Datasættets første parameter "customer_id" kan udelades, da det er en identifikator, der ikke tjener noget formål i en model, der kan forudsige, om en bruger er konverteret eller ej.
Parameteren "credit_account_id" er en hash og vil derfor være en kategorisk variabel. Den har 148 unikke værdier, hvor de fleste opstår 1-4 gange, men en enkelt af værdierne kan findes i datasættet 687 gange. Det er umiddelbart svært at se, hvordan den skulle kunne bidrage til modellen, og parameteren ekskluderes derfor fra datasættet.

Det er fundet at parameteren "age" mangler 177 værdier. Ud af 891 rækker, er det en ret stor andel, og derfor er de manglende værdier i parameteren imputed. Da mean imputation vil gøre at 177 af rækkerne får samme alder, hvilket formegentlig ikke vil være retvisende i et datasæt hvor parameteren spænder mellem 0.42 og 80, er missForest valgt som imputation-metode, da den imputer værdier ved at kigge på hele datasættet.

Da target-parameteren er binær, laves der en logistisk regressionsmodel på datasættet. Et summary af modellen viser at parameterene "customer_segment", "gender", "age" og "related_customers" har lave p-værdier, hvilket antyder, at der er en stærk relation mellem disse parametre og hvorvidt en bruger er konverteret eller ej.

```
> model <- glm(dataset$converted~., data = dataset, family = binomial(link = "logit"))
>
> summary(model)

Call:
glm(formula = dataset$converted ~ ., family = binomial(link = "logit"),
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7572  -0.5950  -0.4023   0.6242   2.5094

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.060e+01  6.042e+02   0.051 0.959602
customer_segment  -1.239e+00  1.516e-01  -8.171 3.06e-16 ***
gendermale        -2.668e+00  2.022e-01 -13.190  < 2e-16 ***
age               -4.857e-02  8.169e-03  -5.945 2.77e-09 ***
related_customers -3.853e-01  1.111e-01  -3.468 0.000524 ***
family_size       -8.818e-02  1.211e-01  -0.728 0.466488
initial_fee_level  7.343e-04  1.181e-03   0.622 0.534218
branchHelsinki    -1.276e+01  6.042e+02  -0.021 0.983152
branchTampere     -1.239e+01  6.042e+02  -0.021 0.983642
branchTurku       -1.240e+01  6.042e+02  -0.021 0.983625
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Summary af den logistiske regressionsmodel, som viser, at modellen har en lav p-værdi for parametrene "customer_segment", "gender", "age" og "related_customers".

En ANOVA test, altså en analyse af variansen, viser, at forskellen mellem *null deviance* og *residual deviance*[1] er størst, når parameterene "customer_segment", "gender", "age" og "related_customers" tilføjes modellen.

```
> anova(model, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: dataset$converted

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                890    1186.66
customer_segment   1  102.254       889    1084.40 < 2.2e-16 ***
gender             1  257.206       888     827.20 < 2.2e-16 ***
age                1   28.946       887     798.25 7.442e-08 ***
related_customers  1   19.861       886     778.39 8.328e-06 ***
family_size        1    0.485       885     777.90    0.4862
initial_fee_level  1    1.023       884     776.88    0.3117
branch             3    3.707       881     773.17    0.2948
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: ANOVA test af den logistiske regressionsmodel, som viser at modellens *residual deviance* falder mest, når parametrene "customer_segment", "gender", "age" og "related_customers" introduceres.

En anden metode er at kigge på forskellen i modellens forklaringsgrad. Hvis forklaringsgradens forøgelse er stor, når en parameter tilføjes, sammenlignet med en model, hvor alle parametre undtaget den undersøgte er inkluderet, så fortæller det, at parameteren er vigtig for modellen. Da det ikke er muligt at lave en standard R^2-mål på en logistisk regressionsmodel, bruges i stedet McFadden R^2 indeks, som er den mest brugte erstatning herfor.

customer_segment	gender	age	related_customers	family_size	initial_fee_level	branch
0.05819323	0.1820019	0.03276349	0.012201	0.0004543226	0.0003408775	0

Figure 3: Tabel som viser, hvor meget McFadden R^2 indekset stiger, når den enkelte parameter er tilføjet modellen, sammenlignet med en model, hvor den undersøgte paramter er holdt ude.

[1] *Deviance* er et "goodness-of-fit"-mål, som man bruger til statistiske modeller, som man ofte bruger i generaliserede lineære modeller, som en logistisk regression er.

Resultaterne bekræfter, hvad der blev fundet i ANOVA testen; at de vigtigste parametre for forudsigelse om en bruger er konverteret eller ej er "gender", "customer_segment", "age" og "related_customers" i den nævnte rækkefølge.

Kode

```
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))


library(dplyr)
library(pscl)
library(missMethods)
library(randomForest)
library(missForest)

#Load data and exclude the parameter "customer_id"
dataset <- read.csv("case_data.csv", header = T)
dataset <- as.data.frame(dataset[,2:10])

#make sure, that the target variable is a boolean
dataset[,1] <- as.factor(dataset[,1])

#Investigate the dataset
nrow(dataset)
sapply(dataset, n_distinct)

data.frame(table(dataset[,8]))
#Exclude the parameter "credit_account_id" according to the resons
    mentioned in the report
dataset <- dataset[,-8]

#Handle missing data
colnames(dataset)[colSums(is.na(dataset)) > 0]
nrow(dataset[is.na(dataset$age),])
#dataset <- impute_mean(dataset, type = "columnwise")
set.seed(1234)
dataset <- missForest(dataset)$ximp

#Build the logistic regression model
model <- glm(dataset$converted~., data = dataset, family = binomial(link
    = "logit"))

summary(model)
anova(model, test = "Chisq")

#Find the McFadden R^2 index for when all parameters are used
McFadden_full <- pR2(model)[4]

#Initiate a table for the increasements in the McFadden R^2 indexes
```

```
r_sqrd_increasement <- matrix(0, nrow=1, ncol = ncol(dataset)-1)
colnames(r_sqrd_increasement) <- colnames(dataset[,2:ncol(dataset)])

#Loop that build a logistic regression model where each of the parameters
    are excluded and then find the increasement i the McFadden R^2 index
for(i in 2:(ncol(dataset)-1)){
  reduced_model <- glm(converted~., data = dataset[,-i], family =
      binomial(link = "logit"))
  #print( McFadden_full - pR2(reduced_model)[4])
  r_sqrd_increasement[1,i-1] <- McFadden_full - pR2(reduced_model)[4]
}

r_sqrd_increasement
```

case_code.R