Mining the Web Data for Classifying and Predicting Users' Requests

Girish S, Ramamurthy B, Senthilnathan T
Department of Computer Science, Christ University, Hosur Road, Bengaluru, Karnataka, India

Article Info

Article history:

Received Jan 31, 2018
Revised Apr 2, 2018
Accepted Apr 20, 2018

Keyword:

Customer Segmentation
Classification
Prediction
Trend
E-commerce
Product Recommendation

ABSTRACT

Consumers are the most important asset of any organization. The commercial activity of an organization booms with the presence of a loyal customer who is visibly content with the product and services being offered. In a dynamic market, understanding variations in client's behavior can help executives establish operative promotional campaigns. A good number of new consumers are frequently picked up by traders during promotions. Though, several of these engrossed consumers are one-time deal seekers, the promotions undeniably leave a positive impact on sales. It is crucial for traders to identify who can be converted to loyal consumer and then have them patronize products and services to reduce the promotion cost and increase the return on investments. This study integrates a classifier that allows prediction of the type of purchase that a customer would make, as well as the number of visits that he/she would make during a year. The proposed model also creates outlines of users and brands or items used by them. These outlines may not be useful only for this particular prediction task, but could also be used for other important tasks in e-commerce, such as client segmentation, product recommendation and client base growth for brands.

Corresponding Author:

Girish S,
Department of Computer Science,
Christ University,
Hosur Road, Bengaluru, Karnataka, India.
Email: girish.s@cs.christuniversity.in

1. INTRODUCTION

E-commerce with the help of the World Wide Web has taken world businesses to the next level, where the traditional approach is overshadowed by the seamless, quick, efficient and mind-blowing assistance of technology. It has paved the way for a convenient form of conducting business. Now, that E-commerce is the trend in business organizations and it is here to stay for the coming years, firms in general must gear to automating their consumer data and the varied information on their preferences in order to retain their existing clients and attract the new clients. 'Customized Servicing' is predicted to be the only way forward in these times of demanding consumers, who today not only look forward to the best service in the least possible time but cost-effective service too. Data-mining techniques analyse large amount of data without any pre-defined hypothesis to extract meaningful information, rules and constraints. Data Mining is a process of extracting knowledge from databases, involving pattern extraction algorithms. Data-mining has a collection of various techniques to extract patterns and to build models from large data-sets. The challenging part for most of the businesses today is to understand the needs of their customers in a dynamic environment. In such a situation, change mining is used by analysts to understand customer's needs.As decision-tree algorithm is a classification-based algorithm, it doesn't involve a complete change in the data-set. 'Rule Extraction' was the most widely used techniques in understanding the relationship between various product items purchased by the customers proposed a system that can identify changes in customers' behaviour. The system uses previous purchase records to determine and identify products that a customer may like. Customer behaviour gives us demographic variables used to analyse patterns. Recency, Frequency and monetary(RFM) can be used to differentiate between consumers.

This study integrates a classifier which can predict the needs of the customers and the type of purchases they might make. In this paper, we have categorized both products and customers based on their previous records and then similar users were grouped together. The extracted information can be used for various e-commerce applications such as product recommendation, Customer Insights, Customer segmentation and User-based brand identification. The generated user profiles or features can be used by merchants to know their customers better.

2. RELATED WORK

Mining is the process of extracting knowledge or information from the web [1]. Web Mining also has its own types or ways in which it is treated according to what kind of data it contains [2]. Content Mining is knowledge or information gained from the content of the site [3]. Structure Mining is the topology of the site or a way in which the references or links are put at the site; Usage Mining is extraction of Information from the user login-in credentials and stored as user details accordingly. Also the concept of web mining from server log details was thrown into light [4]. An attempt was made to classify users based on the site's visitors but it lacked accuracy because content wasn't taken into consideration. Classification attempts were also made on text contents of the users visited sites with the help of the local cache and cookies [5]. However, as this was only based on the recent visits and clicks and user's intentions are subject to change at any time, it wasn't easy to give exact prediction each time as users taste or interests tend to change rapidly. It became difficult to cache pages and predict them. Clustering Based on pages' access and page sequence was also an attempt made where results were drawn based on the session timings [6]. There are some already existing systems which help the web designers in organizing their websites accordingly in both recommendations method and offline methods [7]. Recommendations are generally based on a previous user's interest and if the pattern- match occurs, a recommendation is put forward to the user. According to [8] there are few ways in which 'Content Mining' can happen; Pre-mining, where the sessions only involve contents from the site and Post-mining, where the content and the results are independent [9]. Mining is the process of extracting knowledge or information from the web. Web Mining also has its own types or ways in which it is treated according to what kind of data it contains. Content Mining, is nothing but knowledge or information gained from the content of the site [10]. Also the concept of web mining from server log details was thrown into light [11]. Which is a combination of different systems put together for better results [12], [13]. Web mining helps in improving the scalability and effectiveness of a site. An approach of using semantic data gathered from web mining and show how semantic data can be used to personalize one's website. Also shows how to use semantic data to improve the traffic attracted towards site [14], [15]. A rule based page classification was proposed [16]. A model where user navigation profiles are generated with the help of web mining from the data acquired from the servers. This approach is based on byte-level and language is independent, the profile size is limited and the accuracy rate is based the language inputted [17]. The record of events occurring over a period of time is collected from the server domains [18]. The problems in recording the sequential occurrences of events and each of this actions are split into sessions they show that one such sessions have the data which we can use to form rules for describing the next occurrence of an event. They provide an algorithm which is used to help record the events and provide description for it [19], [20].

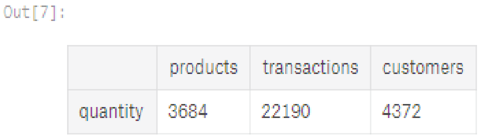
3. PROPOSED METHODOLOGY

This paper aims at analysing the content of an E-commerce database. Based on the analysis, a model was built to predict the purchases of a new customer based on his/her earlier purchasing track record.

3.1. Data preparation

The dataset was selected from an E-commerce dataset comprising 400,00 entries. Figure 1 shows that the data contained 4372 users and they had purchased about 3684 products and the total number of transactions carried out were 22000.

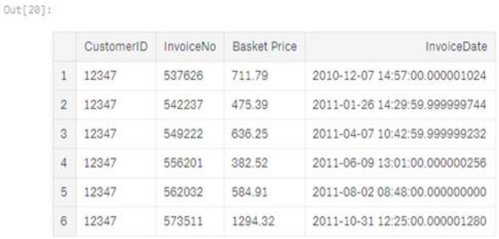
The next step was to arrive at the, number of products bought per transaction and after this all the null values and transactions where orders had been cancelled were removed from the data set. A variable was then created to show the total price of each purchase made by the customer. Figure 2 shows a sample of how basket price is calculated for each transaction.



Out[7]:

	products	transactions	customers
quantity	3684	22190	4372

Figure 1. Summary of the whole data set



Out[20]:

	CustomerID	InvoiceNo	Basket Price	InvoiceDate
1	12347	537626	711.79	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	2011-08-02 08:48:00.000000000
6	12347	573511	1294.32	2011-10-31 12:25:00.000001280

Figure 2. A sample of basket price for each transaction

3.2. Understanding product categories

To begin with, we first extract information regarding products from the description Variable. Figure 3 shows how the basket price of each transaction is divided for the whole dataset. The process of extraction of information is as follows:

a. Extract the terms from the description

b. For each of these terms, try to find out the root and collection of set of terms related with it

c. Calculate the occurrence of term in the data set

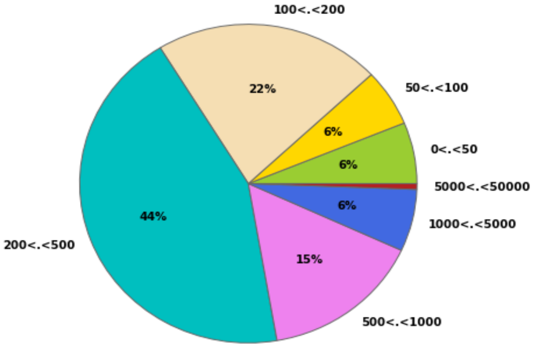


Figure 3. A pie-chart displaying the price against each invoice

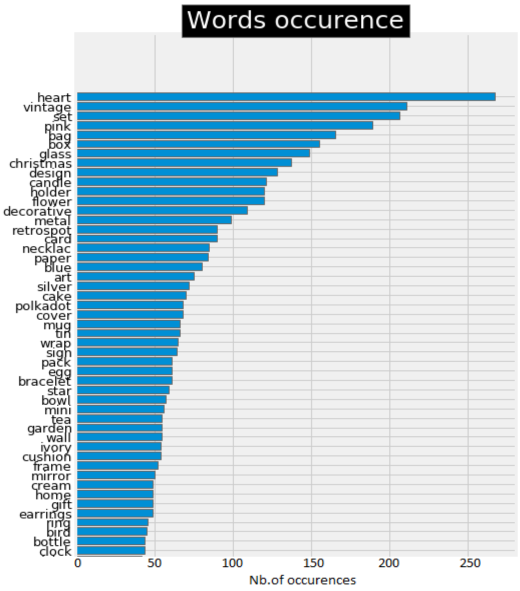
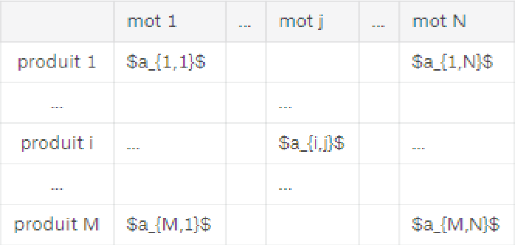


Figure 4. Count of each word in the product description

Figure 4 shows the number of occurrences of each word in the dataset. With the help of these keywords, we could now group the products. Firstly, the products were grouped into an m*n matrix: where the a{m,n} coefficient is 1 if the description of the product m contains the word n, and 0 otherwise. Figure 5 is a matrix showing how the keywords are mapped with the products.

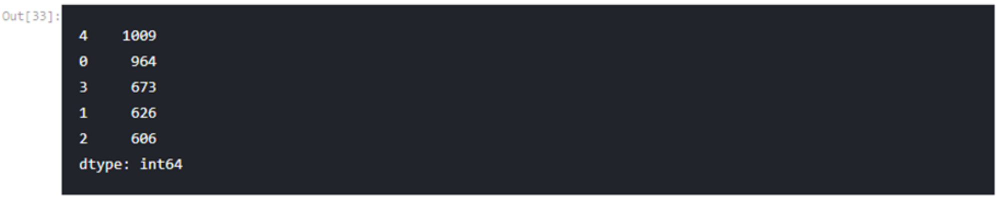


	mot 1	...	mot j	...	mot N
produit 1	$a_{1,1}$				$a_{1,N}$
...			...		
produit i	...		$a_{i,j}$		...
...			...		
produit M	$a_{M,1}$				$a_{M,N}$

Figure 5. A matrix depicting how keywords are mapped with respective products

3.3. Cluster products

In this section, products of similar kind were grouped into respective classes. For grouping of these products, we used Kmeans technique, where it used Euclidean Distance to calculate the distance and group the products accordingly. While Clustering, it was found that when the number of clusters had gone beyond 5, the number of elements in each cluster became very low. Therefore, it was decided to segregate them into 5 clusters. Figure 6 shows number of elements in each cluster.



```
Out[33]:
4    1009
0     964
3     673
1     626
2     606
dtype: int64
```

Figure 6. Number of elements in each Cluster

We noticed that each cluster contained objects that could be associated with terms which had meaning. Figure 7 shows word cloud with various elements in each cluster.

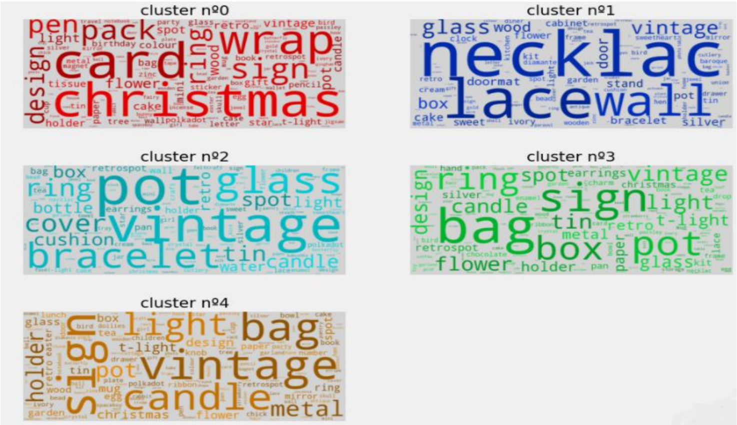


Figure 7. A word cloud of various elements in different clusters

3.4. Principal component analysis

In order to check the uniqueness of elements in the cluster. a PCA was performed to understand its distinctness. Figure 8 shows how elements are scattered within their groups. Figure 9 shows the products clustered after PCA.

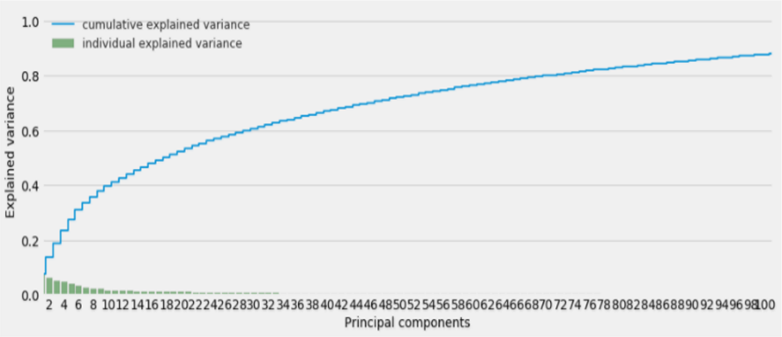


Figure 8. Uniqueness of the elements with the cluster

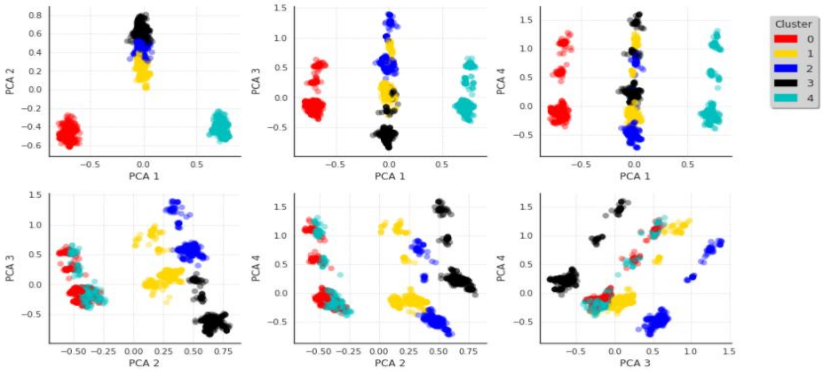
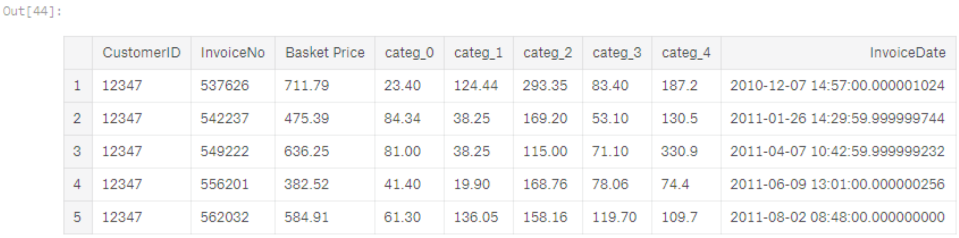


Figure 9. Cluster of various products

3.5. Customer categories

We needed to first group the various products into 5 clusters. For analysis, this information was added to the dataset in the form of cat_prod, where each cluster was denoted. Next, Cat_N variables were created, which indicated the amount spent in each product category. By doing this, we had all the data that was required in one Data frame or table. Figure 10 shows how each customer's investment in different categories.



Out[44]:

	CustomerID	InvoiceNo	Basket Price	categ_0	categ_1	categ_2	categ_3	categ_4	InvoiceDate
1	12347	537626	711.79	23.40	124.44	293.35	83.40	187.2	2010-12-07 14:57:00.000001024
2	12347	542237	475.39	84.34	38.25	169.20	53.10	130.5	2011-01-26 14:29:59.999999744
3	12347	549222	636.25	81.00	38.25	115.00	71.10	330.9	2011-04-07 10:42:59.999999232
4	12347	556201	382.52	41.40	19.90	168.76	78.06	74.4	2011-06-09 13:01:00.000000256
5	12347	562032	584.91	61.30	136.05	158.16	119.70	109.7	2011-08-02 08:48:00.000000000

Figure 10. A sample showing how each customer's investment varies in various categories

Having done all this, we could now find the other variables like max, min, mean and last_purchase for every user which helped in identifying an individual user who had made just one single purchase. One of the objective was to target these customers and try to retain them. Figure 11 shows Uniqueness of customers after clustering.

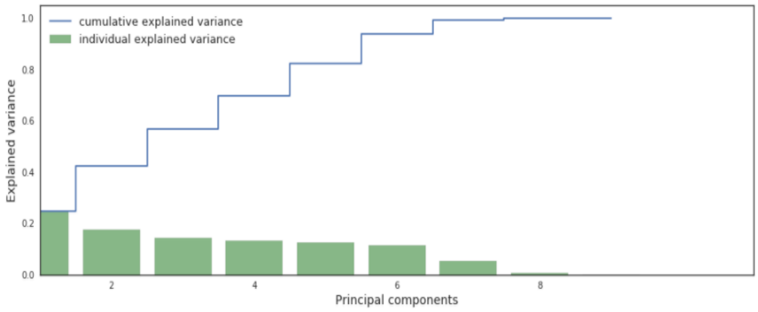


Figure 11. Uniqueness of customers when clustered

The number of clusters chosen were based on the silhouette score and the best score was obtained with 11 clusters. It is evident that customers in the clusters were distinct and the contents were averaged by selecting various clients. When checked after clustering the total number of customers were 3608.

4. EVALUATION AND RESULTS

4.1. Classification of customers

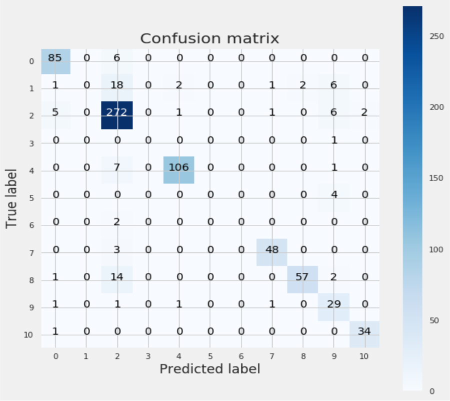
The objective here was to build a model that would classify customers into different customer categories based on the recognition of their earlier purchasing patterns. The main aim was to make sure this classification happened at the very first visit itself. For this, we took the help of the attributes of each class and classified based on these attributes. As the aim was to define a class at the first visit, only the content of the item was considered and variables such as frequency and all were ignored.

1) Support Vector Machine classifier*:* At first, SVC classifier was used to create an Instance of CLASS_FIT and then call grid_search (). Provided Parameters were:

a. Hyperparameters with optimal values.
b. Number of folds for cross validation.

When Tested the model gave a precision of 87.40%.

2) Confusion Matrix: Figure 12, is a confusion matrix that shows the predicted values are plotted on the x axis, the true values are plotted on the y axis and the elements plotted across the diagonal are rightly classified. For example, when seen in the above matrix for True Label ‘2’ and for Predicted Label ‘2’ the classified count of elements is 272.



	0	1	2	3	4	5	6	7	8	9	10
0	85	0	6	0	0	0	0	0	0	0	0
1	1	0	18	0	2	0	0	1	2	6	0
2	5	0	272	0	1	0	0	1	0	6	2
3	0	0	0	0	0	0	0	0	0	1	0
4	0	0	7	0	106	0	0	0	0	1	0
5	0	0	0	0	0	0	0	0	0	4	0
6	0	0	2	0	0	0	0	0	0	0	0
7	0	0	3	0	0	0	0	48	0	0	0
8	1	0	14	0	0	0	0	0	57	2	0
9	1	0	1	0	1	0	0	1	0	29	0
10	1	0	0	0	0	0	0	0	0	0	34

Figure 12. A confusion matrix showing the true values and predicted values

4.2. Learning curve

Figure 13 is a learning curve and learning curve is a typical way to test the quality of a fit. In particular, these type of curves allow to detect possible drawbacks in a model, linked for example to over-or under-fitting. This also shows to which extent the mode could benefit from a larger data sample.

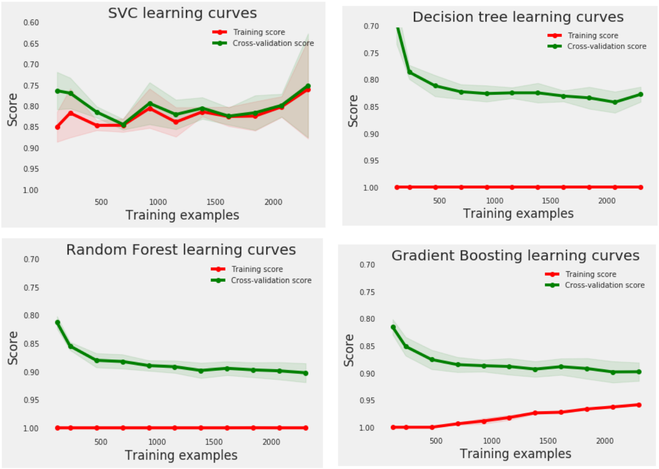
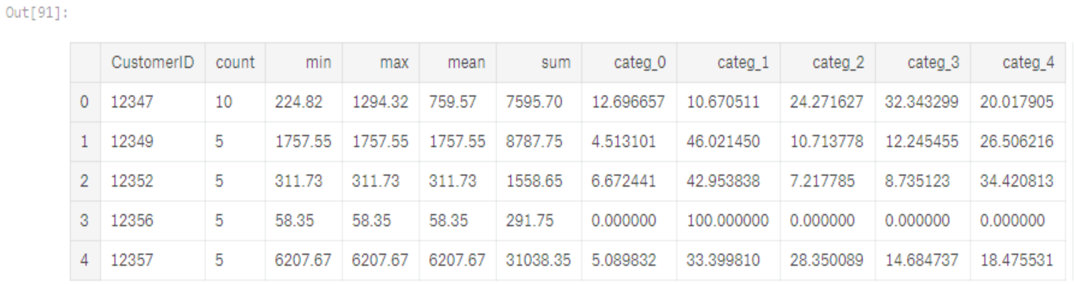


Figure 13. A set learning curves differentiating between various classifiers

Finally, the results of the different classifiers presented in the previous sections could be combined to improve the classification model. This could be achieved by selecting the customer category as the one indicated by the majority of classifiers. To do this, the 'Voting Classifier' method of the sklearn package was used. As a first step, the parameters of the various classifiers using the best previous parameters were adjusted. Then, a classifier was defined that merged the results of the various classifiers and trained them.

A few classifiers were trained to categorize customers. Until that point, the whole analysis was based on the data of the first 10 months. In this section, the model for the last two months of the dataset that had been stored in the set_test dataframe was tested.

In a first step, this data was regrouped and reformatted according to the same procedure as was used on the training set. However, to take into account the difference in time between the two datasets and weigh the variables count and sum to obtain equivalence with the training set, the data was corrected. Figure 14 shows the consolidated dataframe taken for testing.


Out[91]:

	CustomerID	count	min	max	mean	sum	categ_0	categ_1	categ_2	categ_3	categ_4
0	12347	10	224.82	1294.32	759.57	7595.70	12.696657	10.670511	24.271627	32.343299	20.017905
1	12349	5	1757.55	1757.55	1757.55	8787.75	4.513101	46.021450	10.713778	12.245455	26.506216
2	12352	5	311.73	311.73	311.73	1558.65	6.672441	42.953838	7.217785	8.735123	34.420813
3	12356	5	58.35	58.35	58.35	291.75	0.000000	100.000000	0.000000	0.000000	0.000000
4	12357	5	6207.67	6207.67	6207.67	31038.35	5.089832	33.399810	28.350089	14.684737	18.475531

Figure 14. A sample data taken for testing

Then, the dataframe was converted into a matrix and only variables that define the category to which consumers belong were retained. At this level, the method of normalization used on the training set was recalled.

Each line in this matrix contained a consumer's buying habits. At this stage, it was a question of using these habits in order to define a category to which this consumer belongs. Figure 15 shows the precision value of our model when tested with various classifiers.

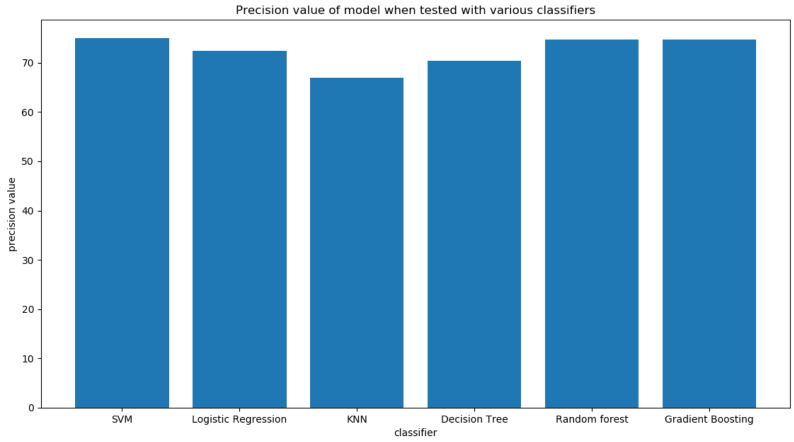


Figure 15. Precision value of model when tested with various classifiers

5. CONCLUSION

The work described in this paper was based on a database providing details on purchases made on an E-commerce platform over a period of one year. Each entry in the dataset described the purchase of a product, by a particular customer and at a given date. Given the available information, a model was developed that allowed the prediction of the type of purchase that a customer would make as well as the number of visits that he/she would make during a year, right from its first visit to the E-commerce site.

Finally, the quality of these predictions of the different classifiers were tested over the last two months of the dataset. The data was then processed in two steps: first, all the data was considered (over the 2 months) to define the category to which each client belongs, and then, the classifier predictions were compared with this category assignment. It was found that 75% of clients were awarded the right classes. The performance of the classifier therefore seemed correct given the potential shortcomings of the current model. In particular, a bias that had not been dealt with, were the concerns on the seasonality of purchases and the fact that purchasing habits would potentially depend on the time of year (for example, festival times like Christmas, Diwali etc.). In practice, this seasonal effect may cause the categories defined over a 10-month period to be quite different from those extrapolated from the last two months. In order to correct such bias, it would be beneficial to have data that would cover a longer period of time.

REFERENCES

[1] O. Etzioni, “The world-wide web: Quagmire or gold mine”, *Communications of the ACM*, vol. 39, no. 11, 1996, pp. 65-68.

[2] M. Eirinki, M. Vazirgiannis, “Web mining for web personalization”, *ACM Transactions on Internet Technology*, vol. 3, no. 1, 2003, pp. 1-27.

[3] M. Henzinger, “Link analysis in web information retrieval”, Bulletin of the technical committee on data engineering”, IEEE Computer Society, vol. 23, 2000, pp. 3-9.

[4] D. Shen, Y. Cong, J.-T. Sun, Y.-C. Lu, “Studies on Chinese web page classification”, in: *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics*, vol. 1, 2003, pp. 23-27.

[5] P. Vahdani Amoli and O. Sojoodi Sh., “Scientific Documents Clustering Based on Text Summarization”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 4, pp.782-787, 2015.

[6] H. Mannila, H. Toivonen, A.I. Verkamo, “Discovering frequent episodes in sequences”, in: *Proceedings of the First International Conference on Knowledge and Data Mining*, 1995, pp. 210-215.

[7] Ravi kumar V., and K. Raghuveer, “Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation”, *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 2, no. 1, pp. 27-35, 2013.

[8] R. Burke, “Hybrid recommender systems: survey and experiments”, *User Modelling and User-Adapted Interaction*, vol. 12, no. 4, 2002, pp. 331-370.

[9] H. Dai, B. Mobasher, “A road map to more effective web personalization: Integrating domain knowledge with web usage mining”, in: *International Conference on Internet Computing*, 2003, pp. 58-64.

[10] W. Lin, S.A. Alvarez, C. Ruiz, “Collaborative recommendation via adaptive association rule mining”, in: *WEBKDD2000 – Web Mining for E-Commerce – Challenges and Opportunities, Second International Workshop*, Boston, MA, USA, 2000.

[11] O.R. Zaiane, M. Xin, J. Han, “Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs”, in: Advances in Digital Libraries, Santa Barbara, CA, USA, 1998, pp. 19-29.

[12] B. Zhou, S.C. Hui, K. Chang, “An intelligent recommender system using sequential web access patterns”, in: *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems*, Singapore, 2004, pp. 1-3.

[13] H. Ishikawa, T. Nakajima, T. Mizuhara, S. Yokoyama, J. Nakayama, M. Ohta, K. Katayama, “An intelligent web recommendation system: A web usage mining approach”, in: *ISMIS*, 2002, pp. 342-350.

[14] R. Meteren, M. Someren, “Using content-based filtering for recommendation”, in: *Proceedings of MLnet/ECML2000 Workshop*, Barcelona, Spain, 30 May 2000.

[15] J. Li, O.R. Zaı¨ane, “combining usage, content, and structure data to improve web site recommendation”, in: EC-Web, 2004, pp. 305– 315.

[16] W. Cohen, A. McCallum, D. Quass, “Learning to understand the web”, *IEEE Data Engineering Bulletin*, vol. 23, 2000, pp. 17-24.

[17] S.K. Madria, S.S. Bhowmick, W.K. Ng, E.P. Lim, “Research issues in Web data mining”, in: *Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery (DaWaK’99)*, 1999, pp. 303-312.

[18] V. Kesˇelj, F. Peng, N. Cercone, C. Thomas, “N-gram-based author profiles for authorship attribution”, in: *Proceedings of the Conference Pacific Association for Computational Linguistics*, Nova Scotia, Canada, 2003.

[19] Das, S., Mathew, M. and Vijayaraghavan, P. (2017). “An Efficient Approach for Finding near Duplicate Web pages using Minimum Weight Overlapping Method”. *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 1, pp. 187-194, 2011.

[20] Zhang, L., Yang, S. and Zhang, M. (2018). “E-commerce Website Recommender System Based on Dissimilarity and Association Rule”. *Indonesian Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, pp. 353-360, 2014.