Carl: A Sports Award Recommender

Martin Pichl, Bernward Pichl
Pichl Medaillen GmbH
Schießstand 10
Inzing, Austria
firstname.lastname@pichl.com

Eva Zangerle
Department of Computer Science, Universität Innsbruck
Austria
eva.zangerle@uibk.ac.at

ABSTRACT

Due to the rise of the web, today's huge open source community and the numerous publications of industry as well as academia in the field of computer science, nowadays even small and mid-sized companies can access state-of-the-art machine learning technologies, that can be leveraged for their businesses. In this paper, we present Carl, a hybrid recommender system utilizing content-based filtering combined with a context-aware sales model trained via XGBoost to recommend sports awards to customers. The computed recommendations are sent via e-mail to regular customers, who have already bought sports awards before. Hence, this systems aims to increase customer satisfaction by simplifying the decision which sports awards to buy every season. In offline experiments we observe, that XGBoost compared to other state-of-the-art approaches as Factorization Machines and Neural Networks provides the best recommendation performance. However, more importantly, in the complementary online evaluation, we monitor that the interaction- and conversion rates of the e-mails sent via Carl are a magnitude higher compared to our corporate newsletter, relying on a non-personalized most popular approach.

ACM Reference format:
Martin Pichl, Bernward Pichl and Eva Zangerle. 2018. Carl: A Sports Award Recommender. In *Proceedings of The 2018 SIGIR Workshop On eCommerce, Ann Arbor, Michigan, USA, July 2018 (eCom'18),* 5 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The Pichl Medaillen GmbH is a family business founded in 1846 and specialized in producing custom medals and mints. Since the 1980s, Pichl also sells sports awards. Today, this segment is responsible for about 20% of the whole annual turnover. Due to the highly standardized products in this segment and the fact that customers demand those products regularly to have different awards for their events, Pichl decided to implement the company's first recommender system in this segment as part of the digital innovation agenda. The system Carl, named after Karl Pichl introducing the industrial manufacturing in the company, aims at helping the company to (i) design the workflow of selling sports awards more efficiently and (ii) increase customer satisfaction by suggesting products, as these suggestions ease the choice overflow by decreasing the search time for sports awards. Because customers demand new products for each season, this segment is moreover characterized by a regularly changing product assortment. Particularly, every year, about one-third of the complete assortment is replaced by new products. This is the reason why collaborative-filtering approaches or model-based approaches leveraging a user-item matrix as SVD [19], which are already available in shop applications and known to work well in other domains, fail in the presented use case: for collaborative filtering-based systems, the estimation of a user- or item-similarity by leveraging user-item interactions is difficult with always changing items as no interactions for new items are available. This cold-start problem is referred to as the new item problem [26]. To avoid this problem, together with the Databases and Information System Group at the University of Innsbruck, the Pichl Medaillen GmbH decided to develop a hybrid approach recommendation facilitating content- and contextual information.

The developed approach is a hybrid recommender system facilitating both, content-based filtering and predictive modeling. The first component based on content-based filtering covers the personalization aspect, whereas the second component is a global (not personalized) classification model that is capable to predict whether a certain product (with certain features) is likely to be sold in a certain period (i.e., winter or summer seasons) or not. We refer to this as *saleability*.

Using an offline evaluation, we show that eXtreme Gradient Boosting [10] provides the best performance for the prediction task, compared to Factorization Machines [21] and Multilayer Perceptron Neural Networks [20]. Moreover, we are able to show that such a hybrid system overcomes the limitation of collaborative filtering for domains where the product assortment is changing regularly. Along with that, we show that the recommendations computed by the proposed system are not only highly precise if evaluated offline but also deliver a high conversion rate in the real-life application. Hence, our proposed hybrid recommendation model is capable of recommending new products to customers and thus is applicable for domains with regularly changing product assortments.

The remainder of this paper is structured as follows: We introduce the used machine learning methodologies in Section 3 and give the reader an overview about the whole recommender system in Section 4, where we also present the underlying recommendation model in more detail. Next, we introduce the dataset and the conducted offline evaluations in Section 5. We present the real-life application in Section 6 and finally conclude this work Section 7.

2 BACKGROUND

Since the 2000s, recommender system research focused on collaborative filtering approaches and in particular, on matrix-factorization techniques such as singular value decomposition (SVD) as these approaches have been shown to achieve the best recommendation

In: J. Degenhardt, G. Di Fabbrizio, S. Kallumadi, M. Kumar, Y.-C. Lin, A. Trotman, H. Zhao (eds.): Proceedings of the SIGIR 2018 eCom workshop, 12 July, 2018, Ann Arbor, Michigan, USA, published at http://ceur-ws.org

accuracies [7, 17, 25] and to be useful for implicit feedback [14, 22]. However, as outlined in the introduction, collaborative filtering-based approaches fail in our setting due to the new item problem. To handle the new item problem, content-based approaches or hybrids facilitating content-based information to find similar items to the new item are suitable [26]. In this work, we present a hybrid approach leveraging content-based information. Generally, content-based recommender systems focus on item characteristics to find similar items. In particular, these systems recommend items that are similar to the items a user already interacted with in the past. This is why these are also called content-based filtering approaches: they filter items based on previous user-item interactions. These approaches have their roots in the field of information retrieval [4, 9, 24] and initially focused on recommending items containing text, for instance, news articles, websites or UseNet messages [1]. In this work, we use content-based filtering to derive an initial set of recommendation candidates. We rank the computed candidates using a context-aware classification approach similar to the approaches introduced next.

In the late 2000s, research shifted towards hybrid approaches additionally integrating contextual information on top of the presented content- and collaborative filtering-based approaches: Several extensions of matrix factorization techniques have been introduced, e.g., time-aware SVD++ [18]. As context is a broad concept, subsuming any circumstances that influencing the perceived usefulness of an item [2], a variety of additional contextual information has been exploited in the field recommender systems, for instance, the current time [6, 18], the current emotion and mood of a user [5, 8, 13, 23] or the user's location [3, 11, 15, 16]. Due to the success of context-aware approaches, we follow up this research and incorporate the current month as a proxy for the current season as contextual information into our recommender system, allowing us to estimate a product bias for certain seasons. To incorporate context along with content-based features, we rely on a classification approach (cf. Section 3).

3 METHODOLOGY

To select the best methodology for classifying whether a product will be successfully sold or not, we evaluate three state-of-the-art classification approaches in an offline evaluation. We evaluate eXtreme Gradient Boosting (XGBoost) [10], Factorization Machines (FM) [21] and Multilayer Perceptron Neural Networks (MLP) [20]. Using these three approaches, we cover a wide range of methodologies. In particular, we cover trees, factorization approaches leveraging latent features and neural networks. To contextualize the performance of the different classification approaches, we conduct an offline evaluation (Section 5). In this evaluation, we require the classifiers to predict whether a certain product will be successful or not. For this two-class classification task, we consider products as successful if they exceed a certain turnover threshold.

4 SYSTEM OVERVIEW

As already outlined in the introduction, our proposed recommender system is based on two major components (C): C1 is responsible for finding the top-n new and similar products to the products that are found in the customers' purchase history. C2 is responsible for predicting whether a certain product is likely to be sold. Hence, it additionally ranks the top-n new products by the probability whether these will be sold. Finally, the top-1 recommendation is sent by a personalized e-mail to the customer periodically. In particular, we send the e-mails two weeks prior to a potential customer order. A potential customer order is computed using last order plus one period. For example, if a customer buys once a year sports awards, he or she will get the e-mail 351 days after their last order. An overview of the system is given in Figure 1. We describe both components along with their interaction next.

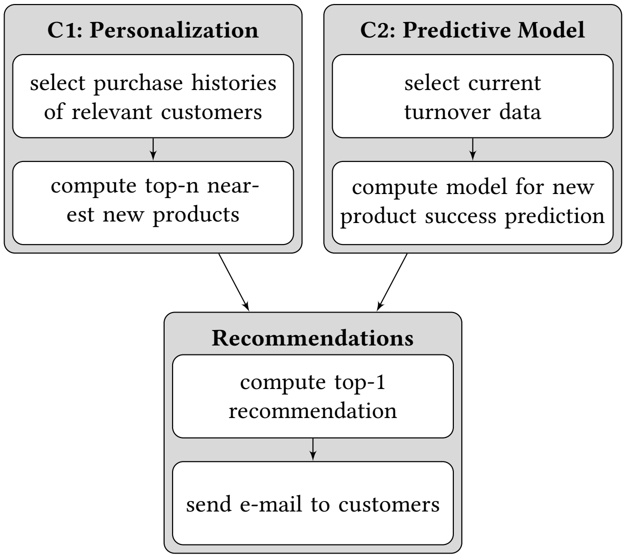
C1: Personalization
select purchase histories of relevant customers
compute top-n nearest new products
C2: Predictive Model
select current turnover data
compute model for new product success prediction
Recommendations
compute top-1 recommendation
send e-mail to customers

Figure 1: Workflow for Computing Recommendations

C1 finds the top-n nearest new products based on the products contained in a user's purchase history. For the computation of product similarity, we utilize a generalization of the Gower coefficient [12]. We use this distance, as in contrast to Pearson correlation, a measure that is widely used in the field of recommender systems, the Gower coefficient allows us to incorporate factor variables into the similarity computation. In Equation 1, we show the computation of the Gower similarity G between two products i and j. We denote $g_{i,j,k}$ to the contribution provided by the feature k weighted by $w_{i,j,k}$. The computation of the feature contribution $g_{i,j,k}$ for numeric features as price or height is shown in Equation 2, where we denote r_k to the range of feature k. For the factorial features in our dataset (color, design, ...), $g_{i,j,k}$ is computed as depicted in Equation 3.

$$G_{i,j} = \frac{\sum_k w_{i,j,k} g_{i,j,k}}{\sum_k w_{i,j,k}} \tag{1}$$

$$g_{i,j,k} = 1 - \frac{|x_{i,k} - x_j, k|}{r_k} \tag{2}$$

$$g_{i,j,k} = \begin{cases} 1 & if\ x_{i,k} = x_{j,k} \\ 0 & if\ x_{i,k} \neq x_{j,k} \end{cases} \tag{3}$$

For the final similarity computation of products that is used in the real-life system, we set all weights $w_{i,j,k} = 1$ and hence let each feature equally contribute to the product similarity. We leave a weighting scheme for future work. Using the presented computation of the Gower distance, we derive the set of recommendation candidates for each user by computing the top-n similar new items. A new item is an item that is (i) newly added to the assortment in the current year and simultaneously an item that is (ii) not found in a user's buying history. We rank this set of recommendation candidates using C2 as described in the remainder of this section.

C2 is responsible for estimating the saleability of a product (in a certain season) and hence computes the probability whether a certain product will be sold or not. As we observe that XGBoost delivers the best performance for this prediction task (cf. Section 5.3), we use XGBoost to compute the saleability of the recommendations candidates computed by C1. The saleability is computed by applying the pre-trained XGBoost model (trained with the turnover data of the last three years as described in Section 5) to the recommendation candidates. For each candidate, we get a saleability value s scaled between 0 and 1.

Using both, the product similarity based on the Gower coefficient g and the saleability s, we compute the final ranking of the new items for each user using the average of both values as depicted in Equation 4.

$$r_{i,j} = w_1 g_{i,j} + w_2 s_j \tag{4}$$

In Equation 4, we denote $r_{i,j}$ as the ranking coefficient for an item i (contained in a user's purchase history) and a new item j. Furthermore, we denote $g_{i,j}$ as the Gower coefficient between item i and j and s_j as the predicted saleability of an item j. For the real-life application, we set $w_1 = w_2 = 0.5$ and hence, consider the personalization aspect represented by the Gower coefficient and the saleability aspect equally.

5 EXPERIMENTS

As already outlined in the previous section, we evaluate the classification accuracy of the predictive model using k-fold cross-validation. Before describing the experimental setup and discussing the results, we introduce the reader to the used dataset.

5.1 Dataset

For evaluating the different classification methods, we use the turnover data of the previous three years (2015, 2016 and 2017) as training- and test data. Please note that we are not able to use the turnover data of the current year (2018), as we cannot estimate the success of the products yet. The dataset contains 538 main products and 1,939 product variations. A main product is available in different sizes, where each size is considered as a product variation. Hence, a product variant shares the same features besides the price and the height. As stated in Table 1, we characterize each product by 13 features.The features price and height are self-explanatory. Color, accent color 1 and accent color 2 specify the main color along with two accent colors of a product,i.e., silver or gold. Handle is a boolean feature, considering whether a cup has handles. Analogously, cap, emblem, and emblem holder are boolean features defining whether a cup or trophy features an emblem (holder) or a cap. Please note that an emblem can be mounted on an emblem holder or on a cap. Stand indicates the material of a cup's stand, i.e., marble, wood or plastic. Decorative states whether there is a decorative element and the type, for instance, a colored orb.

Feature	Type
height	numeric
price	numeric
color	factor
accent color 1	factor
accent color 2	factor
handle	boolean
decorative	factor
emblem holder	boolean
emblem	boolean
design	factor
stand	factor
cap	boolean
material	factor

Table 1: Product Features Overview

5.2 Experimental Setup

Utilizing the previously introduced dataset, we conduct a k-fold cross-validation to determine the most accurate model for the new product success prediction. For this, we randomly split the dataset into 5 folds of equal size where we use each fold as the test set once and the remaining folds for the training. For this evaluation, we use the packages' default parameters but vary the number of latent features for the FM ($k \in \{1, 5, 10, 25, 50\}$) and the number nodes per layer n_l as well as layers l of the neural network ($n_l \in 1, 2, 5, 10, 20$, $l \in \{1, 2, 3\}$). To measure the classification performance, we rely on the accuracy measure and the Kappa statistic. While the first measure solely considers the number of correctly classified instances, the Kappa statistic compares an observed accuracy with an expected accuracy. The expected accuracy is based on the inter-rater agreement. Due to this, the Kappa takes the possibility of correctly classifying a product to be successful by chance into account and hence is the more meaningful measure in our experiments.

5.3 Experimental Results

The results of the conducted offline evaluations are stated in Table 2. For the FM and the MLP classifiers, we only state the best result of our evaluations with different k and different n as well as l values respectively.

Algorithm	Accuracy	Kappa
XGBoost	0.74	0.37
FM ($k = 25$)	0.69	0.02
MLP ($n_1 = 10$, $n_2 = 5$, $n_3 = 5$)	0.60	0.29

Table 2: Prediction Accuracy

We observe that though the accuracies of XGBoost and the FM only differs by 6.76%, the Kappa value of the FM is only 0.03. Hence, we cannot see a substantial performance difference to the random baseline or rather an approach always predicting a product to be

Source	Website Visits	Avg. Session Duration	Viewed Pages	Conv./E-Mail	Conv./Users
Carl	20.61%	6.57	8.32	4.61%	22.34%
Corporate Newsletter	1.38%	2.37	4.41	0.03%	2.50%

Table 3: Recommender Key Metrics

not successful. This is, as the FM predicts a success only for 0.72% of the products. In contrast, XGBoost classifies 39.62% of the products as a success, a more realistic number.

For the MLP-based classifier, using a grid search, we find that a neural network with three layers containing 10, 5 and 5 nodes performs well with an accuracy of 0.60. However, though a good classification accuracy, according to the Kappa value, XGBoost works substantially better. In particular, the Kappa value is 27.59% higher.

To conclude, we see that according to the Kappa statistic, both XGBoost and MLP classifiers show a fair agreement in contrast to the FM which shows only a slight agreement. In addition, we observe that to predict whether a product will be sold or not, XGBoost works best in terms of prediction accuracy and the Kappa statistic. This is why we use XGBoost's computed probability that a product will be sold for our proposed recommender system. We refer to this probability as the *saleability* of the product. In the next section, we show how the computed saleability is leveraged for sports award predictions.

6 REAL-LIFE APPLICATION

The go-live of Carl was on January, the 4th 2018 and until April, the 31st, more than 2,000 e-mails have been sent. As our business is highly seasonal with a peak in the beginning and the end of a year, we consider the current analysis as late-breaking results and aim to perform a more detailed online evaluation using the sales data of a complete year in a future work. Nevertheless, in the current stage, we observe a very high interaction rate with the personalized recommendations in the e-mails sent via Carl compared to the corporate newsletter. The latter follows a simple most popular approach, which suggests the most popular products of the current season and the globally most popular sale articles to our customers. In Table 3, we state the relative number of users who have visited the website via a link in the sent e-mail, which is the relative portion of the recipients actually visiting the website, the corresponding average session duration in minutes, the average number of viewed pages as well as the relative number of conversions per user and per e-mail. For the latter two measures, we divided the number of conversions by the number of e-mails sent and by the number of users respectively. Please note that for this analysis, we only tracked orders made in the online shop as a conversion, no offline conversions as orders via telephone or an informal mail. We observe, (cf. Table 3) that the conversion rate of the personalized e-mail is a magnitude higher compared to the standard (unpersonalized) newsletter. The 8.94 times higher conversion rate per user is accompanied by a 2.77 times higher session duration with 1.89 as many page views per session. Moreover, we observe that the relative number of website visits is already a magnitude higher. However, we assume that this is not only rooted in the recommendations but also in the personalized the e-mail is sent as well as the precise timing. This will be a subject for further research in the next months.

Summing up, we see an excellent conversion rate in the e-mails sent via Carl. Our results show that for selecting products to be promoted in e-mail newsletters, combining the predictor for the saleability of products with a traditional content-based recommender system allows for a substantial improvement in a diverse set of quality measures. Hence, in a future work, we will run experiments on fine-tuning the recommender system and aim to implement a latent feature approach for the content-based part. Besides that, we aim to conduct a profound online analysis after a whole year to capture all seasonal effects.

7 CONCLUSION

In this paper, we present Carl, a recommender system that aims to improve customer satisfaction by suggesting sports awards to customers of the Pichl Medaillen GmbH, an Austrian SME. In particular, the presented system aims to increase the customer satisfaction by sending the recommendations via e-mail to our regular customers who buy sports awards every season (i.e., yearly) and hence helps to find suitable sports awards out of a set of more than 200 awards that change regularly. The recommendations are computed using a hybrid approach that leverages content-based filtering combined with a context-aware sales model. The latter is trained via XGBoost and estimates a general saleability coefficient for each product based on product features and contextual information as the current season approximated by the current month. In an offline evaluation, we show that XGBoost delivers the best performance compared to Factorization Machines and multilayer perceptron neural networks. In a complementary online study can show that the conversion rate is substantially higher than the conversion rate of the unpersonalized corporate newsletter, promoting the most popular articles.

REFERENCES

[1] Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (Jun 2005)

[2] Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, chap. 7, pp. 217–253. Springer-Verlag, New York, NY, USA, 1st edn. (2010)

[3] Ankolekar, A., Sandholm, T.: Foxtrot: a soundtrack for where you are. In: Proceedings of Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications (IwS 2011). pp. 26–31. ACM (2011)

[4] Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)

[5] Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., Lke, K.H., Schwaiger, R.: Incarmusic: Context-aware music recommendations in a car. In: Huemer, C., Setzer, T. (eds.) E-Commerce and Web Technologies, Lecture Notes in Business Information Processing, vol. 85, pp. 89–100. Springer (2011)

[6] Baltrunas, L., Ludwig, B., Ricci, F.: Matrix factorization techniques for context aware recommendation. In: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011). pp. 301–304 (2011)

[7] Bell, R.M., Koren, Y.: Lessons from the netflix prize challenge. ACM SIGKDD Explorations Newsletter - Special issue on visual analytics 9(2), 75–79 (Dec 2007), http://doi.acm.org/10.1145/1345448.1345465

[8] Braunhofer, M., Kaminskas, M., Ricci, F.: Recommending music for places of interest in a mobile travel guide. In: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011). pp. 253–256. ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2043932.2043977

[9] Celma, O.: Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer Publishing Company, Incorporated, 1st edn. (2010)

[10] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)

[11] Cheng, Z., Shen, J.: Just-for-me: An adaptive personalization system for location-aware social music recommendation. In: Proceedings of the 16th ACM International Conference on Multimedia Retrieval (ICMR 2014) (2014)

[12] Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics 27(4), 857–871 (1971)

[13] Han, B.j., Rho, S., Jun, S., Hwang, E.: Music emotion classification and context-based music recommendation. Multimedia Tools and Applications 47(3), 433–460 (2010)

[14] Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008). pp. 263–272 (2008)

[15] Kaminskas, M., Ricci, F.: Location-adapted music recommendation using tags. In: User Modeling, Adaption and Personalization, pp. 183–194. Springer Berlin Heidelberg (2011)

[16] Kaminskas, M., Ricci, F., Schedl, M.: Location-aware music recommendation using auto-tagging and hybrid matching. In: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys 2013). pp. 17–24 (2013)

[17] Kim, D., Yum, B.J.: Collaborative filtering based on iterative principal component analysis. Expert Systems with Applications 28(4), 823–830 (May 2005)

[18] Koren, Y.: Collaborative filtering with temporal dynamics. In: Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2009). pp. 447–456. ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1557019.1557072

[19] Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer Journal 42(8) (2009)

[20] Popescu, M.C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems 8(7), 579–588 (2009)

[21] Rendle, S.: Factorization machines with libFM. ACM Intelligent Systems and Technology 3(3), 57:1–57:22 (May 2012)

[22] Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009). pp. 452–461. AUAI Press, Arlington, Virginia, United States (2009), http://dl.acm.org/citation.cfm?id=1795114.1795167

[23] Rho, S., Han, B.j., Hwang, E.: Svr-based music mood classification and context-based music recommendation. In: Proceedings of the 17th ACM International Conference on Multimedia (MM 2009). pp. 713–716 (2009)

[24] Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989)

[25] Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.T.: Application of dimensionality reduction in recommender systems: A case study. In: Proceedings of the WebKDD Workshop at the 6th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD 2000) (2000)

[26] Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th International Conference on Research and Development in Information Retrieval (SIGIR 2002). pp. 253–260 (2002)