Conference Paper No: PF-01

Product recommendation model for supermarket industry based on machine learning algorithms

U. B. P. Shamika* and A. D. De Silva

Department of Statistics and Computer Science, Faculty of Science, University of Kelaniya, Sri Lanka
shamikapawani137@gmail.com*

Abstract

Recommendation activities automatically display products or content that might interest customers based on previous user activity. Recommendations help customers directly to identify the relevant items that they might otherwise not know. The product recommendation model determines which products are suggested to a consumer, depending on that consumer's shopping history. The main objective of this research was to develop a product recommendation model by considering the shopping history of consumers. The supermarket data used in the study contain customer details, transaction details, and product details. The product recommendation model was built using three machine learning techniques such as the Long Short-Term Memory algorithm, Convolutional Neural Network algorithm, and Collaborative Filtering algorithm. The obtained accuracies of the proposed model with respect to Collaborative Filtering, Long Short-Term Memory and Convolutional Neural Networks are 78%, 54% and 56% respectively. According to the accuracy values the Collaborative Filtering algorithm is more suitable to build the product recommendation model than the Long Short-Term Memory algorithm or Convolutional Neural network.

Keywords: Collaborative Filtering, Convolutional Neural Network, Long Short-Term memory, Machine learning, Shopping history

Introduction

Most people fulfill their daily needs from supermarkets. A grocery list is an integral part of the shopping experience of many consumers (Tahiri§∗, Bogdan, & Makarenkov, 2019). Many online stores use product recommendation systems for their consumers such as Amazon online shopping store by considering product similarity. But according to the conducted literature review, it does not exist any product recommendation system for Sri Lankan supermarket industry based on consumer shopping history. The main objective of the research is to recommend shopping list to consumer by considering his/her past shopping behavior. This research developed the product recommendation model using different machine learning techniques and then identified the most suitable machine learning method to develop a product recommendation model for supermarkets. Then using the best machine learning algorithm, identifying the most frequently bought products of each consumer.

There is an existing product recommendation model that built using combination of the Long Short-Term Memory (LSTM) algorithm and Convolutional Neural Network (CNN) algorithm. (Tahiri§∗, Bogdan, & Makarenkov, 2019), introduced a product recommendation model using a combination of LSTM and CNN. But the accuracy of that model was 49%. This research develops a product recommendation model using the

LSTM algorithm and CNN machine learning algorithm separately. The Amazon company currently uses a product recommendation system for its consumers. Amazon examines for each of the user's purchased and rated items where the algorithm attempts to find similar items, then aggregates the similar items and recommend them to the users (Linden, Smith, & York, 2003). (Linden, Smith, & York, 2003) built the recommendation model using the Collaborative Filtering (CF) algorithm. Further they suggested to apply these recommendation algorithms for targeted marketing, both online and offline in the future.

Methodology

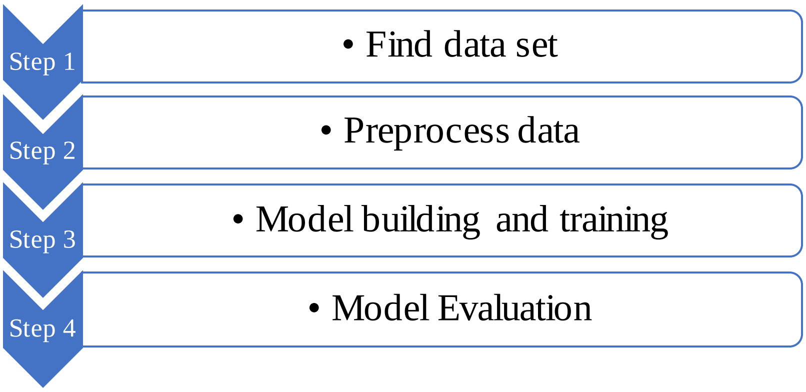


Figure 1. *Method that is been used in the study.*

Methods

Machine learning is the study of mimic the human brain into a machine. Machine learning facilitates computers in building models from sample data to automate decision-making processes based on data inputs. This research used three machine learning techniques to build the recommendation model separately.

Long Short-Term Memory Algorithm

The LSTM is a type of Recurrent Neural Network that achieves a state-of-the-art result on challenging prediction problems. Sequence prediction problems have been around for a long time with the recent breakthroughs that have been happening in data science, is found that for almost all these sequence prediction problems, LSTM networks (Brownlee, 2017). The LSTM model has three regulators(gates) as input gate, forget gate, output gate. The Input gate controls the extent to which a new value flows into the cell and the forget gate controls the extent to which a value remains in the cell. The output gate controls the extent to which the value in the cell use to compute the output activation of the LSTM unit. Recurrent Neural Network is suitable to develop a recommendation model as that model gets current input as past output. But the Recurrent Neural Network model cannot keep the long duration of memory but, the LSTM can keep the long track of memory. As a result, develop a recommendation model using LSTM Algorithm.

Convolutional Neural Network Algorithm

The CNN is a deep neural network designed for processing structured arrays of data. CNN are widely used in computer vision. While focused on feedforward networks, CNN are more often utilized for classification and computer vision tasks. They are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network (Education, 2020).

Collaborative Filtering Algorithm

The CF is a way that the recommendation systems filter information by using the preferences of other people. CF algorithm on the other hand, doesn't need anything else except users' historical preference on a set of items. Because it's based on historical data, the core assumption here is that the users who have agreed in the past tend to also agree in the future. In terms of user preference, it usually expressed by two categories. Explicit Rating is a rate given by a user to an item on a sliding scale, like five stars for Titanic. This is the most direct feedback from users to show how much they like an item. Implicit Rating suggests user's preference indirectly, such as page views, clicks, purchase records, whether to listen to a music track, and so on (Luo, 2018).

Materials

This research used Python 3 as the programming language to preprocess the data set and build the model and used the Jupiter framework to develop the model.

Data set

Download the online data set from (vasudeva, 2019) to develop the model. Data Contain in three tables named customers, transactions, and item data. Customers table contains Customer ID, the age range of customers, marital state, home type, family size, number of children, and income of customers. The Transaction table contains transaction date, customer ID, item ID, quantity, selling price, and discounts. The Item data table contains item id, brand, category, and item name. There are 761 customers on the customer details spreadsheet and 74067 products in the product details spreadsheet.

Data Preprocessing

Data preprocessing refers to using the processed data to build the recommendation machine learning model. An online data set is used to analyze the shopping history of customers individually. After cleaning the data set with eliminating missing and null values the remaining 482 customers' data was considered. Then the remaining customer details were merged with their transaction details to prepare the required dataset. There remain 89401 data and considered only 30000 transaction details to build the product recommendation model. The dataset was split into 80% of data as training data and 20% of data as testing data. The product recommendation model using the training data set and testing data set was used to evaluate the model.

The Neural Network algorithm gets only numeric values and must provide features with data set. For example, the customer features as in non-numeric data types, and in order to build the Neural Network machine learning models it requires to change those non-numerical values into numerical values. Therefore, python 3 was used to implement the above-mentioned task as follows.

Table 1. *Assigned values for age ranges*

Age ranges	Assigned value
18-25	0
26-35	1
36-45	2
46-55	3
56-70	4
70+	5

Table 2. *Assigned values to marital state*

Marital State	Assigned value
Single	0
Married	1

Collaborative Filtering algorithm considered Customer ID, product ID, and how frequently bought that product as only inputs. Therefore, the data set was updated by getting the frequency of products to each consumer using python script.

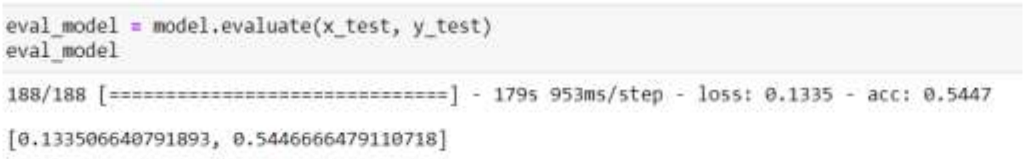
Build the model and evaluate to take the output

Build Product recommendation model using three machine learning techniques LSTM, CNN, and CF with 10 epochs and evaluate the models using testing data set.

Results and Discussion

Model evaluation outcome for LSTM

By adjusting validation split to 0.2, it was able to employ 10 epochs with the default batch size of 50. For every validation loss inside the model, the units were set to 20 in the early callback method. Figure 1 depicts the validation of LSTM. The validation accuracy of this model was 54%.



```
eval_model = model.evaluate(x_test, y_test)
eval_model

188/188 [==============================] - 179s 953ms/step - loss: 0.1335 - acc: 0.5447

[0.133506640791893, 0.5446666479110718]
```

Figure 2. *Model test of LSTM*

Figure 3 depicts the connection between the accuracy of the training set and the validation set for each epoch. The graph shows that the accuracy of the training set has risen and after gone down and in the same value, and validation sets has in the same value with each epoch. The greatest accuracy of the model was reached each epoch in our investigation.

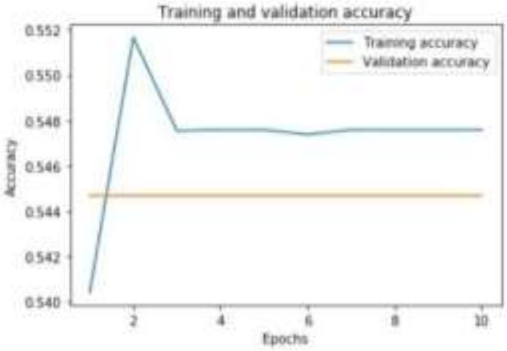
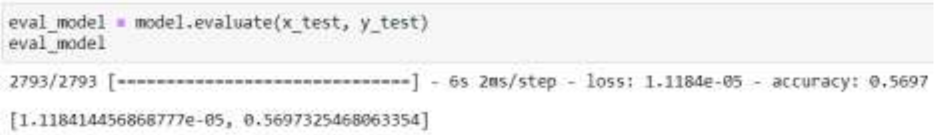


Figure 3. *Model test of LSTM*

Model evaluation outcome for CNN

By adjusting validation split to 0.2, we were able to employ 10 epochs with the default verbose of 1. For every validation loss inside the model, the dense were set to 100,50, and 2 in the early callback method. Figure 2 depicts the evaluation of CNN. The validation accuracy of the model was 56%.



```
eval_model = model.evaluate(x_test, y_test)
eval_model
```

```
2793/2793 [==============================] - 6s 2ms/step - loss: 1.1184e-05 - accuracy: 0.5697

[1.118414456868777e-05, 0.5697325468063354]
```

Figure 4. *Model test for CNN*

Figure 5 depicts the connection between the accuracy of the training set and the validation set for each epoch. The graph shows that the accuracy of the training set is in the same value with each epoch and validation set has risen and after that has gone down and then risen value. It is not always necessary to consider the validation learning curve's last data point with the best accuracy of the model. The greatest accuracy of the model was reached each epoch in the investigation.

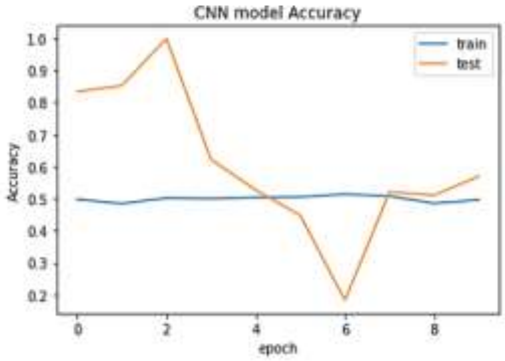
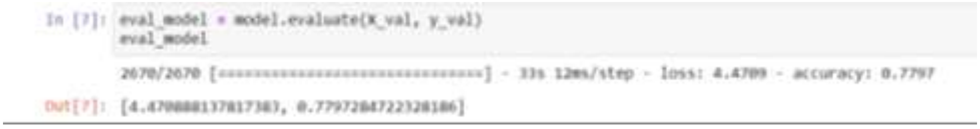


Figure 5. *Accuracy graph, training vs validation in CNN*

Model evaluation outcome for CF

By adjusting validation split to 0.2, it was able to employ 10 epochs Figure 3 depicts the evaluation of CF. The validation accuracy of CF is 77%.



```
In [7]: eval_model = model.evaluate(X_val, y_val)
        eval_model

        2670/2670 [==============================] - 33s 12ms/step - loss: 4.4709 - accuracy: 0.7797

Out[7]: [4.470888137817383, 0.7797284722328186]
```

Figure 6. *Validation of CF*

Figure 7 depicts the connection between the accuracy of the training set and the validation set for each epoch. The graph shows that the accuracy of the training set has risen and after that in the same value, and validation set is in the same value with each epoch. The greatest accuracy of the model was reached each epoch in our investigation.

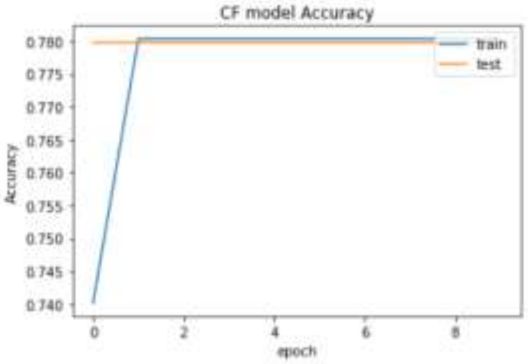


Figure 7. *Accuracy graph, training vs validation in CF*

Conclusion

The main objective of the study is to provide a recommended product list to Sri Lankan supermarket consumers considering their shopping behavior. This research developed product recommendation model used in machine learning Algorithms. In addition, compared different machine learning techniques and identified the most suitable machine learning algorithm to build the product recommendation model. Since the objective was to develop the model focusing on the Sri Lankan community, it was required to collect data from the Sri Lankan supermarkets. But unfortunately, due to the COVID 19 pandemic situation and the stipulated policies in the supermarkets, they rejected to provide the required data. As a result, used an online dataset as a dummy dataset to build the product recommendation model. The product recommendation model developed using three machine learning techniques, LSTM, CNN, and CF algorithms using online data set. The validation accuracy of the LSTM model is 54%, the accuracy of the CNN model is 56%, the accuracy of the CF model is 78%. According to the validation values we may suggest that CF as the most suitable machine learning algorithm to develop a product recommendation model in Sri Lanka. (Tahiri§∗, Bogdan, & Makarenkov, 2019) built product recommendation model by combining CNN and LSTM, but the accuracy was 49%. This research conclude that the CNN product recommendation model was more accurate than combination of CNN and LSTM recommendation model Future work suggests developing a product recommendation model for the supermarket by combining

CF with another machine learning technique and train the CF model used in Sri Lankan supermarket transaction dataset.

References

Brownlee, J. (2017, 05 25). *A Gentle Introduction to Long Short-Term Memory Networks by the Experts*. (Machine Learning Mastery) Retrieved 06 10, 2020, from https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

Education, I. c. (2020, 10 20). *IBM*. (IBM) Retrieved 08 20, 2021, from https://www.ibm.com/cloud/learn/convolutional-neural-networks

Linden, G., Smith, B., & York, J. (2003). *Amazon.com Recommendations Item-to-Item Collaborative Filtering.* Washington: IEEE INTERNET COMPUTING.

Luo, S. (2018, 12 10). *Introduction to Recommender System*. (towards datascience) Retrieved 03 11, 2021, from https://towardsdatascience.com/intro-to-recommender-system-collaborative-filtering-64a238194a26

Tahiri§∗, N., B. M., & Makarenkov, V. (2019). *An intelligent shopping list based on the application of partitioning and machine learning algorithms.* (SCIPY: PROC. OF THE 18th PYTHON IN SCIENCE CONF.

vasudeva. (2019, 11 15). *Predicting Coupon Redemption*. Retrieved from kaggle: https://www.kaggle.com/vasudeva009/predicting-coupon-redemption