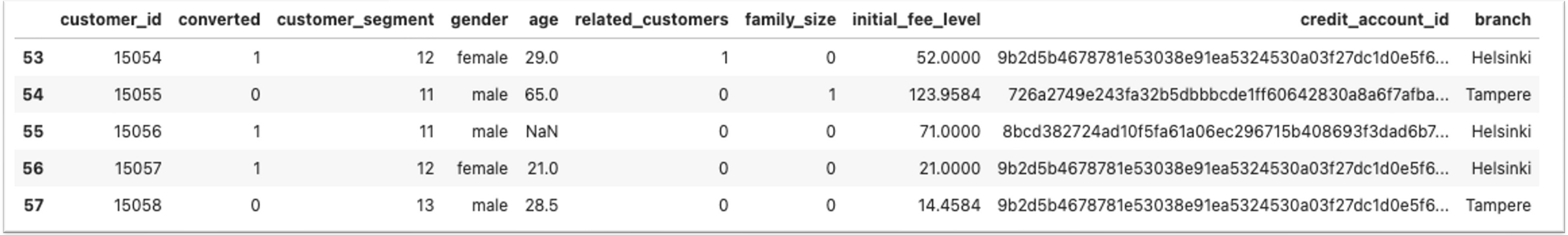
Dhouha Grissa

Customer Segmentation Chalenge Analysis

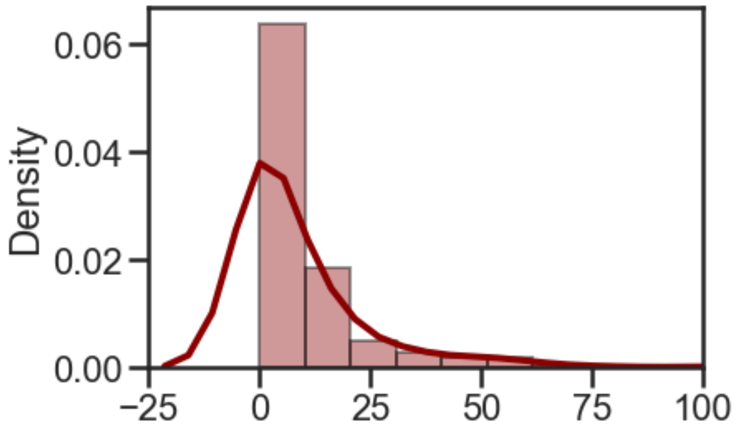
Challenge

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

Input data

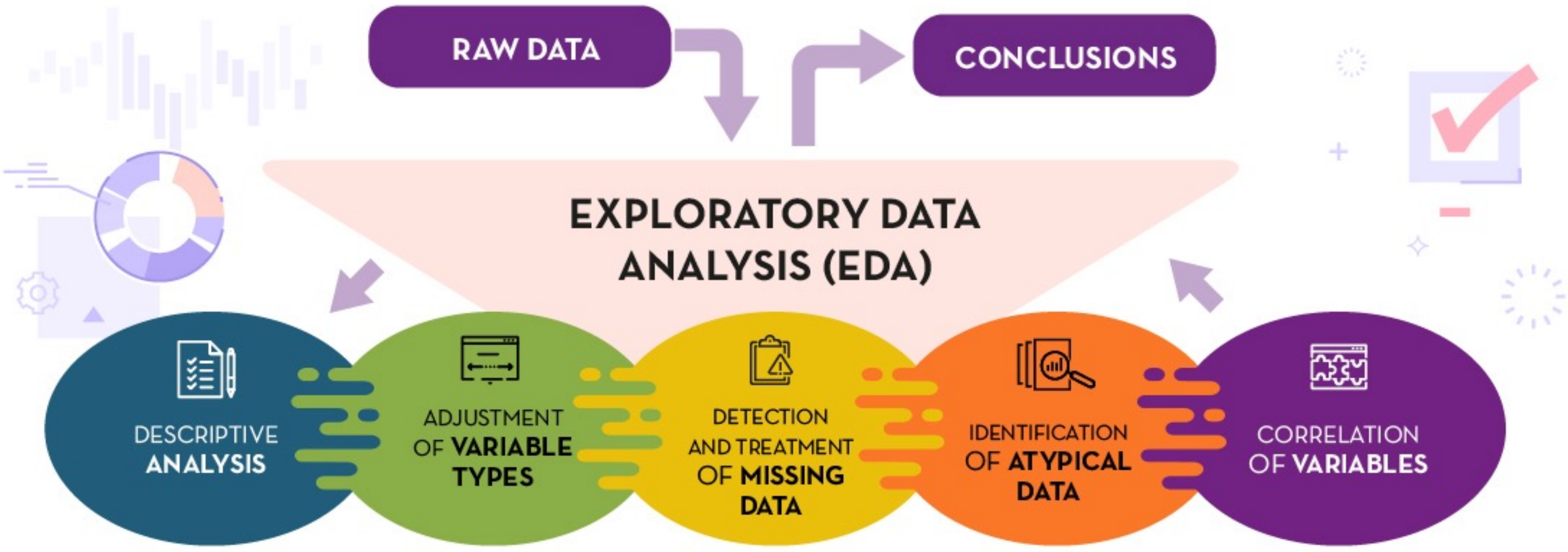


	customer_id	converted	customer_segment	gender	age	related_customers	family_size	initial_fee_level	credit_account_id	branch
53	15054	1	12	female	29.0	1	0	52.0000	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f6...	Helsinki
54	15055	0	11	male	65.0	0	1	123.9584	726a2749e243fa32b5dbbbcde1ff60642830a8a6f7afba...	Tampere
55	15056	1	11	male	NaN	0	0	71.0000	8bcd382724ad10f5fa61a06ec296715b408693f3dad6b7...	Helsinki
56	15057	1	12	female	21.0	0	0	21.0000	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f6...	Helsinki
57	15058	0	13	male	28.5	0	0	14.4584	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f6...	Tampere



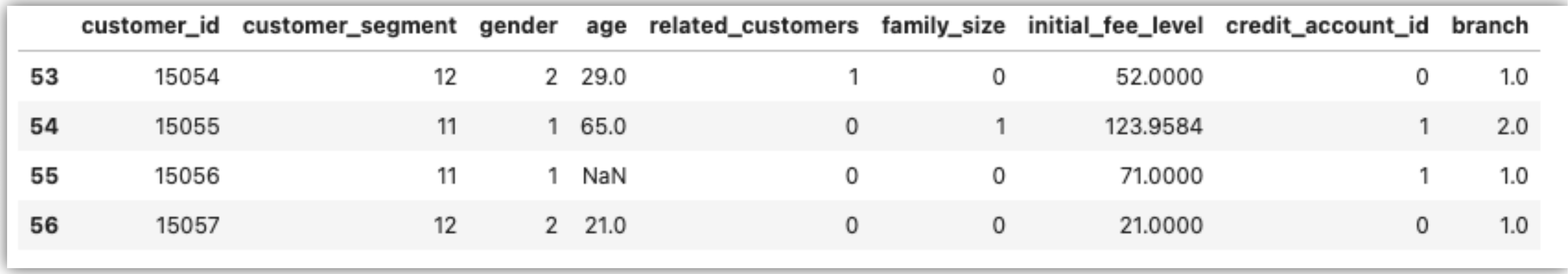
- 9 variables describing every customer
- Heterogenous data: numeric, categorical, etc.
- Data with missing values
- Data is not normally distributed

Exploratory data analysis


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

https://datos.gob.es/

Adjustment of variable types

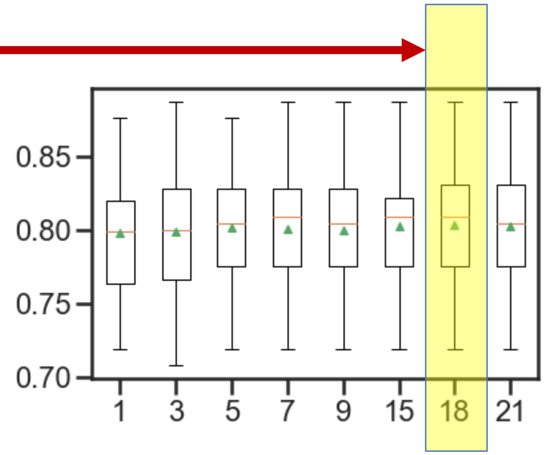


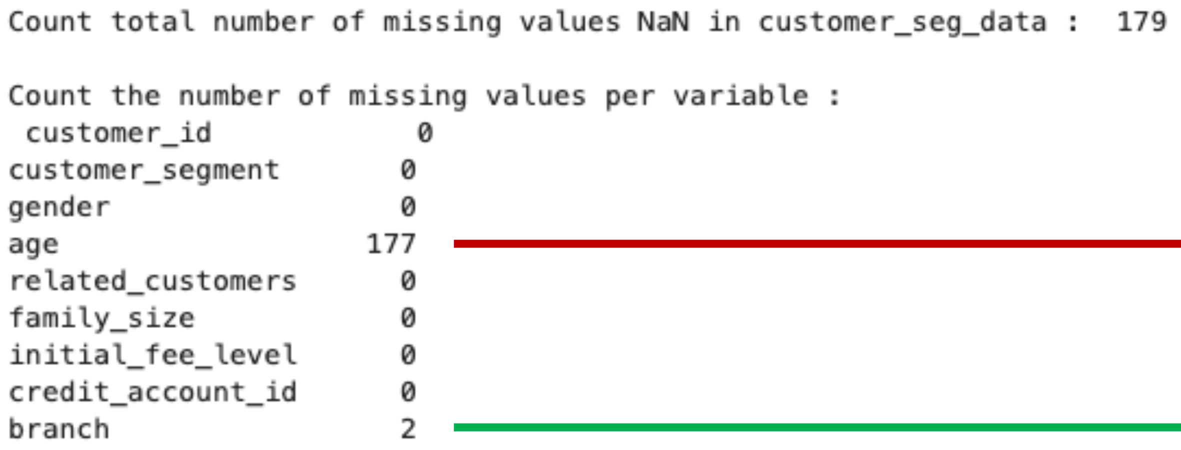
	customer_id	customer_segment	gender	age	related_customers	family_size	initial_fee_level	credit_account_id	branch
53	15054	12	2	29.0	1	0	52.0000	0	1.0
54	15055	11	1	65.0	0	1	123.9584	1	2.0
55	15056	11	1	NaN	0	0	71.0000	1	1.0
56	15057	12	2	21.0	0	0	21.0000	0	1.0

- Converting the following variables to numerical variables:
  - gender
  - credit_account_id
  - branch

Detection and Treatment of missing values

Choice of k

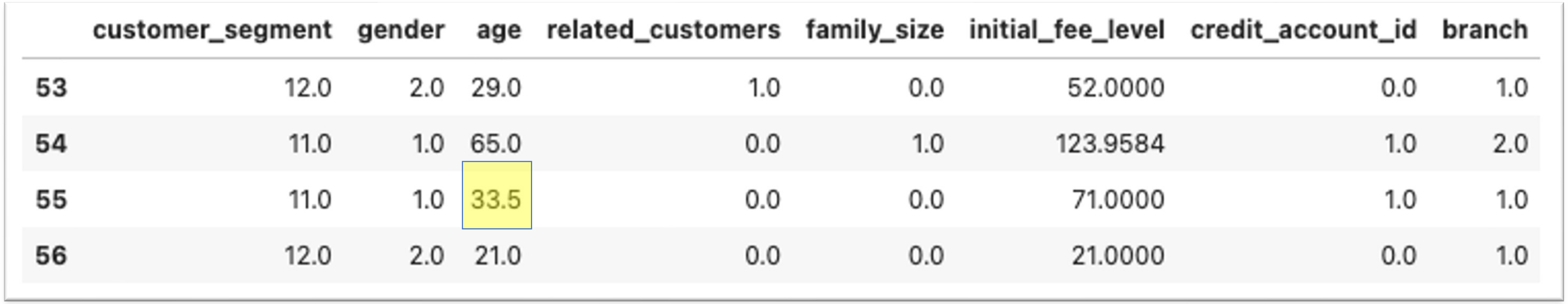




```
Count total number of missing values NaN in customer_seg_data :  179

Count the number of missing values per variable :
 customer_id             0
customer_segment        0
gender                  0
age                   177
related_customers       0
family_size             0
initial_fee_level       0
credit_account_id       0
branch                  2
```

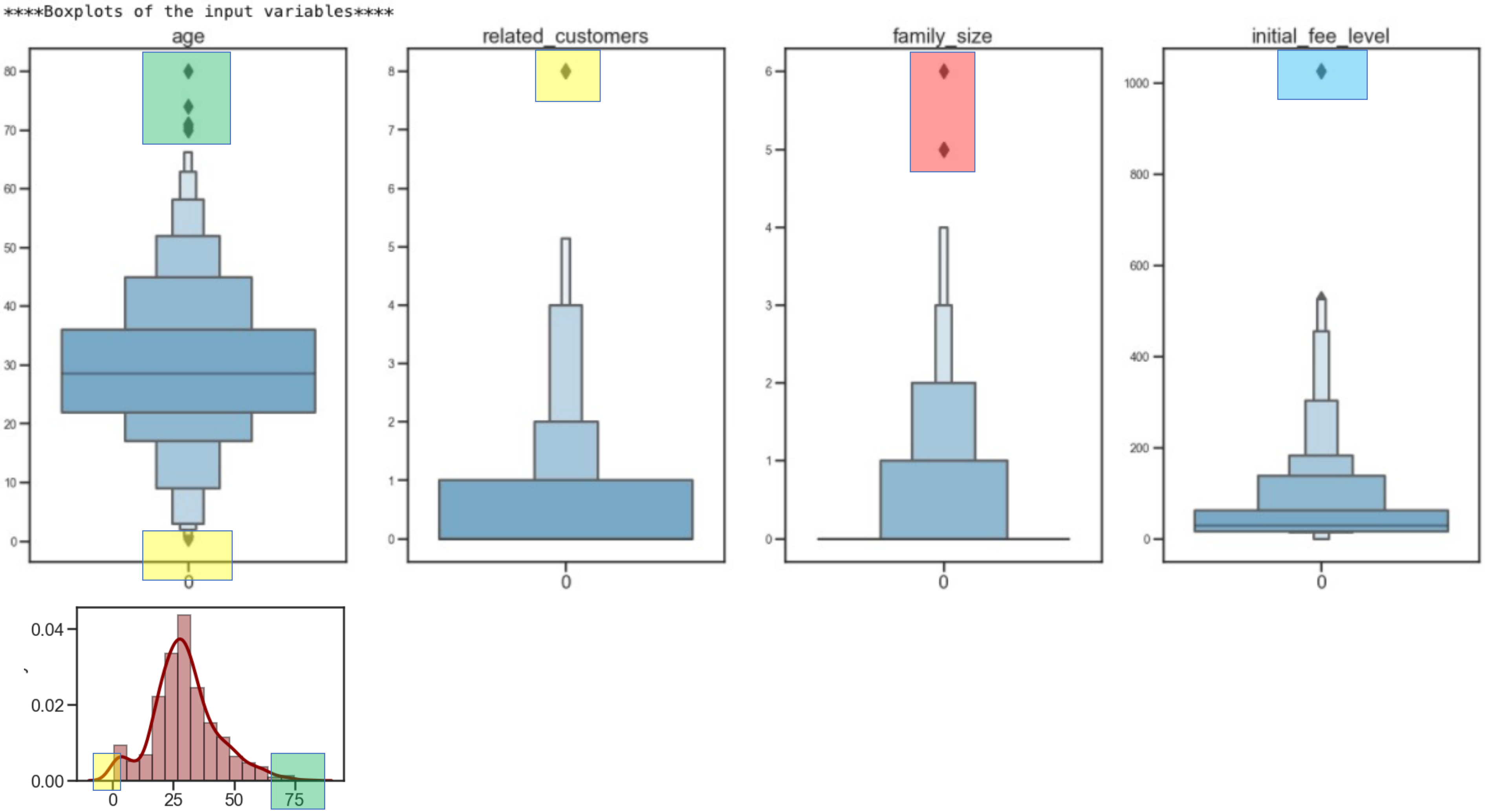
Imputation using knn

Imputation based on the ‘most frequent’ value



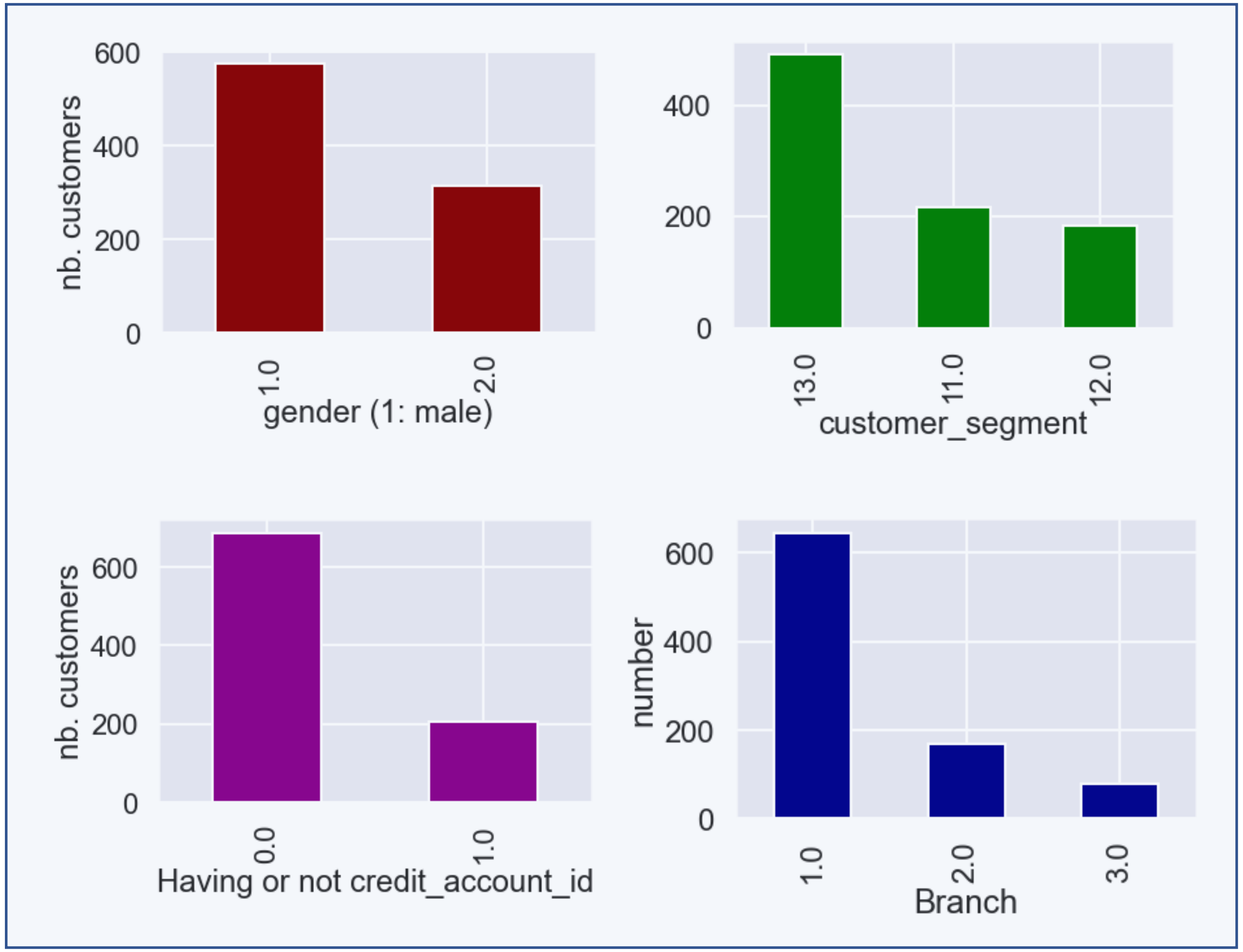
	customer_segment	gender	age	related_customers	family_size	initial_fee_level	credit_account_id	branch
53	12.0	2.0	29.0	1.0	0.0	52.0000	0.0	1.0
54	11.0	1.0	65.0	0.0	1.0	123.9584	1.0	2.0
55	11.0	1.0	33.5	0.0	0.0	71.0000	1.0	1.0
56	12.0	2.0	21.0	0.0	0.0	21.0000	0.0	1.0

Identification of Atypical Data



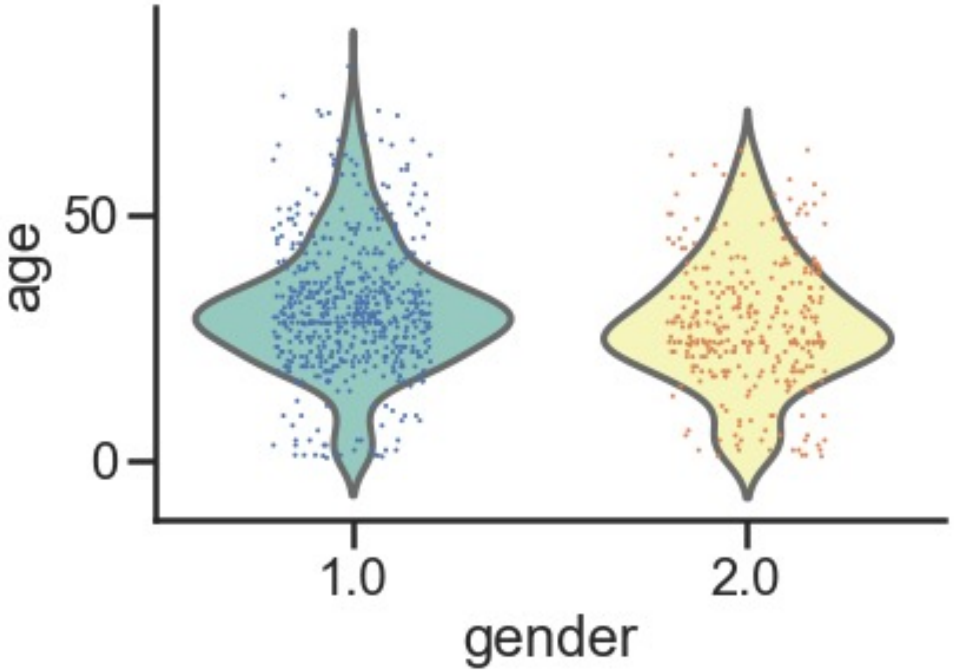
```
****Boxplots of the input variables****
```

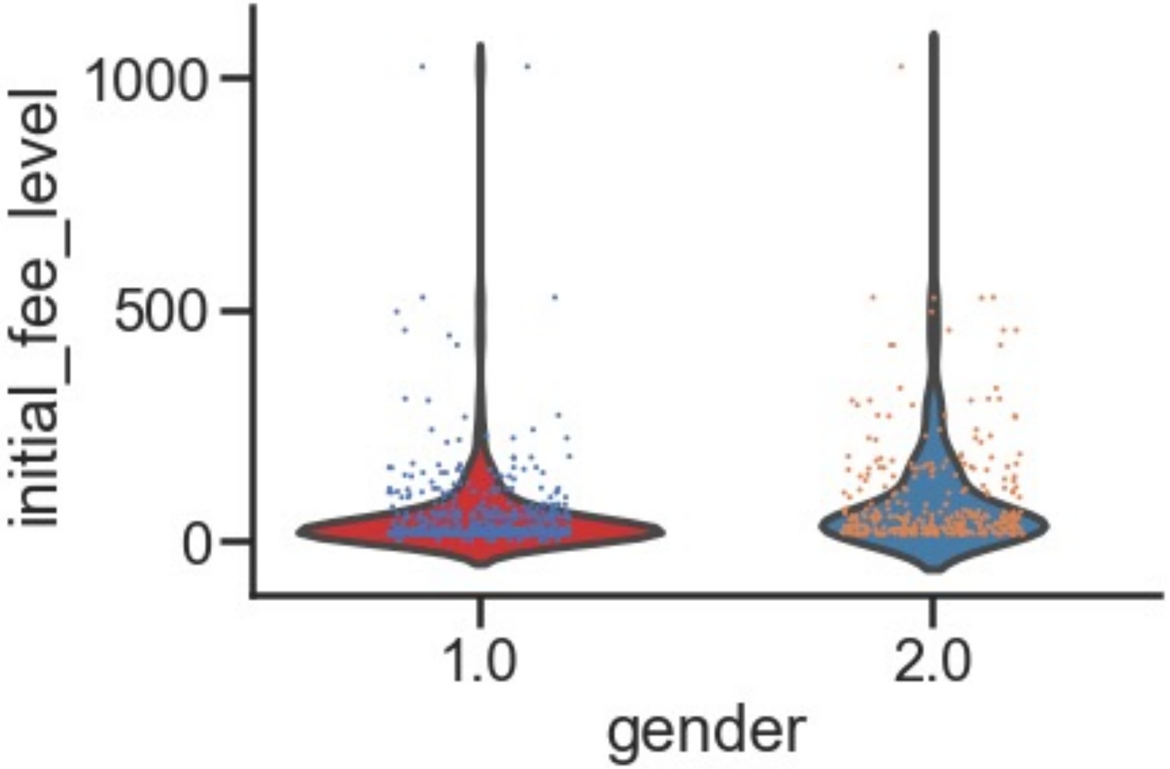
Distribution of the remaining Categorical variables



Does ‘gender’ variable influence the results?

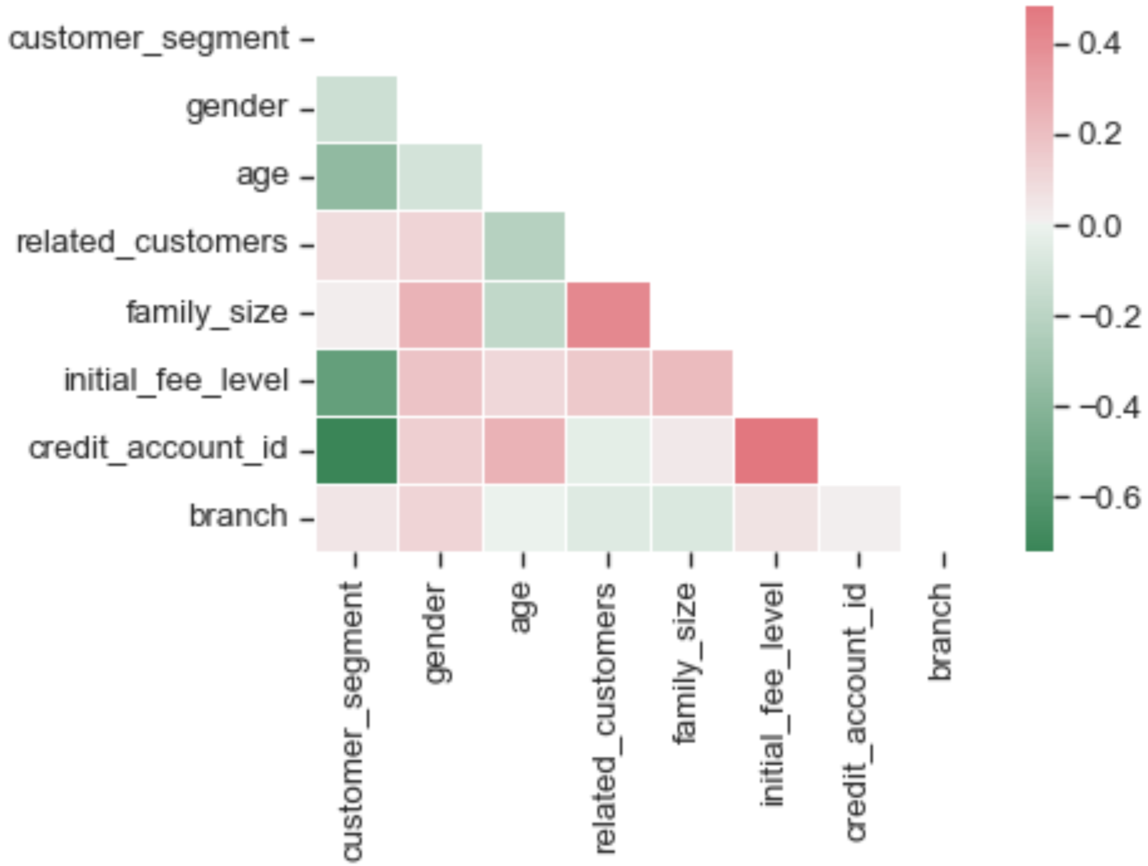
Relationship of ‘gender’ with other variables



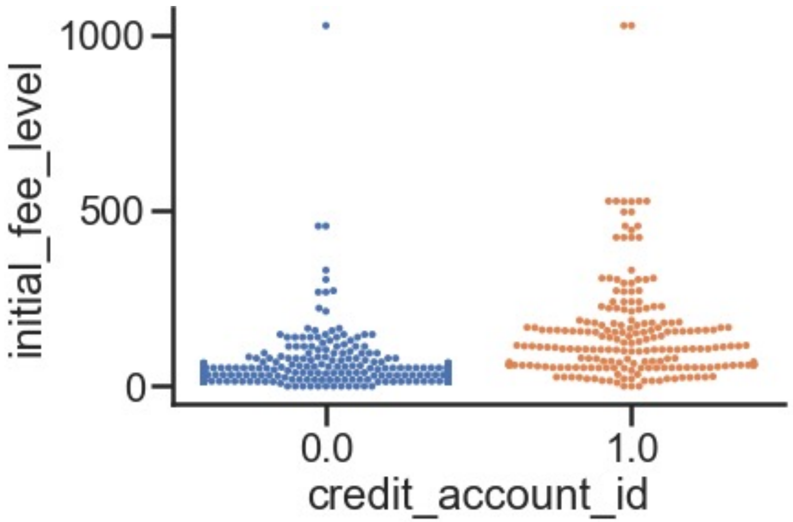


Does ‘gender’ variable influence the results? Response: No

Correlation between the variables

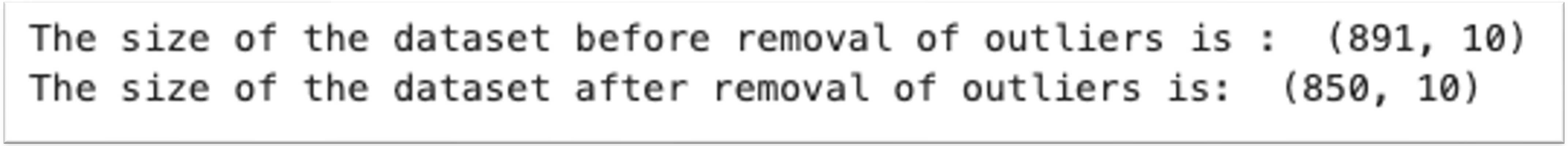


(family_size, related_customers) → cor = 0.48
(credit_account_id, initial_fee_level) → cor = 0.48



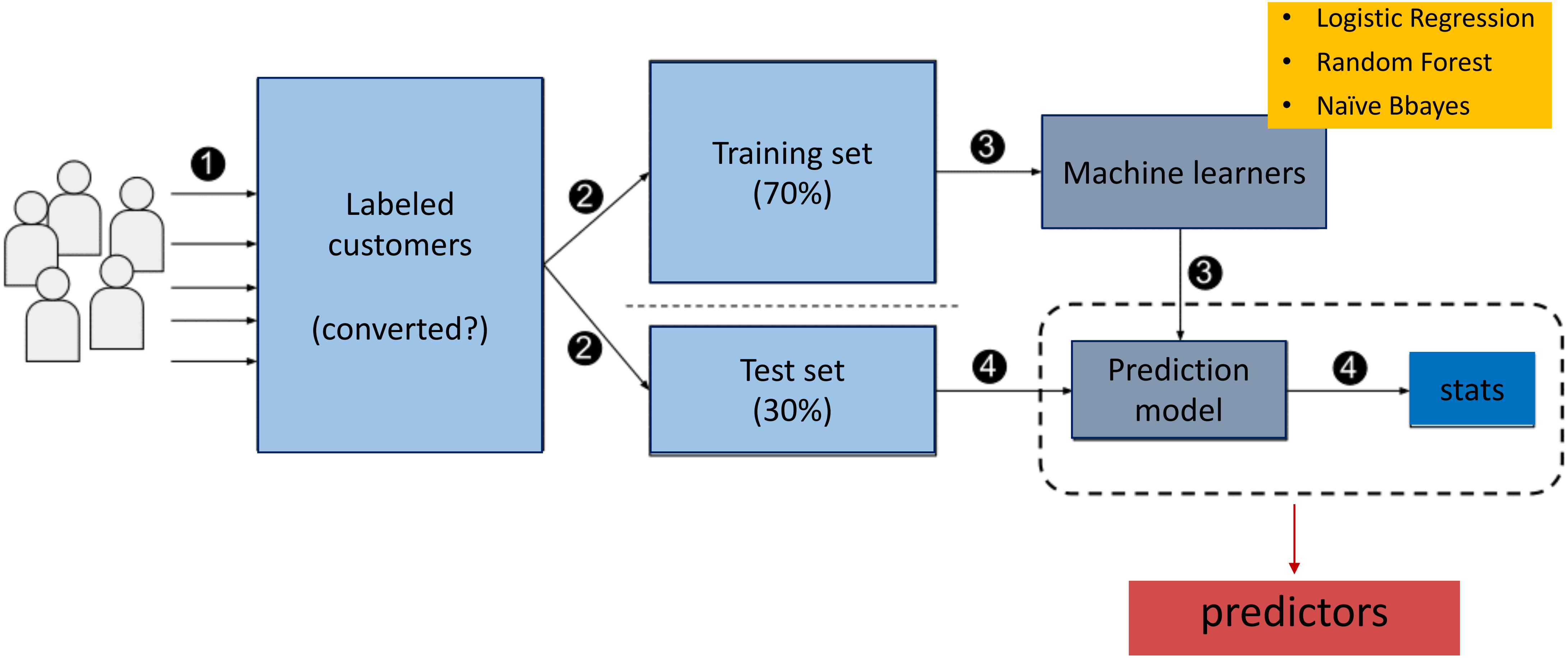
Removal of outliers

- age <= 1 and age => 65
- Initial_fee_level>1000
- Family_size>=5
- Related_customers>= 8



```
The size of the dataset before removal of outliers is :  (891, 10)
The size of the dataset after removal of outliers is:  (850, 10)
```

Supervised Classification and Prediction: Approach



- Logistic Regression
- Random Forest
- Naïve Bbayes

Labeled customers

(converted?)

Training set (70%)

Test set (30%)

Machine learners

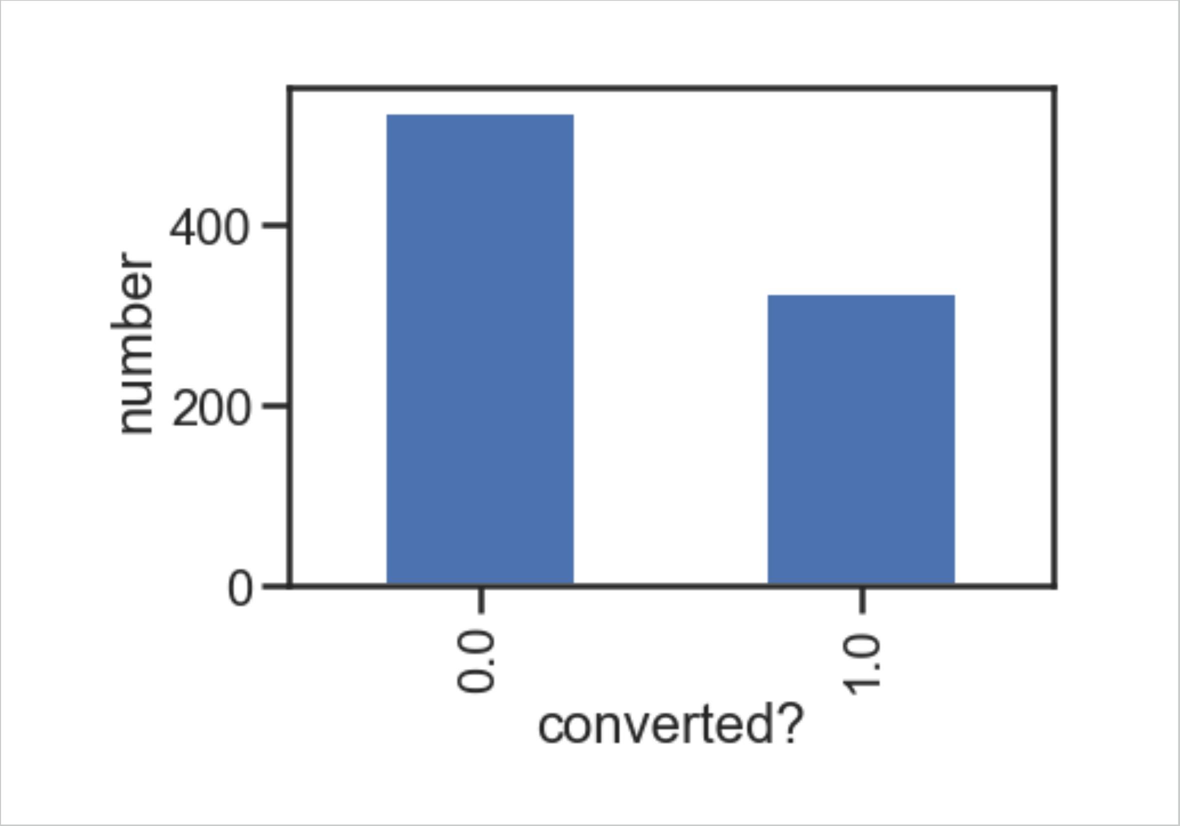
Prediction model

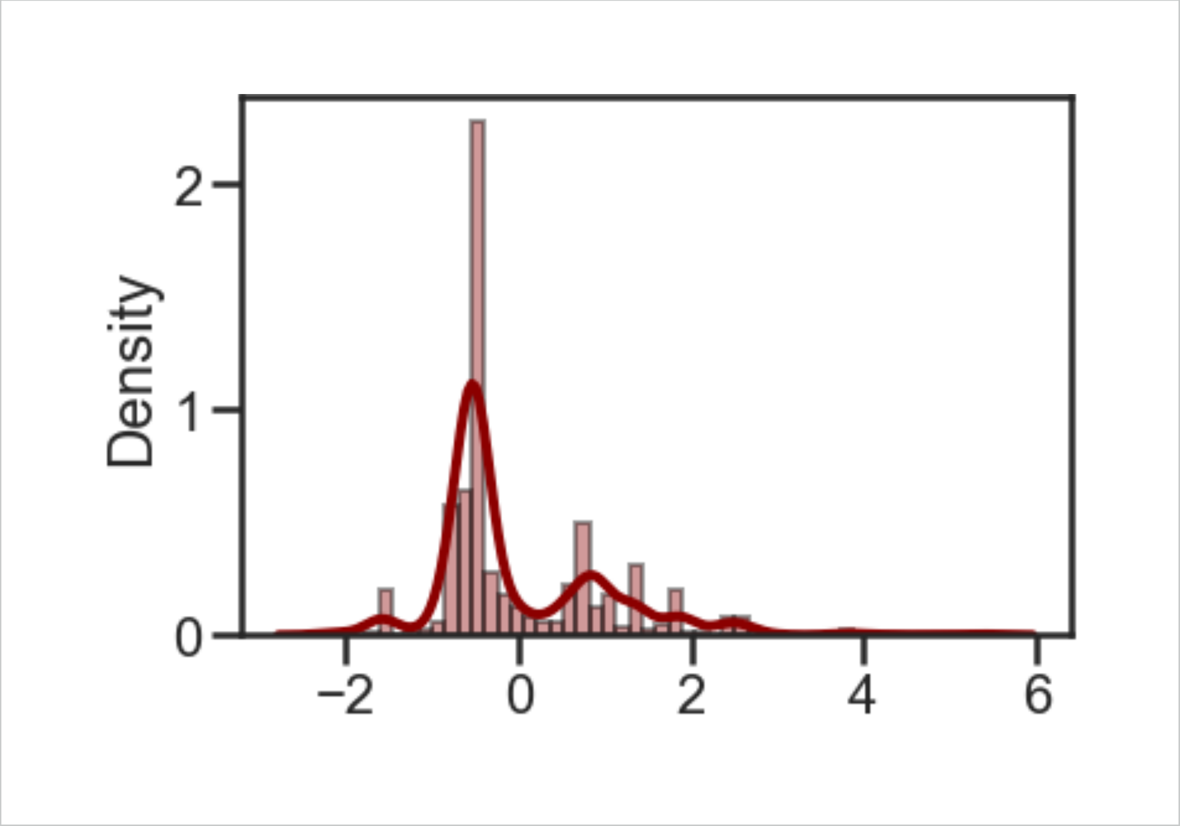
stats

predictors

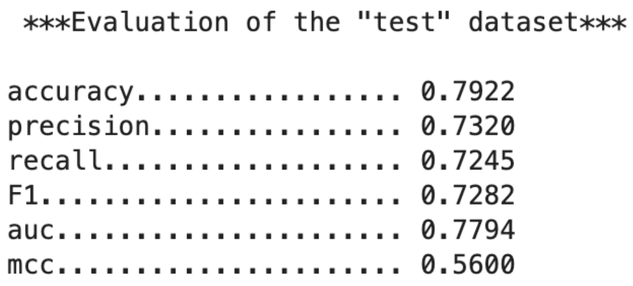
Data: balanced, scaling?

- The data is balanced
- Applied StandardScaler()

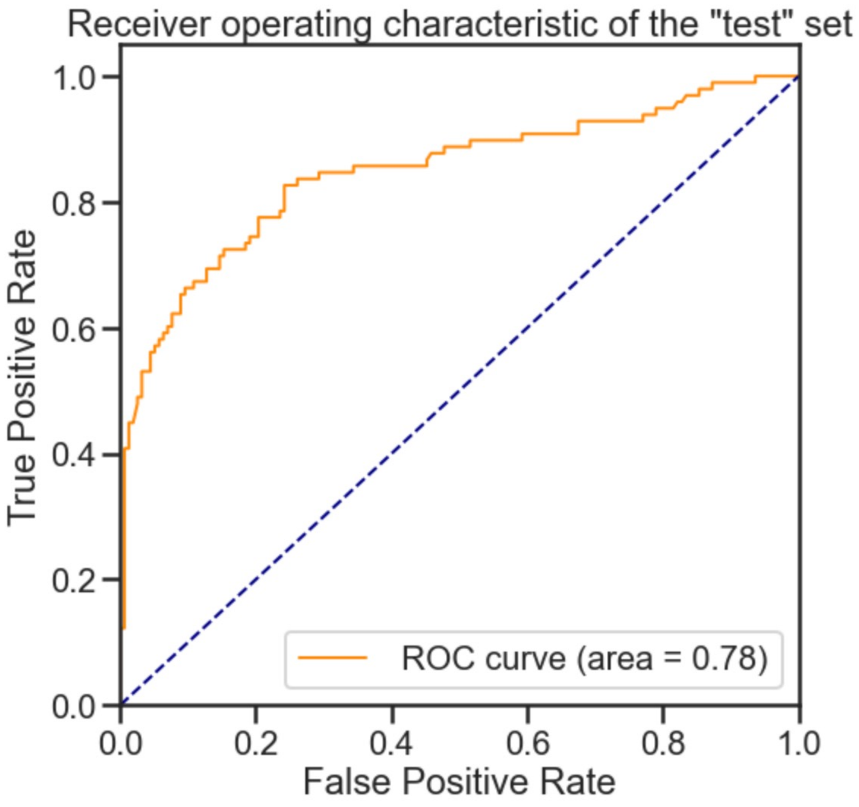




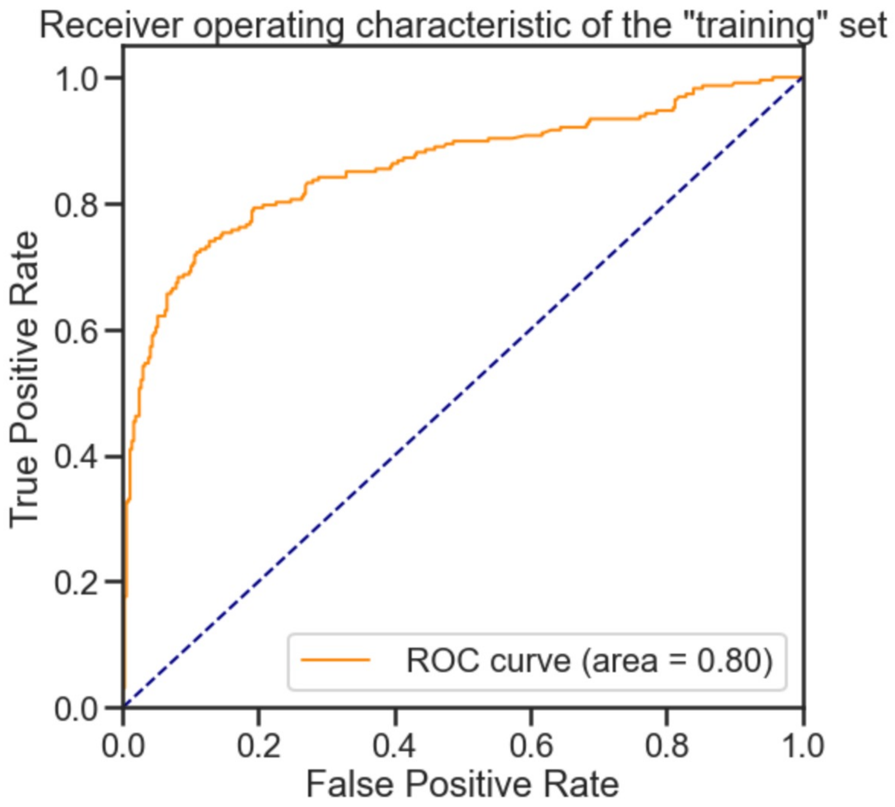
Results of Logistic Regression



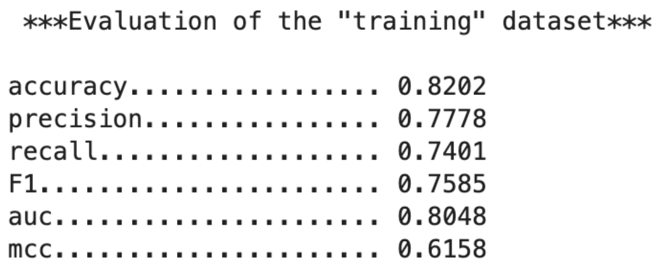
```
 ***Evaluation of the "test" dataset***

accuracy................. 0.7922
precision................ 0.7320
recall................... 0.7245
F1....................... 0.7282
auc...................... 0.7794
mcc...................... 0.5600
```



Receiver operating characteristic of the "test" set

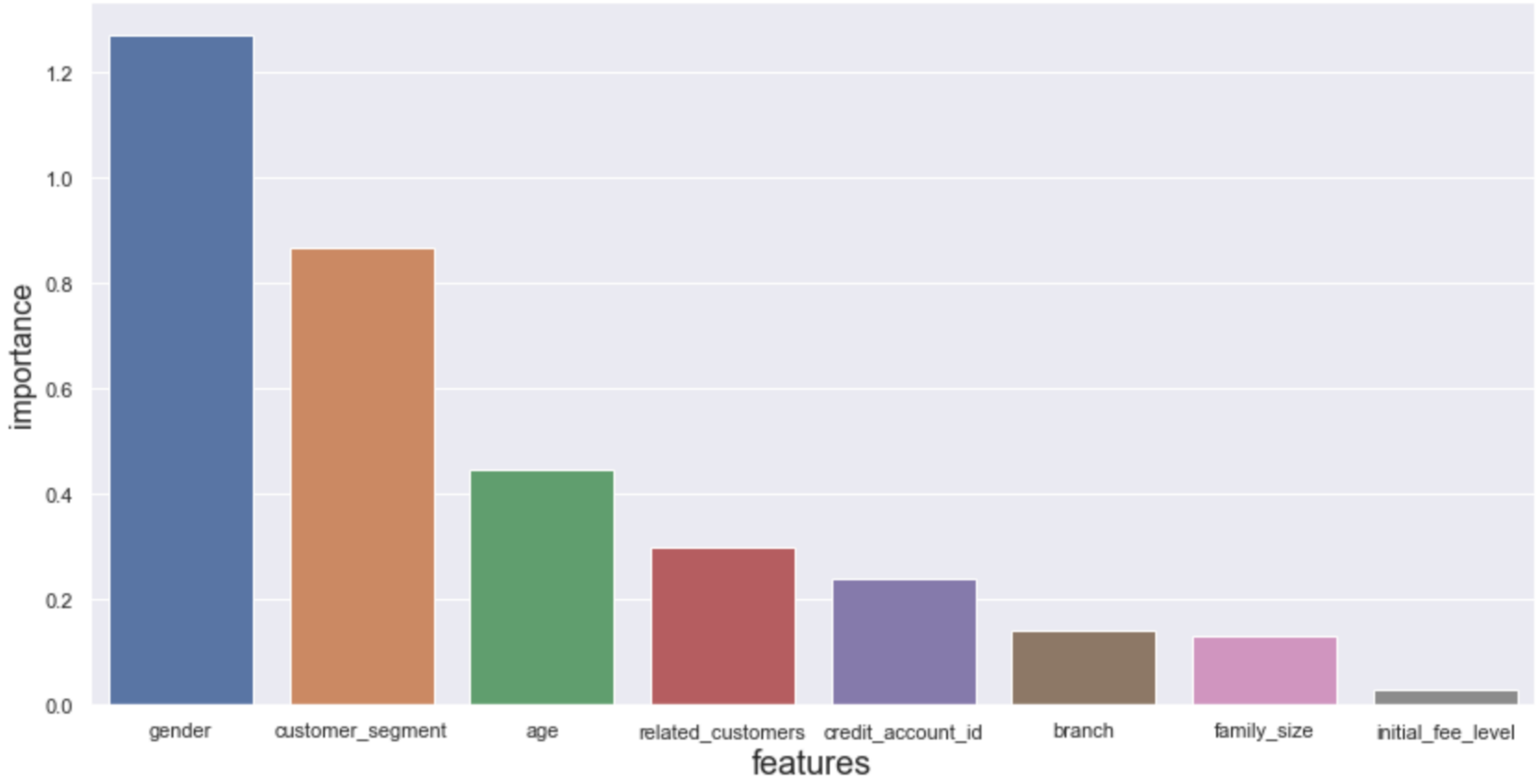


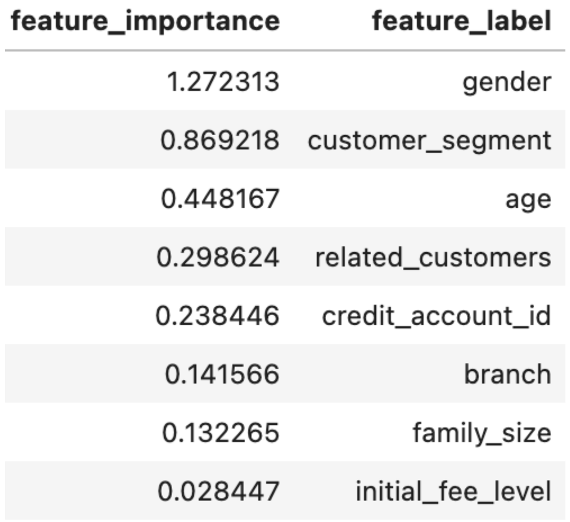
Receiver operating characteristic of the "training" set



```
 ***Evaluation of the "training" dataset***

accuracy................. 0.8202
precision................ 0.7778
recall................... 0.7401
F1....................... 0.7585
auc...................... 0.8048
mcc...................... 0.6158
```

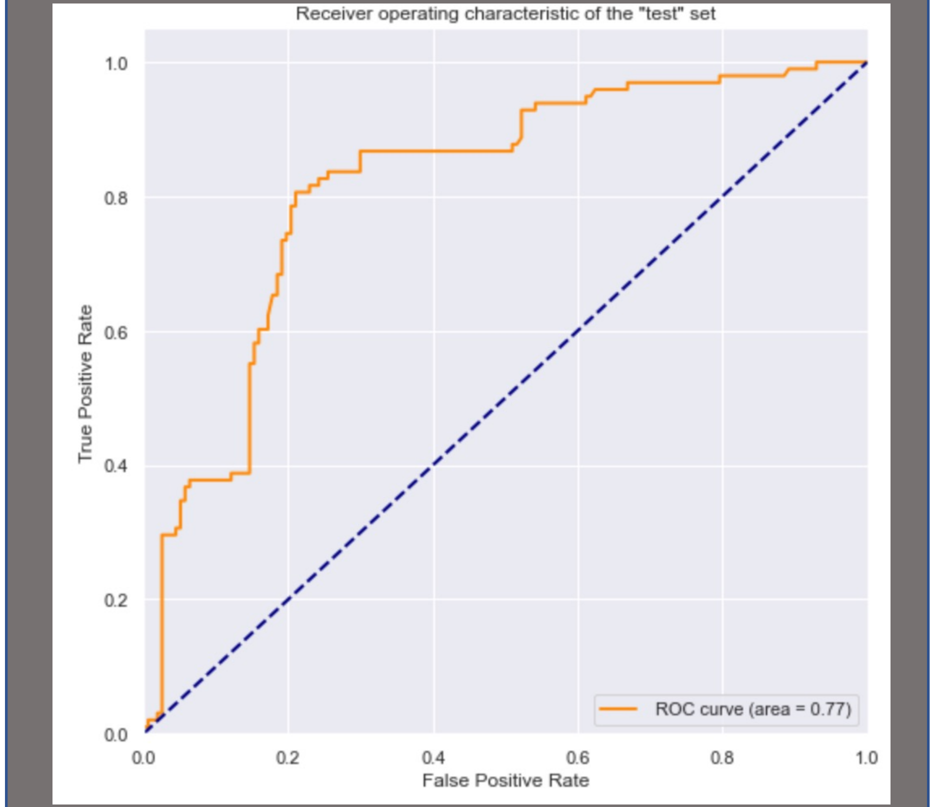
Predictors according to Logistic Regression

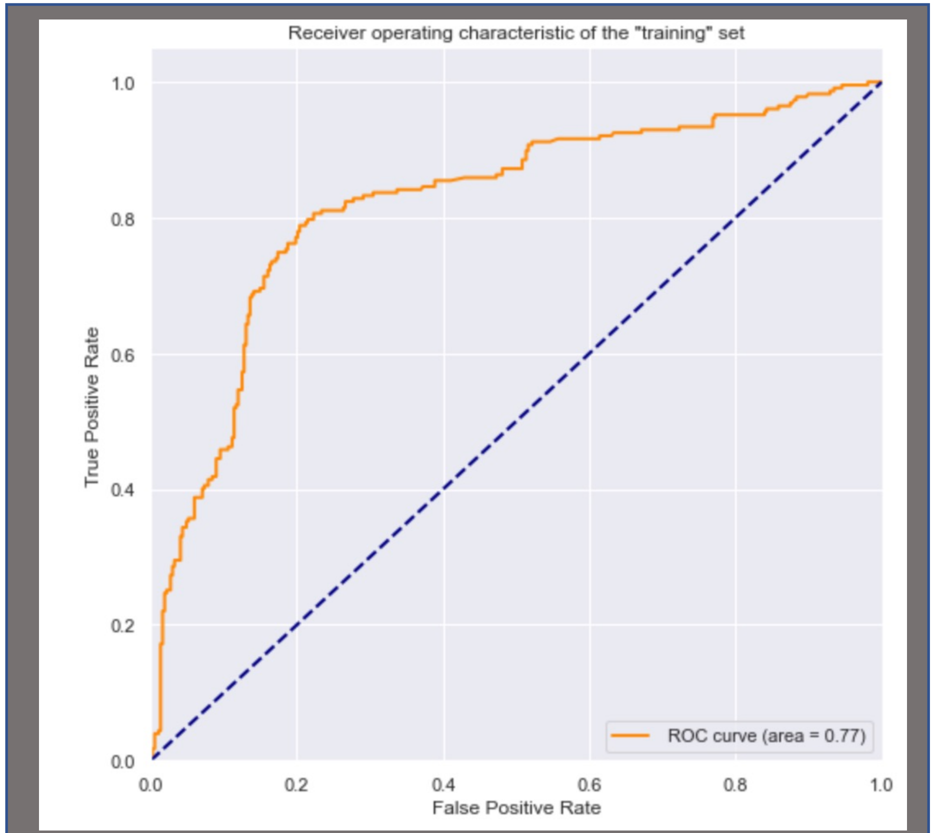


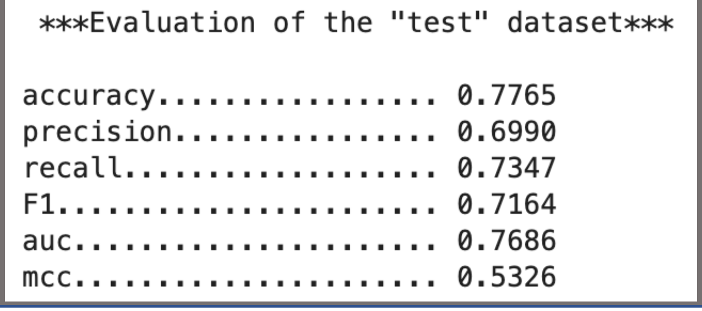


feature_importance	feature_label
1.272313	gender
0.869218	customer_segment
0.448167	age
0.298624	related_customers
0.238446	credit_account_id
0.141566	branch
0.132265	family_size
0.028447	initial_fee_level

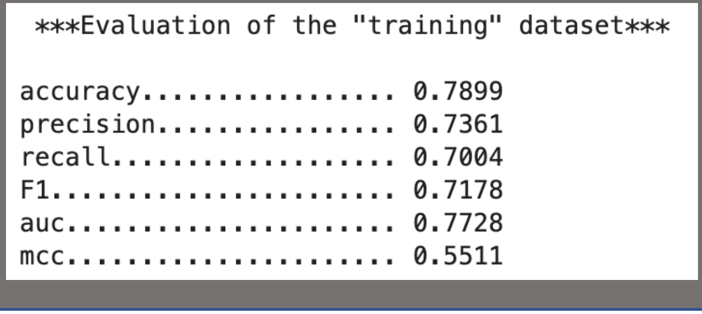
Performance of Naïve bayes







```
 ***Evaluation of the "test" dataset***

accuracy................. 0.7765
precision................ 0.6990
recall................... 0.7347
F1....................... 0.7164
auc...................... 0.7686
mcc...................... 0.5326
```



```
 ***Evaluation of the "training" dataset***

accuracy................. 0.7899
precision................ 0.7361
recall................... 0.7004
F1....................... 0.7178
auc...................... 0.7728
mcc...................... 0.5511
```

Final Results

- The results of LR are promising with an AUC = 0.80, compared to an AUC=0.78 for Naïve Bayes ➔ I rely on the results of LR.
- I have also applied Random Forest but I have noticed an overfitting when comparing the performance of the training vs test sets.

- The major factors that predict if the customer will convert or not are:
  - **Gender**
  - **Customer_segment**

Thank you for your confidence!