Mau, S., Pletikosa, I. and Wagner, J. (2018), International Journal of Bank Marketing, Vol. 36 No. 6, pp. 1125-1144.

Forecasting next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments

May 31, 2017

Abstract

Purpose - This paper demonstrates the value of enriched customer data for analytical CRM in the insurance sector. In this study, online quotes from an insurer's website are evaluated in term of serving as a trigger event to predict churn, retention and cross-selling.

Design/methodology/approach - For this purpose, the records of online quotes from a Swiss insurer are linked to records of the existing customers in the period 2012-2015. Based on the data from automobile and household insurance policyholders, random forest prediction models for classification are fitted.

Findings - Enhancing traditional customer data with such additional information boosts the accuracy for predicting future purchases substantially. The models identify customers who have a short-termed demand to adapt their insurance coverage with a high probability.

Research implications/limitations - The findings of the study imply that enriching traditional customer data with online quotes yields a valuable approach to predicting purchase behavior. Moreover, the quote data provide supplementary features that contribute to improving the prediction performance. In future studies, the authors recommend investigating the value of other data sources and extending the samples with other insurance products.

Practical implications - This study highlights the importance of selecting the relevant data sources to target the right customers at the right time and thus to benefit from analytical CRM in practice.

Originality/value - This paper is one of the first to investigate the potential value of data-rich environments for insurers and their customers. It provides insights on how to identify relevant customers for ensuing marketing activities efficiently and thus avoid irrelevant offers. Hence, the study creates value for both parties, firm and customer.

Keywords Insurance, Customer relationship management (CRM), Research shopping, Data mining, Random forest, Case study

Paper type Research paper

https://doi.org/10.1108/IJBM-11-2016-0180

1 Introduction

Repeated purchases of products and services is a key motivation for firms to manage their relationships with their existing customers. Hence, over the last decades, companies have established the practice of customer relationship management (CRM) to utilize this business potential, which should contribute to their growth and profitability (Reimer and Becker, 2015). In practice, firms often launch product-oriented sales campaigns, and their service representatives would pro-actively contact customers to make the offer. Still, this common approach is not efficient and imposes the risk of "over-touching" the customer with irrelevant offers (Kamakura, 2008). To avoid these implications, marketing practitioners utilize analytical models within the CRM process to identify the best prospects for a certain offer (e.g. Fader and Hardie, 2009). Still, according to Reimer and Becker (2015), in practice, the majority of CRM projects have been considered failures, and their objectives were not achieved. One major driver for this development was the inability of firms to feed the relevant data into their analytical CRM systems.

The use of data-rich multichannel environments is seen to have the potential to overcome existing challenges in analytical CRM (e.g. MSI, 2014; Verhoef et al., 2010). From novel data sources, highly predictive features and trigger events for customer behavior can be derived, indicating the scarce purchase incidents in the customer universe. Such an approach could contribute to a desired objective for CRM, contacting "the right customer, with the right offer, at the right time" (e.g. Verhoef et al., 2010). Still, the design of such analytical models remains highly domain-specific (Wu et al., 2005), and the firm-customer relationship needs to be considered (Fader and Hardie, 2009).

In the insurance context, targeting the right customer becomes even more relevant because customers perceive insurers "mostly as a necessary evil" (Gidhagen and Persson, 2011), and contact between customers and the firm occur very infrequently. In the past, carriers have focused strongly on acquiring new customers and have neglected the value of their existing customer base (McKinsey & Company, 2013). Hence, insurers that have traditionally strong capabilities in pricing risks need to improve their analytical capabilities to identify the changing demands of the customers. Consequently, insurers are advised to foster the usage of novel data sources and predictive techniques to detect policyholders who could switch to another carrier or provide opportunities to sell additional products (e.g. Accenture, 2011). This strategy would be further motivated by the fact that each percentage of increased retention or cross-sales would be equivalent to a boost of several points in revenue (e.g. McKinsey & Company, 2013; Prinzie and Van Den Poel, 2006). However, insurers are only starting to explore the potential of predictive analytics for marketing purposes, and their utilization is largely preliminary (e.g. Earnix, 2013; Swiss Re, 2014).

In this case study, we present an approach for insurers to utilize enriched CRM data to identify customers who (1) could switch their current policy to another carrier, and (2) provide opportunities for cross-selling an additional product. We make use of the *research-shopping phenomenon*[1] (see Verhoef et al., 2007), and link records from an insurer's customer database with anonymous records of online quotes from the companies' own website. We provide evidence that a prediction model based on such enriched data generates accurate forecasts for the two stated scenarios. Moreover, the approach results

[1] Research shopping refers to a behavioral pattern in multichannel distribution environments, where customers search for products in one channel and purchase in another. Recently, insurance customers were observed as often researching coverage online on insurers' websites but making the purchase offline at the agency.

https://doi.org/10.1108/IJBM-11-2016-0180

in an improved prediction performance compared to baseline models within this study and models from previous studies. Thus, the presented approach is able to separate customers who currently shop for insurance and would be receptive to offerings from those who are not. Based on this evidence our study contributes to research in the domain of analytical CRM in data-rich multichannel environments, with a focus on the insurance industry.

The remainder of this paper is organized as follows. First, we discuss the current challenges in CRM and present an insurance-specific data mining approach. We present the obtained results and discuss the implications of our work for research and practice. Finally, we conclude with a summary of the main findings, discuss the imposed limitations, and indicate directions for future work.

2 Literature Review and Research Questions

CRM is defined as "the practice of analyzing and utilizing data of customers, with the objective of maximizing their individual lifetime value" (Kumar and Reinartz, 2006, p. 5). The practical application of CRM is motivated based on the understanding that it is less costly to retain and expand business with existing customers compared to acquiring new ones (e.g. Bhat and Darzi, 2016). Previous studies among various industries presented scenarios where statistical models supported this strategy and obtained insights as to whether the customer-firm relationship is still alive (e.g. Wübben and von Wangenheim, 2008), whether a customer would churn, i.e., terminate his contract and switch to another vendor (e.g. Lemmens and Croux, 2006), or whether he would be interested in cross-buying an additional product (e.g. Staudt and Wagner, 2016). In the various scenarios, the objective remains similar: identify sales opportunities for an individual customer and subsequently target him with an adequate offer (e.g. Li and Montgomery, 2011; Shankar and Malthouse, 2006). Therefore, firms can make informed decisions regarding when to approach a certain customer and avoid sending him irrelevant marketing messages, which could lead to reactance towards these messages (e.g. Godfrey et al., 2011; Rust and Verhoef, 2005).

According to Payne and Frow (2004), the success of analytical CRM systems relies heavily on the input data. However, firms still struggle to select the relevant data and fail to achieve their CRM objectives (e.g. Forrester, 2009; Reimer and Becker, 2015). To overcome this situation, recent studies (e.g. Reimer and Becker, 2015; Verhoef et al., 2010) propose to enrich the traditional personal data (e.g., age, gender) from information on active customer participation in a firm's offerings (e.g., inquiries from websites), which were found to be stronger predictors for CRM purposes. By doing so, firms would leverage so-called *data-rich multichannel environments*, which refers to the ubiquity of customer data on an individual level and the ability of companies to amass these data across several channels (e.g. Thomas and Sullivan, 2005; Verhoef et al., 2010). According to Verhoef et al. (2010), the current research issues for analytical CRM are as follows: How can data from different sources be combined to improve marketing decisions? How can the predictions be improved through data from new sources? The integration of novel data and the design of analytical CRM systems implies a certain complexity since the approaches need to be evaluated on a domain-by-domain basis (Wu et al., 2005).

In the insurance sector, as in other contractual settings, the CRM activities of firms are focused on preventing churn and engaging in cross-selling (e.g. Fader and Hardie, 2009). A market-related study

https://doi.org/10.1108/IJBM-11-2016-0180

shows that an insurers customer base consists of loyalists and shoppers (McKinsey & Company, 2013). When renewing or adapting an existing policy, the loyal customers remain with their carrier. On the other hand, over 30% of customers shop and compare offers before making purchase decisions and are at risk of switching to another company at that time. In addition, on average, less than 20% of consumers have purchased more than one product from their insurer (Swiss Re, 2014). Hence, firms are advised to refocus their marketing efforts from acquisition towards identifying customers who have changing insurance demands, and protecting them from competitors with an adequate offer. These efforts should be supported through analytical models to avoid the discussed negative consequences, such as inefficiency and over-touching (Accenture, 2011). Still, an industry-related study shows that only a minority of insurers (9%) use predictive models permanently for marketing purposes (Earnix, 2013).

Academic articles that engage with the discussed topic for insurers are scarce. For example, the case study of Smith et al. (2000) presents an approach to predicting the churn and renewal decisions of automobile insurance policyholders. Subsequently, the predicted churners should be targeted and persuaded into retaining their contract. The sample consists of 7% churn vs 93% retention observations. The final prediction model classifies 25% of the churn cases correctly, whereas for retention cases, a 99% ratio is achieved. Thus, the model is mainly able to capture renewal cases with high accuracy. Still, the objective of the study is to predict churn cases, which would provide an opportunity for an intervention to retain the customer. Another study by Guelman et al. (2015) aims to predict an optimal treatment rule for a marketing campaign that aimed to cross-sell household insurance to automobile insurance policyholders. As one outcome, the predicted personal treatments for the campaign have not led to a significant increase of the cross-sell rate in the treatment sample (2.55%) compared to a control group (2.21%). The prediction models in both studies rely solely on traditional personal data and, in summary, provide evidence that forecasting insurance customers' purchases based on such data is challenging.

The very sparse records of customer data result from the specific customer-firm relationship; insurers provide products that customers hope they will not have to use, and contact between the insurers and customers occurs on an infrequent basis (Järvinen et al., 2003). Then again, continuous digitalization changes insurance distribution fundamentally and creates additional sources of data to identify the changing demand of customers (Swiss Re, 2014). As a concrete example of this trend, the current study utilizes anonymous online quotes from an insurer's website, and addresses the following research questions:

1. *Do online quote data improve analytical models to identify customers who are currently shopping for insurance?*
2. *Is the approach applicable for the relevant insurance-related scenarios: (1) churn or renewal of an existing policy and (2) cross-selling of an additional product?*

Therefore, traditional data of policyholders would be enhanced with quote records to identify research shoppers in the customer portfolio, and in a next step, the future purchases of such shoppers would be predicted. Thus, the analytical approach of previous insurance-related studies, which is solely based on traditional data, would be extended. Furthermore, such an approach may be more adequate for insurers to reveal the crucial information that policyholders are currently shopping for new or adapted coverage and thus are at risk to decide for another carrier.

3 Research Design

The objective of our study is to demonstrate how online quotes from an insurer's own website could facilitate the forecast as to whether a customer is shopping to adapt or extend his coverage in the near future. Consequently, we enrich the records of customer and policy data with records of anonymous online quotes. Based on such samples, we fit a random forest classification model to predict the purchase activities. In contractual customer-firm relationships like the insurance setting (see Fader and Hardie, 2009), a firm's option to enhance business with existing customers is two-fold - first, to retain existing contracts of active policyholders, and second, to sell additional products to them. Therefore, the study focuses on predicting the following customer activities: (1) the *churn and retention* of an existing policy and (2) the *cross-selling* of a further product. To assess the value of the enriched data samples for prediction, we compare the results to a baseline case where no online quotes are available. The details of our approach are presented in the following sections.

3.1 Original Data Sets

The data sets used for this case study are obtained from a Swiss insurer. The insurer is one of the top three non-life insurers in the national market and offers non-life and life insurance products in all regions of Switzerland. Furthermore, the company pursues a multichannel strategy and sells its product via agencies, independent brokers, and its own website.

Customer Data For the purposes of this study, we collect the records of customers owning an active automobile or household policy in 2012 and 2015. These traditional CRM data contain general personal covariates (e.g., birthday, gender, etc.), general policy covariates (e.g., policy ID, policy version, inception date, termination date), and pricing relevant covariates of the specific insurance product (e.g., household value insured for a household insurance policy). The sample contains observations of automobile and household insurance policies, including approx. 2.5 million and 3 million observations.

Online Quotes Additionally, we include anonymous records of online quotes from the insurer's own website for the two mentioned insurance products in the same period. These data sets contain solely completed enquiries from the insurer's website, i.e., in case a customer had entered all the requested details in the browser and was shown an insurance offer that included the coverage and price for the searched product. Thus, a quotes record includes pricing relevant covariates for the specific insurance product and its creation date. The samples include approx. 275,000 quotes for automobile insurance and 90,000 quotes for household insurance.

3.2 Data Sets for Analysis

Research-Shopper Samples Within our case company, the original data sets of policies and online quotes were not integrated within the CRM database. Despite the importance of data integration, this situation is still common in practice, and many companies have not yet completed data integration

https://doi.org/10.1108/IJBM-11-2016-0180

across all distribution channels (e.g. Neslin et al., 2006; Swiss Re, 2014). Hence, we link records of active customers and online quotes to identify *research shoppers.*

For the prediction task *churn and retention*, the record of a policy version is linked with an online quote over a set of joint pricing relevant covariates with the condition that the quote's creation date occur between the policy's inception and termination date. For the automobile insurance product, the following common covariates are used for record linkage: *date of birth*, *postal code of residence*, *gender*, *issue date of the driver's license*, and *vehicle model.* The creation date of the online quote serves as a potential trigger event, which could portend good or bad future customer activity (see Verhoef et al., 2010). Thus, we observe the successive actions of the customer on the active policy version a posteriori for the following six months. Based on our sample we derive the three classes for the response variable Y: $Y = 1$, churn of the policy; $Y = 2$, retention of the policy; $Y = 3$, no action taken. We use a period of six months based on insurance-specific empirical evidence of previous research (see e.g. Guelman et al., 2015; Mau et al., 2015). The approach is similar for the household insurance product, and a policy version is linked to a quote using the common covariates *date of birth*, *postal code of residence*, *family status*, *home ownership status*, and *household value insured.*

For the prediction task *cross-selling*, a record of an automobile insurance policy is linked with an online quote for household insurance and vice versa with the condition that the quote's creation date occur between the policy's inception and termination date.[2] The match for the case *Automobile* → *Household* includes the following attributes: *date of birth*, *postal code of residence*, *family status*, and *home ownership status.* In the opposite case, *Household* → *Automobile*, the match involves the following features: *date of birth*, *postal code of residence*, and *gender.* Again, the creation date of the online quote is taken as a trigger, and future cross-buying is observed a posteriori for the next six months. Therefore, the response variable Y is derived with values $Y = 1$ for cross-selling and $Y = 2$ when no action is taken.

Non-Research-Shopper Samples To derive the corresponding *non-research-shopper* samples, we gather active insurance customers to whom no online quotes could be linked in the same period. Consequently, for these observations, no trigger event is available. Thus, we choose a random observation date during the policy tenure as substitution and observe future customer activities a posteriori within the next six months. Overall, these samples represent the vast majority in the customer base, with a ratio greater than 90%. For computational reasons, we apply random sampling in this case and derive a representative subsample of observations (see Knott et al., 2002). For the cross-selling case, the chosen subsample size is larger than in the corresponding research-shopper sample. Therefore, the subsamples include a greater nominal amount of positive observations, where $Y = 1$, which is relevant for the model fitting and prediction on the holdout sample.

An extensive list of all covariates utilized for the record linkage and for the prediction is provided in Table I. In addition, Table II includes details of the resulting *research-shopper (RS)* and *non-research-shopper (non-RS)* samples. For illustrative purposes, Table V in the appendix shows the basic data structure and the linkage approach based on exemplary records.

[2] We only consider customers who do not own an insurance policy of the cross-selling product at the time of observation.

https://doi.org/10.1108/IJBM-11-2016-0180

Table I: Overview of the covariates in the original data sets and the data sets for analysis

Covariates used for record linkage in the samples of active policies and online quotes

General	Automobile insurance	Household insurance
Inception date of the policy version	Date of birth	Date of birth
Termination date of the policy version	Postal code of residence	Postal code of residence
Transaction type of the policy version, either new policy, retention of existing policy or churn of existing policy	Gender, either male or female	Family status, either single person or multiple persons household
Product, either automobile or household insurance	Issue date of driver's license	Home ownership status, either owner or tenant
Creation date of online quote	Vehicle model	Household value insured in Swiss Francs

Covariates in the different samples used for prediction

Covariate	Description	Churn and retention		Cross-selling	
		AM	HH	AM → HH	HH → AM
Transaction type	Transaction of active policy version, either new or retained policy	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Distribution channel	Distribution channel through which the policyholder bought the active policy, either online (Website) or offline (Agency, Broker)	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Contract duration	Ratio of elapsed time since inception and planned contract duration measured at the observation date	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Age (inc)	Age of the policyholder measured at the inception date of the policy version	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Age (obs)	Age of policyholder measured at the observation date	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Gender	Gender of the policyholder (male / female)	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Nationality	Nationality of the policyholder (Swiss / non Swiss)	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Urbanicity of residence	Indicator whether a policyholder lives in an urban or rural area	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Number of other products (obs)	Count of policies for other insurance products (e.g., life, travel, or legal) measured at the observation date	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Number of claims (obs)	Count of all registered claims for a policyholder measured at the observation date	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Period since last claim (obs)	Time since the last registered claims for a policyholder measured at the observation date	RS, non-RS	RS, non-RS	RS, non-RS	RS, non-RS
Number of online quotes (onl)	Count of online quotes linked to a policy holder within 180 days before the observations date	RS	RS	RS	RS
Drivers license since (inc)	Time since the policyholder received his/her drivers license measured at the inception date of the policy version	RS, non-RS		RS, non-RS	RS
Vehicle age (inc)	Age of the insured vehicle measured at the inception date of the policy version	RS, non-RS		RS, non-RS	
Vehicle age (obs)	Age of the insured vehicle measured at the observations date			RS, non-RS	
Vehicle age (onl)	Age of the vehicle entered in the online quote	RS			RS
Type of Vehicle (inc)	Type of the insured vehicle (car / motorcycle) measured at the inception date of the policy version	RS, non-RS		RS, non-RS	
Leasing status of vehicle (inc)	Leasing status of the insured vehicle (yes / no) at the inception date of the policy version	RS, non-RS		RS, non-RS	
Home ownership status (inc)	Status whether the policyholder is home owner or tenant measured at the inception date of the policy version		RS, non-RS	RS, non-RS	
Home ownership status (onl)	Status whether the policyholder is home owner or tenant entered in the online quote		RS	RS	
Family status (inc)	Status whether the policyholder lives in a single person or multiple persons household measured at the inception date of the policy version		RS, non-RS	RS, non-RS	RS, non-RS
Family status (onl)	Status whether the policyholder lives in a single person or multiple persons household entered in the online quote		RS	RS	
Household value insured (inc)	Value of the insured household goods in Swiss Francs (CHF) measured at the inception date of the policy version		RS, non-RS		RS, non-RS
Household value insured (onl)	Value of the insured household goods in Swiss Francs (CHF) entered in the online quote		RS	RS	

Note: AM - automobile insurance, HH - household insurance, RS - research-shopper sample, non-RS - non-research-shopper sample

Mau, S., Pletikosa, I. and Wagner, J. (2018), International Journal of Bank Marketing, Vol. 36 No. 6, pp. 1125-1144.

Table II: Overview of data sets for analysis in the period 2012 - 2015

Churn and retention model			*Cross-selling model*		
Product	Sample	Sample Size	Product	Sample	Sample Size
Automobile	RS	26 097	Automobile → Household	RS	1 127
	non-RS	26 000		non-RS	110 000
Household	RS	11 589	Household → Automobile	RS	14 840
	non-RS	11 500		non-RS	100 000

3.3 Prediction Model

To evaluate the enriched customer information in a prediction model, we apply the random forest algorithm for classification to the derived datasets and compare the prediction performance. According to the study of Lemmens and Croux (2006), classification trees are a suitable technique for predicting customer behavior. The random forest models in this study forecast the probability of future purchases, which is coded in the response variable Y. To consider the imbalance of Y in the samples; see Tables III and IV; we apply class weights to the prediction model. Thus, according to their ratio, the model weights observations of a scarce class higher and those of a frequent class less. Moreover, we choose the validation set approach, also referred to as the holdout method, to objectively evaluate the performance of the models and randomly split the customer samples into a training set containing 2/3 of the sample and a validation set that includes the remaining 1/3 of the sample[3] (e.g. Han et al., 2011, Chapter 8).

To fit an optimized random forest on each training set, we follow the approach of Genuer et al. (2010). First, we eliminate unimportant features based on the variable importance (VI). The VI within each tree is measured using the *Gini Index (G)*, defined as

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk}), \tag{1}$$

where K is the number of classes in the response variable and $\hat{p}_{jk}$ represents the proportion of training observations in the jth region that belong to the kth class. To obtain the VI for the total random forest, we measure the mean decrease in G. Therefore, G is added and averaged over the number of trees B (Hastie et al., 2009, Chapter 9). Second, we select the most predictive features, and finally, we optimize the two random forest parameters: the number of trees B and the number of input variables m that are randomly chosen at each split. In the last two steps, we apply a 10-fold cross-validation and select the best-fitting models based on the prediction accuracy and F-score, which are defined as

$$\text{accuracy} = \frac{\#\text{ correctly classified}}{N}; \qquad F\text{-score} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{2}$$

with N referring to the sample size of the validation set in the cross-validation (for details, see Fawcett, 2006; Genuer et al., 2010). Furthermore, we generate Receiver Operator Characteristic curves (ROC henceforth) and compare the corresponding Area Under the Curve (AUC henceforth) to evaluate the

[3] For the Automobile → Household case, we use an 80/20 split, which led to more stable results during the model fitting. In practice, there is no general rule and the split ratio should be chosen according to the actual application (see Hastie et al., 2009, p. 222)

https://doi.org/10.1108/IJBM-11-2016-0180

prediction performance of different classifiers. For a binary classification model, the ROC curve is a graph, which compares the true positive rate on the y-axis against the false positive rate on the x-axis as the discrimination threshold is varied. The AUC represents the area under the ROC curve and reduces the graph to a numerical measure (Anderl et al., 2016). For the three-class response variable Y in the churn and retention models, the evaluation is based on the one-versus-all method, whereas for the binary response variable Y in the cross-selling models, the standard approach is applied (see Fawcett, 2006). For the implementation of the prediction models, we use the package "randomForest"[4], and for the plot of the ROC curves and AUC computation, we use the package "ROCR"[5] in the statistical software R.

4 Results

In this section, we present the results of the designed prediction models. First, we provide insights of classifiers to forecast the churn and retention of an existing policy, and second, we present details of the cross-selling models. For both cases, we compare the performance of models fitted on the research-shopper samples against those fitted on the non-research-shopper samples, which are used as a baseline.

4.1 Churn and Retention Models

As a consequence of the three-class response variable Y in the churn and retention model, the results for the F-score, AUC value and ROC curve are presented for each category of Y using the one-versus-all comparison; see Section 3.3. A detailed presentation of the results is provided in Table III.

Automobile Insurance Splitting the research-shopper and non-research-shopper samples leads to a training set of approx. 17,300 and a test set of approx. 8,600 data points each. For the research-shopper case, the final model fitted on the training set includes 14 features and the parameter values $B = 500$ and $m = 4$; see Table III for further details. When applied to the respective test set, the model achieved an accuracy rate of 0.844, an increase of 0.160 compared to the non-research-shopper case (0.684). This improvement is driven by the correct classification of the classes churn ($F\text{-}score = 0.732$) and retention ($F\text{-}score = 0.792$), whereas in the non-research-shopper case, both classes are basically not captured by the prediction model, and the accuracy value is based on the correct classification of the class no action (F-score=0.810). The results are confirmed through the corresponding AUC values and ROC curves. The AUC values of the research-shopper model in Table III are increased for each class of Y compared to the non-research-shopper case, which indicates the superiority of this model for the given classification. This fact is visualized by the graphs of the ROC curves in Figure 1, which contain the greater area under the curve. The most important variables for both prediction models are shown in Table III. For the research-shopper case, we obtain important features from the policy at the inception date, e.g., *vehicle age (inc)* (VI = 1568.49) and *driver's license since (inc)* (VI = 1063.35), then from the policy at the observation date, e.g., *contract duration (obs)* (VI = 1168.67) and *age (obs)* (VI = 1047.61), and from the online quote at the observation date, e.g., *vehicle age (onl)* (VI = 1075.88).

[4] Available for download at: https://cran.r-project.org/web/packages/randomForest/index.html

[5] Available for download at: https://cran.r-project.org/web/packages/ROCR/index.html

https://doi.org/10.1108/IJBM-11-2016-0180

Table III: Overview of the results for the churn and retention models

Prediction performance of the final models

	Automobile insurance				Household insurance			
Samples	RS		non-RS		RS		non-RS	
Data sets	Training	Test	Training	Test	Training	Test	Training	Test
Sample Size	17 398	8 699	17 334	8 666	7 726	3 863	7 667	3 833
Distribution of Y								
- Churn	9.80%	9.25%	5.41%	5.40%	2.80%	3.03%	2.14%	2.14%
- Retention	34.03%	33.12%	21.59%	22.78%	28.41%	29.25%	15.73%	15.58%
- No Action	56.17%	57.63%	72.00%	71.90%	68.79%	67.72%	82.13%	82.29%
Prediction results								
Accuracy	0.833 (0.002)	0.844	0.697 (0.003)	0.684	0.855 (0.005)	0.859	0.775 (0.003)	0.787
Correctly classified								
- Churn	0.556 (0.007)	0.595	0.015 (0.004)	0.011	0.412 (0.036)	0.496	0.032 (0.011)	0.037
- Retention	0.757 (0.005)	0.768	0.084 (0.004)	0.087	0.651 (0.011)	0.666	0.125 (0.006)	0.147
- No Action	0.928 (0.002)	0.927	0.928 (0.002)	0.923	0.956 (0.005)	0.959	0.919 (0.004)	0.928
F-score								
- Churn	0.699 (0.006)	0.732	0.026 (0.007)	0.020	0.576 (0.038)	0.652	0.045 (0.016)	0.055
- Retention	0.786 (0.003)	0.792	0.126 (0.006)	0.130	0.737 (0.006)	0.750	0.162 (0.008)	0.195
- No Action	0.877 (0.002)	0.885	0.820 (0.002)	0.810	0.902 (0.004)	0.906	0.873 (0.002)	0.880
AUC								
- Churn		0.921		0.577		0.897		0.6720
- Retention		0.935		0.550		0.903		0.5657
- No Action		0.939		0.559		0.907		0.5665

Note: For the training sets the mean and standard error (SE) in parenthesis of prediction performance of the final model during 10-fold cross-validation are reported. For the test sets the performance of the final model is reported.

Variable importance (VI) and random forest (RF) parameters for the final models

Automobile insurance				Household insurance			
RS		non-RS		RS		non-RS	
Vehicle age (inc)	1568.49	Age (inc)	2492.78	Contract duration (obs)	659.46	Age (inc)	1574.23
Contract duration (obs)	1168.67	Drivers license since (inc)	2434.06	Age (inc)	572.72	Contract duration (obs)	715.45
Vehicle age (onl)	1075.88	Vehicle age (obs)	2311.96	Age (obs)	544.21		
Drivers license since (inc)	1063.35			Household value insured (onl)	427.46		
Age (inc)	1047.61			Household value insured (inc)	247.41		
Age (obs)	1009.99			Period since last claim (obs)	204.17		
Period since last claim (obs)	602.58			Number of online quotes (onl)	180.44		
Number of claims (obs)	519.75			Family status (onl)	171.92		
Number of online quote (onl)	428.49			Number of other products (obs)	137.48		
Number of other products (obs)	398.60			Number of claims (obs)	124.20		
Type of vehicle (inc)	318.79			House ownership (onl)	86.74		
Urbanicity of residence	173.99			Urbanicity of residence	77.71		
Transaction type	161.78						
Gender	145.64						
(RF parameters: $B = 500$; $m = 4$)		(RF parameters: $B = 750$; $m = 3$)		(RF parameters: $B = 500$; $m = 6$)		(RF parameters: $B = 500$; $m = 2$)	

Note: A description of all covariates is provided in Table I.

Household Insurance For this product, the sample splits result in a training set of approx. 7,700 and a test set of approx. 3,800 data points each. The final model, fitted on the training set, in the research-shopper case includes 12 features and the parameter values $B = 500$ and $m = 6$; see Table III for further details. In this case, the prediction accuracy of the random forest model fitted on the research-shopper sample reaches 0.859, which is an increase of 0.072 compared to the non-research-shopper case (0.787). Similar to the observations for automobile insurance customers, the increase is based on the improved correct classification of churn (F-score = 0.652) and retention (F-score = 0.750). Moreover, the AUC values in Table III and the graphs of the ROC curves in Figure 1 reveal the superiority of the random forest fitted on data from the research-shopper customers compared to one fitted on the non-research-shopper sample. For the research-shopper model, the list of the most predictive features includes variables from the policy at the inception date, e.g., *age (inc)* (VI = 572.72) and *household value insured (inc)* (VI = 247.41), then from the policy at the observation date, e.g., *contract duration (obs)* (VI = 659.46) and *age (obs)* (VI = 544.21), and from the online quote at the observation date, e.g., *household value insured (onl)* (VI = 427.46). For further details, we refer to Table III.

4.2 Cross-Selling Models

Based on the binary response variable Y in the cross-selling model, the prediction results are presented using the standard method; see Section 3.3. The detailed overview is provided in Table IV.

Automobile Insurance → Household Insurance Within these samples, we observe the cross-selling case, where active automobile policy holders purchase their initial household contracts at the observation time. For the research-shopper case, the sample split resulted in a training set of 902 and a test set of 225 data points. Because of the low ratio of actual cross-selling cases (approx. 0.5%) in the non-research-shopper sample, we generate a larger subsample that includes 110,000 observations, which was divided into a training ($n = 73,334$) and a test set ($n = 36,666$).[6] For the research-shopper case, the final model fitted on the training set includes 11 features and uses the parameter values $B = 250$ and $m = 11$. This results in a prediction accuracy of 0.893, which is a decrease of 0.101 compared to the non-research-shopper model (0.994). Comparing the values of the F-score for both models, provides an opposite picture (research-shopper: 0.818; non-research-shopper: 0.025). In the non-research-shopper case, the model does not predict the cross-selling, and the high accuracy is based on the correct classification of the no action class (0.998). The AUC values provided in Table IV and the plots of the ROC curves in Figure 2 confirm the findings. Table IV shows the most important features in the two random forest models. The research-shopper model contains variables from the policy at the inception date, e.g., *age (inc)* (VI = 64.17) and *vehicle age (inc)* (VI = 63.21), then from the policy at the observation date, e.g., *age (obs) (*VI = 62.78*)* and *contract duration (obs)* (VI = 51.86), and from the online quote at the observation date, e.g., *household value insured (onl)* (VI = 40.48).

Household Insurance → Automobile Insurance In these samples, the cross-selling scenario is the opposite compared to the previous paragraph. The sample split for the research-shopper case results in a training set of 9,894 and a test set of 4,946 data points. For the non-research-shopper sample, the datasets include 66,667 (training) and 3,333 (test) observations. In the non-research-shopper sample, the

[6] The low cross-selling rate is common in the insurance business (see Guelman et al., 2015) but is further decreased in our samples by a data quality issue. In practice, sales agents would often register a customer twice or more using a new customer ID in case a new product is purchased. When gathering cross-selling cases for this study, we observed new policies for a given customer ID. Thus, we missed customers who had been (re-)registered with a new customer ID.

https://doi.org/10.1108/IJBM-11-2016-0180

actual cross-selling rate is very low, which is due to similar reasons, as stated in the paragraph above.[6] For the research-shopper case, the best model, fitted on the training set, includes five features and the parameter values $B = 250$ and $m = 2$. For further details see Table IV. The random forest fitted on the research-shopper sample achieved a prediction accuracy of 0.879, which is a decrease of 0.119 compared to the non-research-shopper case (0.998). In contrast, the latter model is not able to classify the actual cross-selling instances (F-score = 0.000), whereas the research-shopper model achieves a high F-score of 0.786. The findings are confirmed by the increased AUC (see Table IV) and the graphs of the ROC curves in Figure 2, which suggest the superiority of the research-shopper model to classify this customer behavior correctly. For the random forest fitted on the research-shopper sample, the list of the most predictive features includes candidates from policy at the inception date, e.g., *age (inc)* (VI = 1053.39), then from the policy at the observation date, e.g., *contract duration (obs)* (VI = 750.71), and from the online quote at the observation date, *vehicle age (onl)* (VI = 894.30).

Table IV: Overview of the results for the cross-selling models

Prediction performance of the final models

	Automobile insurance → Household insurance				Household insurance → Automobile insurance			
Samples	RS		non-RS		RS		non-RS	
Data sets	Training	Test	Training	Test	Training	Test	Training	Test
Sample Size	902	225	73 334	36 666	9 894	4 946	66 667	33 333
Distribution of Y								
- Cross-selling	29.38%	31.56%	0.52%	0.49%	30.21%	30.93%	0.21%	0.22%
- No Action	70.62%	68.44%	99.48%	99.51%	69.79%	69.07%	99.79%	99.78%
Prediction results								
Accuracy	0.887 (0.008)	0.893	0.994 (0.001)	0.994	0.870 (0.005)	0.879	0.997 (0.001)	0.998
Correctly classified								
- Cross-selling	0.695 (0.032)	0.761	0.011 (0.005)	0.017	0.694 (0.012)	0.722	0.000 (0.000)	0.000
- No Action	0.966 (0.006)	0.955	0.998 (0.001)	0.998	0.945 (0.003)	0.949	0.999 (0.001)	0.999
F-score	0.776 (0.024)	0.818	0.016 (0.008)	0.025	0.762 (0.010)	0.786	0.000 (0.000)	0.000
AUC		0.935		0.549		0.927		0.555

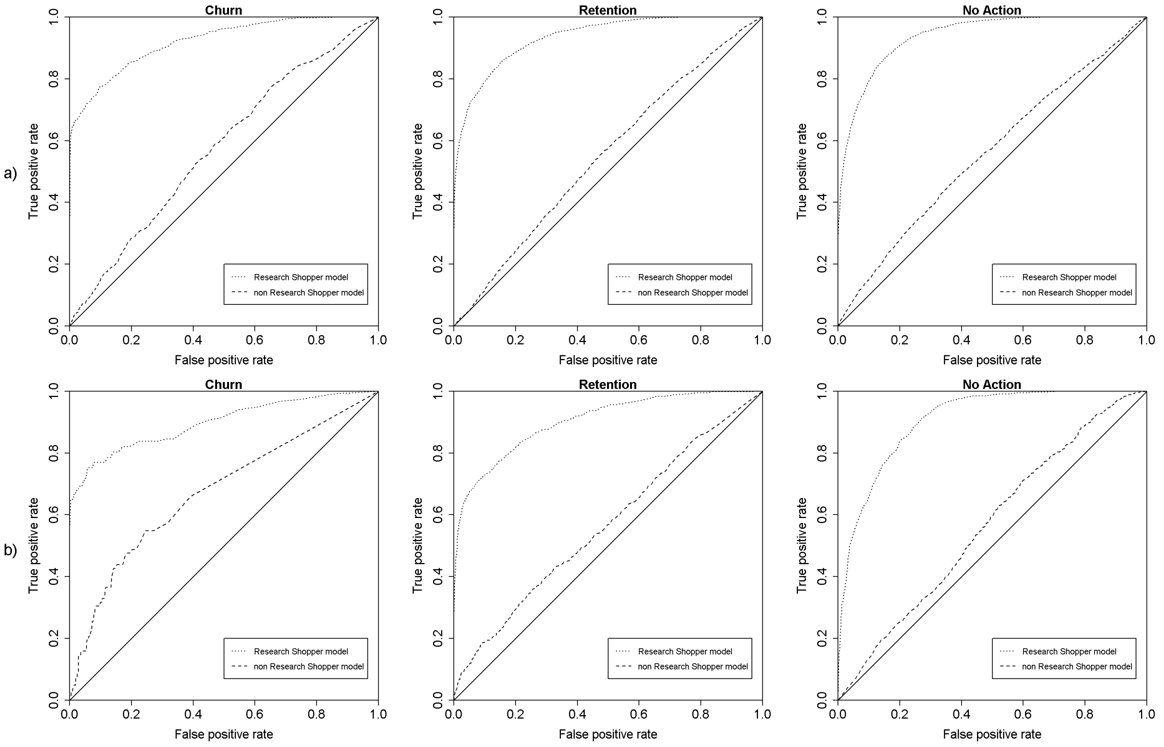
Note: For the training sets the mean and standard error (SE) in parenthesis of prediction performance of the final model during 10-fold cross-validation are reported. For the test sets the performance of the final model is reported.

Variable importance (VI) and random forest (RF) parameters for the final models

Automobile insurance → Household insurance				Household insurance → Automobile insurance			
RS		non-RS		RS		non-RS	
Age (inc)	64.17	Age (obs)	331.09	Age (obs)	1119.62	Age (inc)	206.22
Vehicle age (inc)	63.21	Age (inc)	324.69	Age (inc)	1053.39	Contract duration (obs)	47.56
Age (obs)	62.78			Vehicle age (onl)	894.30	Household value insured (inc)	11.65
Contract duration (obs)	51.86			Contract duration (obs)	750.71	Number of other products (obs)	10.88
Household value insured (onl)	40.48			Household value insured (inc)	345.79		
Period since last claim (obs)	40.20						
Number of claims (obs)	13.88						
Urbanicity of residence	13.38						
Number of online quotes (onl)	9.56						
Number of other products (onl)	9.05						
Gender	5.31						
(RF parameters: $B = 250$; $m = 11$)		(RF parameters: $B = 100$; $m = 2$)		(RF parameters: $B = 250$; $m = 2$)		(RF parameters: $B = 100$; $m = 4$)	

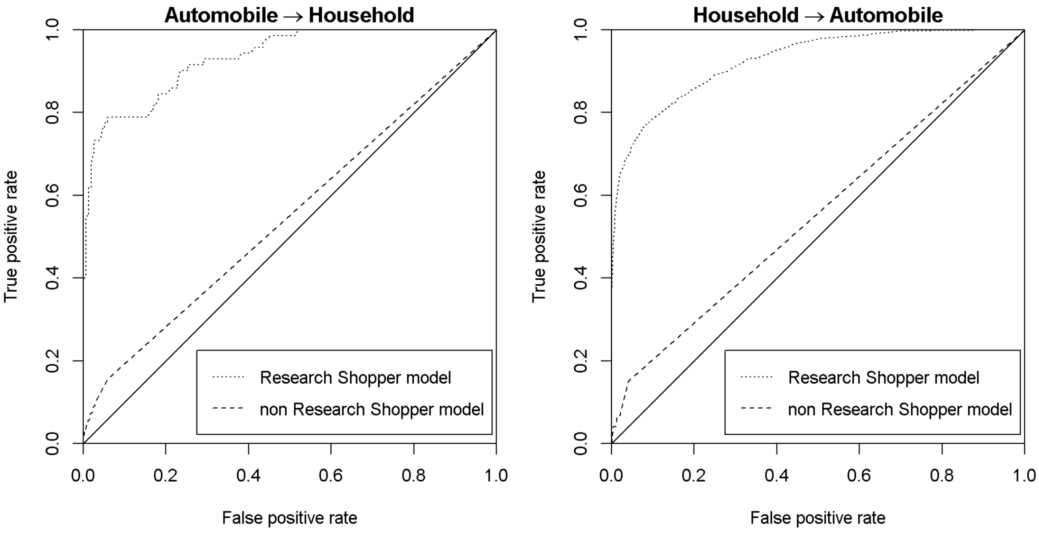
Note: A description of all covariates is provided in Table I.

Figure 1: ROC curves for the churn and retention models



Notes: a) Automobile insurance, b) Household insurance

Figure 2: ROC curves for the cross-selling models



5 Discussion and Implications

To address a recent trend in CRM, previous studies have discussed the value of rich sources of customer data for analytical purposes (e.g. Reimer and Becker, 2015; Verhoef et al., 2010). Our study focuses on the insurance sector and investigates the potential of this trend to forecast churn and retention, and cross-selling opportunities of active policyholders. Therefore, records of automobile and household insurance policies have been enriched with anonymous records of online quotes from the insurer's website. Subsequently, these enhanced data sets have been fed into a random forest prediction model. As a result, we find that models fit on these enriched data produce accurate forecasts of future purchases and outperform the baseline models, which are fit on traditional customer data. Thus, we can extend data mining approaches and improve the prediction performance of existing studies (see e.g. Smith et al., 2000; Guelman et al., 2015). Moreover, our findings show evidence for practitioners in terms of how to detect insurance customers who are currently shopping for coverage. Such knowledge would be crucial for carriers to protect their customer base against competitors' offers (McKinsey & Company, 2013).

Similar to previous work in the financial services sector, the ratio for positive responses of customers (e.g., cross-buying, retaining, churning) in the variable Y is very low in the non-research-shopper samples of our study. For example, the churn rate for automobile insurance customers is approx. 5%, which is comparable to the study by Smith et al. (2000). This value increases in the research-shopper sample to almost 10%, indicating the relevance of online quote data as a trigger for potential purchases. The effect is analog for the churn prediction. In the research-shopper case, 55.6% of potential churners can be classified correctly through the prediction model. This is a substantial improvement compared to 1.1% in the non-research-shopper case and 25% in the study by Smith et al. (2000). Moreover, previous studies predicting cross-selling provide low response rates of customers that range between 0.5% and 2.5% (e.g. Guelman et al., 2015; Knott et al., 2002), similar to the ratios in our non-research-shopper sample. This value increases to 30% in the research-shopper samples within this study and led to a correct classification of over 70% for future cross-buying. This, again, is an improvement compared to the results from the non-research-shopper models and previous studies. Within the study of Guelman et al. (2015), the prediction model to forecast the cross-selling of household insurance policies achieved no significant increase when compared to a control group. Furthermore, the study of Knott et al. (2002) achieved approx. 50% correct classification of the next banking products that a customer would purchase. We acknowledge that the prediction scenarios in the mentioned studies vary in certain aspects to those in the current study and that the results are not fully comparable. Still, our findings provide evidence that our approach to enrich insurance customer data through online quotes could extend approaches of existing studies and result into more accurate forecasts of whether a customer is about to purchase insurance coverage.

Online quote data can be labeled as customer-initiated contacts (see Reimer and Becker, 2015) or inbound customer contacts (see Kamakura, 2008) which hint to customers who already have a purchase mindset on a firm and its services. Thus, it seems simple to turn a shopping customer into a business transaction. Still, a customer might not be loyal and shop for offers at other insurers in parallel, e.g., on price comparison websites. In the churn and retention scenario, it is possible to distinguish disloyal churners, who are likely to switch carriers after searching for coverage, from loyalists, who retain their

https://doi.org/10.1108/IJBM-11-2016-0180

policies. In the cross-selling scenario, the view is limited, and we only observe customers who cross-buy with our case company, and we cannot draw a conclusion about customers who accept a competitor's offer. In this case, one could interpret the false discovery rate[7] as the ratio of customers, who were predicted to shop for an additional product with our insurer but might have decided in favor of another carrier. This limitation can only be solved through data augmentation. Regarding relevant features for prediction, we can confirm relevant covariates to predict purchases of customers in previous studies, e.g., occurrence of a claim (Kamakura, 2008) and product ownership (Knott et al., 2002). Moreover, our results provide evidence that the creation of features from the additional data contributes to the prediction of customer behavior (Verhoef et al., 2010). Within our models, features from the online quotes records, e.g., age of vehicle and sum insured, have been included into the random forest models and show high variable importance. Overall, our findings are in line with previous studies (see Reimer and Becker, 2015), which stated among various industries that data from a customer-initiated contact are a superior predictor compared to the traditional personal data of customers.

The results of this study provide implications for insurance firms as well. When implemented in practice, the presented approach would enable carriers to identify customers who are shopping for insurance coverage. Hence, insurers can exert CRM activities efficiently to retain potential switchers and up-sell and cross-sell loyalists, thus protecting their customer base against competitors with aggressive growth strategies, which is critical to the survival of many carriers (McKinsey & Company, 2013). Based on the high correct classification of the presented models for research-shoppers (see Table III and IV for details) even personal contact through sales agents can be performed efficiently and would not be a waste of human resources. Despite the customers' preference to search for product information online, the personal interaction remains important for the majority of customers when purchasing insurance (Swiss Re, 2014). Moreover, through the presented prediction approach, carriers could benefit from the correct timing for a marketing intervention, i.e., in advance of the customer's purchase decision. This is of particular relevance given the scarce contact between an insurers and its customers (Gidhagen and Persson, 2011). During a campaign at our case company, over 90% of customers who churned a policy could not be won back and approx. 25% of customers stated that they would perceive a call from the insurer afterwards as disturbing. Such customers could have been identified through the presented data mining approach and contacted directly for the retention of his business. Thus, the insurer could decrease marketing activities at other times, which would contribute to avoiding the over-touching of customers.

Overall, insurers are advised to access additional sources of customer data and foster advances in analytical CRM to manage existing customers instead of focusing disproportionately on acquisition (e.g. Prinzie and Van Den Poel, 2006; Accenture, 2011; Swiss Re, 2014). Therefore, carriers would benefit from accurate customer insights, such as those presented in the current study. However, the applications of novel analytical approaches in the insurance sector remain preliminary (Swiss Re, 2014). Still, from marketing professionals at our case company and its partner firms across Europe, we received feedback that some insurers experiment with similar data, whereas others are in an earlier stage.

[7] In data mining, the false discovery rate (FDR) of a prediction model for classification refers to cases that were predicted to be positive ($\hat{Y} = cross-selling$) but where their actual condition was negative ($Y = no\ action$). In both our cross-selling models, this rate is approx. 12%.

6 Conclusions and Future Research

The collection of additional customer data and their utilization for analytical CRM are shown to gain more valuable insights when predicting customers' responses compared to approaches solely based on traditional customer data (e.g. Verhoef et al., 2010). In this study, we demonstrate the value of online quotes from an insurer's website when linked with customer data and fed into a prediction model to forecast future purchases. The data mining approach of this paper provides insights in two ways. First, we document how anonymous records of online quotes can be linked to records of existing customers in the CRM database. Within our approach, the quote data serve as trigger events to identify customers who may be currently shopping for insurance coverage. Second, we show how these linked data, when applied to a machine learning algorithm, lead to accurate forecasts of which customers are adapting their active policy or purchasing a new insurance product in the short-term. Moreover, the analytical approach substantially improves the prediction accuracy when compared to forecasts based solely on traditional customer data, which contributes to the findings of existing studies apart from the insurance sector (see e.g. Reimer and Becker, 2015). Thus, our research design is aligned to the specific nature of the customer-firm relationship in the insurance sector, where interactions are infrequent and forecasting purchases on traditional data is challenging. When turned into practice, the presented approach provides value for both parties. The insurer would profit from an efficiency gain in their CRM process when retaining their customer base and expanding business through up-selling and cross-selling additional coverage. The customer would benefit from receiving the right offer at the right time and thus, would not be bombarded with numerous marketing offers in which the customer is usually not interested in at all.

Though the presented study provides insights, it is also limited in certain respects. First, the study includes only data for two non-life insurance products, and the results may vary for other products, such as life insurance. Furthermore, the analysis of cross-selling can not be conducted as next-product-to-buy (NPTB) approach (see Knott et al., 2002), because we did not have access to online quote data for additional products. Second, the results are based on data of one insurer only and could be biased by the specific structure of the organization and the firm's marketing strategy. Moreover, our database is not augmented by online quote data from other insurers. Thus, in the cross-selling case, we cannot say, whether a false positive prediction of cross-buying was either related to the fact that the customer was not shopping for an additional insurance product or that he purchased this product at another insurance company. Furthermore, the results of the cross-selling scenario could have been influenced through the quality issue of the customer data, which we discussed; see Footnote 6. Still, the results provide meaningful insights.

Future studies could extend the focus of our study by including other data sources, e.g., from external environments, to derive trigger events for the purchase of insurance policies. Moreover, researchers could attempt to replicate the findings of this study with samples from other insurance products, e.g., life insurance, and with other firms in other geographical regions to generalize the findings. Finally, the context of this paper could be expanded to a field study to validate the value of such predictions in a real business setting.

Mau, S., Pletikosa, I. and Wagner, J. (2018), International Journal of Bank Marketing, Vol. 36 No. 6, pp. 1125-1144.

References

Accenture (2011), “The Path to High Performance in Insurance, Transforming Distribution and Marketing with Predictive Analytics”. Available at: http://insuranceblog.accenture.com/wp-content/uploads/2013/07/Transforming_Distribution_and_Marketing_with_Predictive_Analytics.pdf (accessed 24 April 2017).

Anderl, E., Becker, I., von Wangenheim, F., and Schumann, J. H. (2016), “Mapping the customer journey: Lessons learned from graph-based online attribution modeling”, *International Journal of Research in Marketing*, Vol. 33 No. 3, pp. 457–474.

Bhat, S. A. and Darzi, M. A. (2016), “An approach to competitive advantage in the banking sector by exploring the mediational role of loyalty”, *International Journal of Bank Marketing*, Vol. 34 No. 3, pp. 388–410.

Earnix (2013), “2013 Insurance Predictive Modeling Survey”. Available at: http://earnix.com/2013-insurance-predictive-modeling-survey/3594/ (accessed 24 April 2017).

Fader, P. S. and Hardie, B. G. (2009), “Probability Models for Customer-Base Analysis”, *Journal of Interactive Marketing*, Vol. 23 No. 1, pp. 61–69.

Fawcett, T. (2006), “An introduction to ROC analysis”, *Pattern Recognition Letters*, Vol. 27 No. 8, pp. 861–874.

Forrester (2009), “Answers To Five Frequently Asked Questions About CRM Projects”, *Forrester.*

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010), “Variable selection using random forests”, *Pattern Recognition Letters*, Vol. 31 No. 14, pp. 2225–2236.

Gidhagen, M. and Persson, S. G. (2011), “Determinants of digitally instigated insurance relationships”, *International Journal of Bank Marketing*, Vol. 29 No. 7, pp. 517–534.

Godfrey, A., Seiders, K., and Voss, G. B. (2011), “Enough Is Enough! The Fine Line in Executing Multichannel Relational Communication”, *Journal of Marketing*, Vol. 75 No. 4, pp. 94–109.

Guelman, L., Guillén, M., and Pérez-Marín, A. M. (2015), “A decision support framework to implement optimal personalized marketing intervention”, *Decision Support Systems*, Vol. 72, pp. 24–32.

Han, J., Pei, J., and Kamber, M. (2011), *Data mining: concepts and techniques*, 3rd ed., Elsevier.

Hastie, T., Tibshirani, R., and Friedmann, J. (2009), *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics.

Järvinen, R., Lehtinen, U., and Vuorinen, I. (2003), “Options of strategic decision making in services Tech, touch and customisation in financial services”, *European Journal of Marketing*, Vol. 37 No. 5/6, pp. 774–795.

Kamakura, W. A. (2008), “Cross-selling: Offering the right product to the right customer at the right time”, *Journal of Relationship Marketing*, Vol. 6 No. 3-4, pp. 41–58.

Knott, A., Hayes, A., and Neslin, S. a. (2002), “Next-product-to-buy models for cross-selling applications”, *Journal of Interactive Marketing*, Vol. 16 No. 3, pp. 59–75.

https://doi.org/10.1108/IJBM-11-2016-0180

Kumar, V. and Reinartz, W. (2006), *Customer Relationship Management: A Databased Approach*, Hoboken: Wiley.

Lemmens, A. and Croux, C. (2006), “Bagging and Boosting Classification Trees to Predict Churn”, *Journal of Marketing Research*, Vol. 43 No. 2, pp. 276–286.

Li, S. and Montgomery, L. (2011), “Cross-Selling the right Product to the right Customer at the right time”, *Journal of Marketing Research*, Vol. 48 No. 4, pp. 683–700.

Mau, S., Cvijikj, I. P., and Wagner, J. (2015), “From research to purchase: an empirical analysis of research-shopping behaviour in the insurance sector”, *Zeitschrift fur die gesamte Versicherungswissenschaft*, Vol. 104 No. 5, pp. 573–593.

McKinsey & Company (2013), “Beyond Price: The Rise of Customer-Centric Marketing in Insurance”. Available at: http://oesai.org/wp-content/uploads/2015/01/Beyond_price_The_rise_of_customer-centric_marketing_in_insurance.pdf (accessed 24 April 2017).

MSI (2014), “Research Priorities 2014-2016”, *Marketing Research Institute (MSI).* Available at: http://www.msi.org/uploads/files/MSI_RP14-16.pdf (accessed 24 April 2017).

Neslin, S. a., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., and Verhoef, P. C. (2006), “Challenges and Opportunities in Multichannel Customer Management”, *Journal of Service Research*, Vol. 9 No. 2, pp. 95–112.

Payne, A. and Frow, P. (2004), “The role of multichannel integration in customer relationship management”, *Industrial Marketing Management*, Vol. 33 No. 6, pp. 527–538.

Prinzie, A. and Van Den Poel, D. (2006), “Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models”, *European Journal of Operational Research*, Vol. 170 No. 3, pp. 710–734.

Reimer, K. and Becker, J. U. (2015), “What customer information should companies use for customer relationship management? Practical insights from empirical research”, *Management Review Quarterly*, Vol. 65 No. 3, pp. 149–182.

Rust, R. T. and Verhoef, P. C. (2005), “Optimizing the marketing interventions mix in intermediate-term crm”, *Marketing Science*, Vol. 24 No. 3, pp. 477–489.

Shankar, V. and Malthouse, E. C. (2006), “Moving Interactive Marketing Forward”, *Journal of Interactive Marketing*, Vol. 20, pp. 2–4.

Smith, K. A., Willis, R. J., and Brooks, M. (2000), “An analysis of customer retention and insurance claim patterns using data mining: a case study”, *Journal of the Operational Research Society*, Vol. 51 No. 5, pp. 532–541.

Staudt, Y. and Wagner, J. (2016), “What policyholder and contract features determine the evolution of non-life insurance customer relationships? A case study analysis”, *Working Paper No. 28, Faculty of Business and Economics (HEC), University of Lausanne.*

Swiss Re (2014), “Digital Distribution in insurance. A quiet revolution”, *Swiss Re - sigma*, No. 2.

https://doi.org/10.1108/IJBM-11-2016-0180

Thomas, J. S. and Sullivan, U. Y. (2005), “Managing Marketing Communications with Multichannel Customers”, *Journal of Marketing*, Vol. 69 No. 4, pp. 239–251.

Verhoef, P. C., Neslin, S. a., and Vroomen, B. (2007), “Multichannel customer management: Understanding the research-shopper phenomenon”, *International Journal of Research in Marketing*, Vol. 24 No. 2, pp. 129–148.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., and Ganesan, S. (2010), “CRM in data-rich multichannel retailing environments: a review and future research directions”, *Journal of Interactive Marketing*, Vol. 24 No. 2, pp. 121–137.

Wu, C.-H., Kao, S.-C., Su, Y.-Y., and Wu, C.-C. (2005), “Targeting customers via discovery knowledge for the insurance industry”, *Expert Systems with Applications*, Vol. 29 No. 2, pp. 291–299.

Wübben, M. and von Wangenheim, F. (2008), “Instant Customer Base Analysis: Managerial Heuristics Often Get It Right”, *Journal of Marketing*, Vol. 72 No. 3, pp. 82–93.

A Appendix

Table V: Exemplary data records for the automobile insurance product

Original insurance policies

Policy ID	Policy Version	Inception Date	Termination Date	Trans-action	Product	Birthday	Gender	Postal Code	Issue Date of Drivers License	Vehicle Model
1234567	1	13.02.10	28.10.13	N	AM	21.01.80	M	8004	11.03.98	Audi A4
1234567	2	29.10.13	31.12.99	R	AM	21.01.80	M	8004	11.03.98	Audi A4
1987654	4	10.05.08	31.12.99	R	AM	07.04.84	M	3007	13.05.02	VW Golf
1345678	1	28.02.09	31.03.14	N	AM	24.12.54	M	4001	20.01.74	Volvo V70
1345678	1	28.02.09	31.03.14	C	AM	24.12.54	M	4001	20.01.74	Volvo V70
1678912	2	10.01.14	31.12.99	R	AM	10.11.57	F	1010	02.09.80	VW Passat

Original online quotes (OQ)

OQ ID	Creation Date	Product	Birthday	Gender	Postal Code	Issue Date of Drivers License	Vehicle Model
OQ9876	11.09.13	AM	21.01.80	M	8004	01.03.98	Audi A4
OQ9914	08.03.14	AM	07.04.84	M	3007	13.05.02	VW Golf

Derived data set for the research-shopper (RS) case

Policy ID	Policy Version	Inception Date	Termination Date	Trans-action	Product	OQ ID	Observation Date	Product	Response var. (Y)
1234567	1	13.02.10	28.10.13	N	AM	OQ9876	11.09.13	AM	R
1987654	4	10.05.08	31.12.99	R	AM	OQ9914	08.03.14	AM	NOA

Derived data set for the non-research-shopper (non-RS) case

Policy ID	Policy Version	Inception Date	Termination Date	Trans-action	Product	Random Observation Date	Product	Response var. (Y)
1345678	1	28.02.09	31.03.14	N	AM	15.12.13	AM	C
1678912	2	10.01.14	31.12.99	R	AM	02.06.12	AM	NOA

Note: AM - automobile insurance, N - new policy, R - renewed policy, C - churned policy, NOA - no action