Final Report

Table of Contents

Non-Technical Summary

Our dataset consists of 1723 rows and 14 columns, the latest corresponding to month, credit amount, credit term (in months), age, sex, level of education, type of product, has children or not, region, income, family status, phone operator, client or not, bad client or not. When we analyzed our data we noticed that some of the variables we had were numeric, some corresponded to different categories and some others were binary (just two possible values). Having this in mind, we decided to take three different approaches to study our data which are lasso logistic regression, correspondence analysis and factor analysis.

Lasso logistic regression allows you to predict a binary variable or in this case, to understand if a person is a good or bad customer. In the first model, we could see that if a person buys a cell phone, they are more likely to be a bad customer. On the other hand, a person with a higher income is less likely to be a bad customer. In the second model, additionally to what was mentioned before, we could also notice that people with a higher level of education are less likely to be a bad customer. In summary, our model was able to predict in around 72% of the cases when a person was a bad customer and the variable that was more related to this outcome was whether a person is buying a cell phone or not.

Correspondence analysis is a method that lets you understand how categorical variables are related or connected to each other. We decided to analyze 4 groups of variables: product type vs. family status, sex vs. education, family status vs. education and education vs. product type. In the first place, we could see that married people tend to buy more music, medical or auto related products, unmarried people usually buy more products related to jewelry, tourism or beauty, and the rest of the people tend to buy more fishing or hunting supplies. Secondly, we noticed that women are more likely to have a secondary special education while men are more likely to have a higher education. This can be reflected in Figure 1, where a strong association between variables is represented by the blue color, a negative association by the color red and the absence of association by white. In the third group, we did not find any relationship between the family status and the level of education of a person (Figure 2). Lastly, we noticed that people that had a higher education level tend to buy more products related to health, beauty, and leisure but less products related to tech.

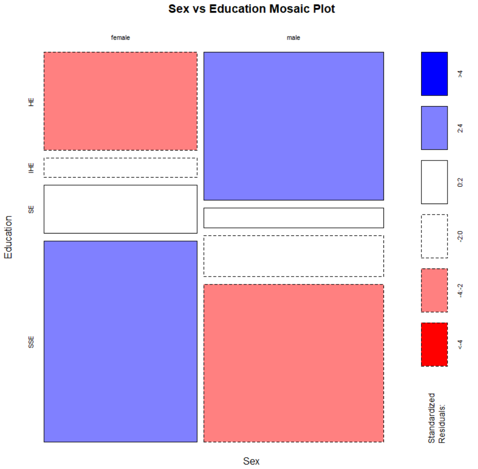


Figure 1. Mosaic Plot of Education vs. Sex

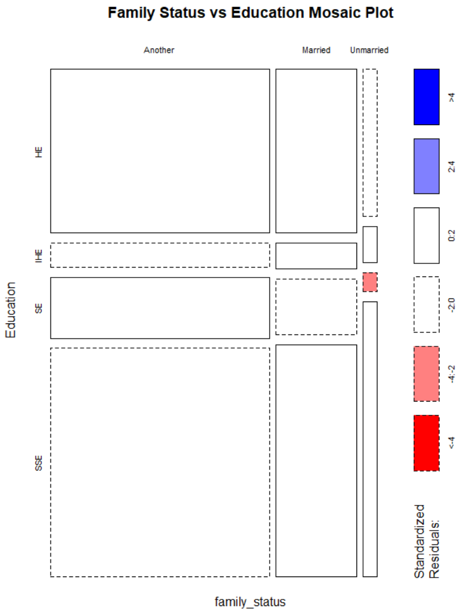


Figure 2. Mosaic Plot of Family Status vs. Education

Finally, factor analysis with logistic regression allows us to discover hidden relationships between numerical variables. Like Lasso regression, logistic regression is used to predict binary variables, which is also used to predict whether a person is a good or bad customer. However, the difference is that we need to factor analyze the continuous variables in the data to determine the principal factors, and then build a logistic model to determine whether the final prediction is accurate or not. Through the prediction, we can know that up to 88.7% of the accuracy can be predicted correctly. Another interesting finding is that age and income are positively correlated, and the larger the data for both, the more likely they are not bad customers.

Technical Summary

Exploratory Analysis

To get a good feel for our data, our initial exploratory analysis checked for missing values and n/a values in the dataset. None were found. For the numerical variables, we decided to visualize their distributions through histograms and found out that three out of four were right skewed. Therefore, we decided to perform logarithmic transformation and got a result an approximately normal distribution (Figure 3, 4).

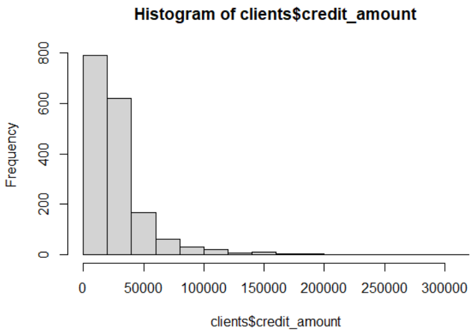


Figure 3. Histogram of Credit Amount

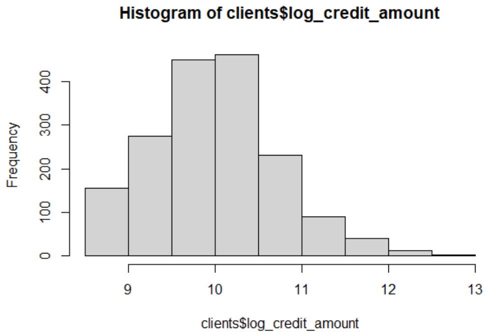


Figure 4. Histogram of Logarithmic Credit Amount

For the categorical variables, we decided to group in the 'product_type' variable to facilitate our analysis. We did this by either grouping the less common products into an 'Other' value or by grouping similar products into one value. Additionally, we used the function 'dummy_cols' included in the 'fastDummies' library to create dummy variables for the categorical data that we had.

Lastly, we just decided to get rid of unnecessary remaining variables and ended up with a total of 31 variables to work with.

Approach 1 Execution - Lasso Logistic Model

In the first place, we split our dataset into train and test sets with a partition of 80% for training and 20% for test. To solve the imbalance of our target variable ('bad_client_target'), we used the upSample function to increase the number of values for '1' and ended up with a total of 1228 observations for both '0's and '1's. Additionally, we separated the x's and y's of both sets and transformed them into matrices before building our model.

To build our model, we first ran the 'cv.glmnet' function to find the ideal value of lambda and changed the family parameter into 'binomial' since our target variable is binary. As we can see in the graph (Figure 5), there is not a significant dip between the two lines, so we can assume that there is not a high regularization in the model. This could be confirmed when we used

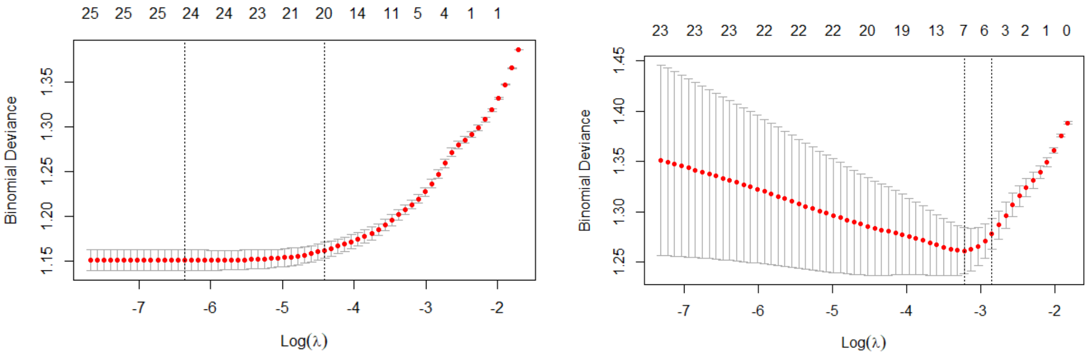
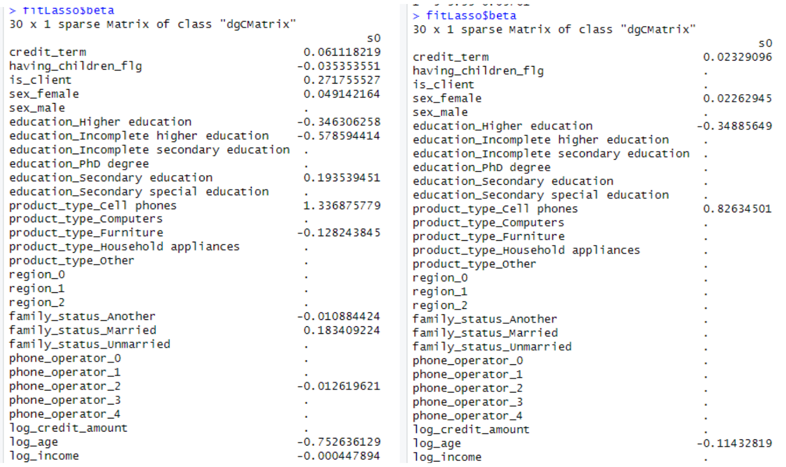


Figure 5. Plot of cv.glmnet for upSample (left) and downSample (right)

‘lambda.1se’, which allows more sparse results than ‘lambda.min’, and we got a model with a total of 14 variables (Figure 6).



```
> fitLasso$beta
30 x 1 sparse Matrix of class "dgCMatrix"
                                                  s0
credit_term                              0.061118219
having_children_flg                     -0.035353551
is_client                                0.271755527
sex_female                               0.049142164
sex_male                                 .
education_Higher education              -0.346306258
education_Incomplete higher education   -0.578594414
education_Incomplete secondary education .
education_PhD degree                     .
education_Secondary education            0.193539451
education_Secondary special education    .
product_type_Cell phones                 1.336875779
product_type_Computers                   .
product_type_Furniture                  -0.128243845
product_type_Household appliances        .
product_type_Other                       .
region_0                                 .
region_1                                 .
region_2                                 .
family_status_Another                   -0.010884424
family_status_Married                    0.183409224
family_status_Unmarried                  .
phone_operator_0                         .
phone_operator_1                         .
phone_operator_2                        -0.012619621
phone_operator_3                         .
phone_operator_4                         .
log_credit_amount                        .
log_age                                 -0.752636129
log_income                              -0.000447894
```

```
> fitLasso$beta
30 x 1 sparse Matrix of class "dgCMatrix"
                                                 s0
credit_term                              0.02329096
having_children_flg                      .
is_client                                .
sex_female                               0.02262945
sex_male                                 .
education_Higher education              -0.34885649
education_Incomplete higher education    .
education_Incomplete secondary education .
education_PhD degree                     .
education_Secondary education            .
education_Secondary special education    .
product_type_Cell phones                 0.82634501
product_type_Computers                   .
product_type_Furniture                   .
product_type_Household appliances        .
product_type_Other                       .
region_0                                 .
region_1                                 .
region_2                                 .
family_status_Another                    .
family_status_Married                    .
family_status_Unmarried                  .
phone_operator_0                         .
phone_operator_1                         .
phone_operator_2                         .
phone_operator_3                         .
phone_operator_4                         .
log_credit_amount                        .
log_age                                 -0.11432819
log_income                               .
```

Figure 6. Output of Lasso-Logistic model for upSample (left) and downSample (right)

Since our data has a lot of observations, we also considered performing a downsample for the ‘0’s of our target variable as this method is proven to perform better in larger datasets. We ended up with a total of 159 observations for both ‘0’s and ‘1’s and then we proceeded with the same process we applied for the upsampled data to build the model. In this case, we got a lambda plot (Figure 5) that showed a deeper dip so it performed a better regularization that was reflected in its output where we had a total of 5 significant variables (Figure 6). Lastly, we decided to run this model 5 more times to analyze its stability and we proved that, although the results were not exactly the same, there was not a lot of variance or difference between the values we got and that the performance of the model was not really dependent on the sample used.

Finally, we performed a relaxed lasso to verify how much regularization was actually needed in this dataset. We got as a result a model with 14 significant variables using a lambda of 0.031 for a gamma of 0.25.

Approach 1 - Results and Analysis

For the upsampled model, we got an RMSE of 0.43 for the training set and of 0.46 for the test set. Therefore, we can say that this model had a good performance. Additionally, as seen in the results (Figure 7), we had an accuracy of 66%, a sensitivity of 68% and a specificity of 52%. On the other side, the downsampled data had an RMSE of 0.46 for both the training and test sets, an accuracy of 72%, a sensitivity of 73% and a specificity of 68%. Lastly, for the relaxed lasso we had an RMSE of 0.44 for the training set and 0.46 for the test set, and accuracy of 71%, a sensitivity of 72% and a specificity of 59%. Also demonstrating a good performance.

When we analyze the results, we can conclude that the upsampled model did not have a performance as high as the other two models. Additionally, even though the relaxed lasso model showed that there was no need to reduce as many variables to 0, its performance still did not surpass the performance of the downsampled model. We can conclude that by using the downSample function, we got better results when performing regression in our dataset and our final model **(model = 0.02*credit_term + 0.02*sex_female - 0.35*higher_education + 0.82*cell_phones - 0.11*log_age)** is able to predict in a high percentage of the cases when a person is a bad customer. This is based mostly on a high positive correlation with the variable 'product_type_cell phones' and a negative correlation with the variable 'education_Higher Education'.

Approach 2 Execution- Correspondence Analysis

For this part of the project, we used the chi-squared test, mosaic plots and correspondence analysis using the CA function in R to map the associations between the categorical variables product_type, family_status, sex, and education.

Because of the large number of levels of the product_type factor, we grouped the products into 5 general categories, recoding all of the products as members of these categories: Leisure, Household Items, Tech, Health and Beauty, and Employment. In addition, the histogram data showed a large bias for the education variable, with 585 values for "Higher Education" and only 3 values for "PhD Degree." Thus, we merged "PhD Degree" into "Higher Education" for better quality analysis.

We performed a mosaic plot analysis to see whether there was a statistical association between the four categorical variables in the dataset mentioned above.

Finally, we used Microsoft PowerPoint line functionality to make factor maps to show associations between each level in the education factor and the levels of the product_type factor.

Approach 2 Execution- Results and Analysis

We started with sex vs education. For these two variables, males were more likely to have a higher education while women were more likely to have a secondary special education as shown in Figure 1. The p-value for this association, given by the chi-squared test, came out to 4.50846e-08, indicating that we can reject the null hypothesis. This supports our hypothesis that there is an association between the two variables.

For product_type vs education, the mosaic plot (Figure 8) demonstrated a negative association between higher education and tech products (computers and cell phones). Health and Beauty and Leisure products were positively associated with higher education. The p-value for this association, given by the chi-squared test, came out to 5.834602e-14, so we can reject the null hypothesis that there is no association between product and education.

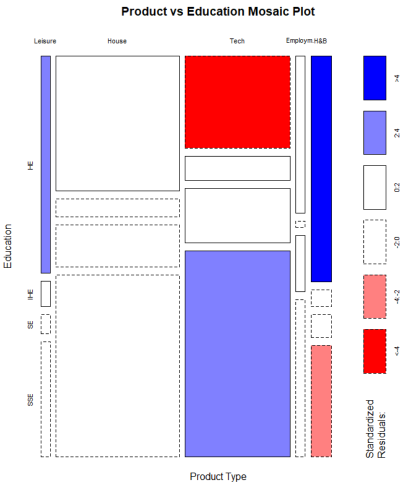


Figure 8. Mosaic plot of product_type vs education

Taking a deeper look at the correspondence between product and education, we first looked at the association between higher education and the 5 product types (Figure 9). Higher education was most positively associated with Health & Beauty, followed closely by Leisure. It had a smaller positive association with Employment, a near zero association with Household Items, and a negative association with Tech.

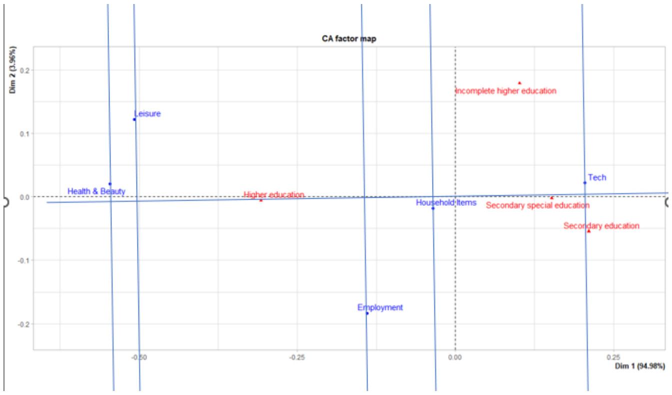


Figure 9. Factor Map of Higher Education vs Product Type

Secondary special education and Secondary education have the opposite associations with the product variables. They are most positively associated with Tech, followed by Household Items, Employment, with Leisure and Health & Beauty coming in last place. An incomplete higher education, shown below (Figure 10), is most positively associated with Tech, followed by Household Items, Leisure, Employment, and Health & Beauty.

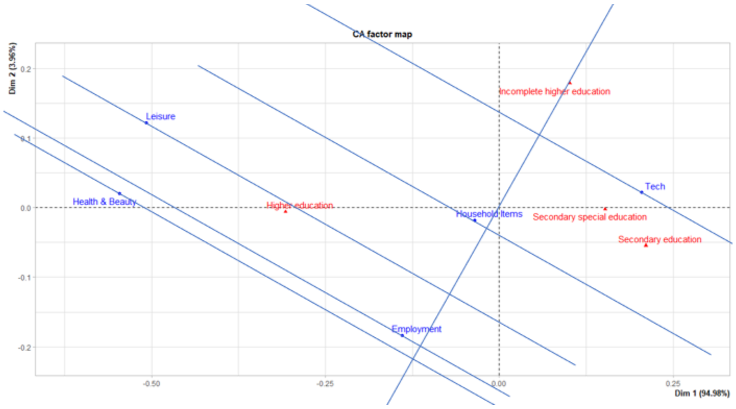


Figure 10. Factor Map of Incomplete Higher Education vs Product Type

For product_type vs family_status, the mosaic plot (Figure 11) shows all rectangles being white which indicates no significant deviations from expected frequencies. This means it shows independence that suggests no strong evidence of an association between product types and family status categories in the dataset.

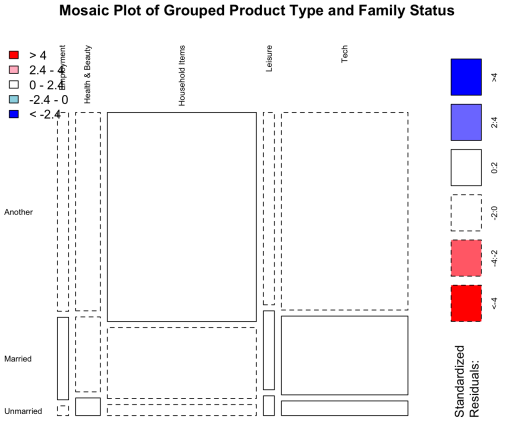


Figure 11. Mosaic Plot Product Type vs Family Status

The correspondence plot below (Figure 12) effectively highlights the relationships between grouped product types and family status categories. The primary dimension (Dimension 1) captures most of the variance, distinguishing major differences between the categories, while the secondary dimension (Dimension 2) adds further granularity to the associations. Looking at Unmarried family status, Leisure is positively associated with it but Household and Employment are negatively associated. Similarly, if we look at Married family status then Tech is most positively associated whereas for Another Family status Household is positively associated. The p-value for these two categories is 0.07577, greater than 0.05, indicating marginal significance.

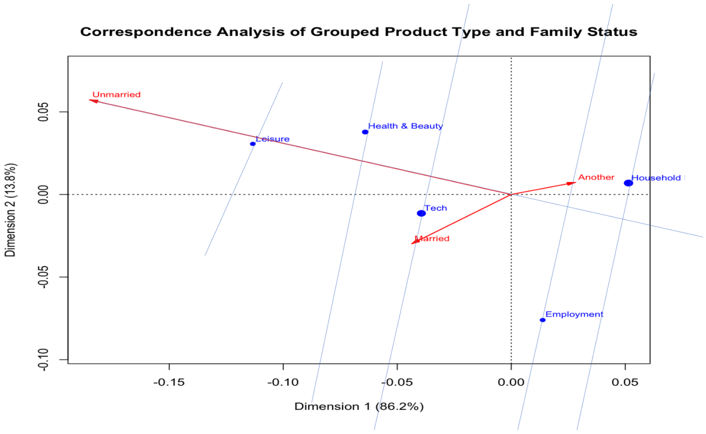


Figure 12. Correspondence plot for Product Type vs. Family Status

Approach 3 Execution - Factor Analysis

For this part of project analysis, we focused on Principal Component Analysis (PCA) and Common Factor Analysis (CFA). We eliminated most categorical variables except for “sex” (“male” and “female”) and made transformations: the variable “credit_amount” became “credit_amount_t” and variable “income” became “income_t”. As a result, we got the independent variables “credit_term age”, “having_children_flg”, “region phone_operator”, “ is_client”, “credit_amount_t”, “income_t”, “sex_numeric” and the target variable “bad_client_target”.

Then we tried PCA using the “prcomp” (with scaling) function in R. By using scree plot, we decided how many factors to analyze, compared them, selected a number of factors, and named all the factors.

Finally, we performed logistic regression to combine the selected factors with the target variables to make predictions and see how well the model fit the dataset.

Approach 3 - Results and Analysis

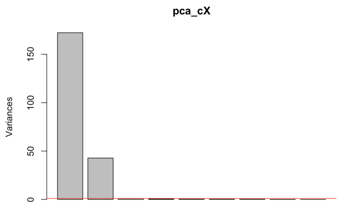


Figure 13. PCA Plot

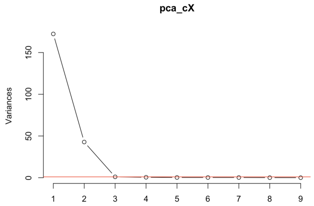
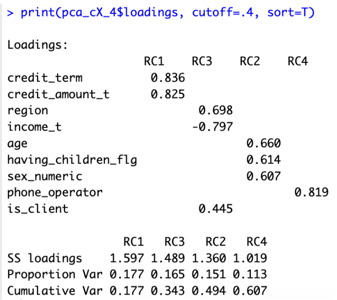


Figure 14. PCA Scree Plot

We noticed, moving left to right, that by the second component we already have a cumulative proportion of 98.92% of the variability, and 99.39% by the third component. From the scree plot, we saw that there may be some disagreement on the number of components to include. As the leveling off doesn’t seem to occur until after 3 factors, we looked at the loadings and compared with 2, 3 and 4 factors.



```
> print(pca_cX_4$loadings, cutoff=.4, sort=T)

Loadings:
                   RC1    RC3    RC2    RC4
credit_term         0.836
credit_amount_t     0.825
region                     0.698
income_t                  -0.797
age                               0.660
having_children_flg               0.614
sex_numeric                       0.607
phone_operator                           0.819
is_client                  0.445

                 RC1   RC3   RC2   RC4
SS loadings    1.597 1.489 1.360 1.019
Proportion Var 0.177 0.165 0.151 0.113
Cumulative Var 0.177 0.343 0.494 0.607
```

Figure 15. CFA Loadings

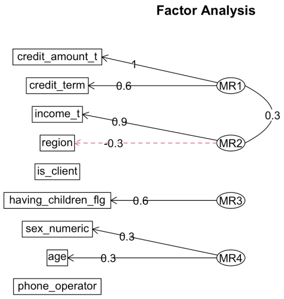
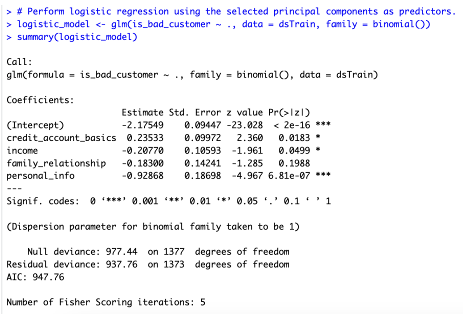


Figure 16. FA Result with 4 Factors

We could see cumulative variance was only 49.4% when it was three factors, which was too low, and the cumulative variance was 60.7% when the number of factors was four, which seemed fair enough. Therefore, four might be the best number of components for the dataset. We named the four factors and target variables after "credit_account_basics", "income", "family_relationship", "personal_info" and "is_bad_customer". Then, we split our dataset into train and test sets with a partition of 80% for training and 20% for test to make a logistic regression model.



```
> # Perform logistic regression using the selected principal components as predictors.
> logistic_model <- glm(is_bad_customer ~ ., data = dsTrain, family = binomial())
> summary(logistic_model)

Call:
glm(formula = is_bad_customer ~ ., family = binomial(), data = dsTrain)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.17549    0.09447 -23.028  < 2e-16 ***
credit_account_basics  0.23533    0.09972   2.360   0.0183 *
income                -0.20770    0.10593  -1.961   0.0499 *
family_relationship   -0.18300    0.14241  -1.285   0.1988
personal_info         -0.92868    0.18698  -4.967 6.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 977.44  on 1377  degrees of freedom
Residual deviance: 937.76  on 1373  degrees of freedom
AIC: 947.76

Number of Fisher Scoring iterations: 5
```

Figure 17. Logistic Regression Result

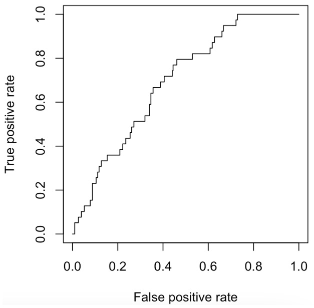
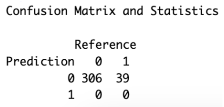


Figure 18. ROC Plot



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 306  39
         1   0   0
```

Figure 19. Confusion Matrix

According to the logistic regression (Figure 17), we could see "credit_account_basics", "income" and "personal_info" were statistically significant, but the p-value of the factor "family_relationship" seemed too large. Based on the confusion matrix, we got 306 true negatives as TN, 345 true positives as TP, 0 false negatives as FN, and 0 false positives as FP. Next, we got the results 0.1130 for RMSE, 0.8870 for accuracy, 0.6949 for AUC, 1.0000 for sensitivity, and 0.0000 for specificity. The reason why the last two results were too extreme is that we didn't have "bad customers" in the prediction part. Except for these two, the measures indicated a robust model. The results were not bad.

Conclusions

We could conclude that when doing lasso-logistic regression in our dataset, performing downsampling gives us the most parsimonious model preserving a low RMSE and a high accuracy. The model is able to predict and differentiate a good customer from a bad one. Additionally, we could notice that a bad customer is more related to buying cell phones and not having a higher education level.

In a similar vein, the correspondence analysis suggested a negative association between tech products and higher education. This surprised us because it seems counter-intuitive. This could indicate that computers and cell phones are owned by people with less education just as much, if not more, than people with higher education. What a huge technological miracle this is! Just twenty years ago, computers and cell phones were the playthings of the ultra rich.

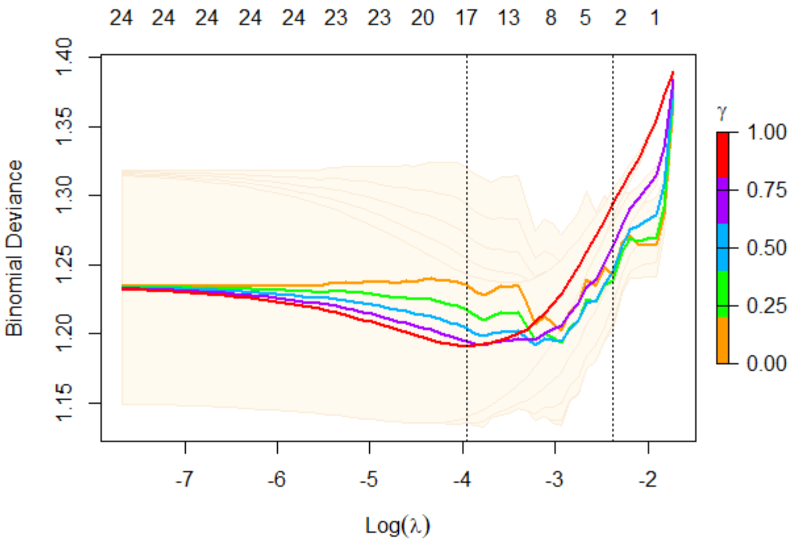
The Employment product category (construction materials, repair services, medical services) was associated most negatively with an Incomplete Higher Education. The reason someone who has dropped out of school early may not be buying employment related materials may be related to the fact that it is harder for that person to find and keep a job in the first place. In the future, it would be interesting to test such a hypothesis.

A person's family status was shown to be likely unrelated to the type of products people buy. The correspondence analysis showed a slight preference of unmarried people for leisure activity products, though not strong enough to reject the null hypothesis. We cannot be sure that these two categories are related, but if they are, it makes sense that unmarried people have more time to relax.

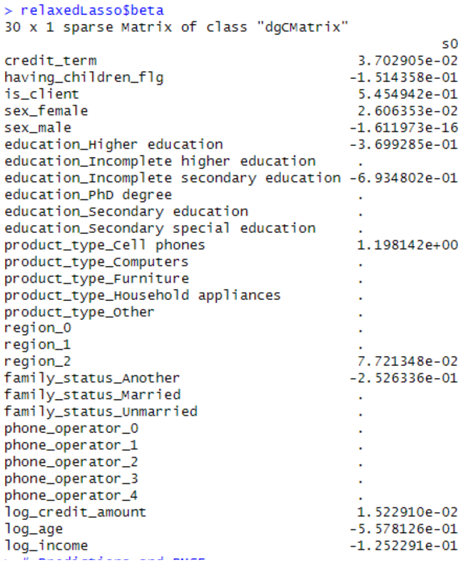
When modeling logistic regression, we were also glad to see the results of high accuracy, high AUC and fair RMSE, which told us that the model fit well. Although we selected 4 factors based on PCA and CFA, there was little relation between "bad customers or not" and family status, which referred to "have children or not". On the other hand, credit account information, income and personal information like age were highly related to our target variable (the goodness of a customer). These factors coincided with our normal perceptions.

In this project, lasso regression and logistic regression were not opposites, but parallel. The lasso regression was used for mainly categorical variables, while the logistic regression used mainly continuous, numerical variables of the original data, which were later analyzed by PCA. In other words, the variables in their respective regions were basically different. But happily, we achieved good results in both.

Appendix



Plot of cv.glmnet for Relaxed-Lasso

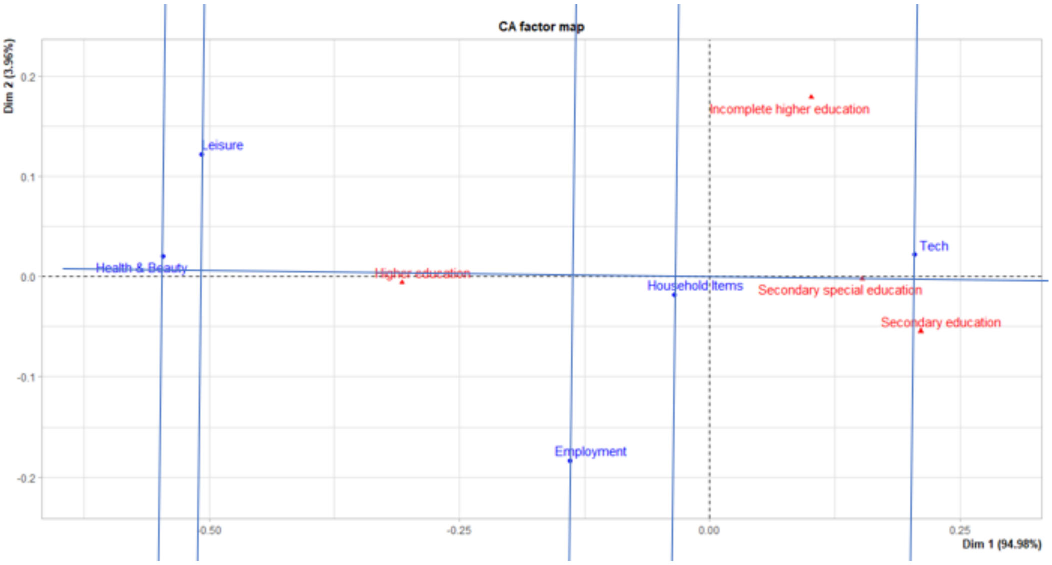


```
> relaxedLasso$beta
30 x 1 sparse Matrix of class "dgCMatrix"
                                                    s0
credit_term                               3.702905e-02
having_children_flg                      -1.514358e-01
is_client                                 5.454942e-01
sex_female                                2.606353e-02
sex_male                                 -1.611973e-16
education_Higher education               -3.699285e-01
education_Incomplete higher education     .           
education_Incomplete secondary education -6.934802e-01
education_PhD degree                      .           
education_Secondary education             .           
education_Secondary special education     .           
product_type_Cell phones                  1.198142e+00
product_type_Computers                    .           
product_type_Furniture                    .           
product_type_Household appliances         .           
product_type_Other                        .           
region_0                                  .           
region_1                                  .           
region_2                                  7.721348e-02
family_status_Another                    -2.526336e-01
family_status_Married                     .           
family_status_Unmarried                   .           
phone_operator_0                          .           
phone_operator_1                          .           
phone_operator_2                          .           
phone_operator_3                          .           
phone_operator_4                          .           
log_credit_amount                         1.522910e-02
log_age                                  -5.578126e-01
log_income                               -1.252291e-01
```

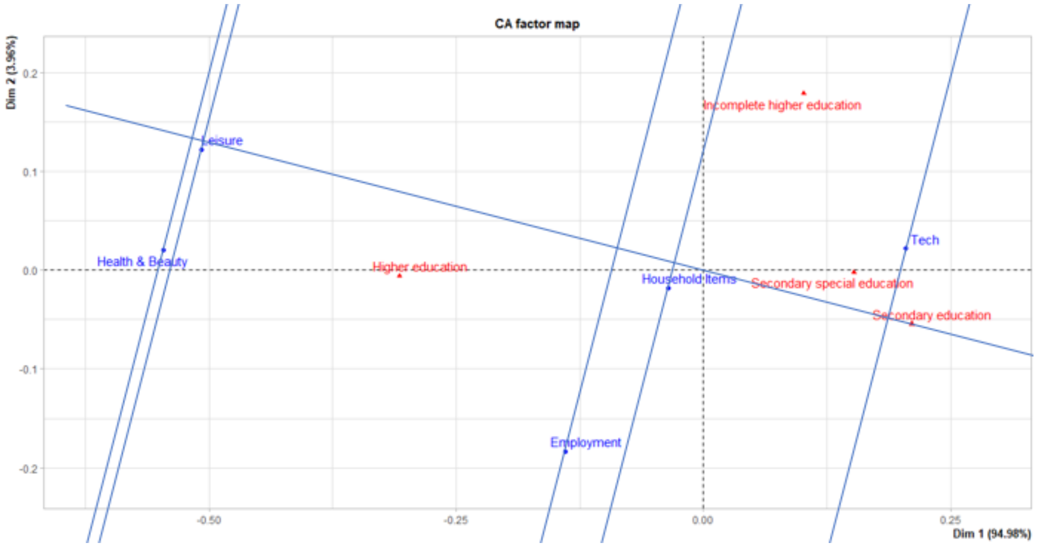
Output of Relaxed-Lasso Model

Model	RMSE Train	RMSE Test	Accuracy (%)	Sensitivity (%)	Specificity (%)
Upsample	0.43	0.46	66	68	52
Downsample	0.46	0.46	72	73	68
Relaxed Lasso	0.44	0.46	71	72	59

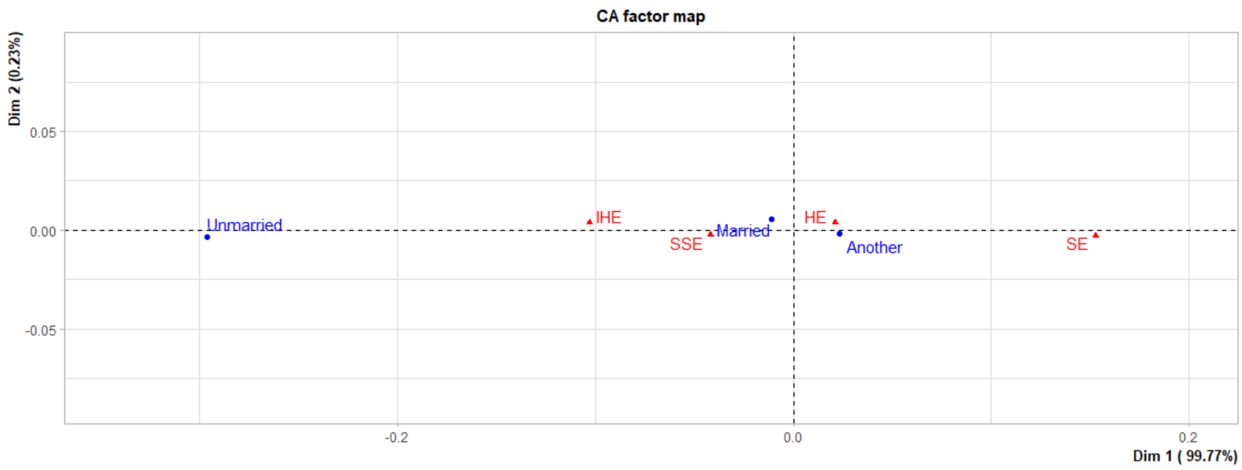
Matric results for upsampled, downsampled and relaxed-lasso



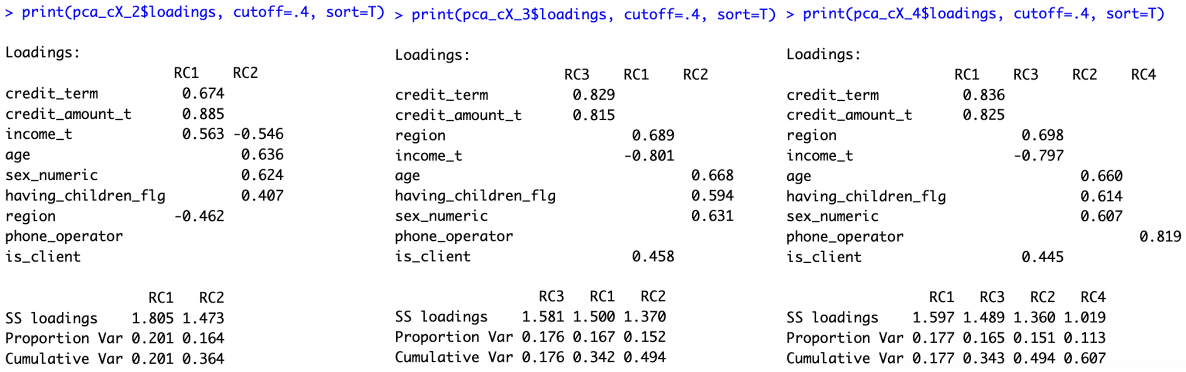
Factor Map of Secondary special education vs Product type



Factor Map of Secondary education vs Product type



Factor map of Family Status vs Education

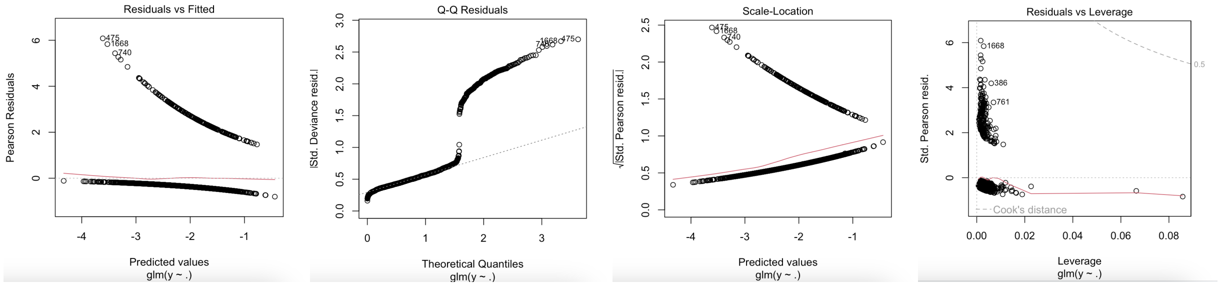


```
> print(pca_cX_2$loadings, cutoff=.4, sort=T)

Loadings:
                    RC1    RC2
credit_term          0.674
credit_amount_t      0.885
income_t             0.563 -0.546
age                         0.636
sex_numeric                 0.624
having_children_flg         0.407
region              -0.462
phone_operator
is_client

                 RC1   RC2
SS loadings    1.805 1.473
Proportion Var 0.201 0.164
Cumulative Var 0.201 0.364
```

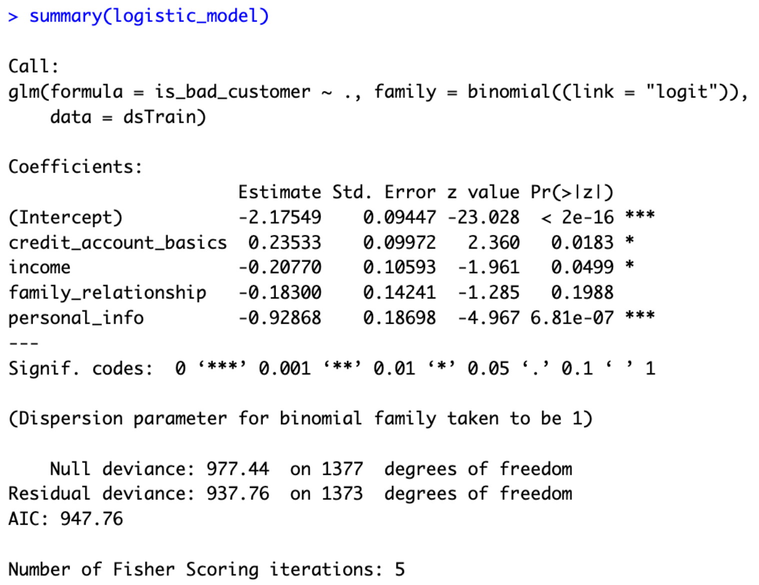
```
> print(pca_cX_3$loadings, cutoff=.4, sort=T)

Loadings:
                    RC3    RC1    RC2
credit_term          0.829
credit_amount_t      0.815
region                      0.689
income_t                   -0.801
age                                0.668
having_children_flg                0.594
sex_numeric                        0.631
phone_operator
is_client                   0.458

                 RC3   RC1   RC2
SS loadings    1.581 1.500 1.370
Proportion Var 0.176 0.167 0.152
Cumulative Var 0.176 0.342 0.494
```

```
> print(pca_cX_4$loadings, cutoff=.4, sort=T)

Loadings:
                    RC1    RC3    RC2    RC4
credit_term          0.836
credit_amount_t      0.825
region                      0.698
income_t                   -0.797
age                                0.660
having_children_flg                0.614
sex_numeric                        0.607
phone_operator                            0.819
is_client                   0.445

                 RC1   RC3   RC2   RC4
SS loadings    1.597 1.489 1.360 1.019
Proportion Var 0.177 0.165 0.151 0.113
Cumulative Var 0.177 0.343 0.494 0.607
```

Factor Analysis with 2, 3 and 4 factors

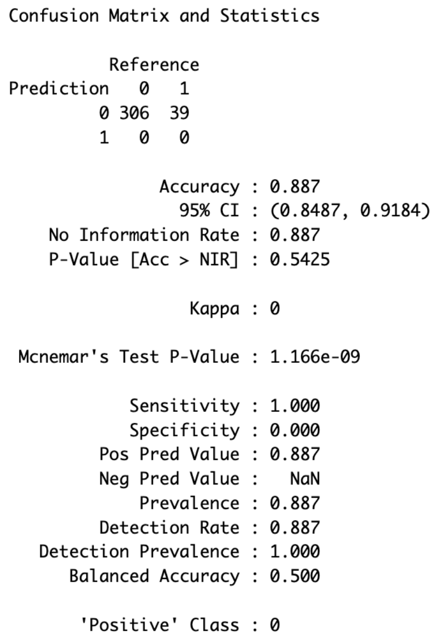


Plot of Logistic Regression of 4 Factors



```
> summary(logistic_model)

Call:
glm(formula = is_bad_customer ~ ., family = binomial((link = "logit")),
    data = dsTrain)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.17549    0.09447 -23.028  < 2e-16 ***
credit_account_basics  0.23533    0.09972   2.360   0.0183 *  
income                -0.20770    0.10593  -1.961   0.0499 *  
family_relationship   -0.18300    0.14241  -1.285   0.1988    
personal_info         -0.92868    0.18698  -4.967 6.81e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 977.44  on 1377  degrees of freedom
Residual deviance: 937.76  on 1373  degrees of freedom
AIC: 947.76

Number of Fisher Scoring iterations: 5
```

Summary of Logistic Model of 4 factors



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 306  39
         1   0   0
                                          
               Accuracy : 0.887           
                 95% CI : (0.8487, 0.9184)
    No Information Rate : 0.887           
    P-Value [Acc > NIR] : 0.5425          
                                          
                  Kappa : 0               
                                          
 Mcnemar's Test P-Value : 1.166e-09       
                                          
            Sensitivity : 1.000           
            Specificity : 0.000           
         Pos Pred Value : 0.887           
         Neg Pred Value :   NaN           
             Prevalence : 0.887           
         Detection Rate : 0.887           
   Detection Prevalence : 1.000           
      Balanced Accuracy : 0.500           
                                          
       'Positive' Class : 0               
```

Confusion Matrix of Logistic Regression