Case 1 Pilgrim Bank

Jonathan Ratschat, Franziska Bülck

20.10.2019

Preparation of dataset

Importing dataset and formating variables

```
library(readxl)

#read xls file
Data <- read_xls("Data_Pilgrim_Case-Part-A.xls")

#change colnames
colnames(Data)[2:7] <-c("Profit","Online","Age","Inc","Tenure","District")

#change data types
Data$ID <- as.factor(Data$ID)
Data$Online <- as.factor(Data$Online)
Data$Age <- as.factor(Data$Age)
Data$Inc <- as.factor(Data$Inc)
Data$District <- as.factor(Data$District)
```

Exploring Data

```
str(Data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    31634 obs. of  7 variables:
##  $ ID      : Factor w/ 31634 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Profit  : num  21 -6 -49 -4 -61 -38 -19 59 493 -158 ...
##  $ Online  : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ Age     : Factor w/ 7 levels "1","2","3","4",..: NA 6 5 NA 2 NA 3 5 4 6 ...
##  $ Inc     : Factor w/ 9 levels "1","2","3","4",..: NA 3 5 NA 9 3 1 8 9 8 ...
##  $ Tenure  : num  6.33 29.5 26.41 2.25 9.91 ...
##  $ District: Factor w/ 3 levels "1100","1200",..: 2 2 1 2 2 3 3 2 2 1 ...
```

```
summary(Data)
```

```
##        ID             Profit       Online         Age            Inc      
##  1      :    1   Min.   :-221.0   0:27780   3      :5390   6      :5413  
##  2      :    1   1st Qu.: -34.0   1: 3854   4      :5376   7      :3152  
##  3      :    1   Median :   9.0             2      :3650   9      :2960  
##  4      :    1   Mean   : 111.5             5      :3236   3      :2571  
##  5      :    1   3rd Qu.: 164.0             7      :2693   5      :2369  
##  6      :    1   Max.   :2071.0             (Other):3000   (Other):6908  
##  (Other):31628                              NA's   :8289   NA's   :8261  
##      Tenure        District    
##  Min.   : 0.16   1100: 3142  
##  1st Qu.: 3.75   1200:24342  
##  Median : 7.41   1300: 4150  
##  Mean   :10.16
##  3rd Qu.:14.75  
##  Max.   :41.16  
## 
```

Findings:

- 31,628 customers
- Profits ranging between -221.0 and 2071.0
- Median 9.0 and Mean 111.5 (right-skewed distribution)
- Data set contains only 12.18% online banking users
- Missing data (8,289 customers do not contain a factor for age and 8,261 customers do not contain a factor for Inc).

Analysis

Backward stepwise regression using Data

```r
library(MASS)

full.model <- lm(Profit ~ Online + Age + Inc + Tenure + District, data = Data, na.action = na.omit)
step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
summary(step.model)
```

```
## 
## Call:
## lm(formula = Profit ~ Online + Age + Inc + Tenure + District, 
##     data = Data, na.action = na.omit)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -522.74 -155.09  -70.72   66.44 1959.25 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -56.3975    13.1960  -4.274 1.93e-05 ***
## Online1       17.0251     5.4976   3.097  0.00196 ** 
## Age2          29.9474    11.9322   2.510  0.01209 *  
## Age3          69.8003    11.7093   5.961 2.54e-09 ***
## Age4          74.6121    11.7941   6.326 2.56e-10 ***
## Age5          79.4354    12.2504   6.484 9.10e-11 ***
## Age6         100.0856    12.7054   7.877 3.49e-15 ***
## Age7         135.7456    12.5281  10.835  < 2e-16 ***
## Inc2           0.9934    11.6513   0.085  0.93206    
## Inc3          10.9358     8.3948   1.303  0.19270    
## Inc4          10.8613     8.5525   1.270  0.20411    
## Inc5          15.9018     8.5367   1.863  0.06251 .  
## Inc6          39.6959     7.4708   5.313 1.09e-07 ***
## Inc7          60.7904     8.1594   7.450 9.64e-14 ***
## Inc8          78.5513     9.3164   8.432  < 2e-16 ***
## Inc9         146.8121     8.3667  17.547  < 2e-16 ***
## Tenure         4.0877     0.2354  17.363  < 2e-16 ***
## District1200  18.6401     6.3787   2.922  0.00348 ** 
## District1300   7.0957     7.7578   0.915  0.36038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 273.6 on 22793 degrees of freedom
##   (8822 observations deleted due to missingness)
## Multiple R-squared:  0.06488,    Adjusted R-squared:  0.06414
## F-statistic: 87.85 on 18 and 22793 DF,  p-value: < 2.2e-16
```

Data$Online1 is significant and increases profit by ~17 when all other independent variables do not change.

Data$Age is significant. We see that higher age has a positive effect on profitability.

Data$Inc is significant from Inc6 to Inc9 ($50,000 to $125,000 and more). We see that higher income has a positive effect on profitability.

Not best solution to delete 8822 observations.

Backward stepwise regression using imputed dataset (random forest imputation)

```r
library(missForest)

dummy <- as.data.frame(Data)

Data.imp <- missForest(dummy[,-1], verbose = TRUE)
```

```
##   missForest iteration 1 in progress...done!
##     estimated error(s): 0 0.3928233
##     difference(s): 0 0.1107116
##     time: 9.25 seconds
##
##   missForest iteration 2 in progress...done!
##     estimated error(s): 0 0.3911548
##     difference(s): 0 0.03150882
##     time: 9.17 seconds
##
##   missForest iteration 3 in progress...done!
##     estimated error(s): 0 0.3920534
##     difference(s): 0 0.0188721
##     time: 9 seconds
##
##   missForest iteration 4 in progress...done!
##     estimated error(s): 0 0.3899024
##     difference(s): 0 0.0155924
##     time: 8.89 seconds
##
##   missForest iteration 5 in progress...done!
##     estimated error(s): 0 0.3926736
##     difference(s): 0 0.0142173
##     time: 9.25 seconds
##
##   missForest iteration 6 in progress...done!
##     estimated error(s): 0 0.3911017
##     difference(s): 0 0.01278687
##     time: 9.11 seconds
##
```

```
##   missForest iteration 7 in progress...done!
##     estimated error(s): 0 0.3921182
##     difference(s): 0 0.01228109
##     time: 9.02 seconds
##
##   missForest iteration 8 in progress...done!
##     estimated error(s): 0 0.3917436
##     difference(s): 0 0.01208352
##     time: 8.98 seconds
##
##   missForest iteration 9 in progress...done!
##     estimated error(s): 0 0.3917431
##     difference(s): 0 0.0123206
##     time: 9.28 seconds
```

```r
Data.imp$OOBerror
```

```
##     NRMSE       PFC
## 0.0000000 0.3917436
```

PFC (proportion of falsely classified) is relatively high.

```r
full.model2 <- lm(Profit ~ Online + Age + Inc + Tenure + District, data = Data.imp$ximp)
step.model2 <- stepAIC(full.model2, direction = "both", trace = FALSE)
summary(step.model2)
```

```
##
## Call:
## lm(formula = Profit ~ Online + Age + Inc + Tenure + District,
##     data = Data.imp$ximp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -523.90 -141.56  -50.21   44.30 1947.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -36.291      6.162  -5.890 3.91e-09 ***
## Online1        14.749      4.536   3.251  0.00115 **
## Age2           27.990      6.556   4.269 1.96e-05 ***
## Age3           68.347      6.577  10.391  < 2e-16 ***
## Age4           76.367      6.795  11.238  < 2e-16 ***
## Age5           78.918      7.413  10.647  < 2e-16 ***
## Age6          111.823      7.892  14.170  < 2e-16 ***
## Age7          146.201      7.287  20.063  < 2e-16 ***
## Inc2          -18.071      8.158  -2.215  0.02675 *
## Inc3          -11.371      6.359  -1.788  0.07374 .
## Inc4           -6.396      6.884  -0.929  0.35288
## Inc5            8.647      7.084   1.221  0.22226
## Inc6           34.253      6.065   5.647 1.64e-08 ***
## Inc7           59.260      6.783   8.737  < 2e-16 ***
## Inc8           51.368      7.581   6.776 1.26e-11 ***
## Inc9          164.285      6.534  25.144  < 2e-16 ***
## Tenure          2.912      0.206  14.134  < 2e-16 ***
## District1200   11.154      5.148   2.167  0.03027 *
## District1300   10.990      6.166   1.782  0.07470 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.9 on 31615 degrees of freedom
## Multiple R-squared:  0.1003, Adjusted R-squared:  0.09974
## F-statistic: 195.7 on 18 and 31615 DF,  p-value: < 2.2e-16
```

Data$Online1 is significant and increases profit by ~14 when all other independent variables do not change.

Data$Age is significant. We see that higher age has a positive effect on profitability.

Data$Inc is significant from Inc6 to Inc9 ($50,000 to $125,000 and more). We see that higher income has a positive effect on profitability. Inc2 is now significant as well having a negative impact on profitability.

Regression (interaction effects) using imputed dataset (rendom forest imputation)

```
interaction.model <- lm(Profit ~ Online*Age + Inc + Tenure + District,data = Data.imp$ximp)
summary(interaction.model)
```

```
##
## Call:
## lm(formula = Profit ~ Online * Age + Inc + Tenure + District,
##     data = Data.imp$ximp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -528.15 -141.38  -50.00   44.93 1949.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -33.5514     6.2912  -5.333 9.72e-08 ***
## Online1      -13.8657    13.7061  -1.012 0.311713
## Age2          24.1733     7.1491   3.381 0.000722 ***
## Age3          63.5898     7.0315   9.044  < 2e-16 ***
## Age4          70.6649     7.2198   9.788  < 2e-16 ***
## Age5          71.2396     7.8167   9.114  < 2e-16 ***
## Age6         107.8733     8.2388  13.093  < 2e-16 ***
## Age7         143.0949     7.5814  18.875  < 2e-16 ***
## Inc2         -16.4661     8.1961  -2.009 0.044543 *
## Inc3          -8.7110     6.4781  -1.345 0.178736
## Inc4          -4.5393     6.9332  -0.655 0.512650
## Inc5          10.8380     7.1288   1.520 0.128441
## Inc6          36.2331     6.1153   5.925 3.16e-09 ***
## Inc7          61.1447     6.8257   8.958  < 2e-16 ***
## Inc8          53.2140     7.6186   6.985 2.91e-12 ***
## Inc9         166.4053     6.5777  25.298  < 2e-16 ***
## Tenure         2.9077     0.2062  14.099  < 2e-16 ***
## District1200  11.0299     5.1490   2.142 0.032191 *
## District1300  11.0022     6.1662   1.784 0.074389 .
## Online1:Age2  24.7718    16.2878   1.521 0.128300
## Online1:Age3  30.1154    16.4200   1.834 0.066653 .
## Online1:Age4  37.5888    16.7881   2.239 0.025162 *
## Online1:Age5  64.1462    20.3048   3.159 0.001584 **
## Online1:Age6  17.7212    27.8242   0.637 0.524196
## Online1:Age7  -6.5844    29.5149  -0.223 0.823471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.8 on 31609 degrees of freedom
## Multiple R-squared:  0.1006, Adjusted R-squared:  0.09994
## F-statistic: 147.4 on 24 and 31609 DF,  p-value: < 2.2e-16
```

Here we see that the interaction effect between Online1 and Age5 (middle-aged) is significant. Lower significance level are present for younger customers while there is no significance level for older customers.

Analysis using only top ten percent of most profitable customers

Reasoning for looking at top ten percent of most profitable customers

10% of the customers generated 70% of the profits. Therefore, these customers deserve special attention since a decision in the strategy has the highest impact on the overall profitability.

```
#Subset data into 10% most profitable and 90% least profitable customers

DataProfit <- Data.imp$ximp[order(Data.imp$ximp$Profit),]
0.1*31634
```

```
## [1] 3163.4
```

```
31634-3163
```

```
## [1] 28471
```

```
VectorProfit <- c(28471:31634)
DataProfit <- DataProfit[VectorProfit, ]
```

Exploring Data

```
str(DataProfit)
```

```
## 'data.frame':    3164 obs. of  6 variables:
##  $ Profit  : num  424 424 425 425 425 425 425 425 425 425 ...
##  $ Online  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : Factor w/ 7 levels "1","2","3","4",..: 4 2 5 3 6 7 7 2 3 5 ...
##  $ Inc     : Factor w/ 9 levels "1","2","3","4",..: 1 4 5 7 1 8 4 9 3 6 ...
##  $ Tenure  : num  2.08 11.41 30 9.5 28 ...
##  $ District: Factor w/ 3 levels "1100","1200",..: 1 3 2 2 2 2 2 2 2 2 ...
```

```
summary(DataProfit)
```

```
##      Profit       Online   Age          Inc           Tenure
##  Min.   : 424.0   0:2737   1:  4    9      :1012   Min.   : 0.16
##  1st Qu.: 519.0   1: 427   2:219    6      : 569   1st Qu.: 6.66
##  Median : 658.0            3:664    7      : 379   Median :12.66
##  Mean   : 771.2            4:807    8      : 250   Mean   :14.35
##  3rd Qu.: 920.0            5:500    4      : 212   3rd Qu.:21.16
##  Max.   :2071.0            6:386    5      : 209   Max.   :41.16
##                            7:584    (Other): 533
##  District
##  1100: 261
##  1200:2553
##  1300: 350
##
##
##
##
```

Findings:

- 3,164 customers
- Profits ranging from 424.0 to 2071.0
- Median 658.0 and Mean 771.2 (still right-skewed distribution, but not as severe as in Data)
- Share of online users is 13.5%

Regression model

```
full.model3 <- lm(Profit ~ Online + Age + Inc + Tenure + District, data = DataProfit)
summary(full.model3)
```

```
##
## Call:
## lm(formula = Profit ~ Online + Age + Inc + Tenure + District,
##     data = DataProfit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -445.8 -243.9 -108.5  145.1 1393.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 660.2850   171.9775   3.839 0.000126 ***
## Online1       4.8697    18.0538   0.270 0.787383
## Age2         -8.1483   170.6622  -0.048 0.961922
## Age3        110.6654   169.6731   0.652 0.514302
## Age4        127.5639   169.6881   0.752 0.452255
## Age5         92.6121   169.9494   0.545 0.585834
## Age6        156.8982   170.2836   0.921 0.356916
## Age7        109.2224   170.1760   0.642 0.521037
## Inc2         25.8347    39.4631   0.655 0.512737
## Inc3        -18.5512    34.1420  -0.543 0.586924
## Inc4        -26.1174    33.8708  -0.771 0.440712
## Inc5        -38.9558    34.4110  -1.132 0.257690
## Inc6          3.1531    29.6028   0.107 0.915183
## Inc7          9.4508    31.8982   0.296 0.767035
## Inc8         13.4034    34.2122   0.392 0.695254
## Inc9         64.0008    29.7368   2.152 0.031454 *
## Tenure        1.1302     0.7243   1.560 0.118746
## District1200 -40.4636    23.6454  -1.711 0.087130 .
## District1300  -9.9034    28.0245  -0.353 0.723824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 337.5 on 3145 degrees of freedom
## Multiple R-squared:  0.0231, Adjusted R-squared:  0.01751
## F-statistic: 4.131 on 18 and 3145 DF,  p-value: 1.066e-08
```

The dependent variable DataProfit$Online1 is not significant. Null-Hypothesis can not be rejected. Therefore, our analysis concerning the top 10% brings no further insights.

Overall conclusion

From our analyses we can derive that it is beneficial for Pilgrim Bank to promote online banking younger and most importantly to middle aged customers. The older aged customers may not buy online products with high margins.

Brief describtion of the shortcomings of our analyses

- Data is not actual (from end of 1999 - one year old). Dataset was constructed under customer self-selection since customers could decide on their own if they want to use online banking or not.
- Data size is small (31,634 out of 5,000,000 obervations).
- Data set contains only 12% online banking users
- Missing data (8,289 customers do not contain a factor for age and 8,261 customers do not contain a factor for Inc). Our missingForest model seems to have a relatively high PFC (proportion of falsely classified).
- Data consists of only a few independent variables. More variables could be of help.
- Complications because column "Online" does not describe how the new channel is actively used. Customers being registered as online banking users can still go most of the times to a branch instead of using the online service.