Lenus Data Scientist Case Brief

Customer Segmentation Challenge

Introduction

Description

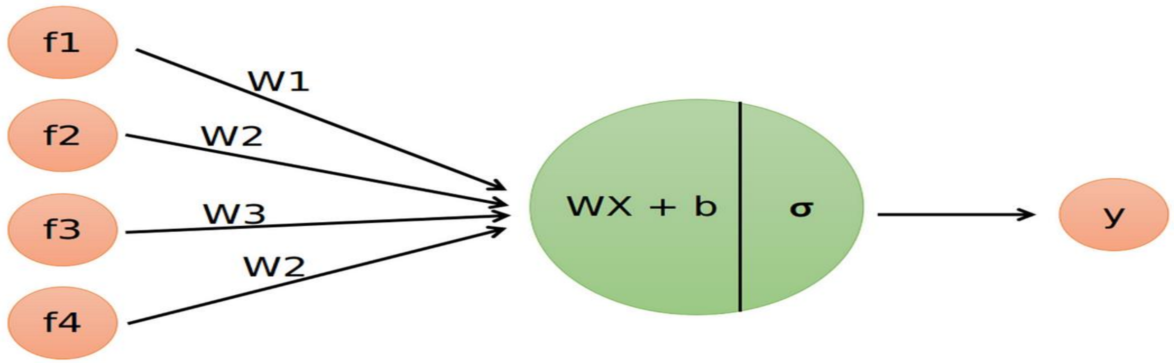
With the data provided in the file customer_data_sample.csv using any method you deem fit, answer the question:

"What are the most important factors for predicting whether a customer has converted or not?"

Converted customer is represented in the data in the field "converted", and the nature of what this conversion means is (intentionally) unknown in the context of the challenge.

Fields

field	explanation
customer_id	Numeric id for a customer
converted	Whether a customer converted to the product (1) or not (0)
customer_segment	Numeric id of a customer segment the customer belongs to
gender	Customer gender
age	Customer age
related_customers	Numeric - number of people who are related to the customer
family_size	Numeric - size of family members
initial_fee_level	Initial services fee level the customer is enrolled to
credit_account_id	Identifier (hash) for the customer credit account. If customer has none, they are shown as "9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f6c9bc9d493a23be9de0"
branch	Which branch the customer mainly is associated with



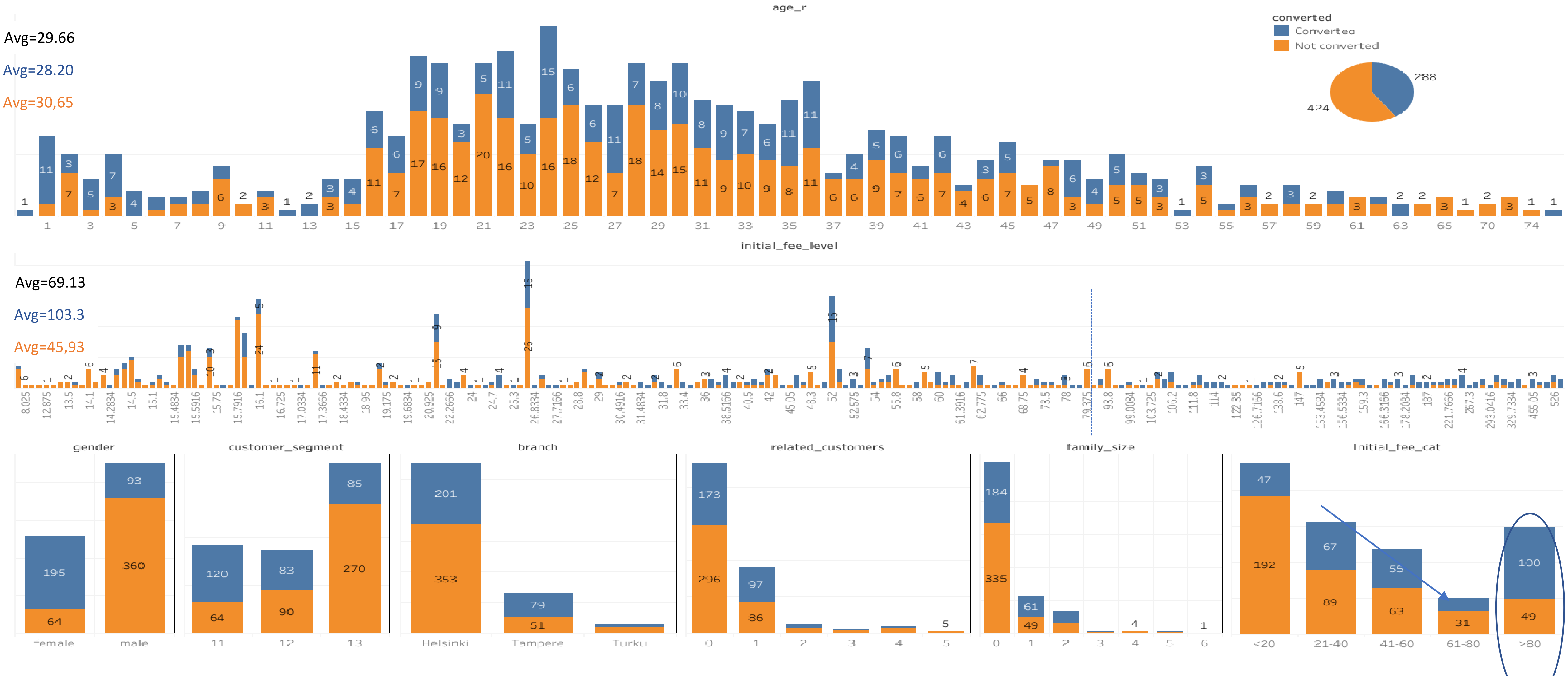
Agenda

- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

Agenda

- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

Descriptive statistics*



*https://public.tableau.com/app/profile/nicola.menale8721/viz/Lenus_Descriptive/Descriptivestatistics

Descriptive statistics – initial observations

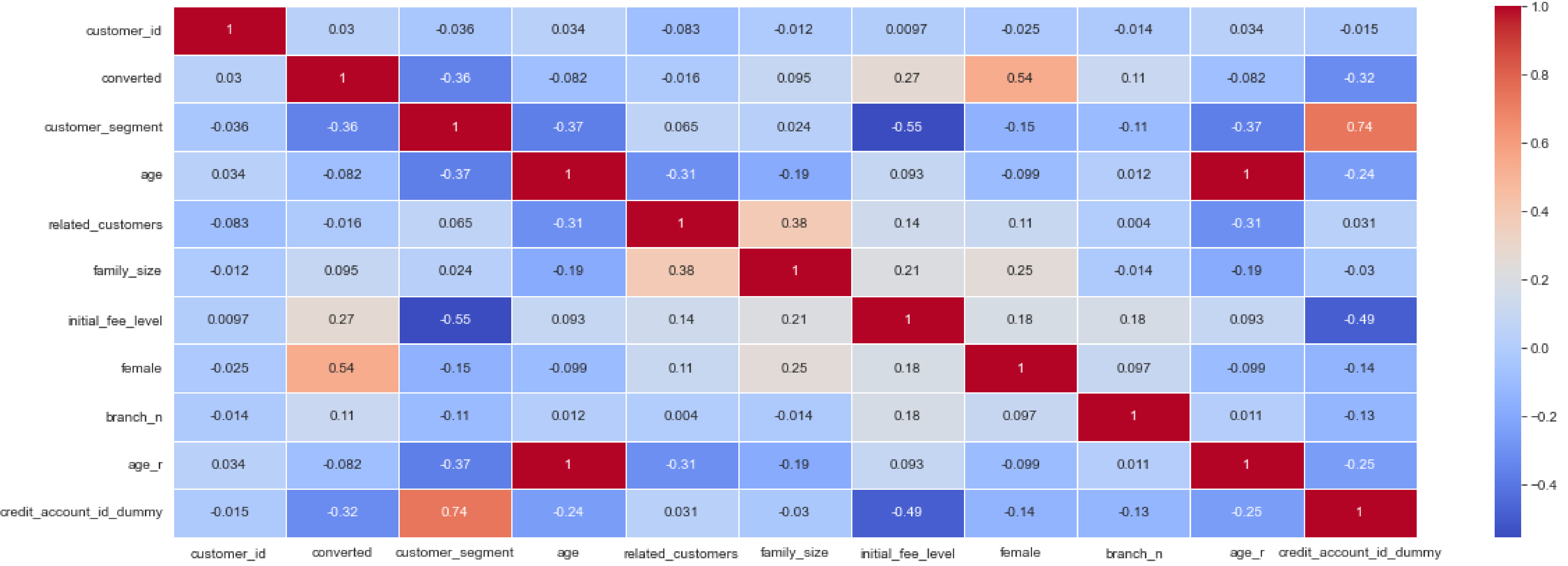
- There are some nulls values in the data that have been excluded from the analysis (mostly in the Age variable)
- In Age there a re 7 customers that have less than one year. (can be typos or customer registered by their parents and have few month only). All those are actually converted*. Normally I would double check these information with the data owner. In this “experiment” I will simply point this out and proceed with the analysis.
- “Converted” seems to be younger and have higher initial fee level compared to “not converted”. There seems to be a tendency for female to be more prone to “convert” despite most customers are male.
- Segment 11 seems to have more “converted” in proportion compared to the other segments (but most customer are in seg 13). Same for Tampere Branch (also here most of customer are in Helsinki branch).
- The number of converted seems to increase when the number of people related to the customer is one and the family size is between 1 and 2 (in their respective groups). In both cases, the majority of customers are in the category 0

* According to danish law, "The Danish government recently passed legislation that prohibits the collection and processing of personal data belonging to children below the age of 13" (https://www.jurist.org/news/2023/06/denmark-to-raise-age-at-which-tech-companies-can-collect-minors-data/).

Agenda

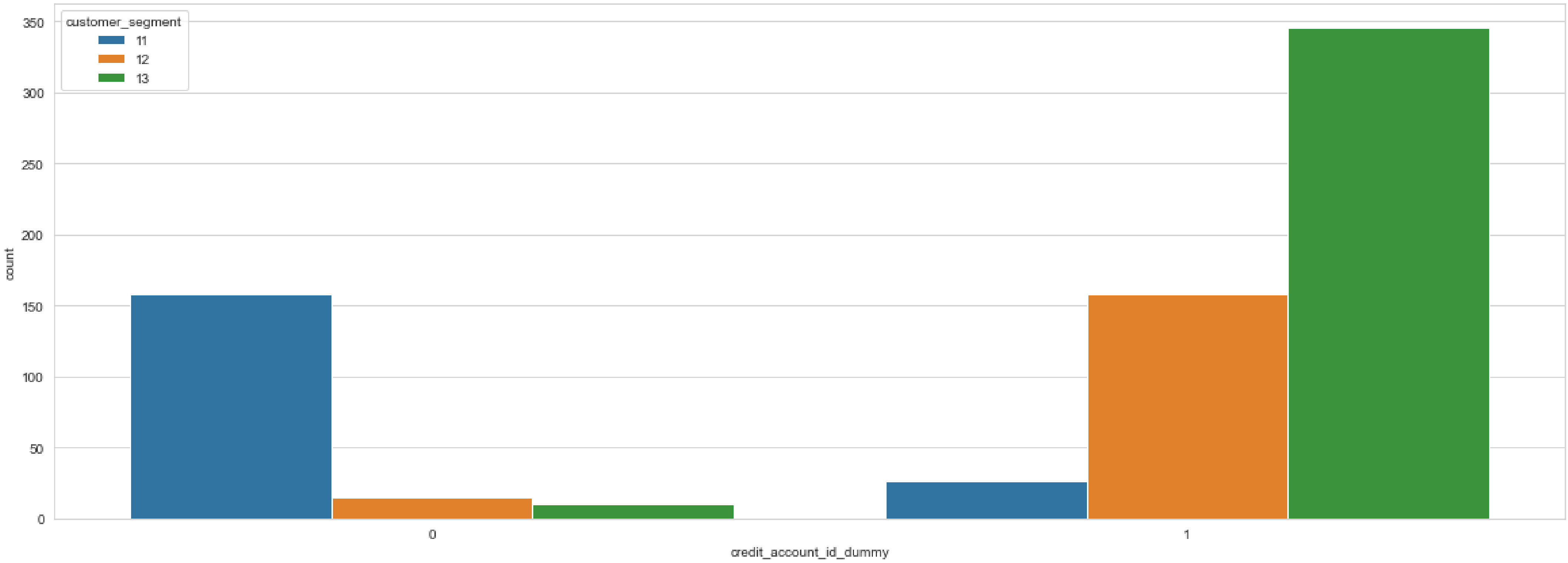
- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

Correlation Analysis



Correlation Analysis

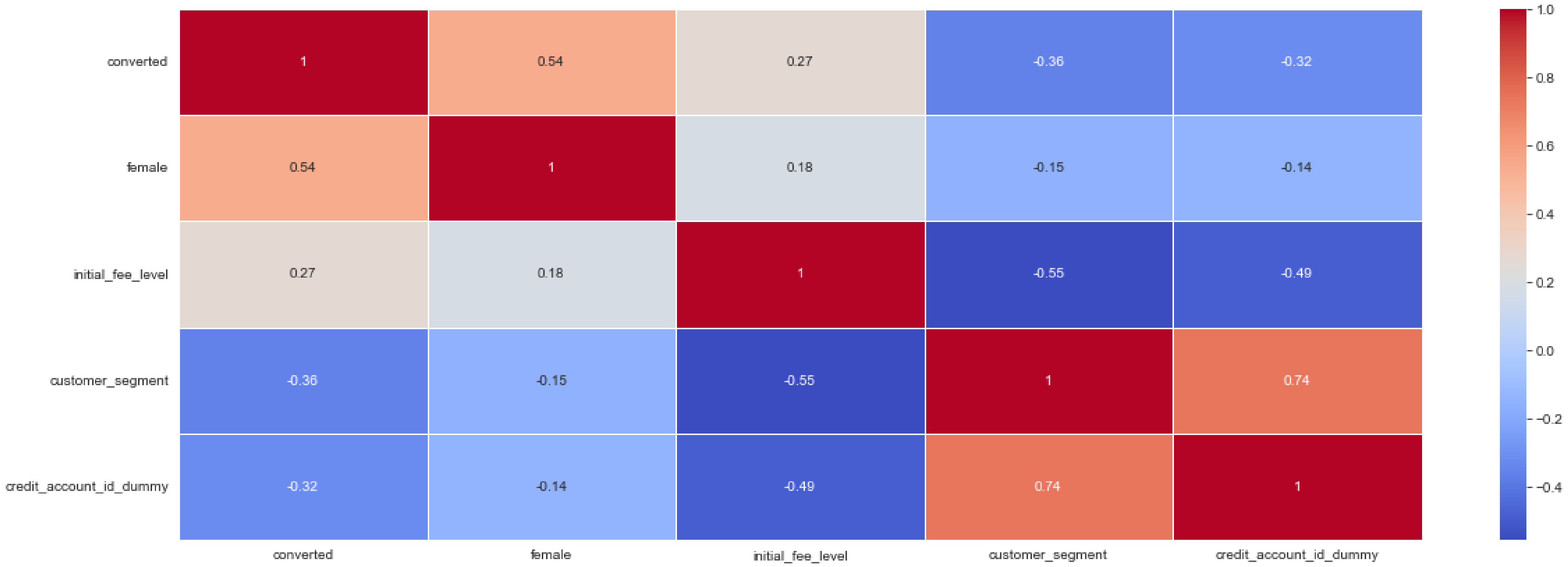
I created a dummy variable for the credit account id. It seems to be a correlation with the segments.
Most of the customers in seg 12 and 13 does not have a customer credit account.



LENUS

Correlation Analysis

Gender, initial fee level and customer segment seems to be relevantly correlated with the conversion rate of the customers.

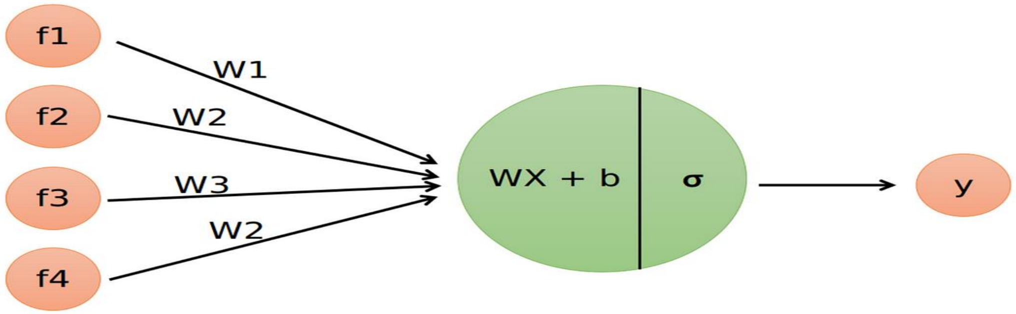


Agenda

- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

What are the factors for predicting conversion?

- In order to find the relevant factors I need to look at the nature of the problem and it´s component: we are trying to understand how much the probability of an event to happen (converted or not) given the behavior of the other variables (independent variables).
- The nature of the dependent variable is categorical (dicotomical) and the nature of the dependent variables is different for each variable (categorical and/or interval (continuous), since we are dealing with different variables.)
- literature* suggests to use a **multiple logistic regression** to identify the variable that best explain the probability to convert.



*https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/; https://stats.oarc.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#logistic

What are the factors for predicting conversion?

- After creating dummy variables for all the explanatory variable we try our first regression inserting all the data that are already dicotomical or continuous: “ 'female', 'age_r’, 'initial_fee_level','credit_account_id_dummy’ ”

```
Optimization terminated successfully.
         Current function value: 0.475354
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      708
Method:                           MLE   Df Model:                            3
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                  0.2956
Time:                        12:00:28   Log-Likelihood:                -338.45
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 2.886e-61
===========================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
female                      2.4697      0.198     12.470      0.000       2.082       2.858
age_r                      -0.0172      0.005     -3.589      0.000      -0.027      -0.008
initial_fee_level           0.0026      0.001      2.320      0.020       0.000       0.005
credit_account_id_dummy    -1.4308      0.173     -8.272      0.000      -1.770      -1.092
===========================================================================================
                         Odds Ratio  95% CI Lower  95% CI Upper
female                    11.818802      8.016535     17.424496
age_r                      0.982899      0.973684      0.992201
initial_fee_level          1.002578      1.000400      1.004762
credit_account_id_dummy    0.239116      0.170361      0.335617
```

All p_values are significant

- Gender has a terrific impact on the conversion rate: a female customer increase the probability to convert by a magnitude of ~12
- Much lower impact of the other variables: initial fees has a very minor impact (very close to 0) but positive
- This analysis confirms that younger customers tend to convert, but the magnitude is not as strong as for the gender
- It seems that having a credit_account_id increase the probability to convert with a magnitude of 0.2

What are the factors for predicting conversion?

- Segment analysis

```
Optimization terminated successfully.
         Current function value: 0.621790
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      710
Method:                           MLE   Df Model:                            1
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                 0.07855
Time:                        12:17:39   Log-Likelihood:                -442.71
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 3.701e-18
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
customer_segment_12    -0.0810      0.152     -0.532      0.595      -0.379       0.217
customer_segment_13    -1.1558      0.124     -9.293      0.000      -1.400      -0.912
=======================================================================================
                    Odds Ratio  95% CI Lower  95% CI Upper
customer_segment_12   0.922222      0.684381      1.242719
customer_segment_13   0.314815      0.246712      0.401718
```

- Being in Segment 11 has a positive impact on the conversion compared to segment 13 – Magnitude ~0.3.
- Impact on seg 12 is much smaller and in any case not significant.

What are the factors for predicting conversion?

- Related Customer*

```
Optimization terminated successfully.
         Current function value: 0.662886
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      708
Method:                           MLE   Df Model:                            3
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                 0.01764
Time:                        12:30:28   Log-Likelihood:                -471.98
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 0.0007224
=====================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
related_customers_0    -0.5371      0.096     -5.612      0.000      -0.725      -0.349
related_customers_2    -0.2412      0.403     -0.599      0.549      -1.031       0.549
related_customers_3    -0.6931      0.612     -1.132      0.258      -1.893       0.507
related_customers_4    -1.6094      0.632     -2.545      0.011      -2.849      -0.370
=====================================================================================
                     Odds Ratio  95% CI Lower  95% CI Upper
related_customers_0    0.584459      0.484500      0.705043
related_customers_2    0.785714      0.356702      1.730707
related_customers_3    0.500000      0.150563      1.660437
related_customers_4    0.200000      0.057901      0.690839
```

- Having zero related customer has a negative impact on the conversion compared to having only one related customer – Magnitude ~0.58.
- The other groups are not significant a part from related_customer_4 with a magnitude = 0.2

*related customer 5 have to be dropped from the analysis as it contains only ”not converted” data

What are the factors for predicting conversion?

- Family size*

```
Optimization terminated successfully.
         Current function value: 0.660808
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      710
Method:                           MLE   Df Model:                            1
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                 0.02072
Time:                        12:35:00   Log-Likelihood:                -470.50
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 8.107e-06
=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
family_size_0    -0.5992      0.092     -6.530      0.000      -0.779      -0.419
family_size_2     0.2963      0.245      1.208      0.227      -0.184       0.777
=================================================================================
               Odds Ratio  95% CI Lower  95% CI Upper
family_size_0    0.549254      0.458846      0.657475
family_size_2    1.344828      0.831669      2.174616
```

- Having a family size of 1 has a positive impact on the conversion compared to having zero family size– Magnitude ~0.54.
- The other groups were not statistically significant

*family size >2 have to be dropped from the analysis as it contains only ”not converted” or only ”converted” data

What are the factors for predicting conversion?

- branch analysis

```
Optimization terminated successfully.
         Current function value: 0.659751
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      710
Method:                           MLE   Df Model:                            1
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                 0.02229
Time:                        12:37:36   Log-Likelihood:                -469.74
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 3.691e-06
==============================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
branch_Helsinki    -0.5632      0.088     -6.373      0.000      -0.736      -0.390
branch_Turku       -0.9163      0.418     -2.190      0.028      -1.736      -0.096
====================================================================================
                 Odds Ratio  95% CI Lower  95% CI Upper
branch_Helsinki    0.569405      0.478858      0.677074
branch_Turku       0.400000      0.176188      0.908120
```

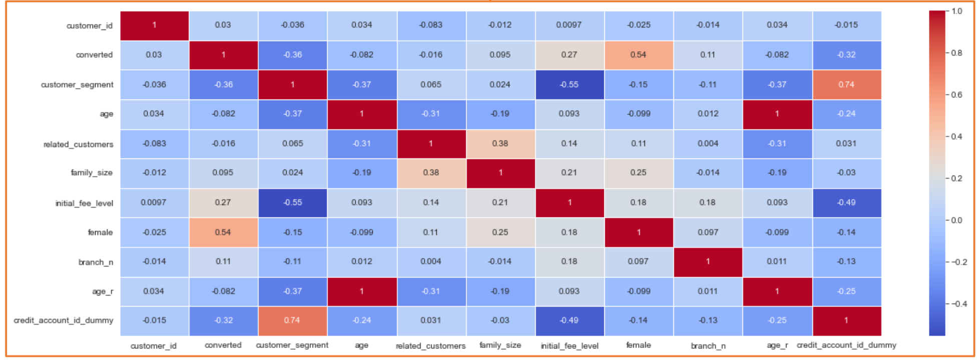
- Being in Tampere branch has a positive impact on the conversion compared to Branch Helsinki– Magnitude ~0.56 and branch Turku – magnitude= 0.4.
- Impact on seg 12 is much smaller and in any case not significant.

What are the factors for predicting conversion?

- Using all relevant variables:

```
Optimization terminated successfully.
         Current function value: 0.466059
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      704
Method:                           MLE   Df Model:                            7
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                  0.3093
Time:                        12:52:41   Log-Likelihood:                -331.83
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 2.355e-60
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
female                     2.4851      0.205     12.134      0.000       2.084       2.887
age_r                     -0.0319      0.007     -4.489      0.000      -0.046      -0.018
initial_fee_level          0.0003      0.001      0.251      0.802      -0.002       0.003
credit_account_id_dummy   -1.2264      0.236     -5.204      0.000      -1.688      -0.764
customer_segment_11        0.8626      0.375      2.298      0.022       0.127       1.598
related_customers_0        0.1925      0.205      0.940      0.347      -0.209       0.594
family_size_1              0.2698      0.264      1.021      0.307      -0.248       0.788
branch_Tampere             0.4292      0.266      1.611      0.107      -0.093       0.952
==========================================================================================
                         Odds Ratio  95% CI Lower  95% CI Upper
female                    12.002753      8.034236     17.931524
age_r                      0.968607      0.955212      0.982190
initial_fee_level          1.000293      0.998010      1.002581
credit_account_id_dummy    0.293350      0.184835      0.465574
customer_segment_11        2.369264      1.135426      4.943884
related_customers_0        1.212276      0.811454      1.811087
family_size_1              1.309682      0.780099      2.198781
branch_Tampere             1.536072      0.911147      2.589613
```

- Logistic regression cannot be performed in case of multicollinearity. I need to get read of the independent variables that may be correlated



What are the factors for predicting conversion?

- Using all relevant variables – controlling for the multicollinearity:

```
Optimization terminated successfully.
         Current function value: 0.509169
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:              converted   No. Observations:                  712
Model:                          Logit   Df Residuals:                      708
Method:                           MLE   Df Model:                            3
Date:                Sun, 03 Mar 2024   Pseudo R-squ.:                  0.2454
Time:                        13:08:54   Log-Likelihood:                -362.53
converged:                       True   LL-Null:                       -480.45
Covariance Type:            nonrobust   LLR p-value:                 7.524e-51
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
female                  2.1776      0.188     11.599      0.000       1.810       2.546
related_customers_0    -0.5699      0.164     -3.482      0.000      -0.891      -0.249
customer_segment_11     0.9954      0.200      4.984      0.000       0.604       1.387
branch_Helsinki        -1.2365      0.161     -7.677      0.000      -1.552      -0.921
=======================================================================================
                     Odds Ratio  95% CI Lower  95% CI Upper
female                 8.825198      6.108289     12.750563
related_customers_0    0.565568      0.410361      0.779477
customer_segment_11    2.705847      1.829349      4.002302
branch_Helsinki        0.290387      0.211777      0.398178
```

Using only the variables that are not correlated among them (avoiding multicollinearity) we have that:

- Gender is strongly the most significant discriminator for predicting the conversion rate
- Having at least one related customer, increase the probability to conversion.
- Being part of segment 11 increase the
- Branch Toperne seems to increase the probability to covert a customer compared to Helsinki

Agenda

- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

Conclusions I

From a descriptive analysis, “Converted” customers seems to be younger, pay higher initial fee, be Female, be in Segment 11, be from Tampere Branch, have at least one person related, have family size is between 1 and 2

From **our logistic regressions*** independently run we learned:

1. Gender has a terrific impact on the conversion rate: a female customer increase the probability to convert by a magnitude of ~12
2. Younger customers tend to convert, but the magnitude of ~0.98
3. Having a credit_account_id increase the probability to convert with a magnitude of 0.2
4. Being in Segment 11 has a positive impact on the conversion compared to segment 13 – Magnitude ~0.3.
5. Having **zero related customer** has a negative impact on the conversion compared to having only one related customer – Magnitude ~0.58.
6. Having a family size of 1 has a positive impact on the conversion compared to having zero family size– Magnitude ~0.54.
7. Being in Tampere branch has a positive impact on the conversion compared to Branch Helsinki– Magnitude ~0.56 and branch Turku – magnitude= 0.4.

* All these results are statistically significant for a P_value<0.01

Conclusions II

Due to some correlation among some of the independent variable I had to select only some of the variables to the overall logistic regression.

Specifically I had to choose one between:

- age, related customers and family size (as they seems to be somehow correlated)
- Initial_fee_level, credit_accoiunt_id, customer segment (as they seems to be somehow correlated)

From **our overall logistic regression*** we learned:

1. Being Female is the most significant discriminator for predicting the conversion rate magnitude of ~8.8
2. Having at least one related customer, increase the probability to conversion magnitude of ~0.56
3. Being part of Segment 11 increase the probability to conversion with a magnitude of ~2.7
4. Branch Toperne seems to increase the probability to covert a customer compared to Helsinki with a magnitude of ~0.29

Changing the variable used (e.g. “age” instead of “related customers”) will give similar results to the ones we saw in the independent logistic statistics run before.

* All these results are statistically significant for a P_value<0.01

Further discussion

Since I couldn´t communicate with the **data/product owner**, there are some assumption I had to make. A deeper knowledge of the data would lead to a different analysis.

Selecting different independent variables would also push the results in slightly different direction, based on **the point of view** we want to give to our data.

We were "lucky" this time as the data fit all the assumption for a logistic regression. But what if it was not the case?
Based on the different assumptions that were not satisfied we could have used a different "solution".

In the worst case scenario, and there was no way to fix the assumptions, we could have made a "one by one" analysis of the independent variables.

Specifically to Test whether the variable converted is correlated with any of the other categorical variables using a **chi-square analysis**.

While to test the correlation with age and/or initial_fee_level, I would have use a simple **T-test**, in case that all the assumptions were satisfied.
If the T-test assumption were not satisfied I would have used the Non-parametric version of it: "**Wilcoxon-Mann Whitney test**“*

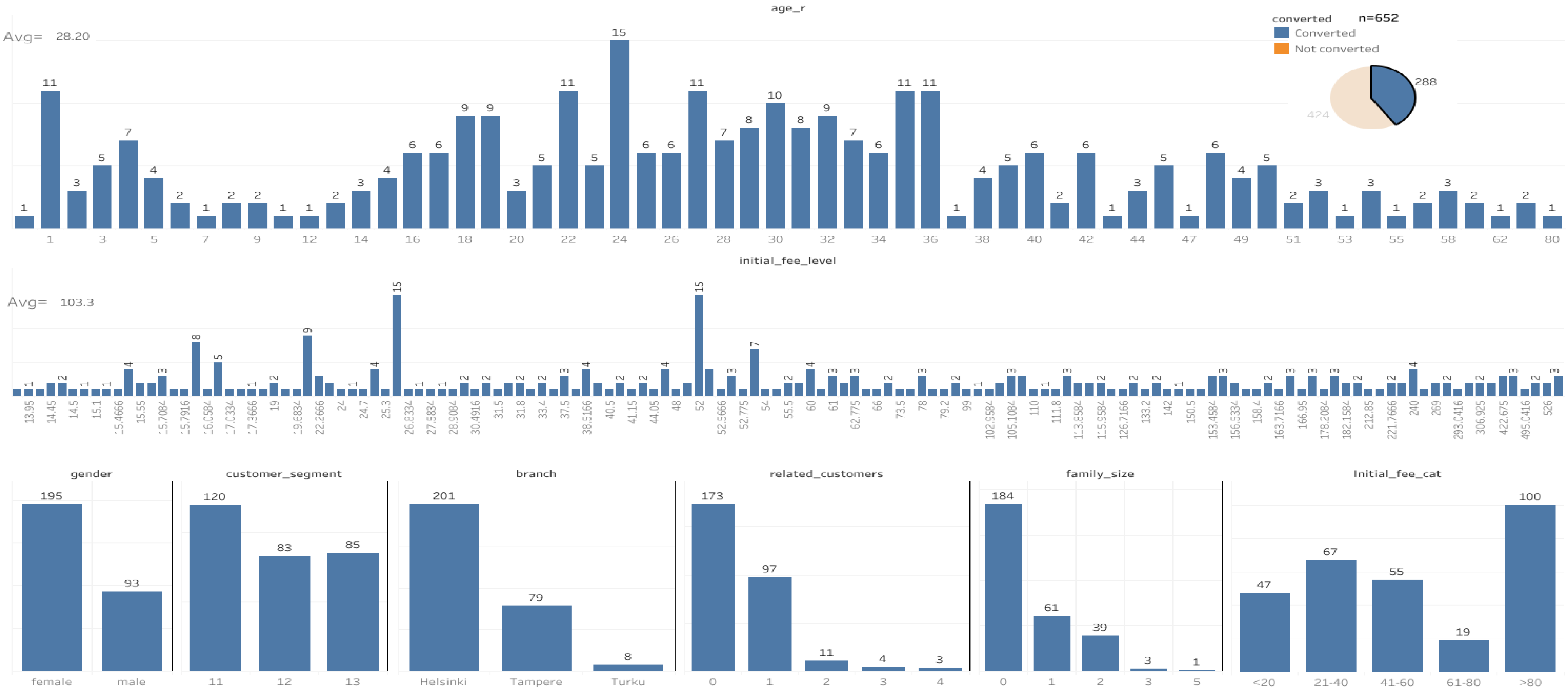
Running the tests as described above, they all confirm the logistic regression results

*https://stats.oarc.ucla.edu/spss/whatstat/what-statistical-analysis-should-i-usestatistical-analyses-using-spss/#wilc

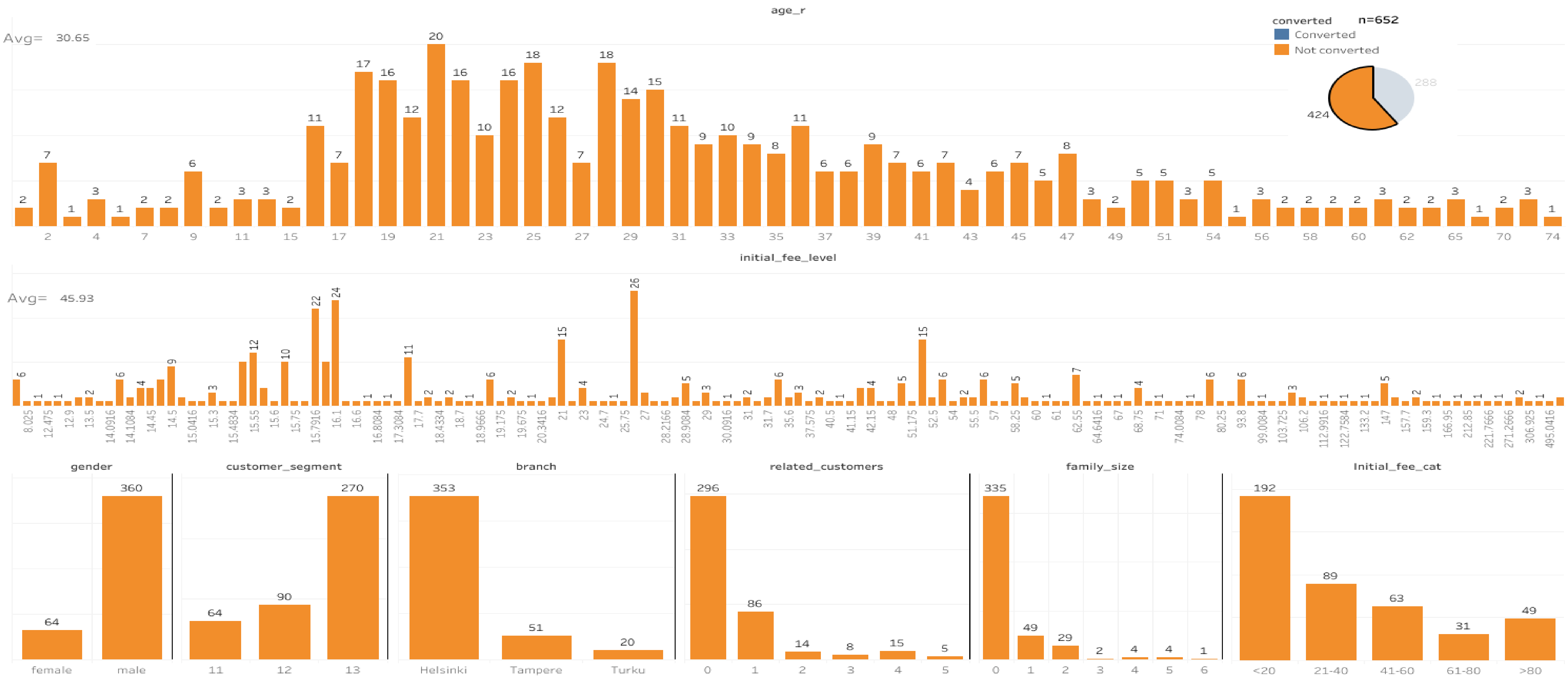
Agenda

- Descriptive statistics
- Correlation analysis
- What are the important factors?
- Conclusions and further discussion
- Appendix

Descriptive statistics - Converted

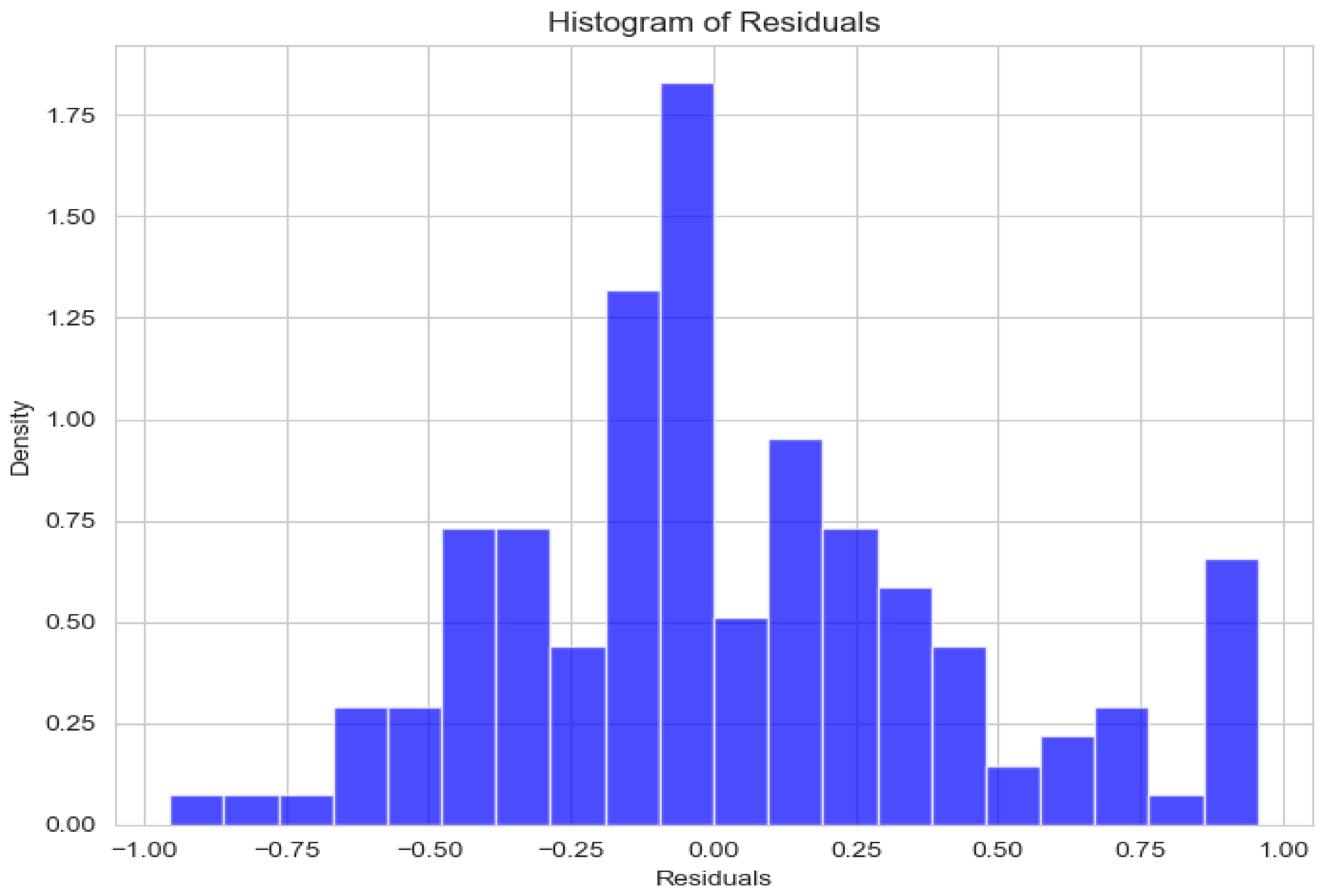


Descriptive statistics – Not Converted

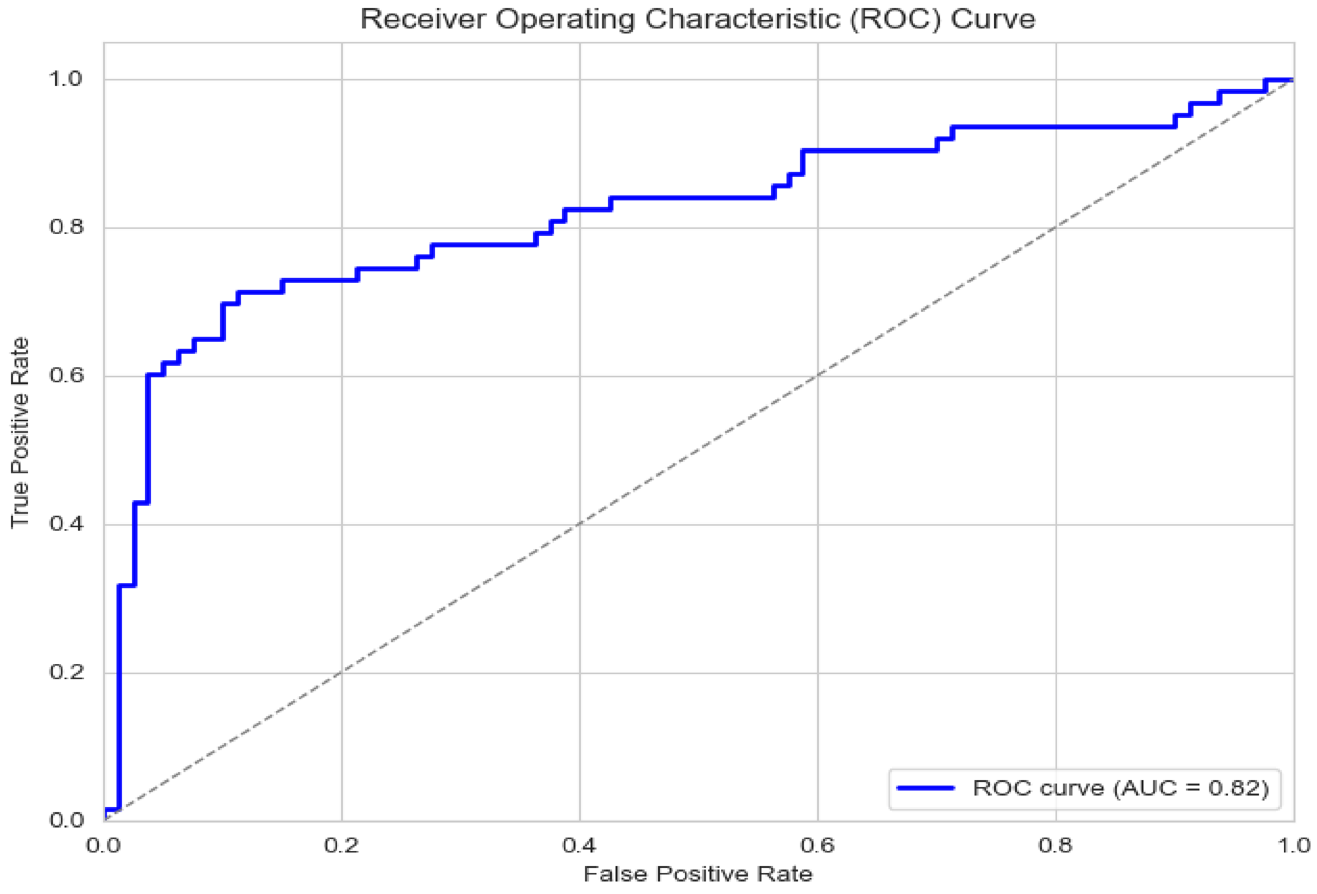


Correlation

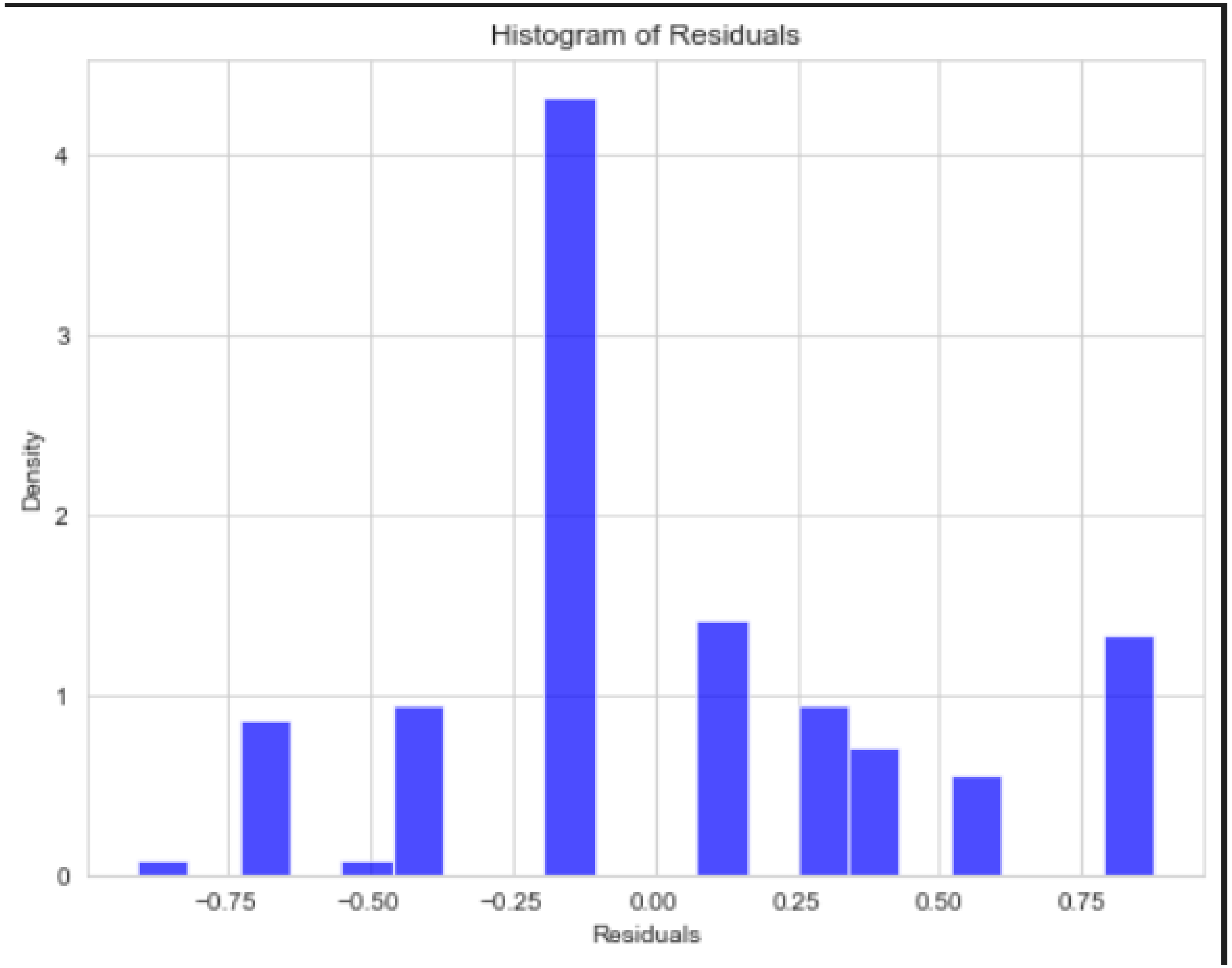
Logistic analysis



Logistic analysis



Logistic analysis – final graph



LENUS

Logistic analysis – final graph