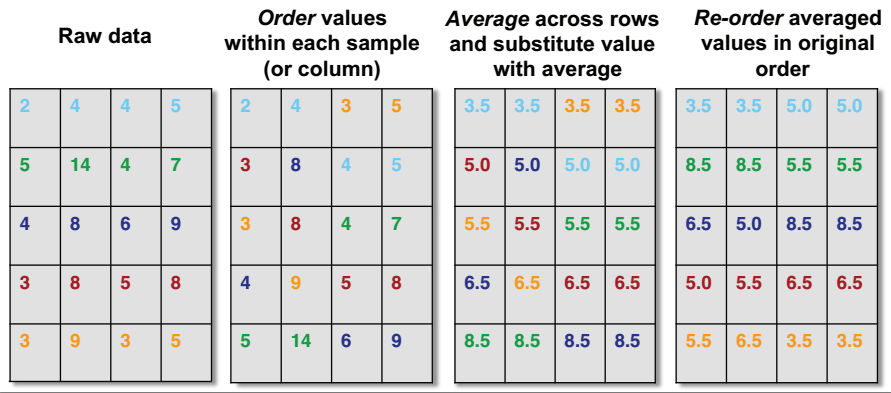
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9773 on 39962 degrees of freedom

Multiple R-squared: 0.1187, Adjusted R-squared: 0.1185

F-statistic: 598.1 on 9 and 39962 DF, p-value: < 2.2e-16

Explanation of quantile normalization:



DATE: 1/24/2018

Z-score normalization also called gaussian normalization didnt improve the accuracy..

lm(formula = price ~ vehicleType + yearOfRegistration + powerPS +

kilometer + fuelType + notRepairedDamage, data = tdata1)

Residuals:

Min 1Q Median 3Q Max

-1.722 -0.044 -0.006 0.024 116.874

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.002769 0.004102 -0.675 0.49960

vehicleType -0.050330 0.004960 -10.148 < 2e-16 \*\*\*

yearOfRegistration 0.026857 0.005032 5.338 9.46e-08 \*\*\*

powerPS 0.021423 0.004119 5.202 1.99e-07 \*\*\*

kilometer -0.029339 0.004179 -7.021 2.24e-12 \*\*\*

fuelType 0.020163 0.004187 4.815 1.47e-06 \*\*\*

notRepairedDamage -0.012099 0.004137 -2.925 0.00345 \*\*

---

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

Residual standard error: 0.8204 on 39993 degrees of freedom

Multiple R-squared: 0.006522, Adjusted R-squared: 0.006373

F-statistic: 43.76 on 6 and 39993 DF, p-value: < 2.2e-16

When applied quantile+guassian, the outcome its showing 11% accuracy

Min-Max normalization:

Call:

lm(formula = price ~ vehicleType + yearOfRegistration + powerPS +

kilometer + fuelType, data = tdata1)

Residuals:

Min 1Q Median 3Q Max

-0.01224 -0.00053 -0.00004 0.00032 0.99445

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.0013422 0.0005225 2.569 0.01020 \*

vehicleType -0.0358383 0.0029815 -12.020 < 2e-16 \*\*\*

yearOfRegistration 0.0323645 0.0050918 6.356 2.09e-10 \*\*\*

powerPS 0.0121578 0.0038414 3.165 0.00155 \*\*

kilometer -0.0009904 0.0001726 -5.739 9.60e-09 \*\*\*

fuelType 0.0027983 0.0005629 4.971 6.70e-07 \*\*\*

---

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

Residual standard error: 0.009456 on 40086 degrees of freedom

Multiple R-squared: 0.006183, Adjusted R-squared: 0.006059

F-statistic: 49.88 on 5 and 40086 DF, p-value: < 2.2e-16

First Applying Min\_Max then going for quantile normalization:

lm(formula = price ~ vehicleType + yearOfRegistration + powerPS +

kilometer + fuelType, data = tdata1)

Residuals:

Min 1Q Median 3Q Max

-0.64666 -0.04965 0.01311 0.06195 0.64465

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.233436 0.003713 62.86 <2e-16 \*\*\*

vehicleType -0.113931 0.004417 -25.80 <2e-16 \*\*\*

yearOfRegistration 0.416367 0.004170 99.85 <2e-16 \*\*\*

powerPS 0.347522 0.003134 110.89 <2e-16 \*\*\*

kilometer -0.379927 0.008955 -42.42 <2e-16 \*\*\*

fuelType 0.063786 0.003329 19.16 <2e-16 \*\*\*

---

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

Residual standard error: 0.1074 on 40075 degrees of freedom

Multiple R-squared: 0.4807, Adjusted R-squared: 0.4806

F-statistic: 7419 on 5 and 40075 DF, p-value: < 2.2e-16

With this we are getting 48% accuracy.

DATE: 1/25/2018

Standardization

DATE: 1/29/2018

Normalization of Square root on dataframe observed 28%

Residual standard error: 43.95 on 39914 degrees of freedom

Multiple R-squared: 0.2835, Adjusted R-squared: 0.2834

F-statistic: 3158 on 5 and 39914 DF, p-value: < 2.2e-16

Normalization using square root first, and then applying quantile is giving 58%

Residual standard error: 727.5 on 39934 degrees of freedom

Multiple R-squared: 0.5821, Adjusted R-squared: 0.582

F-statistic: 1.112e+04 on 5 and 39934 DF, p-value: < 2.2e-16

Root Mean Square Method:

RMS = sqrt(mean(x^2))

DATE: 1/30/2018

Demo preparation

Standardization

DATE: 1/31/2018

Least Absolute Shrinkage and Selection Operator (LASSO) performs regularization and variable selection on a given model

a measure of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).

DATE: 2/1/2018

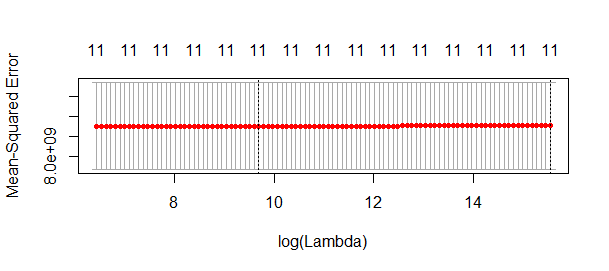
Regularization is shrinking some of the variable values to near possible zero or to near irrelavant

x a matrix with 10 columns

y a numeric vector x2 a matrix with 64 columns

DATE: 5/2/2018

Ridge Regression is a regularization method that tries to avoid overfitting, penalizing large coefficients through the L2 Norm. For this reason, it is also called L2 Regularization.



Terminology details:

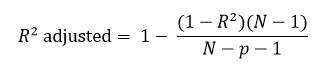
**R-Square**: It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). Mathematically, it can be written as:



The only drawback of R2 is that if new predictors (X) are added to our model, R2 only increases or remains constant but it never decreases. We can not judge that by increasing complexity of our model, are we making it more accurate?

That is why, we use “Adjusted R-Square”.

The Adjusted R-Square is the modified form of R-Square that has been adjusted for the number of predictors in the model. It incorporates model’s degree of freedom. The adjusted R-Square only increases if the new term improves the model accuracy.

where

R2 = Sample R square

p = Number of predictors

N = total sample size

Lasso plot

DATE: 2/8/2018

Linear regression is a classic supervised statistical technique for predictive modelling which is based on the linear hypothesis:  
  
y = mx + c  
where y is the response or outcome variable, m is the gradient of the linear trend-line, x is the predictor variable and c is the intercept. The intercept is the point on the y-axis where the value of the predictor x is zero.

Linear regression is a classic supervised statistical technique for predictive modelling which is based on the linear hypothesis:

## ***y = mx + c***

where *y* is the **response** or outcome variable, *m* is the **gradient** of the linear trend-line, *x* is the **predictor** variable and *c* is the **intercept**. The intercept is the point on the y-axis where the value of the predictor *x* is zero.

there are various methods of transforming non-linear data to make them appear more linear such as log and square root transformations

in order to fit a good model, appropriate values for the intercept and slope must be found.

In finding the appropriate values, the goal is to reduce the value of a statistic known as the mean squared error (MSE):  
  
MSE = Σ(y – y\_preds)² / n

The goal of the algorithm is to find the intercept and gradient values which correspond to the lowest possible MSE