Project 10: Market Basket Insights

CODING :

{

"cells": [

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"# Market Basket Analysis (MBA)\n",

"<br>\n",

"This repo contains the Market Basket Analysis (MBA) project as part of my data science portfolio. There are two parts in this project:\n",

"1. [Understanding MBA](#understanding-mba)\n",

"2. [Implementation in Python](#implementation-in-python)\n",

"\n",

"The first part - Understanding MBA is for beginners who are new to this technique. It explains the theory, applications and workings of MBA with a case study of a supermarket. The second part - Implementation in Python contains the Python code to implement this technique using public dataset from [Kaggle](https://www.kaggle.com). References are provided at the end of this project to give due credit to the authors of journal articles and resources that contributed to the completion of this project."

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"\*\*\*\n",

"# <a name=\"understanding-mba\">Understanding MBA</a> \n",

"Market basket analysis (MBA), also known as association-rule mining, is a useful method of discovering customer purchasing patterns by extracting associations or co-occurrences from stores' transactional databases (Chen et al., 2005). It is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in a supermarket and you buy a loaf of Bread, you are more likely to buy a packet of Butter at the same time than somebody who didn't buy the Bread. Another example, if you are buying a XiaoMi Power Bank in an online store, you are more likely to also buy a carrying case to go with the power bank. [Amazon](https://www.amazon.com) knows this well from the transaction data of its millions of customers and thus recommends a case to you as seen below:"

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"\n",

"<table><tr><td><img height=\"400\" width=\"800\" src=\"images/amazon.jpg\"></td></tr></table>\n",

"<p style=\"text-align: center\">Credit: Amazon</p>"

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"\n",

"The set of items a customer buys is known as an itemset, and MBA tries to identify relationships from the purchases of itemset. The output of MBA consists of a series of product association rules. From the transaction data extracted from the shopping carts of online retailers or the point of sales system of retail stores, we can use MBA to extract interesting association rules between products. For example, if customers buy product A they also tend to buy product B.\n",

"\n",

"Typically we can extract the relationship between products in the form of a rule, an example of association rule:\n",

"\n",

" IF {bread} THEN {butter}. \n",

"\n",

"In this example, if customers buy Bread they also tend to buy Butter. Some people often link products with high association to \"complementary goods\". In Economics 101, complementary good or service is consumed or used in conjunction with another good or service. Usually, the complementary good has little to no value when consumed alone, but when combined with another good or service, it adds to the overall value of the offering. For example a car and petrol. It would be of little value to buy petrol without owning a car. Complementary goods often have a negative cross-price elasticity of demand coefficient (Farnham, 2014). However, it is worth pointing out that, while complementary goods tend to have high association, not all products with high association rules are complementary goods. In MBA, we are more interested in product-pairs with high association rules i.e. products that are frequently purchased together. For example, in a retail store, MBA findings may show that Barbie dolls and candy are frequently purchased together, even though they are not technically complementary goods. In short, complementary goods are fairly obvious and common sense, but MBA seeks to uncover product associations that may not be so obvious and straighforward. In doing so, it is attempting to convert the abstract consumer tastes and preferences into association rules that are more insightful and actionable, from business perspective.\n",

"\n",

"\*\*\*\n",

"## Applications ##\n",

"There are many real-life applications of MBA:\n",

"- \*\*Recommendation engine\*\* – showing related products as \"Customers Who Bought This Item Also Bought\" or “Frequently bought together” (as shown in the Amazon example above). It can also be applied to recommend videos and news article by analyzing the videos or news articles that are often watched or read together in a user session.\n",

"<br>\n",

"<br>\n",

"- \*\*Cross-sell / bundle products\*\* – selling associated products as a \"bundle\" instead of individual items. For example, transaction data may show that customers often buy a new phone with screen protector together. Phone retailers can then package new phone with high-margin screen protector together and sell them as a bundle, thereby increasing their sales.\n",

"<br>\n",

"<br>\n",

"- \*\*Arrangement of items in retail stores\*\* – associated items can be placed closer to each other, thereby invoking \"impulse buying\". For example it may be uncovered that customers who buy Barbie dolls also buy candy at the same time. Thus retailers can place high-margin candy near Barbie doll display, thereby tempting customers to buy them together.\n",

"<br>\n",

"<br>\n",

"- \*\*Detecting fraud\*\* – identifying related actions whenever a fraudulent transaction is performed. For example, in a fraudulent insurance claim for stolen vehicle, it may be analyzed (from historical data) that claimant frequently report the incident a few days late (action 1) and often refuse to cooperate with insurer on investigation (action 2). Insurers can identify these red flags once certain behaviours or actions are displayed by the claimants.\n",

"<br>\n",

"<br>\n"

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"\*\*\*\n",

"## Case Study ##\n",

"For simplicity we are analyzing only 2 items – Bread and Butter. We want to know if there is any evidence that suggests that buying Bread leads to buying Butter.\n",

"\n",

"\*\*Problem Statament:\*\* Is the purchase of Bread leads to the purchase of Butter?<br><br>\n",

"\*\*Hypothesis:\*\* There is significant evidence to show that buying Bread leads to buying Butter.\n",

"\n",

"\n",

"Bread => Butter\n",

"\n",

"Antecedent => Consequent\n",

"\n",

"Let's take the example of a supermarket which generates 1,000 transactions monthly, of which Bread was purchased in 150 transactions, Butter in 130 transactions, and both together in 50 transactions.\n",

"\n",

"In set theory it can be represented as Bread only – 100, Butter only – 80, Bread and Butter – 50, as shown in the Venn diagram below:\n",

"\n",

"![alt text](images/set.jpg \"Example in a set\")\n",

"\n"

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"## Analysis and Findings ##\n",

"We can use MBA to extract the association rule between Bread and Butter. There are three metrics or criteria to evaluate the strength or quality of an association rule, which are support, confidence and lift.\n",

"\n",

"### 1. Support ###\n",

"Support measures the percentage of transactions containing a particular combination of items relative to the total number of transactions. In our example, this is the percentage of transactions where both Bread and Butter are bought together. We need to calculate this to know if this combination of items is significant or negligible? Generally, we want a high percentage i.e. high support in order to make sure it is a useful relationship. Typically, we will set a threshold, for example we will only look at a combination if more than 1% of transactions have this combination.\n",

"\n",

"\n",

"Support (antecedent (Bread) and consequent (Butter)) = Number of transactions having both items / Total transactions\n",

"\n",

"![alt text](images/support.jpg \"Support\")\n",

"\n",

"Result: The support value of 5% means 5% of all transactions have this combination of Bread and Butter bought together. Since the value is above the threshold of 1%, it shows there is indeed \*\*\_support\_\*\* for this association and thus satisfy the first criteria."

]

},

{

"cell\_type": "markdown",

"metadata": {},

"source": [

"\*\*\*\n",

"### 2. Confidence ###\n",

"Confidence measures the probability of finding a particular combination of items whenever antecedent is bought. In probability terms, confidence is the conditional probability of the consequent given the antecedent and is represented as P (consequent / antecedent). In our example, it is the probability of both Bread and Butter being bought together whenever Bread is bought. Typically, we may set a threshold, say we want this combination to occur at least 25% of times when Bread is bought.\n",

"\n",

"Confidence (antecedent i.e. Bread and consequent i.e. Butter) = P (Consequent (Butter) is bought GIVEN antecedent (Bread) is bought)\n",

"\n",

"![alt text](images/confidence.jpg \"Confidence\")\n",

"\n",