31-10-2019

Report; ***Customer Brand Preferences***

Gerhard Westerbeek

Blackwell data analyst department

INDEX

[Summary 3](#_Toc23427380)

[Key findings 3](#_Toc23427381)

[Conclusion 4](#_Toc23427382)

[Appendix 5](#_Toc23427383)

# Summary

The purpose of this report is to generate an insight into the prediction of “Brand” using classification analysis.

# key findings

1. ***Data structure & Data quality***

The used dataset for training the model consists out of 9898 xisting products and includes 7attributes (features). The Dataset Complete Responses does not contain missing data. The datatypes of “Car”, “Zipcode” and “Brand” were changed to Integer using as.factor function. The datatype “elevel” was changed to Integer using the as.ordered function. The data of “Salary”, ”Age”, “Credit”, “Zipcode”, and “Elevel” are equally distributed. Brand 1 (6154) is more sold then brand 0 (3744). See Appendix for related histograms. Running scatterplots based on various feature combinations it seems that there is a relation between the choice of brand and the features age and salary. See Appendix for the related scatterplot.

Furthermore reviewing the data in combination with a boxplot, it should be marked that the distribution of salary between brand 0 and 1 is significantly different. See Appendix.

1. ***Train and assess the models***

For predicting the brand that customers prefer four algorithms (C5.0, Random Forest, KNN and SVM) were used and assessed using the performance indicators Accuracy and Kappa. The models were trained on a set of 75% of the total dataset. Engineering was done based on Tunelenght (C5.0, KNN and SVM) and the mtrygrid for Random Forest. See below the result matrix regarding the algorithms in combination with the performance metrics Accuracy and Kappa after running the prediction function on the testset (25% of total dataset).

|  |  |  |
| --- | --- | --- |
|  | No Filter | |
|  | Accuracy | Kappa |
| C5.0 | 0,9215 | 0,8348 |
| RF | 0,9223 | 0,8357 |
| Knn | 0,6546 | 0,2095 |
| Svm | 0,6216 | 0,000 |

It seems that algorithm RF will predict the “Brand” most accurate based on the used metrics.

Using the VarImp function it shows that the features “Salary” and “Age” correlate most with the predicted feature “Brand”. See appendix for overview of the VarImp for the RF model.

1. ***Brand prediction using the selected model***

The “Brand” prediction has been based on the algorithm with the best performance metrics; Random Forest. We used this model

# Conclusion

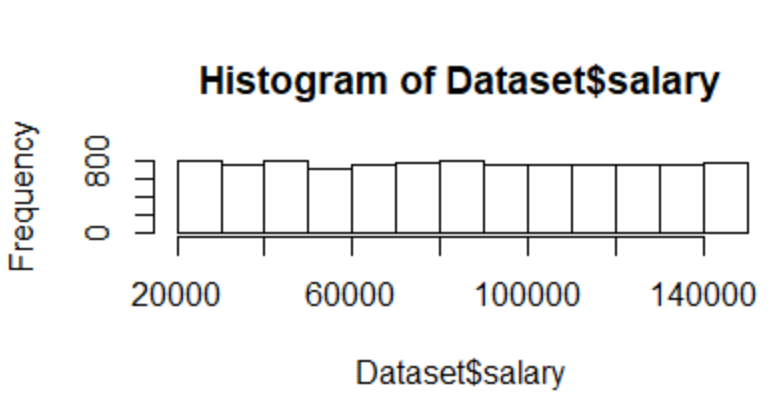
# 

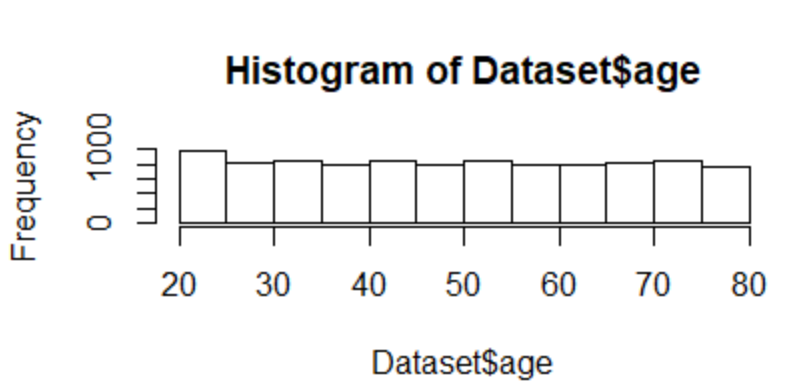
# Appendix

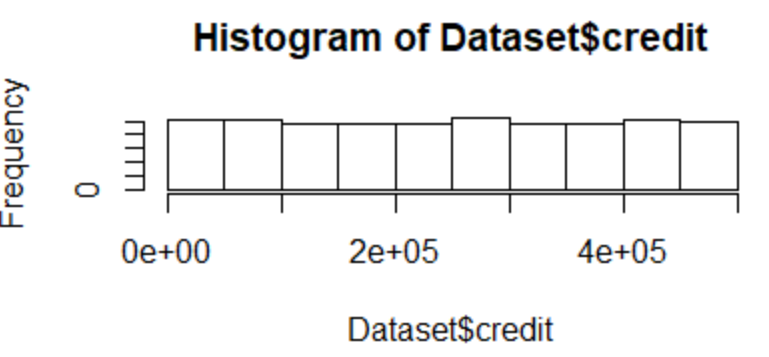
APPENDIX

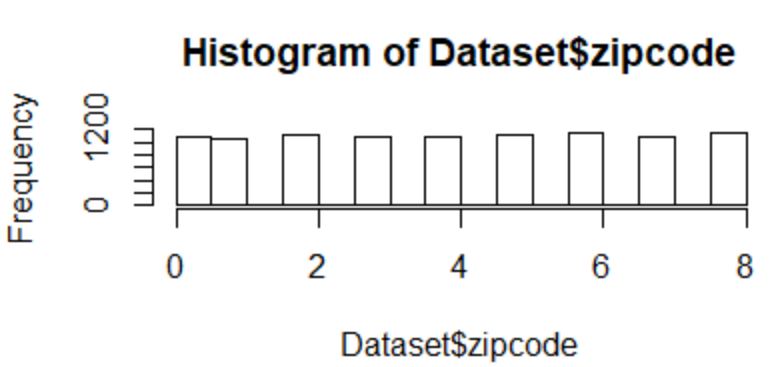
Preprocessing data;

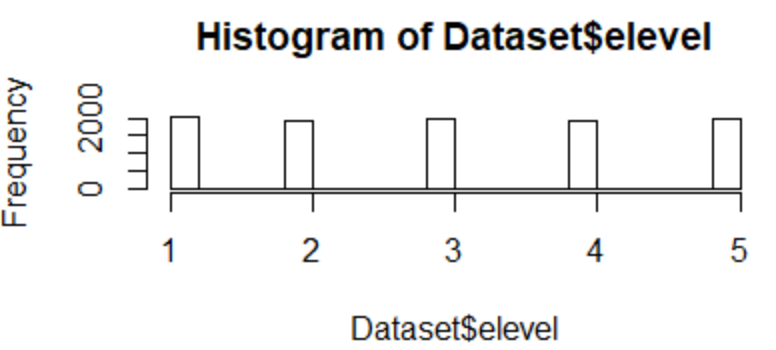
Histograms;



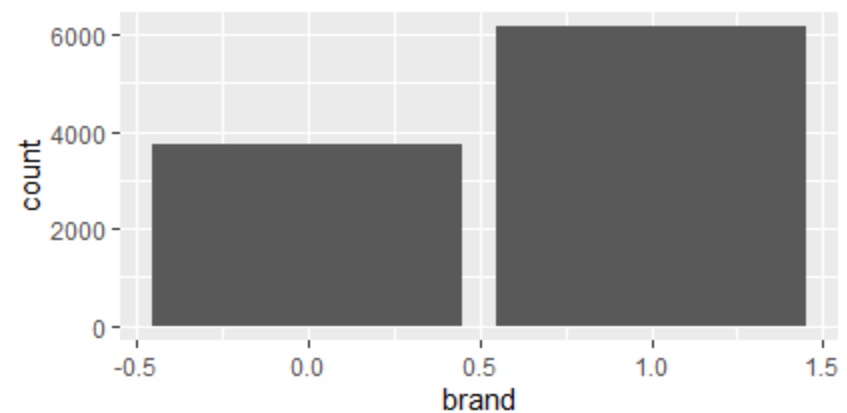




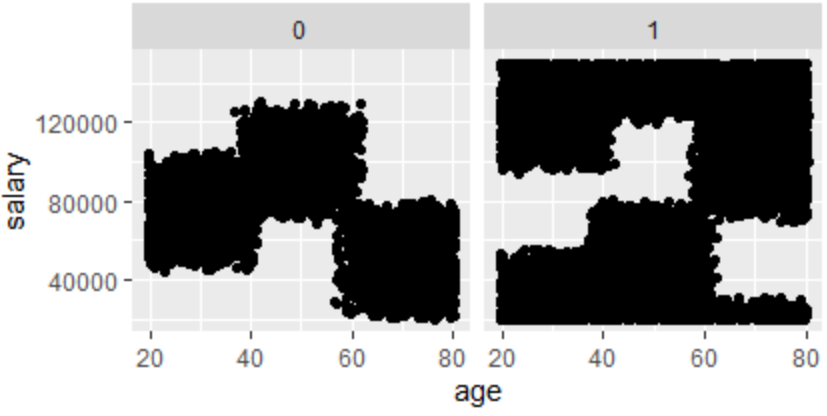




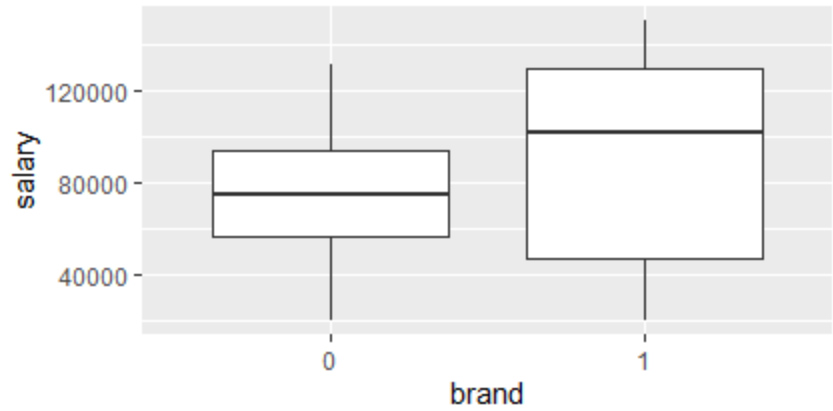
Bargraph Brand count



Scatterplot Age, Salary and brand



Boxplot Brand – Salary



Changing datatypes;

# Change datatypes ####

Dataset$elevel<- as.ordered(Dataset$elevel)

str(Dataset)

Dataset$car<- as.factor(Dataset$car)

str(Dataset)

Dataset$zipcode<- as.factor(Dataset$zipcode)

str(Dataset)

Dataset$brand<- as.factor(Dataset$brand)

Train model C5.0

str(Dataset)

library(caret)

library(lattice)

install.packages("C50")

library(C50)

install.packages("inum")

library(inum)

set.seed(123)

inTrain <- createDataPartition(y = Dataset$brand, p=.75,

list=FALSE)

training <- Dataset[ inTrain,]

testing <- Dataset[-inTrain,]

ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 1)

C5.0Fit <- train(brand ~ .,

data = training,

method = "C5.0",

tuneLength = 2,

trControl = ctrl)

C5.0Fit

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 1 times)

Summary of sample sizes: 6681, 6683, 6681, 6681, 6683, 6682, ...

Resampling results across tuning parameters:

model winnow trials Accuracy Kappa

rules FALSE 1 0.8445581 0.6853060

rules FALSE 10 0.9155452 0.8190757

rules TRUE 1 0.8650456 0.7236350

rules TRUE 10 0.9170281 0.8220150

tree FALSE 1 0.8420000 0.6759178

tree FALSE 10 0.9171621 0.8241264

tree TRUE 1 0.8623518 0.7136691

tree TRUE 10 0.9158166 0.8213564

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were trials = 10, model = tree and winnow

= FALSE.

# Train model Random Forest

library(randomForest)

ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 1)

mtryGrid <- expand.grid(mtry = 1)

RFFit <- train(brand ~ .,

data = training,

method = "rf",

tunegrid = mtryGrid,

trControl = ctrl)

RFFit

mtry =1

mtry Accuracy Kappa

2 0.6216326 -0.0002689458

18 0.9193192 0.8287900374

34 0.9171636 0.8241845844

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

mtry=2

mtry Accuracy Kappa

2 0.6219020 0.0006141832

18 0.9202590 0.8306250942

34 0.9175657 0.8248849793

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

mtry=3

mtry Accuracy Kappa

2 0.6219019 0.0004421814

18 0.9209294 0.8320118737

34 0.9162149 0.8221800997

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

mtry=4

mtry Accuracy Kappa

2 0.6221710 0.001668515

18 0.9190457 0.828002282

34 0.9143310 0.818051971

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

mtry=5

mtry Accuracy Kappa

2 0.6220371 0.001229642

18 0.9197179 0.829452178

34 0.9176980 0.825077449

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

mtry=6

mtry Accuracy Kappa

2 0.6219022 0.0004439929

18 0.9217413 0.8337726968

34 0.9179713 0.8256666833

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 18.

Variable Importance Random Forest

> RFFitImp<- varImp(RFFit, scale = FALSE)

> RFFitImp

rf variable importance

only 20 most important variables shown (out of 34)

Overall

salary 1744.23

age 1158.22

credit 169.59

elevel.L 32.15

elevel.C 31.68

elevel^4 29.41

elevel.Q 22.43

zipcode4 16.01

zipcode6 15.25

zipcode2 14.59

zipcode7 14.18

zipcode3 13.78

zipcode1 13.61

zipcode5 12.98

car7 12.48

zipcode8 12.43

car15 12.17

car17 12.05

car10 11.58

car12 11.45

