Propensity models for KBC test data

D. Lontkovskyi

February 4, 2020

Technical report with the outline of the task solution.

1. Introduction.

This document describes the proposed solution for the optimization of the marketing campaign. In Section 2, mathematical formulation for the estimation of the expected revenue of the campaign is given. Sec. 3 outlines basic information about input dataset. Sections 4, 5 contain the details of the multivariate analysis.

2. Mathematical model for the expected revenue.

To calculate the expected revenue of the direct marketing campaign, the following simplified ansatz is used:

$$\mathrm{E}\left[R_a\right] = \int d\mathbf{x}\, R_a\left(\mathbf{x}|\text{respond}\right)\pi_a\left(\text{respond}|\mathbf{x}\right) + R_a(\mathbf{x}|\text{not respond})\left(1 - \pi_a\left(\text{respond}|\mathbf{x}\right)\right), \tag{1}$$

where $R_a\left(\mathbf{x}|\text{respond}\right)$ ($R_a\left(\mathbf{x}|\text{not respond}\right)$) is the revenue from the client characterized by the feature vector $\mathbf{x}$ that responds (does not respond) to the marketing offer; the subscript index a corresponds to the sort of marketing offer (mutual fund (MF), consumer loan (CL), credit card (CC)); $\pi\left(\text{respond}|\mathbf{x}\right)$ is the conditional probability that this client will respond to the advertisement, given she has the feature vector $\mathbf{x}$.

Assuming that clients that do not wish to respond to the offer do not generate revenue,

$$R_a(\mathbf{x}|\text{not respond}) \equiv 0, \tag{2}$$

the equation simplifies to

$$\mathrm{E}\left[R_a\right] = \int d\mathbf{x}\, R_a\left(\mathbf{x}|\text{respond}\right)\pi_a\left(\text{respond}|\mathbf{x}\right). \tag{3}$$

Since the prediction will be made using finite number of clients, the integral is approximated by the finite sum, where the index i runs over individual clients.

$$\mathrm{E}\left[R_a\right] = \sum_i R_a\left(\mathbf{x}_i|\text{respond}\right)\pi_a\left(\text{respond}|\mathbf{x}_i\right). \tag{4}$$

The conditional probability, $\pi\left(\text{respond}|\mathbf{x}\right)$, and the revenue function, $R_a\left(\mathbf{x}|\text{respond}\right)$, can be estimated from the data using machine learning techniques.

The total expected revenue of the marketing campaign is the sum of the expected revenues from individual classes

$$\mathrm{E}\left[R\right] = \sum_a \mathrm{E}\left[R_a\right] = \sum_a\sum_i R_a\left(\mathbf{x}_i|\text{respond}\right)\pi\left(\text{respond}|\mathbf{x}_i\right). \tag{5}$$

In order to optimize the revenue of the campaign, the clients that provide the largest contribution to the sum have to be identified. Following sections outline main steps to achieve this task.

3. Data.

The data were provided in the form of the .xslx spreadsheet with several tables. In order to facilitate further processing, the file was converted to .csv format and processed with dedicated scripts developed for this analysis. The scripts utilize standard for the field data processing and machine learning libraries ***pandas*** and ***scikit-learn***.

The input tables contained records corresponding to the clients that responded to different classes (Mutual fund (MF), Credit Card (CC), Consumer Loan (CL)) of marketing advertisement and the information about their profile, such as socio-demographic status and financial activity. For a fraction of clients, for which the information about the success of the targeting is not available, the probability of the positive outcome has to be predicted.

It is important to ensure that the data in two subsets is drawn from the same population. Otherwise, the developed propensity model will give biased results. The example plots (see Fig. 1) below demonstrate distributions of the parameters (components of the n-dimensional vectors $\mathbf{x}_i$) characterizing the clients. Remaining plots for all distributions can be found in the Appendix A. As can be seen from the plots, within statistical uncertainties of the datasets the data in the training and prediction subsets follow the same probability distributions.

Majority of the clients did not accept the offer or responded only to one type of the advertisement (see Tab. 1), therefore it was decided to label each client according to the type of the advertisement that was successful. These classes are mutually exclusive.

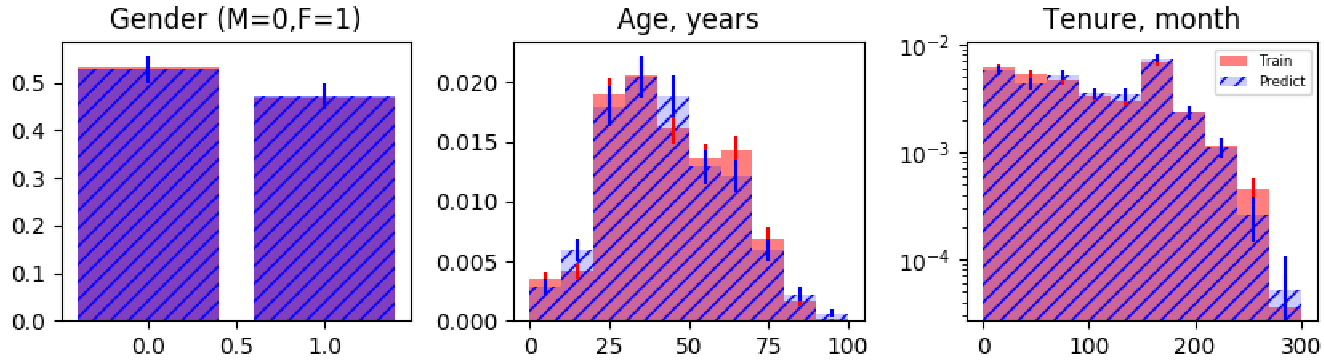

Figure 1: Normalized distributions of the client's feature variables for the subsets of the data with known (Train) and unknown (Predict) target variables. Last bin contains overflow entries.

Description	Class label	Number of clients in the training sample
Rejected	0	387
Buy only Mutual Fund (MF)	1	106
Buy only Credit Card (CC)	2	137
Buy only Consumer Loan (CL)	3	177
Buy two or more products	4	142

Table 1: Description of the clients class labels and the number of clients in each class.

4. Propensity model for different classes.

4.1. Classification algorithm.

Several multivariate techniques were tried in order to build the propensity model for different clients classes. Artificial Neural Network (ANN) was prone to overtraining, while Support Vector Machine (SVM) was too slow during the training and therefore difficult to optimize. These algorithms are not described in the report.

Eventually, I used a popular random forest algorithm implemented in ***scikit-learn***, to train the classifier for the class client class prediction. The classifier was trained using one-vs-rest approach.

4.2. Feature selection and feature importance.

Clients that rejected the marketing offer have characteristics very similar to the customers of different products. For example, the comparison of the feature distributions in two classes[1] is presented in Fig. 2. The clients that responded to the

[1] All distributions for all pairs of classes can be provided for inspection on request.

credit card (CC) marketing offer on average have lower balance on actual current account and have more live current accounts.

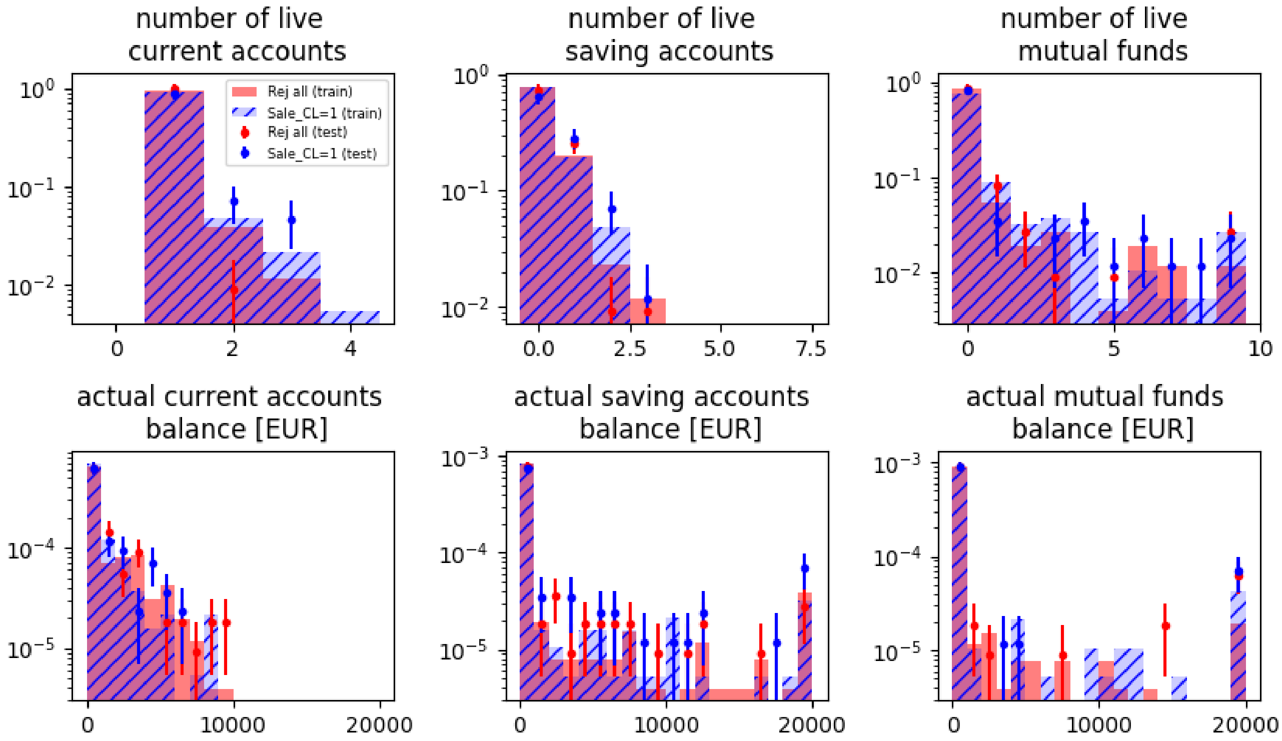

Figure 2: Comparison of the training and hold-out distributions for two different classes of clients.

In order to identify features that correlate with the propensity to buying certain class of products, the intrinsic impurity-based feature importance of the random forest algorithm and generic permutation importance measures were used. Fig. 3 demonstrates the comparison of two measures for different classes. The impurity-based feature importance of random forests suffers from being derived from statistics of the training dataset, therefore permutation importance can be used as a useful cross check. As can be seen from the plots, in general, two importance ranking algorithms share the same features among their top ranking variables. These include **Tenure, Age, current and savings accounts balance**. Selecting only subset of the features reduces the dimensionality of the problem and therefore makes the model more robust to overtraining. It was found that using only 5 top ranked variables was sufficient to achieve optimal performance.

4.3. Classifier parameters optimization.

In order to optimize the parameters of the random forest, grid search with cross validation was performed. Area under the Receiver Operating Characteristic (ROC) curve for multi-class classification was used as the optimization metric.

The validation plot is shown in Fig. 4. Variations of the parameters of the algorithm, such as **number of trees, maximum depth, minimum number of**

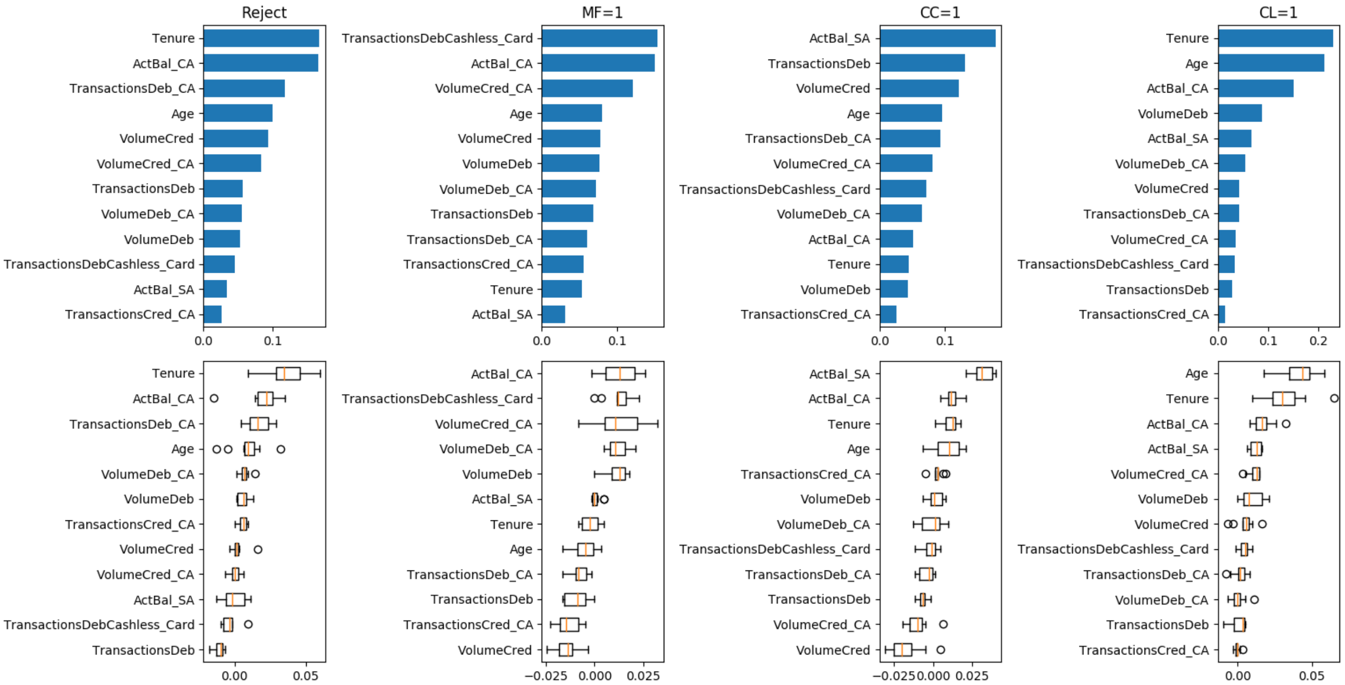

Figure 3: (Top) Impurity-based feature importance for different classes marketing offer outcome. (Bottom) Same for permutation importance. Only top 12 features are shown for each class.

samples in a leaf node and split criterion were investigated. It was found that the forests with higher complexity, e.g. with large number of trees or significant depth, exhibit larger generalization gap, while having approximately the same generalization error. Therefore, low complexity models were preferred. The final configuration of the classifier is outlined in Tab. 2.

In order to mitigate potential problems due to imbalanced population of different classes, intrinsic balancing mechanism of the random forest algorithm was enabled.

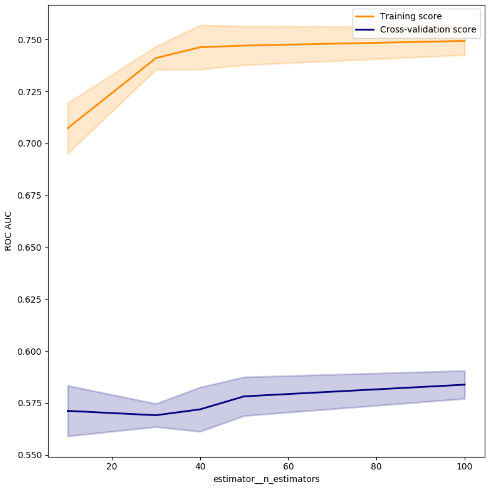

Figure 4: Validation plot for the optimization of the number of trees in the random forest.

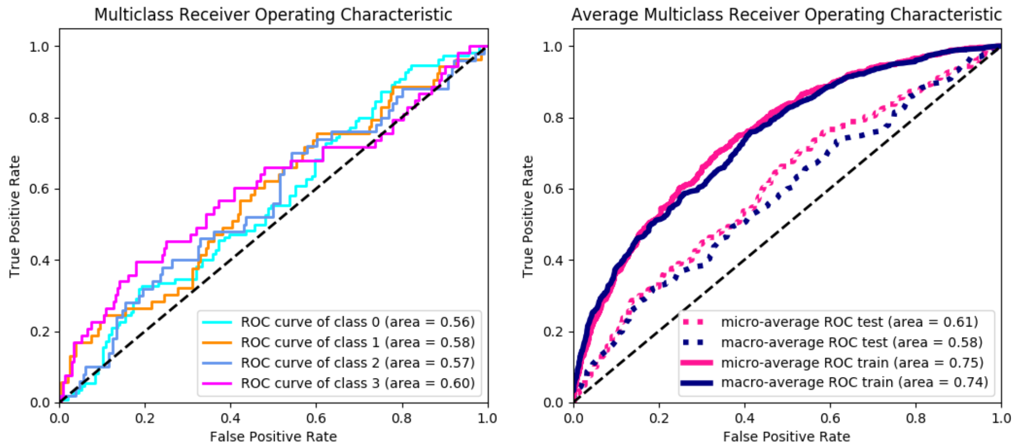

Figure 5: Receiver Operating Characteristic curves for the optimal classifier derived from the hold-out dataset. (Left) ROC curves for individual classes discriminated in one-vs-rest fashion. (Right) Micro- and macro- averaged ROC curves indicating overall classification performance for training and validation datasets.

The ROC curves of the optimal classifier for individual classes as well as micro- and macro-average ROC curves are presented in Fig. 5. **The optimal classifier performs approximately 10% better than random guessing.** Signs of overtraining are visible from the gap between ROC curves calculated using the training and hold-out datasets. Larger training dataset may help to improve the model classification accuracy.

Description	***scikit-learn*** parameter	Value
The number of trees in the forest	n_estimators	106
The minimum number of samples in a leaf node	min_samples_leaf	20
The maximum depth of the tree	max_depth	2
The function to measure the quality of a split	criterion	'gini'
Weights associated with classes	class_weight	'balanced'

Table 2: Configuration of the best discriminating classifier.

The trained random forest model can be used for the prediction of the client's propensity to buying different products.

5. Regression model for class revenue.

Regression function for the expected return from the client accepting a marketing offer is also determined using machine learning technique. For each class, except for the trivial case, when the offer was rejected, an independent regression model for the revenue is constructed using **xgboost** algorithm, which is a fast numerical implementation of the boosted decision tree regressor.

Fig. 6 shows distributions of the revenue generated by clients from different classes. The first bin at 0 corresponds to the clients, that rejected the offer. It can be seen that there are entries that fall outside of the bulk of the distributions. In order to not bias the regression model, the local outlier factor was computed for every entry in the histogram. The client records with the highest ranked score were removed and remaining sample was passed to the regression algorithm.

Early stopping was used in order to regularize models. The loss curves for the three classes can be found in the Appendix B.

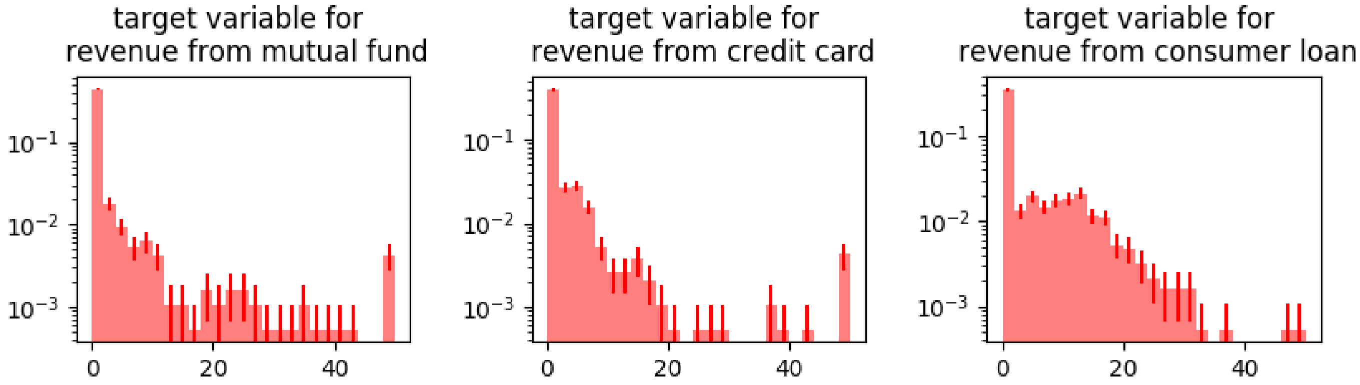

Figure 6: Normalized distributions of the client's revenue.

6. Summary.

An analytical model for the optimization of the targeting and revenue of the marketing campaign is proposed. The solution consists of the client propensity and the revenue regression models that can be used for the clients ranking. The propensity model prediction accuracy, as measured by the area under ROC curve, is approximately 10% better than random choice. The accuracy of the model may be improved using larger dataset.

The executive summary of the study, the file with the table summarizing, which offer has to be targeted to which client as well as the expected revenue for each client are provided separately.

A. Additional material

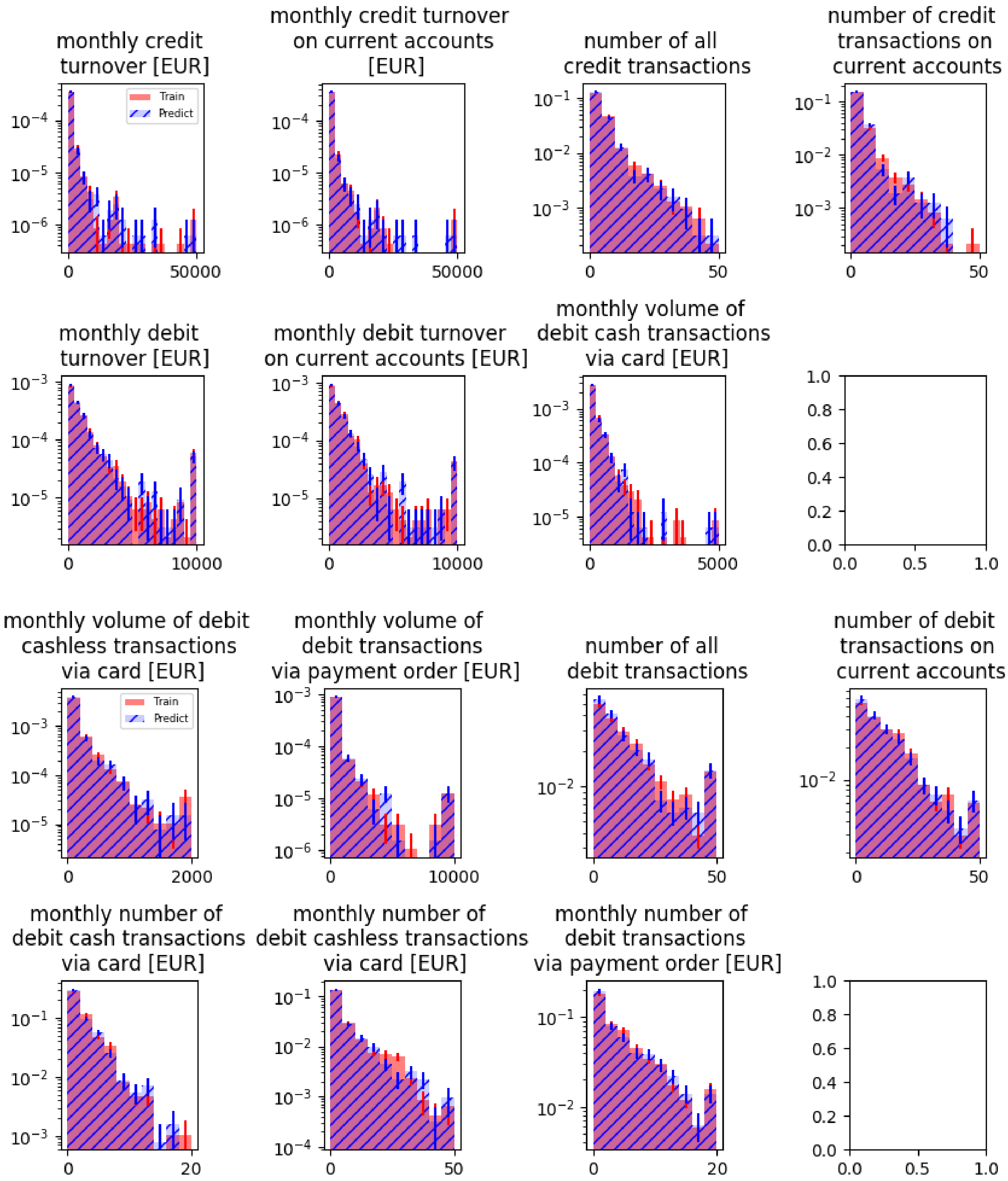

Figure 7: Normalized distributions of the client's feature variables for the subsets of the data with known (Train) and unknown (Predict) target variables. Last bin contains overflow entries.

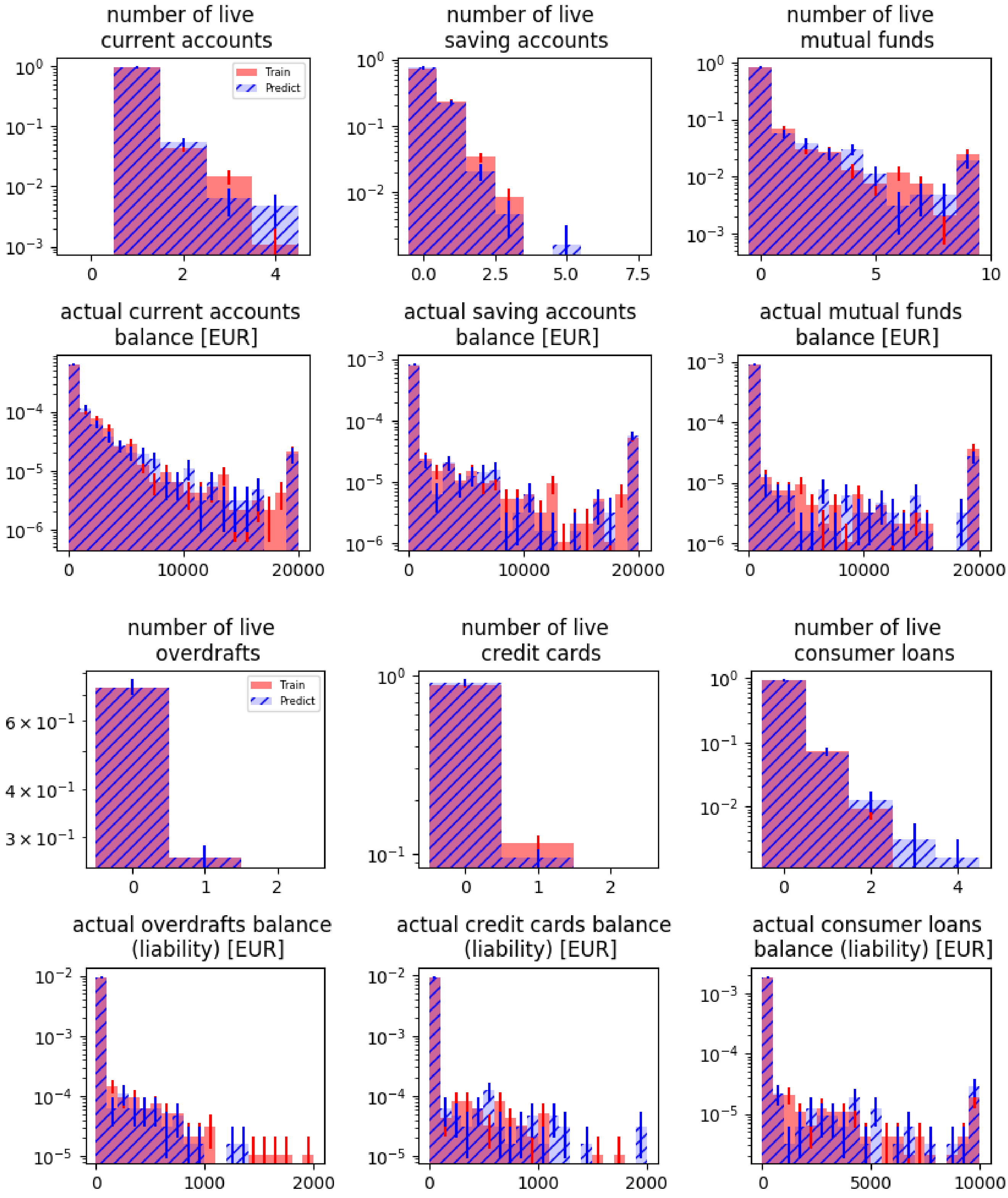

Figure 8: Normalized distributions of the client's feature variables for the subsets of the data with known (Train) and unknown (Predict) target variables. Last bin contains overflow entries.

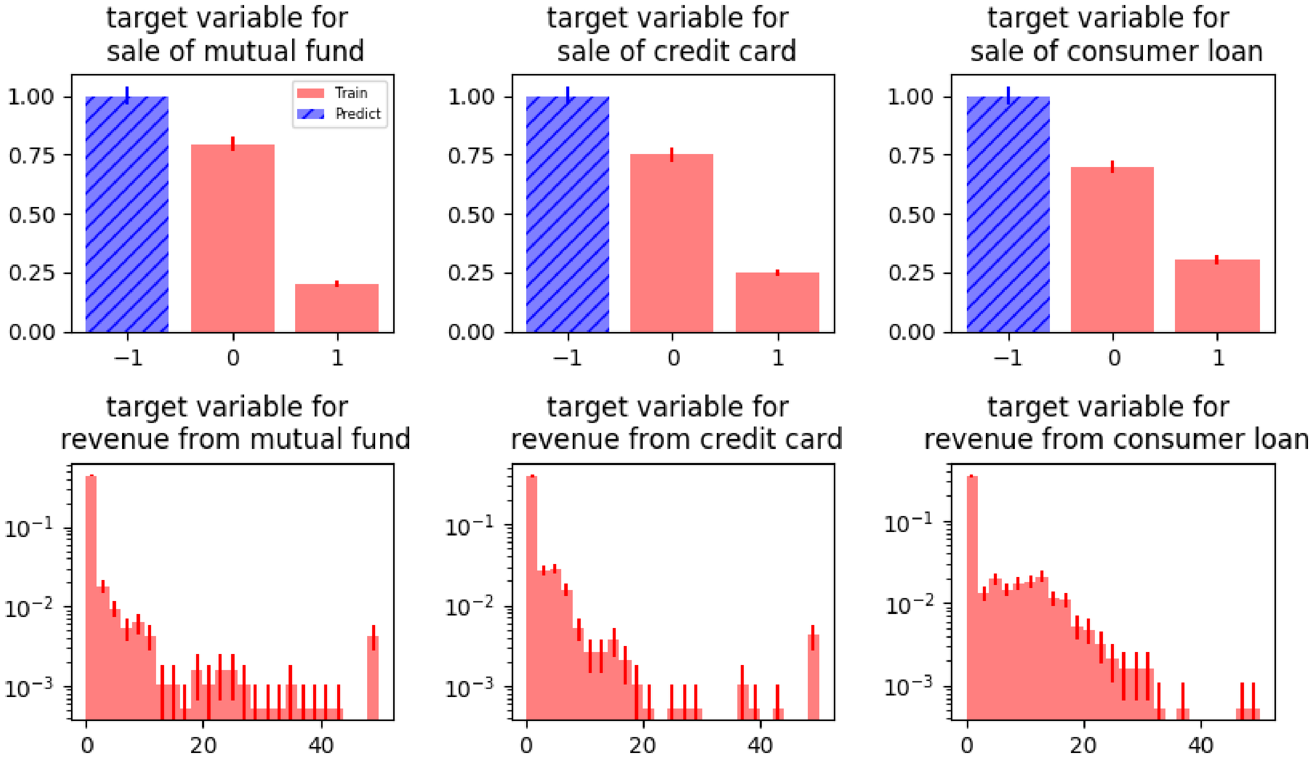

Figure 9: Normalized distributions of the client's target variables for the subsets of the data with known (Train) and unknown (Predict) target variables. Target variables for the samples for which the outcome must be predicted were set to -1. The revenue for this subset is unknown and therefore is not shown on the plot. Last bin contains overflow entries.

B. Regression validation curves

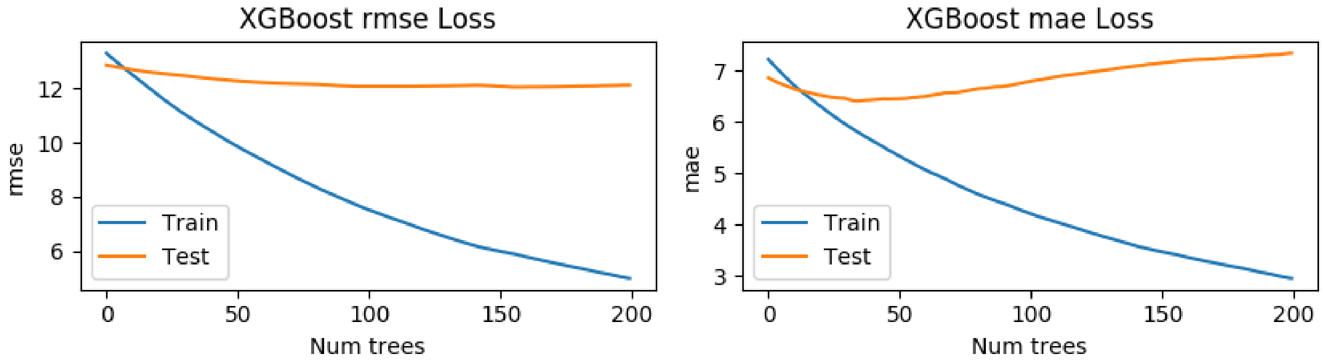

Figure 10: Loss curves showing mean squared error and mean absolute error for the MF class.

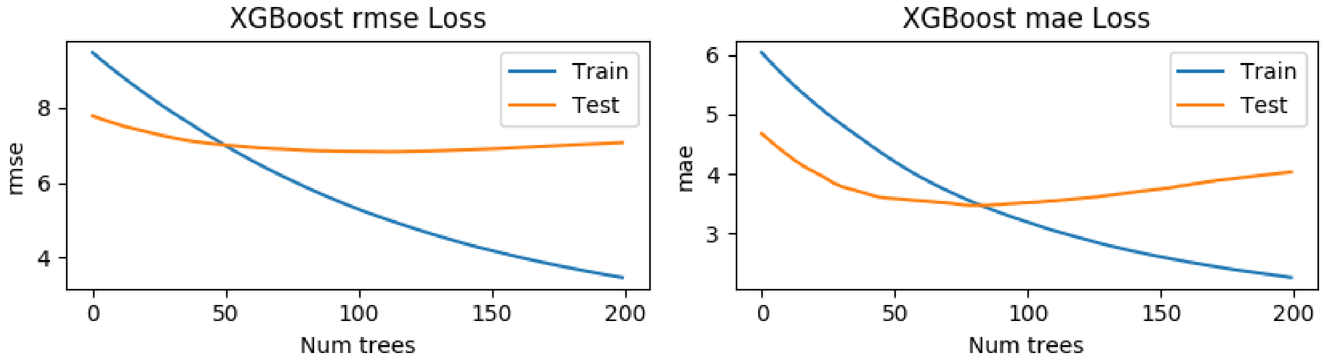

Figure 11: Loss curves showing mean squared error and mean absolute error for the CC class.

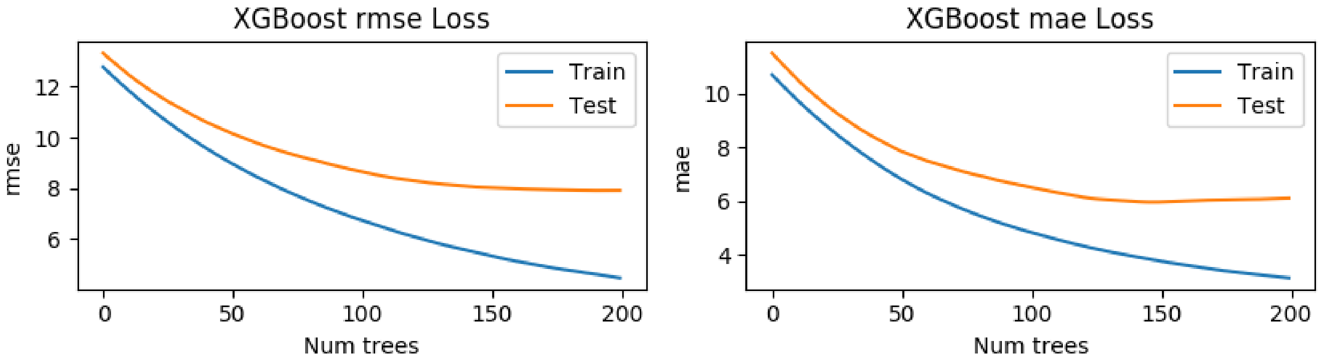

Figure 12: Loss curves showing mean squared error and mean absolute error for the CL class.