Anticipating insurance customers' next likely purchase events

Stefan Mau*, Daniel Mueller, Irena Pletikosa Cvijikj, Joël Wagner †

Note: This is work in progress, paper version as of May 13, 2016.

Abstract

In the current study we present an approach to utilize the digital traces of customers, in order to detect the next likely purchase event for a customer on an individual level. We make use of internal and external data sources, present a method for data matching, event detection and design a predictive model to forecast sales chances with high accuracy.

Keywords: insurance industry · customer relationship management · data matching · predictive analytics · empirical analysis

*Corresponding author.

†Stefan Mau (smau@ethz.ch), Daniel Mueller (danielmueller@ethz.ch) and Irena Pletikosa Cvijikj (ipletikosa@ethz.ch) are with the Chair of Information Management, Swiss Federal Institute of Technology Zurich, Weinbergstrasse 56-58, 8092 Zurich, Switzerland. Joël Wagner (joel.wagner@unil.ch) is with the Department of Actuarial Science, Faculty of Business and Economics (HEC), University of Lausanne, Quartier UNIL-Dorigny, Bâtiment Extranef, 1015 Lausanne, Switzerland.

1 Introduction

"Developing Marketing Analytics for a Data-Rich Environment" is named as one of the top research priorities for marteking researchers in the respective study of the Martking Sciense Institute (MSI, 2014). In the age of digitalization and Big Data and with the different types of data available, companies are more empowered than ever to make use of advanced data analytics, which generate accurate customers insights. Moreover, enterprises need to explore the opportunities of these individual-level data within their customer relationship management (CRM) operations and activities. Recent research has already shown, that action and reaction data of customers have much greater potential to forecast customers' purchase activities compared to the tradional personal characteristics (Reimer and Becker, 2015).

Similar to others industries, the insurance sector is adopting to the digitalization trend, but according to the sigma study of SwissRe (2014) with a smaller velocity. In 2012 in the European Union (EU) about 5% of insurance policies were purchased digitally, whereas the average share over all other products was almost three times as high with about 14%. In contrast, the digital distribution channel is most utilized by European customers to perform research for insurance products (SwissRe, 2012). Thereby for insurance customers the digital channel is an important source for gathering information through online quotes in the product research phase, leading to new shopping pattern known as research-shopping (Verhoef et al., 2007). On the other hand carriers need to recognize the importance of their customers' digital traces during the search process and leverage the potential of these online quotes. Moreover, carriers should expand their view towards external data sources, which could provide enhanced insights about their customers' current situation.

Insurers and their customers find themselves in a contractual setting (Fader and Hardie, 2009), where companies know their active customers and seek to prevent churn or possibilities to up and cross-sell to customers. The challenge here is to design probability models, which anticipate these possibilities with high accuracy. The studies of Kamakura (2008) and Li et al. (2011) addressed this challenge as predicting to sell "the right product to the right customer at the right time". In the current study we present an approach to address this challenge for the insurance sector.

2 Literature Review

According to Fader and Hardie (2009) there has been a strong tradition of probability models in marketing since the 1950's. The motivation was and still is to make accurate forecasts of customers' purchase activities based on the observed behavioral characteristics from the past.

Moreover, the need for adoption of data mining techniques to support decision-making in a customer driven industry was already recognized as essential for targeting customers effectively. Current research regarding the use of predictive models for CRM purposes ranges along many different types of machine learning methods. For example, decision trees were applied to address a common task of customer classification. Yet, this task was shown to be highly domain-specific, thus indicating a need for sector specific implementation and domain knowledge as key factors for success (Wu et al., 2005).

For the insurance industry the documentation for the application of data mining methods for targeting

customers is scarce. However, we present a selection of relevant studies in this paragraph. The paper of Smith et al. (2000) had the objective to predict the retention probabilities of motor insurance customers using a sample of an Australian carrier. The goal was to increase efficiency and achieve market growth through data mining by detecting non-retention customers through a predictive model. Only personal customer data, e.g. age of vehicle, gender etc., was used. The authors demonstrated three different data mining techniques for classification, logistic regression, decision trees and neural networks. The later one outperformed the others and was applied to a holdout sample (test set). Still the chosen neural net model suffered from a low recall (<25%) when classifying customers who terminated their policy. In practice 3 out of 4 non-retention customers would not be identified by the model and not targeted through marketing activities. A further study was conducted by Wu et al. (2005), who applied classification trees and decision rules (IE3) to identify potential customers for life insurance products within the portfolio of Taiwanes insurer. Again only traditional CRM covariates were included in the model. The authors obtained over 3000 decision rules during training and achieved 80% accuracy. A major challenge within the study, as mentioned by the authors, was the communication of these complex decision rules to the management of the participating carrier. The study of Guelman et al. (2015) predicted the cross-selling of household insurance to motor insurance customers based on a sample from a Canadian insurer. Predictions were made using random forest and solely personal customer data were included. In a field test new motor insurance customers were predicted for the cross-selling of household insurance, but the comparision to a baseline group showed that the uplift of the model about 2% from 11% to 13% was not significant.

A further step for preditcing customers' purchase activities is taken in the marketing literature, where studies by Kamakura (2008) and Li et al. (2011) formulate the target to sell "the right product to the right customer at the right time". The studies see retaining customers, up-selling and cross-selling as established techniques, but innovations in CRM have changed their application in the practice. Nowadays human intuitution of the sales personell for this task is complemented by analytical tools and information technology. Moreover, since insurance policies, like other financial services, represent utilitarian products, which involve rationality in purchase decision making (Yang, 2015), these are purchased only when there is a need for reducing future risks. So the timing dimension of selling amongst others is defined by a change in financial maturity or life situation where new risks need to be insured (Kamakura, 2008).

Another According to Reimer and Becker (2015) companies and researchers need to consider adequate data sources to achieve their CRM objectives. The study hinted that customers' action and reaction data possess greater predictive power compared to their personal data. For the insurance case, online quotes from a carriers website represent such reaction data. In combination with the utilitarian nature of insurance policies these online quote data could lead to high accurate predictions of purchases.

Following the recommendation given by Reimer and Becker (2015) in our study we combine personal data, i.e. a customer's covariates from the CRM database, with reaction data, i.e. online quotes from the carriers website and external transaction data of insured goods, and predict purchase behavior. Our objective is to achieve high prediction accuracies for purchase behaviour of insurance customers.

3 Research Design

3.1 Dataset

This analysis is based on three distinct datasets. First, we extracted personal customer and policy data from the data warehouse of a Swiss carrier. The insurer is one of the top three non-life insurers in the Swiss market (SVV, 2014) and offers non-life as well as life insurance products in all regions of Switzerland. For the purposes of this study, only active policies of private customers owning a motor or household insurance were chosen. Second, we added a dataset of online quotes as customer reaction data from the insurers website for the two mentioned products. Third, we included two datasets of external data, 1) advertisments from an online car sales platform in Switzerland, and 2) advertisments from a Swiss online appartment rental platform. These external data could be considered as reaction data, as those activities of customers have a direct effect on the purchased insurance product. All details regarding the datasets, such as sample size, period and included covariates, are described in Table 1.

Table 1: Details of included data samples

Sample	Policies		Online quotes		External data	
Insurance product	Household	Motor	Household	Motor	Household	Motor
Sample size	$\sim$ 2 million	$\sim$ 2.5 million	$\sim$ 90 000	$\sim$ 270 000	$\sim$ 2 600	$\sim$ 15 000
Period	2012 - 2014	2011 - 2014	2012 - 2014	2011 - 2014	2013	2014
Covariates						
Inception date	✓	✓				
Termination date	✓	✓				
Occurence date			✓	✓	✓	✓
Name	✓	✓			✓	✓
Address	✓	✓			✓	✓
Postal code	✓	✓	✓	✓	✓	✓
Date of birth	✓	✓	✓	✓	✓	✓
Household insurace specific						
Family status	✓		✓			
Home ownership	✓		✓		✓	
Sum assured	✓		✓			
Motor insurace specific						
Vehicle brand & model		✓		✓		✓
Vehicle registration date		✓		✓		✓
Issue date of drivers licencse		✓		✓		✓
Gender of driver		✓		✓		

3.2 Methodology

Matching and detection

Our objective in the study is to use customers' reaction data as predictors for short-term purchase behaviour of insurance policies. In order to accomplish this goal, we had to match the instances within three distinct dataset chronolocally. Therfore, we applied a three step approach. For each given month

in the observation period:

1. We extracted active contracts from the policy sample.
2. We matched online quotes conducted in this month and external data advertisments created in this month using common covariates.
3. We gathered the next successive action (change or termination) for each contract within a period of six months, or NA in case no transcation occured.

To link online quotes to existing policies, we compared vectors of joint covariates from both samples, e.g. the covariates date of birth, postal code, family status, home ownership and sum assured for the household insurance product. For the matching of the external vehicle and appartment advertisements to the existing contrats we used the Jaro-Winkler distance function (Winkler, 2006) to match free text covariates, such as name, street and brand and model of vehicle. As well, we compared vehicle registration date, issue date of drivers licencse and home ownership. Moreover, we chose a period of six months to detect the next action, as a similar period was applied by a previous study (Mau et al., 2015).

Prediction

As a final step, we aimed to achieve high prediction accuracy for the forecasts of policy change using random forest (Guelman et al., 2015). In the current stage of the study we applied the random forest to the sample of household insurance customers, where an online quote could be matched. Therefore, the overall data set was split into a training sample and a test or holdout sample. For all included observations the following covariates were selected from the CRM database and used for the tree split: age of customer at inception of existing policy (AGE_E), ratio of contract duration elapsed at time of online quote (CDR_E), time since the last claim (TLC_E), number of online quotes performed (NOQ_D), age of customer at online quote (AGE_D), and sum assured in online quote (SAS_D). The response variable (CAC) measured the customer's action on his existing policy within then next six month using the binary coding: 0 - Action (change or terminate policy), 1 - No action. Our sample was unbalanced (Action: 25.6%, No action: 74.4%) and therefore we used weigthed approach for the random forest. We performed gridsearch over the parameters number of tree ($NTREE$) and the number of covariates randomly chosen for each tree split ($MTRY$) to optimize our random forest. We used the package "randomForest" in the statistical software R for computation.

4 Preliminary results

Matching and detection

Overall, we were able to match online quotes and external data to existing policies and detect the next action (change or termination) of the policy, if occuring, within 6 months. After the matching we divided the resulting observations into three distinct groups. In the first group we were able to link an active policy with an online quote. For household insurance $\sim 11\,000$ cases and for motor insurance $\sim 27\,000$ cases were found. The second group consists of cases where external advertisement could be matched to

active contracts and included 121 cases for household insurance and 288 cases for motor insurance . For the third group (no-match), which represented the majority of the customer portfolio no reaction data could be matched. The results of the comparison for all three groups are shown in Table 2.

Table 2: Comparision of customers' action ratios (change or termination) for the three groups

	No-match	**Online quote**	**delta to no-match**	**External data**	**delta to no-match**
Household insurance					
mean	10.7%	29.5%	+18.8%	57.7%	+47.0%
max	11.6%	36.5%	+24.9%	67.6%	+56.0%
min	8.2%	5.3%	-2.9%	42.9%	+34.7%
Motor insurance					
mean	15.7%	41.7%	+26.0%	87.0%	+71.3%
max	16.7%	49.3%	+32.6%	92.6%	+75.9%
min	12.3%	32.1%	+19.8%	80.0%	+67.7%

We found that the matching of customers' reaction data, either online quotes or external advertisements, lifted the likelyhood for the occurence of a next policy action (change or termination) within the next six months. For household insurance customers the likelyhood of changing or terminating the contract increased to 29.5% when an online quote could be matched and even to 57.7% when an external advertisement could be match. This is an increase of 18.8% respectice 47.0% compared to the ratio of the "no-match" group. For motor insurance the results were similar. In this case 41.7% of the online quote group and 87.0% of the external data group adapted their policy, which is an increase of 26.0% respectice 71.3.0% compared to the "no-match" group. Still, we observed that for online quotes group 70.5% of the household insurance and 58.3% of the motor insurance customers had no action despite the matching. To improve the forecast accuracy of action or no action we applied the data of the household insurance customers where an online quote was matched to a random forest.

Prediction

Therefore we split the online quote group of household insurance customers into a training sample (∼10000 observations) and a test or holdout sample (1000 observations). We grew a random forest to predict, whether a customer will undertake an action (change or terminate) or no action on his existing policy within the next six month.

Table 3: Variable importance for the training set

Covariate	MeanDecreaseGini
AGE_E	1195.5297
AGE_D	1071.2282
CDR_E	1043.3805
SAS_D	854.9983
NOQ_D	231.8069
TLC_E	204.0579

We included the covariates presented in the methodology chapter. The importance of each covariate

for the trees is shown in Table 3. We found the latter two not to impact the overall accuracy, but to impact the balance of precision and recall. This point will be a subject to further investigations in order to optimize the prediction model. During the training the best random forest was found for the parameter $NTREE = 200$ and $MTRY = 5$, which lead to an training accuracy of 86.45%. The figures show that the first four variables have the greatest impact on the tree splits.

This random forest was then used to predict the action of customers' in the test sample. The results are shown in Table 4. Overall, a prediction accuracy of 88.5% was achieved, which was the objective of the study. An area of improvement is the recall (73.24%), as currently about 24% of the customers who are likely to chance their contract are not predicted.

Table 4: Confusion martrix for test set

	Predicted - Action	Predicted - No action	
Actual - Action	219	80	$TPR = 73.24\%$
Actual - No action	35	666	$TNR = 95.01\%$
	$PPV = 86.22\%$	$NPV = 89.28\%$	$Accuracy = 88.50\%$

Note: TPR - true positive rate or recall, TNR - true negative rate or specifity, PPV - positive predictive value or precision, NPV - negative predictive value

5 Summary and Outlook

The objective of the study was to verify whether customers' reaction data could make a contribution to highly accurate forecasts of purchase behaviour for insurance customers. Therefore, we matched online quotes and external advertisment data to the personal data of customers from a Swiss insurer and detected succesive actions (change or termination) on their policy within six months. We found that customers who could be matched successfully to such reaction data, had a higher likelyhood of changing their contract. In a further step, we incorporated these data into a random forest to forecast the customers' action even more accurate. Finally, for our test set we achieved a prediction accuracy of almost 90%. When applying this prediction within the daily business this would mean that an insurance agent would succesfully closes nine deals out of ten predicted customer contacts.

Despite the achieved results our objective is to improve and extend the current work. One goal is to optimize between recall and precison in the random forest. Therefore, we plan to include further covariates as well as adjust the threshold for the prediction probability for action or no-action similar to Abrahams et al. (2009). Moreover, the focus of the customer action will be extended by defining our response variable with three action classes: 1) contract termination, 2) contract change and 3) no action. Such details could provide even more valuable insights of the customers purchase activities for insurers. As well, we plan to conduct the prediction for motor insurance and for case with matched extenal data. Further, we aim to apply additional data mining techniques for classificaton such as neural networks, support vector machines and logistic regression. This approach would allow us to select the best method

for such a domain specific prediction task. At last, we intent to include matching and prediction for cross-sellinig applications, whereas currently our focus is limited to up-selling and retention.

Overall, the usage of customers' reaction data seems to have an impact to generate accurate forecast of purchase activities in the insurance sector and could become a straightforward application in the future.

References

Abrahams, A. S., A. B. Becker, D. Sabido, R. DSouza, G. Makriyiannis, and M. Krasnodebski (2009), Inducing a marketing strategy for a new pet insurance company using decision trees, *Expert Systems with Applications*, 36(2), p.1914–1921.

Fader, P. S. and B. G. Hardie (2009), Probability models for customer-base analysis, *Journal of interactive marketing*, 23(1), p.61–69.

Guelman, L., M. Guilln, and A. M. Prez-Marn (2015), A decision support framework to implement optimal personalized marketing intervention, *Decision Support Systems*, 72, p.24–32.

Kamakura, W. A. (2008), Cross-selling: Offering the right product to the right customer at the right time, *Journal of Relationship Marketing*, 6(3-4), p.41–58.

Li, S., B. Sun, and A. L. Montgomery (2011), Cross-selling the right product to the right customer at the right time, *Journal of Marketing Research*, 48(4), p.683–700.

Mau, S., I. P. Cvijikj, and J. Wagner (2015), From research to purchase: an empirical analysis of research-shopping behaviour in the insurance sector, *Zeitschrift für die gesamte Versicherungswissenschaft*, 104(5), p.573–593.

MSI (2014), Research Priorities 2014-2016, *Marketing Research Institute (MSI)*.

Reimer, K. and J. U. Becker (2015), What customer information should companies use for customer relationship management? Practical insights from empirical research, *Management Review Quarterly*, 65(3), p.149–182.

Smith, K. A., R. J. Willis, and M. Brooks (2000), An analysis of customer retention and insurance claim patterns using data mining: A case study, *Journal of the Operational Research Society*, pages 532–541.

SVV (2014), Schadenversicherung Marktanteile, *Schweizerischer Versicherungs Verband (SVV)- Zahlen und Fakten 2014*.

SwissRe (2012), European Insurance Report 2012: Customers for Life.

SwissRe (2014), Digital distribution in insurance: a quiet revolution, *sigma*, (2).

Verhoef, P. C., S. a. Neslin, and B. Vroomen (2007), Multichannel customer management: Understanding the research-shopper phenomenon, *International Journal of Research in Marketing*, 24(2), p.129–148.

Winkler, W. E. (2006), Overview of record linkage and current research directions, In *Bureau of the Census*. Citeseer.

Wu, C.-H., S.-C. Kao, Y.-Y. Su, and C.-C. Wu (2005), Targeting customers via discovery knowledge for the insurance industry, *Expert Systems with Applications*, 29(2), p.291–299.

Yang, A. S. (2015), Measuring self-service technology latent difficulties: insurance decisions on utilitarian and hedonic influences, *Asia-Pacific Journal of Risk and Insurance*, 9(1), p.1–33.