Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Allocating sales capacity to existing B2B customer accounts by margin at risk: a decision-date evaluation

Joona Mäntyvaara

joeman@utu.fi

University of Turku Faculty of Science and Engineering: Turun yliopiston Teknillinen tiedekunta
https://orcid.org/0009-0000-8764-9563

Jukka Heikkonen

University of Turku https://orcid.org/0000-0002-2468-5708

Research Article

Keywords: Sales resource allocation, Prescriptive analytics, Customer churn, Business-to-business, Account management, Temporal validation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816818/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Allocating sales capacity to existing B2B customer accounts by margin at risk: a decision-date evaluation

Joona Mäntyvaara[1*] and Jukka Heikkonen[1]

[1*]Department of Computing, University of Turku, Turku, Finland.

*Corresponding author(s). E-mail(s): joeman@utu.fi;
Contributing authors: jukhei@utu.fi;

Abstract

Business-to-business (B2B) firms must decide how to spread account managers' limited time over their existing customer accounts. A natural rule weights each account's margin by its risk of loss, but evidence on what the risk estimate adds to a plain margin ranking, tested as the decision is made, is scarce. We evaluate a prescriptive system that allocates a fixed appointment capacity in closed form by margin at risk, using records of 793 accounts of a European equipment-rental company. The decision is replayed at three year-ends using only information available on each date, and each decision is scored against the losses of the following year. Account size alone ranks next-year loss about as well as behavioral signals do. Once a year of losses is available, the system's list holds more of the accounts later lost than margin, size or RFM (recency, frequency, monetary value) rankings of the same size at the lower capacities, but not more of the lost margin, which follows account size. The leading risk model changed between years, an equal-weight combination of models stayed close to the best, and random cross-validation within a year overstated next-year accuracy in one of two years. An autoencoder representation added nothing over the raw features. A synthetic benchmark shows that a risk term protects more margin only above an AUC of about 0.62 to 0.69, and that ranking accuracy alone does not guarantee it. Whether the attention retains margin requires a randomized trial, for which the system can provide the assignment.

Keywords: Sales resource allocation, Prescriptive analytics, Customer churn, Business-to-business, Account management, Temporal validation

1 Introduction

On 27 October 2025, a sales director responsible for the existing customers of a European business-to-business equipment-rental company asks a question: where should the account managers' attention go? The company's account management can hold a limited number of appointments with the roughly 800 existing customer accounts it looks after. Which accounts should receive that attention, and how much, so that the margin they bring is protected? The firm's customer relationship management (CRM) system holds the data such a decision can draw on: yearly sales and margin per account, a dated log of appointments and the account records. It does not, however, tell the sales director how the appointments should be allocated, and whether an allocation was right shows only later, when the losses are known.

Decisions of this kind are routine in business-to-business (B2B) firms, which spend substantial resources on customer relationships (Palmatier, Dant, Grewal, & Evans, 2008; Reinartz, Krafft, & Hoyer, 2004) and must spread the same people's time over protecting existing margin, developing promising accounts and keeping costs in check (Ryals, 2005; Zoltners, Sinha, & Lorimer, 2004). Where the resources for winning new customers and for keeping existing ones can be separated, as in many B2B settings (Reinartz, Thomas, & Kumar, 2005), the allocation among existing accounts is a decision of its own. It is the decision studied here; the sales function that wins new business is outside the scope. In a survey of sales executives and of the sales literature, Zoltners, Sinha, and Lorimer (2008) found customer targeting among the issues that practitioners raise far more often than researchers study, and noted that firms are reluctant to let researchers experiment with their salespeople.

A natural analytical answer weights each account's margin by an estimated risk of losing the account and directs most attention to the accounts with most margin at risk. The system studied here, the Sales Operations Optimizer (SOO), does this. It combines a behavioral risk score with a risk estimate learned from earlier losses and allocates a fixed appointment capacity in closed form, so that its protection list is the set of eligible accounts with the largest margin at risk. SOO prioritizes accounts; it does not forecast the effect of attention. Whether the attention retains the margin is a causal question that requires randomized assignment (Ascarza et al., 2018), and the system can provide that assignment rather than answer the question. The question that can be answered from the firm's records is a prior one: does weighting margin by risk point the capacity at the accounts that are about to be lost, and how much of the margin do those accounts carry?

Two literatures address which customers should receive scarce attention. Value-based allocation studies rank customers by forecast lifetime value and allocate marketing contacts accordingly (Kumar, Venkatesan, Bohling, & Beckmann, 2008; Venkatesan & Kumar, 2004); they aim at growth and test the ranking at a single cutoff. Churn-prediction studies rank customers by their risk of leaving and compare classifiers, increasingly with profit-based measures (Coussement, Lessmann, & Verstraeten, 2017; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012; Verbraken, Verbeke, & Baesens, 2013). Most are set in B2C markets, and most evaluate by splitting the customers of one period at random or by cross-validation (De Caigny, Coussement, & De Bock, 2018; Tamaddoni, Stakhovych, & Ewing, 2016; Verbeke et al., 2012), a

practice that benchmark studies caution may overstate the advantage of more complex models (Lessmann, Baesens, Seow, & Thomas, 2015); tests on later periods are fewer (Gattermann-Itschert & Thonemann, 2021; Neslin, Gupta, Kamakura, Lu, & Mason, 2006; Risselada, Verhoef, & Bijmolt, 2010). In B2B markets, profit-driven studies weight churn risk by customer value (Janssens, Bogaert, Bagué, & Van den Poel, 2024; Rahman, Janssens, & Bogaert, 2025) and evaluate the ranking across all contact rates (Rahman et al., 2025), but by cross-validation within one period and without a ranking by value alone as the baseline. A randomized field experiment in a B2B wholesaler found that contacting the customers with the highest predicted churn reduced churn more than contacting randomly chosen ones, with customers selected by risk alone (Gattermann-Itschert & Thonemann, 2022). What remains open is how much a loss-risk estimate adds to a plain margin ranking when account managers' time is allocated among existing customer accounts, across the capacities a sales organization may have, and whether that answer holds when it is tested as the decision is made: on a fixed date, with only the information available then, and against the losses that followed. Whether attention changes those losses is a separate question that requires randomized assignment (Ascarza et al., 2018; Devriendt, Berrevoets, & Verbeke, 2021; Lemmens & Gupta, 2020).

This paper therefore asks three research questions:

RQ1 When a fixed capacity of account managers' time is allocated by margin at risk, how many of the accounts lost in the following year, and how much of the margin they carried, does it cover compared with rankings by margin, size and RFM (recency, frequency and monetary value), across the range of capacities?

RQ2 Using only information available on each decision date, how well can existing customer accounts be ranked by loss in the following year? What do engagement behavior, learned representations and more flexible models add beyond account size, and does the best choice of model hold from one year to the next?

RQ3 How good must a loss-risk estimate be before weighting margin by it protects more margin than a margin ranking, and how does that depend on the relation between account size and loss and on the capacity?

To answer them without information the decision maker could not have had, the paper replays the system's decision at three past year-ends, 31 December 2022, 2023 and 2024, with every input and every learned parameter restricted to what was known on that date, and scores each decision against the losses that followed (Kaufman, Rosset, & Perlich, 2011). The data are the company's records of 793 accounts from January 2021 to the export date, and a synthetic benchmark with planted ground truth extends the analysis beyond the one regime that a single firm provides.

The paper makes three contributions. First, it provides decision-date evidence on what loss risk adds to a plain value ranking when account managers' time is allocated among existing customer accounts. Each decision is scored on the losses that followed, across the whole range of capacities and against margin, size and RFM rankings, comparisons that the profit-driven B2B churn studies, evaluated by cross-validation within one period, do not make. Risk weighting found more of the departing accounts at the lower capacities but not more of the lost margin, which follows account size; value and

churn are negatively related in other B2B data as well (Rahman et al., 2025). Second, it evaluates loss-risk models forward over successive decision dates. Account size alone ranked next-year loss about as well as behavioral signals. Earlier forward tests found churn models losing accuracy over time, with the same method best throughout (Risselada et al., 2010), and a split within one period overstating the accuracy found on a later one (Gattermann-Itschert & Thonemann, 2021); here the model ranked first in one year was clearly behind in the next, an equal-weight combination stayed close to the leading model, and a random within-year split overstated next-year accuracy in one of two years. Comparing the chosen model's list with the combination's list at the same capacity, a two-model case of the top-k ambiguity studied for allocation under resource constraints (Watson-Daniels, Barocas, Hofman, & Chouldechova, 2023), marks on the decision day the accounts whose priority depends on the choice of model. An autoencoder representation added nothing over the raw features at this scale. These results corroborate or qualify findings of churn-prediction benchmarks (Coussement et al., 2017; De Caigny et al., 2018; Lessmann et al., 2015; Tamaddoni et al., 2016) in a setting those studies did not cover. Third, a synthetic benchmark with planted ground truth shows how good a risk term must be before it protects margin. A risk score whose errors are unrelated to account size starts to add lost margin over a margin ranking at an AUC of about 0.62 to 0.69, depending on how loss relates to account size, and at a lower AUC at the smallest capacity tested; the behavioral composite falls short of that line at the same AUC. Consistent with the observation that classification accuracy and campaign profit can diverge (Lemmens & Gupta, 2020; Verbraken et al., 2013), ranking accuracy alone cannot certify a risk term for margin protection; a replay on the firm's own history can.

Section 2 reviews the related work. Section 3 describes the decision, the allocation rule, the risk scores and the replay protocol, and Section 4 the case and data. Section 5 reports the decision of 27 October 2025 and answers the three research questions. Section 6 discusses the findings, their relation to earlier work and their limitations, including the randomized trial that would estimate the effect of attention, and Section 7 concludes.

2 Related work

The decision studied here draws on three bodies of work: sales resource allocation and value-based customer selection, churn prediction and retention targeting, and the study of customer loss in B2B settings. Two methodological strands, learned representations of customer data and the temporal evaluation of predictive models, shape the design of the evaluation. Table 1 summarizes how representative studies evaluated their recommendations.

2.1 Sales resource allocation and value-based selection

Allocating selling effort across accounts by mathematical programming has a long history. The CALLPLAN system (Lodish, 1971) allocated a salesperson's call time across accounts by optimizing over the salesperson's own estimates of how each account would respond to alternative call policies, and Zoltners and Sinha (1980) set out a general

Table 1 How representative studies evaluated customer selection, allocation and churn models

Study	Setting	Decision or outcome	Evaluation design	Compared with	Value measure
Venkatesan and Kumar (2004)	B2B hardware and software maker	Customer selection, contact levels by lifetime value	One later period; allocation by simulation	Past revenue, past value, lifetime duration	Realized profit (selection); lifetime value (allocation)
Kumar et al. (2008)	B2B (IBM, midmarket)	Reallocation of marketing contacts by lifetime value	One later period; non-randomized field roll-out	Spending score, RFM, past value	Profit (ranking); revenue (roll-out)
Verbeke et al. (2012)	Telecommunications (11 operator data sets)	Churn	Single random split per data set	21 classifiers	Maximum profit
Tamaddoni Jahromi, Stakhovych, and Ewing (2014)	B2B online retail	Churn (no purchase in the next half-year)	Customers split within one period	Logistic regression, trees, boosting	Campaign profit (assumed acceptance)
Coussement et al. (2017)	B2C telecommunications	Churn	Single split	Eight single and ensemble learners	–
Gordini and Veglio (2017)	B2B online retail	Churn	Random split within one year	Logistic regression, neural network, SVM	–
De Caigny et al. (2018)	14 data sets, several sectors	Churn	Cross-validation	Logistic regression, trees, ensembles	–
Lemmens and Gupta (2020)	B2C subscription and membership	Retention targeting by profit lift	Random splits of two campaigns (one randomized)	Churn and uplift models	Campaign profit
Devriendt et al. (2021)	Financial services (one European institution)	Churn and uplift	Random stratified split	Churn and uplift models	Maximum profit
Gattermann-Itschert and Thonemann (2022)	B2B wholesaler, field representatives (about 5,000 customers)	Retention contacts by churn risk (partial churn, three months)	One later period; randomized field test, model vs. random selection	Logistic regression, SVM; random selection	Revenue change (field test)
Janssens et al. (2024)	B2B beverage retailer (41,739 customers)	Profit-driven churn targeting	Cross-validation within one period	Six churn classifiers, incl. profit-driven	Maximum profit (B2B)
Rahman et al. (2025)	B2B beverage retailer and construction wholesaler	Targeting by churn × customer value	Cross-validation within one period	Churn and profit-driven models	Expected profit at every contact rate
Carvalho et al. (2025)	Construction-equipment seller (Canada)	Sales territories balanced on RFM value classes	Implemented; outcome not measured	Company's plan	–
This study	B2B equipment rental, account-managed	Account-management capacity by margin at risk	Decision-date replays on three dates; fitting on earlier dates only	Size, margin, RFM, negative control, eight risk scores	Lost margin; maximum profit (stylized)

framework and integer programming models for sales resource allocation with multiple resources, periods and carryover effects. Territory alignment has been optimized for total profit contribution (Skiera & Albers, 1998), and Carvalho et al. (2025) recently designed and implemented sales territories for a construction-equipment seller, balanced on value classes derived from recency, frequency and monetary value (RFM), expecting rather than measuring an effect on sales. The system studied here keeps the

structure of allocating effort under diminishing returns, which have been estimated for account-level effort in business markets (Bowman & Narayandas, 2004), but replaces judgmental response estimates with observed margin and a risk term tested against later losses.

Value-based selection ranks customers by a forecast of their future value. In a B2B setting, Venkatesan and Kumar (2004) selected customers and set contact levels by lifetime value, showed that a lifetime-value ranking selected more profitable customers than rankings by past revenue or past value, and noted that a field experiment against a control group would provide a stronger justification for such strategies. Reinartz et al. (2005) modeled how much to spend on acquiring and on retaining customers, and noted that the two resources can be separated in many B2B settings. At IBM, Kumar et al. (2008) reallocated marketing contacts by lifetime value, found that it selected customers better than RFM and past value over one later period, and reported revenue gains from a field implementation compared with a non-randomized group. These studies maximize future customer value, estimate the response to contacts from observational data, and test their rankings, where they do, at a single later period. Protecting the margin of existing accounts against loss, with account managers' time as the resource, is not their target.

2.2 Churn prediction, profit-based evaluation and uplift

Churn-prediction research ranks customers by their risk of leaving and compares classifiers, increasingly with measures of the profit a retention campaign would earn (Höppner, Stripling, Baesens, vanden Broucke, & Verdonck, 2020; Verbeke et al., 2012; Verbraken et al., 2013). Two findings of this literature bear directly on the present study. Logistic models, well prepared or combined with trees, are often competitive with ensembles (Coussement et al., 2017; De Caigny et al., 2018), and which technique is best depends on the data context (Tamaddoni et al., 2016); in a benchmark of credit-scoring classifiers, a simple average of all models ranked among the best (Lessmann et al., 2015). When several models predict about equally well, the choice among them can still change the predictions for many individuals (Marx, Calmon, & Ustun, 2020) and, when a fixed number of cases is selected under a resource constraint, which cases are selected (Watson-Daniels et al., 2023). Most of these studies are set in B2C markets, and they evaluate by splitting the customers of one period at random or by cross-validation (Table 1).

A second strand questions whether risk is the right target. The customers at highest risk are not necessarily those whose loss a retention action would prevent (Ascarza et al., 2018). Targeting by the predicted effect of an intervention can yield more profitable campaigns than targeting by risk (Devriendt et al., 2021; Lemmens & Gupta, 2020), even though such models predict churn itself less accurately (Lemmens & Gupta, 2020); estimating that effect needs variation in the intervention, ideally randomized. Operating a churn model also changes the data on which the next model is trained, a feedback loop that can be taken into account when interventions are recorded and control groups kept (Devriendt et al., 2021). Lemmens and Gupta (2020) further note that the number of customers to target has received little attention compared with the ranking itself; in the present setting, that number is set by the sales capacity.

2.3 Customer loss in B2B settings

B2B customer churn has received less attention than B2C consumer churn, although business customers are fewer and the loss of each weighs more (Gordini & Veglio, 2017; Tamaddoni Jahromi et al., 2014). Early studies predicted the inactivity of the business customers of online suppliers: Tamaddoni Jahromi et al. (2014) from recency, frequency and monetary value among more than 11,000 customers, and Gordini and Veglio (2017) with support vector machines among 80,000, without studying how the models held up in later periods. Gattermann-Itschert and Thonemann (2022) predicted a partial loss of revenue three months ahead among the roughly 5,000 business customers of a convenience wholesaler, served by phone agents and field representatives, from purchase records and the CRM log of contacts, and tested the model on a later period. In a randomized field experiment, calling and visiting the customers with the highest predicted churn reduced churn and kept more revenue than contacting the same number of randomly chosen customers, although the effect of the contacts on the called customers themselves remained unclear. Later work has turned to profit. Janssens et al. (2024) built a profit-driven churn model for a beverage retailer with 41,739 business customers, valuing each customer by its past purchases, and found that models with an AUC around 0.80 did not yield profitable campaigns under their cost assumptions. Rahman et al. (2025) compared, among other methods, a ranking by predicted churn times predicted customer value across all contact rates in the same retailer and a construction-material wholesaler, found value and churn negatively correlated in both, and excluded the customers served by key-account managers. De Caigny, Coussement, Verbeke, Idbenjra, and Phan (2021) estimated the effect of a retention campaign on the churn of 6,432 software contracts. These three studies evaluate by cross-validation within one period, and none of the B2B studies above compares its rankings with a ranking by customer value alone. Within B2B analytics, RFM has been extended with loyalty and periodicity (Lau & Wasesa, 2025), with the trend of customer value over time (Hosseini & Shabani, 2015) and with Net Promoter Score survey data (Rusdarmono, Sofwan, & Batu, 2025), but these studies segment or predict without testing their rankings against later losses. Evidence on the value of customer analytics in B2B firms is largely cross-sectional: firms that report more use of customer big data also report better sales growth (Hallikainen, Savimäki, & Laukkanen, 2020), and a review of B2B customer-behavior research finds a lack of longitudinal data and field experiments (Bilro, Loureiro, & Souto, 2023). Where field staff and their CRM records do appear (Gattermann-Itschert & Thonemann, 2022), the decision is which customers to contact, chosen by risk alone, rather than how account managers' time should be spread over the existing accounts by margin at risk.

2.4 Learned representations of customer data

Representation learning aims to discover useful features from data rather than to specify them by hand (Bengio, Courville, & Vincent, 2013). Autoencoders, trained to reconstruct their inputs, are an established route to compact non-linear representations (Hinton & Salakhutdinov, 2006; Wang, Huang, Wang, & Wang, 2014) and have been applied to bank-customer data (Mancisidor, Kampffmeyer, Aas, & Jenssen,

2021); related neural embeddings have been learned for online-retail customers (Chamberlain, Cardoso, Liu, Pagliari, & Deisenroth, 2017). The gains reported for customer data come from large B2C data sets and are small or confined to part of the data: embeddings of browsing sessions improved prediction at an online fashion retailer only slightly (Chamberlain et al., 2017), and model-based embeddings improved recommendations at eBay mainly in the smallest customer segment, on a random split (Jagabathula, Subramanian, & Venkataraman, 2018). On medium-sized tabular data, tree-based models remain stronger than deep learning, while very small data sets remain an open question (Grinsztajn, Oyallon, & Varoquaux, 2022). Whether a learned representation adds anything at the scale of one firm's accounts therefore has to be tested rather than assumed.

2.5 Temporal evaluation and leakage

A model that is used on a decision date meets data from after the date on which it was fitted. Evaluating it on customers of the same period, as random splits and cross-validation do, can give a different answer from evaluating it on the following period. Gordini and Veglio (2017) left the staying power of their model unstudied, and Verbeke et al. (2012) stressed that continuous evaluation on newly gathered data is crucial while evaluating on a single random split. Lessmann et al. (2015) caution that laboratory experiments may overestimate the advantage of advanced classifiers, and population drift is one reason why the advantage of a more sophisticated classifier can fade once it is applied to later data (Hand, 2006). Tests on later periods exist but are few. In a churn-modeling tournament, accuracy fell little between customers of the calibration period and customers scored about three months later (Neslin et al., 2006), whereas churn models re-estimated over successive periods lost accuracy within one or two periods ahead, with the same method best in every period (Risselada et al., 2010). In a B2B wholesaler, a split of the customers of one period overstated the accuracy found on a later period (AUC 0.75 against 0.70), and training on several earlier time slices improved it (Gattermann-Itschert & Thonemann, 2021). Leakage, the use of information that would not be available at the time of prediction, inflates estimated performance; a frequent source is database records that are updated without time stamps, and the remedy is a separation of learning and prediction by time (Kaufman et al., 2011). The value-based selection studies above test their rankings over a later period, but on a single cutoff and without intervals. The present study applies the learn–predict separation to the whole decision: features, representations and every fitted parameter are built as of each decision date, and each decision is scored on the losses that followed.

3 Decision problem and method

Table 2 collects every setting of the method, how it was set and where its effect is tested. Online Resource 1 specifies every step precisely enough to rebuild the system, and Online Resource 2 contains the code of every analysis and a synthetic dataset with planted ground truth on which it runs.

Table 2 Settings: value, how each was set, and where its effect is tested

Setting	Value	How set	Tested in
Composite weights (w_v, w_d, w_h)	0.40, 0.30, 0.30	design default	Sec. 5.3 (per component); Online Resource 1 (five alternative weightings)
Contact expectation T_c by class	XXL 30, XL 45, L 60, M 90, S 120, XS 180 days	design default	Online Resource 1 (dormancy-only variant)
Dormancy ramp	0 at T_c, 100 at $4\,T_c$	design default	–
Trend-to-risk mapping	-100% change $\to$ 100, 0% $\to$ 50, $+100\% \to 0$	design default	–
Autoencoder	51–128–64–32, dropout 0.25, early stopping (patience 15); trained at each decision date	design default	Sec. 5.4; Online Resource 1 (seeds)
Logistic heads	ℓ_2, $C = 0.3$ (embedding, principal components, size plus behavior), $C = 0.1$ (raw features); standardized inputs	fixed before the replays	Sec. 5.3
Gradient boosting	depth 3, learning rate 0.05, 200 iterations, minimum leaf 20, ℓ_2 1.0	fixed before the replays	Sec. 5.3
Blend weight w	0.5	pre-specified system risk term	Sec. 5.3 (vs. combination)
Combination	equal weights over five scores (Eq. 5)	added after the first replay run (model-choice check)	Secs. 5.1, 5.3
Diminishing returns λ	0.05	design default	Sec. 5.6 (0.01–0.20)
Capacity A; cap $a_{\max}$	5,000 appointments; 24 per account	planning scenario (Sec. 4)	Sec. 5.2 (1,200–10,000); Sec. 5.6 (cap)
Protection eligibility	margin in the decision year > 0; on the real decision date also status purchasing	added after the first replay run	Sec. 5.2 (without the rule)
Annual margin M_i	margin per year as a customer within the window	adopted after the first replay run	Online Resource 1 (whole-window average)
Outcome	complete loss of margin in $T + 1$	data-derived (Sec. 4)	Secs. 5.2, 5.3
Intervals	2,000 bootstrap resamples, percentile	standard	–
Expected maximum profit	customer value 200, incentive 10, contact cost 1, acceptance Beta(6, 14)	telecom defaults (Verbraken et al., 2013)	Sec. 5.3

3.1 The decision on a date

The decision can be made on any day. On that day, the sales director responsible for existing customers asks who should get attention and how much: which of the firm's existing customer accounts, that is, the customer organizations with sales history, should receive it, and how much each should get. The system answers with one aim, to protect the margin these accounts bring, and from the history accumulated in the CRM up to that day. New business is handled by a separate function with its own capacity and is not part of the decision. The answer has three parts: a protection list of accounts, each with an allocated number of appointments a year; a ranking of accounts by loss risk that can serve as retention alerts; and a separate win-back list of accounts that have stopped buying.

The method consumes three exports of the firm's CRM: one row per account (class, engagement model, rating, status), one row per account and calendar year (sales and margin) and one row per appointment (account, start time, status). Every quantity is computed as of the decision date (Figure 1). The dated appointment log can be cut

at any past day, but the export used here aggregates sales by account and calendar year. On a decision date, current-year sales and margin are therefore year-to-date. On 27 October 2025 they cover almost ten months, but early in a year they would cover little, and few accounts would yet meet the eligibility rule below. A past date can be replayed without leakage only at a year-end, where the ledger gives complete years before and after it. Sales recorded with their dates would allow every quantity to be measured over the twelve months before the decision date, whatever the day, and any past day to be replayed in the same way (Online Resource 1). At a year-end, the same code reproduces what the system would have seen on that date. The account fields exist only as they stood at the export date; Section 3.4 describes how their use is checked.

Annual margin M_i is the account's cumulative margin in the observation window up to the decision date divided by the time it has been a customer within the window. Because the ledger dates sales only by year, that time is counted from the start of the year of the account's first sale, or from the start of the window if the account bought earlier. Absolute margin rather than margin percentage is used so that large accounts with moderate percentages, which contribute most absolute profit, rank above small accounts with high percentages. Because M_i averages over the whole relationship, an account that has already stopped buying can still carry a large M_i. The system therefore applies an *eligibility rule*: only accounts with positive margin in the decision year can enter the protection list. Accounts with past margin but none in the decision year form the win-back list, which is reported separately and receives no capacity from the allocation. The rule and the averaging over the time as a customer were both adopted after the first replay run, which showed the protection list mixing the two groups and, with margin averaged over the whole window, ranking accounts that had started buying recently below their current business. Results without the rule, with every account that has past margin competing for the capacity, are reported alongside (Section 5.2), and results with the whole-window average in Online Resource 1.

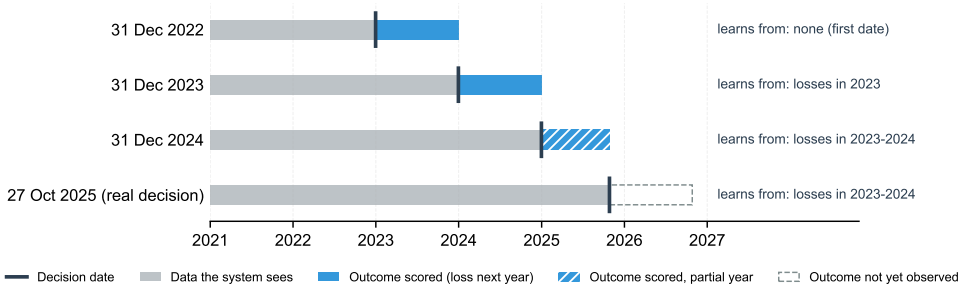


Fig. 1 Decision dates and the information each one uses. At each replay date the system sees only the data up to that date (light bars) and is scored on account loss in the following year (dark bars; hatched for the end-2024 replay, whose outcome covers 2025 up to the export date). Fitted parts learn only from the outcomes of earlier dates. The real decision of 27 October 2025 follows the same procedure; its outcome is not yet observed (dashed outline).

3.2 Allocation rule

Let a_i denote the annual appointment allocation for account i. The allocation spreads a fixed appointment capacity over the eligible accounts in proportion to each account's margin and the risk to it:

$$\max_{a_1,\ldots,a_N} \sum_{i=1}^{N} M_i \cdot E_i \cdot Q(a_i) \qquad \text{subject to} \qquad \sum_{i=1}^{N} a_i = A, \quad 0 \le a_i \le a_{\max}, \tag{1}$$

where M_i is annual margin, $E_i \in [0,1]$ is the loss-risk term of Section 3.3.2, A is total capacity, $a_{\max}$ is a per-account cap and $Q(a_i) = \ln(1+\lambda a_i)$ encodes diminishing returns from appointment intensity, in line with the decreasing returns to account-team contact hours that Bowman and Narayandas (2004) estimated, on customers' ratings of the account team, for the business customers of an industrial vendor. Q represents improvement *potential*, not an estimated effect of an appointment on retention.

The objective multiplies margin and risk rather than adding them, so that priority requires both margin and signs of deterioration. A consequence is that an account with near-zero risk receives little priority however large it is. A floor on E_i, an additive objective, revenue in place of margin and single-component risk terms are therefore compared as alternatives (Section 5.6; full comparison in Online Resource 1).

The problem belongs to the class of continuous nonlinear resource allocation problems with a separable objective and a single capacity constraint (Patriksson, 2008). Because the objective is separable and strictly concave in each a_i, the Karush–Kuhn–Tucker conditions (Karush, 2014; Kuhn & Tucker, 1951) give the optimum in closed form:

$$a_i^* = \operatorname{clip}\left(\frac{1}{\lambda}\left(\frac{M_i E_i}{\mu} - 1\right),\ 0,\ a_{\max}\right), \tag{2}$$

where $\mu > 0$, the multiplier on the capacity constraint divided by λ, is found by bisection so that $\sum_i a_i^* = A$. The solution is exact; Sequential Least Squares Programming (Kraft, 1988; Virtanen et al., 2020) serves only as a numerical check. Equation (2) also shows what the optimizer decides. An account receives a positive allocation exactly when $M_i E_i > \mu$, so the protection list is the top-N of eligible accounts by $M_i E_i$, which we call the account's *margin at risk*, and within the list the allocation rises linearly with $M_i E_i$ until it reaches the cap. Because E_i is a risk score rather than a calibrated probability, margin at risk is a priority index, not an expected loss in euros. Capacity and cap set how many accounts are prioritized and how intensively, not which come first. The system is therefore prescriptive in the sense of Bertsimas and Kallus (2020): it uses data to prescribe a decision. In their terms it makes a point-prediction-driven decision, in which an estimated loss risk enters the optimization in place of the unknown outcome, so the evaluation scores the decision as well as the estimate. Which accounts are on the list can be tested against later losses; the intensity of attention cannot, because the appointments delivered after a past decision date were the company's, not the system's.

3.3 Loss-risk scores

Three terms are kept apart below. A *risk score* ranks accounts by their risk of being lost in the following year, and a *risk model* is the method that produces it, weighted by hand like the composite or fitted to earlier losses. The *risk term* E_i is the risk score that enters the allocation of Eq. (1); for the system it is the blend of Eq. (4).

3.3.1 Composite engagement risk

The composite answers one question: *is this customer relationship showing signs of deterioration?* It combines three signals defined on different quantities, the appointment trend, contact recency and the margin trajectory, so that no input enters the score twice. Each component is expressed on a 0–100 risk scale, and the composite is their weighted sum:

$$E_i^{\text{comp}} = w_v \cdot \text{Velocity Risk}_i + w_d \cdot \text{Dormancy Risk}_i + w_h \cdot \text{Health Risk}_i. \qquad (3)$$

Velocity risk is the percentage change in appointments between the twelve months before the decision date and the twelve months before that, mapped linearly so that vanished activity scores 100, stable activity 50 and doubled activity 0. Accounts without activity in either window receive the neutral value.

Dormancy risk measures contact recency against a class-specific expectation: risk is zero while the days since the last appointment are within the class threshold T_c and rises linearly to 100 at four times the threshold, with never-contacted accounts at 100.

Health risk is the percentage change between the margin of the last complete calendar year and the average of the preceding complete years, taken from the sales ledger and mapped as velocity is. Accounts without margin in any complete year receive the neutral value, and margin that appears only in the last complete year counts as growth.

The weights are defaults, not fitted values. In the replays the account-status field is not used, because it is known only as of the export date. On the real decision date, status is current information, and accounts no longer recorded as purchasing are routed to the win-back list.

Negative control. A risk score that is inverted, or that counts one signal twice, still produces a plausible-looking allocation, because margin dominates the priority weight. The evaluation therefore includes an *inverted score*: the same weighted sum applied to the components on their stability scale rather than their risk scale. It counts contact recency twice and defines velocity only for accounts active in three of the last four quarters (Online Resource 1). Such a negative control should rank accounts below chance on subsequent loss.

3.3.2 Learned representation and the system's risk term

The representation input comprises 51 numeric features in five groups: behavioral, performance, risk (including the composite of Eq. 3), structure and potential, and categorical indicators built from the labels present in the export. The purchasing-status indicator is held out. A feedforward autoencoder (Table 2) compresses the 51

standardized features into a 32-dimensional embedding $\mathbf{z}_i$. At each decision date it is trained on that date's features only and without outcomes. The feature list and the treatment of the CRM's categorical fields are given in Online Resource 1.

An unsupervised embedding has no orientation: nothing in it says which direction corresponds to risk. The system therefore learns the direction from outcomes. An account is recorded as *lost* in year $T+1$ if it generated positive margin in year T and none in year $T+1$. Year-to-year margin is lumpy in equipment rental (Section 4), so a partial decline is not a reliable signal, whereas a complete loss is. A regularized logistic regression $\hat{p}_i = \sigma(\mathbf{v}^\top \tilde{\mathbf{z}}_i + b)$ on the standardized embedding is fitted to the accounts and outcomes of earlier decision dates and gives the embedding-derived risk $E_i^{\text{emb}} = \hat{p}_i$.

The system's risk term blends the two estimates on rank scale,

$$E_i = w \cdot \text{rank}(E_i^{\text{comp}}) + (1 - w) \cdot \text{rank}(E_i^{\text{emb}}), \tag{4}$$

where rank$(\cdot)$ is the percentile rank across accounts, which puts a hand-weighted score and a fitted probability on a common scale. The blend with $w = 0.5$ was fixed as the system's risk term before the replays were run; choosing the best-performing score afterwards would select on the test data. Equal weighting is a stated default rather than an estimate. In the forecast-combination literature, simple averages have often performed as well as or better than estimated weights (Clemen, 1989), which Timmermann (2006) attributes mostly to estimation error in the weights when few data are available; equal weights are also close to optimal when the combined estimates are about equally accurate (Timmermann, 2006). On the first decision date there is no loss history from which to learn a projection, and the system runs on the composite alone, as a new deployment would.

3.3.3 Comparison scores

Every other score serves as a comparison, with settings fixed before the replays (Table 2). Three are rankings a manager could use without a model: account *size*, measured by revenue in the decision year; *margin* M_i; and an RFM-type ranking that sums the percentile ranks of appointment recency, appointment count and annual margin. As a loss-risk score, size ranks small accounts first. Four learned scores are fitted on the same earlier outcomes as the embedding projection. They are a regularized logistic regression on the 51 raw features, gradient boosting on the same features (Grinsztajn et al., 2022), a logistic projection of 32 principal components, and a logistic model on size and the three composite components (*size plus behavior*). In the allocation comparison each risk score is used as margin times score.

The last comparison tests the system's own choice of model. The *combination* is the equal-weight mean of the percentile ranks of every learned and composite score available on the date,

$$E_i^{\text{comb}} = \frac{1}{K} \sum_{k=1}^{K} \text{rank}\big(E_i^{(k)}\big), \tag{5}$$

over the composite, the embedding projection, the principal-component projection, the raw-feature logistic regression and gradient boosting. Size is excluded because it

already enters through the margin. Unweighted averages of many models are a strong benchmark in classifier comparisons (Lessmann et al., 2015), which single model is best is known to depend on the data context (Tamaddoni et al., 2016), and equal weights hedge against changes over time in which model performs best (Timmermann, 2006). The combination was added after the first replay run as a check of the pre-specified blend, not as its replacement: if the protection list changes materially when the combination replaces the blend, the recommendation depends on which model was chosen. Comparing the selections of two models at the same capacity is a two-model case of the top-k ambiguity defined for allocation under resource constraints (Watson-Daniels et al., 2023). Such a comparison can show that the selection depends on the choice of model, but not that the dependence is small, since further candidate models could change more accounts (Marx et al., 2020).

3.4 Replay protocol and evaluation

The procedure is replayed at 31 December 2022, 2023 and 2024 (Figure 1). At each date the features, the autoencoder, the scaler and the principal components are built from data up to that date only. Every fitted part is trained only on earlier dates whose outcome year has ended by the decision date: on nothing at the first date, on the 2023 outcomes at the second and on the 2023 and 2024 outcomes at the third. Pooling the earlier dates in this way is a yearly form of training on several time slices, which improved later-period accuracy in a B2B churn study (Gattermann-Itschert & Thonemann, 2021). This is the learn–predict separation recommended against leakage by Kaufman et al. (2011): no information from after the decision date enters any input or any fitted parameter. Model settings were fixed in advance and not tuned on the replays. For the end-2024 replay the outcome is observed over 2025 up to the export date, so an account that buys again later in 2025 counts as lost (Section 4). On the real decision date the same rules apply, with the fitted parts learning from the 2023 and 2024 outcomes, the two complete outcome years.

The account fields that exist only as of the export date (engagement model, account rating, customer type and the scores built from them) could carry information from after a past decision date back to it, because records are overwritten rather than time-stamped (Kaufman et al., 2011). The replays are therefore repeated without them. In the terms of Kaufman et al. (2011) this is a check for leakage rather than a separation, which the export does not allow. Customer class, which sets the contact thresholds, is checked separately against revenue at the decision date.

Allocation (RQ1). At capacities of 1,200, 2,500, 5,000 and 10,000 appointments the system's protection list has a size N, and every comparison ranking is cut at the same N among the same eligible accounts. For each list we count the accounts lost in the following year and the share of their decision-year margin that the list covers; capture curves repeat the count at every list size. That the system's list equals the top-N eligible accounts by margin at risk is checked in every replay and at every capacity.

Ranking and model choice (RQ2). Among accounts with margin in the decision year, each risk score is evaluated by the area under the receiver operating characteristic (ROC) curve (AUC) for next-year loss, with bootstrap intervals, and by

paired bootstrap differences against account size. Retention alerts are evaluated by the share of the riskiest 10% and 20% of accounts that leave and by its ratio to the base rate (lift); the cut-offs are illustrative (Verbraken et al., 2013). A profit-based measure, the expected maximum profit for customer churn (Verbraken et al., 2013), is reported with the parameter values of its telecommunications application (Table 2). No B2B estimates of these parameters exist, so it is a stylized check that ranking quality would translate into campaign value, not a profit estimate. In each replay with fitted parts, the combination is compared with the blend by the paired AUC difference. On the real decision date, where no outcome exists yet, the two protection lists are compared directly, both at the same list size and at the same capacity: accounts on both lists, accounts on one only, and their share of the allocated appointments. Finally, to see how much the evaluation design itself matters, the learned scores are also evaluated by repeated stratified cross-validation within the decision year (5×5 folds), the design common in the churn-prediction literature, and compared with their forward result on the same accounts.

Conditions (RQ3). Three replays of one firm show one regime of signal strength and of the relation between account size and loss. A synthetic benchmark with planted ground truth varies both (Section 5.5).

3.5 What the evaluation can and cannot show

The evaluation separates two claims that need different evidence. The first is that the system identifies the accounts where margin is at risk; the replays test it against three years of subsequent outcomes. The second is that allocating appointments to those accounts changes outcomes. Observational CRM data cannot establish this, because past appointment intensity is itself a response to perceived risk, and the appointments delivered after each past decision date were the company's, not the system's. It requires a randomized trial among prioritized accounts, for which the system can serve as the assignment mechanism (Section 6; design in Online Resource 1). The output should therefore be read as a ranking of where attention is warranted, tested against subsequent losses, together with an allocation of intensity whose effect on retention remains to be measured.

4 Case and data

The study uses the records of a European business-to-business technology-equipment rental company that serves corporate and public-sector customers through account managers responsible for customer portfolios. Sales cycles range from 3 to 18 months, and annual contracts are typical. The data cover 793 existing customer accounts from January 2021 to the export date of 27 October 2025 (Table 3). Margin is concentrated: the largest quarter of the accounts holds 83% of it.

The appointment log holds 5,868 appointments up to the export date, about 1,200 a year; the median account received four, and 139 accounts (18%) none. The capacity of 5,000 appointments a year used in the analysis is a planning scenario: it corresponds to the company's whole annual appointment volume across all of its accounts, about four times the logged rate to these 793 accounts, directed at the managed portfolio.

Table 3 Accounts by customer class, January 2021 to 27 October 2025

Customer class	Accounts	Revenue 2021–2025 (€M)	Mean per account (€M)
XXL (¿3000 users)	24	108.4	4.5
XL (1000–3000)	58	125.4	2.2
L (500–1000)	60	103.9	1.7
M (200–500)	123	59.1	0.5
S (100–200)	155	40.1	0.3
XS (¡100)	369	43.9	0.1
Unclassified (no sales in the window)	4	0.0	0.0
Total	793	480.9	0.6

The cap of 24 appointments a year, twice a month, is the most frequent contact the company would plan for a single account; no account received more than 19 in the year before the export. Because capacity and cap set the size of the list but not its order, the results are reported for capacities from 1,200 to 10,000 appointments.

Table 4 lists the decision dates. Of the accounts with margin in a decision year, 13 to 23% bought nothing in the following year. The end-2024 outcome covers 2025 up to the export date. Its loss rate is close to the full-year rate of the year before, so the shorter window does not visibly inflate losses, although an account that resumes buying in November or December 2025 is counted as lost. Year-to-year margin is lumpy: among accounts with margin in two consecutive years, the log ratio of the later to the earlier year has a standard deviation of 1.2, which is why the outcome is complete loss rather than a partial decline. The win-back pool, the accounts with past margin but none in the decision year, grows from 37 to 194 accounts over the three dates, and few of them buy again in the following year.

Table 4 Decision dates, outcomes and what the system learns from

Decision date	Accounts with margin	Lost in the following year	Lost margin (€M)	Fitted parts learn from	System's risk term	Win-back pool (bought again)
31 Dec 2022	504	65 (12.9%)	0.24	– (no loss history)	composite	37 (13)
31 Dec 2023	522	119 (22.8%)	0.59	losses in 2023	blend	89 (14)
31 Dec 2024	484	105 (21.7%)[a]	0.67	losses in 2023–2024	blend	194 (21)
27 Oct 2025	463[b]	not yet observed	–	losses in 2023–2024	blend	278

Lost margin: decision-year margin of the accounts lost. [a]2025 up to the export date. [b]Accounts eligible for protection: margin in 2025 to date and status purchasing.

5 Results

Section 5.1 reports the decision of 27 October 2025. Sections 5.2 to 5.5 answer the research questions: what a fixed capacity covers (RQ1), how well accounts can be

ranked by loss and how stable the choice of model is (RQ2), and when the risk term pays (RQ3). Section 5.6 summarizes the robustness checks.

5.1 The decision of 27 October 2025

On the export date, 463 accounts were eligible for protection: they had margin in 2025 to date and were recorded as purchasing. With 5,000 appointments and a cap of 24, the system places 278 of them on the protection list (Table 5). The listed accounts hold 90% of the eligible accounts' margin, whether measured as annual margin or as margin earned in 2025 to date, and 153 of them are allocated the cap. A further 278 accounts form the win-back list: 250 are no longer recorded as purchasing, and 28 bought in 2024 but not in 2025 to date. Together they held €1.75M of annual margin, which the protection allocation does not try to defend.

Table 5 The decision of 27 October 2025: protection list by intensity, and the accounts outside it

Group	Accounts	Appointments	Annual margin (€M)
At the cap (24 appointments)	153	3,672	3.64
High (12–23)	52	920	0.43
Medium (1–11)	73	407	0.49
Protection list	278	4,999	4.56
Eligible, not on the list	185	0	0.51
Win-back list	278	0	1.75

Capacity 5,000 appointments, cap 24. Appointments are rounded tier sums; about one appointment falls on accounts allocated less than half an appointment. Annual margin is M_i, margin per year as a customer. The 52 accounts without positive margin in the window are not shown.

The list directs attention to accounts that have gone quiet. The median time since the last appointment is 245 days for the listed accounts, against 138 days for a margin ranking of the same size and 75 days for the same-size list of the negative control.

Between the two risk models compared, the decision depends little on which is used. A list of the same size built with the equal-weight combination of all risk models in place of the blend shares 261 of the 278 accounts, so 17 accounts change. At the same capacity of 5,000 appointments the combination's list holds 287 accounts, among them 262 of the system's 278, which receive 98.5% of the system's appointments. The accounts on one list only are the decisions that depend on the choice of model, and they are natural cases for an account manager's judgment. Because only two models are compared, these counts are a lower bound on how much the choice of model matters. Without the eligibility rule, that is, if accounts that had stopped buying could also enter the protection list, a list of the same capacity would have contained 294 accounts, 124 of them without margin in 2025, and would have shared 170 accounts with the protection list. Of the 63 eligible accounts that first bought in 2025, 32 are on the list; with margin averaged over the whole window instead of over the time as a customer, 9 would be.

Whether this list protects the margin it covers cannot be observed yet. The following sections show how lists made by the same procedure fared at earlier dates.

5.2 What a fixed capacity covers (RQ1)

Table 6 and Figure 2 compare the lists the system would have produced on the three replay dates with rankings of the same size by margin, account size and RFM, all applied to the same eligible accounts. Two patterns hold on both dates on which the system had a loss history to learn from.

Table 6 What a fixed capacity covers: accounts lost in the following year (panel A) and their share of the lost decision-year margin (panel B) inside lists of the size the system produces

	31 Dec 2022 (65 lost)			31 Dec 2023 (119 lost)			31 Dec 2024 (105 lost)		
Capacity (appointments)	1,200	2,500	5,000	1,200	2,500	5,000	1,200	2,500	5,000
A. Lost accounts on the list									
List size N	79	149	263	88	163	278	86	157	281
SOO, protection list (system)	2	9	15	20	36	62	28	39	56
SOO without the eligibility rule	2	9	15	18	35	58	23	30	49
Combination of models × margin	2	9	14	17	36	61	31	42	58
Margin ranking	3	5	14	14	26	49	16	32	55
Size ranking	3	6	19	11	19	43	14	27	49
RFM ranking	3	5	18	14	30	53	17	29	57
Inverted score × margin (control)	3	6	16	13	26	44	18	32	54
B. Share of the lost margin on the list (%)									
SOO, protection list (system)	8	24	59	45	75	87	65	75	88
SOO without the eligibility rule	8	24	59	44	75	85	62	67	86
Combination of models × margin	5	24	56	42	78	87	67	77	89
Margin ranking	37	45	71	56	70	86	62	77	89
Size ranking	38	48	75	57	68	87	62	81	92
RFM ranking	29	32	67	54	75	88	62	70	87
Inverted score × margin (control)	37	48	73	55	71	86	66	78	89

N is the size of the system's protection list at the capacity; every other ranking is cut at the same N among the accounts with margin in the decision year. On the first date the system's risk term is the composite. SOO without the eligibility rule (Section 3.1) ranks all accounts with past margin by annual margin times risk, including accounts that had already stopped buying in the decision year.

First, at the lower capacities, 1,200 and 2,500 appointments a year, where the protection list holds about a sixth to a third of the eligible accounts, the system's list contains more of the accounts lost in the following year. At 1,200 appointments it holds 20 of the 119 accounts lost in 2024 and 28 of the 105 lost in 2025, against 14 and 16 on a margin ranking of the same size, and the advantage persists at 2,500 appointments. As the list grows toward the whole eligible base, the two dates differ: at 5,000 appointments the system holds 62 lost accounts against 49 on the end-2023 date, but 56 against 55 on the end-2024 date. Second, on lost margin the system has no clear advantage. It is within a few points of the margin ranking at most capacities, ahead at some and behind at others, and eleven points behind at 1,200 appointments on the end-2023 date, because lost margin follows account size: a ranking that puts large accounts first captures most of the lost euros. The risk term's contribution is to

find the accounts about to leave that such a ranking passes over. Pooled over the two dates with loss history, the system's list holds 18 more lost accounts than the margin ranking at 1,200 appointments, 17 more at 2,500 and 14 more at 5,000, clearly beyond the resampling noise ($z = 3.2$, 3.5 and 2.9). None of the differences in lost margin can be distinguished from zero (largest $z = 0.8$); Section 6.4 discusses why.

On the first date, without loss history, the system runs on the composite alone. Its list finds about as many leavers as the margin ranking at 1,200 appointments (2 against 3) and more at 2,500, but it captures far less of the lost margin at every capacity.

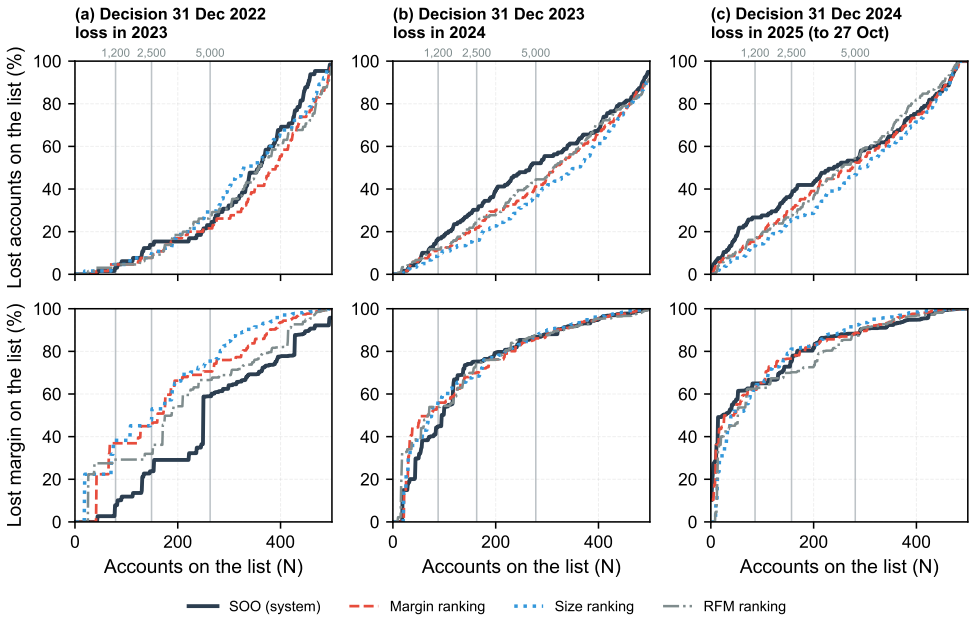


Fig. 2 Accounts lost in the following year (top) and their decision-year margin (bottom) on lists of increasing size, as a share of all accounts lost, for the decisions of 31 December 2022 (a), 2023 (b) and 2024 (c). Solid line: the system's protection list; dashed: margin ranking; dotted: size ranking; dash-dot: RFM ranking, all among the accounts with margin in the decision year. Vertical lines mark the system's list size at capacities of 1,200, 2,500 and 5,000 appointments.

The negative control behaves as intended. At the lower capacities, margin times the inverted score finds fewer leavers than the system on the dates with loss history, yet it captures as much lost margin as the margin ranking, so euro capture alone cannot validate a risk term. The eligibility rule never does worse than the same system without it and helps most on the end-2024 date, when the pool of accounts that had stopped buying is largest and, without the rule, takes places on the list. Averaging margin over the whole window, as in the first replay run, changes the lists by at most four lost accounts on the dates with loss history and captures less of the lost margin at 5,000 appointments (83 and 80% against 87 and 88%; Online Resource 1).

The combination of all risk models produces lists close to the system's. Appointments delivered after each decision date may themselves have affected the outcomes, so the comparison shows which accounts were lost under the company's actual sales activity, not the effect of the recommended allocation (Section 3.5).

5.3 Ranking accounts by next-year loss and the choice of model (RQ2)

Account size alone ranks next-year loss with AUCs of 0.68, 0.66 and 0.61 on the three dates (Figure 3, Table 7): small accounts are lost more often. On the first date, without loss history, the behavioral composite is clearly weaker than size (0.54; difference −0.14, 95% interval −0.22 to −0.06). Once a year of losses is available, the system's blend ranks level with size and slightly above the composite on the end-2023 date (0.68; +0.02 over the composite, −0.01 to +0.06), and above the composite but not clearly above size on the end-2024 date (0.63; +0.07, +0.03 to +0.10). The negative control stays below chance on every date.

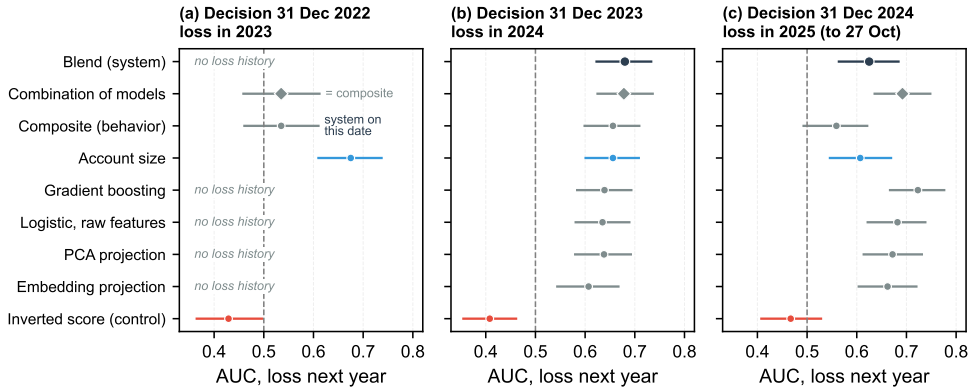


Fig. 3 Area under the ROC curve (AUC) for loss in the following year, with 95% bootstrap intervals, by score and decision date, among the accounts with margin in the decision year. Diamonds mark the equal-weight combination of models; the dashed line marks chance. Learned scores need earlier outcomes and are absent on the first date, on which the system runs on the composite.

As retention alerts, the riskiest tenth of accounts by the system's risk term leaves at about twice the base rate once loss history exists: 46% against 23% on the end-2023 date and 35% against 22% on the end-2024 date. The riskiest third is lost two to three times as often as the safest third on these dates, while the first date shows no gradient. The alerts point mostly to smaller accounts and hold little of the lost margin, consistent with the capture results. The profit-based measure ranks the scores much as the AUC does (rank correlations of 1.00, 0.84 and 0.99 on the three dates) and puts the same score first, or two scores level, on each date; every score except the negative control is above its no-information level once loss history exists. Under its

telecommunications parameters it would target half or more of the accounts, which reflects those parameters rather than this setting.

Accounts in their first year show where the behavioral composite falls short. Among the accounts with margin in a decision year, those that first bought in that year were lost in the following year more often than older accounts (32, 26 and 28% against 9, 22 and 21% on the three dates). The composite nevertheless rates them as healthy, with a median of 6 or 7 on its 0–100 scale against about 50 for older accounts, because they have been met recently, their activity is rising and they have no complete year of margin that could decline. The learned half of the blend partly corrects this: on the two dates with loss history the system's risk term gives first-year accounts a median of 36 against 38 and of 24 against 36.

Table 7 Ranking accounts by loss in the following year: AUC (panel A), retention alerts (panel B) and expected maximum profit (panel C), by decision date

	31 Dec 2022	31 Dec 2023	31 Dec 2024
A. AUC, loss in the following year			
Account size (small first)	0.68	0.66	0.61
Composite (behavior)	0.54	0.66	0.56
Blend (system)	–	0.68	0.63
Combination of models	0.54	0.68	0.69
Logistic regression, raw features	–	0.64	0.68
Gradient boosting	–	0.64	0.72
PCA projection	–	0.64	0.67
Embedding projection	–	0.61	0.66
Inverted score (control)	0.43	0.41	0.47
B. Share of the riskiest 10% that leave (lift over the base rate)			
Base rate (all accounts with margin)	13%	23%	22%
Account size (small first)	30% (2.3)	42% (1.9)	42% (1.9)
Composite (behavior)	12% (0.9)	42% (1.9)	38% (1.7)
System's risk term	12% (0.9)	46% (2.0)	35% (1.6)
Combination of models	12% (0.9)	52% (2.3)	54% (2.5)
Gradient boosting	–	38% (1.7)	54% (2.5)
Riskiest vs. safest third (system)	13% vs. 11%	40% vs. 14%	30% vs. 16%
C. Expected maximum profit per account (telecommunications parameters)			
No information	0.26	4.57	3.92
Account size (small first)	1.54	5.63	4.50
Composite (behavior)	0.60	5.74	4.32
Blend (system)	–	5.98	4.95
Combination of models	0.60	5.98	5.50
Gradient boosting	–	5.73	6.01

Accounts with margin in the decision year: 504, 522 and 484. – : needs earlier outcomes. On the first date the system's risk term and the combination equal the composite. Expected maximum profit after Verbraken et al. (2013), in units of customer lifetime value 200; "No information" is the best of targeting all accounts or none.

The order of the models is not stable between the two dates with loss history. On the end-2023 date the blend ranks first, though not clearly ahead of the runner-up ($z = 1.3$), and the scores fitted on the raw features are not ahead of size. On the end-2024 date, with two years of losses, gradient boosting ranks first (0.72; +0.12 over size, +0.05 to +0.19), clearly ahead of the blend (+0.10, $z = 3.7$) but not of the logistic regression on the raw features ($z = 1.6$). Selecting a model by the end-2023 ranking

would therefore have kept the blend, which was clearly behind a year later. The equal-weight combination came close to the first-ranked single score on both dates without knowing in advance which would lead (0.68 against 0.68, and 0.69 against 0.72), and pooled over the two dates it ranks loss better than the blend (+0.03, $z = 2.7$). Its difference from the blend was nil on the first of these dates (−0.003; −0.04 to +0.03) and positive on the second (+0.07; +0.04 to +0.10): the system's pre-specified model passed the model-choice check on the end-2023 date and failed it on the end-2024 date. Fitting the composite's weights to earlier outcomes does not help either: weights chosen to maximize the AUC on the earlier dates score below the defaults in the following year on both dates (0.61 against 0.66, and 0.55 against 0.56).

Random cross-validation within the year, the design common in the churn-prediction literature, gives a different picture (Figure 4a). On the end-2023 accounts it credits the learned scores with AUCs of 0.73 to 0.77, whereas the same scores fitted on earlier outcomes and tested on the following year reach 0.61 to 0.64, an overstatement of 0.09 to 0.13. On the end-2024 accounts the two designs agree within 0.03. On the first date, when no forward model could be fitted at all, random cross-validation would have reported about 0.75. On synthetic data whose generating process does not change between years, random cross-validation shows no such overstatement: it is below the forward result in six of eight comparisons and at most 0.03 above it (Online Resource 1). The gap is therefore a property of the real data rather than of the comparison, and a random split gives no warning of the year in which it misleads.

5.4 What the learned representation adds

The autoencoder's embedding carries information about loss, but only a projection learned from outcomes reads it out. Fifty random directions through the embedding rank next-year loss with AUCs centered on 0.5 on every date, whereas the learned projection beats 49 of them on the end-2023 date and all 50 on the end-2024 date (Figure 4b). The embedding is not a better representation than the inputs it compresses: on both dates with loss history, a logistic regression on the raw features and a projection of principal components rank loss as well as or slightly better than the embedding projection (Table 7), and across ten training seeds the embedding projection ranges from 0.59 to 0.67 and from 0.61 to 0.69. Its role in the system is as half of the blend, which is above the composite in every seed on both dates. At the scale of one firm's accounts, the autoencoder is therefore not necessary: a projection of principal components or of the raw features would fill the same role.

5.5 When the risk term pays: synthetic benchmark (RQ3)

Three replays of one firm show one regime of signal strength and of the relation between account size and loss. To characterize when weighting margin by risk captures more than a margin ranking, a synthetic benchmark generates CRM data with planted ground truth. Each synthetic account carries a latent deterioration propensity θ_i. Appointment counts, contact recency and margin trajectory are generated from θ_i with noise whose strength is set by a *signal* parameter, and loss in the following year is drawn from θ_i. The generator never uses the risk formula, so the composite's ability to

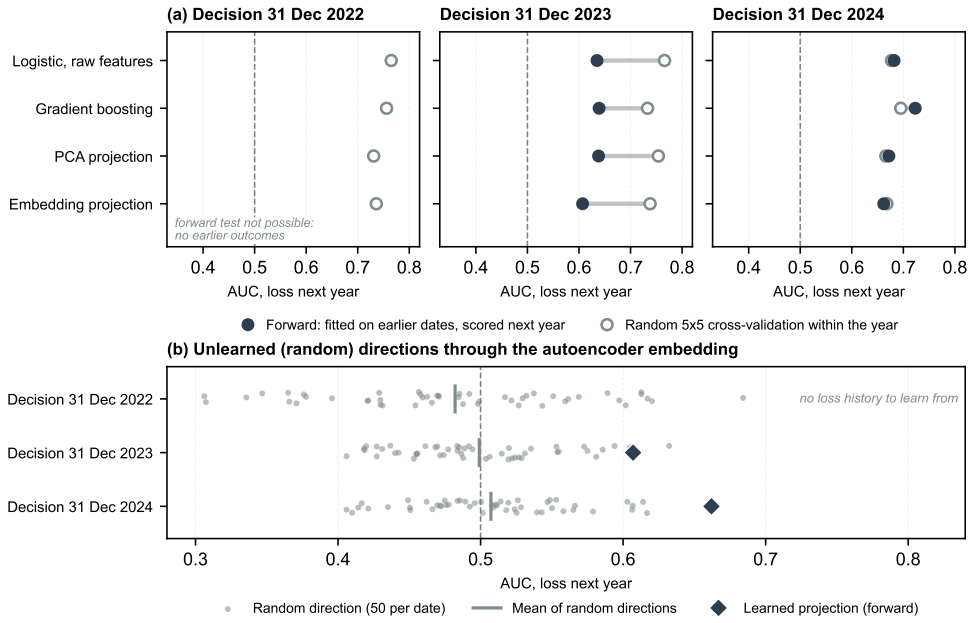


Fig. 4 (a) AUC of the learned scores under random 5×5 cross-validation within the decision year (open circles) and under the forward test, fitted on earlier dates and scored on the following year (filled circles). (b) AUC of 50 random directions through the autoencoder embedding (dots; vertical bar: mean) and of the projection learned from earlier outcomes (diamond).

recover θ_i is not built in. A second parameter sets the correlation between loss propensity and account size: negative as in the company data, where small accounts are lost more often, zero, or positive. Base loss rates of 10, 20 and 35% and twenty replications complete a grid of 720 datasets. Each is run through the unchanged pipeline at one decision date, with the composite as the risk term and lost margin measured, as in the replays, on the decision-year margin of the accounts lost. To see how good a risk term must be, the same datasets are also run with risk scores of known quality: the planted propensity plus random noise of increasing strength, whose AUC for the planted loss runs from 0.83 down to 0.57. Every list is compared with a margin ranking of its own size (generator, grid and full results in Online Resource 1).

Figure 5 summarizes the result. The composite's AUC for the planted loss rises from about 0.5 without signal to about 0.7 at full signal. With composite AUCs of 0.54 to 0.66 and size AUCs of 0.61 to 0.68, the company data correspond to the regime in which small accounts are lost more often and the signal is moderate. On lost accounts, the composite adds 3 to 4 percentage points over a margin ranking once the signal is informative and up to 10 points at full signal, in every size regime (panel a). On lost margin it adds nothing at this capacity: averaged over the base loss rates, the margin ranking captures more of the lost margin in every combination of signal strength and size relation, by 1 to 9 points (panel b). The planted propensity itself adds 2 to 8 points, although its AUC is only 0.82 to 0.84, because loss is drawn at random from it.

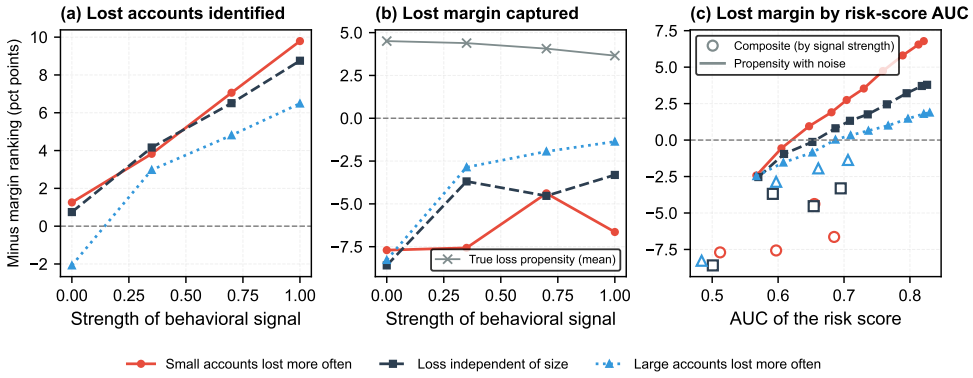


Fig. 5 Synthetic benchmark at 5,000 appointments: a risk-weighted list minus a margin ranking of the same size, in percentage points of (a) the accounts lost in the following year and (b) their decision-year margin, by strength of the behavioral signal, and (c) of the lost margin against the AUC of the risk score. Circles and solid line: small accounts lost more often; squares and dashed: loss independent of size; triangles and dotted: large accounts lost more often. In (a) and (b) the risk term is the composite; crosses in (b): the planted propensity itself (AUC 0.82 to 0.84), averaged over the size settings. In (c), filled markers and lines: the planted propensity with increasing noise; open markers: the composite at the four signal strengths. Each point averages three base loss rates and twenty replications.

Panel (c) shows how good a risk score must be. With noise unrelated to account size, the gain in lost margin turns positive at an AUC of about 0.62 when small accounts are lost more often, 0.66 when loss is independent of size and 0.69 when large accounts are lost more often. At 1,200 appointments the threshold is lower: where small accounts are lost more often, every tested score adds lost margin, and the composite does too at the two stronger signal levels, by 1 to 2 points. At 2,500 appointments it is no lower than at 5,000. At 5,000 appointments, however, the composite falls below the curves: at the same AUC it captures less lost margin, and at full signal it trails the margin ranking although its AUC (0.69 to 0.71) is above the thresholds. Neither its margin-weighted AUC nor its AUC among the largest accounts explains the gap (Online Resource 1). Ranking accuracy is therefore necessary but not sufficient for a risk term to protect margin; whether a given risk score does so has to be checked on the firm's own history.

5.6 Robustness

Four sets of checks leave these conclusions unchanged; Online Resource 1 reports them in full.

Formulation and solver. On the real decision date, the protection list overlaps with the list from the composite alone by 91%, with that from the embedding projection alone by 88%, with blend weights of 0.3 and 0.7 by 97%, and across five alternative weightings of the composite by 94–99%. An additive objective spreads the capacity over more accounts and depends on its weights. A floor on the risk term adds twelve

low-risk accounts holding €0.16M of margin and removes five holding €0.01M. Capacity, cap and λ change how many accounts are prioritized and how intensively, not their order. The closed form and SLSQP select the same list on eight of nine formulations, with objectives within 10^{-7}; on the revenue-weighted objective two accounts at the threshold differ.

As-of-date fields. Removing the account fields that exist only as of the export date moves the learned scores by at most 0.03, in both directions, raises the blend on both dates, and leaves the system's capture within two lost accounts. Customer class, which these runs keep, ranks loss less well than revenue on every date and carries no signal within revenue terciles (AUC 0.36 to 0.56), so it does not record later losses.

Training seeds. Across ten autoencoder seeds, the blend is above the composite on both dates in every seed, and the system's list holds more lost accounts than the margin ranking at 1,200 and 2,500 appointments in every seed.

Perturbed data. On an unchanged copy and seven perturbed copies of a synthetic dataset with three decision dates (shuffled rows; 20% of accounts or 30% of appointments removed; blank fields; outliers; a missing year of appointments; gaps in the yearly ledger), the chain runs to completion in every case. Under the first five perturbations, the negative control stays below chance and the protection list holds at least as many lost accounts as the margin ranking at 1,200 and 2,500 appointments in nearly every comparison, and never fewer by more than one account. A missing year of appointments brings the composite and the negative control close to chance on that date. Gaps in the ledger are the damaging case: a missing account-year is read as a complete loss, which corrupts the outcome and brings every score toward chance. A complete yearly ledger with explicit zero rows is therefore a data requirement.

6 Discussion

6.1 Answers to the research questions

RQ1. At the lower capacities, weighting margin by a risk term learned from earlier losses puts more of the accounts about to leave on the list than a margin, size or RFM ranking does, but it does not put more of the lost margin there. Lost margin follows account size, so any ranking that puts large accounts first captures most of it. This qualifies a finding of the value-based selection literature, in which forecasts of future value select better customers for growth than past value does (Kumar et al., 2008; Venkatesan & Kumar, 2004). For protecting the margin a firm already earns, the current size of an account is a hard benchmark to beat.

RQ2. Account size ranks next-year loss as well as or better than the behavioral composite, and learned scores add to it only once a loss history accumulates. The order of the learned scores changed between the two years, and the score ranked first in one year was clearly behind in the next, so a single model chosen in one year can be the wrong one in the next. An equal-weight combination was close to the best in both years, and comparing it with the chosen model on the decision day shows at least some of the recommendations that depend on that choice. The design of the evaluation matters as much as the model: a random split within the year overstated the learned scores in one of the two years, and nothing in the random split would have signaled it.

RQ3. The synthetic benchmark explains the pattern. A risk score adds lost margin over a margin ranking only when it ranks loss well: at the capacity used here, from an AUC of about 0.62 to 0.69 when its errors are unrelated to account size, and the behavioral composite falls short even where its AUC is higher. The replays show the same shortfall. In the company data small accounts are lost more often, where the threshold is about 0.62, and the system's risk term exceeded it on both dates with loss history (0.68 and 0.63), yet it found more of the departing accounts without capturing more of the lost margin. A gain in protected margin has to come from a better risk estimate rather than a different allocation rule, and because ranking accuracy is not enough, whether an estimate is good enough can be checked only by replaying the allocation on the firm's own history.

Table 8 sets these results against earlier findings. A single firm with two years of loss history can corroborate or qualify such findings, not establish them, and the table is worded accordingly.

6.2 Risk targeting versus uplift

The system targets risk, not the effect of attention. The accounts most likely to leave are not necessarily those whose loss attention would prevent (Ascarza et al., 2018), and targeting by the predicted effect of an intervention can yield more profitable campaigns than targeting by risk (Devriendt et al., 2021; Lemmens & Gupta, 2020), even though such models predict churn itself less accurately (Lemmens & Gupta, 2020). Estimating that effect requires variation in treatment that is unrelated to the outcome, which the company's records do not contain: account managers chose whom to meet, in part because they saw risk. The replays therefore show where losses occurred under the company's own sales activity, not what the recommended attention would have changed.

Operating the system would also change its own data. Once accounts on the list receive more attention, a loss model fitted on later years learns from outcomes that the list helped to shape (Devriendt et al., 2021; Gattermann-Itschert & Thonemann, 2022). A deployment should therefore record which accounts were recommended and which received attention, and keep a randomly chosen part of the list as a control.

Both points lead to the same next step: a randomized trial in which the system provides the assignment. A fresh protection list would be randomized within strata of intensity tier and risk, with the recommended attention for the treated accounts and business as usual for the controls, change in margin as the primary outcome with prior margin as a covariate, complete loss as a secondary outcome and compliance recorded, over twelve months. A list of about 270 accounts can detect only a large effect, so such a trial is a pilot that sizes a full one (design in Online Resource 1); Lemmens and Gupta (2020) likewise base their targeting on a randomized pilot on a small sample of customers. Field implementations evaluated against non-randomized comparison groups (Kumar et al., 2008) cannot separate the effect of attention from the selection that the system itself makes. Gattermann-Itschert and Thonemann (2022) randomized the selection rule in a B2B wholesaler: contacting the customers with the highest predicted churn reduced churn more than contacting randomly chosen ones, but the

Table 8 Earlier findings and what the decision-date evaluation shows

Earlier finding	Source	Here	Reading
Which model predicts churn best depends on the data context	Tamaddoni et al. (2016)	The order of the learned scores changed between the two years (blend first, then gradient boosting)	Consistent
An unweighted average of many models is close to the best	Lessmann et al. (2015)	The combination was close to the best score in both years (AUC 0.68 against 0.68, and 0.69 against 0.72)	Consistent
Laboratory experiments may overestimate the advantage of advanced classifiers	Lessmann et al. (2015)	Random cross-validation credited every learned score, gradient boosting included, with an advantage over size that the forward test did not show (end-2023 date)	Consistent, in one of two years
Churn models lose accuracy within a few periods; the best method stays the same	Risselada et al. (2010)	Next-year accuracy fell below the within-year estimate in one of two years; the order of the models changed between the two years	Consistent on accuracy; differs on the order of models
Little loss of accuracy about three months ahead	Neslin et al. (2006)	One year ahead, the forward test was 0.09 to 0.13 below random cross-validation in one year and level with it in the other	Qualified at a one-year horizon
A split within one period overstates the accuracy found on a later period	Gattermann-Itschert and Thonemann (2021)	Random cross-validation overstated next-year AUC by 0.09 to 0.13 on the end-2023 date and agreed within 0.03 on the end-2024 date	Consistent, in one of two years
Learned customer embeddings give small gains over hand-crafted features	Chamberlain et al. (2017)	The autoencoder was no better than raw features or principal components	Consistent
A well-prepared logistic regression matches more complex learners	Coussement et al. (2017); Tamaddoni Jahromi et al. (2014)	Level with gradient boosting on the end-2023 date, behind it on the end-2024 date	Holds in one of two years
A profit-based measure can change which model is chosen	Verbraken et al. (2013); Höppner et al. (2020)	Profit and AUC rankings agreed closely (rank correlation 0.84 to 1.00) and named the same best score, or two tied ones	Not reproduced for the best score
Forecast value selects customers better than current value	Venkatesan and Kumar (2004); Kumar et al. (2008)	For protection, margin and size rankings capture most of the lost margin	Different target: growth vs. protection
Customer value and churn are negatively related in B2B data	Rahman et al. (2025)	Size ranks next-year loss with AUCs of 0.61 to 0.68: small accounts are lost more often	Consistent
Accurate churn models need not yield profitable retention campaigns	Janssens et al. (2024); Lemmens and Gupta (2020)	The risk term found more leavers but not more lost margin; in the synthetic benchmark, AUC alone did not predict the margin gain	Consistent
RFM-type segmentation of B2B customers, without an outcome test	Lau and Wasesa (2025); Hosseini and Shabani (2015)	Tested against later losses, an RFM-type ranking performs about as well as a margin ranking	Supports RFM as a baseline

effect of the contacts on the called customers remained unclear. The trial sketched here randomizes the attention itself within the protection list.

6.3 Implications for practice

The decision splits into two lists that call for different actions. The protection list covers accounts that still buy, whose margin can still be lost; the win-back list covers accounts that have already stopped. Mixing them, as the system does without the eligibility rule, fills the protection capacity with accounts that can no longer be lost.

The win-back pool grows each year and few of its accounts return on their own (35%, 16% and 11% in the year after the three replay dates), so win-back effort should be selective and judged on its own outcomes.

The protection list values accounts by their margin alone. A firm may keep an account at a thin or even negative margin for reasons that margin does not record, such as its reputation or its value as a reference, and earn the margin from other accounts. In these data all but one small account with revenue in a year also had positive margin, so no buying account was excluded on these grounds, but accounts with large revenue at a thin margin rank low: one of the 50 largest accounts by revenue in 2025 earned a margin of under 3% and is not on the list of 27 October 2025. An additional trigger, for example large revenue at little or no margin, can flag such accounts for a management decision. An account the firm decides to keep for such reasons can be given a minimum number of appointments, which the closed form accommodates as a lower bound on that account, while the rest of the capacity is still allocated by margin at risk.

The risk ranking itself is a second output. Once a year of losses is available, the riskiest tenth of accounts leaves at about twice the base rate, which makes the ranking useful as a retention alert. The alerts point mostly to smaller accounts, whereas the protection list is driven by margin, so the two serve different purposes and do not compete for the same capacity. Accounts in their first year need attention of their own: they are lost more often than older accounts, yet a behavioral score reads their recent contact and rising activity as health (Section 5.3), and a trigger for new accounts, like the one for accounts kept for their reputation, can put them before an account manager.

A firm considering such a system should first ask whether a risk model is worth having. A margin or size ranking is a strong baseline for protecting margin, a behavioral composite needs at least a year of outcomes before it can be checked, and learned models need more. Where several models are available, an equal-weight combination is a safer default than any single model chosen on one year's results, and comparing its list with the chosen model's list on the decision day identifies accounts on which the choice matters. In this study 17 of the 278 accounts changed between the two lists of the same size, and they are the natural cases for an account manager's judgment.

Five steps follow for implementation. First, build the risk score from raw exports and check its sign against a negative control, because an inverted or double-counted score still produces a plausible-looking allocation. Second, replay the decision on the firm's own history at past decision dates, scoring each replay on the following year's losses, rather than evaluating models on random splits. Third, start with the composite and the size and margin baselines, and add learned scores only when the replays show that they help. Fourth, supply a complete yearly ledger with explicit zero rows, because a missing account-year is read as a loss. Fifth, record recommendations and attention and keep a control group, which is what turns the system from a ranking tool into a source of evidence about the effect of attention.

At the scale of one firm's accounts, an autoencoder is not needed for this. Its embedding contributes only through a projection learned from outcomes, and a projection of principal components or of the raw features serves equally well; random directions through the embedding, which need no outcomes, carry no information on

average. Because the allocation is available in closed form, the list can be re-planned at once when capacity changes, and the quantity that decides membership, the margin at risk $M_i E_i$, can be inspected account by account. The same closed form answers the reverse question, how many appointments a target share of the attainable objective requires, which supports capacity budgeting (Online Resource 1).

6.4 What one firm's history can show

The precision of these results is set less by the method than by the data that a B2B portfolio provides. Business customers are fewer in number and more valuable than consumers (Gordini & Veglio, 2017; Tamaddoni Jahromi et al., 2014), which makes both training a model and running a field experiment harder (Gattermann-Itschert & Thonemann, 2022), and their values differ more widely (Janssens et al., 2024). In the company studied, about a hundred accounts were lost in each outcome year. Outcomes counted in accounts can be judged on that basis, because every lost account counts once: the lead of the system's list over a margin ranking is clear at all three capacities once the two years with loss history are pooled (Section 5.2). Outcomes counted in margin cannot. Lost margin sits in a few large accounts, so its estimate moves with each of them. The differences in lost margin observed here are a few points at the lower capacities and close to nil at 5,000 appointments; detecting the former with the usual 80% power would take roughly 2,900 to 5,300 losses, thirteen to twenty-four times the 224 of the two years with loss history, or decades of this firm's history; separating the blend from account size by the AUC difference observed would take about 1,900 losses (Online Resource 1). These projections assume that the observed effects hold, and effects observed in small samples tend to be overstated, so the numbers are optimistic.

One firm's history can therefore show whether a risk term finds more of the accounts about to leave. Whether it protects more margin has to be judged on longer histories, on evidence pooled over several firms, or on synthetic benchmarks such as the one in Section 5.5. A firm that evaluates its own risk model should accordingly judge it on lost accounts, pool the evidence over its decision dates, and treat differences in lost margin of a few points as unresolved.

6.5 Limitations and future research

The replays show that the system's list contains the accounts that were subsequently lost; they do not show that the recommended attention would have prevented the losses. Past appointment intensity is itself a response to perceived risk, accounts that receive more attention differ from those that receive less in other ways, and competitive and economic conditions affect engagement and margin alike. The effect of attention, and with it the intensity rule of the allocation, which assumes diminishing returns rather than estimating them, can only be measured in the trial described above.

The evidence comes from one firm, three decision dates and two years of loss history; Section 6.4 sets out what that amount of data can and cannot resolve. The last outcome year is partial, and accounts that resumed buying late in 2025 are counted as lost. Because sales were exported by calendar year, past decisions could be replayed only at year-ends, not on the other days on which a plan might be revised, and on

the decision date itself current-year margin is year-to-date, so the answer depends on where in the year the decision falls. Some account fields exist only as of the export date; removing them does not change the conclusions, but their history could not be reconstructed. Engagement is measured by scheduled appointments only, without e-mails, calls or customer-initiated contacts. The profit-based measure uses the parameters of its telecommunications application and assumes a constant retention rate, an assumption that has been criticized (Devriendt et al., 2021). Three elements were added after the first replay run: the eligibility rule and the averaging of margin over the time as a customer, whose results are reported together with those without them, and the combination of models, which is used as a check of the pre-specified system rather than as its replacement.

Five extensions follow. The randomized trial would estimate the effect of attention and the response to its intensity. Repeating the replays as outcome years accumulate would show whether the choice of model settles, and whether learned scores move ahead of the size and margin baselines. With sales recorded by date, eligibility, margin, health and loss could all be measured over twelve-month windows before and after the decision date, so that the system answers alike on every day, and the replays could be run at every month-end; Online Resource 1 sets out this design and the data it needs. Profit-driven churn models, which take each customer's value into account when the model is fitted (Janssens et al., 2024; Rahman et al., 2025), would fit the risk term to the margin at risk rather than to the loss alone, and are the natural next comparator. Applying the protocol to other firms' histories, with richer interaction data than appointments, would test how far the findings travel beyond one equipment-rental company.

7 Conclusion

The paper began with a sales director's question: where should the account managers' attention go, and how much of it should each existing account receive, so that the margin the accounts bring is protected? The system answers with a protection list ranked by margin at risk and an allocation of appointments within a fixed capacity. Replaying that answer at three earlier year-ends, with only the information available on each date, showed how it would have fared against the losses that followed.

Weighting margin by a risk term learned from earlier losses put more of the accounts about to leave on the list than margin, size or RFM rankings did at the lower capacities, a lead that holds when the two years with loss history are pooled. It did not put more of the lost margin on the list, because lost margin follows account size and a margin ranking already covers most of it (RQ1). Account size proved a hard benchmark for ranking loss; the leading learned model changed between the two years, an equal-weight combination of models stayed close to the leading one, and random cross-validation within a year overstated next-year accuracy in one of the two years (RQ2). In the synthetic benchmark, a risk term protected more margin only above an AUC of about 0.62 to 0.69, and the behavioral composite fell short even where its AUC was higher (RQ3).

The answer to the sales director therefore has two parts. Margin decides most of where protective attention should go, because the accounts that carry the margin are those whose loss would cost most. A risk model earns its place by finding the departing accounts that a margin ranking passes over, and comparing its list with that of a combination of models on the decision day shows which accounts depend on the choice of model. For research, the study shows what one B2B firm's history can resolve: whether a risk term finds the accounts about to leave can be judged from a few years of outcomes, whether it protects more margin cannot, and a random split within a year can change which model appears best.

The study does not show that the recommended attention retains margin. A randomized trial, with the system's protection list providing the assignment, would answer that question, and replays repeated as outcome years accumulate would show whether the choice of model settles.

Supplementary information. Online Resource 1 is a replication guide with additional results: the definitions of every feature and score, including the negative control; the models, the allocation algorithm and the replay protocol in pseudo-code; capacity budgeting; the comparison of objective formulations and the solver check; the evaluation-design comparisons; the robustness checks (as-of-date fields, training seeds, perturbed data); the synthetic generator and benchmark grid; and the design of the randomized trial. Online Resource 2 is the replication package: the code of every analysis and a synthetic dataset with planted ground truth on which the whole chain runs.

Acknowledgements. The authors thank participating sales organizations for providing authentic CRM data. This research was part of the PATA co-research project (AI-enabled customer experiences), in cooperation with the University of Turku and the Haaga-Helia University of Applied Sciences, Finland.

Statements and Declarations

- **Funding**: Business Finland supported this co-research project. Funding number: 6684/31/2023. The funding institution had no role in study design, data collection and analysis, decision to publish, or manuscript preparation.
- **Competing interests**: The authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate**: Not applicable. The study analyzes anonymized business records of customer organizations; it involves no human participants and no personal data.
- **Consent for publication**: Not applicable.
- **Data availability**: The company's customer records were anonymized by the company before they were provided to the researchers under the consortium agreement of the research project. They are confidential and cannot be shared. Aggregate results supporting the findings are available from the corresponding author on reasonable request, subject to the company's approval. The synthetic dataset, on which every analysis of the paper runs, and its generator are included in Online Resource 2.
- **Materials availability**: Not applicable.

- **Code availability**: The code that produces every table and figure is provided as Online Resource 2 for peer review and will be released in a public repository with a persistent identifier on publication.
- **Author contribution**: **Joona Mäntyvaara**: Conceptualization, Writing - Original Draft, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Funding acquisition. **Jukka Heikkonen**: Validation, Writing - Review & Editing, Funding acquisition, Project administration, Supervision.

References

Ascarza, E., Neslin, S.A., Netzer, O., Anderson, Z., Fader, P.S., Gupta, S., ... others (2018). In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, *5*(1-2), 65–81, https://doi.org/10.1007/s40547-017-0080-0

Bengio, Y., Courville, A., Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828, https://doi.org/10.1109/TPAMI.2013.50

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044, https://doi.org/10.1287/mnsc.2018.3253

Bilro, R.G., Loureiro, S.M.C., Souto, P. (2023). A systematic review of customer behavior in business-to-business markets and agenda for future research. *Journal of Business & Industrial Marketing*, *38*(13), 122–142, https://doi.org/10.1108/JBIM-07-2022-0313

Bowman, D., & Narayandas, D. (2004). Linking customer management effort to customer profitability in business markets. *Journal of Marketing Research*, *41*(4), 433–447, https://doi.org/10.1509/jmkr.41.4.433.47011

Carvalho, A.M., Lopes, C., Cruz, M., Santos, J., Ramos, S., Vieira, F., Louro, P. (2025). Optimizing commercial teams and territory design using a mathematical model based on clients' values: A case study in Canada. *Mathematics*, *13*(18), 2940, https://doi.org/10.3390/math13182940

Chamberlain, B.P., Cardoso, A., Liu, C.B., Pagliari, R., Deisenroth, M.P. (2017). Customer lifetime value prediction using embeddings. *Proceedings of the 23rd*

acm sigkdd international conference on knowledge discovery and data mining (pp. 1753–1762).

Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583, https://doi.org/10.1016/0169-2070(89)90012-5

Coussement, K., Lessmann, S., Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36, https://doi.org/10.1016/j.dss.2016.11.007

De Caigny, A., Coussement, K., De Bock, K.W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772, https://doi.org/10.1016/j.ejor.2018.02.009

De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. *Industrial Marketing Management*, *99*, 28–39, https://doi.org/10.1016/j.indmarman.2021.10.001

Devriendt, F., Berrevoets, J., Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, *548*, 497–515, https://doi.org/10.1016/j.ins.2019.12.075

Gattermann-Itschert, T., & Thonemann, U.W. (2021). How training on multiple time slices improves performance in churn prediction. *European Journal of Operational Research*, *295*(2), 664–674, https://doi.org/10.1016/j.ejor.2021.05.035

Gattermann-Itschert, T., & Thonemann, U.W. (2022). Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests. *Industrial Marketing Management*, *107*, 134–147, https://doi.org/10.1016/j.indmarman.2022.09.023

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, *62*, 100–107, https://doi.org/10.1016/j.indmarman.2016.08.003

Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems 35 (NeurIPS 2022), datasets and benchmarks track* (pp. 507–520).

Hallikainen, H., Savimäki, E., Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, *86*, 90–98, https://doi.org/10.1016/j.indmarman.2019.12.005

Hand, D.J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1), 1–14, https://doi.org/10.1214/088342306000000060

Hinton, G.E., & Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507, https://doi.org/10.1126/science.1127647

Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., Verdonck, T. (2020). Profit driven decision trees for churn prediction. *European Journal of Operational Research*, *284*(3), 920–933, https://doi.org/10.1016/j.ejor.2018.11.072

Hosseini, M., & Shabani, M. (2015). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics*, *3*(3), 110–121, https://doi.org/10.1057/jma.2015.10

Jagabathula, S., Subramanian, L., Venkataraman, A. (2018). A model-based embedding technique for segmenting customers. *Operations Research*, *66*(5), 1247–1267, https://doi.org/10.1287/opre.2018.1739

Janssens, B., Bogaert, M., Bagué, A., Van den Poel, D. (2024). B2Boost: Instance-dependent profit-driven modelling of B2B churn. *Annals of Operations Research*, *341*(1), 267–293, https://doi.org/10.1007/s10479-022-04631-5

Karush, W. (2014). Minima of functions of several variables with inequalities as side conditions. G. Giorgi & T.H. Kjeldsen (Eds.), *Traces and emergence of nonlinear programming* (pp. 217–245). Basel: Birkhäuser. (Reprint of the author's M.Sc. thesis, Department of Mathematics, University of Chicago, 1939)

Kaufman, S., Rosset, S., Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 556–563).

Kraft, D. (1988). *A software package for sequential quadratic programming* (Tech. Rep.). Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht.

Kuhn, H.W., & Tucker, A.W. (1951). Nonlinear programming. J. Neyman (Ed.), *Proceedings of the second berkeley symposium on mathematical statistics and probability* (pp. 481–492). Berkeley, CA: University of California Press.

Kumar, V., Venkatesan, R., Bohling, T., Beckmann, D. (2008). Practice prize report—the power of CLV: Managing customer lifetime value at IBM. *Marketing Science*, *27*(4), 585–599, https://doi.org/10.1287/mksc.1070.0319

Lau, U., & Wasesa, M. (2025). Incorporating loyalty and periodicity into a data-driven RFM customer segmentation model: An empirical study in the B2B retail sector. *Ieee international conference on technology, informatics, management, engineering and environment (time-e)* (pp. 261–266).

Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science*, *39*(5), 956–973, https://doi.org/10.1287/mksc.2020.1229

Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136, https://doi.org/10.1016/j.ejor.2015.05.030

Lodish, L.M. (1971). CALLPLAN: An interactive salesman's call planning system. *Management Science*, *18*(4, Part II), P25–P40, https://doi.org/10.1287/mnsc.18.4.P25

Mancisidor, R.A., Kampffmeyer, M., Aas, K., Jenssen, R. (2021). Learning latent representations of bank customers with the Variational Autoencoder. *Expert Systems with Applications*, *164*, 114020, https://doi.org/10.1016/j.eswa.2020.114020

Marx, C., Calmon, F., Ustun, B. (2020). Predictive multiplicity in classification. *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 6765–6774).

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., Mason, C.H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211,

Palmatier, R.W., Dant, R.P., Grewal, D., Evans, K.R. (2008). Interfirm relational drivers of customer value. *Journal of Marketing*, *72*(4), 76–89, https://doi.org/10.1509/jmkg.72.4.076

Patriksson, M. (2008). A survey on the continuous nonlinear resource allocation problem. *European Journal of Operational Research*, *185*(1), 1–46, https://doi.org/10.1016/j.ejor.2006.12.006

Rahman, S., Janssens, B., Bogaert, M. (2025). Profit-driven pre-processing in B2B customer churn modeling using fairness techniques. *Journal of Business Research*, *189*, 115159, https://doi.org/10.1016/j.jbusres.2024.115159

Reinartz, W., Krafft, M., Hoyer, W.D. (2004). The customer relationship management process: Its measurement and impact on performance. *Journal of Marketing Research*, *41*(3), 293–305, https://doi.org/10.1509/jmkr.41.3.293.35991

Reinartz, W., Thomas, J.S., Kumar, V. (2005). Balancing acquisition and retention resources to maximize customer profitability. *Journal of Marketing*, *69*(1), 63–79, https://doi.org/10.1509/jmkg.69.1.63.55511

Risselada, H., Verhoef, P.C., Bijmolt, T.H.A. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198–208, https://doi.org/10.1016/j.intmar.2010.04.002

Rusdarmono, I., Sofwan, A., Batu, K.L. (2025). Predictive analytics for B2B customer retention: a machine learning random forest based on net promoter score model. *2025 12th international conference on information technology, computer, and electrical engineering (icitacee)* (pp. 1–6).

Ryals, L. (2005). Making customer relationship management work: The measurement and profitable management of customer relationships. *Journal of Marketing*, *69*(4), 252–261, https://doi.org/10.1509/jmkg.2005.69.4.252

Skiera, B., & Albers, S. (1998). COSTA: Contribution optimizing sales territory alignment. *Marketing Science*, *17*(3), 196–213, https://doi.org/10.1287/mksc

.17.3.196

Tamaddoni, A., Stakhovych, S., Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: A contingent perspective. *Journal of Service Research*, *19*(2), 123–141, https://doi.org/10.1177/1094670515616376

Tamaddoni Jahromi, A., Stakhovych, S., Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258–1268, https://doi.org/10.1016/j.indmarman.2014.06.016

Timmermann, A. (2006). Forecast combinations. G. Elliott, C.W.J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting* (Vol. 1, pp. 135–196). Elsevier.

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, *68*(4), 106–125, https://doi.org/10.1509/jmkg.68.4.106.42728

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229, https://doi.org/10.1016/j.ejor.2011.09.031

Verbraken, T., Verbeke, W., Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, *25*(5), 961–973, https://doi.org/10.1109/TKDE.2012.50

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272, https://doi.org/10.1038/s41592-019-0686-2

Wang, W., Huang, Y., Wang, Y., Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 496–503).

Watson-Daniels, J., Barocas, S., Hofman, J.M., Chouldechova, A. (2023). Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. *Proceedings of the 2023 acm conference on fairness, accountability,*

and transparency (pp. 297–311).

Zoltners, A.A., & Sinha, P. (1980). Integer programming models for sales resource allocation. *Management Science*, *26*(3), 242–260, https://doi.org/10.1287/mnsc.26.3.242

Zoltners, A.A., Sinha, P., Lorimer, S.E. (2004). *Sales force design for strategic advantage*. New York: Palgrave Macmillan.

Zoltners, A.A., Sinha, P., Lorimer, S.E. (2008). Sales force effectiveness: A framework for researchers and practitioners. *Journal of Personal Selling & Sales Management*, *28*(2), 115–131, https://doi.org/10.2753/PSS0885-3134280201