Supervised learning - Expenditure

Lorenzo de Sario

17/8/2020

ANALYSIS OF THE EXPENDITURE

Abstract

The Marketing campaign is one of the most relevant data sources for a company that need to be dug deeper when the issues are effectiveness, allocation of the budget among different actions or return on investment. Through this report, I'm going to evaluate how predictors, related to the customers characteristics, and different strategies are going to predict the amount spent ($). Starting with an Exploratory Data Analysis to understand main patterns of the data, I proceed evaluating models in the regression analysis, and applying different resampling methods in order to understand the goodness of the models. This analysis is related to building models of Supervised Learning, in which for each independent variables the response variable is measured, with the scope of evaluating the best model in terms of predictability.

Dataset and goals

The original dataset was extracted from Kaggle and modelled in reason of the goal of this analysis. It could be useful for any classification problem or regression analysis. Indeed, it is composed by 11 variables mixed between categorical, ordinal and continuous predictors. The dataset includes 500 observation and is characterized by several missing values which could be generated by the human error when recording each observation. Specifically, the greatest amount of missing data came from the categorical variable History of purchases for which the null values have been omitted wherefore the mode of this variable is Na. In contrast, for the ordinal variables the NA values are substituted with the median that is a robust measure to outliers. The objective of this analysis is to exploit the accuracy of different supervised learning models such as linear regression, multiple and non linear regression, decision trees and infer some of the significative measures of quality of the fit with different resampling methods.

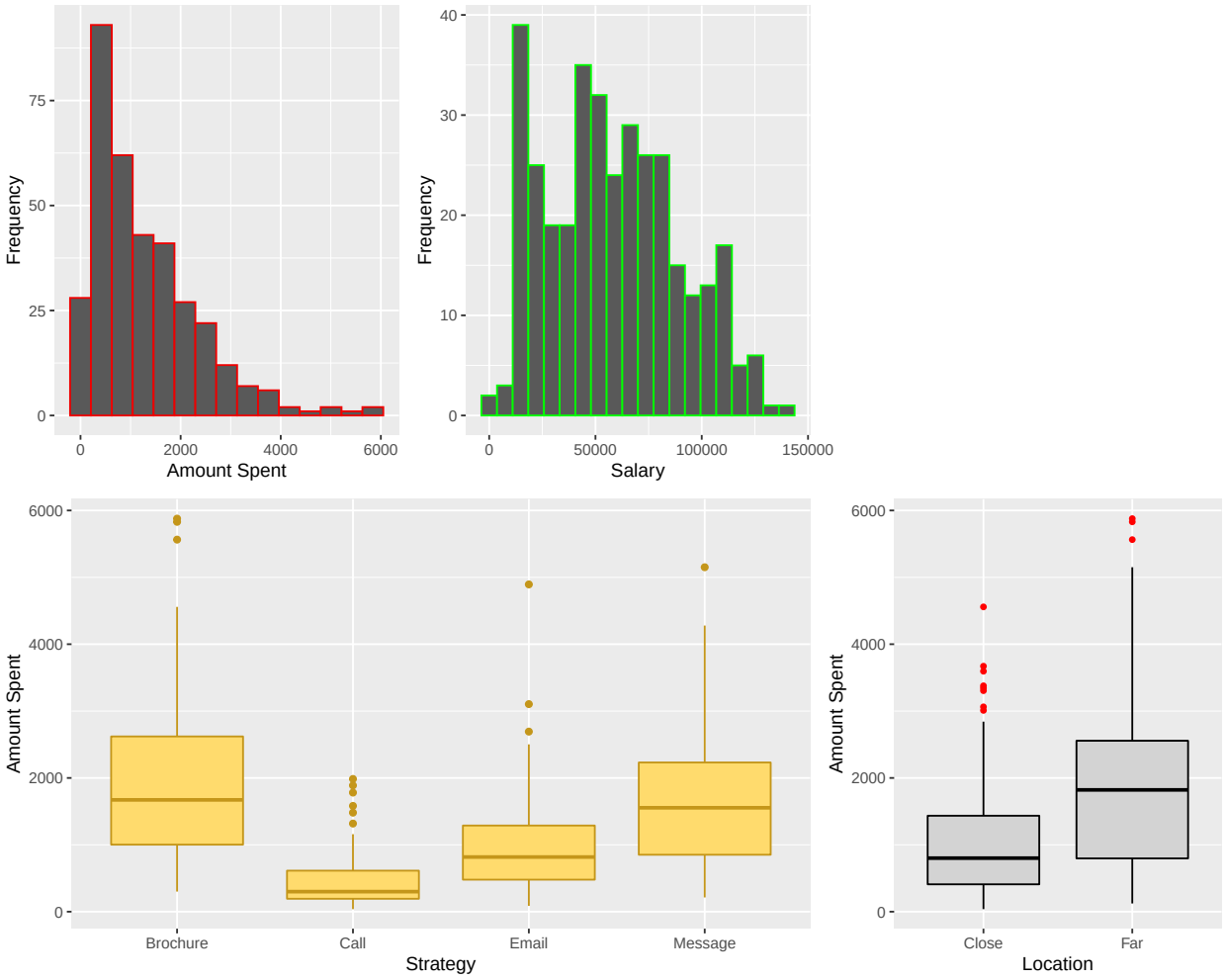
- GOALS:
  - Determine the degree of freedom of the natural splines toward finding the best fitting line between the dependent variable (Amount Spent) and the dependent Salary;
  - Run bootstrap in order to establish the variability of the adjusted R-squared and to identify the best model according to different fitted regression models;
  - Use of the full linear regression model so as to predict the Amount Spend, determine the RMSE and check for the multicollinearity between variables;
  - Employ forward stepwise selection for the feature selection task and choose a model with the maximum adjusted R-square or the minimum validation error;
  - Apply decision tree and determine its adjusted R-squared in order to make a comparison with the previous applied methods. Prune the tree and evaluate the RMSE.

```
##      Age         Gender    OwnHome       Married     Location
##  Middle:183   Female:186   Own :193   Married:191   Close:243
##  Old   : 80   Male  :163   Rent:156   Single :158   Far  :106
##  Young : 86
##
##
##
##      Salary          Children      History     AmountSpent       Strategy
##  Min.   :     0   Min.   :0.00   High  :140   Min.   :  38   Brochure:90
##  1st Qu.: 33000   1st Qu.:0.00   Low   :112   1st Qu.: 470   Call    :82
##  Median : 55500   Median :0.00   Medium: 97   Median : 962   Email   :86
##  Mean   : 58216   Mean   :0.86                Mean   :1278   Message :91
##  3rd Qu.: 80800   3rd Qu.:2.00                3rd Qu.:1841
##  Max.   :140000   Max.   :3.00                Max.   :5878
##   Health_index
##  Min.   :1.00
##  1st Qu.:1.00
##  Median :2.00
##  Mean   :2.11
##  3rd Qu.:3.00
##  Max.   :3.00

## [1] 704
```

Including Plots

By looking at the histogram plotting the distribution of the dependent variable (Amount Spent), is possible to note right skeweness, that is mean, medain and mode differs and moreover that the mean is greather than the mode ($1278 > 704$). Skeweness could affect the statistical model by the presence of outliers. Indeed, by looking at the residual plots of the linear relation between the Amount Spent and the Salary, note that there is heteroscedasticity, that is non constant variance of the error terms. This is described as an increase in the variance of the error terms when the value of the response increase. Is it possible to observe graphically that when the response is log transformed or by taking the square root of the variable, the heteoscedasticity is reduced. As an useful information for the business owner, by looking at the boxplot of the qualitative variables, the median expenditure for those insividuals that have received a brochure and that lives faraway from the shop is higher as well as their maximum capacity of expense.

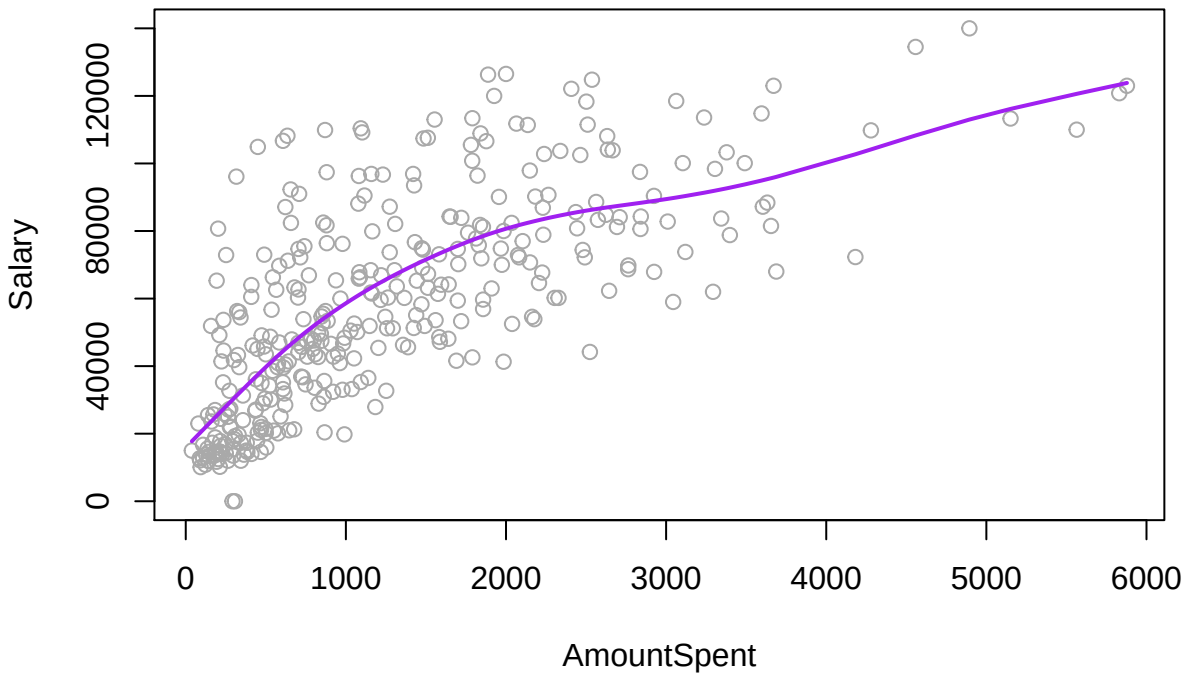


Smoothing splines

Smoothing splines, differently from the natural splines, are characterized when is applied, in order to find the function g(smoothing spline), a minimization of the residual sum of squares

$$min \sum (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

subject to a smoothness nonnegative penalty term over the variability of g. When the matter is choosing the Smoothing Parameter the best method is the leave-one-out cross validation. In such a case the smoothness level is choosen by cross validation and thus a lambda with 5,3 degree of freedom (thought as level of flexibility of the model: higher lambda corresponds to more flexible models; while in the bias-variance trade-off, for higher values of lambda we get lower bias and high variance).




```
## [1] 5.34
```

```
## [1] 0.00838
```

Bootstrap analysis over the Adjusted R-squared

$$R^2_{adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

In order to select the best regression model that is able to explain the largest proportion of variance in the Amount Spent using the p predictors and to estimate the accuracy of this statistic, is applied the bootstrap resampling method. The bootstrap generates B new bootstrap datasets from the original one by sampling with replacement (the same n observations can appear more the once in the dataset), and for each one is possible to generate a new esitmate of the parameter. Thus is possible to estimate measure of precision of the parameter by the standard error. Without any assumption about the distribution of the k samples, the result will be a nonparametric bootstrap. To do that build a function that takes as parameters the data and the index by which the subset, namely the bootstrap sample, will be generated. From each bootstrap sample (k = 5000) will be carried out the statistic of interest, that is the Adjusted R-squared and relative SE(R2-adj). By choosing this measure, instead of the R-squared, is possible to consider more robust results since it has been adjusted for the number of p predictors considered in the model. The models that are selected for this kind of analysis evaluate the predictor and relative response variables in differents functional forms. The function takes into account as models for the Amount Spent: - regression of a polynomial function of the Salary; - regression with the multilevel categorical variable (Strategy); - regress the salary plus an interaction term between the Salary and the Strategy; - regress natural cubic splines with five degree of freedom of the Salary. Splines guarantee greater flexibility than the polynomial regression and works by dividing the range

of X (Salary) in K regions to which is applied a constrained polynomial function in order to fit the data also at the frontier (knots). The number of knoots (or degree of freedom) is chosen by looking at the best curve, while a more technical approach would be use the cross-validation as did for the smoothing spline regression. - regression with increasing number of variables up to the full model. What is possible to observe is that the the adjusted-R2 is increasing in the number of variables that are added to the model due to more accuracy of the fitting. But if the increse in the R2, due to more predictors, is not substantial, including more variables will be likely to lead to overfitting (eg. model_S.S.C and model_S.S.C.G gets the same adjusted R-squared).

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = data, statistic = Adj_R_squared, R = 5000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*    0.554 -0.000377      0.0336
## t2*    0.567  0.002138      0.0340
## t3*    0.261  0.006828      0.0341
## t4*    0.644  0.005493      0.0311
## t5*    0.653  0.011225      0.0313
## t6*    0.835  0.002291      0.0140
## t7*    0.835  0.002724      0.0139
## t8*    0.894  0.003665      0.0103
```

Linear regression for prediction

So as to estimate the test error, that is the average error arising from the fitted model to predict the dependent variable (Amount Spent) on new observations, is adopted the Validation Set Approach. This method implies random splitting the dataset in two parts: training and validation set. The model is fitted on the training set and used for predict the response for the observation of the validation set. Then the perfomance are evaluated adopting the Mean Squared Error (MSE). For this purpose is adopted the root of the MSE (RMSE), a measure that gives relatively high weight to large errors. In order to check for multicollinearity beteween the data is computed the variance inflation factor (VIF) for each variable. The presence of collinearity, that is values of the VIF that exeeds 5 would give rise to problem of redundancy of the variables for which the result is verified. This values is given by the ratio between the variance of the coefficient when the variable is fitted to the full model and when it is fitted alone. Is possible to observe that the variable History presents high VIF, therefore if this variable is dropped from the model, from the previous results, the reduction in adjusted R-squared would be of the order of 0.877-0.835 = 0.04 thus, affecting the quality of the model.

```
##              GVIF Df GVIF^(1/(2*Df))
## Age          1.80  1            1.34
## Gender       1.18  1            1.09
## OwnHome      1.36  1            1.17
## Married      1.83  1            1.35
## Location     1.13  1            1.06
## Salary       3.63  1            1.91
## Children     1.24  1            1.11
## History      1.60  1            1.26
## Strategy     1.39  3            1.06
```

```
## Health_index 1.03  1           1.02
```

```
##                GVIF    Df GVIF^(1/(2*Df))
## Age           FALSE FALSE           FALSE
## Gender        FALSE FALSE           FALSE
## OwnHome       FALSE FALSE           FALSE
## Married       FALSE FALSE           FALSE
## Location      FALSE FALSE           FALSE
## Salary        FALSE FALSE           FALSE
## Children      FALSE FALSE           FALSE
## History       FALSE FALSE           FALSE
## Strategy      FALSE FALSE           FALSE
## Health_index FALSE FALSE           FALSE
```

```
## [1] 1633
```

Forward Stepwise Selection

An automatic procedure computationally efficient in order to perform feature selection so as to gain in interpretability of the model in multiple regression are the stepwise methods. In particular, the forward method include in the model, starting from the null model with no predictors (when the coefficients of the multiple regression are all sets to 0, therefore considering the null hypotesis

$$\beta_1 = \beta_2 = \beta_3 = ... = \beta_p = 0$$

), one variables at the time that gives the lowest residual sum of square(RSS). The problems of this method are those variables that might be included in a stage and at the later ones becomes redundant. This matter could be easily solved by the Mixed selection that starting from a forward procedure, combine a backward selection in order to remove those variables that give rise to high p-values. While the R squared cannot be used for selection with different number of p predictors, the model selection is performed with a validation Set Approach, that is by choosing the minimum validation error obtained from prediction, with a maximum size of the subset of 10. By looking at the minium validation error the resulting model includes 9 variables, excluding the variable that stand for the Age of the consumer, while the model that owns the higher adjusted R- squared is the one that takes into account all the 10 variables to predict the amount spent. Other methods for selecting among models with a different number of variables are the Mallows' Cp, BIC and AIC. For example by applying a minimization of the Cp (penalty of

$$2d\hat{\sigma}$$

to the training RSS) to the fitted least squares models, is selected a model of 8 variables and are excluded predictors such as Age, OwnHome, Married Status.

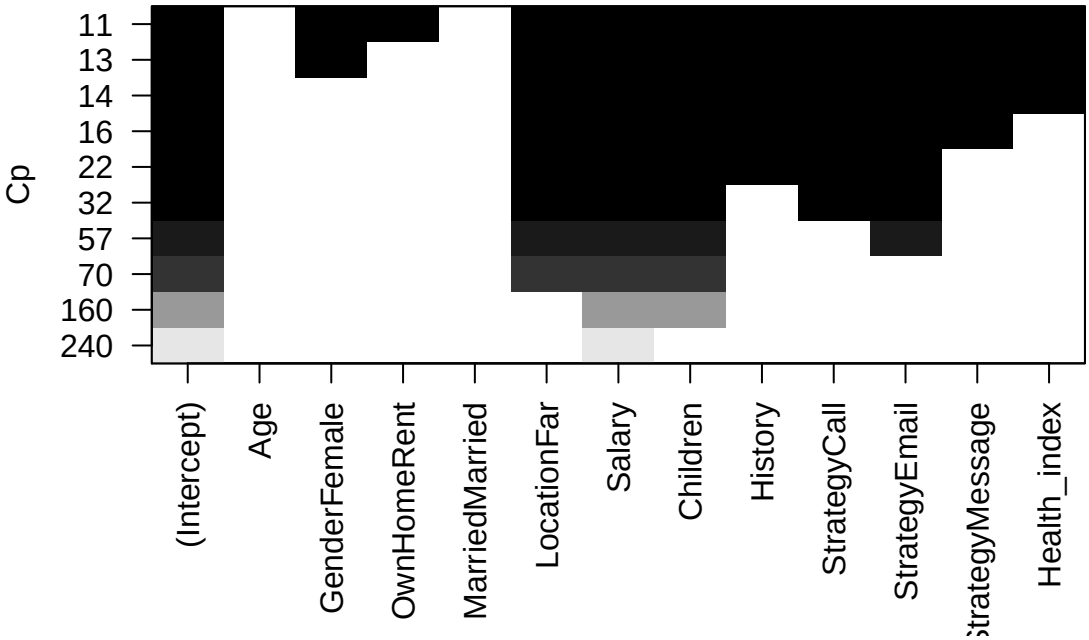
```
## [1] 9
```

```
##     (Intercept)    GenderFemale     LocationFar          Salary        Children
##        657.8399        137.0958        667.5355          0.0196       -303.0200
##         History    StrategyCall   StrategyEmail StrategyMessage    Health_index
##        -172.0612       -704.0356       -582.1934       -280.6880         82.1077
```

```
## [1] 10
```

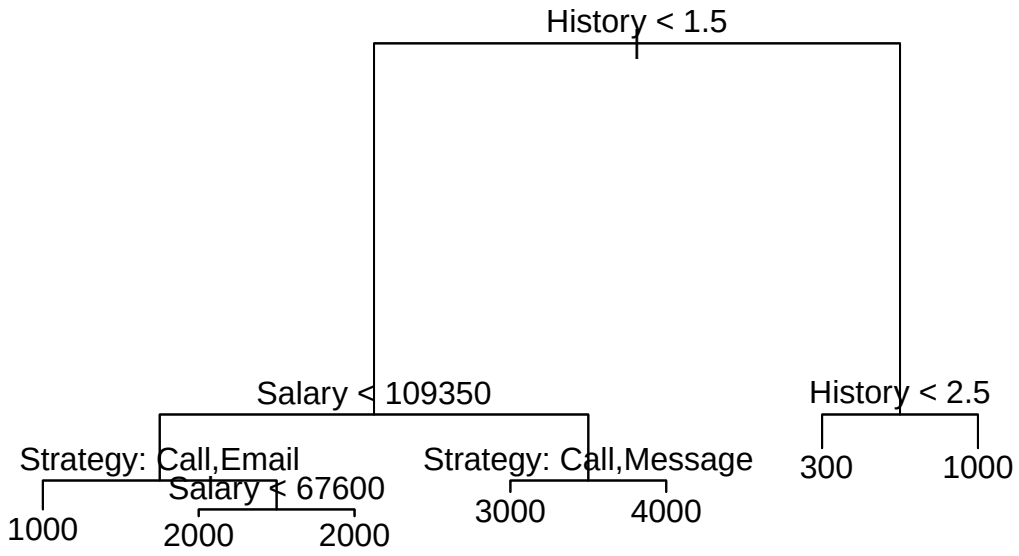
```
## Subset selection object
## Call: regsubsets.formula(AmountSpent ~ ., data = data[train, ], method = "forward",
##     nvmax = 10)
## 12 Variables  (and intercept)
##                  Forced in Forced out
## Age                  FALSE      FALSE
## GenderFemale         FALSE      FALSE
## OwnHomeRent          FALSE      FALSE
## MarriedMarried       FALSE      FALSE
## LocationFar          FALSE      FALSE
## Salary               FALSE      FALSE
## Children             FALSE      FALSE
## History              FALSE      FALSE
## StrategyCall         FALSE      FALSE
## StrategyEmail        FALSE      FALSE
## StrategyMessage      FALSE      FALSE
## Health_index         FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: forward
##           Age GenderFemale OwnHomeRent MarriedMarried LocationFar Salary
## 1  ( 1 )  " " " "          " "          " "             " "         "*"
## 2  ( 1 )  " " " "          " "          " "             " "         "*"
## 3  ( 1 )  " " " "          " "          " "             "*"         "*"
## 4  ( 1 )  " " " "          " "          " "             "*"         "*"
## 5  ( 1 )  " " " "          " "          " "             "*"         "*"
## 6  ( 1 )  " " " "          " "          " "             "*"         "*"
## 7  ( 1 )  " " " "          " "          " "             "*"         "*"
## 8  ( 1 )  " " " "          " "          " "             "*"         "*"
## 9  ( 1 )  " " "*"          " "          " "             "*"         "*"
## 10  ( 1 ) " " "*"          "*"          " "             "*"         "*"
##           Children History StrategyCall StrategyEmail StrategyMessage
## 1  ( 1 )  " "      " "     " "           " "           " "
## 2  ( 1 )  "*"      " "     " "           " "           " "
## 3  ( 1 )  "*"      " "     " "           " "           " "
## 4  ( 1 )  "*"      " "     " "           "*"           " "
## 5  ( 1 )  "*"      " "     "*"           "*"           " "
## 6  ( 1 )  "*"      "*"     "*"           "*"           " "
## 7  ( 1 )  "*"      "*"     "*"           "*"           "*"
## 8  ( 1 )  "*"      "*"     "*"           "*"           "*"
## 9  ( 1 )  "*"      "*"     "*"           "*"           "*"
## 10  ( 1 ) "*"      "*"     "*"           "*"           "*"
##           Health_index
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  " "
## 7  ( 1 )  " "
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"

## [1] 10
```



Decision tree Thought as a series of splliting rules made according to the best split, in an approach top-down, the decision tree is a widely applied method due to its great interpretability. To exploit it once more is applied the Validation Set Approach, splitted the dataset in training and validation set, and saved the response in order to check for the difference between the prediction made by the tree based model and the test value. Adopting 6 variables, by looking at the RMSE of the decision tree for regression against the full regression linear model, notice that it is minimized when the tree method is applied. This could be due to non-linear and complex relationship between response and predictors. For further predictive accuracy of the decision tree, methods such as bagging, random forest and bagging could be introduced. For interpretation of a split the following holds: left hand branch is pursued if the condition is true. Thus, for example, those individuals who have an History of expenditure low will spend 400$, while those with an high history of expenses in the shop and with a salary greater than 109350 with at least 1 child will spend 3000 dollars. Moreover is possible to note that strategies such as call and email will not maximize the expenditure so as the brochure and messages strategies.

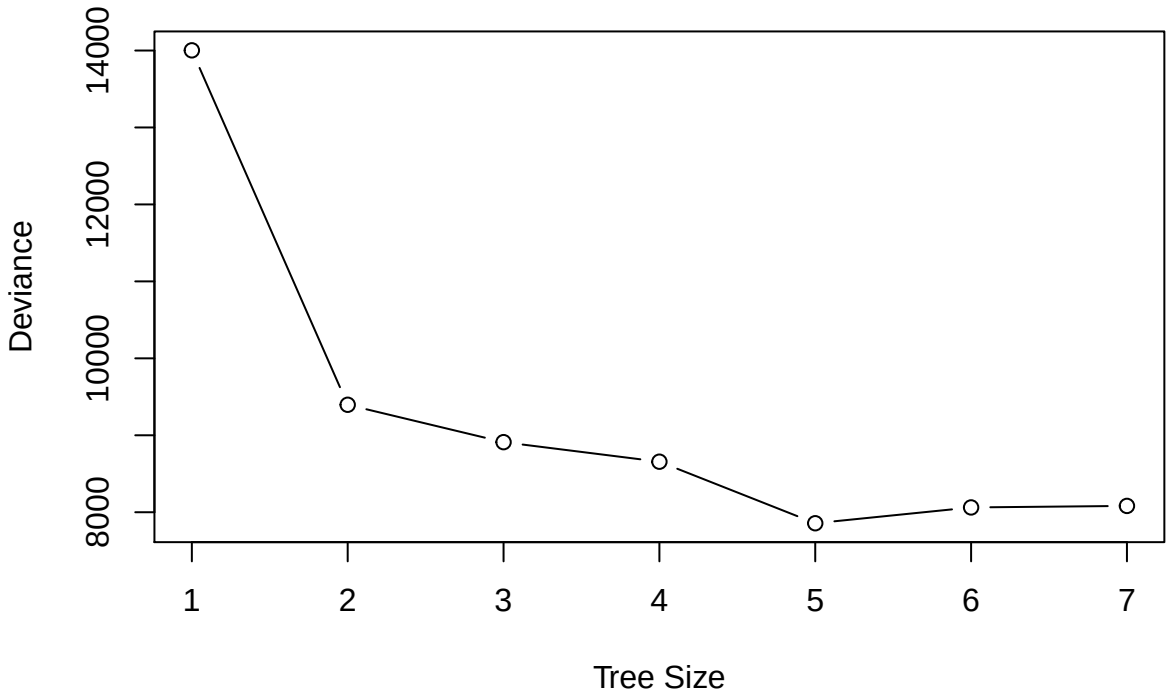
```
##
## Regression tree:
## tree(formula = AmountSpent ~ ., data = data[train, ])
## Variables actually used in tree construction:
## [1] "History"  "Salary"   "Strategy"
## Number of terminal nodes:  7
## Residual mean deviance:  236000 = 39200000 / 166
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1570    -248     -35       0     244    1850
```




```
## [1] 572
```

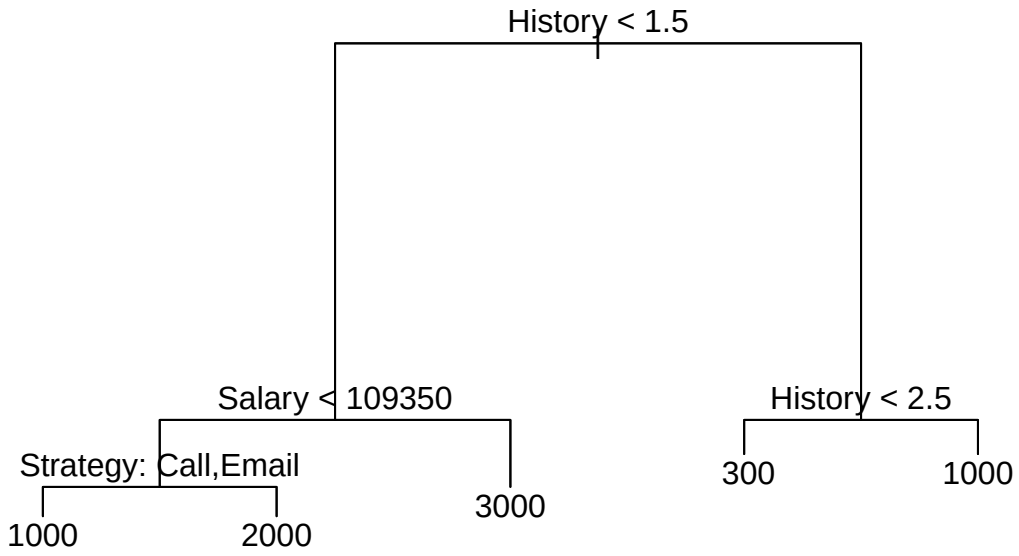
Pruning

In order to select a subtree that possess the lowest test error given the one builded in the previous section, the pruning technique is adopted. In particular, the test error is estimated using the validation set approach and the optimal level of tree complexity is performed with cross-validation. The plot of the of the deviance (SSE) with respect to the size tell us that the minimum SSE is at the tree with size 5 that corresponds to index 3 RMSE. Thus the pruning is at the size 5, for which is reached a RMSE of 572 which is higher than the tree builded before. Thus the best model in term of accuracy of the prediction of the Amount spent is the tree based one.




```
## [1] 3
```

```
## [1] 5
```




```
## [1] 572
```

Take aways- Conclusions

As the smoothing splines results in ~5 degree of freedom and a value of lambda that tends to 0, the curvature of the curve in the relation between the Amount Spent and the Salary is not relevant. Thus with the LOOCV is reached a function that interpolate between the selected data points and is choosen the ones with minimum average curvature. Moreover, by running bootstrap on different regression models in order to make inference on the adjusted R-squared coefficients notice that this value is increasing in the number of dependent variables added to the model and its variability (standard deviation) is decreasing in the increasing number of parameters. The worst model that results in the lowest adjusted R-squared is the ones that simply regress the amount spent over the strategy. By looking at the linear regression model that take into account all the predictors to determine the response is reached the 89.4% of variance explained. Another technique, is the feedforward regression for feature selection for which towards the lowest validation error select 9 variables to predict the response. Therefore from a comparison of the fitted model with the results gained from the standard decision tree and the pruned one, untill this point is possible to determine that the best model for RMSE is the linear regression. Different results or improvement in the quality of fitting of the decision tree could be reached with applications of the random forests or boosting method.

Appendix

```
library(tidyr)
library(ggplot2)
```

```r
library(leaps)
library(tidyverse)
library(dummies)
library(boot)
library(splines)
library(tree)
#load data
data = read.csv("C:/Users/Lorenzo de Sario/Desktop/marketing/MarketingData1.csv",
                header = TRUE, sep = ",")
summary(data)
#switch off scientific notation
options(scipen=999, digits = 3)

data = data %>%
  drop_na(History)

#replace missing int with the median
data = data %>%
  group_by(Children) %>%
  mutate(AmountSpent=ifelse(is.na(AmountSpent),
                            median(AmountSpent, na.rm=TRUE), AmountSpent)) %>%
  as.data.frame()

attach(data)
#Exploratory Data Analysis
#Quantitative variables
hist = ggplot(data, mapping = aes(x = data$AmountSpent))
hist +
  geom_histogram(color = "red2", bins = 15)+
   ylab('Frequency') + xlab('Amount Spent')

hist1 = ggplot(data, mapping = aes(x = data$Salary))
hist1 +
  geom_histogram(color = "green1", bins = 20) +
  ylab('Frequency')
#Qualitative variables and amount spent
box = ggplot(data, mapping = aes(x = data$Strategy, y = data$AmountSpent))
box +
  geom_boxplot(fill = "#FFDB6D", color = "#C4961A") +
  ylab('Amount Spent') + xlab('Strategy')

box1 = ggplot(data, mapping = aes(x = data$Location, y = data$AmountSpent))
box1 +
  geom_boxplot(fill = "lightgray", color = "black", outlier.colour = 'red', outlier.shape = 16) +
  ylab('Amount Spent') + xlab('Location')

box2 = ggplot(data, mapping = aes(x = data$History, y = data$AmountSpent))
box2 +
  geom_boxplot(fill = "#D16103", color = "#4E84C4") +
  ylab('Amount Spent') + xlab('History')

#choosing reference categories. That is modelling with respect to
#those individuals who are targetted by a Brochure, that own a home and are single men.
```

```r
data = data %>% mutate(Strategy = relevel(Strategy, ref = "Brochure"))
data = data %>% mutate(Gender = relevel(Gender, ref = "Male"))
data = data %>% mutate(Married = relevel(Married, ref = "Single"))
data = data %>% mutate(OwnHome = relevel(OwnHome, ref = "Own"))

 #Encoding Ordinal categorical variable. Assume equidistance between the levels
data$Age = as.numeric(data$Age,
                      labels = c("Old", "Middle", "Young"),
                      levels = c(1, 2, 3))
data$History = as.numeric(data$History,
                          labels = c("High", "Medium", "Low"),
                          levels = c(1, 2, 3))

library(gridExtra)
library(grid)
hist =  ggplotGrob(hist)
hist1 = ggplotGrob(hist1)
box = ggplotGrob(box)
box1 = ggplotGrob(box1)
grid.arrange(hist, hist1, box, box1, ncol = 4, nrow = 4,
             layout_matrix = rbind(c(1, 2, NA), c(3, 3, 4)))

#fitting splines and selecting the smoothing parameter by the LOO CV
fit <- smooth.spline(AmountSpent, Salary, cv = TRUE)
plot(AmountSpent, Salary, col="darkgrey")
lines(fit, col="purple", lwd=2)

model_sal <- lm(log(AmountSpent)~ Salary, data = data)

#Bootstrapping Adjusted R-squared linear regression
library(car)

set.seed(123) #reproducibility of the experiment

Adj_R_squared = function(data, indices){
  marketing = data[indices,] #creating my sample with replacement

  model_sal <- lm(log(AmountSpent)~ Salary, data = marketing)
  AR.s = summary(model_sal)$adj.r.squared

  model_sal1 <- lm(sqrt(AmountSpent)~ Salary + I(Salary^2) + I(Salary^3) , data = marketing)
  AR.s1 = summary(model_sal1)$adj.r.squared

  model_strat <- lm(AmountSpent~ Strategy, data = marketing)
  AR.st = summary(model_strat)$adj.r.squared

  model_child <- lm(AmountSpent~ Salary + Salary:Strategy, data = marketing)
  AR.c = summary(model_child)$adj.r.squared

  model_S.S <- lm(AmountSpent~ ns(Salary, df = 5) + Strategy + Salary:Strategy, data = marketing)
  AR.s.s = summary(model_S.S)$adj.r.squared

  model_S.S.C <- lm(log(AmountSpent)~ Salary + Strategy + Children, data = marketing)
```

```
  AR.s.s.c = summary(model_S.S.C)$adj.r.squared

  model_S.S.C.G <- lm(log(AmountSpent)~ Salary + Strategy + Children + History, data = marketing)
  AR.s.s.c.g = summary(model_S.S.C.G)$adj.r.squared

  full_model <- lm(log(AmountSpent)~. , data = marketing)
  AR.full = summary(full_model)$adj.r.squared

  Adjusted_r_sq = c(AR.s, AR.s1, AR.st, AR.c, AR.s.s, AR.s.s.c, AR.s.s.c.g, AR.full)
  return(Adjusted_r_sq)
}

#use the boot() function to compute variability
#SE(adjR^2) of 5000 bootstrap estimates for the Adjusted R squared
boot(data, Adj_R_squared, R = 5000)

#make prediction
set.seed(456)
train = sample(c(TRUE, FALSE), nrow(data), rep = TRUE)
test = (!train)
full_model <- lm(log(AmountSpent)~. , data = data[train,])
vif(full_model)
sqrt(vif(full_model))>1.5
#MSE sqrt between real data and test set
sqrt(mean((AmountSpent -predict(full_model, data))[-train]^2))

#Forward stepwise regression - model selection  (p.245)
library(leaps)
#Choosing a model with validation set approach: test error
set.seed(789)
#Forward Regression
fwd <- regsubsets(AmountSpent ~., data = data[train,], method = "forward", nvmax = 10)
test_matrix = model.matrix(AmountSpent~., data = data[-train,])
validation_error = rep(NA, 10)
for (i in 1:10){
  coeff = coef(fwd, id = i)
  predicted = test_matrix[,names(coeff)]%*%coeff
  validation_error[i] = mean((data$AmountSpent[-train]- predicted)^2)
}

which.min(validation_error) #model with best validation error
coef(fwd,9) #the best model contains 9 variables
summary(fwd)
plot(fwd ,scale = "Cp")

reg.summary = summary(fwd)
which.max(reg.summary$adjr2)

#decision tree
set.seed(101112)
test = data[test,]
testY = test$AmountSpent
tree <- tree(AmountSpent~., data[train,])
```

```r
summary(tree)
plot(tree)
text(tree, pretty = 0)
tree_prediction = predict(tree, test)
sqrt(mean((tree_prediction - testY)^2)) #is better MSE sqrt in regression
#high sqrt since high values of Y

#cv for pruning the tree
cv_tree = cv.tree(tree)
plot(cv_tree$size, sqrt(cv_tree$dev), type = 'b', xlab = 'Tree Size', ylab = 'sqrt MSE')
which.min(sqrt(cv_tree$dev)) #index 3
cv_tree$size[3] #size 5

# prune the tree to size 5
prune <- prune.tree(tree, best = 5)
plot(prune)
text(prune, pretty = 0)
tree.pred = predict(prune, test)
sqrt(mean((tree.pred - testY)^2)) #by pruning there is no increase in the mean squared error
```