Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests

Preprint · October 2022

CITATION
1

READS
840

2 authors:

Theresa Gattermann-Itschert
University of Cologne
5 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Ulrich W. Thonemann
University of Cologne
134 PUBLICATIONS **2,711** CITATIONS
SEE PROFILE

Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests

Theresa Gattermann-Itschert, University of Cologne
Ulrich W. Thonemann, University of Cologne

October 6, 2022

Abstract

Customer churn prediction enables companies to target customers at risk with proactive retention measures. We develop a churn prediction model for a non-contractual business-to-business (B2B) wholesale setting and apply it in a field study. Our experiment shows that compared to random targeting, contacting the customers with the highest predicted churn probabilities reduces churn in the population significantly. We demonstrate that this also entails a positive financial impact in terms of revenue development.

In addition to validating B2B churn prediction and retention in the field, we contribute to the literature by identifying the most important features. On top of the common recency, frequency and monetary value features, we show that features specific to customer relationship management such as the recency of the last contact with a field representative are important. We provide a concept on how to integrate proactive churn management into operations by leveraging existing customer care processes.

1 Introduction

Building and fostering relationships with customers is a crucial success factor for companies. One aspect of customer relationship management (CRM) is proactive churn management to identify customers at risk of churning and taking appropriate retention measures. With recent advances in technology, machine learning classifiers can be used to effectively predict which customer are at risk of churning. By training on historical customer data, such models learn to identify patterns of customers who churn. A trained model is able to predict the future churn probability of current customers. This probability can be used to target proactive churn prevention measures.

There exists a solid body of literature on modeling choices (Glady et al. 2009, Miguéis et al. 2012), data pre-processing techniques (Crone et al. 2006, Coussement et al. 2017), and classification algorithms (Verbeke et al. 2012) for churn prediction. Models are usually evaluated by testing their capability to detect customer churn behavior on historical data and various literature exists on how to construct good churn classification models. However, literature on actual use of such prediction models to carry out targeted retention campaigns in practice is scarce and the results of only few field experiments have been reported so far (Table 1).

Table 1: Previous churn prediction studies using field experiments to analyze retention success.

Authors	**Industry**	**Relationship**	**Type**
Burez & Van den Poel (2007)	Pay-TV	Contractual	B2C
Ascarza et al. (2016, 2017), Ascarza (2018)	Telecommunications	Contractual	B2C
Ascarza (2018)	Professional memberships	Contractual	B2C
Godinho de Matos et al. (2018)	Telecommunications	Contractual	B2C
Ringbeck et al. (2019)	Retail	Non-contractual	B2C
Our Study	*Wholesale*	*Non-contractual*	*B2B*

Burez & Van den Poel (2007) use a churn prediction model to identify

the customers of a pay-TV company with the highest probability to churn. They test three retention measures in the field which all reduce the churn rate. Asking customers about their satisfaction achieves the largest effect. Ascarza et al. (2016, 2017) investigate proactive churn prevention in telecommunications. They do not use churn predictions but select customers based on other criteria. For example, they target customers that would benefit from changing their plan if their usage pattern continues (Ascarza et al. 2016) or customers whose accounts are suspended shortly (Ascarza et al. 2017). Ascarza (2018) conducts field experiments in the telecommunications sector and regarding professional memberships. She predicts which customers have the highest risk of churning and which customers have the highest sensitivity to the retention campaign, demonstrating that both groups do not necessarily coincide. Godinho de Matos et al. (2018) use knowledge about the customers' social network connections to develop proactive churn management for a telecommunications company and show that calling the friends of likely churners reduces churn. Ringbeck et al. (2019) investigate non-contractual relationships in retail. They conduct a large-scale field experiment with 400,000 customers, targeting the top 10% of customers with the highest predicted churn probabilities in the treatment group. They find that proactive churn management with coupons decreases the churn rate and increases revenue.

In the pay-TV and telecommunications industries, contractual relationships are typical whereas in retail, non-contractual relationships prevail. The two different types of relationships need to be distinguished as they influence how easily churn can be identified. Churn in contractual relationships occurs when the contract period ends or the customer notifies the company of his or her cancellation of the subscription. Therefore, both the intention to churn

and the actual churn date can be clearly identified and in most settings, the company has time to react and make an effort to retain customers before the relationship ends. Non-contractual relationships are not formally defined and therefore, churn is more difficult to identify as it does not go along with a contract cancellation but with a customer making less or no more purchases with the company. The time window to evaluate such changes in customer behavior varies with the expected purchasing frequency. Detecting churn in time to react is particularly challenging in non-contractual settings which makes developing churn prediction models more complex but also highly valuable for companies.

The studies mentioned above consider business-to-consumer (B2C) settings. Data analytics has been used extensively to support CRM in B2C settings and has received less attention in business-to-business (B2B) settings (Wiersema 2013) while there is a large potential to support managing customer relationships and gain competitive advantages (Hallikainen et al. 2020). Purchase volumes of B2B customers tend to be larger, hence maintaining a good relationship with each individual customer is even more important than in a B2C context (Rauyruen & Miller 2007). Studies on B2B churn such as Tamaddoni Jahromi et al. (2014), Chen et al. (2015) and Gordini & Veglio (2017) focus on building predictive models for companies and analyzing their expected impact on retention activities. They investigate which models and features are valuable for predicting B2B churn but do not conduct a field experiment like the other authors stated in Table 1.

Recently, the focus of churn studies is shifting from predicting the customers which have the highest risk to leave towards additionally considering which customers can most likely be persuaded to stay. De Caigny et al. (2021) are the first to move to prescriptive analytics for a B2B setting by

developing an uplift model with the goal of identifying the customers for whom retention activities have a positive impact. The value of uplift modeling lies in increasing the efficiency of retention activities. A drawback is that it requires data on how well customers respond to retention activities, which might not always be available.

In general, the literature on retention analytics in B2B settings is scarce (De Caigny et al. 2021). To our best knowledge, we are the first study to conduct a field experiment in a non-contractual B2B setting, providing practical evidence on the benefit of conducting retention activities based on predicted B2B customer churn probabilities. This might be due to the fact that compared to B2C settings, the number of customers in B2B settings is typically small (Lilien 2016), which makes training a churn prediction model as well as running a field experiment more challenging.

We investigate the B2B relationships of a European convenience wholesaler (company). The company sells beverages, tobacco, food, and other essential supplies to small convenience stores (customers). Out of all active customers, we consider the subset of approximately 5,000 customers who are not tied to master agreements of a chain, but manage their sourcing decisions individually. Customers can choose the wholesaler they order from and can churn without canceling a contract.

In our study, we develop a machine learning model that predicts customer churn with high precision based on features capturing customer characteristics and behavior. We identify the most important features. Besides known important indicators such as recency, frequency, and monetary value (RFM) (Buckinx & Van den Poel 2005, Fader et al. 2005), we find that features specific to the B2B setting are important, such as the time since the last contact with a field representative.

We evaluate model performance not only on past data, but validate it in a field study predicting future churners. We contact customers with the highest churn probabilities and compare the actual number of churns with that of an approach where customers are contacted randomly.

We find that targeting retention measures based on the model's predictions reduces overall churn. Contacting the customers with the highest predicted churn probabilities reduces the number of churners in the population compared to the control group (random targeting) significantly. The revenue development of customers in the treatment group is also higher than in the control group, demonstrating a positive financial impact for the company.

2 Case study

Our research was conducted at a large European wholesale company with sales revenue of over 12 billion EUR in 2018. The product segments include tobacco, beverages, food and deep-frozen products, electronic value (e.g. phone and gift cards) and other daily needs. Since the products require different storage temperatures, the company is also specialized in logistics. It aspires to function as a one-stop supplier, with customers getting all products in one delivery and within 24 hours. The market is highly competitive with alternative companies offering similar product palettes but other delivery concepts. When customers churn, they do not notify the company, since they are not bound to a contract. Therefore, churn is not straightforward to identify.

2.1 Customer churn

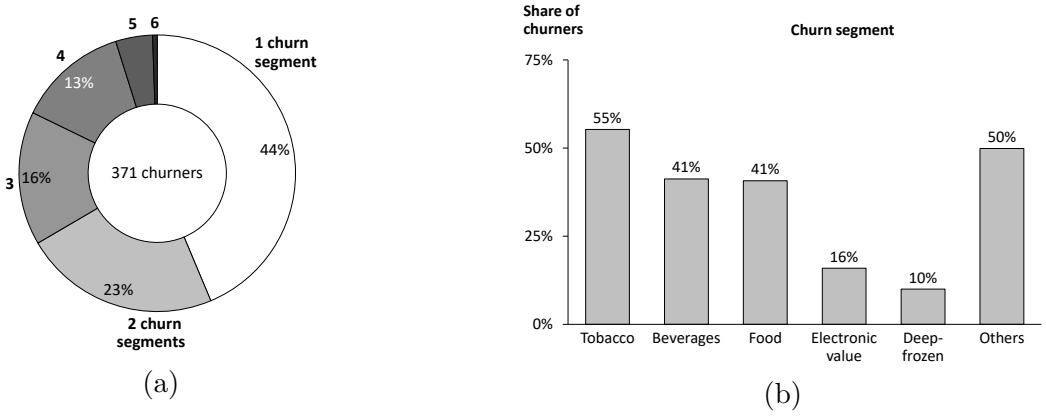
In the non-contractual B2B setting that we consider, customers do not necessarily stop abruptly to buy from the wholesale company, but gradually buy less or reduce the frequency of their orders. Typically, churning customers start buying products in a specific segment elsewhere and then switch more segments to the competition. In such settings, there is no specific end date of the customer relationship, which makes the churn definition particularly challenging. In other studies in non-contractual settings, a substantial decrease in purchase volumes, known as partial churn, is used to identify customer churn (Buckinx & Van den Poel 2005, Miguéis et al. 2012, 2013). Such a drop in sales is also the prediction focus of our study, since the wholesale company considers such a customer behavior as churn and grounds for retention measures.

We use the partial churn definition by the company that considers a customer to be a churner, if the revenue over three months fulfills two conditions:

1. In at least one segment, revenue decreased by 50% or more compared to the previous three months and

2. overall revenue decreased by 30% or more compared to the previous three months.

Figure 1 shows the product segments that play a role in identifying churn. The analysis is based on customer churn behavior from April to June, 2019. We focus on the subgroup of independent customers who can decide to buy from a different supplier at any time. We consider customers who bought regularly throughout the last year, with a mean interpurchase time of at most two weeks and at least one purchase per month. On April 1st, 2019, this applies to 4,952 customers, out of which 371 are considered churners.



(a) (b)

Figure 1: Churn segment analysis for April to June, 2019.

Figure 1 (a) shows that among the 371 churners, 44% have a decrease in revenue of 50% only in one segment. 23% of churners exhibit such a decrease in revenue in two of the product segments. 33% of churners have three or more churn segments. Figure 1 (b) shows that tobacco is the primary churn segment. Beverages, food, and others are also segments in which more than 40% of churners decrease their revenue substantially. The segments of electronic value and deep-frozen products play a minor role in determining churn.

2.2 Customer retention

Previous to our field study, the company did not use a churn prediction model. The sales department monitored the revenue development of customers and flagged customers with negative trend. These cases were discussed in monthly reviews, but there was no established process for dealing with churning customers.

In our study, we establish proactive churn management based on churn predictions. For retention actions, we rely on the customer communication

methods that already exist. Customer relationships are managed mainly by phone calls and visits. Phone agents are assigned to a region, managing a group of approximately 1,000 customers by taking orders and responding to complaints. Field representatives are assigned to a group of around 100 customers with short geographic distances to allow for routes with multiple visits a day. Field representatives maintain a close relationship to their customers, solving issues and advising on product portfolio and display of goods.

For retention measures, we introduce a two-staged process. A customer suspected to churn is called first and asked about the general satisfaction with the company. If the customer appears to be dissatisfied and the phone agent shares the churn suspicion, a visit is scheduled. By this procedure, unnecessary visits are avoided and only confirmed suspicions are dealt with in person.

3 Churn prediction model

black For selecting the best churn prediction model for our field experiment with the wholesale company, we benchmark three different algorithms (Section 3.1) and use an out-of-period model evaluation process (Section 3.2). We give an overview of the features used (Section 3.3) and the performance measures applied to evaluate the models (Section 3.4). We select the best model based on the prediction results (Section 3.5) and analyze feature importances (Section 3.6).

3.1 Model training and classifier selection

In Figure 2, the process of data collection, feature and label creation, model building, and model evaluation is shown. In the first step, we collect customer

master data and relevant transactional data from different sources within the company: invoicing, CRM activities, and delivery performance. This data is combined, aggregated and matched to individual customers. On customer level, we then derive informative features such as order frequency, revenue development, and recency of interaction with a field representative. Data from a 12-month period is taken into account to calculate the features. The corresponding churn label for each customer is derived from the purchase behavior during the subsequent 3 months according to our churn definition. In total, the required time span for creating a set of features and labels is 15 consecutive months.

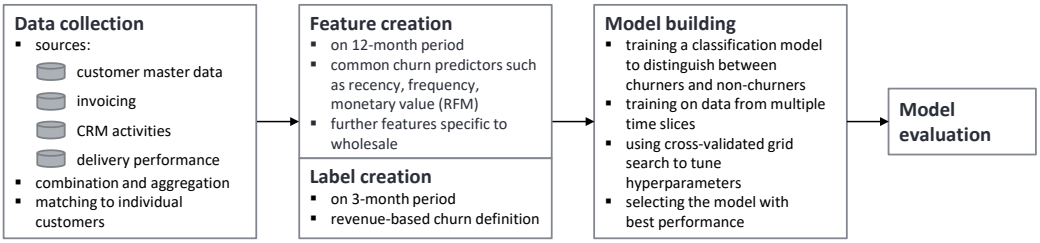


Figure 2: Process of model training and evaluation.

During model building, we train a churn prediction model which classifies customers as future churners or non-churners. For a comparison of different classification algorithms, we refer to Verbeke et al. (2012). The most popular choices in churn prediction studies are logistic regression, support vector machines (SVMs), and tree models, in particular random forests (De Caigny et al. 2018). We benchmark all three machine learning techniques and use the one with the highest prediction quality for our field experiment. The hyperparameters of each classification algorithm are tuned by conducting grid search. For an overview on grid search parameters of each algorithm see Appendix A.

Logistic regression is a common benchmark in churn prediction studies

(Coussement et al. 2017) because it is a well-known technique from statistics with low complexity. Samples are classified by fitting a linear model and we apply L2 regularization which penalizes the squared feature weights to avoid overfitting.

Support vector machines (SVMs) aim to maximize the distance of samples to the decision boundary (Vapnik 1995) so that the trained model can generalize well when classifying new samples. The algorithm has been the focus in several churn prediction studies (Coussement & Van den Poel 2008, Lessmann & Voß 2009, Chen et al. 2012), also in B2B settings (Gordini & Veglio 2017).

Random forests are frequently applied in churn prediction studies because of their robustness and low run times compared to other techniques while delivering good predictive results (Buckinx & Van den Poel 2005, Coussement & Van den Poel 2008, 2009, Burez & Van den Poel 2009). The algorithm classifies by taking the most frequent class prediction of a high number of decision trees which are built on bootstrap samples (Breiman 2001). Burez & Van den Poel (2009) recommend weighted random forests which perform even better than classic random forests in their setting. The technique deals with class imbalance by placing more weight on samples from the minority class (Chen et al. 2004). We therefore use balanced class weights to build a weighted random forest model and vary the maximum tree depth and the maximum number of features during grid search. Low values for both parameters decrease the risk of overfitting by reducing model complexity.

We report results for all three algorithms and select the best performing one. In other studies in the retail and wholesale field such as the retail churn prediction study by Ringbeck et al. (2019) who tested different algorithms (logistic regression, decision trees, SVMs, neural networks, boosted trees, and

random forests), weighted random forests delivered the best results. In the B2B wholesale industry, Gattermann-Itschert & Thonemann (2021) also find that weighted random forests outperform other classification techniques.

3.2 Model evaluation

For model evaluation, performance is assessed on a test set of unseen data. There are two ways to construct such a test set. Out-of-sample testing uses a subset of data from the same time period as hold-out data. Out-of-period testing uses data from a time period not yet seen during training and has gained popularity in churn prediction studies (Burez & Van den Poel 2008, Coussement & De Bock 2013, Óskarsdóttir et al. 2017, 2018). It reflects the inherent time shift between building a model and applying the model in practice. Since we conduct a field experiment using our trained model, this is particularly relevant. We therefore evaluate the model with out-of-period testing.

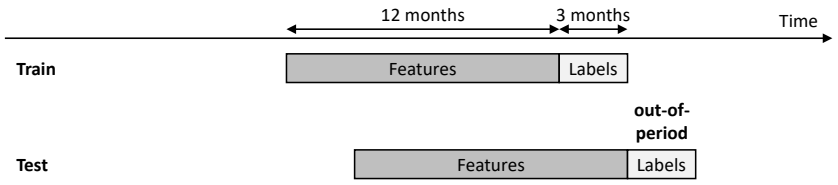


Figure 3: Training and testing time slices for out-of-period testing.

We offset the testing period by three months to the future so that the test label period has not been used to train the model (see Figure 3). For this technique, two time slices of data are required, one for training and one for testing. A time slice refers to data from a certain time window. Transactional data is organized relative to a reference point within that window in order to compute the corresponding features and labels.

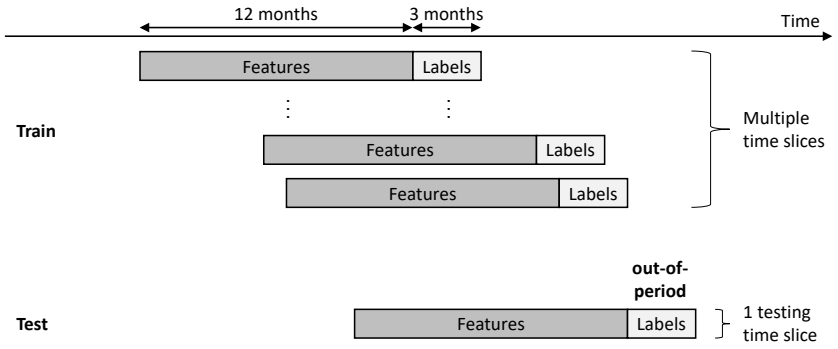


Figure 4: Time slices for multi-slicing (adapted from Gattermann-Itschert & Thonemann (2021)).

Gür Ali & Arıtürk (2014) and Gattermann-Itschert & Thonemann (2021) recommend to use more than one time slice of data as training data. Figure 4 illustrates the concept of using multiple time slices for training which is referred to as multi-slicing. In our study, we shift each time slice by one month and apply multi-slicing, an approach that enhances performance compared to using only one time slice.

3.3 Feature engineering

For constructing informative features, we rely both on results from the literature and on experience from the company business experts. For an overview on the features created, see Table 2. In the following, we focus on selected features from the first two groups, highlighting the established features and features we created specifically for the B2B case.

3.3.1 Recency, frequency and monetary value

Several studies in the churn prediction B2C domain suggest using features regarding recency, frequency, and monetary value (RFM) (Buckinx & Van den Poel 2005, Bose & Chen 2009, Miguéis et al. 2013). Tamaddoni Jahromi

Table 2: Features used in this study.

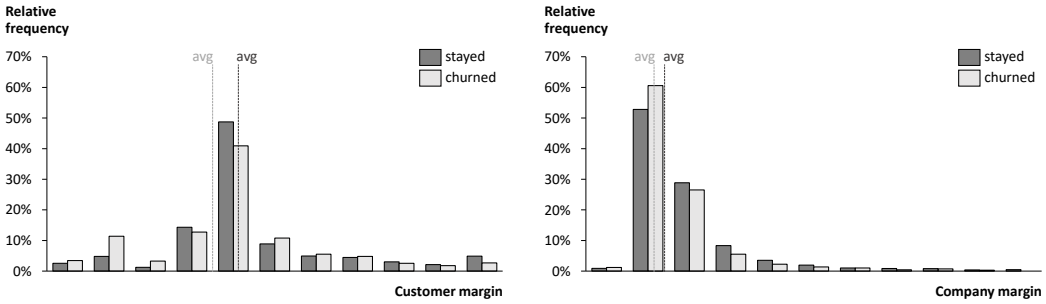
Type	Feature category	Feature variations	Source
Recency, frequency and monetary value			
Recency	Interpurchase Time (IPT)	Last IPT / mean IPT	Invoice data
		Mean, std, coefficient of variation (CV), max	Invoice data
Frequency	Number of days ordered	Monthly and quarterly mean, std, CV, per segment	Invoice data
		X-to-last month / monthly mean, last 1-3 weeks	Invoice data
		X-to-last quarter / quarterly mean, per segment	Invoice data
Monetary value	Revenue	Monthly and quarterly mean, std, CV, per segment	Invoice data
		X-to-last month / monthly mean	Invoice data
		X-to-last quarter / quarterly mean, per segment	Invoice data
Features specific for B2B relationships			
Customer margin	Margin	recommended selling prices - wholesale prices	Invoice data
Company margin	Margin	wholesale prices - internal purchase prices	Invoice data
Delivery performance	Number of days with missing quantities	total, monthly mean, last and second last month	Invoice data
	Product value not delivered	total, monthly mean, last and second last month, ratio to revenue	Invoice data
	Number of days with delivery issues	total, early, late, monthly means, ratio to transactions	Invoice data, delivery tracking
	On-time in-full (OTIF)	total, last, second last and third last month	Invoice data, delivery tracking
CRM	Number of contacts	with sales representative in person, with phone agent	CRM activity log
	Days since last contact	with sales representative in person, with phone agent	CRM activity log
Other			
	Number of segments ordered from	Monthly mean, std, CV, last month / mean	Invoice data
	Credit / Debit / Returns	total amount / revenue, total days / days ordered, days since last	Invoice data
	Region	first two digits of postal code	Customer master data
	Basic price level	conditions of framework agreement	Customer master data
	Type of store	categorical: convenience shop, beverage shop, tobacco shop, etc.	Customer master data
	Length of relationship	since customer account creation	Customer master data
	Length of framework agreement	since latest renewal	Customer master data
	Payment method	categorical: direct debit, cash etc.	Customer master data
	Assigned warehouse	different facilities	Customer master data

et al. (2014) confirm this in the B2B context when predicting churn for business customers of an online fast moving consumer goods retailer. They validate that the frequency of purchases and the time since the last purchase are important.

For calculating RFM features, we use 12 months of transactional data from invoicing. We include the recency of the last purchase relative to the mean interpurchase time (IPT), the frequency of orders and its coefficient of variation. Mean and coefficient of variation of monthly and quarterly revenue as well as revenue development in the recent months are computed. Quarterly revenue is also evaluated per product segment.

3.3.2 Features specific for B2B relationships

We include further features that the company believes to be good predictors of churn. Figures 5-7 show the histograms of selected industry specific features for both groups of customers (stayed and churned).



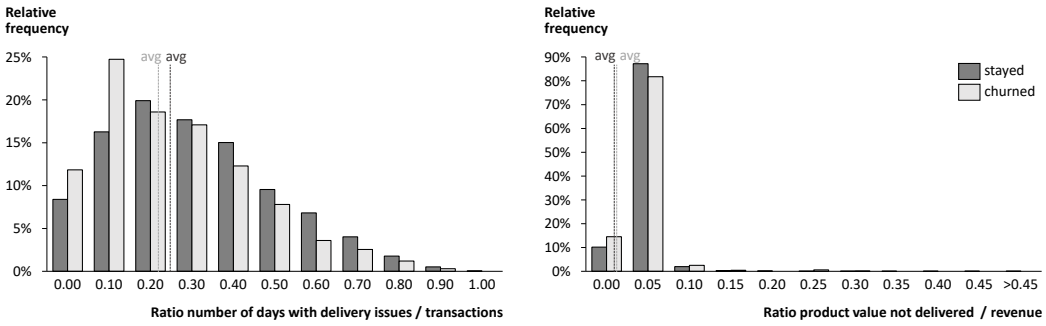
(a) (b)

Figure 5: Histogram of relative frequency for margins of company and customer (stayed/churned).

Customer margin Company experts advised to take the difference between recommended selling prices and actual wholesale prices as an indicator for the customer margins. Due to confidentiality issues, we cannot disclose the actual margins achieved, but can report the general distribution (Figure 5 a). The average margin among churners is lower than that of non-churners. As expected by the employees of the company, customers with low margins are more likely to churn and customers with high margins are more likely to stay.

Company margin The company margin can be approximated by taking the difference between wholesale prices and purchasing prices. The company hypothesized that customers with whom the company makes a high margin are more likely to churn. However, Figure 5 (b) shows that the average company margin is in fact lower for churners than for non-churners. A possible explanation is that customers with whom the company earns a good profit margin are also those that are more loyal to the company. The relative frequency of churners is especially high in the low margin categories.

Delivery performance In previous company surveys, customers reported delivery performance as a factor of their satisfaction. There is insufficient evidence from literature on the importance of features capturing delivery performance. For a large Taiwanese B2B logistics company, Chen et al. (2015) explore various features regarding delivery performance in addition to RFM, longevity and profit variables, but only one delivery feature is important enough to be included in the final model: number of delivery failures divided by total transactions. They conclude that delivery problems only play a minor role in their case because the company's service level is very high in general. In our case, the wholesale company offers complex multi-temperature logistics services and suspected delivery performance to have an effect on churn behavior. The quantity and product value ordered but not



(a) (b)

Figure 6: Histogram of relative frequency for delivery performance features (stayed/churned).

delivered can be extracted from the invoicing data source. We also evaluate tracking data to determine whether the communicated delivery time window was missed. As a combined measure, we compute the ratio of number of days with delivery issues divided by the total transactions. Figure 6 (a) shows that the averages are close but lower for churners, contrary to what the company expected.

In line with expectations is the slightly higher average in the ratio of product value not delivered divided by revenue, as shown in Figure 6 (b). Judging from these descriptive statistics, the informative value of the delivery features seems limited but we still include them in the model for potential interaction effects with other variables.

CRM For a Belgian newspaper company, Coussement & Van den Poel (2008) showed that CRM variables such as number and recency of interactions are important when predicting churn. Close relationships with customers are particularly relevant in B2B settings, so we analyze the activity log of both field representatives and phone agents to determine the number of contacts in the past 12 months and the days since the last contact (set to 365 if no contact was recorded in the past year).

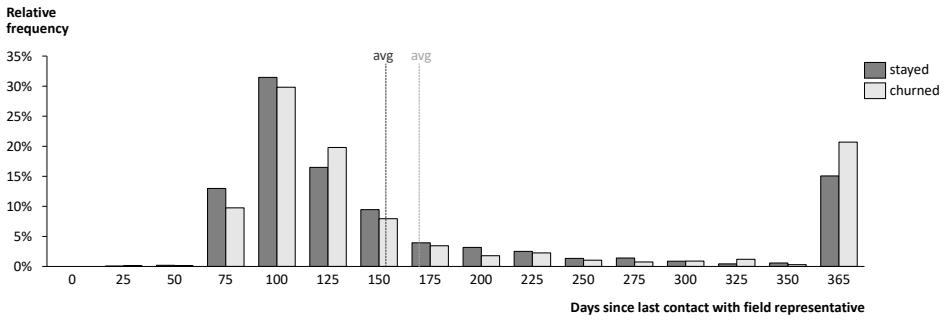


Figure 7: Histogram of relative frequency for days since last contact with field representative (stayed/churned).

Figure 7 visualizes that for the customers not contacted for 12 months, the relative frequency of churners is higher. For customers with more recent contacts ($\leq$ 100 days ago), the relative frequency is higher for staying than for churning customers.

3.4 Performance measures

Classification performance of the churn prediction model can be evaluated with measures that are based on the confusion matrix. The confusion matrix shows how many customers predicted to be churners turned out to be actually churners (true positives: TP) or were non-churners (false positives: FP) and how many of those predicted to be non-churners were correctly classified (true negatives: TN) or were undetected churners (false negatives: FN). The distribution of customers into the four fields in the confusion matrix depends on the decision threshold. Customers are assigned to the positive class if their predicted class probability is above this threshold.

3.4.1 Threshold-dependent measures

In practice, only one specific model with a set decision threshold can be applied. When choosing how many customers to target and thereby setting the threshold, two curves can inform the decision by indicating the model performance for different scenarios. Most common is the receiver operating characteristic (ROC) curve which displays the trade-off between the true positive rate $\left(\text{TPR} = \frac{TP}{TP+FN}\right)$ and the false positive rate $\left(\text{FPR} = \frac{FP}{FP+TN}\right)$. For situations such as ours with high class imbalance, the alternative precision-recall curve also has benefits since it provides more detail than the ROC curve, especially regarding the performance among top-ranked samples (Saito & Rehmsmeier 2015). Precision $\left(\frac{TP}{TP+FP}\right)$ indicates the fraction of predicted churners that are actually churners. Recall $\left(\frac{TP}{TP+FN}\right)$ is the fraction of actual churners that are correctly predicted as such. The precision-recall curve shows that finding more churners (increasing recall) by lowering the decision threshold comes with less confidence about churn predictions (decreasing precision).

3.4.2 Threshold-independent measures

To measure general model capability regardless of a specific decision threshold, performance can be summarized by the areas under the curves. Hence, we report AUC as the area under the ROC curve and average precision (AP) as the area under the precision-recall curve.

3.4.3 Domain-specific measures

In the churn prediction domain, a typical approach is to select the customers with the highest churn probabilities for retention activities. A popular measure is the top-decile lift (TDL) (Coussement et al. 2017, De Bock & Van Den Poel 2012, De Caigny et al. 2018, Verbeke et al. 2012), which captures how much better a model identifies churners in the top decile compared to randomly targeting 10% of the customers. It is calculated by dividing precision among the top 10% by the prevalence.

3.5 Prediction results

We compare performance of random forests, logistic regression and SVMs on test data which exhibits a medium churn rate of 10%. Out of 5,155 customers, 512 are churners. Table 3 reports performance of the algorithms regarding AUC, AP and TDL measures. We select the random forest model as it performs best regarding all three measures. In the following, we report more detailed results on the prediction performance specifically for this model.

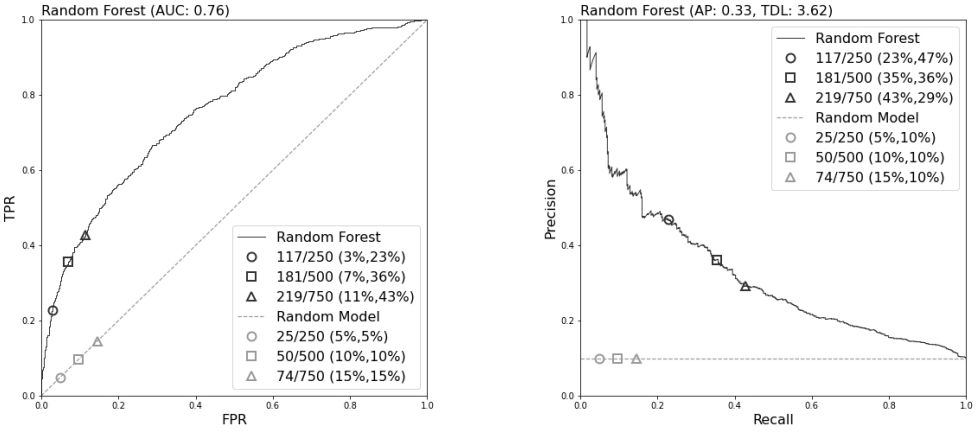
We evaluate our random forest model by examining the ROC curve and the precision-recall curve. Since the wholesale company only has limited resources to spend on retention activities, different scenarios of contacting 250, 500 or 750 customers with the highest churn probabilities are analyzed.

Rank	Model	AUC	TDL	AP
1	Random Forest	0.758	3.625	0.331
2	Logistic Regression	0.739	3.423	0.307
3	SVM linear	0.736	3.323	0.290
Baseline	Random classification	0.500	1.000	0.099

Table 3: Performance of different ML classifiers compared to random baseline

Figure 8 was used in discussions with CRM managers to visualize the trade-off between

- How many churners are found and can potentially be retained?
- How many customers are incorrectly predicted as churners resulting in wasted resources?
- How precise are the predictions determining the effectiveness of retention activities?



(a) ROC curve

(b) Precision-recall curve

Figure 8: Model performance across thresholds and results for classifying 250, 500 or 750 customers as churners. Details in legend: # correct churn classifications / # churn classifications (x-axis%, y-axis%).

Figure 8 (a) shows that the random forest model with an AUC score of 0.76 is much more capable than a random model (AUC 0.5). The marks represent model performance at the threshold levels corresponding to predicting 250 (circle), 500 (square) and 750 (triangle) customers as churners. The corresponding marks on the dashed line show how well a random model would perform. When contacting 5% of customers (250), then by chance also 5% of churners are found. With the random model, the TPR and FPR are always as high as the fraction of customers contacted. With the random forest model, a much larger fraction of customers can be found when contacting 250 customers (TPR 23%) while the FPR is kept very low with 3%. By increasing the number of customers contacted to 750, the fraction of churners found is raised to 43%, while still maintaining a lower FPR of 11% compared to the random model with 15%.

The AP of the random forest model is 33%. The TDL indicates that the random forest model is 3.52 times more precise than a random model among the top decile. Figure 8 (b) allows a closer look at the precision that goes along with different recall levels. When contacting 250 customers, a moderate recall level of 23% is achieved while precision is very high with almost every second churn prediction being true. Increasing the number to 750 targeted customers decreases precision to 29% but enables finding 219 of the 512 churners (43% recall).

CRM managers from the company opted for this scenario which they considered a good trade-off between the effort to target 750 customers and the potential to identify up to 43% of the churners.

3.6 Feature importances

Next, we evaluate the importance of the features included in the random forest model. For an overview on model interpretability and different methods, we refer to Molnar (2020). We apply the SHAP (SHapley Additive exPlanations) method by Lundberg & Lee (2017). The method provides insights both on a local level (per sample) and on a global level (across samples). With SHAP, individual predictions can be explained by evaluating the impact of each feature value on the model output.

Overall feature importance can be evaluated with a summary plot (Figure 9), where each point represents the SHAP value for one specific feature and sample, with the color indicating the feature value (the higher the darker). The features are ordered by descending importance.

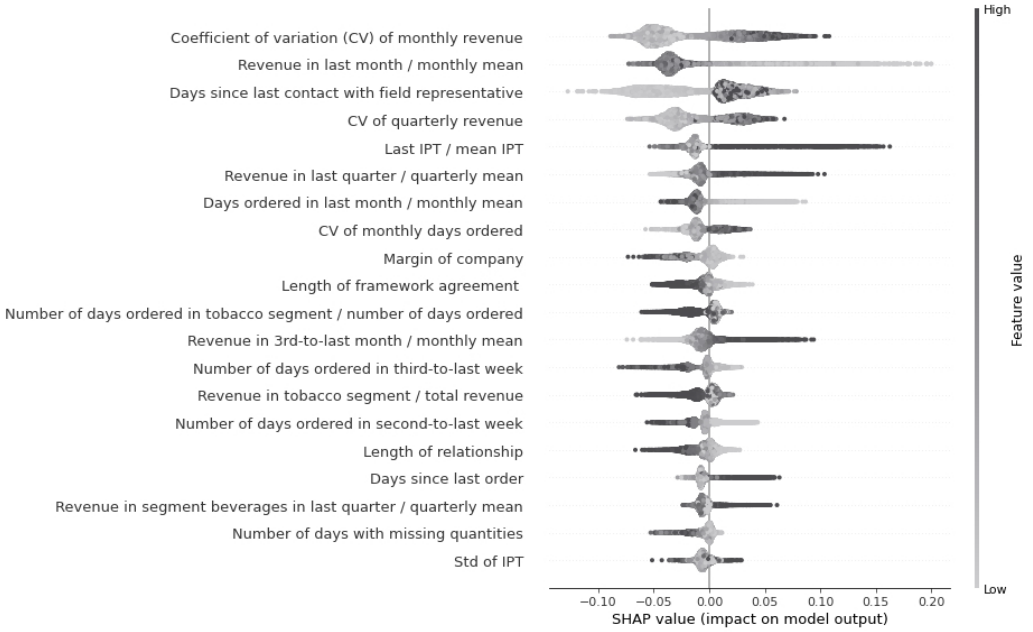


Figure 9: SHAP summary plot to evaluate global importance of features.

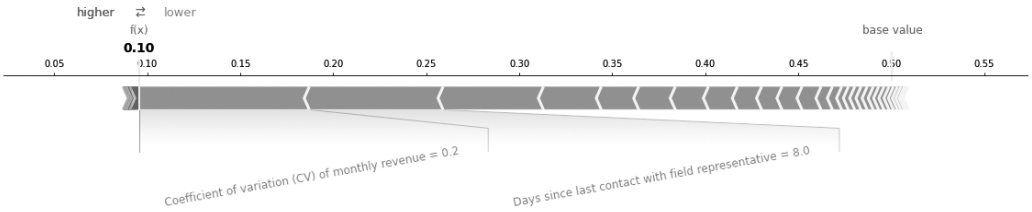
Figure 9 shows that features from various categories are relevant. The three most important features are the coefficient of variation (CV) of monthly

revenue, the development of revenue in the last month compared to the monthly mean, and the days since the last contact with a field representative. Also important are the CV of quarterly revenue, the recency of the last transaction (last IPT/ mean IPT), and the development of revenues (last quarter / quarterly mean). The ranking is in line with findings from literature that RFM features are important churn indicators. In addition, it shows that variables capturing CRM performance such as the third-ranking feature are important. In B2B, close relationships with customers are important and directly link to whether a customer stays with the company or not. Also, the margin of the company and the length of the framework agreement (time since last renewal of basic conditions with the customer) contribute to the predictive performance. Features regarding the delivery performance such as number of days with missing quantities are of marginal importance. Several variables regarding the revenue and order frequency in the most important segment tobacco are important.

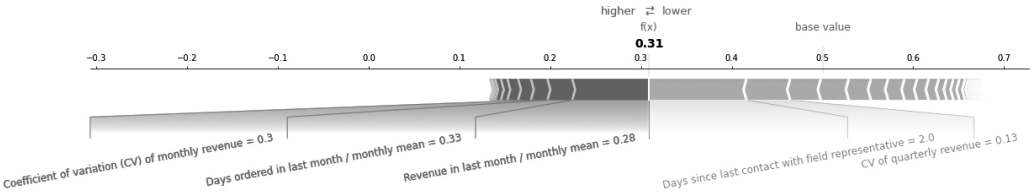
For individual samples, force plots visualize which features have a large impact on the prediction and whether the feature values increase or decrease the prediction value. For the CRM department and sales representatives, force plots can provide valuable insights on why a certain customer has a low or high predicted churn probability. In Figure 10, we evaluate three examples of customers with a low (a), medium (b), or high churn probability (c).

The first example shows that a low coefficient of variation of monthly revenue and only eight days since the last contact with a field representative lead to a low predicted churn probability of 0.10 for the specific customer. In the second example, the customer has a medium predicted churn probability of 0.31 because his or her feature values have both increasing and decreasing effects. Low revenue and a low number of days ordered in the last month

(a) Customer with a low predicted churn probability.



(b) Customer with a medium predicted churn probability.



(c) Customer with a high predicted churn probability.

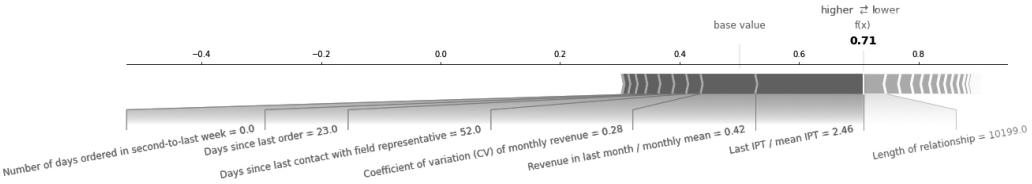


Figure 10: SHAP force plots for three different customers.

compared to the mean are indicators of an increased risk to churn. At the same time, the last contact with a field representative has been two days ago and the coefficient of variation of quarterly revenue is very low. In the third example, the churn probability is 0.71 due to many features indicating an increased risk to churn. Revenues have dropped and the last contact was over seven weeks ago. Only the length of relationship points to a very loyal customer. Based on this information, a sales representative can contact the customer in order to maintain the relationship.

4 Experimental setting

Training a churn prediction model is only the first step in proactive customer retention activity. A good model is a requirement, but no guarantee for successfully retaining customers. To analyze the effectiveness of basing retention actions on churn predictions, we conducted a field experiment with the wholesale company. Our goal was to answer the following questions:

1. Can good predictive quality be maintained when applying the model in practice?

2. Do retention measures based on churn predictions lower the overall churn rate?

3. Do the measures positively influence the revenue development of customers?

We set up a customer retention program integrating established customer relationship management processes. The program was designed to leverage the work of two groups of employees already assigned to customer care and support: Phone agents and field representatives.

Figure 11 shows the steps in this process: Input data is updated with the latest transactional data to determine the relevant group of currently active and loyal customers and to calculate their features based on data from the last 12 months. This data is then fed into the developed model, which has been trained before. The output is a prediction of customer churn behavior for the next three months.

Customers are ranked by descending churn probability and the top ranked customers are selected. These selected customers are then called by phone

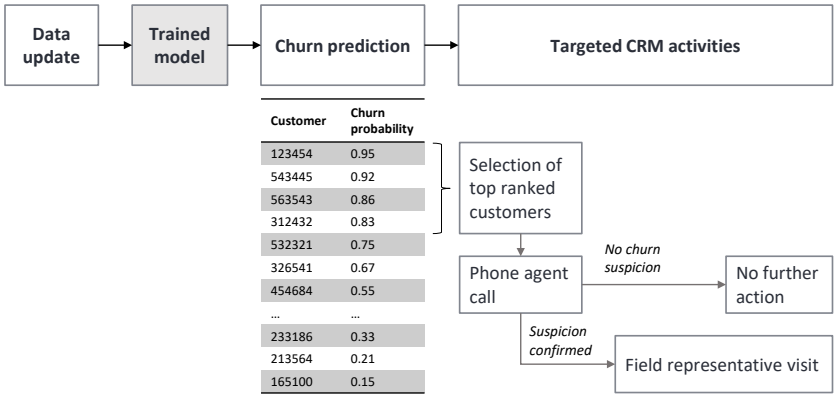


Figure 11: Customer retention program integrating the churn prediction model with customer support units.

agents. If the agents confirm the churn suspicion, they initiate a field representative visit to the customer. The subsequent measures address individual customer needs. They range from resolving issues with product shelf life or service to consulting on product range. Both the phone agent call and the field representative visit also aim to generally strengthen the relationship by increased interaction.

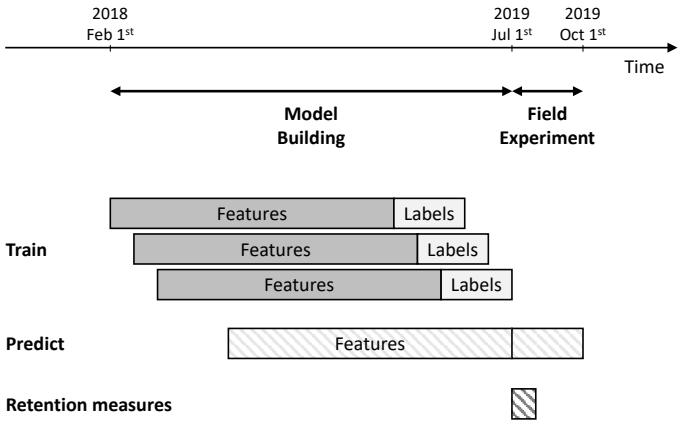


Figure 12: Model building and field experiment timeline.

In Figure 12, the timeline of the field experiment is displayed. Almost one and a half year of historical data from February 1st, 2018 to June 30th,

2019 are available for model training. With the time span of data available at the time of our field experiment, the training set can include up to three time slices.

The field experiment took place during July, August and September, 2019. For predicting the churn behavior during these months, the latest trained model was applied. It was fed with new input data by computing features on the last 12 months (July 2018 until end of June 2019). Churn predictions were available at the beginning of July, followed by retention measures. The focus was on the first weeks of July with potential consecutive visits. After a period of three months, beginning October 1st, data on the customer transactions during July, August and September 2019 could be extracted to analyze the impact of our field experiment on churn behavior.

5 Experimental design

Our objective is analyzing the value of using a machine learning (ML) churn prediction model to target retention efforts. Customers are split randomly into two groups: A treatment group, in which *ML-based selection* is used to determine the customers to call by decreasing predicted churn probability and a control group, in which the same fraction of customers is called, but with *random selection*. Hence, the treatment groups do not differ by whether or not customers are called or by the fraction of customers that are called, but by the method that is used to select the customers to call. This allows to assess the value of developing and using a churn prediction model.

On July 1st, 2019, in total 4,920 customers from the subgroup of independent customers met the criteria of being active and loyal and hence entered the field experiment. The company agreed to investing resources for con-

tacting 750 customers, which is a fraction of 15.2% customers of the dataset. One third of customers were assigned to the random selection treatment and two thirds were assigned to the ML-based selection treatment.

Customers were not aware that they take part in an experiment as calls were disguised as customer satisfaction surveys. To ensure proper randomization, the assignment of customers to the treatment groups was conducted without company involvement.

In both treatments the same fraction of customers was selected, resulting in 250 and 500 calls respectively. In order to prevent biases, the full list of 750 customers to call was provided to the company without the phone agents knowing whether a customer was selected randomly or based on their churn prediction score. Based on the call list, the two-step targeted CRM activities took place.

Table 4: Descriptive statistics on retention measures.

	Control Random selection	Treatment ML-based selection
Number of customers	1,640	3,280
Number of calls	250	500
Number of visits	62	132

6 Results and discussion

Table 4 gives an overview on the number of calls per group and the number of visits triggered by confirmed churn suspicions.

After a three months period, customer data was gathered to determine churners. Table 5 reports the churn rate for called and not called customers and for the full group. There is a difference between the control and the

Table 5: Churn rate.

	Control Random selection	Treatment ML-based selection
Called	14.00%	28.40%
Not called	12.73%	8.53%
Full group	12.93%	11.55%

treatment group regarding the churn rates among called and not called customers. In the control, the churn rates in the subgroups are on a similar level while in the treatment, the churn rate among called customers is more than three times higher than among those not called. This suggests that the ML-based selection is capable of detecting churners. As a result, the churn rate among not called customers is reduced to 8.53% compared to 12.73% in the corresponding control group.

In the control group, the churn rate among randomly called customers is slightly higher than among those who did not receive a call. A possible explanation is that calling customers could trigger churn because customers become aware of their usage or satisfaction level, as in Ascarza et al. (2016). However, the difference is not significant. Since the group of called customers is much smaller than the group of called customers, its churn rate is more sensitive to a randomly elevated number of absolute churners.

In the treatment, the churn rate among called customers is 28.40%. This is a sign for high precision of the ML-based selection but the effect of the retention activities is unclear. For evaluating how well the retention measures succeed at reducing churn, we must compare the churn rate in the entire population. Looking at the full groups, the difference in churn rates between the random selection and ML-based selection is 1.38 percentage points. This can be interpreted as an overall baseline uplift of 1.38% when targeting a

15.2% proportion of customers with ML-based retention measures instead of random targeting. The churn rate difference between treatment and control translates to a 10.67% reduction from 12.93% to 11.55% in the ML-based selection group.

6.1 Churn prediction

We next analyze the churn prediction quality of the model. A model which shows good performance on historical data does not automatically perform well when applied in practice.

Since retention measures were initiated based on the churn predictions, the predictive performance of the model is difficult to assess for the entire set of customers considered in the experiment. If retention measures are successful, customers are non-churners at the end of the field experiment even if they were correctly predicted as would-be churners in the beginning.

We therefore evaluate model performance with two approaches. First, we calculate predictive performance on the group of customers that is not affected by retention measures: the customers in the control group that have not been called. Second, we estimate the relationship between predicted probability to churn and actual churn with a regression model.

With the first approach, we consider the 1,390 customers that have been randomly selected to not be targeted with retention measures in the control group. Out of these, 177 customers churned. We then consider how well the prediction model would have performed on this control subgroup. We rank customers in descending order by their ex ante predicted churn probabilities and identify the 212 customers that would have been targeted (applying the share of 15.2% targeted customers from the overall experiment). Among these customers, 64 are true positives, which results in a precision of 30.2%

and a recall of 36.2%. Recall is lower than the value obtained on historical data, but the model is successful in delivering the same level of precision. Predictive performance is good, considering that we evaluate model performance in a real-life setting.

The second approach to estimate predictive performance is a conservative assessment, since we take into account all customers including those that have been targeted by retention measures. We estimate the churn behavior of customers depending on their predicted churn probability with the following model and apply logistic regression (Logit) to estimate the effects.

$$y_i = \beta_0 + \beta_1 \cdot P_i + \epsilon_i \tag{1}$$

P_i denotes the predicted churn probability. The dependent variable y_i is binary (churn: 1 or non-churn: 0). The results from the Logit model in Table 6 Model (1) show that customers with a higher predicted probability to churn are significantly more likely to churn. The results are in line with our previous findings on historical data and on the control subgroup of not called customers, indicating that the ML model is able to predict future churners well.

6.2 Churn prevention

The treatments are designed to facilitate a comparison between random retention measures and measures based on the ML churn predictions. We examine the effect of the treatment with the following model:

$$y_i = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot T_i + \epsilon_i \tag{2}$$

Table 6: Effect of the treatment on churn controlling for the predicted churn probability.

	Dependent variable:	
	churn	
	(1)	(2)
ML-based selection treatment		-0.161^{*}
		(0.096)
Predicted churn probability	8.564^{***}	8.582^{***}
	(0.483)	(0.483)
Constant	-6.070^{***}	-5.973^{***}
	(0.247)	(0.253)
Observations	4,920	4,920
Log Likelihood	−1,636.285	−1,634.875
AIC	3,276.570	3,275.750

*p<0.1; **p<0.05; ***p<0.01

We include an independent variable T_i for the treatment and control for the predicted churn probability. The coefficient of the treatment variable indicates whether being in the ML treatment group has a positive or a negative effect on later churn. Model (2) of Table 6 shows that being in the ML-based selection treatment reduces the probability of becoming a churner significantly.

The coefficient of the treatment indicates, that customers in the ML treatment are 15% ($=100\% - e^{-0.161} \cdot 100\%$) less likely to churn. Hence, retention activities based on ML predictions are more successful than random targeting.

6.3 Revenue change

We analyze whether ML-based proactive churn management also has an impact on revenue development. Calls and visits are not retention measures that directly affect revenue like coupons or special offers would. However, successful retention measures should have a positive financial impact for the company. For testing the effect of treatments on revenue, we use a suitable dependent variable. Since we cannot disclose absolute revenues due to data privacy, we calculate the change of three-month revenues per customer R_i^c (revenue in next three months divided by revenue in previous three months). We estimate the regression model

$$R_i^c = \beta_0 + \beta_1 \cdot P_i + \beta_2 \cdot T_i + \epsilon_i \tag{3}$$

where T_i denotes the treatment and and we control for the predicted churn probability P_i. Regression results in Table 7 show that the ML-based selection treatment has a significant positive effect on revenue development.

The change in three-month revenue is 4.2 percentage points higher for customers in this treatment, meaning that 4.2 percent more of their revenue is maintained. For the customers considered in the experiment, this translates to over 6 million euros higher revenue for the three-month period of the experiment.

Table 7: Change in 3-month revenue.

	Dependent variable:
	Change in 3-month revenue
ML-based selection treatment	0.042**
	(0.021)
Predicted churn probability	0.390***
	(0.106)
Constant	0.740***
	(0.050)
Observations	4,920
Log Likelihood	-5,138.031
AIC	10,284.060

*p<0.1; **p<0.05; ***p<0.01

7 Conclusions

We complement previous literature by building a churn prediction model for a B2B wholesale setting with non-contractual relationships and conducting a field experiment showing that retention measures based on the model's predictions successfully reduce the churn rate.

Evaluated on historical test data, our random forest model achieves an AUC score of 0.76 and a top-decile lift of 3.52. Compared to a random baseline model with AUC 0.5 and top-decile lift of 1, our model is good at identifying customers that are likely to churn. When basing the selection of customers to target with retention measures on the model predictions, the ROC and precision-recall curves can be evaluated. We set the decision threshold so that around 15% of customers are targeted and achieve recall and precision levels of more than 30%.

We verify the capability to identify churners not only on historical data, but also with a field experiment and find that customers with a higher predicted churn probability are in fact more likely to churn. Targeting a fraction of around 15% of customers with customer satisfaction calls and follow-up visits significantly reduces churn when selecting customers based on churn predictions compared to random selection. In our field experiment, the churn rate in the treatment group was reduced to 11.6% compared to 12.9% in the control group. When controlling for the predicted probability to churn, we find a significant treatment effect of decreasing the likelihood to churn by 15%. The retention activities also achieve better revenue developments in the treatment group with a 4.2 percentage points higher change ratio of three-month revenue.

7.1 Theoretical implications

Our research has implications for both churn prediction model development and proactive customer retention management.

First, we contribute to existing literature on churn prediction modeling in the retail and wholesale domain. We develop a churn prediction model for a B2B setting and offer new insights into which features are relevant in such a setting and how models can be trained when the number of customers (samples) is relatively small compared to B2C settings.

In the retail industry, there are several studies on churn prediction for B2C relationships (Buckinx & Van den Poel 2005, Miguéis et al. 2012, 2013), but the B2B setting has received little attention, with the exceptions of Tamaddoni Jahromi et al. (2014) and Gordini & Veglio (2017), who examine online retailers of fast moving consumer goods (FMCG) with business customers as consumers. The studies in the retail domain have shown that

partial churn can be predicted with RFM features (recency, frequency and monetary value).

Our model includes RFM features as well as features specific to the B2B setting. We observe that RFM variables are most important but CRM activity logs also contain valuable information, such as the recency of the last contact of the field representative with the customer. This feature ranks third in terms of importance to our model, which highlights how critical a good customer relationship is, especially in a B2B setting. Our model performance also benefits from information about the margins of both company and B2B customers. We find that there is merit in including features based on different data sources and that features based on invoicing data and CRM data have a high informational value.

Our second contribution is further empirical evidence on the effectiveness of basing retention actions on churn predictions. By conducting a field experiment in the B2B wholesale sector, we fill a gap in literature. There are only a few field experiment studies on churn in general and to our knowledge, none in a B2B setting. In comparison with previous field experiment studies, our customer base is substantially smaller. Our study shows, that successful proactive churn management is possible even in a B2B setting with a few thousand customers. We give evidence for both a reduction of the churn rate and better revenue development in the treatment group. We believe that developing churn prediction models for targeting retention activities also has high potential in further B2B cases.

7.2 Managerial implications

From a managerial perspective, our results also have important implications. Developing a churn prediction model pays off in multiple ways: Firstly, fea-

ture importances give insights into the reasons of customer churn. Secondly, the ROC and precision-recall curves assist in determining an economically viable number of customers to call. Thirdly, the identification of customers to target can be highly automated and integrates well with established CRM retention processes.

In our setting, there is a positive effect on revenues during the three-month period which is analyzed. For the customer base considered, these short-term effects amount to over 6 million euros higher revenue. Under the assumption, that part of the effect persists and the retention measures have a lasting effect on customers in terms of revenue, the direct financial impact of proactive customer relationship management can be substantial.

In our experiment, we could reduce the churn rate in the treatment group by more than 10%. Since it is much more costly to acquire new customers compared to retaining current ones, the company benefits highly from lowering their churn rate. To achieve a long-term effect, the company plans to quarterly perform proactive churn management according to the process outlined in this paper. With churn predictions renewed every three months, they repeatedly target customers at risk and take retention measures.

7.3 Limitations

One limitation of our study is that it relies on data from a one-time experiment with one wholesale company. Since customer retention is an ongoing effort and requires repeated churn predictions and measures, there are methodological questions that arise in connection with it. In order to adapt to changing conditions over time, a churn prediction model needs to be retrained on updated data. This means that the input data can potentially contain customers who have been targeted before. Churn prediction quality

could worsen if the model is trained on customers who would have churned but in fact did not because they were successfully targeted with retention measures. One option is to exclude these customers from the training set, another one is to include features capturing whether and how long ago a customer has been targeted and what the churn probability was at that time. Further research is needed on how to approach this issue.

When targeting a random selection of customers in the control group, we observed a slightly higher churn rate. While the deviance was not significant, further research needs to verify whether this is random or whether retention measures can also increase the probability to churn for certain types of customers.

In the treatment group, we target the customers with the highest probabilities to churn, but research by Ascarza (2018) and De Caigny et al. (2021) shows that these are not necessarily the customers that respond most to retention measures or for whom retention activities bring the highest value. Further discussion is necessary on how to use churn prediction most effectively for churn retention activities. Uplift modeling is a promising method that should be investigated in B2B non-contractual settings. In our set-up, relying on data from the wholesale company, data availability and the overall number of customers limited the ability to develop such a model.

In conclusion, more research is needed in the B2B field, focusing on field experiments regarding customer retention measures based on traditional churn prediction models as well as uplift models. Our study contributes to existing literature by offering new insights and practical evidence from a field experiment in a non-contractual B2B setting and future studies could investigate how our findings generalize to churn in other B2B relationships.

References

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, *55*, 80–98.

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the Target Customer: Social Effects of Customer Relationship Management Campaigns. *Journal of Marketing Research*, *54*, 347–363.

Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The Perils of Proactive Churn Prevention Using Plan Recommendations: Evidence from a Field Experiment. *Journal of Marketing Research*, *53*, 46–60.

Bose, I., & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*, 1–16.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*, 252–268.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, *32*, 277–288.

Burez, J., & Van den Poel, D. (2008). Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. *Expert Systems with Applications*, *35*, 497–514.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626–4636.

Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, *110*, 12.

Chen, K., Hu, Y.-H., & Hsieh, Y.-C. (2015). Predicting customer churn from

valuable B2B customers in the logistics industry: a case study. *Information Systems and e-Business Management*, *13*, 475–494.

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of Operational Research*, *223*, 461–472.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, *66*, 1629–1636.

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.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*, 313–327.

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, *36*, 6127–6134.

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*, 781–800.

De Bock, K. W., & Van Den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications*, *39*, 6816–6826.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classi-

fication algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*, 760–772.

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.

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, *42*, 415–430.

Gattermann-Itschert, T., & Thonemann, U. W. (2021). How training on multiple time slices improves performance in churn prediction. *European Journal of Operational Research*, *295*, 664–674.

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, *197*, 402–411.

Godinho de Matos, M., Ferreira, P., & Belo, R. (2018). Target the ego or target the group: Evidence from a randomized experiment in proactive churn management. *Marketing Science*, *37*, 793–811.

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.

Gür Ali, Ö., & Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications*, *41*, 7889–7903.

Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, *86*, 90–98.

Lessmann, S., & Voß, S. (2009). A reference model for customer-centric data min-

ing with support vector machines. *European Journal of Operational Research*, *199*, 520–530.

Lilien, G. L. (2016). The B2B Knowledge Gap. *International Journal of Research in Marketing*, *33*, 543–556.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4766–4775). https://github.com/slundberg/shap/.

Miguéis, V., Camanho, A., & Falcão e Cunha, J. (2013). Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines. *Expert Systems with Applications*, *40*, 6225–6232.

Miguéis, V., Van den Poel, D., Camanho, A., & Falcão e Cunha, J. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, *39*, 11250–11256.

Molnar, C. (2020). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, *85*, 204–220.

Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., & Vanthienen, J. (2018). Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Systems with Applications*, *106*, 55–65.

Rauyruen, P., & Miller, K. E. (2007). Relationship quality as a predictor of B2B customer loyalty. *Journal of Business Research*, *60*, 21–31.

Ringbeck, D., Smirnov, D., & Huchzermeier, A. (2019). Proactive Retention Management in Retail: Field Experiment Evidence for Lasting Effects. *SSRN Electronic Journal*, .

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative

than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*.

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*, 1258–1268.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York, NY.

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*, 211–229.

Wiersema, F. (2013). The B2B Agenda: The current state of B2B marketing and a look ahead. *Industrial Marketing Management*, *42*, 470–488.

A Details on model training

Hyperparameter tuning for the classification algorithms such as regularized logistic regression, support vector machine (SVM) and random forest is done with a grid search using 5-fold cross-validation. The best performing combination scored by AUC is used for predicting on the test set. See Table 8 for an overview on the hyperparameter settings depending on the number of features F for grid search.

Table 8: Hyperparameters and ranges used during grid search.

Classifier	Hyperparameter	Values
Logistic regression	penalty	L2 regularization
	solver	liblinear
	regularization C	$[10^{-5}, 10^{-4}, \ldots, 10^{2}]$
	class weight	balanced
SVM	kernel	linear
	regularization C	$[0.2, 0.3, \ldots, 1.2]$
	class weight	balanced
Random Forest	max. tree depth	$[3, \ldots, \min(F, 15)]$
	max. number of features	$[\sqrt{F}, 2\sqrt{F}, 3\sqrt{F}, \min(F, 4\sqrt{F})]$
	min. samples per leaf	2
	min. samples split	2
	number of estimators	200
	class weight	balanced