ISSN: 2521-0874 (Print)
ISSN: 2521-0505 (Online)

Contents List available at RAZI Publishing
Acta Informatica Malaysia(AIM)
Journal Homepage: http://www.razipublishing.com/journals/acta-informatica-malaysia
https://doi.org/10.26480/aim.01.2017.26.28

A CLASSIFICATION APPROACH FOR NAÏVE BAYES OF ONLINE RETAILERS.

Aida Mustapha, Shazwani Mustapa, Nurfarahim Md.Azlan, Noor Fatin Ishmah Saifarrudin, Shahreen Kasim, Mohd Farhan Md Fudzee, Azizul Azhar Ramli, Hairulnizam Mahdin, Seah Choon Sen

Faculty of Computer Faculty Computer Science & Information Technology, University Tun Hussein Onn Malaysia, Johor, Malaysia
seanseah0702@gmail.com, {hana, shahreen, farhan, azizulr, hairuln}@ uthm.edu.my

ARTICLE DETAILS

Article history:

Received 22 January 2017
Accepted 03 February 2017
Available online 05 February 2017

Keywords:
consumer-centric marketing,
online retailer

ABSTRACT

Many small online retailers and new entrants to the online retail sector are keen to practice data mining and consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so. In this article a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. On the basis of the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using the classification and naïve bayes algorithm, and the main characteristics of the consumers in each segment have been clearly identify ed. Accordingly a set of recommendations is further provided to the business on consumer-centric marketing.

1. Introduction

Data mining is the process of discovering interesting knowledge, such as associations, patterns, changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories. It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found in fields such as Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management and Decision Support.

The online retailer under consideration in this dataset is a UK-based and registered non-store business with some 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogues, and orders were taken over phone calls. It was only 2 years ago that the company launched its own web site and shifted completely to the Web. Since then the company has maintained a steady and healthy number of customers from all parts of the United Kingdom and Europe, and has accumulated a huge amount of data about many customers. The company also uses Amazon. co.uk to market and sell its products [1].

The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. So, customers of the business have been segmented into various meaningful groups using the naïve bayes classification algorithm the main characteristics of the consumers in each segment have been clearly identified. Accordingly, a set of recommendations is provided to the business on customer-centric marketing and further data analysis tasks. The dataset was collected by using mark reports.

On the basis of the RFM model, customers of the business have been segmented into various meaningful groups using the clustering algorithm and decision tree induction, and the main characteristics of the consumers in each segment have been clearly identified.

2. Related Work

Researchers have proposed many different approaches for sentiment analysis. In this previous technical article for the online retails industry is used k-means clustering to create some nested segments internally inside the cluster. In other words, these nested segments form some sub-clusters inside cluster, and make it possible to categorize the consumers concerned into some sensible subcategories.

With the prepared target dataset they intended to identify whether consumers can be segmented meaningfully in the view of recency, frequency and monetary values. The k -means clustering algorithm was employed for this purpose, and it can be easily performed by using the Cluster node in SAS Enterprise Miner. As well-known, the k -means clustering algorithm is very sensitive to a dataset that contains outliers (anomalies) or variables that are of incomparable scales or magnitudes.

3. Methodology

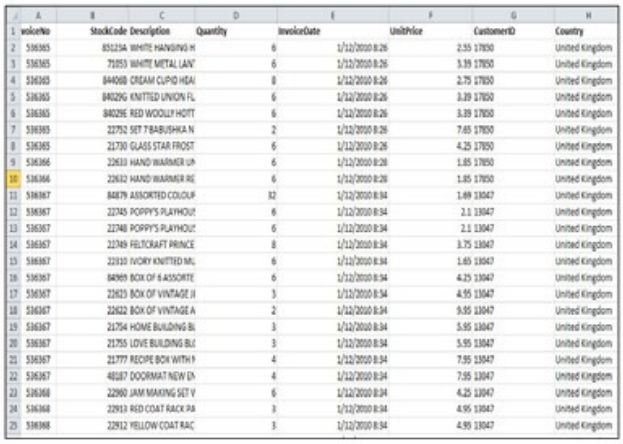
For this online retailer dataset, we choose classification methodology. Classification is a function of data mining, distributed collection in a project goal category or class. Classification is the purpose of accurately predicting each case of target class data. Classification task begins in a data set of class assignments are known Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Credit rating would be the target, the other attributes would be the predictors, and the data for each customer would constitute a case.

Advantages

- Users are thus better able to find precisely what they want, whether they wish to search in one discipline or across all. Works need not but can be coded by discipline.
- While other classification systems provide specific instructions in multiple places for coding by time or place or people, this system has a universal coding for such elements. This renders both classification and searching easier.
- The classification is able to use natural language but yet there are very precise definitions associated with each concept in the BCC.

Following these steps a target dataset for the analysis has been generated. The original dataset was in MS Excel format, and was Part of the target dataset is shown in figure 1.



	A	B	C	D	E	F	G	H
1	voiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING H	6	1/12/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LAN	6	1/12/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEA	8	1/12/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FL	6	1/12/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTT	6	1/12/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA N	2	1/12/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROST	6	1/12/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UN	6	1/12/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RE	6	1/12/2010 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR	32	1/12/2010 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOU	6	1/12/2010 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOU	6	1/12/2010 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCE	8	1/12/2010 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MU	6	1/12/2010 8:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTE	6	1/12/2010 8:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JI	3	1/12/2010 8:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE A	2	1/12/2010 8:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BL	3	1/12/2010 8:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BL	3	1/12/2010 8:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH M	4	1/12/2010 8:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW EN	4	1/12/2010 8:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET V	6	1/12/2010 8:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PA	3	1/12/2010 8:34	4.95	13047	United Kingdom
25	536368	22912	YELLOW COAT RAC	3	1/12/2010 8:34	4.95	13047	United Kingdom

Figure 1: sample of the target dataset

Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification.

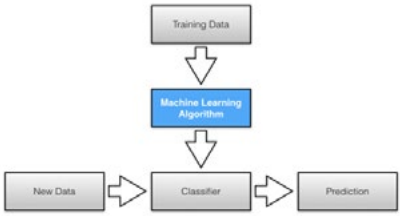


Figure 2: A simplified diagram of the general model building procedure for pattern classification.

The chosen datasets were downloaded from the UCI repository (University of California, Irvine) [4]. Table I shows their characterization. The name by which they are known in the literature, the variables data type with nominal grouped into categorical, the number of Attributes and the number of possible with numeric type. In the following sections, we will take a closer look at the probability model of the naive Bayes classifier and apply the concept to a retailer online data.

Table 1: Variables in the target dataset

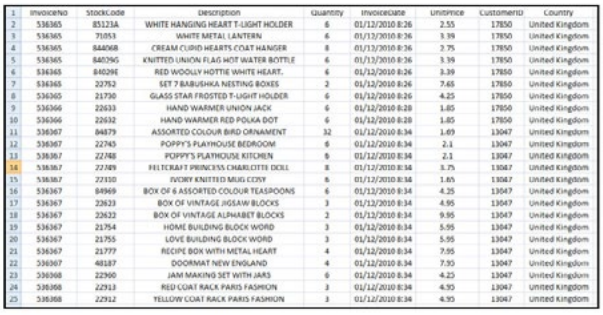
Variable Name	Data Type
Voice No	Nominal
Stock_Code	Numeric
Description	Nominal
Quantity	Numeric
Invoice_Date	Nominal
Unit_Price	Numeric
Customer_Id	Numeric
Country	Nominal

In this paper, we try to give the basic concept of classification technique using naïve bayes Support vector that can be used for classifying the unknown records [2-4].

4. Experiment

a. Dataset

The experiments that were carried out in this research were done using by Rapidminer. Rapidminer is a java-based open source data mining and machine learning software. It has a graphical user interface (GUI) where the user can design his machine learning process without having to code. Here, the classification task for a given classifier is designed as a process. From the data we gathered, Algorithm that we use in online retailer is naïve bayes classification algorithm. We provided with a training dataset consisting of information about online retailers. Figure 3 show the dataset was in the form of a Microsoft Excel 2003 spreadsheet and had details of country, invoice number, stock code, quantity, description, quantity, invoice date, unit price and customer id. For ease of performing data mining operations, the data was change into a csv format.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 8:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	01/12/2010 8:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 8:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 8:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 8:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01/12/2010 8:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01/12/2010 8:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	01/12/2010 8:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	01/12/2010 8:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01/12/2010 8:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01/12/2010 8:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01/12/2010 8:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01/12/2010 8:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	01/12/2010 8:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01/12/2010 8:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	01/12/2010 8:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	01/12/2010 8:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	01/12/2010 8:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	01/12/2010 8:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	01/12/2010 8:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	01/12/2010 8:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	01/12/2010 8:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	01/12/2010 8:34	4.95	13047	United Kingdom
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	01/12/2010 8:34	4.95	13047	United Kingdom

Figure 3: Dataset for Online Retailers

The next step was to feed the pruned online retailer database as input to RapidMiner. This helped us in evaluating interesting results by applying classification algorithms on the online retailer training dataset.

b. Algorithm of naive Bayesian Classifier

Naïve Bayes:The design view of Naïve Bayes Classifier is given below in Figure 4 which includes the UCI dataset with 1000 instances and 8 attributes, a cross validation Operator with 10 fold cross validation, Apply Model and Performance evaluation operator with necessary parameter values for execution.

Figure 4 : Design view of Naive Bayes

Figure 5 shows the example dataset online retailers when apply the naive bayes. It include 1 special attribute which is country and 7 regular attribute.

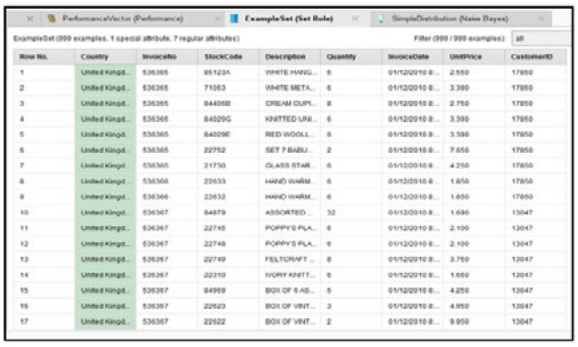


Figure 5 : Set role of Online Retailers

c. Evaluation Matric

Evaluation Metric is a process of assessing progress towards the goals and objectives that have been defined. It including information on the efficient use of resources in producing to compare with the results of the intended meaning and effectiveness in achieving the objectives of the action. This evaluation metric will be proof to show that the online retailer classification are correctly or incorrectly classified using classification accuracy, coefficient metric and time taken test model on dataset. The Confusion Matrix obtained is given below in instances and 8 attributes is fed as input to the Naïve bayes figure 6and figure 7 is plot view for analysis.

Cite this article as: Aida Mustapha, Shazwani Mustapa, Nurfarahim Md.Azlan, Noor Fatin Ishmah Saifarrudin, Shahreen Kasim, Mohd Farhan Md Fudzee, Azizul Azhar Ramli, Hairulnizam Mahdin, Seah Choon Sen, A CLASSIFICATION APPROACH FOR NAÏVE BAYES OF ONLINE RETAILERS / Acta Inform. Malay (1) (2017) 26-28

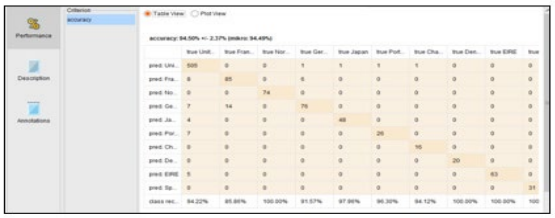


Figure 6 : Naïve Bayes Confusion matrix

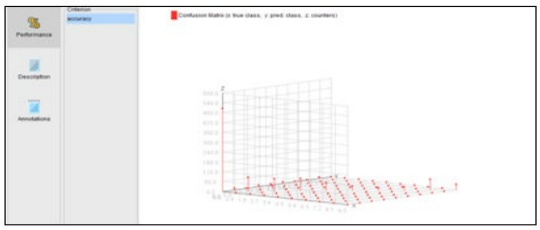


Figure 7: Plot View

5. Result

Results of the empirical study show that prediction the country in online retailer dataset. The classification of this datasets was conducted to prove that the histograms in figure 8 of the variables country the target dataset of online retailer it is evident that there are a few instances having quite different monetary and frequency values compared to the majority of the country in the dataset. These instances are valid from the business point of view as they are genuine transaction records; however, they are outliers from the data analysis point of view. United kingdom is the highest ranking country for delivery the item.

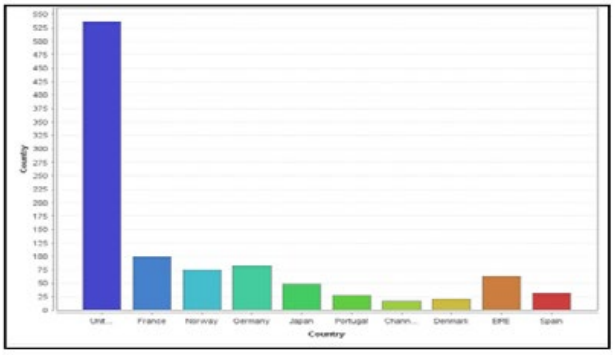


Figure 8: Bar chart of country

RapidMiner helped significantly in finding hidden information from the training dataset. These newly learnt predictive patterns for predicting country performance in online retail. In figure 9 it show the simple distribution for label attribute country and figure 10 show the distribution table. We can see each country have 7 distribution.

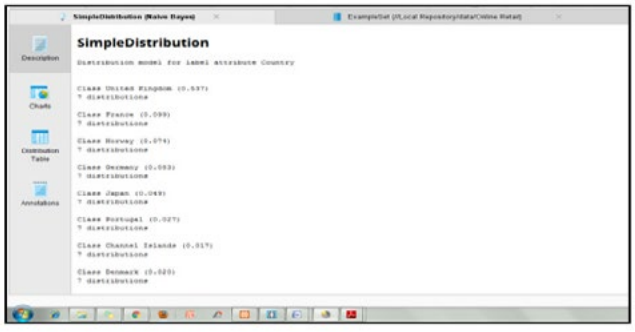


Figure 9: Simple Distribution for Country

Simple distribution included distribution model for label attribute Country which is United Kingdom, France, Norway, Germany, Japan, Portugal, Chanel Island and Denmark.

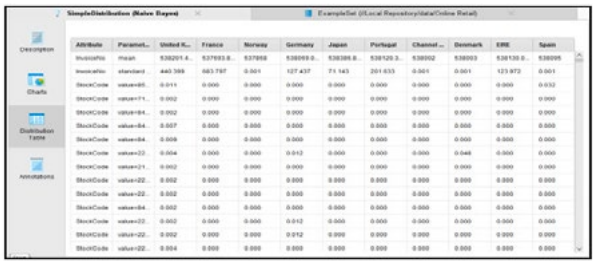


Figure 10: Distribution table

In figure 11, it shows the example chart of country based on the quantities of each product (item) per transaction. We can see that portugal is the highest density with 0.068.

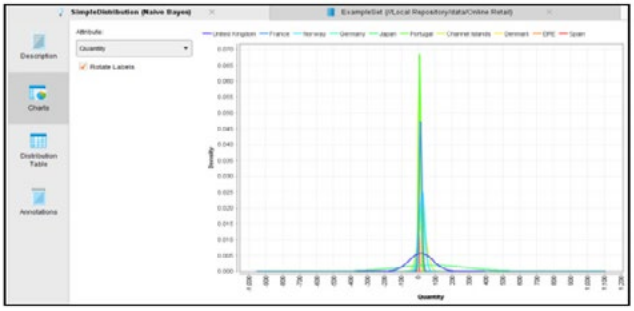


Figure 11 : Chart of Country based on Quantity

6. Conclusion

A case study has been presented in this article to demonstrate how customer-centric business intelligence for online retailers can be created by naïve bayes of data mining techniques. The distinct customer groups characterized in the case study can help the business better understand its customers in terms of their profitability, and accordingly, adopt appropriate marketing strategies for different consumers. It has been shown in this analysis that there are two steps in the whole datamining process that are very crucial and the most time-consuming: data preparation and model interpretation and evaluation.

7. References

1. Kumar , V . and Reinartz , W . J . (2006) Customer Relationship Management: A Databased Approach ,Hoboken, NJ: John Wiley & Sons .
2. “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998 by C. Burges.
3. Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.
4. Sebastian Raschka (2013) Naive Bayes and Text Classification