International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Customer Gender Prediction Model based on E-Commerce Data

[1]Mrs. S. Sasi, [2]Dr. K. Guru, [3]Dr. J. Anand

Submitted: 08/03/2024 **Revised:** 20/04/2024 **Accepted:** 30/04/2024

Abstract: Demographic data of Customers such as gender, age, etc. provide valuable information for marketing and personalization of web applications for e-commerce service providers. However, online consumers often do not supply this sort of information because of privacy concerns and other safety reasons. In this article, we propose a method for gender prediction based on their catalogue browsing data on e-commerce systems, including the date and time of access, list of categories and items displayed, etc. We use a machine learning approach and explore many features derived from catalogue viewing data to predict the sex of viewers. Experiments on PAKDD Data Mining Competition datasets were carried out and the successful results were achieved. Results 92.3% for balanced accuracy and 91.2% for macro F1 indicate that fundamental features such as time viewing, product/category features used in combination with advanced features such as product/category sequence and transfers effectively promote customers' prediction.

Keywords: *Data analytics, Machine learning, Predictive modelling, Decision-making, Advanced analytics, Data management, Data forecast, Business Value, Big data*

1 INTRODUCTION

It is essential, Many Web apps today use personalization to enhance the user experience, such as e-commerce systems, search engines, online marketing systems. The information shown with a good customised service is uniquely tailored to each user instead of staying the same for all users. E-commerce systems, for example, may show promotions or suggest items that are connected to a visitor instead of random promotions or items. The value of data in today's world has reached new levels where businesses depend on data sets to understand results and to make business decisions. The study of data is particularly important in the fields of e-commerce and retail. They can predict purchasing, earning, loss and even trick customers into buying products by observing their behaviour. Retail brands study consumer profiles to discover their pitfalls to advertise their product to inspire consumers to purchase. Applications and algorithms in data science Engines Suggestion recommendation engines are the most important instruments in an arsenal of retailers. Retailers use these motors to force a consumer to purchase the product. It helps retailers increase revenue and dictate patterns by making recommendations. The engines consist of complex components of machine learning and deep learning algorithms. They are structured so that they can monitor the online activity of any customer and analyse the trends to offer shows based on these results. That's how Netflix recommends movies or television shows every time, which is focused on your past searches and shopping history, and Amazon offers tips and discounts on them. Because face it, who can refuse to buy anything they've always wanted, particularly in the case of discount. This whole process requires a lot of data filtering and analysing with machine learning algorithms.

[1]*Research Scholar, Department of Business Administration, Annamalai University, India.*

[2]*Associate Professor and Head, Department of Management Studies, Takshashila University, India.*

[3]*Associate Professor, Department of Management Studies, SRM Valliammai Engineering College, India.*

Review of Consumer Basket

Consumer basket analysis is one of the most conventional data analytics tools retailers have used for years. Consumer basket Analysis operates on the premise – if a customer purchases a set of papers, he or she will more or less buy other papers. For example, you are more likely to order main course or desserts if you went to the restaurant and ordered

starters or drinks free. The bundle of goods bought by customers is known as an item, the state that a customer is confident after starters. Customers buy products on a pulse basis in retail, and consumer basket analysis operates on this theory by estimating the buying chances of a buyer. This covers a lot of how retailers sell the product and consumer data in the e-commerce environment is the best place to search for future purchasing impulse. Business basket analysis operates like search suggestions with a machine learning or profound learning algorithm.

Analytics of warranty

Guarantee data analysis allows retailers and suppliers to track their goods, their lifetimes, their issues, returns and even to check for fraudulent behaviour. The analysis of warranty data depends on the calculation of the fault distribution based on data that include age and number of returns and the age and number of field surviving units. Distributors and distributors monitor the number of units sold and how many were returned due to problems. They also focus on the identification of irregularities in guarantee statements. This is an ideal way to make warranty challenges a realistic perspective for retailers. Not only for the buyer, but also for the manufacturer or supplier, it is an essential job to sell a commodity at the right price. The price shall not only include the expense of manufacturing the commodity but also a customer's willingness to pay for it, taking into account competitive costs. All this is measured with machine learning algorithms that evaluate a series of parameters such as price versatility, the venue, consumer attitudes and competition pricing. It then provides the best price that both parties will benefit from. This is a powerful tool for retailers to sell their goods at the right price.

Machine Learning.

Inventory refers to inventory of products for subsequent use in crisis periods. In order to maximise capital and increase revenues, inventory management is therefore essential for companies. Retailers need to control inventories efficiently so that supply does not suffer even if sales unexpectedly spike. The supply chains and stock chains are extensively analysed to accomplish this. Effective machine learning algorithms analyse in great detail data between elements and deliver and identify trends and associations between transactions. Then, the analyst analyses these data and has a plan for revenue growth, timely distribution and inventory management. A significant aspect of data processing is position processing. Until a company can decide where to start its business, it is important to evaluate plausible locations in order to settle the best. The algorithm used is simple but efficient in this case. The analyst analyses data that offer demographic importance. The coincidences of zip codes and sites provide a framework for understanding the business opportunity. Competitor markets are often taken into account in the study of places. The algorithm also analyses the most appropriate alternative for retail networks.

Analysis of consumer perception

For a long time, consumer sentiment analysis has been present in the business world. However, machine learning algorithms are now helping to simplify, automate and save time with accurate performance. Social networking is the simplest and most readily accessible method for an analyst to evaluate consumer feelings. It uses language analysis to classify terms with a customer's negative or positive attitude to the brand. This feedback allows companies to develop their products. Merchandising is an integral part of any retail business. The concept is to build techniques to maximise product sales and promotions. Merchandising aims to influence the decision making of customers via visual channels. While attractive packaging and branding maintain the customer's attention and boost consumer appeal, changing items help to keep the products fresh and new. Data sets, observations and priority set of customers take account of seasonality, importance and patterns are used to build merchandising algorithms.

At retail, the customer's value for the entire customer-business relationship is the net value of customer benefit for the company. The revenue is given special consideration, as far as it is not predictable by costs. Via direct purchase assessment, organisations can recognise two critical methodologies for consumer life; historical and predictive. All predictions are made with past data leading to the current transactions. The algorithms typically gather, identify

and clean data regarding consumer desires, prices, recent transactions, and behaviour. After the data is analysed, the potential value of the current and possible customer is viewed linearly. This algorithm also recognises interconnections between the attributes of the consumer and their choices. Data science has applications in all technology fields and allows enterprises to make better data-driven decisions. The above 9 applications are among the most common and significant in the field of e-commerce. Check the positions in data science in order to understand the various possibilities available.

2 RESEARCH METHODOLOGY

The personalization of the data relates specifically to two forms of data: historical (e.g. previously viewed or purchased items) and demographic (e.g. sex, age, education, etc.) of the consumer. Only if the user has used the system previously and logs into the system will historical data be accessed. Historical approaches based on data are also useless for visitors or new users. Even if the consumer never used the device before, demographic methods are useful. This knowledge is, however, not easily accessible because anonymous texts have been studied for decades by internet users but not by anybody on the framework. Another way to predict users' demographic details is through their activity on systems such as browsing traffic on websites and data from the catalogue. The biggest benefit of this approach is that data is mostly accessible as users need to do something on the system such as access pages, click objects or search the catalogue. This research addresses the problem of forecasting demographic consumer details on the basis of their catalogue data such as browsing time / duration, categories / products viewed, etc. FPT Corporation presented datasets of our experiment at the PAKDD (Pacific-Asia Knowledge Exploration and Data Mining). Competition for Data Mining. Apart from fundamental features such as display time, categories and individual products seen in the session, we analyse features containing information on relation of products/ categories that have been seen in the session, such as product/category sequence and transition, etc. (we refer to them as 'advanced features'). We used a tree-based feature representation with a multi-level hierarchic category / product structure, which offers a better view than the list-based style for function extraction. Common learning methods such as Random Forest, Support Vector Machine (SVM) and Baiesian Network (BayesNet) were used for experiments and we used support techniques such as resampling, cost-sensitive learning to improve prediction overall. The findings are positive, although more features and techniques need to be studied in order to improve performance.

The paper is structured as follows. At the beginning, most investigations in this field concentrated on author's studies, tasks to evaluate or predict the characteristics of the author by analysing texts he / she has made. Methods used by researchers in these studies are primarily focused on writing style analysis using different feature types such as lexical, syntactic, or contents. The previous related researches concentrated frequently on literature texts such as novels or posts. Recently, the focus was shifted to computer media content, such as email, blogs and comments, due to the growth of the Internet and online media outlets. In addition to the methods used to evaluate textual data, researchers recently studied the use of user activity on web applications to predict their demographic details. They used information on the website as input variables to distribute users' demographic information. The SVM approach was used on features consisting of content-based features (web pages) and category-based features (web definition hierarchy), resulting in 89.7% of gender and 73.3% of age. Some study also explored machine learning methods for the prediction of websites' demographic attributes through content and linked structure information. Some research aimed to draw the population of users on the basis of their mobile contact habits on a regular basis. Their research was carried out over 7,000,000 individuals and over 1,000,000,000 communications records on a wide mobile network in the world. They used individual features, friendly features and circles, and obtained the best results from 80 percent for sex and 70 percent for age. We explored a way of predicting demographics for consumers based on their catalogue navigation habits on e-commerce systems in this paper. As far as we know, there is no detailed research on this subject.

3 OVERVIEW OF THE FRAMEWORK

In this paper, we built a framework that allows users with known gender, extract features and class labels to construct a training dataset with data from product viewing logs. A model is constructed from a classification based dataset and can then be used to predict the sex of unknown users based on their product viewing behaviours. The training data file provides documents that fit the product viewing logs. A single log provides details of products viewing a user's data, such as the session start time, session end time, list of products and category IDs. The class labels are male and female for each training sample. The job is therefore a binary classification problem with two labels. In the following sections we explain detailed features and techniques used for prediction. The feature set used in this work can be divided into two groups, which are referred to as basic and advanced. Temporary and individual product / category features are important features. Temporary features are related to timestamps and visual frequency. Hour a day, day a week, holidays, the duration of the display, the number of items displayed at a session etc. can be used to determine a customer's gender. We used 98 binary features and 3 such numerical features as shown in Table I.

TABLE I: TEMPORARY FEATURES

Duration	Features
Day	47
Month	18
Week	8
Duration	1
Number of products	1
Average time per product	2

Features of each product / category consist of all device categories and products. Since all the categories and product IDs are given, we have only extracted them and used them as a feature. We count the number of times that the user browses it and uses that number as the feature value for each category or product. Since each entire product ID can be divided into four different IDs, from the most general groups (starting with 'A') to the subcategories (starting with 'B' and 'C') and the individual product (Starting with 'D'), we have four types of features, which equal 2,057 features. We only select the IDs that appear at least five times for the large number of individual product IDs. In addition to individual categories / products characteristics, we believe that the relationship between the session categories / products also represents the sex of the audience. The categories / products are described in different styles and patterns. Because the lists can trigger problems for extracting all the relationship information between individual categories / products, we have suggested a tree-based presentation, in which the root of the tree is the most general category, the products are located on the tree leaves and the subcategories are situated at intermediate levels. For example, as in the Figure, the above mentioned category / products presentation can be converted to a tree-based fig1.

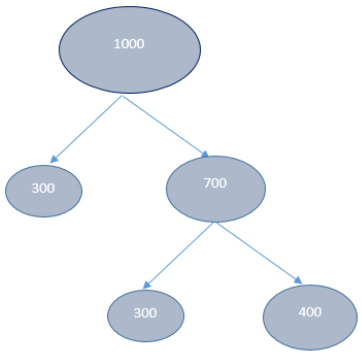


Fig. 1. Tree Based product presentation

Based on this tree, the list of categories / products can easily be obtained by travelling the tree deep and from the leftmost. In addition, from the tree view, we can extract information related to the categories / products by analysing tree properties such as nodes, levels, pathways, siblings, etc. The overall potential sequences and move pairs can be very big because of the vast number of categories and products. Therefore, we only pick sequences and pairs that appear in the training dataset at least five times.

We used three Random Forest, SVM, and BayesNet algorithms to learn the model. However we employed such supporting strategies such as cost-sensitive learning, resampling, class balance to improve accuracy, since the training data is unequalled (around 80% female and 20% male). Resampling approaches are also used to deal with the issue of

class imbalance. The fundamental concept is to add or delete instances in order to make the dataset more balanced. There are also two resampling methods that are under-sampled (reduce the number of large classes) and over-sampled (reproduction of small class cases). The main downside of under-sampling is that this approach may discard possible task-related data while over-sampling can cause additional calculation costs and over-fit in the event of a random replication. The class balance strategy is another tool for re-evaluating the instances of each class to achieve the same overall weight instead of duplicating or reducing instances. We used the class balance approach in conjunction with the cost-sensitive learning algorithm in our experiments. Although re-evaluation is the data-level approach, cost-sensitive learning is a method based on algorithms to overcome the class-imbalance classification issue. Cost-sensitive learning is a process that recognises the cost of misclassification, which means it handles the various erroneous classifications differently. Finally, since the total number of features (5,300 features) remains very high, we apply a technique of selection to minimise complexity and exclude non-discriminatory features. We used the information gain in our work to pick 3,500 features with the highest shared information. We used datasets from FPT Corporation's PAKDD'15 Data Mining Competition. The data were split into training and test sets. Every collection contains 15,000 records that match the product viewing logs. Four forms of knowledge consist of a single log in the training data file (ID of the session, Start time, End time, Product IDs List). The product IDs list includes the product IDs that the user accessed during the session. Due to the fact that goods which belong to various categories, category information is also included in IDs. As already stated, the balanced accuracy measure (BAC), due to the class imbalance problem, was used to test the model. Average accuracy in both classes is characterised as a balanced accuracy that can prevent inflationary output estimates of imbalanced data sets.

TABLE II: RESULTS OF COST-SENSITIVE LEARNING WITH CLASS BALANCING

	Basic features		Basic and Advanced features	
	BAC	Macro F1	BAC	Macro F1
Random Forest	87.3	85.8	92.3	86.8
SVM	88.8	86.4	89.8	89.2
BayesNet	86.2	84.4	88.4	86.6

TABLE III: RESULTS OF EXPERIMENTS WITH COST-SENSITIVE LEARNING

	Basic features		Basic and Advanced features	
	BAC	Macro F1	BAC	Macro F1
Random Forest	86.6	87.4	90.8	91.2
SVM	86.2	86.4	89.0	88.4
BayesNet	85.2	85.6	88.8	84.4

As shown in Table III, Random Forest achieved best results in cost-sensitive class-balancing, while BayesNet produced the lowest output on both BAC and Macro F1 scores in which the best BAC scoring of 92.3% is higher than the Macro F1 score of 78.8%. But Table V reveals that the best Macro F1 score increased to 91.2 percent by utilising cost-sensitive schooling, and the BAC score decreased to 80.8 percent. The advanced features in conjunction with basic features also increase the prediction result considerably compared to only using basic features. However, there are several sessions in which users view only one product in the given data sets, and advanced features have little impact on such instances. In balanced precision (which is equal to

0.5 * tp) is 0.5 * tn. TP allegedly fn & tn allegedly, users see over a few goods as they surf in e-commerce systems. We therefore think that the change will increase more when our approach is implemented. Where tp is true positive, tn is true negative, fp is false and fn is incorrect. In order to assess results in the PAKDD'15 data mining competition, this metric was also used. In this paper, we report the score along with macro F1 to make comparisons with earlier works easier.

4 FINDINGS

In order to determine the efficiency of basic and advanced features, we have carried out experiments with various sets of features, including only fundamental features and combination of both features. The common master learning methods, namely Random Forest, SVM and BayesNet, have been used to test all features. The training data and evaluation datasets are supplied separately (15,000 samples per dataset). Our model was developed and evaluated on a different dataset based on the training dataset. We also examined the effects of supporting techniques such as cost sensitive learning, resampling and class balancing, using different methods we experimented with and found that the best Marcro F1 score 91.4 percent was obtained using cost sensitive learning alone with cost matrix 1:4, but in combination we used more practical class balancing data sets.

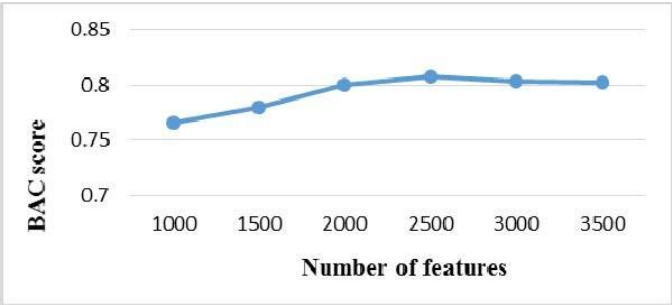
We have chosen 2,500 for the number of characteristics since we have carried out experiments with various features ranging from 1000 to 3,500 and have found that the predicted results increase and hit a top of 2,500 characteristics. Fig. 2 displays the BAC method scores by using various functions.

Fig. 2. BAC Score

The basic results of demographic prediction tasks on the basis of text analysis are 80% for gender (precision). The BAC and Macro F1 gender forecast score in our work can therefore be considered promising. The F1 score of our work have predicted the sex of users based on their Web browsing data, while browsing activities produce more relevant training data than product viewing activities. In addition, web pages contain words, so they can use more features, including content words or page topics, etc. The outcome is the top 10 researchers compared to other solutions from the teams that participate in the PAKDD'15 Data Mining Competition. The best team's score is 87.9% and the top 10 positions obtained the results of 81%. However, most of their solutions used unique data sets characteristics such as product ID prefix, session alignment between sessions in training and test sets, etc., to achieve the highest possible scores in the contest. Thus, when applying to other data sets, such solutions could not produce such good results. We are investigating a general approach that can be applied to any dataset in this paper and therefore we have not tried to use those features.

5 CONCLUSION

We tested an approach in this study to predict consumer gender based on product viewing data on e-commerce systems. We have suggested an approach that utilises simple features such as viewing time and length, various categories / products, combined with advanced features, such as the category / product sequences and transmission pairs. List of products. /products list. This feature design works best on the Random Forest algorithm with supporting techniques such as cost-sensitive learning and resampling. Moreover, the advantage of the approach is that it can easily be extended to other datasets since it does not use unique dataset functions. In the future, the feature set can be further explored. The tree-based presentation will deduce further features to leverage the link between the items shown in the same session. We also intend to collect data from other sources in order to increase overall efficiency and to extend to other demographic attributes including age, location, work etc.



References

[1] S. Argamon, J. Good, M. Koppel, A. Shimoni, "Gender, genre, and style of writing in formal texts," Text 23(3), August 2003.

[2] S. Argamon, J. Pennebaker, M. Koppel, J. Schler, "Profiling an anonymous text automatically," ACM Communications, v.52 n.2, February 2009.

[3] J. C. A. Culotta, R. N. Kumar, J. Cutler, "The Demographics of website traffic data users on Twitter," Proceedings of 29th AAAI Artificial Intelligence Conference, Jan 2015.

[4] O. A. Anderson, de Vel, M. Corney, and G. M. Mohay, "Mining forensic authors' email material," SIGMOD Record 30(4), p. 55-64, 2001.

[5] And. Y. Yang, J. Tang, Y. Yang, N. V. Chawla, "Inferring demographics of consumers and networking tactics in mobile social networks." ACM. ACM. 15–24, 2014. 2014.

[6] D. T. Duc, T. Duc, P. B. Hanh, In: Recent Advances in Intelligent Information and Database Systems, pp. 286–295: "Using content based features for profiling Vietnamese forum articles" Berlin, 2016, Springer International Publishing.

[7] J. Hu, H. J. Zeng, H. Li, C. Niu, C. Chen, "Demographic forecast based on consumer surfing behaviour," 16th International World Wide Web Conference Proceedings, pp. 151-160, 2007.

[8] F. M. Debbabi, Iqbal, B. C. M. Fung, L. A. Khan, 'Verification of forensical inquiry by email authorship,' Proceedings for the 2010 ACM Applied Computing Symposium, ser. SAC '10. SAC' New York, NY, United States: ACM, pp. 1591-1598.

[9] S. Kabbur, E. H. Han, G. Kabbur. Karypis, "Website-based predicting demographic attributes content based approaches," ICDM Proceedings, pp. 863-868, 2010.

[10] M. S. Argamon and Koppel, and A. R. Shimoni, "Automatic gender categorization of written texts," Literary and Linguistic Computing, 17(4), pp: 401-412, 2002.

[11] S. Kotsiantis, D. and Kanellopoulos, P. The GESTS International Transactions on Computer Science and Engineering 30(1), pp. 25-36, 2006, "Making Unbalanced Datasets: A Study."

[12] C. X. Ling, V. S. Sheng, "Cost-sensitive learning and the problem of class imbalance." In: Sammut C(ed) Machine Learning Encyclopaedia. Berlin, Springer, 2008.

[13] M. And A, Pennachiotti. M. Popescu, "A Twitter user classification machine learning solution." AAAI prosecutions, 2011.

[14] T. M. Phuong, D. V. Phuong, "Gender history prediction," KSE 2013 Fifth International Conference Proceedings, volume 1. 271-283, 2013. 2013.

[15] D. R. Gravel, D. Trieschnigg, and T. R. Gravel. Meder: "How old do you think I am? language and age research in twitter," Seventh International AAAI Weblog and Social Media Conference 2013.

[16] F. Rangel and P. Rosso, "Use of language and profiling by authors: gender and age identity," in natural language processing and cognitive sciences, p. 177, 2013.

[17] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, In Proceedings of the Weblog Computational Approaches Symposium of the AAAI Spring, pp. 191-197,2006.

[18] R. E. Schapire, 'Machine Learning Enhancement: Summary,' Proc. MSRI Non-linear Estimation and Classification Workshop, 2001.

[19] J J. J. C. Ying, Y. J. Chang, C. M. Huang, and V. S. Tseng, In Nokia Mobile Data Challenge, 'demographic forecast based on mobile users,' 2012.

[20] C. Zhang, and P. Zhang, "Predicting gender from blog posts," Technical Paper, Massachusetts University Amherst, USA, 2010.