International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Business Analysis Tool for Investment Measures Using Recommendations

Amulya Bodla

Department of Information Technology, CBIT

Abstract: ***Business Predictions and recommendations are majorly important for any organization to proceed further and succeed. The basic problem that occurs is on what basis one should invest. Most of the times what happens is one invests in wrong ventures and that leads to a decline of the profits. One important aspect for any business to flourish is proper investments in the right field. This paper is going to suggest the proper fields of investment by examining the current trends. These conclusions can be made by analyzing a recommendation system. The efficiency of our recommendation system is proportional to the profits generated by using this as our analysis tool. Our analysis is done using some of the data mining strategies which we apply on the result of the recommendation algorithm.***

1. Introduction

Business Analysis, is required especially when it comes to marketing a certain product or analyzing the current trends in order to know where and when to invest. This basically requires two approaches, the quantitative and qualitative investigations. Both of them together will make the ideal analysis for the problems encountered , this may not give you 100 percent accurate results but will give you a solution much closer to real life scenarios.

But to do the analysis we need to get a clear picture of the transactions and the products being sold and the type of customers interested in purchasing these, so for this we need a tool. This tool can be constructed by using algorithms like collaborative filtering algorithm.[To implement these algorithms we can use the R TOOL or MATLAB or JAVA or any programming language which does the basic functions, but if we use R TOOL or MATLAB the pre-processing gets easier as it has some pre-defined functions which can be directly implemented]. This further can be classified into item based collaborative filtering and user based collaborative filtering. One of the above algorithms have to be chosen according to the situation in the organization.

These algorithms when implemented will give you certain classified output results in the form of clusters or categories which will enable us to take this further for the business analysis.

2. Methods and Procedures

a) Methods and Concepts

Baselining technique is used to manage the scope of the solution to a project and associate the requirements that are necessary for developing the project. In this technique we lock the requirements once the initial phase is done and as a result the estimations have to be done much early before we start implementing our idea.

b) Algorithms used.

Initially before applying our business tool we need to classify our data in such a way that it enables us to perform business analytics in a much easier way. For this we take the dataset and perform data cleaning methodologies to ensure the data set is perfect for us to implement the algorithms. The missing data values should either be filled manually or we can use any of the data mining strategies for filling it with either the average attribute value or a default value.

Once this is done we need to find a pattern in the purchases i.e if people on one locality are purchasing similar kind of products or if people from with same mindset or same kind of profession are planning to buy the same product or not.

Collaborative filtering algorithm can be implemented in two ways, that is using the item based collaborative filtering and user based collaborative filtering.

2.1 Item Based Collaborative Filtering

In this recommendations or similarities are found based on the items being purchased by a customer. If customer A purchases a product X, then the probability of him purchasing another product Y which is very similar to X is calculated.

2.2 User Based Collaborative Filtering

In this recommendations or similarities are found based on the similarity in users purchasing similar products. If customer A purchased products X, Y and Z, then the probability of customer B to purchase product Z who already purchased products X, Y is calculated.

Based on the results from either of these algorithms we can generate a list of products liked by a certain group of people.

a) Implementation of Algorithms

To implement them we first generate a similarity matrix. There are several methods to calculate similarity between two objects. It also is based on the attributes defined, if the attributes are binary then different methods are used and if the attributes are numeric then different implementation techniques are used.

For purchases we can take two different situations, if the attributes are binary then ‘1’ across a person A and object X would tell that, person A has purchased the object X and ‘0’

would imply that he did not purchase the object. But the problem here is that if a person purchases the object more than once then such things cannot be accounted in such a table. So it is better if we consider the numeric dataset where the number of purchases by each person is also noted.

Similarity between two people and their purchases will give us the matrix for both the algorithms which we have mentioned above. To find the similarity various distance measures can be used or the cosine similarity matrix can be generated. The distance measures that can be used are Euclidian distance, Minkowski Distance and Manhattan Distance. But a better distance measure is cosine similarity. Cosine similarity function gives the similarity between two objects and this value lies in the range from 0 to 1. The similarity value if 0 says that the objects are completely dissimilar and if 1 says that they are completely the same.

$$cos\ (a, b) = \sum(x*y) / (\sqrt{(\sum(x*x))} * \sqrt{(\sum(y*y))})$$

Applying this formula we generate a matrix which tells us the similarity between two objects. This matrix will be a 2D matrix which will give the similarity value between any two objects. From this when a person purchases an object X we can generalize the other items he might buy with the highest similarity value with whichever other person he might have. For simplicity we can sort the database in the descending order of similarity values calculated between this person A and every other person included in the initial database.

Similarly we can implement the user based collaborative algorithm also by considering all the users who have purchased X and have also purchased other products purchased by X and predict the probability of purchase of every other product by the similarity values of each and every person when compared with A. We can compare the efficiency of both the algorithms which each other by having a predefined trained data set. But better results are obtained when a hybrid of both these algorithms is used. In this the user-item based matrix is taken and Euclidian distance measure is used to find the distances and similarities between objects. Then sparse users are removed from it to generate the dense matrix after the user based filtering is done and then on the dense matrix generated item based filtering is done. Hence this ensures more efficiency of the algorithms.

b) Business Tool Implementation

Now each of the user gets a certain number of recommendations or so called neighbors which are similar to the object being purchased or which has more likelihood of such objects being purchased by him. Since in any market scenario we have a pre-existing database of the users and when this database is thoroughly examined and when these algorithms are applied on this, some extremely interesting patterns can be generated.

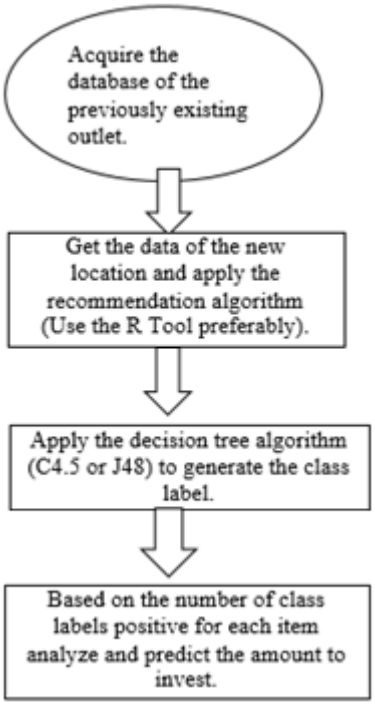

Figure 1: Flow Chart representing the process.

The sole reason behind these patterns is what a business analyst needs to do, so one such person is going to see the trend in these patterns. If it is because of the place they live in, the type of profession they are in or if it is because of their very own personal culture and taste. This basis criteria can be decided by the decision tree mechanism of DATA MINING. Many algorithms exist which will generate the decision tree and confusion matrix can also be drawn with this data. This will give us a clear picture on what basis the sales are progressing or on what basis one performs a certain action. Then based on these results we can know how much investment we can make in each sector or on each product. Another thing to keep in mind while making investments is the per unit price of the product along with the per capita income of the product. The ratio between these two will give you a value, if this value is closer to one then the likelihood that a product might be purchased by that group of people with that particular income is high. If not the chances that they might buy that product is low. Decision Trees can be generated by following the C4.5 Algorithm. The main steps are as follows

1) First take all the recommendations or similar objects from the algorithm and then put them aside for the decision tree construction.
2) Divide them into classes, if all the attributes belong to the same class then put them with a class label and make it as a leaf node.
3) If they do not belong to the same class then move up by a level and generate a separate node with this class label. This separation is done using the splitting criterion as mentioned above, which tells us where to split and create a new branch.

But we need a criterion for choosing the splitting criterion and this can be done using the information gain function which will tell us which attribute to be split based on the value generated for each attribute.

Entropy(X) = -∑P (value i) $\log_2$ P (value i)
I.G(X, A) = Entropy(X) - ∑ ((| X_i |/ |X|)*Entropy (X_i)).

From the above formulae for each attribute calculate the gain ratio, and split that attribute which has the highest gain ratio. The P (value i) tells us the probability of the number of objects that come under one class out of the total number of objects. By this we classify the recommendations into different classes as to which class of objects is bought by which type of people and hence it enables us to make our prediction even better and increases the efficiency of our business tool.

We generate a class label to each of the recommendations and based on this class label we prioritize and analyze the option of either investing more in the particular item or not. And also based on the number of items that fall under a particular class label we can decide the proportion of the total new investments which can be diverted to this section.

3. A Real Life Example

Consider a company opening its branch outlet at a new location, or consider a first outlet being opened. Then it needs to decide on what kind of items it has to store in what amount, since it cannot waste resources on storing something which is not going to be in demand in that particular location. So for this it can be implemented using the above mentioned idea. We initially take the database X of an existing company outlet in another location and also the database Y of the people living in this new location. We train our recommendation algorithm by database X . In database X we have attributes such as occupation, monthly income, culture, age, level of education, etc. Now take the database Y and compute similarities between these two using the similarity function and implement the hybrid user-item collaborative algorithm, with this we get a generalized view for each and every member in database Y a set of similar customers from database X and also a list of similar items which he is likely to purchase. From this list we generate a decision tree and prune it for efficiency. The tree generated using the C4.5 algorithm will in the end give a class label specifying whether a particular person is likely to purchase that object and the probability of him purchasing it is mentioned. When this methodology is applied for the entire database we get the approximate number of items in demand for each kind of object. Based on this one can easily know how much to invest and the amount of stock to be purchased before hand to meet the demands without incurring any losses.

4. Conclusion

Here finally we calculate the amount to be invested in each item using the following formula.

$I_m = X/N_{m\ where}$ X denotes the total investment available for the requirements, N_m denotes the number of items with positive purchases under each class label and I_m gives you the investment for this category of products.

The rate of accuracy of this method is 84%. So this method gives a clear picture of the investment mechanism and makes it easier for new ventures to develop without incurring losses.

References

[1] Deepak K Agarwal, "Statistical Methods for Recommender Systems."
[2] Pradeep Hari, "Business Analysis: Visualizing Business Processes and Effective Software Solutions".
[3] R and Data Mining Case Studies by Yachang Zhao.
[4] Han and Kamber- Data Mining Concepts and Techniques 3rd Edition.
[5] Kargupta and Joshi , "Data Mining: Next Generation Challenges and Future Directions.
[6] Data Mining for Business Analytics by Nitin R.Patel.

Author Profile

Amulya Bodla –Student, 3rd year of Engineering (B.E), Department of Information Technology, Chaitanya Bharati Institute of Technology