Customer Behaviour Analysis for Recommendation of Supermarket Ware

Stavros Anastasios Iakovou, Andreas Kanavos(✉), and Athanasios Tsakalidis

Computer Engineering and Informatics Department,
University of Patras, Patras, Greece
{iakovou,kanavos,tsak}@ceid.upatras.gr

Abstract. In this paper, we present a prediction model based on the behaviour of each customer using data mining techniques. The proposed model utilizes a supermarket database and an additional database from Amazon Company, both containing information about customers' purchases. Subsequently, our model analyzes these data in order to classify customers as well as products; whereas being trained and validated with real data. This model is targeted towards classifying customers according to their consuming behaviour and consequently propose new products more likely to be purchased by them. The corresponding prediction model is intended to be utilized as a tool for marketers so as to provide an analytically targeted and specified consumer behavior.

Keywords: Supervised learning · Data analytics · Customer behaviour · Knowledge extraction · Personalization · Recommendation system

1 Introduction

During the last years, more and more companies store their data into large data-centers so as to initially analyze them and to further understand how their consumers behave. Every day, a large amount of information is accessed and processed by companies in order to get a deeper knowledge about their products' sales and consumers' purchases. From small shops to large enterprises, owners try to record information that probably contains useful data regarding consumers.

In addition, the rapid development of technology provides high quality network services. A large percentage of users utilize the Internet for information of each field. For this reason, companies try to take advantage of this situation by creating systems that store information about users who entered their site in order to provide them with personalized promotions. Companies concentrate on their desired information and personal transactions. Businesses provide their customers with cards so they can record every buying detail. This procedure has led to a huge amount of data and search methods for data processing.

Historically, several analysts have been involved in the collection and processing of data. In modern times, the data volume is so huge that it requires the use

L. Iliadis and I. Maglogiannis (Eds.): AIAI 2016, IFIP AICT 475, pp. 471–480, 2016.
DOI: 10.1007/978-3-319-44944-9_41

of specific methods so as to enable analysts to export correct conclusions. Due to the increased volume for automatic data analysis, methods use complex tools; along with the help of modern technologies, data collection can be now considered as a simple process. Analyzing a dataset is a key aspect to understanding how customers think and behave during each specific period of the year. There are many classification and clustering methods which can be successfully used by analysts to aid them broach in consumers' mind. More specifically, supervised machine learning techniques are utilized in the present manuscript in the specific field of supermarket.

Besides the development of web technologies, the abundance of social networks has created a huge number of reviews on products and services, as well as opinions on events and individuals. Concretely, consumers are used to being informed by other users' reviews in order to carry out a purchase of a product, service, etc. One other major benefit is that businesses are really interested in the awareness of the opinions and reviews concerning all of their products or services and thus appropriately modify their promotion along with their further development. As a previous work on opinion clustering emerging in reviews, one can consider the setup presented in [5].

Furthermore, the emotional attachment of ample customers to a brand name is a topic of interest in recent years in the marketing literature; it is defined as the degree of passion that a customer feels for the brand [13]. One of the main reasons for examining emotional brand attachment is that an emotionally attached person is highly probable to be loyal and pay for a product or service [15]. In [6], authors infer details on the love bond between users and a brand name being considered as a dynamic ever evolving relationship. More concretely, users that have demonstrated emotional connection to the brand through their tweets are considered. Thus, the aim is to find those users that are engaged and rank their emotional terms accordingly.

In this paper, we present a work on modeling and predicting customer behavior using information concerning supermarket ware. More specifically, we propose a new method for product recommendation by analyzing the purchases of each customer. With the use of category of the current dataset, we were able to classify the aforementioned data and subsequently create clusters. According to viral marketing [9], clients influence each other by commenting on specific fields of e-shops. Practically, this method appears in real life when people communicate in real time and affect each other on the products they buy. The aim of this model is to analyze every purchase and propose new products for each customer. More to the point, we want to perform the following steps in the below mentioned order: firstly the analysis of the sales rate is utilized, then the distances of each customer from the corresponding supermarket is clustered and finally the prediction of new products that are more likely to be purchased from each customer separately is implemented.

The remainder of the paper is structured as follows: Sect. 2 presents the related work. Section 3 presents our model, while in Sect. 4 we utilize our experiments. Moreover, Sect. 5 presents the evaluation experiments conducted and the

results gathered. Ultimately, Sect. 6 presents conclusions and draws directions for future work.

2 Related Work

In recent years, a large percentage of companies maintain an electronic sales transaction system aiming at creating a convenient and reliable environment for their customers. On the other hand, retailers are able to gather significant information for the corresponding customers. Moreover, since the number of data is significantly increasing, more and more researchers have developed efficient methods as well as rule algorithms for market basket analysis [2]. Researchers have also developed applications for optimal product selection on supermarket data; one of them being the "Profset model". Using cross-selling potential, this model selects the most interesting products from a variety of ware. Additionally, Li et al. [10] analyzed and designed a model of E-supermarket shopping recommender. Lu et al. [12] developed a personalized recommendation system for government to business e-services.

Furthermore, according to [8], researchers have invented a new recommendation system where supermarket customers were able to get new products. This recommendation system was first presented as a part of "SmartPad", which was a system that allowed customers to prepare their shopping list in advance. In this system, matching products and clustering methods are used so as to provide new products to less frequent customers. The analysis of this method showed 1.8 % increase in the income of the supermarket.

In addition, since consumers started to resist to traditional methods of marketing, it was necessary for companies to invent a new advertising method based on alternative strategies. Leskovec et al. [9] presented an analysis of a person-to-person recommendation network which included 4 million people along with 16 million recommendations. This model illustrated how effective the recommendation network for both sender and receiver was. Despite the fact that average recommendation networks are not very effective in increasing purchases, this model had successfully managed it.

Another significant model was presented by Dickson and Sawyer [4], where they created a model for a grocery shop so as to analyze how customers respond to price and other point-of-purchase information. Moreover, according to [1], consumers know exactly the price of the product they purchase; the results of this work showed that customers did not pay attention to the price of the products they buy. More specifically, they did not know if the price was reduced since it was on special offer.

In addition, Yao et al. [16] created a recommendation system targeted towards supermarket products for consumers since supermarkets are too big to find the desirable product; using RFID technology with mobile agents, they constructed a mobile-purchasing system. Furthermore, Kim et al. [7] presented another recommendation system based on the past actions of individuals, where they provided their system to an Internet shopping mall in Korea.

In point of fact, in [3], authors showed a new method on personalized recommendation in order to get further effectiveness and quality since collaborative methods presented limitations such as sparsity. Regarding Amazon.com, they used for each customer many attributes, including item views and subject interests, since they wanted to create an effective recommendation system. This view is echoed throughout [11], where authors analyzed and compared traditional collaborative filtering, cluster models and search-based methods. Also, Weng and Liu [17] analyzed customers' purchases according to product features and as a result managed to recommend products that are more likely to fit with customers' preferences.

Finally, authors in [14] analyzed the product range effect in purchase data. Since market society is affected by two factors (e.g. rationality and diversity in the price system), consumers try to minimize their spending and maximize the number of products they purchase. So, researchers invented an analytic framework based on big customers' transaction data. They observed that customers did not always choose the closest supermarket and then they tried to answer why consumers buy specific products.

3 Model Overview

In our model, we want to predict whether a customer will purchase a product or not based on a supermarket ware dataset using data analytics and machine learning algorithms. We can classify this problem as a classification one, since the opinion class consists of specific options. Furthermore, we have gathered the reviews of Amazon Company and in particular the reviews of each customer, in order to analyze the affection of person-to-person influence in each product's market.

The overall architecture of the proposed system is depicted in Fig. 1 while the proposed modules and sub-modules of our model are modulated in the following steps.

3.1 Customer Metrics Calculation

From the supermarket dataset, we randomly sampled 10000 records regarding customers' purchases, containing information about sales over a vast period of 4 years. More specifically, the implementation of our method goes as follows: initially, we sample the customers as well as the products. Subsequently, the clustering of the products based on the sales rate takes place. Then, we cluster our customers related on the distance of their houses from the supermarket. Furthermore, a recommendation model, with new products separately proposed to each customer based on their consumer behavior, is utilized. Furthermore, we sampled the customers of Amazon Company and then using the rates of the reviews, we came up with the fraction of the satisfied customers.

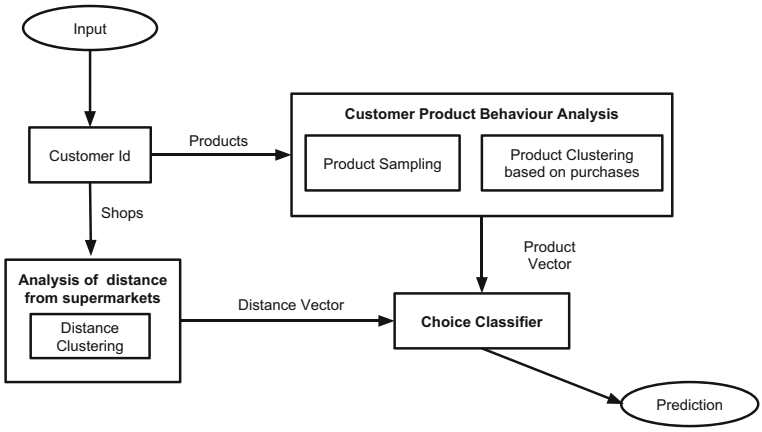

Fig. 1. Supermarket model

The training set of the supermarket data contains 8 features, as presented in following Table 1, including customer ID, the category of the product, the product ID, the shop, the number of items purchased, the distance of each supermarket, the price of the product as well as the choice.

3.2 Decision Analysis

In this subsection, the choice analysis based on classification and clustering tools is described. This method gathers eight basic features of the supermarket database and eleven different methods of classification in order to further analyze our dataset, as shown in the next section. In [8], researchers used clustering to find customers with similar spending history. Furthermore, as [18] indicate, the loyalty of customers to a supermarket is measured in different ways. Specifically, a person is considered as loyal to a specific supermarket if they purchase specific

Table 1. Training set features

Features	Description
Customer ID	The ID of the customer
Product category	The category of the product
Product ID	The ID of the product
Shop	The shop where the customer makes the purchase
Number of items	How many products he purchased
Distance cluster	The cluster of the distance
Product price	The price of the product
Choice	Whether the customer purchases the product or not

products and visit the store several times. Despite the fact that the percentage of loyal customers seems to be less than 30 %, they purchase more than 50 % of the total amount of products.

Since the supermarket dataset included only numbers for each category, we created our own clusters for customers and products. More concretely, we measured the sales of each product as well as the distances and in following we created three clusters for products and two classes for distances.

4 Implementation

The present manuscript utilizes two datasets, e.g. a supermarket database [14] as well as a database from Amazon Company [9] which contains information about the purchases of customers.

Initially, we based our experiments on the supermarket database [14] and we extracted the data using C# language so as to calculate customer metrics. We have implemented an application with which we have measured all the purchases of the customers. In following, a sample of the customers, so as to further analyze the corresponding dataset, was collected. The final dataset consists of 10000 randomly selected purchases with all the information from the supermarket dataset as previously mentioned.

The prediction of any new purchase is based on the assumption that customers are affected by each other. Consumers communicate every day and exchange reviews for products. On the other hand, since their budget is tight, they select products that correspond better to their needs. Therefore, a model that recommends new products to every customer from the supermarket they mostly prefer, is proposed.

By analyzing the prediction model, information about consumers' behavior is extracted. We measured the total amount of products that customers purchased and then categorized them accordingly. Several classifiers are trained using the dataset of vectors. We separated the dataset and we used 10-Fold Cross-Validation to evaluate training set and test set. The classifiers that were chosen, are evaluated using TP (True Positive) rate, FP (False Positive) rate, Precision, Recall, as well as F-Measure metrics. We chose classifiers from five categories of Weka library[1] including "bayes", "functions", "lazy", "rules" and "trees". The classifiers from Weka are used with their default settings and the results are introduced in following in Table 3.

Additionally, we evaluated a model using the results of our experiments on Amazon Company [9] since we wanted to measure the number of contented customers regarding five product categories, namely music, book, dvd, video and toy. In Table 2, we show the number of delighted and on the other hand, the number of not satisfied customers.

Next Fig. 2 illustrates the amount of customers who are satisfied with products of every category. We can observe that the number of satisfied customers

[1] Weka toolkit: http://www.cs.waikato.ac.nz/ml/weka/.

Table 2. Measurement of satisfaction of customers

Product category	Satisfied customers	Not satisfied customers
Music	80149	15377
Book	235680	68152
DVD	41597	16264
Video	38903	13718
Toy	1	1

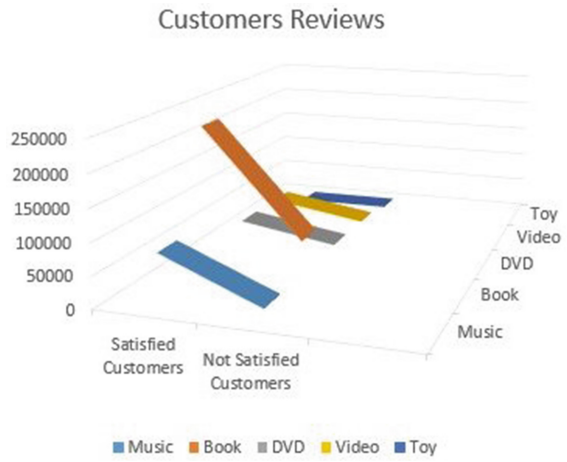

Fig. 2. Customers Reviews

is much bigger than the one of not satisfied ones in four out of five categories (regarding category entitled toy, the number is equal to 1 for both category of customers). With these results, one can easily figure out that Amazon customers are loyal to the corresponding company and prefer purchasing products from the abovementioned categories.

5 Evaluation

The reported values in the charts for the classification models are recorded as AdaBoost, J48, JRip, Multilayer Perceptron, PART, REPTree, RotationForest and SMO. The results for each classifier for their several values are illustrated in Table 3. Depicted in bold, are the selected best classifiers for each value. What is more, Fig. 3 depicts the values of F-Measure for each classifier.

We can observe that J48 achieves the highest score in every category except FP rate. Subsequently, REPTree follows with almost 89 % TP rate and F-Measure, whereas JRip has value of F-Measure equal to 84 %. In addition, concerning F-Measure metric, the other algorithms range from 42 % of Multilayer

Table 3. Classification for predicting

Classifiers	TP rate	FP rate	Precision	Recall	F-Measure
AdaBoost	0.605	0.484	0.598	0.605	0.564
J48	**0.922**	0.095	**0.924**	**0.922**	**0.921**
JRip	0.847	0.187	0.854	0.847	0.843
Multilayer perceptron	0.596	**0.539**	0.66	0.596	0.482
PART	0.748	0.325	0.783	0.748	0.728
REPTree	0.892	0.119	0.892	0.892	0.891
RotationForest	0.785	0.285	0.83	0.785	0.768
SMO	0.574	0.574	0.33	0.574	0.419

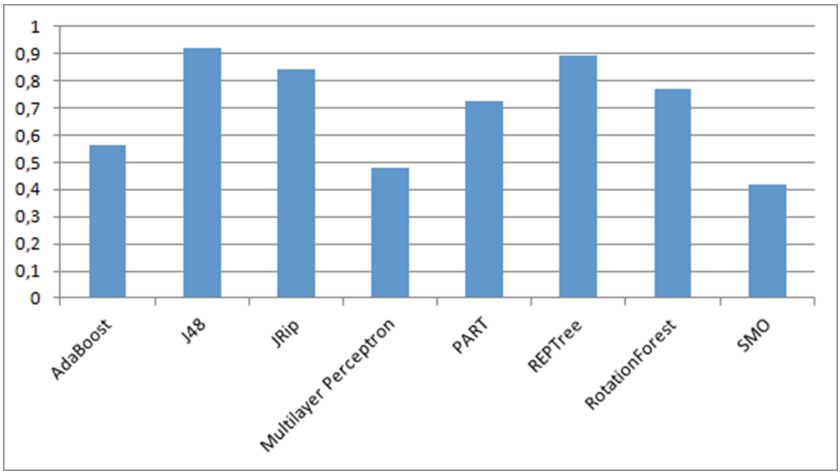

Fig. 3. Customers classification

Perceptron to 77 % of Rotation Forest. Moreover, we see that almost all classifiers achieve a TP rate value of above 60 %, while the percentages for FP rate are relatively smaller. Precision and Recall metrics have almost the same values for each classifier, ranging from 60 % to 92 %.

6 Conclusions and Future Work

In our work we present a methodology to model and predict the purchases of a supermarket using machine learning techniques. More specifically, two datasets are utilized; a supermarket database as well as a database from Amazon Company which contains information about the purchases of customers. Given the analysis of the dataset from Amazon Company, a model that predicts new products for every customer based on the category and the supermarket they prefer

is created. We also examine the influence of person-to-person communication, where we found that customers are greatly influenced by other customer reviews.

As future work, we plan to create a platform using the recommendation network. Customers will have the opportunity to choose among many options on new products with lower prices. Next, we can take into consideration more features of the supermarket dataset, in order to improve classification accuracy. In conclusion, we could use a survey to have further insights and get an alternative verification of user's engagement.

References

1. Allen, J.W., Harrell, G.D., Hutt, M.D.: Price Awareness Study. The Food Marketing Institute, Grocer, Washington, D.C. (1976)
2. Brijs, T., Goethals, B., Swinnen, G., Vanhoof, K., Wets, G.: A data mining framework for optimal product selection in retail supermarket data: the generalized PROFSET model. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 300–304 (2000)
3. Cho, Y.H., Kim, J.K., Kim, S.H.: A personalized recommender system based on web usage mining and decision tree induction. Expert Syst. Appl. **23**(3), 329–342 (2002)
4. Dickson, P.R., Sawyer, A.G.: The price knowledge and search of supermarket shoppers. J. Mark. **54**, 24–53 (1990)
5. Gourgaris, P., Kanavos, A., Makris, C., Perrakis, G.: Review-based entity-ranking refinement. In: WEBIST, pp. 402–410 (2015)
6. Kanavos, A., Kafeza, E., Makris, C.: Can we rank emotions? a brand love ranking system for emotional terms. In: IEEE International Congress on Big Data, pp. 71–78 (2015)
7. Kim, J.K., Cho, Y.H., Kim, W.J., Kim, J.R., Suh, J.H.: A personalized recommendation procedure for Internet shopping support. Electron. Commer. Res. Appl. **1**(3–4), 301–313 (2002)
8. Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., Duri, S.S.: Personalization of supermarket product recommendations. In: Applications of Data Mining to Electronic Commerce, pp. 11–32 (2001)
9. Lescovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (TWEB) **1**(1), 5 (2007)
10. Li, Y., Meiyun, Z., Yang, B.: Analysis and design of e-supermarket shopping recommender system. In: ICEC 2005 Proceedings of the 7th International Conference on Electronic Commerce, pp. 777–779 (2005)
11. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)
12. Lu, J., Shambour, Q., Xu, Y., Lin, Q., Zhag, G.: BizSeeker: a hybrid semantic recommendation system for personalized government-to-business e-services. Internet Res. **20**(3), 342–365 (2010)
13. Malar, L., Krohmer, H., Hoyer, W.D., Nyffeneggeri, B.: Emotional brand attachment and brand personality: the relative importance of the actual and the ideal self. J. Mark. **75**, 35–52 (2011)
14. Pennacchioli, D., Coscia, M., Rinzivillo, S., Pedreschi, D., Giannotti, F.: Explaining the product range effect in purchase data. In: IEEE International Conference on Big Data, pp. 648–656 (2013)

15. Thomson, M., MacInnis, D.J., Park, C.W.: The ties that bind: measuring the strength of consumers' emotional attachments to brands. J. Consum. Psychol. **15**(1), 77–91 (2005)
16. Yao, C., Tsui, H., Lee, C.: Intelligent product recommendation mechanism based on mobile agents. In: 4th International Conference on New Trends in Information Science and Service Science, pp. 323–328 (2010)
17. Weng, S.S., Liu, M.J.: Feature-based recommendations for one-to-one marketing. Expert Syst. Appl. **26**(4), 493–508 (2004)
18. West, C., MacDonald, S., Lingras, P., Adams, G.: Relationship between product based loyalty and clustering based on supermarket visit and spending patterns. Int. J. Comput. Sci. Appl. **2**(2), 85–100 (2005)