Shortening Delivery Times by Predicting Customers' Online Purchases: a Case Study in the Fashion Industry

Jennifer Weingarten
WHU - Otto Beisheim School of Management
jennifer.weingarten@whu.edu

Stefan Spinler
WHU - Otto Beisheim School of Management
stefan.spinler@whu.edu

Abstract

Online retailers still struggle with the disadvantage of delivery times compared to traditional brick and mortar stores. With the emergence of big data analytics, it has become possible to extract meaningful knowledge from the volume of data that online retailers collect on their website. Nevertheless, limited research exists that investigates how this data can be used to optimize delivery times for customers. The goal of this paper is to develop a prediction model for anticipatory shipping, which predicts customers' online purchases with the aim of shipping products in advance, and subsequently minimize delivery times. Different forecasting methods in combination with k-means clustering are applied to test if, and how early, it is possible to predict online purchases. Results indicate that customer purchases are, to a certain extent, predictable, but anticipatory shipping comes at a high cost due to wrongly sent products. The proposed prediction model can easily be implemented and used to predict purchases, which can also be leveraged for other areas of application besides anticipatory shipping.

1. Introduction

In recent years, the interest for big data (BD) has been growing from both academia and the industry [1]. BD is often defined in terms of 5 V's: volume, velocity, variety, veracity, and value [2]. Volume refers to the quantities of data, which require a massive amount of storage. Variety refers to the different types of data collected, which can be structured (e.g. customers' demographic data) and unstructured (e.g. likes, tweets). Velocity stands for the speed of data generation and processing in (near) real-time. Veracity stresses the importance of data quality. Lastly, value refers to the process of extracting value from BD to aid decision-making [1, 2]. BD provides tremendous opportunities as it is widely available and nowadays much less expensive to access and store [3]. Due to the large volume of data, the variety of data sources and the speed at which data needs to be collected and analyzed, big data analytics (BDA) has emerged. BDA involves the application of advanced analytics techniques, such as statistics, simulation or optimization, to gain insight from big data to enhance decision-making and increase business value and firm performance [4]. Businesses that already use BDA report a 5% increase in productivity and 6% increase in profitability, compared to those that do not [3]. In supply chain management (SCM), analytics and data-driven decision-making are not novel. Techniques such as statistics and simulation have frequently been used in the past to optimize the supply chain [4]. However, the exponential increase in big data generated from end-to-end supply chain management creates new opportunities, as well as challenges, as companies are faced with the difficulty of mining large datasets [4]. As supply chain performance depends to a large degree on information, BDA could be especially beneficial for SCM. Nevertheless, the research of the application of BDA in SCM is still in its infancy [5].

BDA has also been emphasized in the e-commerce context, where big data enables online sellers to track each user's behavior, which provides companies with opportunities such as real-time customer service, dynamic pricing or personalized promotion activities [1]. While the time between purchase and product arrival used to be the main disadvantage of e-commerce players compared to brick and mortar stores, last mile solutions, such as same day or 2-hours delivery, enable almost instant gratification for consumers [6]. To enable nearly instant delivery services, products need to be stored close to the consumer [7]. The large assortment of many e-commerce players, such as Amazon or Alibaba, makes this especially difficult. While many online retailers have been forward-deploying inventory to enable fast delivery [7], Amazon has been using BDA to predict customers' purchase behavior and as a result, ship products closer to the customer before they place their order online. Amazon has patented this approach

URI: https://hdl.handle.net/10125/63898
978-0-9981331-3-3

as 'anticipatory shipping' (AS) [8]. Whether Amazon is successful with this, and whether predicting costumers' purchase behavior is possible to the extent that it enables the successful shipping of products in advance, is, to the best of our knowledge, not known. To better understand possibilities for AS, this paper investigates the predictability of customers' purchase behavior using BDA, and tests in a case study how AS would impact delivery time and cost. Specifically, structured data, for instance customer age and gender, as well as unstructured data, such as customers' online browsing behavior, are used to predict customers' purchases. The research questions that guide this study are:

1. To what extent can customer information and browsing behavior be used to anticipate consumer purchases to ship products in advance and subsequently decrease delivery time?
2. What is the optimal point in time to predict customer purchases?
3. What is the operational value of using predicted purchases for AS?

The structure of the paper is as follows: Section 2 reviews literature on the application of BDA in SCM and e-commerce and outlines approaches for AS. Section 3 introduces the case study context and explains the applied research approach. Section 4 presents the results and discusses managerial implications. Section 5 concludes the paper and gives an outlook on areas of future research.

2. Literature review

2.1. BDA and its application in SCM

A widely adopted taxonomy of BDA is the classification into descriptive, predictive and prescriptive analytics. Descriptive analytics gives insights into past events, predictive analytics makes predictions about future events and prescriptive gives recommendations for future actions to support decision-making [2]. Applications of all three types of BDA can be found across the entire spectrum of SCM. For a detailed list, see Nguyen and Zou [2].

BDA is currently a vividly discussed topic among scholars, due to its wide area of application. Its usage in SCM still provides many areas for future research. A Delphi study from Kache and Seuring [5] that investigated opportunities and challenges of BDA, showed that 'customer behavior' and 'logistics' are two key opportunity areas of big data analytics. To further investigate opportunities in this area, this paper is positioned at the intersection of customer behavior and logistics.

2.2. BDA in e-commerce

E-commerce players typically deal with two types of data: structured (e.g. customer age, gender) and unstructured (e.g. clicks, likes, tweets), where the challenge in BDA lies in creating meaningful insights from the combination of the two [1]. Typical applications of BDA in e-commerce are the identification of customer needs, market segmentation, or making relevant information available at the right time [1]. An example of the latter is Amazon's recommendation system that recommends products to customers, based on an understanding of their preferences [9]. Studies from various disciplines offer insights into customers' online shopping behavior and conversion likeliness, but usually not with the aim of improving supply chain performance. Moe [10] developed a model to predict purchase probabilities for a given site visit, to re-direct visits with a high purchase probability to a better performing server. Overall, applications where the combination of structured and unstructured data has been used to predict customers' online behavior and subsequently improve supply chain performance are scarce. An example can be found in Huang and Van Mieghem [11], who used click and order data to predict the propensity, amount, and timing of offline orders to improve inventory management.

2.3. Anticipatory shipping

Amazon has patented an approach for AS, in which the company uses big data, including order history and data from its e-commerce portal, to predict a customer's online purchases and ships products to a geographical area close to the customer. The final delivery address is not completely specified until the customer places the order online [8].

Not much research regarding AS can be found in the literature. Lee [12] developed a model for AS in an omni-channel context. The study uses associate rule mining based on the Apriori algorithm to predict orders within pre-defined clusters of demand points, to ship products to the nearest distribution center in advance. A genetic algorithm is then applied to optimize AS in the distribution network. Viet, Behdani, and Bloemhof [13] present a model for AS in the agro-food industry. They also apply associate rule mining but add a time threshold to take into account product perishability. Both papers use historical orders as input to associate rule mining to identify potential products and volumes for AS, assuming that association rules (e.g. 'if product A is purchased, product B is likely to be purchased later as well') can be found in the

historic data, that are applicable to future orders. We believe that this approach is not suitable for the fashion industry where retailers have enormous, frequently changing assortments with few data points (e.g. past orders) available for each product, limiting possibilities to find association rules. Therefore, this paper uses classification instead of association methods and leverages data such as customers' browsing behavior to determine products for AS.

3. Methodology

3.1. Case study context and data

The data for analysis is provided by an online retailer in Europe that mainly sells fashion items. As most online retailers, the case company tries to minimize delivery time. Consequently, the case company is interested in using predictive analytics to explore opportunities to decrease delivery times. For confidentiality reasons, no further information regarding customer and warehouse locations can be given. The data received spans over a time period of one year and includes five types of datasets, which can all be linked via pseudonymized customer IDs:

- Customer information: gender, sign-up year, segment (mainly dependent on profitability)
- Order information: Order date, products ordered, total number of orders per customer
- View information: Number of product page visits of a customer, date and length of visit
- Event information (information on where a customer clicked on a product page): event type (e.g. 'click on image', 'add to cart'), event date, total number of clicks of a customer
- Product information: product category

Additionally, the season, month and weekday on which a customer viewed a product for the first time were included as variables. Moreover, the number of times a customer opened a product page, and the total number of events that occurred on a product page by a customer were calculated. Lastly, the average decision time per customer was calculated, which is the average time between the first date a product was viewed and the order date.

The data protection principles of the General Data Protection Regulation (GDPR) were strictly followed so that any personal data received was in a form which does not permit identification of data subjects. Data was maintained and encrypted using Advanced Encryption Standard 256-bit encryption.

As the decision for repeat purchases (e.g. due to the wrong size) is assumed to be different from the decision to order a product for the first time, all data that occurred after a costumer purchased a product for the first time, were excluded from analysis.

In order to predict whether a customer will buy a product, a prediction on a customer-product level is essentially made. One observation in the constructed dataset thus contains the views and events that happened between a customer and a product page, combined with the general information on that specific customer and product. The response variable 'purchase decision' is a binary variable that falls into one of two categories, yes (1) or no (0), indicating whether an observation led to a purchase. We are thus faced with a binary classification problem. Due to the size of the dataset, a random sample of 100 thousand (k) customers was selected and led to a total of 8.3 million observations. From those, only 3.8% resulted in a purchase, indicating that the dataset is imbalanced as the classification categories are not equally represented.

3.2. Research approach

This paper follows a three-step approach (Figure 1), with the goal to predict online purchases as early and accurately as possible to enable advanced shipment of products, while minimizing the number of products that are erroneously sent in advance with no subsequent order. In the first step, five different forecasting methods are applied to different datasets to evaluate which forecasting method and dataset yield the best results in terms of prediction accuracy. The first dataset consists of all observations, the second one contains only observations from customers with frequent purchases (at least 12 per year) and lastly, the dataset is split into clusters and each cluster is predicted separately.

In this first step, the whole one-year period of the dataset is used, and observations are split into training, validation, and test data. This essentially means that the prediction for purchases is made at the end of the one-year time period. To actually achieve delivery time savings, the best performing forecasting method and dataset from step one are used to predict purchases at an earlier point in time, namely 1) at the end of the first day a customer viewed a product and 2) right after a first 'add to cart' click occurred. Lastly, the impact of the forecasting methods is estimated in terms of packages sent (in-)correctly to better understand the real-life application of such a prediction model.

As the dataset is imbalanced, techniques for dataset balancing, such as over- and undersampling, were investigated. Dataset balancing would cause an increased bias towards the minority class. However,

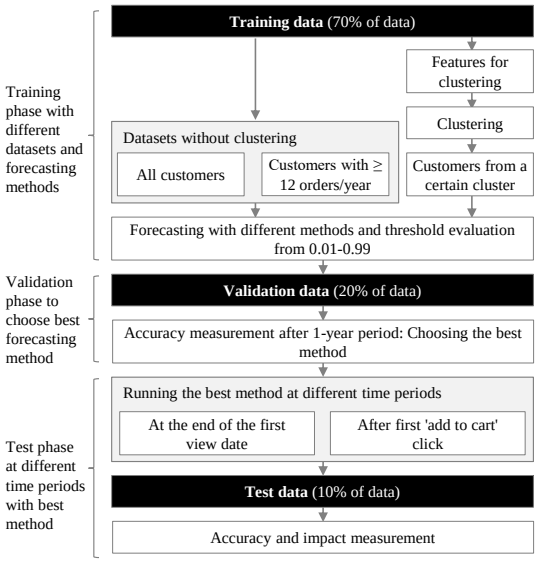

Figure 1. Research approach

the aim is to predict as many purchases as possible while trying to predict observations that do not lead to a purchase as accurately as possible, to avoid erroneously sending a large number of products in advance. In this case, a bias towards the majority class is beneficial, which is already achieved by the currently imbalanced dataset [14]. Balancing techniques were hence not applied.

3.3. Forecasting methods and accuracy measure

In the following, the different forecasting methods that were applied will be explained, namely logistic regression (LG), random forest (RF), neural network (NN) and (one-class) support vector machine (SVM). This is followed by an explanation how variables were selected (feature selection) and lastly, how the accuracy of the methods was assessed and which assumptions were made.

3.3.1. Forecasting methods LG is a statistical method that can be used for classification. It estimates the probability of an observation belonging to one of two classes, using the logistic function [15]. RF is a supervised learning method that constructs an ensemble of decision trees. At each split in a decision tree, a random subset of variables is considered and the data is split in a way that homogeneity of the daughter nodes is improved compared to the parent node. For classification, each tree casts a vote for the predicted class and a majority vote is taken [16]. RFs are a popular learning method as they are simple to train while yielding high accuracy [17]. In the class of feed-forward neural networks, the multilayer perceptron was applied, which is the most commonly used form of neural networks. A multilayer perceptron consists of multiple neurons (nodes) arranged in several layers. It learns the relationship between the variables and the response variable through backpropagation. SVM is another supervised learning method used primarily for binary classification. SVMs plot each observation as a point in an n-dimensional space, where 'n' is the number of variables. They create an optimal hyperplane that separates data points into two classes. If the data points are not linearly separable, SVMs map the data points to a higher dimensional space to enable separation [15]. The four methods were selected as they are well-known in the context of two-class classification, but yield different results in terms of accuracy and training times depending on the input data (e.g. size of training data, number of features, etc.) [17].
Most methods for classification do not work well if any class is heavily undersampled. For this reason, one-class SVM was tested as additional method. One-class SVMs construct a decision boundary around the majority class to differentiate it from observations in the minority class, which are considered outliers or anomalies [18].
All methods were implemented in R-3.5.1 and constructed to predict the probability that an observation belongs to one of the two classes. The threshold, which determines at which probability an observation is considered to lead to a purchase, was evaluated between 1-99%, to evaluate which value leads to the highest overall accuracy. All predictor variables were standardized, except for RF.

3.3.2. Feature selection Correlation analysis was performed to remove highly correlated variables ($>$0.7). Additionally, lasso linear regression and RF were performed. Lasso uses a penalty term (lambda) for regression coefficients and can drive coefficients of non-relevant variables to zero, hence essentially excluding variables. RF assesses the mean decrease in accuracy if a variable is excluded from analysis. Additionally, a set of variables was selected that excludes categorical variables with many categories as they substantially increase training times. This resulted in five different sets of variables that were tested with each forecasting method:

- Set 1: All variables
- Set 2: All variables except categorical variables with $>$53 categories
- Set 3: All variables in lasso output using

lambda.min (minimum error observed)

- Set 4: All variables in lasso output using lambda.1se (error is within 1 standard error of minimum error)
- Set 5: All variables in random forest output with mean decrease in accuracy $>0.01\%$ (as this already led to a reduction of $\sim 50\%$ of variables)

3.3.3. Accuracy measure A confusion matrix is often used to evaluate the performance of classification models. In this study, five measures from the confusion matrix were used to assess model performance: accuracy, sensitivity, specificity, precision and prevalence. Accuracy measures the overall proportion of correct classifications. Sensitivity assesses the proportion of observations resulting in a purchase that the classifier correctly predicted as such, while specificity measures the proportion of observations not resulting in purchases that the classifier correctly predicted as such. Precision measures from all observations the classifier predicted as purchase, the proportion that resulted in a purchase. Lastly, prevalence assesses the proportion of observations that resulted in a purchase. As the accuracy measured with a confusion matrix is often not appropriate for imbalanced datasets, the area under the precision-recall curve (AUPR) was calculated as additional performance measure, which assesses the trade-off between precision and sensitivity (also called recall) [19].

3.4. Clustering

According to Chen and Lu [20], clustering can improve the performance of classification models. Therefore, k-means clustering was performed with the aim of grouping together customers with similar purchasing patterns, into a number of k pre-specified clusters. From the predictor variables, those that describe the customer as such were used for clustering (e.g. average decision time). To represent the relation between the set of 100k customers and 8.3 million observations, the variable 'purchase decision average' was used as additional input. It assesses the share of a customer's observations that led to a purchase. A good clustering is achieved when the within-cluster variation is as small as possible [15]. The within cluster variation was assessed for k running from 1 to 10 to determine the optimal number of clusters. Subsequently, forecasting methods were used to train and predict each cluster separately.

3.5. Assumptions

This paper assumes that the purchase behavior of a customer does not substantially change over time, hence one observation could be assumed to be from a time period outside that of the dataset. That is why the split into training, validation and test data in this paper does not take into account any temporal order of observations, as would be done for time-series forecasting.

4. Results and managerial implications

4.1. Feature selection

Six variables showed high correlation: The number of times a customer opened a product page and the total time a customer viewed a product (r = 0.78), the total events and total views of a customer during the one-year time period (r = 0.80), and the number of times a customer opened and closed the image gallery on a product page (r = 0.84). Lasso using lambda.1se was most aggressive in terms of feature selection and resulted in a set of 23 variables, while lasso using lambda.min resulted in 34 variables. In order to perform RF, the variables relating to product category had to be reduced to 53 categories for implementation in R. Those product categories with a small amount of observations were hence removed until the maximum number of 53 categories was reached, resulting in a 25% decrease in dataset size. Applying RF resulted in 24 variables with a mean decrease in accuracy $>0.01\%$. The variable with the largest decrease in accuracy is, as can be expected, the 'add to cart' click (2.23%). As incorporating the product category variables with more than 53 categories led to a such a substantial decrease in dataset size, those two variables were excluded from analysis using RF, resulting in variable set 5b (e.g. Table 1).

4.2. Discussion of results from different forecasting methods

4.2.1. Dataset 1: all customers The size of the training dataset was too large for most of the forecasting methods, hence the required training size to obtain meaningful results was estimated first using logistic regression, resulting in $\sim$115k observations. Using less observations led to an outcome where each observation was predicted with 'no purchase'. To be more conservative, a training size of 175k observations was used across all methods for comparison reasons. Afterwards, training size was increased for each method, using the optimal choice of parameters and variable set, until no more significant improvements

in accuracy were achieved or model training resulted in an error, for example due to non-convergence of algorithms.

First results from one-class SVM indicated that the model is not appropriate for this particular binary classification problem, as prediction accuracy was exceptionally low. A large fraction of non-purchases was not identified, leading to low specificity. One-class classification is typically used if one class is sampled well, while the other class is heavily undersampled [21]. While the dataset is imbalanced, the minority class still has a large number of observations, due to the size of the dataset. This could explain why other models showed better performance. Moreover, observations from the minority class might be too similar to the majority class to be considered as outliers in a one-class SVM model. One-class SVM was therefore not further applied in the analysis.

If all observations were predicted to be non-purchases, an accuracy of 96.23% would be achieved. Any method resulting in accuracy above that was thus considered to be adding value. The best results were achieved by RF, which also had the fastest training times (Table 1). RF achieved an accuracy of 96.95% (AUPR: 58.83%), using 10 variables available for splitting at each tree node and 500 trees. The model is able to predict almost 48% of all purchases and almost 99% of all non-purchases correctly. From all 'yes purchase' predictions, the model was correct approximately 63% of the time. However, the model seems to be overfitting, as the accuracy of training data prediction is 100%. According to Breiman [16], RFs always converge so that overfitting is not an issue. Increasing the minimal size of terminal nodes was tested, which lowered the prediction accuracy of the training data, but did not improve validation data prediction accuracy, thus overfitting indeed seems to be no issue. NN resulted in 96.58% accuracy (AUPR: 48.95%), using one layer of five hidden neurons. SVM achieved 96.51% accuracy (AUPR: 45.73%), while LG resulted in an accuracy of 96.42% (AUPR: 45.90%), using a radial kernel, cost of 1 and gamma of 0.01. As the variable 'add to cart' was determined most important by the RF importance measure, it was tested whether solely using this variable would be sufficient to achieve high prediction accuracy. This led to a much lower accuracy (96.24%) and an AUPR of 7.6%, indicating that the remaining list of predictor variables add substantial value in combination with 'add to cart'.

After increasing training data size, the accuracy of RF improved to 97.20% (AUPR: 63.79%) using $\sim$ 870k observations (Table 2). The performance of NN could not be improved, as larger training data sizes did not produce algorithm convergence. Increasing training size for SVM only showed a small improvement, resulting in 96.54% accuracy (AUPR: 46.89%). Lastly, the accuracy of LG remained as before (96.42%), with a slight increase in AUPR (46.48%). As RF outperformed all other models, it was used for the remaining analyses.

4.2.2. Dataset 2: customers with high order frequency Prediction of customers with high order frequency using RF with an increased training data size shows an overall accuracy of 96.96% and an AUPR of 67.31% (Table 3). The accuracy should not be compared to the accuracy of dataset 1 as this dataset has a much higher prevalence, meaning more observations lead to purchases. Instead, AUPR is used for comparison, showing a higher value than for dataset 1 (Table 2), indicating that customers with high order frequency are easier to predict. In the subsequent sections, however, we continue to use dataset 1 to further test the application of AS across all customers.

4.2.3. Dataset 3: impact of clustering on prediction accuracy Assessing the within-cluster variation showed that for more than five clusters, there is only a small reduction in within-cluster variation. K =

		Validation data						Training data			
Method	Variable selection	Overall accuracy	Sensitivity	Specifity	Precision	Prevalence	AUPR	Overall accuracy	Sensitivity	Specifity	Precision
RF	Set 5b	96.95%	47.51%	98.89%	62.57%	3.77%	58.83%	100.00%	99.94%	100.00%	100.00%
NN	Set 2	96.58%	38.74%	98.84%	56.73%	3.77%	48.95%	97.13%	45.63%	99.09%	65.61%
SVM	Set 2	96.51%	24.06%	99.35%	59.00%	3.77%	45.73%	97.46%	40.89%	99.61%	79.98%
LG	Set 5	96.42%	27.76%	99.11%	54.94%	3.77%	45.90%	96.57%	28.42%	99.16%	56.32%
RF	Add to cart	96.24%	3.47%	99.87%	51.56%	3.77%	7.60%	96.35%	3.76%	99.87%	52.75%

Table 1. Results dataset 1: all customers (175k training observations)

Method	Variable selection	No. of training observations	Validation data						Training data			
			Accuracy	Sensitivity	Specifity	Precision	Prevalence	AUPR	Accuracy	Sensitivity	Specifity	Precision
RF	Set 5b	~870.000	97.20%	50.36%	99.04%	67.22%	3.77%	63.79%	99.98%	99.49%	100.00%	99.94%
NN	Set 2	~230,000	Algorithm did not converge									
SVM	Set 2	~350,000	96.54%	27.67%	99.24%	58.78%	3.77%	46.89%	97.33%	41.00%	99.51%	76.27%
LG	Set 5	~2,300,000	96.42%	28.23%	99.09%	54.95%	3.77%	46.48%	96.41%	28.04%	99.08%	54.38%

Table 2. Results dataset 1: all customers (larger training data size)

5 was thus chosen as optimal number of clusters. Predicting five clusters separately led to an overall prediction accuracy of 97.16% (AUPR: 63.29%) (Table 4), indicating that clustering did not improve model performance. Figure 2 shows how customers in those five clusters differ. For confidentiality reasons, the variable sign-up year was adjusted so that the earliest sign up year corresponds to 'year 1'. Cluster 1 contains customers that signed-up several years ago and show an average order frequency. Decision time ranges from slow to fast. Cluster 2 are rather new customers that have not been buying much yet, while cluster 3 consists of customers with high order frequency that make fast purchase decisions. The remaining two clusters are not noticeable much in Figure 2 as cluster 4 contains customers that have not purchased anything yet and viewed very few products, and cluster 5 consists of customers with few product views that bought 1-2 products on the same day they viewed the products for the first time.

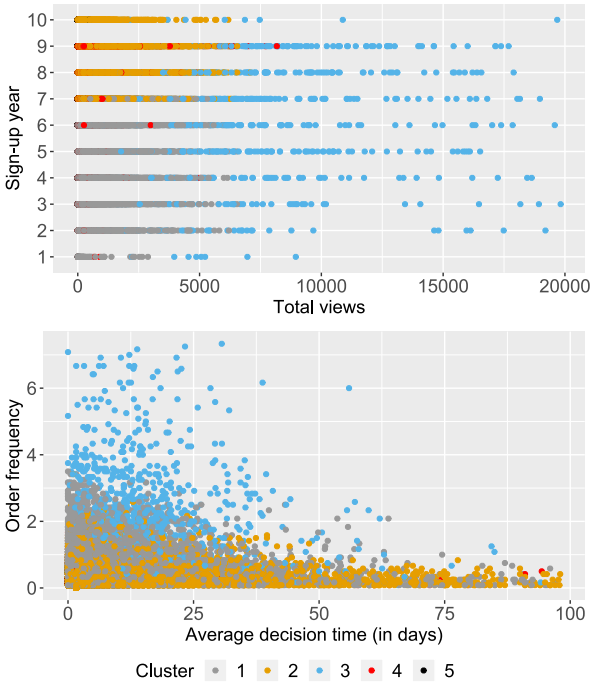

Figure 2. Cluster differences

4.3. Prediction at different points in time

Successfully predicting customer purchases at the end of the first view date is not possible according to the results (Table 5). The prevalence of this dataset is much lower as the data does not contain orders that occurred on the same day as the first view date. Model accuracy is only 98.44% (AUPR: 20.83%), which is almost the same as predicting 'no purchase' for all observations. Predicting a purchase right after an 'add to cart' click yields much better results (Table 5). Accuracy is 76.08% (AUPR: 76.46%) compared to an accuracy of 59.28% if 'no purchase' was predicted for all observations. In this dataset, prevalence is much higher as the data only consists of observations that contain an 'add to cart' click. To further improve model performance, the additional variable 'add to cart conversion' was added, which measures the proportion of a customer's 'add to cart' clicks that led to a purchase. This resulted in an accuracy of 77.56% (AUPR: 81%). Those values outperform all previous results. In all three cases, clustering did not improve model performance.

Method	Variable selection	No. of training observations	Validation data						Training data			
			Accuracy	Sensitivity	Specifity	Precision	Prevalence	AUPR	Accuracy	Sensitivity	Specifity	Precision
RF	Set 2	~870.000	96.96%	56.87%	98.79%	68.32%	4.37%	67.31%	99.10%	82.36%	99.87%	96.61%

Table 3. Results dataset 2: customers with high order frequency (random forest with variable set 2)

			Validation data						Training data			
Cluster	No. of training observations	% of test dataset	Accuracy	Sensitivity	Specifity	Precision	Prevalence	AUPR	Accuracy	Sensitivity	Specifity	Precision
1	~210.000	32.52%	95.74%	54.38%	98.31%	66.68%	5.85%	64.41%	99.97%	99.54%	100.00%	99.95%
2	~230.000	36.36%	97.41%	44.28%	99.14%	62.70%	3.16%	56.15%	99.99%	99.59%	100.00%	100.00%
3	~220.000	25.49%	98.11%	44.25%	99.36%	61.61%	2.26%	55.86%	99.99%	99.63%	100.00%	99.96%
4	~200.000	5.43%	100.00%	25.00%	100.00%	50.00%	0.00%	19.51%	100.00%	100.00%	100.00%	100.00%
5	~12.000	0.20%	87.44%	93.71%	74.31%	88.42%	67.68%	95.17%	99.00%	99.79%	97.42%	98.71%
Weighted average			97.16%	46.61%	98.92%	63.08%	3.77%	63.29%	99.98%	99.61%	99.99%	99.97%

Table 4. Results dataset 3: Clusters (random forest with variable set 5b)

			Test data						Training data			
Dataset	Method	No. of training observations	Accuracy	Sensitivity	Specifity	Precision	Prevalence	AUPR	Accuracy	Sensitivity	Specifity	Precision
First view date	RF	580.000	98.44%	1.39%	99.99%	61.97%	1.56%	20.83%	99.47%	66.00%	100.00%	99.98%
Add to cart	RF	440.000	76.08%	68.18%	81.51%	71.69%	40.72%	76.46%	99.23%	98.73%	99.57%	99.36%
Add to cart*	RF	440.000	77.56%	65.89%	85.57%	75.82%	40.72%	81.00%	99.17%	98.42%	99.68%	99.53%

Table 5. Results at the end of first view date and after 'add to cart' click (random forest with variable set 5b)

4.4. Application to the future

As explained in Section 3.5, the split into training, validation, and test data that has been applied so far assumed that purchase behavior of a costumer does not substantially change over time. To test whether this assumption holds true, the dataset was split into training and test data that respect the temporal order of observations, meaning that the training data now consists of observations with a first view date in the first half of the one-year time period, while the test data contains observations with a first view date in the last quarter. Using RF with the original training data size of 175k observations, an AUPR of 64.60% can be achieved, compared to the 58.53% from dataset 1 using RF. The AUPR shows that the prediction model is also applicable to future data.

4.5. Impact estimation and managerial implications

To estimate the impact, the results have to be translated into products sent (in-)correctly. For 100k customers, the RF results of dataset 1 would have led to 157k products being sent in advance correctly within one year, and 77k products would have been sent without the customer buying it, creating unnecessary logistics cost. Essentially, for every 100 products sent correctly, 49 are sent erroneously. A share of the latter could eventually be bought by a different customer from a similar region, mitigating the cost of products sent incorrectly. On the other hand, shipping products to a different location might result in insufficient stock for purchases from the region those products were originally shipped from. The impact of this could not be estimated with the given dataset. The best impact is achieved when predicting after an 'add to cart' click (including the new variable 'add to cart conversion'). 169k products would have been sent correctly and 54k incorrectly, translating into only 32 products sent by mistake for every 100 products sent correctly. To lower the cost of erroneously sending products, the impact of different thresholds can be tested. A 90% threshold would have led to only 11 products sent erroneously for every 100 products sent correctly. Increasing the threshold, however, also leads to a smaller number of purchases identified (60k instead of 169k).

Studies have shown that faster delivery times lead to a lower number of returns [22]. The logistics cost from wrongly predicted purchases could potentially be reduced through savings in product returns, if time savings are, for instance, leveraged to reach same-day delivery cut-off times. This could also be enabled through anticipatory picking and packaging instead of anticipatory shipping. Model application could also be limited to customers with high order frequency to reduce

the cost of erroneously sending products, as results indicate that those customers are easier to predict.
It should be noted that for 60% of the correctly predicted purchases, a delivery address was not known. The reason for this can be that the delivery addresses of new customers are not known yet. Encouraging website visitors to sign up early and provide a future delivery address, as well as leveraging Google Analytics' location reporting, could mitigate this issue.
When investigating the delivery time savings of predicting after an 'add to cart' click, meaning the time saved between an 'add to cart' click and actual order, it becomes apparent that the time difference is often too short to send a product closer to the customer (Figure 3). Only $\sim$15% of purchases predicted correctly would have resulted in delivery time savings of more than one day. The location of the retailer's warehouses and customers plays a major role in deciding whether the approach presented could be used for AS. For retailers with few warehouses and a wide-spread customer base, resulting in long transportation time, AS could be very difficult to implement. In this paper, transportation time was not taken into account as too many delivery addresses of customers were not available at the time of prediction to be able to assess whether the time for AS would have been sufficient.
The case company will use the results of this research to estimate the business case for AS for varying threshold levels. As many fashion retailers, the case company has a large volume of order movements between warehouses to avoid sending orders containing several items in various parcels. Implementing AS could also have a positive effect on the number of order movements.
In terms of application, we believe that the approach should be equally applicable to other online fashion retailers. Variables that were listed as most important by RF (e.g. 'add to cart' click, average decision time) are all data points that other fashion retailers should be able to obtain. A limitation here is that retailers need to have sufficient website traffic (i.e. customers, clicks and orders) to generate enough observations for the algorithms to deliver meaningful results.
Being able to forecast customer purchases could have many areas of application besides AS. Especially in the fast fashion industry, it could be leveraged to reorder products, which are likely to sell out quickly, in advance. Also, it could be used in returns management to redistribute returns to warehouses where sales are likely to occur in the near future. Whether the proposed prediction model can be used in other fields cannot be answered with this study. The purchase decision for a fashion item might be very different from other types of products. As an example, tracking data collected from a consumer electronics seller might not reveal much of a customer's purchase intentions.

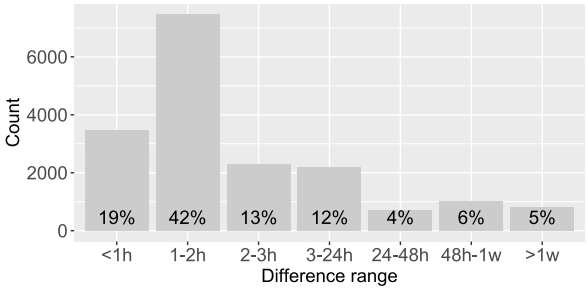

Figure 3. Time difference between 'add to cart' click and purchase for correctly predicted purchases

Lastly, it should be noted that the prediction model is fully dependent on the quality and reliability of the input data. Issues in tracking, for example due to programming errors, customers not being logged in, or click types frequently being renamed, are typical challenges that can arise in trying to obtain reliable data.

5. Conclusion and future research directions

This paper showed how forecasting methods can be applied to predict customers' future purchases to generate delivery time savings. Logistic regression, neural network, (one-class) support vector machine and random forest were applied. Random forest strongly outperformed all other models in terms of accuracy and speed, indicating that the other models are not suitable in this context. When predicting purchases after an 'add to cart' click using RF, $\sim$ 66% of purchases could be correctly predicted. Clustering as input for forecasting did not lead to accuracy improvements. From the results, we conclude that online purchases are, to a certain extent, predictable, but AS still comes at a high cost. Due to the low number of product site visits that convert into purchases, even a 99% accuracy of predicting those non-purchases correctly results in many products wrongly sent in advance. For the case company, the model would have resulted in 169k products correctly sent in advance throughout the year, and 54k products sent incorrectly. As some of the wrongly sent products could be purchased by a different customer in the same area, or create insufficient stock for customer purchases from the original area, the true cost of AS could not be assessed with the given data.
AS generally leads to high logistics cost, despite good prediction accuracy. In combination with the fact that for only 15% of all correctly predicted purchases, the time saved would have been more than a day, the model

could potentially be better leveraged for alternatives to AS, such as anticipatory picking or packaging.

The results of this study are limited by the data quality provided. Improved data quality would likely result in higher prediction accuracy. 18% of the orders in the data did not have an 'add to cart' click, which is necessary to purchase a product. Customers not being logged in or not registered yet are most likely the main causes for this. With 'add to cart' being the most important variable in the application of random forest, such data issues clearly limit model performance.

In terms of future research directions, five areas of interest can be highlighted. First, this research only studied to which extent online purchases can be predicted. Understanding in which quantity and size a customer will buy a product is additional input required to enable AS. Second, being able to predict when a purchase will occur could help determine if delivery time savings are sufficient to send a product in advance. Third, additional data related to pricing, marketing campaigns, fashion trends, and weather data, among others, could be further studied to assess their impact on prediction accuracy. Fourth, increasing the forecast horizon could help better capture seasonality trends. Lastly, further research could be conducted to better understand underlying order patterns of customers and use this information to improve prediction accuracy.

References

[1] S. Akter and S. F. Wamba, "Big data analytics in e-commerce: a systematic review and agenda for future research," *Electronic Markets*, vol. 26, no. 2, pp. 173–194, 2016.

[2] T. Nguyen, L. Zhou, V. Spiegler, P. Ieromonachou, and Y. Lin, "Big data analytics in supply chain management: a state-of-the-art literature review," *Computers & Operations Research*, vol. 98, pp. 254–264, 2018.

[3] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data: the management revolution," *Harvard Business Review*, vol. 90, no. 10, pp. 60–68, 2012.

[4] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: insights to industries," *Computers & Industrial Engineering*, vol. 115, pp. 319–330, 2018.

[5] F. Kache and S. Seuring, "Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management," *International Journal of Operations & Production Management*, vol. 37, no. 1, pp. 10–36, 2017.

[6] S. A. Voccia, A. M. Campbell, and B. W. Thomas, "The same-day delivery problem for online purchases," *Transportation Science*, vol. 53, no. 1, pp. 167–184, 2019.

[7] M. Hu and S. T. Monahan, "US e-commerce trends and the impact on logistics." [Online]. Available: https://www.atkearney.com/web/japan/retail/article/?/a/us-e-commerce-trends-and-the-impact-on-logistics [Accessed: Apr. 2, 2018], 2016.

[8] J. R. Spiegel, M. T. McKenna, G. S. Lakshman, and P. G. Nordstrom, "Method and system for anticipatory package shipping." [Online]. Available: https://patents.google.com/patent/US8615473B2/en [Accessed: Apr. 9, 2018], 2013.

[9] Q. Zhao, Y. Thand, D. Friedman, and F. Tan, "E-commerce recommendation with personalized promotion," in *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 219–226, ACM, 2015.

[10] W. W. Moe and P. S. Fader, "Dynamic conversion behavior at e-commerce sites," *Management Science*, vol. 50, no. 3, pp. 326–335, 2004.

[11] T. Huang and J. A. van Mieghem, "Clickstream data and inventory management: model and empirical analysis," *Production and Operations Management*, vol. 23, no. 3, pp. 333–347, 2014.

[12] C. K. H. Lee, "A GA-based optimisation model for big data analytics supporting anticipatory shipping in retail 4.0," *International Journal of Production Research*, vol. 55, no. 2, pp. 593–605, 2017.

[13] N. Q. Viet, B. Behdani, and J. Bloemhof, "Data-driven process redesign: anticipatory shipping in agro-food supply chains," *International Journal of Production Research*, vol. 9, no. 1, pp. 1–17, 2019.

[14] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2010.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York, NY: Springer New York, 2013.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[17] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009.

[18] S. S. Khan and M. G. Madden, "One-class classification: taxonomy of study and review of techniques," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 2014.

[19] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[20] I. F. Chen and C. J. Lu, "Sales forecasting by combining clustering and machine-learning techniques for computer retailing," *Neural Computing and Applications*, vol. 28, no. 9, pp. 2633–2647, 2017.

[21] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[22] H. Bergmann, "Versand- und Retourenmanagement im E-Commerce 2018." [Online]. Available: https://www.ehi-shop.de/image/data/PDF_Leseproben/EHI-Studie_Versand-Retourenmanagement_im_E-Commerce_2018_LP.pdf [Accessed: Jan. 15, 2018], 2018.