2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8303

Big data analytics in returns management – Are complex techniques necessary to forecast consumer returns properly?

Asdecker, Björn and Karl, David
Chair of Operations Management and Logistics, University of Bamberg, Germany

Abstract

The more people shop online, the more consumer returns e-tailers face. In order to plan the returns management process capacity adequately, it is necessary to forecast the expected amount of returned parcels. Big data analytics provides a vast number of methods to perform such tasks. However, it should be noted that particularly small- and medium-sized e-tailers lack the capabilities and resources to employ such complex techniques. Against this background, this paper analyses the performance of several data analysis methods that differ in application complexitiy using real data from an apparel e-tailer. On the one hand, we find that –as expected– complex methods outperform simple ones. On the other hand, and from a practitioner's perspective probably even more interesting, we also conclude that a binary logistical regression as the simplest analyzed method may already provide satisfactory results. The findings indicate that the use of big data analytics is of great value to effectively and efficiently manage consumer returns – even if not the most sophisticated state-of-the-art method is used.

Keywords: *returns management; product returns; e-commerce; forecast models.*

1. Introduction and motivation

In today's retailing world, more people shop online more frequently. Consequently, e-commerce revenues have been skyrocketing in the last two decades and an end to this development is not foreseeable. In 2015, for instance, US consumers spent more than $340.41 billion online; up from $4.98 billion in 1998 (United States Census Bureau, 2018). While e-tailing clearly offers numerous advantages over traditional brick-and-mortar retailing, there is one major disadvantage: supply and demand are geographically separated. Therefore, consumers are unable to see, feel, and test the products before purchase. In other words, they can not try before they buy, which almost inevitably leads to consumer returns.

From a business perspective consumer returns are a major cost driver and pose a serious threat on an e-tailer's profitability. According to Stock et al. (2006), expenditures can be as high as $30–35 per return, while return rates may well exceed 50 % for fashion items (Asdecker et al., 2017). These already staggering numbers still reflect only part of the problem. In addition to the direct costs there are indirect effects that influence customer value.

Existing research shows that the returns process is part of the post purchase-experience and herein influences customer satisfaction and retention (Petersen & Kumar, 2009). Laseter and Rabinovich (2012) argue that the improvement of the product return experience is based on the following three principles: (1) lower customer efforts to return the product, (2) offer customized solutions that fit the customers' needs, and (3) exceed customer expectations when processing returns. The third principle specifically refers to the desired outcome (e.g., compensation of made payments), ease of contact, and recovery responsiveness (Mollenkopf et al., 2007). The latter can be operationalized with the time necessary to process a return. To speed returns up, operations require the most accurate capacity planning, which, in turn, is based on forecasts. The importance of this task cannot be underestimated: The better the forecasts are, the more effective and efficient will cosumer returns be processed.

Literature provides various econometric, statistical and/or data mining methods that can be employed to predict returns. Some are higly complex while others are more straightforward and easier to apply. In a world where many decision makers strive for the one optimal alternative, complex state-of-the-art methods seem to be the best choice. However, they also demand sophisticated skills, additional capabilities and financial resources, which are confined, particularly in small- and medium-sized e-tailing companies (Coleman et al., 2016). In a challenging paper, Banks (1993, p. 360) concludes: "I would guess that intelligent use of simple tools will achieve 95 % of the knowledge that could be obtained through more sophisticated techniques, at much smaller cost. Also the simple tools can be applied more quickly to all problems, whereas the complex tools are unlikely to be ubiquitously used." While the paper was written in a different era the general message remains. Against this background, we address the following research-leading question:

- **How are simple data analysis methods performing compared to complex, more sophisticated ones when predicting consumer returns?**

Unlike other publications that try to identify factors that influence the likelihood of consumer returns (e.g., Toktay et al., 2004, Asdecker et al., 2017, Srmiti, 2018), this paper compares the performance of different forecasting methods. To the authors' best knowledge, the publication that is the closest to the one at hand originates from Urbanke et al. (2015). They developed a decision support system that identifies transactions with a high likelihood to return before the actual sale takes place and demonstrated the approach's applicability using a large dataset from a German fashion e-tailer. Within their study they compare seven forecasting techniques, namely the principal component analysis, linear discriminant analysis, randomized truncated singular value decomposition, feature selection based on univariate chi-squared statistic, random projection, non-negative matrix factorization, and a specific feature extraction technique that ignores nominal indicators. While they search for the technique with maximum precision, they do not compare simple with complex approaches.

The remaining article will provide an overview of different forecasting approaches, followed by a report on the performance of the previously introduced techniques. Finally, we conclude with a summary and an outlook on future research.

2. Theoretical background and return forecasting techniques

The data science and statistics literature provides a variety of different methods or techniques that can be used to to predict consumer returns. Since consumers decide to either return or keep the delivered items the dependent variable is binary in nature. This article considers five approaches that are briefly described as follows.

2.1. Binary logistic regression

The binary logistic regression is the simplest method to be taken into account. It is an extension of the linear regression, where the dependent variable is binary (1=return, 0=keep). The independent variables can be either continuous (interval/ratio) or categorial (ordinal/nominal) in nature. For each observation, the binary logistic determines the probability that the dependent variable takes value "1" (Hastie et al., 2009).

2.2. Linear discriminant function analysis

The linear discriminant function analysis, which was first performed by Fisher (1936), shares great similarities with the logistic regression. The analysis requires at least two a priori known groups to which observations shall be assigned. The basic idea is to create a linear combination of independent variables, which best classifies the available data. Thereby, it determines a score for each observation which is then compared to a critical discriminant score in order to carry out the classification (return vs. keep).

2.3. Artifical neuronal network: multiylayer perceptron

Artificial neuronal networks are based on a set of connected nodes, the so called artificial neurons, which are organized in layers. Connected artificial neurons can exchange signals with each other. The receiving artificial neuron will then process it and, in turn, signal artificial neurons connected to it. The ultimate goal is to find a function that best assigns input data to the correct output. To achieve this goal in the returns management context, this study uses multilayer perceptrons, a class of feedforward neural networks. Herein, the information flows exclusively from the input layer through hidden layers with a certain amount of units to the output layer with no feedback flow. Training is done through backpropagation, which is a supervised learning technique that compares the outputs of the network with the known actual values (Hastie et al., 2009).

2.4. Decision tree learning: C5.0 algorithm

Decision trees are hierarchical structures of branches, representing conjunctions of certain characterisics, and leafs, representing class labels. The goal of this technique is to create a decision tree that best classifies the available observation. For this purpose many decision tree algorithms have been presented. This analysis refers to the C5.0 algorithm. C5.0 is the faster, more efficient successor of the widely-employed C4.5 algorithm (Pandya, 2015).

2.5. Ensemble learning technique

The ensemble learning method uses several algorithms to improve predictive performance. It determines the result of every single algorithm and interprets it as a hypothesis for the final verdict (Polikar, 2006). In this study, we used the training data to determine the three most accurate techniques from the following selection: artificial neuronal network, different decision trees (C5.0, QUEST, CART, CHAID), binary logistic regression, linear discriminant analysis, Bayesian network, and nearest neighbor. This resulted in three hypotheses that were given a vote proportional to the confidence, which equals the probability that the postulated hypothesis of a specific algorithm is accurate.

With regard to the required expertise and software, the binary logistic regression and the two-group discriminant analysis are quite straightforward and implemented in common statistical programs with simple and intuitive user interfaces. The remaining three are more complex and therefore require more data mining knowledge as well as less intuitive software (e.g., R, Python, MATLAB, or specific data mining tools such as the IBM SPSS Modeler).

3. Comparison of the different forecasting techniques

The introduced prediction techniques will be tested using real data from a medium-sized German fashion e-tailer that requested confidentiality concerning its name. We will first describe the characteristics of the provided data and then determine the predictive performance of the previously introducted techniques.

3.1. The dataset

The dataset contains shipment and returns information from April 2012 to April 2013. More up-to-date data would be desirable but is unrealistic because many merchants consider returns data as proprietary and are unwilling to share this kind of information. During the analyzed period, the e-tailer sent 220,474 shipments and received 131,907 returns, which equals an α-returns rate of 59.8 % (Asdecker, 2015). While the returns rate might change over time due to different customer behavior or successful avoidance strategies, it should be noted that data age is irrelevant to this type of comparative investigation. The e-tailer shared the following information:

- Package ID: unique ID of the shipment
- Price: Total value of shipped goods in Euro
- Number of articles: total number of articles in the shipment
- Delivery time: time between order placement and delivery in days
- Customer type: categorial variable that indicated the customer type (1=female, 2=male, 3=family, 4=company, 5=unknown)
- Customer age: customer age at the time of order in years
- Account age: age of the customer account at the time of order in days
- Return: binary variable to indicate whether the enclosed return label had been used (1=return, 0=no return)

3.2. Predictive performance

We divided the shared dataset in two parts. The first twelve months, that is, the data from April 2012 to March 2013, were used to derive the prediction model. The last month, April 2013, is the actual test data. Within that period, the e-tailer sent 22,069 shipments and received 13,166 returns. Consequently the α-returns rate is 59.7 %. The criterion used to assess the quality of the forecast is the total absolute error (TAE), which is the summed absolute difference between the predicted and the actual value. This equals the total number of cases that were mispredicted, which means they were either classified as a return even though nothing was returned or vice versa. All five predicting methods were implemented in IBM SPSS Modeler 18, leading to the following results:

- Binary logistic regression, TAE=7,329 (33.21 %)
- Linear discriminant function analysis, TAE=7,372 (33.40 %)
- Artifical neuronal network I (MLP, one hidden layer, six units/layer), TAE=7,364 (33.37 %)
- Artifical neuronal network II (MLP, two hidden layers, six units/layer), TAE=7,146 (32.38 %)
- Decision tree learning: C5.0 algorithm, TAE=6,973 (31.60 %)
- Ensemble learning technique, TAE=6,963 (31.55 %)

Accordingly, the ensemble learning technique provides the best results and predicts 68.45 % of the analyzed cases correctly. The three techniques derived from the training data and used in the ensemble were C5.0, CHAID, and QUEST, which are all decision trees. This finding was expected because the utilization of multiple algorithms within an ensemble usually provides better predictive results than any of the algorithms separately (Polikar, 2006). More surprising is the good performance of the simpler, less sophisticated techniques. Binary logistic regression and linear discriminant function analysis correctly predict 66.79 % and 66.60 % of cases, respectively. This is equivalent to the artificial neuronal network with one hidden layer (66.63 %) and only slightly worse than the neuronal network with two hidden layers (67.62 %).

On the one hand, this generally shows that data mining can be helpful when it comes to plan the return processes and determining the necessary capacity in the returns department. On the other hand, this study also highlights that it does not always have to be the most sophisticated method to generate an acceptable consumer return forecast. In fact, a cost-benefit analysis might favor simple methods, since more sophisticated and complex methods require more resources and data mining knowledge. This holds particulary true for the binary logistic regression, which is not only easy to conduct but also allows for a detailed analysis regarding the factors that affect consumer return behavior. Table 1 summarizes the SPSS report for the statistically significant binary logistic regression ($p=0.000$, Nagelkerke’s $R^2=0.212$) derived from the twelve months training data.

In the model, the probability of a return increases with the total value of the shipped goods, the number of items in a shipment and the account age, whereas it decreases with the delivery time. Packages that are delivered to women have the highest return probability, followed by families, companies, men and unknown recipients. The standardized coefficients b and the Wald statistics show the factor’s relative impact: the biggest effect has the price, followed by the number of items and the customer type.

Table 1. Results of the binary logistic regression model

Variable	b	SE	Exp(b)	Wald	Odds Ratio
Constant*	0.655	0.006	1.926	12,514.985	
Price*	1.267	0.010	3.550	15,552.654	3.550
Nr. of articles*	0.325	0.006	1.384	2,532.507	1.384
Delivery time*	-0.026	0.005	0.975	23.871	0.975
Sent to: Mr.[a]*	-0.437	0.026	0.646	272.223	0.646
Sent to: Family [a]	-0.139	0.086	0.870	2.597	0.870
Sent to: Company[a]	-0.303	0.178	0.739	2.904	0.739
Sent to: Unkown [a]*	-0.554	0.190	0.574	8.525	0.574
Customer age*	-0.065	0.005	0.937	172.983	0.937
Account age*	0.036	0.005	1.036	48.129	1.036

Legend: [a] = reference category: Mrs; * = significant on .05-level;
Nagelkerke’s R² = 0.212; all continuous variables standardized

4. Summary and outlook

Overall, the good classification performance of the parametric methods (binary logistic regression and linear discriminant analysis) surprises. In fact, their performance is only 1.66/1.85 percentage points worse than the ensemble technique as the best nonparametric method. However, their results are easy to interpret and understand. Therefore, simple models such as the binary logistic regression might be the better choice in business practice, especially for small and medium-sized e-tailers that face limited data mining capabilities and financial resources. This holds particularly true because they can also be used for the initiation of preventive returns management measures.

This study is based on real data provided by a German e-tailer. Nevertheless, the scope of the analyzed data was very limited. It would be desirable if future studies had access to additional information, e.g., a customer’s order and return history, shopping basket composition, to substantiate the presented results. With a larger amount of data, it is very likely that the investigated complex techniques can better exploit their advantages. Moreover, this analysis focused on the prediction of return shipments which is important to plan the reverse logistics process. In further research, it may be of interest to take a closer look inside the shipments to extend the analysis to single articles/items.

References

Asdecker, B. (2015). Returning mail-order goods: analyzing the relationship between the rate of returns and the associated costs. *Logistics Research* 8(3), 1–12.

Asdecker, B., Karl, D., & Sucky, E. (2017). Examining Drivers of Consumer Returns in E-Tailing with Real Shop Data. *Proceedings of the 50th Annual Hawaii International Conference on System Sciences (HICSS)*, 4192–4201.

Banks, D. (1993). Is Industrial Statistics Out of Control? *Statistical Science* 8(4), 356–377.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorelle, X., & Reis, M. S. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International* 32(6), 2151–2164.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7(2), 179–188.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Laseter, T. M., & Rabinovich, E. (2012). *Internet Retail Operations – Integrating Theory and Practice for Managers.* Boca Raton: CRC Press, Taylor & Francis Group.

Mollenkopf, D. A., Rabinovich, E., Laseter, T. M., & Boyer, K. K. (2007). Managing Internet Product Returns: A Focus on Effective Service Operations. *Decision Sciences* 38(2), 215–250.

Petersen, J. A., & Kumar, V. (2009). Are Product Returns a Necessary Evil? Antecedents and Consequences. *Jounal of Marketing* 73(3), 35–51.

Pandya, R. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications* 117(16), 18–21.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6(3), 21–45.

Srmiti, K. (2018). Predicting Online Returns. In Kumar, A. & S. Saurav (Eds.), *Supply Chain Management Strategies and Risk Assessment in Retail Environments* (pp. 181–194). Hershey: IGI Global.

Stock, J., Speh, T., & Shear, H. (2006). Managing product returns for competitive advantage. *MIT Sloan Management Review* 48(1), 57–62.

Toktay, L. B., van der Laan, E. A., & de Brito, M. P. (2004). Managing Product Returns: The Role of Forecasting. In Dekker, R., Fleischmann, M., Inderfurth, K. & L. N. Van Wassenhove (Eds.), *Reverse Logistics* (pp. 45–64). Berlin: Springer.

United States Census Bureau (2018). U.S. Retail Trade Sales - Total and E-commerce (1998-2015). Retrieved February 22, 2018, from http://www2.census.gov/retail/releases/current/arts/ecommerce.xls.

Urbanke, P., Kranz, J., & Kolbe, L. (2015). Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction. *Proceedings of the 36th International Conference on Information Systems (ICIS)*, 1–12.