Towards Minimizing e-Commerce Returns for Clothing

A. K. Seewald[1], T. Wernbacher[2], A. Pfeiffer[2], N. Denk[2], M. Platzer[3], M. Berger[3] and T. Winter[4]
[1]*Seewald Solutions, Lärchenstraße 1, A-4616 Weißkirchen a.d. Traun, Austria*
[2]*Donau-Universität Krems, Dr.-Karl-Dorrek-Straße 30, A-3500 Krems, Austria*
[3]*yVerkehrsplanung, Brockmanngasse 55, A-8010 Graz, Austria*
[4]*Attribu-i, Nibelungengasse 32d, A-8010 Graz, Austria*

Keywords: Machine Learning, Visualization, Data Mining, Rule Learning, e-Commerce, Returned Goods.

Abstract: The importance of e-commerce including the associated freight traffic with all its negative consequences (e.g. congestion, noise, emissions) is constantly increasing. Already in 2015, an European market volume of 444 billion Euros at an annual growth of 13.3% was achieved, of which clothing and footwear account for 12.7% as the largest category (Willemsen et al., 2016). However, online commerce will only have a better footprint than buying in the local retail shop under optimal conditions (for example: group orders, always present at home delivery, no returns and no same day delivery). Next to frequent single deliveries, CO^2 intensive and underutilized transport systems, returned goods are the main problem of online shopping. The last is currently estimated at up to 50% (Hofacker and Langenberg, 2015; Kristensen et al., 2013). Our research project Think!First tackles these problems in freight mobility by using an unique combination of gamification elements, persuasive design principles and machine learning. Customers are animated, targeted and nudged to choose effective and sustainable means of transport when shopping online while ensuring best fit by compensating both manufacturer and customer biases in body size estimation. Here we show preliminary results and also present a slightly modified rule learning algorithm that always characterizes a given class (here: returns).

1 INTRODUCTION

e-Commerce is nationally and internationally on the rise (Knabl et al., 2015). Especially retail e-commerce in clothes and shoes over the internet – just for Europe – was already at 56 billion EUR in 2015 and is rising at a yearly rate of 13.3%, forecasted to double until 2020 (Willemsen et al., 2016).

In Austria the market volume in 2014 was 2.1 billion EUR corresponding to a growth of 11.6% from 2013 (Knabl et al., 2015). The ten biggest retailers in Austria received 46% of their sales volume from online shops (Hofacker and Langenberg, 2015) and an increasing number of local shops aim to follow their lead. This increase in online shops is also reflected in the shopping behaviour of Austrian customers. In 2013 already 57% of Austrian customers bought items over the internet with a total sales volume of 5.9 billion EUR spent in Austria and abroad (Lengauer et al., 2015). Because of the similarities with the global situation and the early-adopter status w.r.t. internet shopping of Austrian customers we believe that our study and its conclusions – although mainly based on data from a large Austrian store and online shopping chain (one of our project partners) – are also valid in a global setting.

One important issue in retail e-commerce is the inherently high returns rate of up to 50% (Hofacker and Langenberg, 2015; Kristensen et al., 2013). Combined with a higher number of offers for shorter delivery times this results in a corresponding increase of freight traffic and therefore CO^2 emissions at a time when a reduction of these emissions is needed. On the other hand an increasing number of online shoppers want to buy in a sustainable way (Hagemann, 2015; Halbach et al., 2015). Actual shopping decisions are however less sustainable due to ignorance, laziness and missing incentives. Our research project Think!First aims to optimally inform and motivate online customers through three methods.

1. Creating transparency by visualizing the effects of customer decisions on climate and environment and nudging customers into a sustainable direc-

Seewald, A., Wernbacher, T., Pfeiffer, A., Denk, N., Platzer, M., Berger, M. and Winter, T.
Towards Minimizing e-Commerce Returns for Clothing.
DOI: 10.5220/0007568508010808
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 801-808
ISBN: 978-989-758-350-6

tion by drawing on research in persuasive design and gamification.

2. Optimize logistics towards more sustainable transport vehicles (eg. load bicycles).
3. Reduce returns by compensating customer bias (e.g. misjudgment of correct size) and manufacturer bias (inconsistent or even incorrect reported product sizes) on the size matching process. Non-fitting garments are a known factor to strongly drive returns (Kristensen et al., 2013; Singh, 2015).

Within this paper we will only address the last point. More details on the remaining points can be found in Wernbacher et al. (2017) or at the project website https://www.thinkfirst.blog[1].

2 RELATED RESEARCH

Kristensen et al. (2013) present TrueFit, a system to determine precise body measurements which can reduce returns by up to 30%. However it requires much effort by potential customers. TrueFit works by combining extensive information provided by customers on their height, age, weight as well as a set of previously bought fitting clothes with manufacturer, model type and given size to determine best fit. While it therefore tries to compensate both customer and manufacturer bias, in its present form it ignores body size temporal drift.[2]

Colsen (2013) describes StitchFit, a clothes seller which uses a combination of machine learning and styling expertise provided by human fashion stylists. Here, customers regularily receive fitting garments by post. The machine learning system provides a list of fitting garments based on initial customer-provided body size data and the complete history of returned garments. The fashion stylist chooses garments from this list and sends complete ensembles to the customer. However, he completely ignores the potential for unstructured data.[3]

Singh (2015) analyses reasons for returns within Indian online market Flipkart, where mainly womens' garments are sold directly by the manufacturers. Apart from a detailed analysis of returns reasons they also provide a minimal set of measurements for size tables to reduce returns.[4] Simply changing the shown size tables for nine manufacturers according to his recommendations reduced returns dramatically: An average reduction of absolute returns rate of 9% was reported with a maximum of 46% – so the luckiest manufacturer saw their returns rate halved. He also provides an analysis of returns reasons due to product quality issues which are also a major cause for returns within this online market, albeit less relevant within our project.

Ghaffari (2011) analyses the quality of customer decisions depending on the presentation of garments via product images, and finds that customer skill level in online shopping determines the optimal type of product images: Seasoned clothes shoppers obtain better size estimates from product images of garments worn by realistic models, while novice shoppers obtain better estimates from product images showing just the unworn clothes on a flat surface. Therefore it is preferrable to provide both types of images – and restrict shown image types to the optimal type if the skill level of the customer is known. Our partner provides both types of product images as well as a novel combination of both: images of products worn by mannequins where the mannequins are afterwards digitally removed from each image, and additional product information (such as size and washing tags) is digitally added. It would be interesting to test the effect of this new image type on novice and expert shoppers.

Seewald (2007) creates and analyzes a model to predict the response of inactive customers to a postal mailing. He tests the value of feature subset selection in improving the model and finds little to no effect for the tested algorithms (Naive Bayes, Hidden Naive Bayes and robust logistic regression). In this paper we even found that arbitrarily removing features with high predictive value – a type of feature subset selection reminiscent of adversial examples – may in some cases be beneficial (see section 5).

Toktay (2003) analyses different models to predict returns via synthetic data. He differentiates between modelling via periodical data where only the number of sold and returned products is known (i.e. where it is not possible to identify products and determine exactly which products were returned), and modelling via individual data on product level (i.e. where such a identification is feasible). For the second case – which corresponds to our data – he proposes an Expectation

[1]Presently all information on the website is only available in German. We apologize for the inconvenience.

[2]I.e. changes in body size over time

[3]We are intrigued by the possibility of replicating styling expertise via deep learning on product images or garment style graphs... but this is a topic for another paper.

[4]Minimal set of measurements: breast width, waist circumference, shoulder circumference, sleeve diameter at $\frac{3}{4}$ height; provide at least UK, US and EU-Sizes and at least S,M,L,XL,XXL for simple sizes.

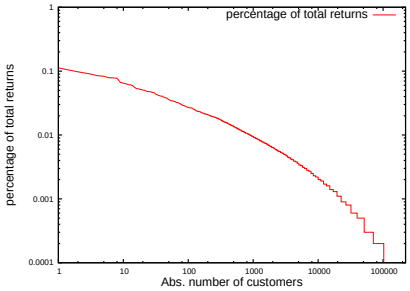

Figure 1: Log-log plot of customers vs. percentage of total returns (top left 1 = 1%).

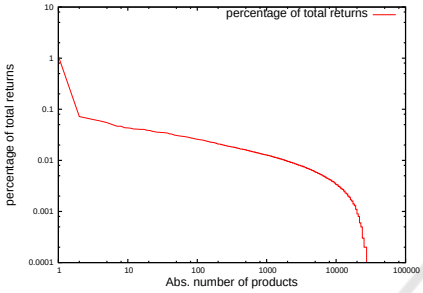

Figure 2: Log-log plot of products vs. percentage of total returns (top left 1 = 1%).

Maximization model. No explicit modelling of the reasons for returns takes place.

3 INITIAL DATA SURVEY

From the data warehouse of our project partner we received a set of a few million samples from a time period of several years, containing detailed information on customers, orders, products, deliveries and returns during this time period. In total there were 312 features – 202 numeric features and 110 nominal features with up to 10 values (avg. 2.75) per feature. Overall returns over all products (including clothing) within this time period were within the ranges reported in the literature.

One trivial way to reduce returns would be to delist products or ban customers which are responsible for a large proportion of returns. So we first checked for power-law distributions as these would indicate that such an approach is feasible. As can be seen in Figures 1 and 2 both products and customers show an approximate line in the log-log plot – with some outliers – and may in first approximation be considered a power law distribution. However note that at the highest point one customer is at most reponsible for 0.1% of total returns and one article is at most reponsible for 1% of total returns so the slope of the power-law distribution is in both cases quite small.

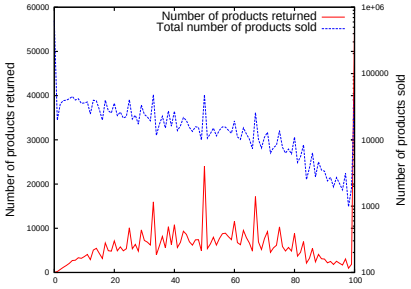

Figure 3: Products sold vs. products returned vs. returns rate (X axis), grouped by customers. More details see text.

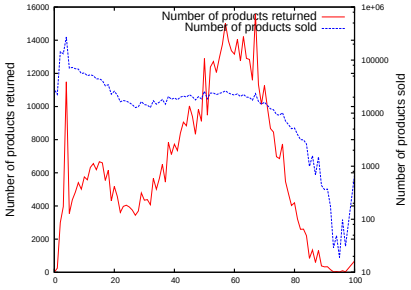

Figure 4: Products sold vs. products returned vs. returns rate (X axis), grouped by products. More details see text.

Due to aliasing effects the number of returns, number of products sold and the returns rate are linked especially for small number of products sold. To allow for a fair comparison, we decided to visualize all three parameters in a single graph. We binned the returns rate in percent into 100 bins at the X axis. For each bin we computed both the number of products sold and the number of returns. The X axis therefore shows the returns rate in percent between 0% (no returns) and 100% (only returns). Absolute number of products returned and absolute number of products sold are shown on the left and the right Y axis. For the number of products sold we had to use a logarithmic scale to enable a comparison.

Figures 3 and 4 show the results. The difference between both figures is that for Figure 3 we computed returns rate on customer level (i.e. each customer was assigned their average returns rate over all their orders) while in Figure 4 we computed returns rate on products level (i.e. each product was assigned its average returns rate over all its sales).

Figure 3 shows a simple pattern. Customers with very low (left in graph) and very high (right in graph) return rates cause comparably small returns – but those with very low return rates tend to buy much more products than those with very high return rates so that the absolute number of returns is almost the same. However the highest number of returns is found in the center at around 40% to 60% returns rate.

Figure 4 shows a much more interesting pattern. The bimodal distribution of returns indicates that returns are driven by two approximate normal distributions (excluding the outliers around 2%): one smaller peak of those products which sell very well at a low relative returns rate – which however translates to a high number of absolute returns – and a larger peak of products which sell less well at a higher returns rate. The latter group seems to contribute about six times the returns as the former group.

We can now ask what would be necessary to reduce the returns rate by e.g. 10%, assuming a well-trained model that can identify customers and products with high returns a priori with high confidence.[5] To achieve this reduction, we would need to ban 0.17% of customers at the cost of selling 4.31% less products overall; or delist 1.13% of products at the cost of selling 13.46% less products overall. Both alternatives were deemed unsatisfactory to our project partner and therefore rejected.

4 APPROACH

For our approach we initially proposed a system that does not rely on time-consuming and cumbersome customer self-reporting[6] while still ensuring a precise quality customer body size and manufacturer size data. In each case we tested the data provided by our partner whether it would be feasible to implement each point.

1. Reconstruct a precise body size from previous orders that were not returned, using as negative set those orders which were returned. This approach is complicated by customers buying for multiple persons (e.g. parents for children, or wife for husbands / vice versa) and of course necessitates precise manufacturer size information.

   Here we evaluated the available data provided by our partner and found this approach to be feasible. We may follow it up in the future.

2. A combination of age (reconstructed from birth date), sociodemographic and other customer data could potentially be used to create a rough estimate of height, weight and other body parameters. For ground truth data customers could be measured when buying products in physical stores of which our partner has several.

   A sufficiently large proportion of the customers in our dataset have a birth date stored, so for these age can be reconstructed. However our partner considered it too costly to measure customers in the stores so we were not able to build a customer-based body size model and had to reject this approach.

3. Unstructured data such as textual comments and feedback including five-star-ratings and continuous rating values may be useful to determine manufacturer-dependent and possibly user-dependent biases in size estimation.

   Our partner provides star-ratings and textual comments for logged-in users which are linked to the product data so it initially looked feasible to build such a model for manufacturer-dependent biases. However, the amount of data was insufficient to build a model of user-dependent biases so for this purpose we had to reject the approach.

   Concerning manufacturer-dependent bias, we found that provides detailed size information almost on the level of garment cuts to their manufacturers who are then producing the garments according to specifications. An appropriate amount of samples chosen by international standards in garment production are then measured to ensure compliance with the initial size information. In this context there is only a single manufacturer and therefore no meaningful manufacturer-dependent biases to analyze. However it also means that high-quality consistent size information was already available which we could use for evaluation (see subsection 4.1).

   We may revisit the use of unstructured data in future work, however for this project it was rejected.

   Due to manufacturing tolerances the actual size may differ about ± 1 unit from the size reported on each garment. We have therefore proposed to produce garments without any size information at all and measure each piece delivered by the manufacturer, adding the correct size information and thereby ensuring perfect size information on each item.[7] However this approach was also deemed to be too costly by our partner and also had to be rejected.

[5] This *may* be an unreasonable assumption.

[6] E.g. the approach described by Kristensen et al. (2013)

[7] This approach is similar to the one used in microprocessor production where maximum operating frequency is automatically measured and each chip tagged accordingly. Changes in production quality influence maximum operating frequency in a complex way – especially at the beginning of a new microprocessor generation – and it is deemed easier to simply measure the produced chips than to model the process.

4.1 Quantitative Size Information

As we mentioned during the evaluation of the 3rd approach, we found out that relatively precise size information is available for most products. Since it is known that providing better size information leads to smaller returns rates (Kristensen et al., 2013; Singh, 2015) and the webshop by our partner only provided a single size table for all products, we followed up on this by extracting all size tables from the backend systems and converted them into a format suitable to be displayed in the webshop. Among the more than 200 different measurements we restricted ourselves to those found by Singh (2015) to perform best. Evaluation is still ongoing.

4.2 Qualitative Size Information

Another observation was that – additionally to the quantitative size information just mentioned – humans communicate size preferences using qualitative concepts related to fit such as figurative, figure-accentuating, casual and straight. Our partner graciously provided such style information for several hundred products which we used to train a model to consistently determine qualitative fit information directly from the size tables.

As features we used arithmetic average, standard deviation, relative standard deviation, median, max and min of breast width, inner leg length, waist circumference and hip circumference over all sizes. For simplicity we used a robust Logistic Regression model.

The best model had an accuracy of 59.47% – much improved over the baseline accuracy of 43.13% – and we are currently working to improve this by adding unstructured textual data from product descriptions as well as unstructured image data from product images. Evaluation is ongoing.

5 CHARACTERIZATION OF RETURNS

As final step we aimed to characterize returns by a well-known rule learning algorithm, JRip, which is an open source implementation of RIPPER (Cohen, 1995) within the data mining suite WEKA.[8] We chose RIPPER for its ability to produce small concise rule sets that are easy to interpret. We also considered Logistic Regression but found that too many features had high weight resp. odds ratio, severly impairing interpretability of the trained model.[9] As we had more than enough samples, we downsampled a 20% subset of the original dataset to 1:1 class distribution between *returns* and *non-returns* and evaluated the model on the remaining 80% data that was not subsampled plus the *non-returns* removed from the training data during subsampling. For initial evaluation we used the training set with twofold crossvalidation.

Table 1: Successive removal of predictive features as estimated by two-fold crossvalidation.

Dataset	#F.	Prec.	Rec.	F_0	AUC
tr3	154	0.813	0.821	0.817	0.840
tr4	150	0.866	0.596	0.706	0.790
tr5	149	0.798	0.772	0.785	0.820
tr6	145	0.822	0.862	0.842	0.857
tr7[11]	303	0.865	0.898	0.881	0.917
tr8	290	0.784	0.827	0.805	0.830
tr9	283	0.765	0.793	0.779	0.792
tr10	282	0.765	0.794	0.779	0.794
tr11	237	0.768	0.789	0.778	0.789
tr11_17[12]	237	0.767	0.775	0.771	0.769

One major problem was that the data dictionary had not been updated for some time. Therefore from 312 features only about a third were actively used and well-known. To prevent inadvertently using features that are changed when returns are entered into the system – which may give the system an unfair advantage and yield results that are too good to be true – we chose two mitigations: 1) training the model on one set of data and evaluating on a later data warehouse export (ongoing), 2) successively removing highly predictive attributes (either by high odds value in logistic regression or by appearing often in the first 5-10 rules from RIPPER).[10] Table 1 shows the results for the second mitigation. It can be seen that the removal of highly predictive attributes does not always reduce the performance of RIPPER but in some cases even proves beneficial. The addition of products base data from `tr7` onward clearly proved beneficial at first and it will turn out that many final rules make use of those features.

[8] See https://www.cs.waikato.ac.nz/ml/weka/

[9] However we stil used it to determine candidate attribute for removal (see later).

[10] The final choice of which attributes to remove was mainly based on these two criteria since for most candidate attributes little or no feedback on their meaning could be obtained from our partner. The cut-off was manually chosen for each dataset by visually inspecting the distribution of odds values and attributes used in top rules by RIPPER.

[11] Adding products base data (VKA.∗); modifying RIPPER to force rules on *returns* (see text)

[12] tr11 with only samples from 2017 by order date.

Initially we ran RIPPER on the complete data from all years containing a few million samples (`tr3-tr11`). However, the data warehouse format had been changed at least three times during the last several years in which the data was collected causing some variables' interpretation to be changed and some new variables to be added, both causing RIPPER to return large sets of around a hundred rules to account for the additional data variance. So we chose to retrain it using only the latest data from 2017 and 2018 (`tr11_17`) containing a few hundred thousand samples. This led to a small set of eleven rules which predict returns on the independent test set with a precision of 0.495 and recall of 0.784 (balanced F-measure: 0.607), comparable to models trained on the whole dataset.[13]

One problem with RIPPER was that according to its internal heuristics its default class[14] was sometimes *returns* and sometimes *non-returns*. However since rules for non-returns are much harder to interpret than rules for returns and we are also much more interested in the latter, we chose to adapt RIPPER to force it to use only *non-returns* as default class and therefore always output rules that predict *returns*. The modified RIPPER was used from `tr7` onwards. `tr1` and `tr2` had minor errors in preprocessing and were therefore removed from the table.

We will now interpret the rules from the rule list in order. Note that `class=1` corresponds to samples with class *returns* and that it is necessary to apply these rules in exactly the given order to get correct results.

```
(VKA.ARSeit_month >= 4) and (VKA.VK_PreisA >=
    49.9) and (VKA.WarennummerCode <=
    62069090) => class=1 (7573.0/1469.0)
```

This rule only utilizes products base data (VKA.∗). All products which have been in the online shop since April (ARSeit_month – primarily excluding some products only available in winter) and the sales price (VK_PreisA) is at least 49.9 EUR and the product group code (WarennummerCode) is smaller than 62069090 (exluding some products for which return patterns are presumably different) are predicted to be returns. The antecedents of this rule cover 7,573 samples in training data of which only 1,469 (19.39%) are **not** of class *returns*.

```
(VKA.ARSeit_year >= 2014) and (VKA.VK_PreisA
    >= 33.9) and (VKA.Laenge <= 0) and (VKA.
    VK_PreisA >= 49.9) and (VKA.
    EingebbarAb_weekday = Di) => class=1
    (730.0/128.0)
```

Again, this rule only utilizes products base data (VKA.∗). All products which have been listed since year 2014 (ARSeit_year) and have a sales price (VK_PreisA) of at least 49.9 EUR[15] and have length (VKA.Laenge) of zero or less (excluding mainly trousers since only these have positive values for length) and have been set active in the webshop (EingebbarAb_weekday) on a Tuesday[16] are predicted to be returns. The antecedents of this rule cover 730 samples of which 128 (17.53%) are **not** returns.

```
(VKA.EingebbarAb_year >= 2016) and (VKA.
    ARSeit_month >= 5) and (VKA.VK_PreisA >=
    34.9) and (VKA.Laenge <= 0) => class=1
    (2569.0/795.0)
```

Again this rule only utilizes products base data (VKA.∗). All products that have been listed since 2016 and in each year listed from month May onwards (ARSeit_month, including mostly those products sold in summer) and which have a sales price of at least 34.9 EUR and a length of zero or less (again excluding trousers) are predicted to be returns. The antecedents of this rule cover 2,569 samples of which 795 (30.94%) are **not** returns.

```
(VKA.EingebbarAb_year >= 2016) and (VKA.
    EingebbarAb_year >= 2017) and (VKA.
    WarennummerCode <= 63012090) and (VKA.
    WarennummerCode >= 61091000) => class=1
    (368.0/120.0)
```

This is the last rule that only uses products base data (VKA.∗). All products that have been listed from 2017 and have a product group code[17] between 63012090 and 61091000 inclusive (this set consists almost exclusively of blankets) are predicted to be returns. The antecedents of this rule cover 368 samples of which 120 (32.60%) are **not** returns.

```
(VKA.WarennummerCode >= 40169997) and (VKA.
    ARSeit_year >= 2015) and (Auftrag.
    Lieferdatum_weekday = Di) and (Auftrag.
    LieferAdresseFl = 1) => class=1
    (68.0/10.0)
```

This rule utilizes products base data (VKA.∗) as well as order data (Auftrag.∗). All orders with products

[13] I.e. a 2.54% reduction in Area-under-ROC-curve (AUC) versus the model trained on whole data (estimated by two-fold CV. Precision and Recall cannot be directly compared since these may be traded off differently in both models.

[14] All initial rules predict the non-default class, followed by a rule with empty antecedents predicting the default class.

[15] Note that this feature appears twice in this rule but of course just using the higher treshold value is equivalent to the redundant form.

[16] We presumed that different product groups were activated on different week days. However our project partner could not confirm this.

[17] This code is used for Intrastat declarations.

having a product group code of at least 40169997[18] and have been listed since 2015 and have been delivered (Auftrag.Lieferdatum_weekday) on a Tuesday via express delivery (Auftrag.LieferadresseFl=1) are returned. The antecedents of this rule cover 68 samples of which 10 (14.7%) are **not** returns. We hypothesize that these orders were made over the weekend and were expected to arrive on Monday but arrived too late and were therefore returned. However the small number of cases did not allow us to validate this.

```
(VKA.WarennummerCode >= 40169997) and (Auftrag
    .ErsterVersPlan_weekday = Do) and (Auftrag
    .LieferAdresseFl = 1) => class=1
    (107.0/13.0)
```

All orders containing products with a product group code of at least 40169997 (see previous rule) and which were last planned to be sent out on a Thursday (Auftrag.ErsterVersPlan_weekday) are predicted to be returns. The antecedents of this rule cover 107 samples of which 13 (12.14%) are **not** returns. We hypothesize this to be a variant of the previous rule.

```
(VKA.WarennummerCode >= 40169997) and (VKA.
    EingebbarAb_year >= 2015) and (VKA.
    ARSeit_year <= 2016) and (Auftrag.
    Auftragsdatum_month <= 2) and (
    Warenausgang.geliefert_weekday = Fr) =>
    class=1 (74.0/20.0)
```

All orders containing products with a product group code of at least 40169997 (see previous two rules) and which have been set active from 2015 onwards and have been available in the webshop until 2016 (i.e. they have only been available for 1-2 years) and which were ordered in January or February (Auftrag.Auftragsdatum_month $<= 2$) and were delivered on a Friday (Warenausgang.geliefert_weekday = Fr) are predicted to be returns. The antecedents of this rule cover 74 samples of which 20 (27.02%) are **not** returns.

```
(VKA.WarennummerCode >= 40169997) and (Auftrag
    .Lieferart = p) and (Auftrag.
    ErsterVersPlan_weekday = Mo) and (Auftrag.
    Kontaktform >= 12) => class=1 (144.0/33.0)
```

All orders containing products with a product group code of at least 40169997 (see previous three rules) and which have been delivered by post (Auftrag.Lieferart=p) and have last been intended to be sent out on Monday and have been ordered in actual shops (Auftrag.Kontaktform$>=$12, i.e. not ordered via Webshop) are predicted to be returns. The antecedents of this rule cover 144 samples of which 33 (22.91%) are **not** returns. Since this rule describes returns outside of webshop orders it is not directly relevant to our project.

```
(VKA.WarennummerCode >= 42023290) and (Auftrag
    .ErsterVersPlan_weekday = Do) and (Auftrag
    .Auftragsdatum_weekday = Di) and (Auftrag.
    AuftragNichtTeilen = 0) => class=1
    (94.0/14.0)
```

All orders containing products with a product group code of at least 42023290 (this seems to basically exclude similar products as the previous three rules with this slightly different threshold) and which have been planned to be sent out on a Thursday and which were ordered on a Tuesday and which did not have the order flag *do not split* (Auftrag.AuftragNichtTeilen=0) are predicted to be returns. The antecedents of this rule cover 94 samples of which 14 (14.89%) are **not** returns.

```
(VKA.WarennummerCode >= 40169997) and (Auftrag
    .Lieferart = p) and (Auftrag.
    ErsterVersPlan_weekday = Mo) and (Kunden.
    DatumErstanlage_year >= 2011) and (Auftrag
    .Landkuerzel = a) and (Auftrag.
    AufnahmeZeit_sec <= 23) => class=1
    (117.0/33.0)
```

This is the first rule that also includes anonymized customers base data (Kunden.$*$). All orders containing products with a product group code of at least 40169997 (see some previous rules) and which have been delivered by post (Auftrag.Lieferart=p) and which have been last planned to be sent out on a Monday, ordered by customers which were initially created in 2011 or later, sent to an address in Austria (Auftrag.Landkuerzel=a) and which were confirmed in at most 23 seconds (Auftrag.AufnahmeZeit_sec $<= 23$) are predicted to be returns. The antecedents of this rule cover 117 samples of which 33 (28.20%) are **not** returns.

```
=> class=0 (10255.0/2538.0)
```

All samples not covered by any previous rule are predicted to be non-returns. This empty antecedent covers 10,255 samples of which 2,538 (24.74%) **are** returns.

We note that the most features of these rules were drawn from products base data (VKA.$*$) followed by orders (Auftrag.$*$) and only the last (non-default-)rule contains customers base data (Kunden.$*$). This may indicate the relative importance of these feature subsets. We showed and explained these rules to our project partner and they were quite surprised and intrigued by these results.

[18]This excludes a small group of free product giveaways, vouchers and made-to-order products and services which are very unlikely or even impossible to return.

6 CONCLUSIONS

We have discussed several approaches to reduce returns in the context of garment e-commerce. A few promising approaches could not be followed up because they proved too costly to implement, such as adding garment size tags only after delivery and subsequent measurement to effectively remove manufacturing tolerances. One simple approach – providing more detailed product-specific measurement tables – is currently in evaluation. We shortly mentioned the usefulness of qualitative size information and presented some preliminary results.

In the main part of our paper, we describe a method to identify and remove highly predictive features from large, mostly undocumented datasets to improve the quality and stability of trained models while also preventing overfitting. We demonstrate the usefulness of this method by describing a rule set of only eleven rules that predicts returns at good precision and recall on a large real-life dataset. To achieve this, it was also necessary to modify the chosen learning algorithm RIPPER in a minor way to ensure it always characterizes returns rather than non-returns. The described rules show some intriguing patterns which are currently investigated by our commercial partner and some may prove to be generally useful.

In the future we hope to follow up on reconstructing precise body size from ordering information – observing that we already obtained reasonably precise manufacturing size information – and finish our preliminary investigations towards a final result.

ACKNOWLEDGEMENTS

This project was funded by the Austrian Research Promotion Agency (FFG) and by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT) as project Think!First (859099)

REFERENCES

Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning, San Francisco, CA, 1995*, pages 115–123. Morgan Kaufmann.

Colsen, E. (2013). Using human and machine processing in recommendation systems. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts*, number 13-01 in CR.

Ghaffari, S. (2011). Will it fit? consumer decision making in online shopping environments. Master's thesis, School of Industrial Design, Georgia Institute for Technology, USA.

Hagemann, H. (2015). *Umweltrelevante Produktinformationen im E-Commerce: Chancen für nachhaltigen Konsum*. Umweltbundesamt, D-06844 Dessau-Roßlau, Germany.

Halbach, J., Stüber, E., and Piepke, M. (2015). *Nachhaltigkeit im Online-Handel - Die Rolle von Ausgestaltung und Kommunikation*. IFH Institut fü Handelsforschung GmbH.

Hofacker, L. and Langenberg, C. (2015). *E-Commerce-Markt Österreich/Schweiz 2015*. EHI Retail Institute. https://www.ehi.org/deStudien/e-commerce-markt-oesterreichschweiz-2015.

Knabl, W., Köb, M., Meszaros, G., Prenger, C., Rischanek, U., Segal, D., and Weigl, A. (2015). *retail - Magazin für den österreichischen Handel. Offizielles Medium des Handelsverbandes*, volume 04. Handelsverband. https://www.handelsverband.at/fileadmin/content/images_publikationen/retail/Retail_2015_04.pdf.

Kristensen, K., Borum, N., Christensen, L., et al. (2013). Towards a next generation universally accessible online shopping-for-apparel system. In *Human-Computer Interaction: Users and Cotexts of Use, Volume 8006 of the series Lecture Notes in Computer Science, pp. 418-427*. Springer.

Lengauer, E., Koll, O., Kreuzer, M., Herry, M., and Sedlacek, N. (2015). *eComTraf - Auswirkungen von E-Commerce auf das Gesamtverkehrssystem*. Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT). Study funded within the search program Mobility for the Future (Mobilität der Zukunft, MdZ) by the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT).

Seewald, A. (2007). Improving the effectiveness of mailings by building a response model for inactive customers. Technical Report 2007-01, Seewald Solutions, Lärchenstraße 1, A-4616 Weißkirchen a.d. Traun, Austria.

Singh, K. (2015). *Reducing Customer Returns in an Online Marketplace*. Dept. of Fashion Technology, National Institute of Fashion Technology, Mumbai, India.

Toktay, L. (2003). Forecasting product returns. In Guide Jr., D. and Van Wassenhove, L., editor, *Business Apects of Closed-Loop Supply Chains, International Management Series*, volume 2. Carnegie Bosch Institute.

Wernbacher, T., Pfeiffer, A., Denk, N., Platzer, M., Berger, M., Seewald, A., Winter, T., and Miller, I. (2017). Minimizing returns through gamification, persuasive design principles & machine learning. In *Poster presentation at the 11th European Conference on Games Based Learning (ECGBL 2017), Oct. 5-6 2017, Graz, Austria*.

Willemsen, R., Abraham, J., and van Welle, R. (2016). *Global B2C E-commerce Report 2016*. Ecommerce Foundation, Raddhuisstraat 22, NL-1060 Amsterdam, The Netherlands.