Lenus eHealth: Customer segmentation challenge

Per Simmendefeldt Schmidt
Jagtvej 215A 3.1. 2100 København Ø
28149439, pers1@hotmail.com

30. oktober 2021

1 Task

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

Outline Solution will address this problem in the following steps:

1) Data investigation/clean up.

2) Only being able to use one factor for predicting the "converted"status, how do the factors rank in terms of accuracy score.

3) Two factors may hold the same information and appear equal in importance in 1) but might be redundant once more factors are included in fitting the predictive model. In that case, including top 3 factors in predicting "converted"might not be top 3 from 1).

The code to create the included figures, models and accuracy-calculations are found in LenusCaseTest.py.

2 Solution

2.1 Data investigation

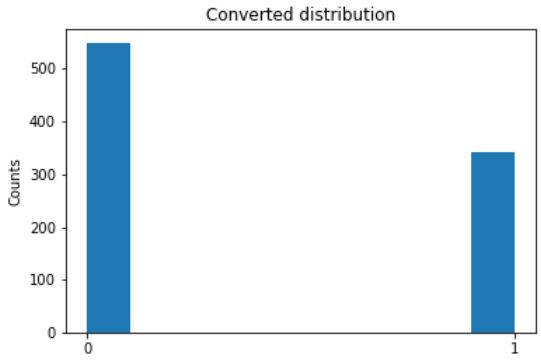
Data as per "Data_Scientist_-_Case_Dataset.xlsx":

Index	customer id	converted	customer segment	gender	age	related customers	family size	initial fee level	credit account id	branch
0	15001	0	13	male	22.0	1	0	14.5	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f…	Helsinki
1	15002	1	11	female	38.0	1	0	142.5666	afa2dc179e46e8456ffff9016f91396e9c6adf1fe20d1…	Tampere
2	15003	1	13	female	26.0	0	0	15.85	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f…	Helsinki
3	15004	1	11	female	35.0	1	0	106.2	abefcf257b5d2ff2816a68ec7c84ec8c11e0e0dc4f342…	Helsinki
4	15005	0	13	male	35.0	0	0	16.1	9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f…	Helsinki

Figur 1: Example rows.

Converted

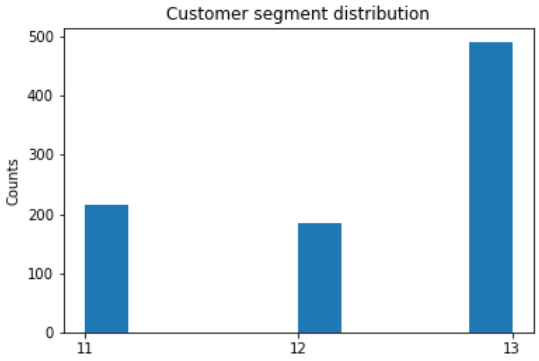
The target variable is slightly skewed towards more zeros. No null values. Note: A model predicting all zeros would be correct 61.6% on this entire dataset. Models should atleast beat this accuracy.



Figur 2: Converted distribution.

Customer segment

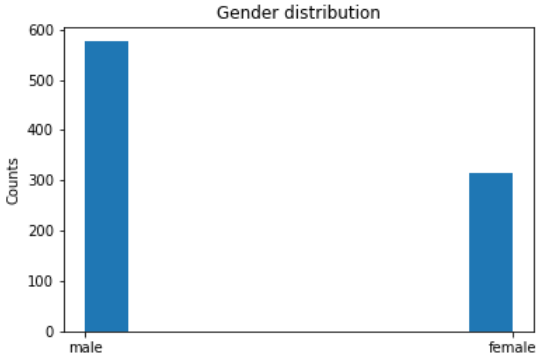
From Fig. 3, customer segment is seen to come in 3 categories (11, 12, 13). No null values found. These are determined to be categories and not numbers as such - they could be named category_11, category_12 and category_13. In order to do logistic regression later, this categorical feature is split into 3 separate columns in the final dataset, containing zeros and ones.



Figur 3: Customer segment distribution.

Gender

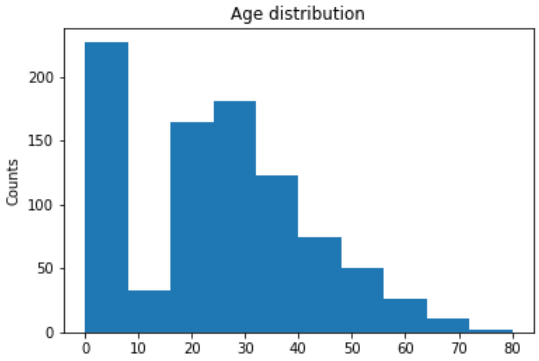
Gender is another categorical feature, no null values found. The data is skewed towards more males.



Figur 4: Gender distribution.

Age

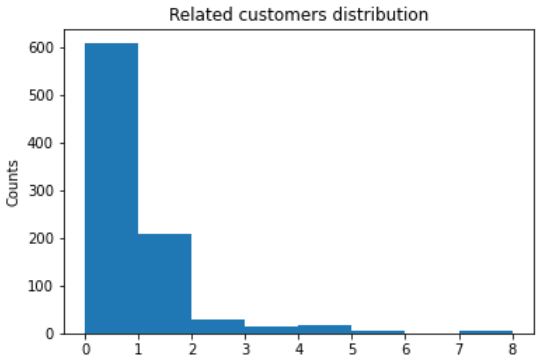
Age comes as an integer. Empty rows replaced with 0 in Fig. 5. 177 rows are missing values for age.



Figur 5: Age distribution.

Related customers

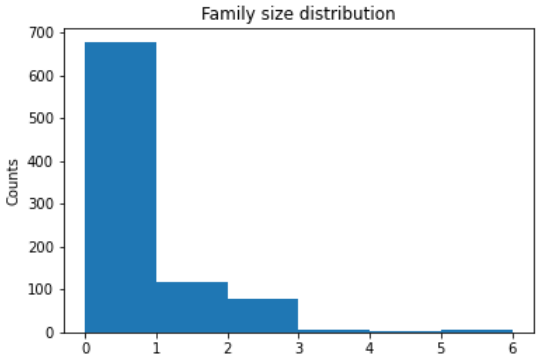
Integer column. No null values. Distribution ranges from 0 to 8. The high values could possibly be classified as outliers and removed from the dataset. This is not done here.



Figur 6: Related customers distribution.

Family size

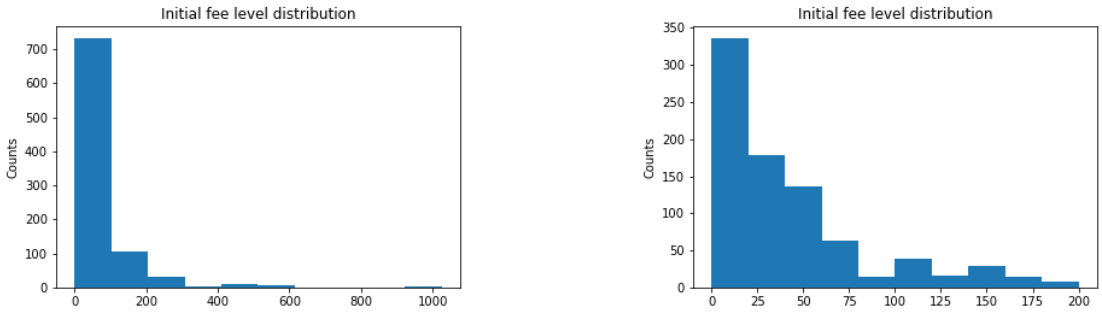
Integer column. No null values. Distribution ranges from 0 to 6.



Figur 7: Family size distribution.

Initial fee level

Float column. No null values. Distribution ranges from 0 to 1025.

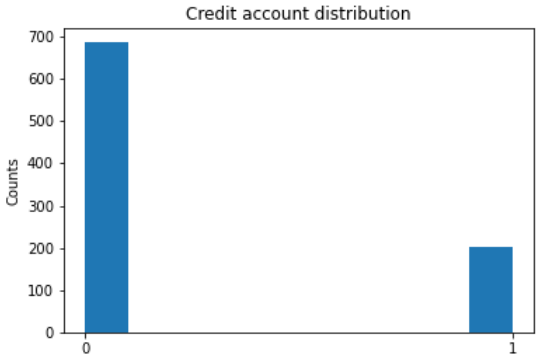


(a) (b)

Figur 8: Initial fee level distributions.

Credit account ID

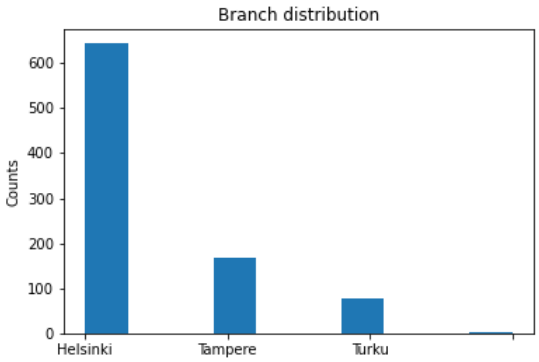
The individual credit account id hash is unique to the customer and as such not relevant but there might be value in creating a new column simple stating if the customer has a credit account id or not (shown with hash "9b2d5b4678..."). Fig. 9 shows that most customers do not have a credit account id.



Figur 9: Credit account distribution. 0 = no credit account id.

Branch

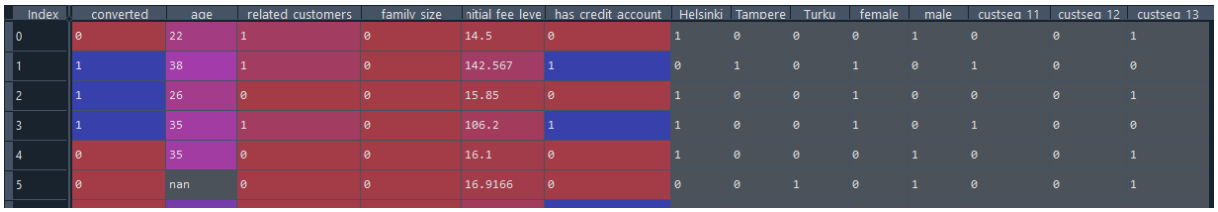
Categorical feature, 3 entries: Helsinki, Tampere, Turku and one empty row.



Figur 10: Branch distribution.

2.2 Ranking factors for predicting "converted"

With the considerations described above, the dataset now has the columns shown in Fig. 11.



Index	converted	age	related customers	family size	nitial fee leve	has credit account	Helsinki	Tampere	Turku	female	male	custseg 11	custseg 12	custseg 13
0	0	22	1	0	14.5	0	1	0	0	0	1	0	0	1
1	1	38	1	0	142.567	1	0	1	0	1	0	1	0	0
2	1	26	0	0	15.85	0	1	0	0	1	0	0	0	1
3	1	35	1	0	106.2	1	1	0	0	1	0	1	0	0
4	0	35	0	0	16.1	0	1	0	0	0	1	0	0	1
5	0	nan	0	0	16.9166	0	0	0	1	0	1	0	0	1

Figur 11: Example rows post clean-up.

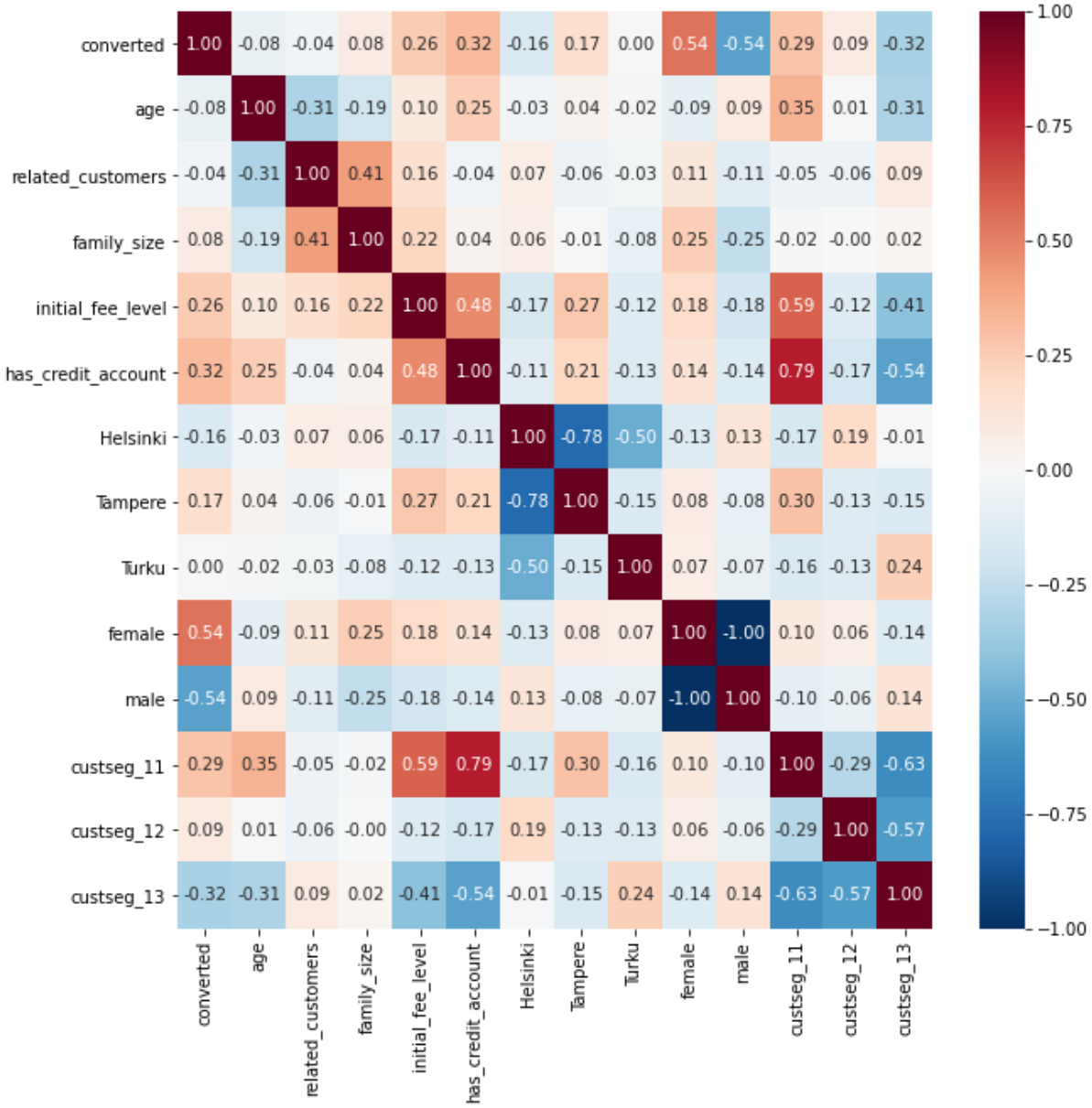
How well a factor can predict the "converted"column depends on how well the two variables are *correlated*. This is visualized in the correlation matrix in Fig. 12. Rows with nulls (age and branch category) are excluded from the calculation. From the first column of the correlation matrix, the ranking of variables to predict "converted"would be:

1. male/female
2. has_credit_account
3. customer_segment
4. initial_fee_level
5. branch
6. age
7. family_size
8. related_customers

It is worth noting that female/male columns are of course fully correlated (you are either one or the other) and one of these columns will be omitted. has_credit_account seems to be somewhat correlated with the customer_segment and it is as such expected that these variables provide most of the same information when fitting a model.

2.3 Simple logistic regression including multiple factors

In the following section, the predictive power of each variable on the "converted"column is tested using a simple logistic regression model. For this, all rows without an "age"value has to be dealt with. For simplicity, they are simply removed here. First however, a "recursive feature elimination"is performed on a transformed dataset (scaled for each variable to have zero mean and unit variance for better performance). This elimination procedure works



Figur 12: Correlation matrix.

by recursively considering smaller and smaller sets of features. Initially the estimator is trained on the entire set of features and the importance of each is obtained. Then the least important feature is removed from the set and this process is recursively continued until the desired number of features is left (here set to 1). The resulting importance ranking is found to be:

1. male/female
2. customer_segment

3. age
4. related_customers
5. has_credit_account
6. branch
7. initial_fee_level
8. family_size

We see that the top 2 spots are in agreement with the correlation matrix analysis. has_credit_account is moved further down the list, most likely because of its correlation with the customer_segment. Age is surprisingly high up the list, but lets check the individual features to see how much is actually gained in a model once we move past the top 2 features.

The dataset is split into a training and test set (75/25 split) and a logistic regression model is created using only 1 variable at a time, trying to predict "converted". The resulting accuracy score (percentage of correct classifications) are listed in Table 1. As expected, this order matches the correlation matrix order because we are fitting a new model for each variable.

Variable	**Accuracy**
male/female	0.7598
has_credit_account	0.6927
customer_segment	0.6760
branch	0.6480
initial_fee_level	0.6480
age	0.6201
related_customers	0.5978
family_size	0.5978

Tabel 1: Individual accuracy score for each variable trying to predict "converted".

In the next step, multiple features are included in fitting of the model. Accuracy scores are shown in Table 2. The best accuracy is found when including the features male/female, customer_segment, has_credit_account and age.

Variables	**Accuracy**
male/female	0.7598
male/female, customer_segment	0.7598
male/female, has_credit_account	0.7598
male/female, has_credit_account, customer_segment	0.7598
male/female, customer_segment, age	0.7654
male/female, customer_segment, has_credit_account, age	0.7765
all	0.7765
customer_segment, has_credit_account, age	0.7374

Tabel 2: Accuracy score for different combinations of variables.

3 Conclusion

On an individual feature level, the variables gender, has_credit_account and customer_segment has the most value. They do however seem to share most of the same information in relation to "converted", hence no improvement in accuracy is seen when including all three rather than gender alone. Age improves the accuracy slightly and all four variables brings the accuracy to the best achieved here. Without gender, the accuracy drops significantly.

Further work could look into which rows are being predicted correctly by the various variables. The same accuracy might be the result of the same number, but different rows, being classified correctly. The dataset could be cleaned for outliers in the various categories. As the age variable seemed to bring information, there could be value in trying to fill out the rows with an empty age.

The accuracy score can be misleading for skewed datasets. The resulting models here do however beat the 61.6% mentioned under the "converted"subsection of Section 2.1. The confusion matrix of the model including gender, has_credit_account, customer_segment and age is printed below. It shows that both zeros and ones are predicted correctly so the slightly skewed dataset towards converted = 0 does not seem to be an issue. There seems to be an even distribution of false positives and false negatives that could be investigated further in detail.

86	21
19	53

Tabel 3: Confusion matrix of best performing model.