Big data applications in business analysis

Wenqiang Huang

China Southern Airlines

Sien Chen

School of Economics, Xiamen University
sandy80@vip.sina.com

Zhenyu Liu

School of Management, Xiamen University

Abstract

Using dataset provided by the airline company, this paper demonstrates how to apply big data techniques to explore passengers' travel pattern and social network, predict how many times the passengers will travel in the future, and segment customer groups based on customer lifetime value.

Keywords： Big Data, Airlines, Data Mining

BACKGROUND

In order to illustrate how the massive passenger data can be utilized for business decision making, this paper presents a real-world case of China southern Airlines. Using Recency-Frequency-Monetary (RFM) analysis, this chapter presents how to apply big data techniques to explore passengers' travel pattern, and predict how many times the passengers will travel in the future. The findings will provide airline companies to make more effective customer relationship management. Airline also could analyze webtrends data to explore passenger behavior of websites and mobile usage (Davenport TH, 2012). Passenger's webtrends information includes mobile number, membership number, identity number, and other web browsing records (Ghee R, 2014). Connecting these webtrends data with other information sources provides an overview and insights on individual passenger's website and mobile usage. The following session demonstrates how the accessing event flow on WebTrends can be configured and incorporated into the sequence analysis of passenger events.

RECENCY-FREQUENCY-MONETARY ANALYSIS

Recency-Frequency-Monetary method is considered as one of the most powerful and useful models to implement consumer relationship management (Khajvand M et al. 2011). Bult and Wansbeek defined the variables as: (1) R (Recency): the period since the last purchase; a lower value corresponds to a higher probability of the customer's making a repeat purchase; (2) F (Frequency): number of purchases made within a certain period; higher frequency refers to greater loyalty; (3) M (Monetary): the money spent during a certain period; a higher value means that the company should focus more on that customer.

This case study adopted an extended RFM model to analyze the airline passenger behavior. The extended RFM model incorporated average discount factor as an additional variable, because average discount factor is an important indicator to measure the price level of passenger's airline purchase. The average discount factor defined here is ratio of purchase price to published price of the airplane seat. Therefore, the extended RFM model involves four variables: number of days from the last order date to modeling (R), number of flight trips (F), sum of consumption (M), and average discount factor (D). In this way, a traveler's ID generates the consolidated data.

Principal component analysis was used to score individual travelers based on the RFMD variables, and 16 consumer groups were identified. The findings would help marketers to recognize those most valuable consumers and establish profitable consumer relationship. The procedure of the RFM analysis is described as below.

EXPLORATORY DATA ANALYSIS

This step involves taking a closer look at the data available for investigation. Exploratory data analysis consists of data description and verifying the quality of data from the airline company's databases. Table 1 and Table 2 provide a general understanding of the passenger data set. Table 1 reveals that difference between the maximum and the minimum of the two variables: number of flight trips and sum of consumption is huge. The data distribution plot also indicates that the original data heavily right-skewed. Therefore, using the original data directly in our modeling will have big problem. In order to fix this data problem, logarithmic transformation is used regarding number of flight trips, sum of consumption and average discount factor. We also take the opposite number regarding the difference of dates from the last order date to modeling date, and then standardize the data to remove dimension's influence.

Table 1 – Descriptive Data Analysis of RFMD Variables

Modeling variables	N		Mean	SD	Minimum	Maximum
R: Number of days	6879711		172.62	166.13444	1	730

from the last order date to modeling						
F: Number of flight trips	6879711		2.27	2.22444	1	214
M: Sum of consumption	6879711		1997	2574	0	318430
D: Average discount factor	6879711		0.62511	0.22226	0	4.33

Table 2 indicates that the number of flight trips positively correlates with income. The more flight trips, the bigger sum of consumption, which corresponds with the flight reality.

Table2 – Corrlation Matrix of RFMD Variables

	R	F	M	D
R: Number of days from the last order date to modeling	1	-0.01536	-0.0396	-0.13952
		<.0001	<.0001	<.0001
F: Number of flight trips	-0.01536	1	0.84767	-0.04863
	<.0001		<.0001	<.0001
M: Sum of consumption	-0.0396	0.84767	1	0.17584
	<.0001	<.0001		<.0001
D: Average discount	-0.13952	-0.04863	0.17584	1
	<.0001	<.0001	<.0001	

Pearson correlation, N=6879711; When H0: Rho=0, Prob>|r|

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is used to determine the weight of each RFMD variables. Figure 1 shows the steps of principal component analysis of RFMD modeling. Through the principal component analysis, the result shows that the three RFM modeling variables account for 95% of the overall variance. The four RFMD variables and weights were determined to further depict the passenger's value model in Table 2. In particular, weight of number of days from the last order date to modeling is 1.24，weight of number of flight trips is 1.19，weight of sum of consumption is 1.42，and average discount factor is 0.55.

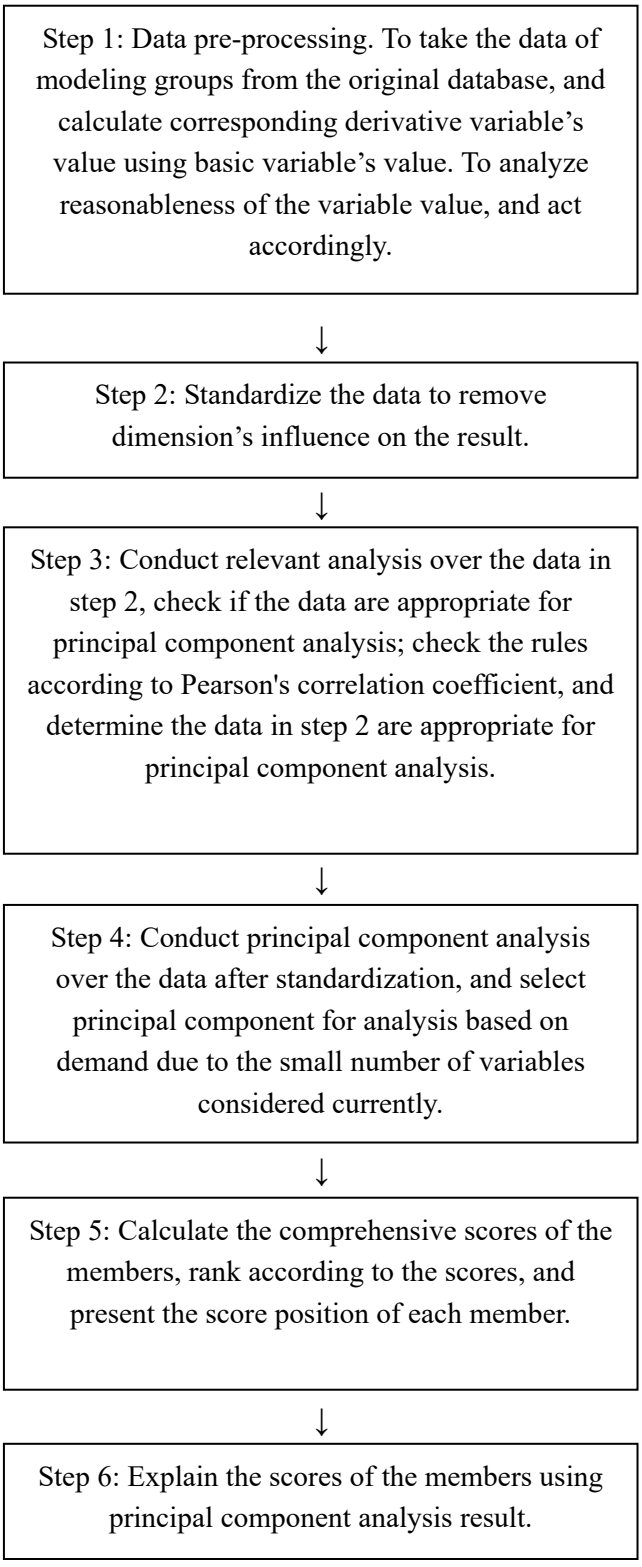

Figure 1 – Principal component analysis steps

Table 3 – Basic Statistics of RMFD Data

Modeling variables	Weight (all principal components)	Weight (the former three principal components)
R: Number of days from the last order date to modeling	1.24	1.24
F: Number of flight trips	1.30	1.19
M: Sum of consumption	1.31	1.42
D: Average discount	0.61	0.55

CLUSTERING ANALYSIS

K-mean value clustering method was applied to generate 16 passenger groups. The four RFMD indicators can be used to analyze specific target groups in more details. The four RFMD indicators can help to rank the level of passenger lifetime values, and determine individual marketing strategy and realize precision marketing in respect of individual high-end travelers.

The concept of Customer lifetime value (CLV) is adopted to evaluate the profitability of each cluster. CLV is the present value of all future profit generated from a customer. In this case study, the average CLV value of each cluster can be calculated with the equation:

$$CLV_{ci} = NR_{ci} \times WR_{ci} + NF_{ci} \times WF_{ci} + NM_{ci} \times WM_{ci} + ND_{ci} \times WD_{ci} \quad (1)$$

NR_{ci} refers to normal recency of cluster ci, WR_{ci} is weighted recency, NF_{ci} is normal frequency, WF_{ci} is weighted frequency, NM_{ci} is normal monetary, WM_{ci} is weighted monetary, ND_{ci} is normal duration of cluster ci, and WDci is weighted duration. The result of clustering analysis is shown in Table 4. Based on the result of clustering analysis, some insights of customer segmentation and corresponding business strategies can be developed. For more than half a year not to seize the opportunity of direct labeling for low value, such as scores of 5 groups of 33.46 points, although the flight times higher than the average, but because has more than 500 days of no on the website to purchase, so as count loss handling group (referred to here as the loss is a loss to the other channels or other companies). The 3 and 9 groups scored higher on average, each index are better than the average value. Belonging to the company is important to keep customers and key customers, to them for continuous customer care, improve service measures, to improve the customer experience, when the group for promotional price of measures needs to be carefully considered. For each customer in each group have a score, the specific size of the order in accordance with the order of priority will be allocated to these groups. Group 1 although the scoring value is relatively low, because their flight number is relatively low, but they have three months there is the opportunity that is active, is the need to pay attention to the group, perhaps it is this one individual constitute the sales site of the long tail.

Table 4 – Result of Clustering Analysis

Group	Number of days from the last order date to modeling	Number of flight trips	Sum of consumption	Average discount	Customer lifetime value	Percentage	Label
1	85	1.00	555	0.604	13.27	5.02	Low value
2	94	1.00	881	0.582	24.86	6.04	Low value
3	70	14.30	14951	0.715	99.55	1.53	Promising
4	534	2.10	1077	0.377	5.64	5.66	Low value
5	523	2.33	2296	0.689	33.46	7.02	Low value
6	151	2.26	2527	0.789	77.92	11.30	Customer to retain
7	130	1.89	4052	1.700	86.75	1.03	Customer to retain
8	89	1.02	893	0.930	42.07	4.48	Low value
9	118	5.17	5395	0.694	95.13	5.80	Customer to develop
10	95	1.01	868	0.732	31.29	7.42	Low value
11	99	3.61	2750	0.498	85.11	7.37	Customer to develop
12	98	1.00	1464	0.649	46.73	4.61	Low value
13	290	2.18	1436	0.474	35.72	9.48	Low value
14	75	2.07	1588	0.535	66.68	11.39	Customer to develop
15	94	1.07	1590	0.882	60.10	4.59	Low value
16	85	2.15	840	0.320	35.20	7.27	Low value
Mean	173	2.27	1997	0.625	50.50		

(↑) indicates the value for the cluster is higher than the mean.

By analogy, through a combination of the specific analysis of the actual business, we can roughly to the website to buy customer groups is important to keep customers, important customer development, important to retain customers, low value customers four parts.

CONCLUSION

Using the data mining techniques discussed in this chapter, the airlines companies can learn important marketing implications of big data analytics. This paper introduced Recency-Frequency-Monetary (RFM) analysis, and the data mining results reveal passenger travel patterns, preference, travel social networks and other aspects of purchase behavior. The law of each individual passenger for scoring, we get 16 types of customers, the average score of each type of customer, according to the score and the average score of each person, to evaluate the level of the value of such customers, so as to get the specific circumstances of the 16 types of customers. Marketing personnel can be taken to maintain, develop, retain, temporarily do not pay attention to other operations in the case of these customer groups.

ACKNOWLEDGEMENTS

Supported by Provincial Science and technology project of China Guangdong Province (2015B010109006).

Bibliography

Harteveldt HH (2012) The future of airline distribution: a look ahead to 2017. s.l.: *special report commissioned by IATA*.

Davenport TH (2013) At the Big Data Crossroads: turning towards a smarter travel experience. *Available via AMADEUS*. http://www.bigdata.amadeus.com/assets/pdf/ Amadeus_Big_Data.pdf. Accessed 1 May 2014

Ghee R (2014) Top 5 in-flight trends to look out for in 2014. *Available via* http://www.futuretravelexperience.com/2014/01/top-5-flight-trends-look-2014/. Accessed 7 May 2014

Chen J, Xiao YB, Liu XL, Chen YH (2006) Airline seat inventory control based on passenger choice behavior. *Systems Engineering–Theory and Practice* 1: 65-75

Fader PS, Hardie, BGS, Lee, KL (2005) "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Sci* 24: 275-284

Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. *Addison Wesley, Upper Saddle River*, NJ

Khajvand M, Zolfaghar K, Ashoori S, Alizadeh S (2011) Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Comput Sci* 3:57-63

Gupta S, Lehman DR (2003) Customers as assets J *Interact Mark* 17(1):9-24