Data Mining Project

Shirin BASHIRIAMID
https://github.com/sh-amid/Data-Mining

supervised by

Fabrice MUHLENBACH

March 2020

Abstract

As I am intent in photography and camera is an essential component of my interest, I have decided to consider crucial features in a camera set. This project aims to find the correlation between properties and the price of a camera and the model which defines the feature's behavior.
It turns out that it is categorized in supervised learning, meaning that We already know what our correct output should look like.
To find the likelihood of price and properties, "Regression" is an appropriate supervised approach that allows estimating the outcome variable (price) with a given data set based on dependent variables (features).

1 Introduction

Data is fuel to run a project and find results. In this project, I select Camera Data set from "*Kaggle.*"

Kaggle is a platform for predictive modeling in which companies and researches post data, and data scientists and data miners compete to find the best model for a given data set.

Petra Isenberg, Pierre Dragicevic, and Yvonne Jansen have gathered this data set, which describes 1000 cameras. Features describe: Model, Release date, Max resolution, Low resolution, Effective pixels, Zoom wide (W), Zoom Tele (T), Normal focus range, Macro focus range, Storage included, Weight (inc. batteries), Dimensions, Price.

The camera data set consists of 1038 unique values in an 84.92 KB file and is accessible by the link below:

https://www.kaggle.com/crawford/1000-cameras-dataset

2 Data Structure

First, I connect the data set with MySQL to RStudio. We use the "MySQL" database to store data; It is useful for Handling more data, Working with the data of different entities, using big data set, and so on. The number of values in the camera data set is not too much, but Using MYSQL is a good practice and experience to be familiar with how it works.

Due to preparation and cleansing the data set, I consider the summary of the data set. There is no NA value, but we have empty cells. It seems the data set has already cleaned; however, the ratio of empty cells to the whole data set is too small, and we can omit them.

The first column, "Model," is a characteristic feature and consists of the name of the brand and a unique number to identify the model of the camera. To make it easy to work, we create a new column and use the three first letters of the "Model" and call it "Brand."

3 Exploring Data

The "Cor" function is an excellent way to make a correlation matrix and deduce if there is any correlation between features (only for numeric features). To follow finding a relation between properties, we can create a scatter plot matrix with details for visual comparing.

The Scatter plot and Matrix show that there is no impressive correlation between some features except "Max.resolution," "Low.resolution," "Effective.pixels," and "Release.date"

By this result, I want to contemplate the role of these features in a model. In separate scatter plots, I examine the relationship between selected features and the "price" of the camera based on the Brand. Correlation between Max resolution and Effective pixels is significant, and it seems that some brands have nothing to present in these domains.

Principal Component Analysis(PCA) is a suitable algorithm to decrease the dimension of the data set. As the number of features is not huge, I expect that PCA would be useless. The created plot confirms this, and in general, if most of the correlation coefficients are smaller than 0.3, PCA will not help.

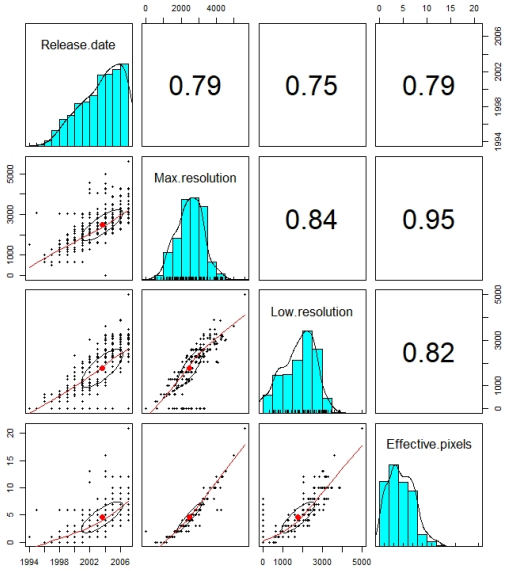

Figure 1: Impressive correlation between four features

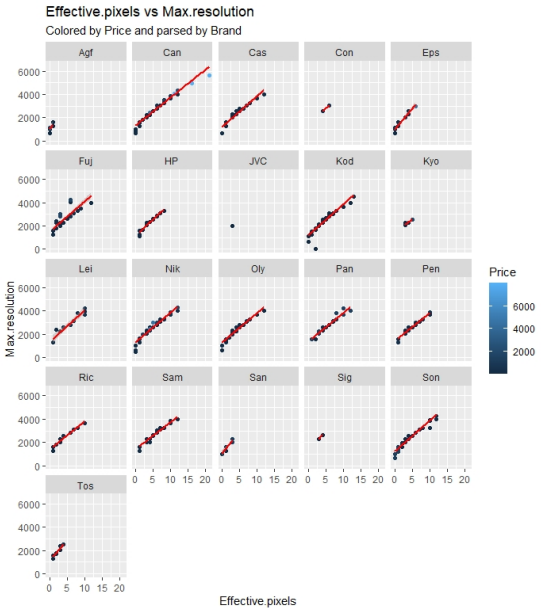

Figure 2: High correlation between price and 2 selected features based on the Brand

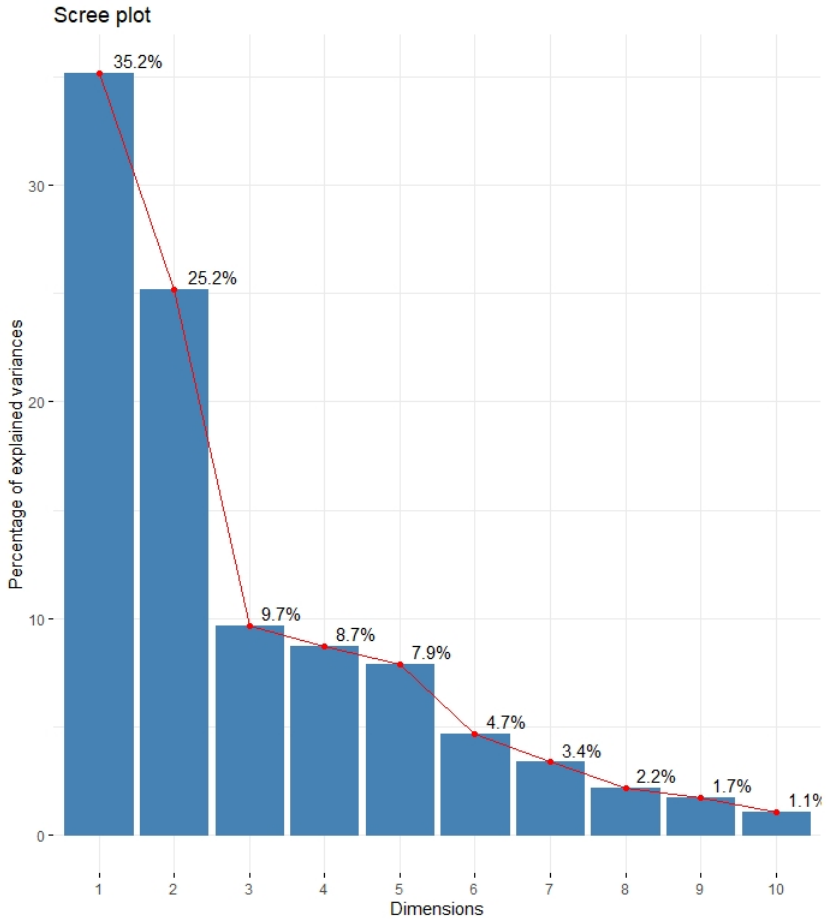

Figure 3: Percentage of explained variances in PCA

4 Regression

It is the time to examine the hypothesis, by a set of statistical process to build a regression model. General linear regression (glm) function fits the data set to gain a linear model.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

where y_i is an independent variable (Price), x_i are dependent variables (properties of the camera) and β_i are unknown parameters that model estimates.

In the summary of model, Values in Estimate column are β_i. The first model is :

Model.1 :

$$Price = (1.426e+04) - (6.747e+00)Release.date - (2.678e-01)Max.resolution +$$
$$(1.321e-01)Low.resolution + (8.396e+01)Effective.pixels + ...$$

Coefficients introduce features that have a significant effect on our model(Figure 4). To recognize their behavior, we look at the p-value that is better to close to zero ($p<0.05$). "*" symbols in the last column help us to recognize powerful features on the model.

The significant relationship between max and low resolutions and effective pixels was understandable from the covariance matrix while the model demonstrates other relevant features.

It gives the impression that meaningful correlation is not the only factor to play an essential role in a model. As if "Release date" has impressive correlation but not effective in the model, versus "Weight" or "Dimensions."

I expected the predicted model, was a time function. Now I understand that technology growth passes the time. Invention and upgrade the features are independent of time.

With all that, I am motivated to build the model based on highly correlated features.

Model.2

$$Price = (2.835e+05) + (1.398e-01)Max.resolution + (1.163e-01)Low.resolution +$$
$$(9.468e+01)Effective.pixels - (1.418e+02)Release.date$$

The conclusion shows "Max resolution" has a weak impact on the model that it is weird (Figure 5).

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2092.6   -248.6    -58.1    154.7   5253.0

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.426e+04  3.044e+04   0.468 0.639676
Release.date    -6.747e+00  1.522e+01  -0.443 0.657746
Max.resolution  -2.678e-01  1.007e-01  -2.659 0.007968 **
Low.resolution   1.321e-01  4.872e-02   2.711 0.006818 **
Effective.pixels 8.396e+01  2.380e+01   3.528 0.000438 ***
Zoom.w          -5.658e+00  3.754e+00  -1.507 0.132073
Zoom.t          -1.706e+00  2.586e-01  -6.596 6.81e-11 ***
Normal.focus     3.816e-01  1.118e+00   0.341 0.732840
Macro.focus      4.963e+00  2.779e+00   1.786 0.074397 .
Storage         -7.365e-02  7.283e-01  -0.101 0.919471
Weight           1.503e+00  1.522e-01   9.872  < 2e-16 ***
Dimensions      -3.060e+00  1.079e+00  -2.835 0.004677 **
BrandCan        -1.108e+01  2.409e+02  -0.046 0.963331
BrandCas        -1.207e+02  2.455e+02  -0.492 0.622901
BrandCon         3.665e+02  4.996e+02   0.734 0.463419
BrandEps         6.491e+01  2.773e+02   0.234 0.814930
BrandFuj        -1.574e+02  2.455e+02  -0.641 0.521464
BrandHP         -1.474e+02  2.526e+02  -0.584 0.559684
BrandJVC        -3.976e+02  4.843e+02  -0.821 0.411843
BrandKod        -5.200e+02  2.440e+02  -2.131 0.033331 *
BrandKyo         5.359e+02  2.803e+02   1.912 0.056173 .
BrandLei        -4.180e+02  2.986e+02  -1.400 0.161895
BrandNik        -3.467e+01  2.428e+02  -0.143 0.886469
BrandOly         1.546e+02  2.423e+02   0.638 0.523630
BrandPan         5.101e+02  2.506e+02   2.035 0.042106 *
BrandPen        -2.040e+02  2.461e+02  -0.829 0.407314
BrandRic         2.268e+02  2.637e+02   0.860 0.389850
BrandSam        -2.523e+02  2.554e+02  -0.988 0.323516
BrandSan         1.971e+02  3.150e+02   0.626 0.531630
BrandSig        -3.728e+02  3.894e+02  -0.957 0.338625
BrandSon         3.657e+01  2.400e+02   0.152 0.878920
BrandTos        -2.310e+02  2.737e+02  -0.844 0.398822
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 359642.7)

    Null deviance: 599685320  on 1037  degrees of freedom
Residual deviance: 361800589  on 1006  degrees of freedom
AIC: 16258
```

Figure 4: fit regression model to whole dataset

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1783.0   -318.5   -155.6     41.5   6457.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.835e+05  2.781e+04  10.197  < 2e-16 ***
Max.resolution   1.398e-01  1.042e-01   1.342 0.179953
Low.resolution   1.163e-01  5.057e-02   2.299 0.021731 *
Effective.pixels 9.468e+01  2.657e+01   3.564 0.000383 ***
Release.date    -1.418e+02  1.391e+01 -10.191  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 508573.9)

    Null deviance: 599685320  on 1037  degrees of freedom
Residual deviance: 525356871  on 1033  degrees of freedom
AIC: 16591

Number of Fisher Scoring iterations: 2
```

Figure 5: fit regression model based on four features

5 Improving model performance

Due to the summary of the second model and the residual and p-values, it does not look well-fitted. The "step-wise" regression function To evaluate the model, is helpful. The structure of this function is considering all features and their behavior to build the best model. The output is a regression model that introduced crucial factors.

Model.3

$$Price = (758.27744) - (0.27808)Max.resolution + (0.13044)Low.resolution+$$
$$(82.22101)Effective.pixels - (5.00030)Zoom.w + ...$$

The model denotes "Release date," "Normal focus range," and "Storage included" do not have meaningful effects on the model, and it omits them.

AIC value and Residual Deviance are essential factors to compares these three models. In comparison, the small amount of AIC and Residual Deviance define the behavior of the dataset well.

More data gives a better result; for the second model, we have cut off lots of information, and the conclusion is not acceptable.So far, the last Model has been the best one.

```
Call:
glm(formula = Price ~ Max.resolution + Low.resolution + Effective.pixels +
    Zoom.w + Zoom.t + Macro.focus + Weight + Dimensions + Brand,
    data = camera_data[2:14])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2113.7   -248.5    -58.9    156.4   5261.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       758.27744  307.04850   2.470 0.013692 *
Max.resolution     -0.27808    0.09764  -2.848 0.004489 **
Low.resolution      0.13044    0.04843   2.693 0.007193 **
Effective.pixels   82.22101   23.41179   3.512 0.000465 ***
Zoom.w             -5.00030    3.47300  -1.440 0.150245
Zoom.t             -1.73475    0.25010  -6.936 7.19e-12 ***
Macro.focus         5.26194    2.68090   1.963 0.049949 *
Weight              1.53327    0.13659  11.225  < 2e-16 ***
Dimensions         -3.03105    1.07637  -2.816 0.004958 **
BrandCan         -17.82188  239.10688  -0.075 0.940599
BrandCas        -121.94087  244.03962  -0.500 0.617412
BrandCon         393.76755  496.79036   0.793 0.428184
BrandEps          73.90686  275.88736   0.268 0.788840
BrandFuj        -151.98470  241.96663  -0.628 0.530066
BrandHP         -153.92057  249.87527  -0.616 0.538040
BrandJVC        -389.98597  482.83677  -0.808 0.419455
BrandKod        -522.79523  240.82082  -2.171 0.030172 *
BrandKyo         540.73778  278.23232   1.943 0.052236 .
BrandLei        -413.25690  297.10626  -1.391 0.164550
BrandNik         -40.96310  241.09189  -0.170 0.865118
BrandOly         156.00092  239.46323   0.651 0.514897
BrandPan         506.66868  247.86461   2.044 0.041199 *
BrandPen        -207.95485  244.07884  -0.852 0.394417
BrandRic         224.34600  262.68147   0.854 0.393274
BrandSam        -247.75098  248.09216  -0.999 0.318216
BrandSan         197.95954  312.89700   0.633 0.527095
BrandSig        -394.51207  384.52171  -1.026 0.305146
BrandSon          31.05378  238.31668   0.130 0.896351
BrandTos        -221.68823  270.66388  -0.819 0.412949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 358696.4)

    Null deviance: 599685320  on 1037  degrees of freedom
Residual deviance: 361924670  on 1009  degrees of freedom
AIC: 16253

Number of Fisher Scoring iterations: 2
```

Figure 6: fit step-wise regression model

6 Evaluating model performance

From the summary of the new model, it has the smallest value of AIC and Residual deviance. It means we are getting closer to predict a line to the camera data set.

Now we can predict the price of a camera with setting the features in our model.

257th data in camera data set presents with :

Max.Resolution = 2592 , Low.resolution = 2048 , Effective.pixels = 5 , Zoom.T = 38 , Zoom.w = 130 , Marco.focus = 9 , Weight = 180 , Dimension = 93 , Brand = Fuj

features are pasted to model and the result is amazing. The output gains 189.7013 while actual value of price is 199. Our model works well.

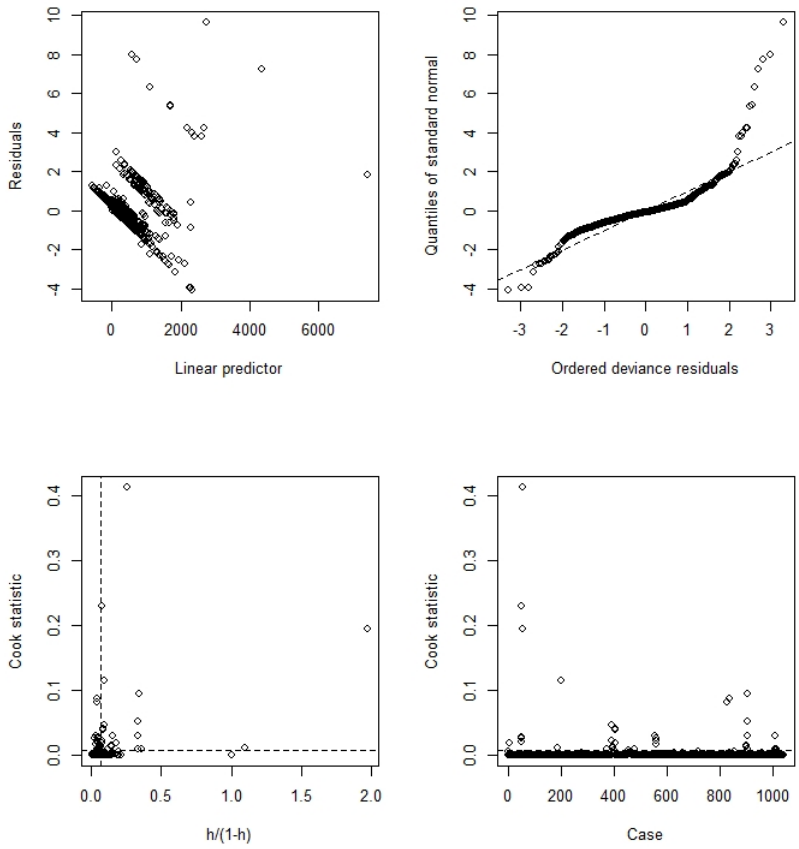

Figure 7: residuals in new model tend to line

7 Conclusion

In supervised learning, we have the idea that there is a relationship between the input and the output. The regression approach is used to predict the exact value for each example. The initial model has based on all variables then it will be improved in terms of the results and summary. Graphs have essential information in a short space that very useful to select the approach and start modeling.

In this case, we observed the high effect of Effective pixels, Zoom(T) and Weight which Zoom has a negative effect on the price. Logically, Weight and Effective pixels are essential components for customers. It makes sense that photographers tend to have a light camera with an easy carry-on. Effective pixel and resolution are part of the most critical components to rank the camera.

Polynomial linear regression for low and max resolution neutralizes the negative effect of Zoom(T). We consider this issue by residuals that in our final model, tend to the line.

The dramatic growth of technology since 2000 and upgrade features fast shows the release date of products is not essential. Camera dataset covers data till 2007 and it will be impressively changed if we gathered data till 2020.

" brand" does not have very profoundly affect our model, and it demonstrates that access to technology becomes possible for all companies, and all have the chance to get the market.In supervised learning, we have the idea that there is a relationship between the input and the output. The regression approach is used to predict the exact value for each example. The initial

model has based on all variables then the model improves in terms of the results and summary. Graphs have essential information in a short space that very useful to select the approach and start modeling.

In this case, we observed the high effect of Effective pixels, Zoom(T) and Weight which Zoom has a negative effect on the price. Logically, Weight and Effective pixels are essential components for customers. It makes sense that photographers tend to have a light camera with an easy carry-on. Effective pixel and resolution are part of the most critical components to rank the camera.

The dramatic growth of technology since 2000 and upgrade features fast shows the release date of products is not essential. Camera dataset covers data till 2007 and it will be impressively changed if we gathered data till 2020.

" brand" does not have very profoundly affect our model, and it demonstrates that access to technology becomes possible for all companies, and all have the chance to get the market. procedure to find and fit the model takes 8920 ms and it considers all features to build a model.The result will be different if we tend to find the effect of features individually.

References

www.kaggle.come

Machine Learning with R - Brett Lantz

www.tutorials.iq.harvard.edu

www.sthda.com

www.cran.r-project.org