Technical Report

Contents

The case

The project's main scope is to find KBC's clients that would be targeted by marketing offers to buy one of the following products: Consumer Loans, Credit Cards, or Mutual Funds. The aim is to find up to 100 clients to maximize the expected revenue incoming from bought products after the proposed marketing offer. For this purpose, an analytical approach with the deployment of predictive modelling is required. To sum up, the main goal is to maximize the efficiency of marketing offers and so to maximize potential expected revenue.

Data

Four datasets containing information about clients are provided. Datasets are combinations of data about clients' products, inflows, and outflows related to their products, socio-demographics data and last dataset provides info about sales and revenues.

All datasets contain a common variable that was used as a key indicator while merging those datasets together into one piece – 'Client', which is an indicator of the client's ID in the bank's systems.

Client IDs have no duplicates in any of provided datasets so each row presents unique data about an individual client.

There are **no severe missing data in datasets**. The only dataset that contains a significant amount of missing information is a dataset about products, however, these missings are caused by the absence of given products in the case of some clients, so they can be confidently replaced by 0, as it means the same information is provided. Variable 'Sex' has 3 missing values, instead of deleting, those are replaced by mode.

However, there were new missing data introduced after the merge of datasets due to their different lengths. 28 missings for the variables about inflows and outflows and 646 missings for the dataset about sales and revenues. Missing data related to I/O were replaced by mean values of given variables to avoid deleting and as the number of observations to fill by means is very small there is no need to be worried about the disruption of the original distributions. Missing information about sales and revenues does not need to be treated as these observations are not used for purposes of model training and are held out as data to make the final predictions about marketing offers.

Distributions of data are checked via Histograms and Boxplots. Data about inflows and outflows are mostly right-skewed, Tenure is very concentrated at around 150 months, and Boxplots showed there are a few unusual observations that could be theoretically outliers (such as Actual Balance CA 171k vs 75^{th} percentile 2174), however, most of these observations seem realistic and does not represent suspicious wrong recorded data so they are kept in the dataset as they may still represent important information.

However, there were several **observations with ages under 18 years**. These observations are **excluded** from the dataset as there is a high chance those are wrong data and additionally products of our interest cannot even be held by u18 person (especially CL and CC). Additionally, one more control of data quality is performed, when **age in month and tenure (also months) is compared**. If the **difference is negative**, observations are also **excluded** from dataset as we would suppose those observations can be fishy (how to have longer tenure with a bank than age, if tenure is meant to have a product's relationship with a bank).

One variable is additionally created and added, representing the **difference between average credit and debit volume turnovers**. It is clear, that clients that went for a Consumer loan or Credit card have negative average balances of inflows in comparison to other clients. Also, clients targeted with MF have significantly higher average positive inflow than others. Both situations make sense – people tend to loan money when they are lacking them, while people tend to invest when they have sparse resources.

Categorical data (only 'Sex') were one-hot-encoded.

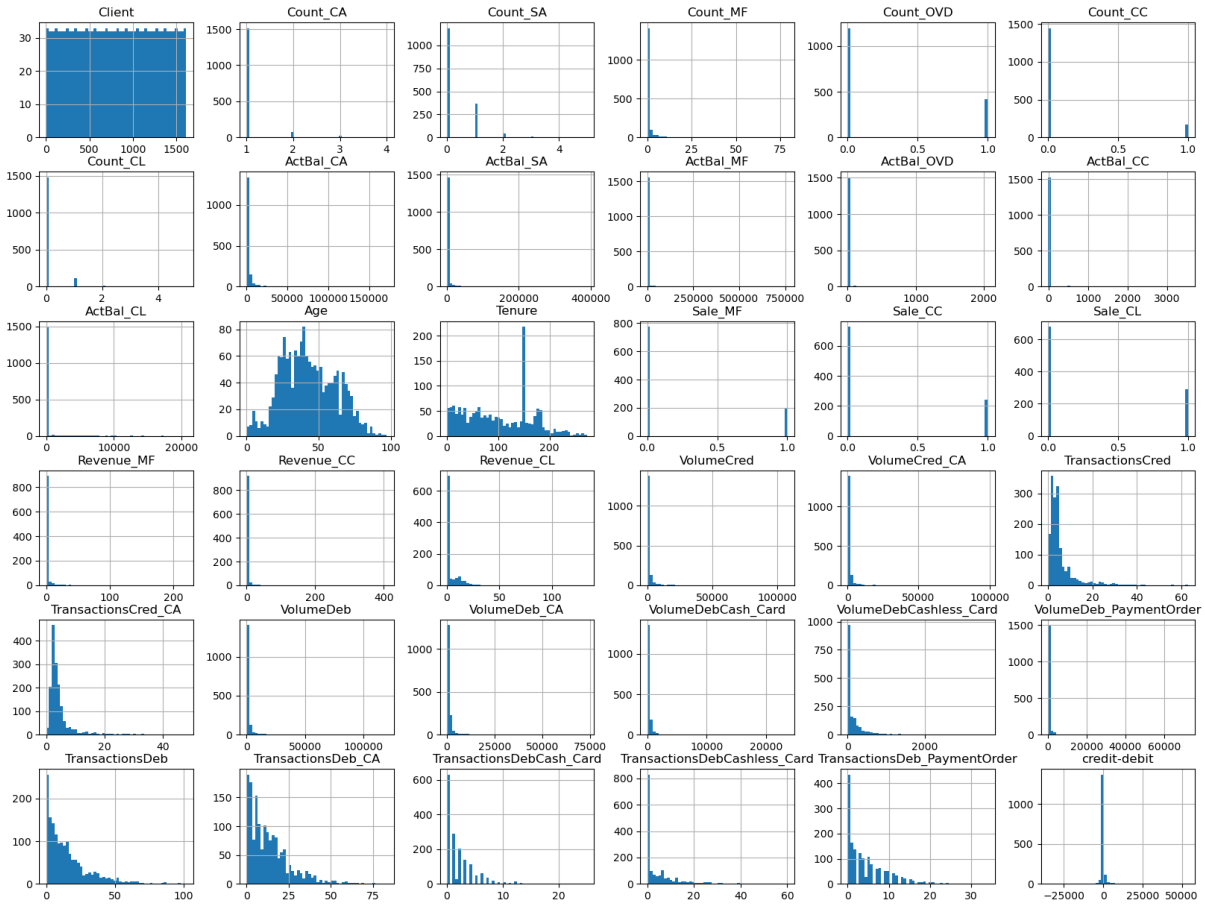
Average differences in credit – debit turnovers:

	Sale_CL	Sale_CC	Sale_MF
0	307	245	75
1	-261	-139	435

Classes (sale/no-sale) are slightly imbalanced in favour of no-sale class:

	Sale_CL	Sale_CC	Sale_MF
0	70%	75%	80%
1	30%	25%	20%

Distributions



More details, including descriptive statistics and more in-depth EDA: data_preparation.py

Modeling:

Training/validation/test:

Initially, only observations with available records about sales and revenues are included in the training dataset (907 obs). 606 records are held out for the final predictions of targeted clients.

Out of 907 observations in the training set, 30 and 15 percent for classification and regression respectively, are excluded from the training dataset to be able to measure the performance of the created model on unseen data.

Data for the classification task are split via Stratified sampling to identical distributions of classification classes in both training and validation sets. For regression, a random selection split is performed.

Models:

There are 6 predictive models built in total. Each product (CC, CL, MF) has two separate models. One for classification problems and obtaining probabilities of no/sale, another one to obtain predictions about possible revenue.

Predicting no/sale:

To predict targeted clients that would buy a product after a marketing offer, Logistic Regression was deployed. The rationale for using LR is that it tends to return **well-calibrated probabilities by default as it directly optimizes Log loss**. Additionally, the algorithm is not data-hungry (relatively small dataset) and predictions can be easily explained and interpreted via coefficients (white-box). Getting valid probabilities is in our case crucial as they are important for calculating expected revenue (p * R) and so for choosing clients to target. To obtain valid probabilities, artificial editing of imbalances (penalization weights towards the majority, up/down sampling…) is avoided, so we do not need to deal with additional calibration as models are built on data with original class distributions.

For each logistic regression model, several model evaluations are calculated, such as ROC AUC, Precision, Sensitivity, F-1 score and to assess the accuracy of forecasted probabilities, Brier Score. As our dataset is imbalanced, there is no point to look at accuracy because it is a positively biased score towards the majority class. To deal with imbalances and avoid discrimination of positive class, which is the class of our main interest, different cut-off values for classifying 0/1 were estimated (range 0.25-0.50) and cut-off selected based on the ideal combination of precision and sensitivity, as we care about potential sales that should be captured (sensitivity) but also possibilities of the bank to contact clients are limited so we would like to have positive clients captured precisely without mess lying in the prediction basket.

For the case of each product, predicted clients that are selected for the final round of selection (explained later), are considered individually based on cut-off value and logic. Firstly**, it would not make sense to count for clients with very low probabilities in favour of sale as it means wasted resources because convincing them is unlikely. Secondly, the same may apply to clients with high probabilities (above 0.9) – these clients are already convinced to buy a product and would likely do it even without a marketing offer**. That means there should be a focus on clients that are convincible by marketing offers. The lower probability interval for the client's selection is the selected cut-off (explained paragraph above) and the upper one is always 0.9.

Models' Evaluation:

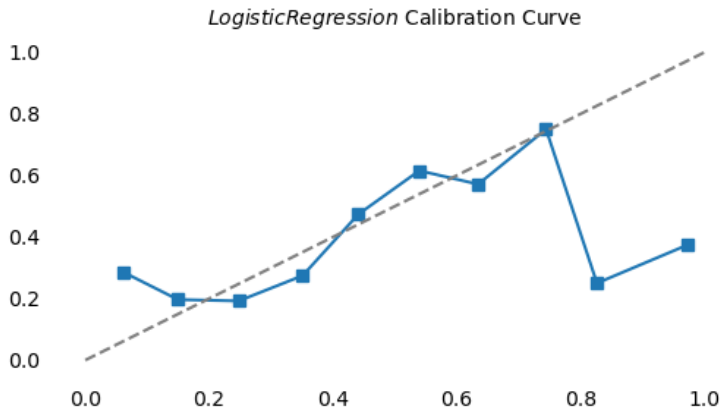
	CL	MF	CC
ROC_AUC	0.58	0.56	0.52
Brier Score	0.27	0.19	0.27
Precision (0.5 cut-off)	0.53	0.16	0.38
Sensitivity (0.5 cut-off)	0.25	0.56	0.13
F1 Score (0.5 cut-off)	0.34	0.25	0.19
Precision (0.3 cut-off)	0.39	0.44	0.29
Sensitivity (0.3 cut-off)	0.54	0.3	0.42
F1 Score (0.3 cut-off)	0.45	0.36	0.35

CL

The most influential predictor in favour of the sale of CL is the number of live current accounts, when with each increase by one account odds in favour of CL sale increase by 55 pcts. Followed by numbers of credit cards, savings accounts, and overdrafts.

features	coef	log-odds delta
Count_CA	[0.43]	55.04866
Count_CC	[0.18]	19.84737
Count_SA	[0.16]	18.39341
Count_OVD	[0.12]	12.8293

Probabilities are well calibrated except for the high ones, however, the amount of probabilities in intervals above 0.8 is really small so it can be a biased view due to lack of data.




```
probabilities_CL
(0.0, 0.1]      74
(0.1, 0.2]     137
(0.2, 0.3]     150
(0.3, 0.4]     112
(0.4, 0.5]      70
(0.5, 0.6]      28
(0.6, 0.7]      14
```

```
(0.7, 0.8]       6
(0.8, 0.9]       5
(0.9, 1.0]      10


cutoff 0.3 precision: 0.39 sensitivity: 0.54 f1: 0.45
```

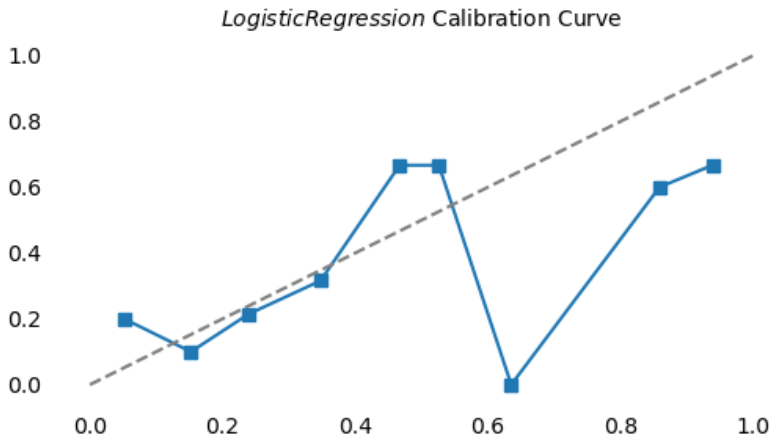
Selected cut-off value and so lower interval for selecting clients = 0.3

Number of selected clients for the final selection procedure = 235

MF

The most influential predictor in favour of the sale of MF is the number of credit transactions on current accounts when with each increase by one unit, odds in favour of ML sale increase by 15 pcts. Followed by the monthly number of debit cashless transactions via card and the number of live mutual funds.

features	coef	log-odds delta
TransactionsCred_CA	[0.14]	15.74
TransactionsDebCashless_Card	[0.12]	13.71
Count_MF	[0.11]	12.21




```
probabilities_MF
(0.0, 0.1]     123
(0.1, 0.2]     237
(0.2, 0.3]     150
(0.3, 0.4]      46
(0.4, 0.5]      20
(0.5, 0.6]       7
(0.6, 0.7]       8
(0.7, 0.8]       3
(0.8, 0.9]       6
(0.9, 1.0]       6


cutoff 0.3 precision: 0.44 sensitivity: 0.30 f1: 0.36
```

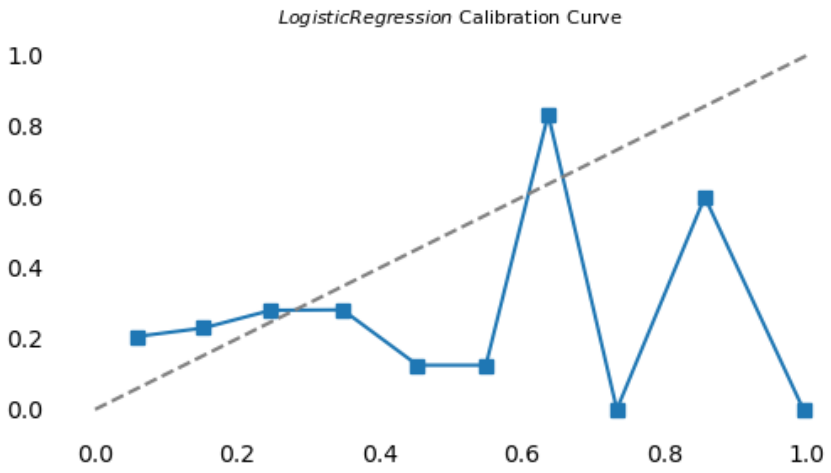
Selected cut-off value and so lower interval for selecting clients = 0.3

Number of selected clients for the final selection procedure = 90

CC

The most influential predictor in favour of the sale of CC is the number of live overdrafts, when with each increase by one unit, odds in favour of CC sale increase by 47 pcts. Followed by the monthly number of debit cash transactions via card and the number of all credit transactions.

features	coef	log-odds delta
Count_OVD	[0.38]	47.58
TransactionsDebCash_Card	[0.16]	18.17
TransactionsCred	[0.15]	16.28




```
probabilities_CC
(0.0, 0.1]     88
(0.1, 0.2]    233
(0.2, 0.3]    146
(0.3, 0.4]     56
(0.4, 0.5]     27
(0.5, 0.6]     17
(0.6, 0.7]     10
(0.7, 0.8]      7
(0.8, 0.9]      7
(0.9, 1.0]     15
```

```
cutoff 0.25 precision: 0.29 sensitivity: 0.42 f1: 0.35
```

Selected cut-off value and so lower interval for selecting clients = 0.25

Number of selected clients for the final selection procedure = 181

Predicting Revenues

To obtain some information about possible revenues coming from clients, regression models for each product are developed. It is worth noting that these calculations can be probably much more accurate in the case of banks' environments as they are aware of their fees and interest rates and so they might be able to state revenue coming from their operation as fixed without dependency on predictions with errors.

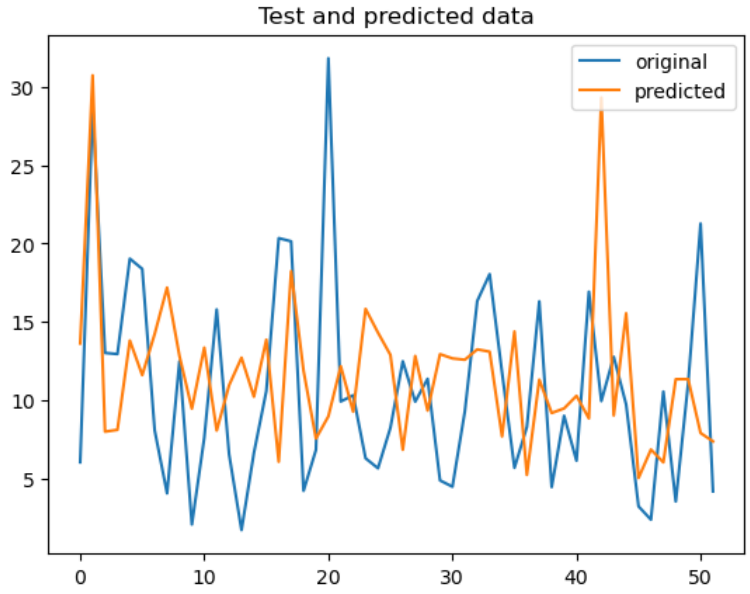
For purposes of this modelling, only positive revenues are selected. The reason is that revenues are highly skewed on zeros, and we are interested in clients generating revenues, so it would make more sense to consider only positive revenues.

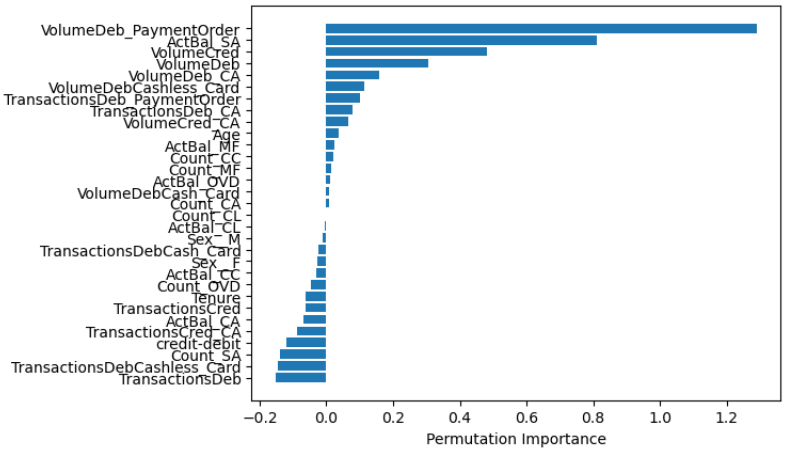
As the only interest is to obtain accurate revenue predictions without the need for explainability and interpretations, it was decided to deploy XGB Regression. This allows skipping treating assumptions of some linear models, tends to be accurate and works well also on smaller datasets. Additionally, despite the complexity of the algorithm, it is computationally not extensive. To find optimal tuning parameters for XGB, Randomized Search with Cross-Validation and scoring criteria RMSE was used. To see the importance of individual variables for accurate predictions, Permutation Importance was deployed and displayed via bar plots.

Errors:

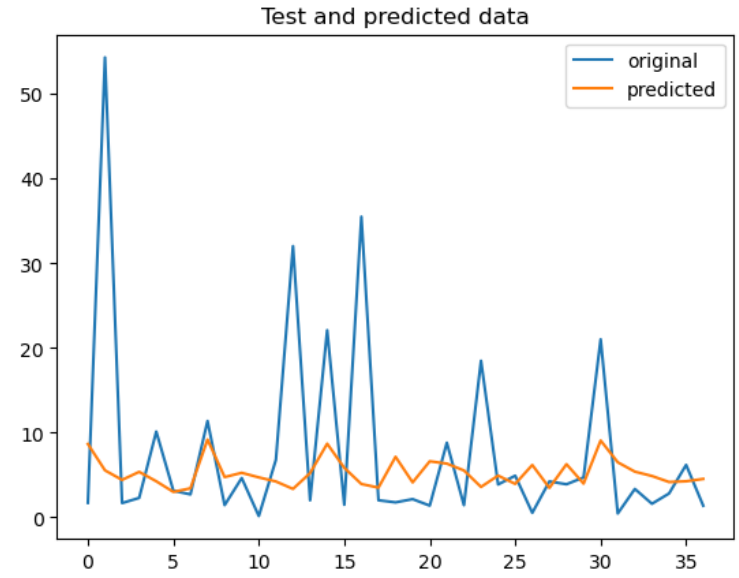
	CL	MF	CC
MAE	5.9	6.5	7.6
RMSE	7.4	11.7	15

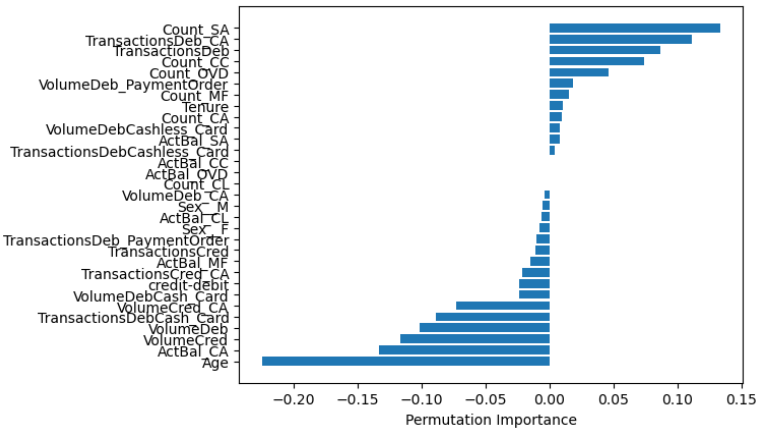
CL



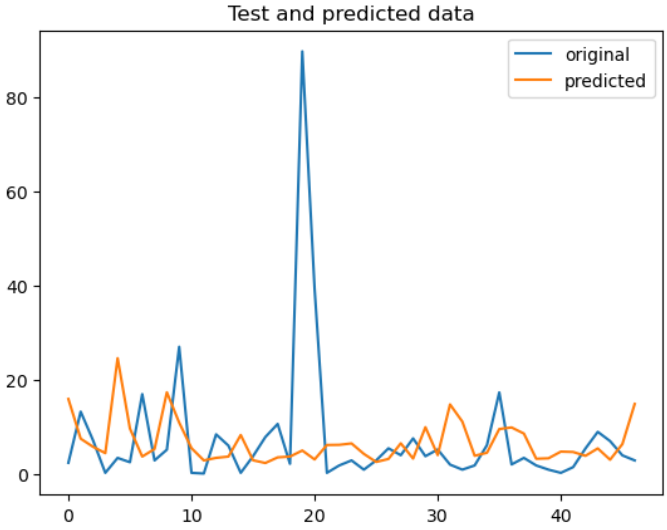

VolumeDeb_PaymentOrder
ActBal_SA
VolumeCred
VolumeDeb
VolumeDeb_CA
VolumeDebCashless_Card
TransactionsDeb_PaymentOrder
TransactionsDeb_CA
VolumeCred_CA
Age
ActBal_MF
Count_CC
Count_MF
ActBal_OVD
VolumeDebCash_Card
Count_CA
Count_CL
ActBal_CL
Sex_M
TransactionsDebCash_Card
Sex_F
ActBal_CC
Count_OVD
Tenure
TransactionsCred
ActBal_CA
TransactionsCred_CA
credit-debit
Count_SA
TransactionsDebCashless_Card
TransactionsDeb
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Permutation Importance

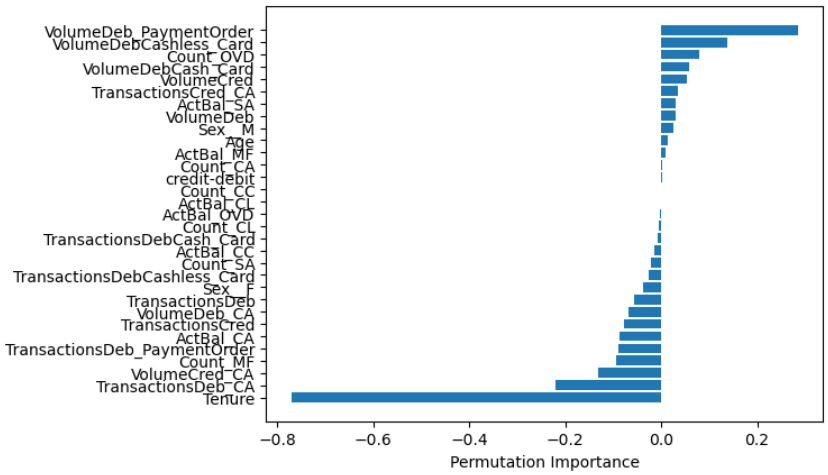
MF


Test and predicted data
original
predicted
0
10
20
30
40
50
0
5
10
15
20
25
30
35


Count_SA
TransactionsDeb_CA
TransactionsDeb
Count_CC
Count_OVD
VolumeDeb_PaymentOrder
Count_MF
Tenure
Count_CA
VolumeDebCashless_Card
ActBal_SA
TransactionsDebCashless_Card
ActBal_CC
ActBal_OVD
Count_CL
VolumeDeb_CA
Sex_M
ActBal_CL
Sex_F
TransactionsDeb_PaymentOrder
TransactionsCred
ActBal_MF
TransactionsCred_CA
credit-debit
VolumeDebCash_Card
VolumeCred_CA
TransactionsDebCash_Card
VolumeDeb
VolumeCred
ActBal_CA
Age
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Permutation Importance

CC





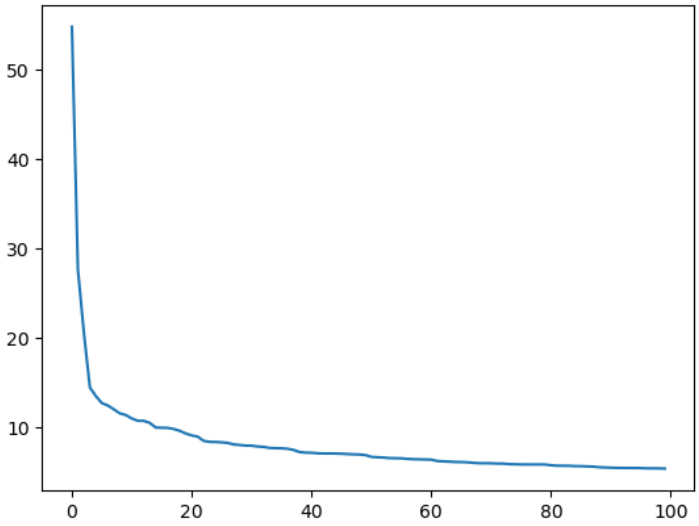
Selecting TOP 100 clients to target

The final procedure is to take selected clients from each part of the process (clients for CC, CL, MF) and to find TOP 100 clients to target with marketing offer, find what product to offer to propose and maximize expected revenue.

For each client, **expected revenue is calculated as ER = p * pE**, where p = probability of sale of given product obtained via LR model and pE = predicted revenue obtained via XGB model. We calculate for expected revenue with probabilities, as we must count for costs of no-sale possibility which occurs into some extent in each case. If we were sure about the sale, ER would be equal to predicted revenue, but it is not the case.

Then, the ideal product for the marketing offer is indicated as the one that yields the highest ER and 100 clients with the highest ERs are selected for the final outcome which is a list of clients to target.

Expected Revenues per individual client – sorted from the highest to lowest for top 100 clients



The total ER is 817 EUR and the highest share out of all offers are shares of Consumer Loans, followed by offers of CCs and the last minor offers are MFs. That means that the marketing offers should mainly focus on consumer loans.

Total ER (EUR)	**817**
Share of CL offers	82%
Share of CC offers	13%
Share of MF offers	5%

Conclusion

There are six models created for three product categories. The aim is to predict clients to be targeted via marketing offer as they would be likely to buy a product. For this, two types of models have to be created. Classification models that would predict who to target and what's the probability of sale and regression model to predict revenues coming from a given client. Combination of these two factors – predicted probabilities and revenues, there is the possibility to calculate expected revenue for each client. Expected revenue states for revenue after adjustment for risk costs of no sale – the higher probability of sale the closer ER is to the actual revenue.

Based on our models and calculation, if selected 100 clients would be targeted by marketing offers it yield ER of 817 EUR. The main focus should be on CL as they count for the majority of offers with high ER.

Limitation

The assignment is mostly an illustrative assignment on how to approach problems. There is probably a chance to get better performing models, for example, if there were clusters created first and then models for individual clusters.