Project Title: Using Market Basket Analysis to Study Purchase Patterns

Market Basket Analysis Report

Author: Robert Ferdinand

Email: UOFLOUISIANA@GMAIL.COM

Project Title: Using Market Basket Analysis to Study Purchase Patterns

Table of Contents

Project Title: Using Market Basket Analysis to Study Purchase Patterns

Section 1: Introduction of Problem

Data on purchases made by customers at an undisclosed food retail store is obtained from the Kaggle website. The data consists of features such as customer ID, date of purchase and items purchased. Based on this data we attempt to predict purchasing trends for the various customers, based on their respective purchases at the store. Using association functions from a PYTHON software package MLXTEND, we utilize machine learning to help us understand what items customers are likely to purchase in the future, given their current purchases. Having this knowledge would help the retail store target customers efficiently by way of advertising for specific products instead of just advertising everything to everyone. Such advertising can take the form of email, flyers and text messages.

To be able to make such a prediction, one would need to analyze the data, to be able to "learn" from it. Hence the term "Machine Learning". This can be done using Machine Learning (ML) techniques which is a Data Science tool. The PYTHON programming language, MLXTEND and its awesome functions will be fully utilized to accomplish the task at hand.

The machine learning will be carried out on a PYTHON Jupyter environment on the IBM Watson Cloud.

Section 2: The Data

The customer data is acquired from the open-source website Kaggle.com. This website has several sets of highly interesting and accurate data sets that have important real-world, industrial applications. The customer data contains information including the collective items purchased by store customers, at visits to the store, over a period of time. In total there are about 40,000 customer purchase records. The features of the data include parameters such as customer ID, date of purchase and items purchased.

In order to be able to predict future customer purchases based on current purchases we need to use machine learning. The PYTHON programming language will be used to implement this. Several software packages in PYTHON can be used to determine the associations among the items purchased as well as the prediction of future purchases.

The data will need to be cleaned of any missing values and then removed of any duplicate records or rows. Finally we will need to make sure that all of the data makes sense. For example the items purchased cannot be just numerical in value. They need to textually describe the products purchased by the customer.

We note here that the data is in a comma-separated version (CSV) file that will need to be read into a technology platform that can be ultimately used for machine learning. The platform we will use for computation is the IBM Watson Cloud.

Project Title: Using Market Basket Analysis to Study Purchase Patterns

Section 3: The Methodology

In this section we use the data set described in the section above, from the Kaggle.com website. Then we proceed with learning from this data set as follows:

Step I: Data Cleaning or Wrangling

(a) Import data set from Kaggle.com and read the .CSV file as a pandas dataframe.

(b) Install PANDAS machine learning package MLXTEND needed for the Market Basket Analysis.

(c) From MLXTEND import the following functions:

- from mlxtend.preprocessing import TransactionEncoder
- from mlxtend.frequent_patterns import apriori
- from mlxtend.frequent_patterns import association_rules

(d) Delete any duplicate rows in the dataframe to remove redundancy.

(e) Rename the dataframe columns as "ID" for customer ID, "DATE" for date of purchase and "PURCHASE" for purchases. This makes the column names easy to work with.

(f) Remove any rows that have missing column information.

(g) Strip the "PURCHASE" column of any leading and trailing white spaces to facilitate ease of using the PYTHON functions imported earlier.

(h) Drop the "DATE" column since the model is only interested in customers identified by their IDs in the "ID" column and purchases titled the "PURCHASE" column.

(i) Ensure that all columns are of data types that makes sense. For example the "PURCHASE" column should contain text that details the items purchased.

(j) Using the PYTHON *value_counts()* function, to obtain the frequency distribution for the "ID" and "PURCHASE" columns in the pandas dataframe

(k) As a results of (a)-(h) above, our data is reduced to about 38,000 rows and 2 columns. After cleaning the data, we have a very large sized data set and would need substantial computing resources for its processing. However, we are ensured that the data is fully cleaned and ready for processing.

Step II: Data Extraction, Transformation and Loading (ETL)

(a) From the cleaned data obtained from Step I, in pandas dataframe form, we proceed by grouping the data. To this end we group the data on the customer ID key "ID" and group together all customer purchases in the "PURCHASE" column for each "ID". This gives us a list of customers, via "ID", and all of their respective purchases over the entire data set. This is converted into a pandas dataframe.

(b) For each customer, listed via customer ID in the "ID" column in the dataframe from part (a), their respective purchases in the "PURCHASE" column is ONE-HOT ENCODED. The `TransactionEncoder` function from MLXTEND as well as the `transform` function from PYTHON is used to accomplish the ONE-HOT ENCODING. This gives us a large dataframe with Boolean (True/False) values for each item of purchase in the store being a column, for each customer ID. These appear as True if the customer purchased the item and False if they did not. This large dataframe has shape of about 4000 x 200.

(c) Our data is now ready in a dataframe format to which Market Basket Analysis can be applied, using the MXLTEND package from PYTHON. This is described in the next step.

Step IV: Modeling

This is the most FUN part of the project. We apply machine learning techniques to associate customer purchases using priority rules. Here we proceed as follows:

(a) The customer purchases in the dataframe obtained from Step III above are prioritized based on the frequencies of purchases of the items. The MLXTEND function `apriori` is used here with minimum frequency or probability of purchase as 0.01. This gives us a dataframe with frequencies of group purchases of items as purchased by customers' visits to the retail store.

(b) The MLXTEND function `association_rules` is applied to the dataframe from part (a) above using the `lift` metric. For theoretical purposes, `lift` is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values: target response divided by average response. The minimum threshold for the `lift` here is taken as 0.7. This creates a dataframe containing grouped item purchases, per customer visit to retail store. These items are paired as antecedents and consequents, where an antecedent would be items purchased in the past or currently that correspond or map to consequents which would be items purchased in the future. So, for example if antecedent X has consequent Y then that would

mean that if item X was purchased in the past or currently then item Y would be purchased by the same customer in the future.

(c) We can change the values of `lift` metric and even use a combination of the `lift` and `confidence` metrics to create a dataframe of antecedents and consequents. The latter metric is a ratio of correctly predicted results to correct results. This dataframe can then be used to find what products a customer is likely to purchase given they have made a certain purchase in the past.

Project Title: Using Market Basket Analysis to Study Purchase Patterns

Section 4: The Results

In this section we discuss results from Section 4 above. The analytical and graphical results can be found in the IPYTHON notebook on Github. URL: https://github.com/DataScienceEsq/PROJECT-4.

(a) Using the dataframe obtained from Step IV in Section 3 above, we can make certain predictions about what purchases a customer is likely to make in the future given they made certain purchases currently or in the past.

(b) For example, the table below shows data regarding what customers purchase in the future as a consequent when they purchase bottled water and whole milk as an antecedent:

Antecedent	Consequent	Confidence	Lift
Bottled water and Whole Milk	Other Vegetables	0.5	1.3
Bottled water and Whole Milk	Rolls/Buns	0.4	1.1
Bottled water and Whole Milk	Soda	0.4	1.1
Bottled water and Whole Milk	Yogurt	0.4	1.3

(c) Next, again for example, the table shows data regarding what customers might have purchased in the past as an antecedent when they currently purchased bottled water and whole milk as a consequent:

Antecedent	Consequent	Confidence	Lift
Brown Bread, Other Vegetables	Bottled water and Whole Milk	0.2	1.8
Curd, Other Vegetables	Bottled water and Whole Milk	0.2	2.0
Frankfurter, Other Vegetables	Bottled water and Whole Milk	0.2	1.8
Rolls/Buns, Yogurt, Other Vegetables	Bottled water and Whole Milk	0.2	1.8

Section 5: Discussion and Conclusion

Pandas DataFrames and Machine Learning Algorithms (stated earlier) are used extensively in this project. The implementation was carried out on a PYTHON notebook on IBM Watson Studio platform which proves to be an outstanding resource.

The two tables in Section 4 above show us the usefulness of Market Basket Analysis as a tool to predict future customer purchases based on their current purchase patterns. This can be used effectively by a retail enterprise to advertise particular goods to specific target audiences, via different media, with laser-like precision, as per predictions from the Market Basket Analysis based on current purchases. This will result in the optimization of advertising costs and sales for the retail merchandise.

Certainly, Market Basket Analysis can also be applied to a retail organization dealing with non-food items. Well-known companies are known to be using this Machine Learning technique. There is a famous story about a well-known retail chain that used Market Basket Analysis to make a stunning (yet True) prediction that can be read about in most predictive analytics texts.

Acknowledgments: The author would like to acknowledge the FREE computing resources provided by the IBM Watson Cloud platform.