**Using Machine Learning Techniques to Predict Potential Customers and Their Purchasing Items**

Short Title: Machine Learning to Predict Customers and Purchases

Zihao Wang

Long Island University, zihao.wang3@my.liu.edu

Sevim Oztimurlenk

Long Island University, sevim.oztimurlenk@my.liu.edu

Ling Zhu

Long Island University, ling.zhu@liu.edu

The need has increased in B2C e-commerce for a more accurate recommendation system that targets customers. The system should be able to provide customers targeted purchase information based on their online behavior. A powerful and meaningful recommendation system can offer enormous value to customers and e-commerce providers. Customers can locate an item of their interest more easily and quickly. On the other hand, an e-commerce provider can achieve a higher customer conversion rate and thus increase sales. To this end, this research developed a two-step machine-learning-based recommendation prototype with two data mining goals: to identify potential customers with purchase intention and to predict the products they may purchase based on an e-commerce log (i.e., a sequential record of products click-through events). In the first module, we used both the K-nearest neighbor (KNN) model and the multi-layer perceptron (MLP) neural network to classify customers with and without purchase intention. In the second module, we developed a first-order Markov chain model to predict products that those potential customers will purchase. We then tested the prototype with the dataset provided in the RecSys 2015 challenge. The results show that in the first module the KNN has a higher true positive rate than the MLP neural network. In the second module, the Markov model also performs well to predict the products purchased by a customer. A recommendation system was developed based on the KNN and first-order Markov chain model. The findings show that the system has a great potential to be developed into a machine-learning-based recommendation system for e-commerce usage. To our knowledge, this system is the first of its kind to combine a classification model with time series analysis.

CCS CONCEPTS • Recommendation system • Neural network • K-nearest neighbor • Markov chain

**Additional Keywords and Phrases:** e-commerce, purchase intention, customer preference, online behavior, targeted marketing.

1. Introduction

Online service personalization has become growing research and commercial interest [5]. Both customers and businesses benefit from personalization systems [1]. For example, through a recommendation system, customers can locate an item of their interest quickly and easily. E-commerce service providers can achieve a higher customer conversion rate and thus increase sales. To find customers with purchase intention, Romov and Sokolov proposed seven features [5]. Pálovics et al. developed a linear model to find potential customers [4]. Chen et al. used the multi-Layer Perceptron (MLP) neural network with a single hidden layer to find potential customers [2]. It is a big challenge, however, as to how to precisely know customers’ needs.

This study developed a recommendation system to identify potential customers who have purchase intention and what products they may purchase. We proposed two functions in the system: 1) identify customers who have an intention to purchase items online in a given session; 2) if a customer does have the intention, identify what products are most likely to be purchased by the customer.

The paper is presented in the following sections: in the section of demo overview, we described the data collection, methodology, and analytical techniques that were developed for the study. In the section of the conclusion, we discussed the results of our predictive model and the future applications.

1. Demo overview
   1. Data collection

This study used e-commerce search logs from RecSys 2015 challenge [1]. It contains two datasets: clicked events and bought events. The columns in the clicked events dataset are session ID, timestamp, item ID, and category. The bought events dataset contains the products bought with session ID, timestamp, item ID, price, quantity. The session id in the bought event dataset is linked to the same column in the clicked event dataset; that is, if a customer had bought an item, her click behavior can be found in the clicked event.

* 1. Mining potential customers with purchase intention

The first step is finding effective features that can be used for training machine learning classifiers. We implemented in this study two machine learning models: MLP neural network and KNN classifier. The feature vector for training is a vector with a length of seven with features of 1) the day of the starting week, 2) the hour of the starting day, 3) the day of finishing week, 4) the hour of the finishing week, 5) maximum dwell time in the session, 6) numbers of clicks on the session, and 7) numbers of items on the session.

* + 1. Neural network method

We implemented the MLP neural network based on the structure of DeepFM [3]. A neural network with 3 hidden layers is constructed with a tanh activation function in each hidden layer. The first and second hidden layers had 32 neurons and the third hidden layer had 16 neurons. The output layer had 2 neurons with a SoftMax activation function and cross-entropy loss function. The neural network was trained 100 epochs with a batch size of 50,000 each. After the training process, the best result appeared on the initial state with the accuracy of 0.89, and the true positive rate of 0.03 for correctly predicting customers with purchase intention on the validation dataset. With the testing dataset, the respondent accuracy and true positive rate are 0.89 and 0.02.

* + 1. K-nearest neighbor method

The K-nearest neighbor (KNN) algorithm is one of the machine learning algorithms for classification. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. The k value is determined by trying each odd number from 1 to 20. The best k value is 1 with the highest accuracy of 0.9 and the highest true positive rate is 0.12 in the validation dataset. The accuracy and true positive rate for the testing set are 0.9 and 0.13.

* 1. Predicting products likely to be purchased by customers
     1. Sequential recommendation method based on Markov chain

We use the first-order Markov chain to predict products that customers may purchase based on their search sequences. There are four steps in this part. The first step is selecting the unique selected product sequence (USP) vector for each customer. If the size of the USP is less than 5, then all items in the USP are seen as the prediction items. Otherwise, the second step is calculating the item probability distribution and the transition matrix for the current state based on the current customer’s unique selected products. Third, the probability distribution of purchasing products is computed based on the inner product of the current probability distribution and the transition matrix. Lastly, the searched products are sorted by the decreasing order of the probability distribution of purchasing products, and the first half of items are treated as the final prediction.

The process for selecting the USP sequence for each customer is described in Algorithm 1.

|  |
| --- |
| ALGORITHM 1: The process of selecting the unique selected products |
| for each customer:  queue=empty  for item in searched sequence based on time series:  if item not in queue:  enqueue(item) |

A file that recorded all customers and their unique selected products is generated from the clicked events dataset. An example of the file is shown in .

Table 1. An example of customers and their unique selected products

|  |  |
| --- | --- |
| Session ID | Unique selected products |
| 1 | 214536502, 214536500, 214536506, 214577561 |
| 2 | 214662742, 214825110, 214757390, 214757407, 214551617 |
| 3 | 214716935, 214774687, 214832672 |

To get the probability distribution of the current state, the actual searched products (ASP) vector is summarized for each customer. The process is described in Algorithm 2.

|  |
| --- |
| ALGORITHM 2: The process to get the actual searched products |
| for each customer:  queue=empty  for item in searched sequence based on time series:  enqueue(item) |

A file that recorded all customers and their actual selected products is generated from the clicked events dataset. An example of the file is shown in . The customer with session ID 2 searched product “214662742” appears once in but twice in .

Table 2: An example of customers and their actual selected products

|  |  |
| --- | --- |
| Session ID | Actual selected products |
| 1 | 214536502, 214536500, 214536506, 214577561 |
| 2 | 214662742, 214662742, 214825110, 214757390, 214757407, 214551617 |
| 3 | 214716935, 214774687, 214832672 |

In current state probability (CSP) distribution is defined as:

()

()

The transition matrix is the size of N × N matrix, where N is the size of USP set for each customer. Rows and columns are the element order in the USP set. The transition probability is defined as:

()

()

where means item x appeared before item y in the USP vector. X is a set of given products and T is a set of transactions.

The process of the first order Markov chain is described in Algorithm 3.

|  |
| --- |
| ALGORITHM 3: The process of the first order Markov chain |
| all\_customers\_USP=Algorithm1  all\_customers\_ASP=Algorithm 2  for current\_customer\_USP in All\_customers\_USP:  if size(current\_customer\_USP)<5:  prediction= current\_customer\_USP  else:  get transistion matrix for current\_customer\_USP  get initial item distribution for current\_customer\_USP  compute the buying probability distribution for current\_customer\_USP  sort current\_customer\_USP by decrease order  prediction=first half items from the sorted current\_customer\_USP  return prediction |

For each customer, the actual bought products (ABP) vector is generated from the bought products dataset by Algorithm 1. The predicted purchasing products (PBP) vector is computed by the first-order Markov chain model. The accuracy for each customer is defined as:

()

The average accuracy among the 1,000 customers is 0.97. Most customers have an accuracy between 0.65 and 1.0. A few customers have an accuracy between 0.35 and 0.65. The result indicates this model can apply to real-world situations.

1. Conclusion

This study proposed a two-step model for mining potential customers and predicting their interesting products. In the first step, the MLP neural network and KNN classifier can recognize customers with purchase intention with an accuracy of 90%. The KNN classifier has a better performance than the MLP neural network due to the higher true positive rate. In the second step, the first-order Markov chain shows an excellent performance in sequential analysis with an accuracy of 97%. A recommendation system that utilizes the two-step model demonstrated in this study is developed and can be applied to real e-commerce cases.

REFERENCES

[1] David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. 2015. RecSys Challenge 2015 and the YOOCHOOSE Dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems*, ACM, Vienna Austria, 357–358. DOI:https://doi.org/10.1145/2792838.2798723

[2] Wenliang Chen, Zhenghua Li, and Min Zhang. 2015. Linear and Non-Linear Models for Purchase Prediction. In *Proceedings of the 2015 International ACM Recommender Systems Challenge*, ACM, Vienna Austria, 1–4. DOI:https://doi.org/10.1145/2813448.2813518

[3] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. *arXiv:1703.04247 [cs]* (March 2017). Retrieved March 1, 2021 from http://arxiv.org/abs/1703.04247

[4] Róbert Pálovics, Péter Szalai, Levente Kocsis, Adrienn Szabó, Erzsébet Frigó, Júlia Pap, Zsófia K. Nyikes, and András A. Benczúr. 2015. Solving RecSys Challenge 2015 by Linear Models, Gradient Boosted Trees and Metric Optimization. In *Proceedings of the 2015 International ACM Recommender Systems Challenge*, ACM, Vienna Austria, 1–4. DOI:https://doi.org/10.1145/2813448.2813513

[5] Peter Romov and Evgeny Sokolov. 2015. RecSys Challenge 2015: ensemble learning with categorical features. In *Proceedings of the 2015 International ACM Recommender Systems Challenge*, ACM, Vienna Austria, 1–4. DOI:https://doi.org/10.1145/2813448.2813510