

# Mining Airline Data for CRM Strategies

LENA MAALOUF, NASHAT MANSOUR

Division of Computer Science and Mathematics, Lebanese American University,  
Mme Curie St., Kreitem, Beirut, LEBANON

*Abstract:* In this paper, we apply data mining techniques to real airline frequent flyer data in order to derive customer relationship management (CRM) recommendations and strategies. Clustering techniques group customers by services, mileage, and membership. Association rules techniques locate associations between the services that were purchased. Our results show the different categories of customer members in the frequent flyer program. For each group of these customers, we can analyze customer behavior and determine relevant business strategies. Knowing the preferences and buying behaviors of customers allow marketing specialists to improve campaign strategy, increase response and manage campaign costs by using targeting procedures, and facilitate cross-selling, and up-selling.

*Key-words:* customer relationship management, data mining, decision support, intelligent information processing.

## 1 Introduction

The variety of offers and availability of communication technologies provide airline customers the power to access information on competitors, products, availability, and prices. Due to these factors, airline business has to become customer centric. Companies have to identify the most valuable customers and the appropriate strategies to use in developing relationships with these customers. Such strategies would include developing one-to-one relationship with customers using market segmentation and Customer Relationship Management (CRM). Lee [4] defines CRM as a concept that has been developed from marketing theory offering an interaction of the entire business with customers. CRM is a management model that has the potential of converting a production-driven airline into a customer-driven airline in order to significantly raise the airline's efficiency and effectiveness.

Customer acquisition deals with profiling, segmentation, and ranking of customers based on tendency to buy, order frequency, and purchasing behavior. Segmentation is the process of separating customers into groups according to common characteristics so that marketing and operational strategies can be targeted to specific populations [3].

The airline data we consider consist of frequent flyer data for which decisions require processing of a large amount of data. Often, airlines use methods based on human expertise and, thus, developing computerized solutions are badly needed. We propose

using data mining techniques for analyzing real-world frequent-flyer data. Previous work in this field is minimal. The objectives of previous works on mining frequent flyer airline data have been: (a) categorizing customers into groups based on sectors most frequently flown, class flown, period of year, hometown compared to sector flown [6]; (b) classifying trip purposes into leisure, business, etc... [5]; (c) addressing airline ticket prices behavior over time [2].

Our objective is to explore the Frequent Flyer database using data mining (DM) methods in order to prepare for CRM implementation. We have also used the Cross-Industry Standard Process for Data Mining (CRISP-DM) process cycle. Our contribution in this paper is based on the following: (a) Selected data mining techniques (clustering and association rules) are applied to Frequent Flyer airline data with new CRM objectives; (b) A preprocessing technique is used for processing the huge amount of data for a feasible application of DM techniques; (c) We use real data from MEA airlines and conduct experimental work for validating our techniques.

## 2 Problem and Data Description

### 2.1 Data Description

The goal of our study is to extract business and CRM strategies for an airline company. The data source is the frequent flyer program. Frequent flyer programs data allow getting a better understanding of customer types and behaviors. The program intends to identify

high value customers and provide them with special services and benefits such as upgrades.

In this study, the Frequent Flyer Program is an air miles reward program including, in addition to the flight services, a financial service of a credit card and a hotel service. Each time a passenger uses the dedicated credit card for any transaction or has a stay in the dedicated hotel, he/she will win additional miles in the reward program. Due to agreements with banks and hotels, the airline generates revenue. Additional services are provided to the passenger such as Adjustment, Miscellaneous, Multi, Program and Reward Claim. Adjustment is used to rectify errors when it occurs with mileage calculation. Miscellaneous covers compensation for delay, survey and others. Multi is available only for Elite and President Club members. It is a mileage bonus given when the passenger uses a group of services such as 3 dedicated flights or 5 flights in a special class. Program groups the mileage received due to promotion packages such as class of service program given double mileage. Customers are divided into four categories of members. The data used in this study are based on 1,322,409 customer activities transactions and 79,782 passengers for a period of 6 years.

## 2.2 Problem Description

The objective of this study is to help market specialists in decision-making concerning some of the key business process questions. For the frequent flyer customer data, these questions are as follows.

### Customer value measurement:

- Which customers are the most valuable? What activities contribute to their value?
- Are the most valuable customers receiving an appropriate allocation of services to retain them?
- Which customers are most promising for a defined campaign?
- What can be done to transform low profit customers to a position of improved profitability?
- What is the predicted lifetime value by customer segment?

**Customer retention:** Define best market segment.

### Customer growth:

- What customer segment has a potential to purchase additional travel segment?
- Identify up-selling and cross-selling opportunities
- Design packages or grouping of services.

### Customer acquisition:

- What constitutes a good customer?

- What are the attributes and characteristics of the most valuable customer segments?
- Can we match new customers to the right services?

## 3 Solution Strategy

The data preparation task includes data cleansing and preprocessing. The resultant data will be the input for the data mining process.

### 3.1 CRISP Implementation

#### Business goals

Our target customers shall be not only those who spend much on the airline ticket, but also the valuable candidates for cross-selling. The main concern is to understand customers in order to implement new strategies to different customer segments. The results can be used for marketing purposes such as promotions and targeted campaigns, and improving customer service such as information availability for call centers.

#### Data mining goals

Our goal is to develop models that generate passenger revenue value, based on the booking history. We use customer transaction data to track buying behavior and create strategic business initiatives. Business can use these data to divide customers into clusters. These clusters highlight marketing opportunities such as cross-selling (selling new products) and up-selling (selling more of what customers currently buy).

#### Data preparation

Data is based on Z-Score Normalization ( $x_{new} = (x_{old} - shift)/scale$ ). The values for shift and scale are computed to be shift = mean, and scale = standard deviation, respectively.

#### Data transformation and aggregation for clustering

Several queries have been built to merge the "Activities" transaction data to the "Individuals" passenger file. These queries create the clustering input record. The queries determine the manipulation done on the transaction data. It includes pivoting, aggregating, and inserting into each passenger record. We discard customer records with missing values. The records remaining are 50,830 records.

#### Data transformation and aggregation for association rules

The result generated by clustering provides customer segmentation with respect to important dimensions of

customers’ needs and value. Two different approaches have been used for association rules application. Each approach is based on different data.

*Approach Based on Original Activities*

In the clustering process; the “Customer ID”, “Flight”, “Financial”, and “Hotel” activities are used as services purchased by customers. A query (Q5) is used to group all the Frequent Flyer Customer information. Query (Q6) is based on a selected Cluster and Q5. It groups best customers information.

*Approach Based on Flight Activities Only*

In the second approach, we consider only the Flight activity from our best customer (Selected Cluster) to study and analyze the sector used taking into account that the originating airport have to be one of our main Hubs. A query (Q7), based on “Activities” table, includes the Customer ID, Sector (concatenation of Origin and Destination), Origin (must be one of our main Hubs only), Destination, and the Activity Type (“Flight” only). It groups Customer information by Sector. Query (Q8) is based on Selected Cluster and Q7. It includes the Customer ID, and sector used only by the Selected Cluster customers. It groups best customers information.

**Model building and evaluation**

Using a data mining tool, we apply clustering and association rules techniques [1] in order to generate marketing and CRM strategies. Behavioral clustering help derive strategic marketing initiatives using the variables that determine customer value. By conducting association rules within behavioral segments, we can define tactical campaigns.

The clustering techniques used are the k-means and O-Cluster algorithms. We use the Apriori algorithm for association rules mining.

**4 Experimental Results**

**4.1 K-Means Clustering**

In our study, we have used the Oracle Tool called Oracle Data Miner (ODM). The first step in the clustering process is to choose the basic run parameters for the K-means algorithm. Different scenarios have been tested. For brevity, we present only a limited set of results. The algorithms are applied on “Behavioral Clustering” query including 50,830 records.

The input variables we selected include:

- Number of services (“Financial”, “Flight”, or “Hotel”) the customer used over lifetime (ACTLIFE).
- Number of services (“Financial”, “Flight”, or “Hotel”) the customer used in the last 12 months (ACTLASTYEAR).
- Customer’s revenue mileage contribution over lifetime (MILEAGE).
- Customer membership period in months. Number of months since customer first enrolled in the program (MEMBERSHIP).
- Revenue Mileage / Membership period (RMM).
- Number of services over lifetime / Membership period (RAM).

The basic parameters available for clustering are:

- Maximum number of clusters.
- Maximum iterations or Maximum number of passes through the data.
- Minimum Error Tolerance. It must be between 0.001 (slow build) and 0.1 (fast build). Increasing minimum error tolerance builds models faster, but with lower accuracy.

The model stops after either the change in error between two consecutive iterations is less than minimum error tolerance or the maximum number of iterations is greater than maximum iterations. For clustering run, we choose a maximum of 9 clusters; a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. Table 1 gives the details of a sample of two clusters.

**4.2 Association Rules**

The result generated by k-means clustering are used as a basis for the association rules algorithm. The first step in the process is to choose the basic run parameters for the Apriori algorithm. Two different scenarios have been applied. The first scenario is based on “Financial”, “Flight”, and “Hotel” activities with 1,896 records. The second scenario is based on the flight activities especially the sectors, with 1,867 records. The results are evaluated using support and confidence attributes [1].

For association rules, we choose our best customers cluster, Cluster 16, which has 1,886 records or customers. The input variables are divided into two different scenarios depending on the cases studied with the association rules. The case presented herein is based on Original Activities using the Query “Original Activities Cluster 16”. The “Original Activities Cluster 16” query includes: The Customer ID; Financial (The value is 1 if the customer has used the service; otherwise the value is “0”); Flight (The

value is 1 if the customer has used the service; otherwise the value is “0”); Hotel (The value is 1 if the customer has used the service; otherwise the value is “0”).

**Table 1: Details of Clusters 8 and 16**

Cluster ID	Cluster Level	Record Count	Attribute	Centroid Value	Attribute	Centroid Value
8	4	9,414	ACTLASTYEAR	1.02-1.05	MILEAGE	12616.56-21492.34
			ACTLIFE	1.02-1.05	RAM	0.0402-0.0469
			MEMBERSHIP	23.6-24.32	RMM	750.6066-1210.9099
16	5	1,239	ACTLASTYEAR	1.98-2.01	MILEAGE	48119.68-56995.46
			ACTLIFE	2.01-2.04	RAM	0.0335-0.0402
			MEMBERSHIP	57.44-58.16	RMM	750.6066-1210.9099

We look for two types of association rules. For the first one, we keep only the sectors that have a percentage of use greater than 10%. For the second one, we keep those greater than 20%.

We implement the Apriori algorithm of ODM to build association models. The algorithm settings in the Apriori algorithm depend on the marketing professional decision. The minimum support controls the rules produced depending on the application percentage of this rule on existing data; minimum confidence controls the production of rules depending on the probability of having this rule in future data.

The algorithm settings are as follows: minimum support was set to 0.1; minimum confidence was set to 0.5; number of attributes in each rule was set to 3.

*Scenario 1:* The run is based on “Original Activities Cluster 16” query. Table 2 displays a sample of the rules.

*Scenario 2:* The second scenario is based on “Activities Cluster 16” query. We keep from the “Activities Cluster 16” query the sectors used by the customer with a percentage greater than 10%. The remaining number of sector field is 17 sectors. Table 3 gives a sample of the rules.

**Table 2: Sample Association Rules for Best Customers Activities (Scenario 1)**

Rule Id	If (condition)	Then (association)	Confidence	Support
3	FLIGHT=1 and HOTEL=0	FINANCIAL=1	1	0.91251326
2	FINANCIAL=1 and HOTEL=0	FLIGHT=1	1	0.91251326
1	FINANCIAL=1 and FLIGHT=1	HOTEL=0	0.9907887	0.91251326

**Table 3: Association Rules for Best Customers Activities (Scenario 2)**

Rule Id	If (condition)	Then (association)	Confidence	Support
498	BEYCAI=1 and BEYDXB=1	BEYAMM=1	0.5799458	0.11462239
494	BEYAMM=1 and BEYCDG=1	BEYCAI=1	0.6005155	0.12479914
1228	BEYCAI=1 and BEYDXB=1	BEYCDG=1	0.8157182	0.1612212

## 5 Discussion Of Results

### 5.1 K-Means Clustering

Table 4 provides a summary of the profile produced by k-means clustering that includes: revenue mileage, number of services used, and customer membership period. The purpose is to quantitatively assess the potential business value of each cluster and rules by profiling the aggregate values of the variables by cluster and rules. We have used the following parameters for evaluation:

- Revenue Mileage percentage = (Total Mileage per cluster \* 100)/ Total Mileage.
- Customer percentage = (Total Customer per cluster \* 100)/ Total Number of Customer.
- Average Service per Cluster = Sum of Act Life / Total Number of Customer.
- Service Index = Average Service per Cluster / Average of Different Services used overall.
- Weight or Mileage per Customer = Revenue Mileage Percentage / Customer Percentage.
- Membership = Sum of Membership per Cluster / Number of Customer.

**5.1.1 Clustering Analysis**

The most profitable cluster is cluster 16. From Table 4, this cluster groups about 8.88% of the mileage with only 3.71% of the passengers and has the highest weight fraction. A valuable business opportunity is shown in this cluster profile based on increasing the number of services used by passengers.

It is obvious that clusters 11, 16, and 17 contain the best customers. These passengers have a higher mileage per passenger than other clusters, as shown by the weight column in Table 4. Some possible CRM strategies would include:

- A retention strategy for best customers (in clusters 11, 16, and 17).
- A cross-selling strategy for cluster 8. Cluster 8 has a service index close to that of cluster 16. Cluster 16 has the highest number of services used. The effort needed to convert passengers from cluster 8 to cluster 16 should be minimal, since both clusters are close in number of services used. The comparison of services bought by the best passengers of cluster 16 to those purchased by cluster 8 passengers would determine services that are candidates for cross-selling.
- The same cross-selling strategies can be applied between: 15 and 11; 13 and 17 because they are close in services value.

- Cluster 9 has to be observed closely during some period of time. It defines a group of new passengers. We have to collect more data to determine the behavior of those new passengers. We have to adopt some marketing efforts to inform cluster 9 passengers of the frequent flyer program’s products and services in order to accelerate profitability.
- Cluster 12 is the worst, since its passengers have very low mileage percentages. These passengers use very few services even though they have been with the company for 37 months. The strategy may be to minimize spending on marketing to this group.

**5.1.2 Best Route from CDG**

The result of clustering was used to prepare data for association rules. As shown before, based on our best customers (Cluster 16) we have prepared the query “Activities Cluster 16”. This query contains 145 sectors flown by our best customers. The percentage of each sector flown by customers with origin CDG shows the preferable routing from the CDG hub. This approach will be applied on the result given by the k-means algorithm.

**Table 4: Clustering Analysis for K-Means Algorithm**

Cluster ID	Mileage %	Customers %	Avg. Services per Cluster	Service Index	Weight	Membership (Sum Membership/ NB. Customer)
17	34.70	17.02	1.00	0.971	2.04	67.87
11	20.62	16.66	1.01	0.977	1.24	40.78
8	12.10	14.38	1.01	0.979	0.84	21.35
16	8.88	3.71	2.01	1.951	2.39	44.53
9	7.67	16.67	1.00	0.976	0.46	9.26
14	5.49	8.76	0.94	0.913	0.63	70.61
15	4.97	9.45	1.00	0.971	0.53	54.73
13	2.92	7.25	0.99	0.961	0.40	22.28
12	2.63	6.10	0.92	0.896	0.43	37.20

Average Number of Services used overall = Sum of Activities used over lifetime / Number of Customers: 1.03

**5.2 Association Rules**

The association rules evaluation is based on the scenarios discussed before. We analyze the rules for each scenario, each with confidence and support values. Future plans have to be based on the confidence. Such plans can be a marketing campaign, or special offers. In this subsection we present an

analysis of two scenarios of the association rules results.

**5.2.1 Scenario 1**

Scenario 1 is based on the “original activities” of cluster 16 – the best customer cluster. The original activities are “Flight”, “Financial”, and “Hotel”. We

conclude from the results that customers are divided into two different categories:

- The customers using the “Flight” and “Financial” services never use the “Hotel” Services.
- The customers using the “Flight” and “Hotel” services never use the “Financial” Services.

A manual inspection of the data has been done. To enhance business, we have to divide customers into two different categories; Flight/Financial customers and Flight/Hotel customers. Hence, two different marketing campaigns have to be launched.

### 5.2.2 Scenarios 2

Scenario 2 is based on the “activities” of cluster 16. The activities are mainly the sectors flown by the customers. The sectors are restricted to the sectors originated from the main hubs. This scenario addresses the sectors with flown percentage over 10%. In the following, we present some interesting rules.

- BEYDXB = 1 and BEYRUH=1 → BEYCDG = 1 with support = 0.1 and confidence = 0.84. That is, 10% of the best customers are traveling to Beirut/Dubai, Beirut/Riyadh, and Beirut/Charles-De-Gaulle. Hence, the airline has an opportunity to enhance its business for customers traveling on the sectors Beirut/Dubai, and Beirut/Riyadh such as marketing campaigns or special offers on the Beirut/Charles-De-Gaulle sector.

## 6 Summary of CRM Recommendations

The best scenario for clustering using k-means algorithm generated 9 different clusters with specific profile for each one. Such information is valuable in determining the resources the airline should commit in order to gain and retain a customer in the event he/she should defect. The cluster profile shows a business opportunity in increasing the number of services purchased by customers.

We track high-value customers. The results show three clusters as best customers. A retention strategy should be applied to these customers. Another result in these clusters is providing opportunities for the airline to produce more revenue from a customer. For example, the airline could apply an up-selling strategy by selling a higher fare seat.

The second type of clusters defined in this study is the mid-range cluster. The analyst of customer behavior would propose an enhanced strategy for customers in these clusters in order to increase services usage and revenue mileage per passenger.

This strategy shall define candidate services for cross-selling.

The third type of clusters identified in this study is new customer cluster. The recommendation is to observe these customers to determine their behavior in order to improve profitability.

The fourth type of clusters includes the bad customers with very low revenue mileage per passenger. The recommendation is to retain marketing campaigns for those customers.

We have found that the best route occurs from the CDG airport-hub. This best route helps in defining new route market, develops marketing strategy for customers to propose the route with low sales, identifies customers’ preferable destinations, and observes the worst route in order to take a decision: stop it or market it more aggressively.

The association rule algorithm based on the best customer cluster provides more results. By analyzing the services used, we characterize services integration. It enables the airline to serve a customer the way the customer wants to be served based on the stated and observed requirements of the customer.

The second use of association rules explored routes. It allowed us to propose to customers additional route flight tailored to the needs, behavior, and values of the airline’s most profitable customers.

**Acknowledgment.** This work was supported in part by the Lebanese American University.

### References:

- [1] Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Prentice Hall.
- [2] Etzioni, O.; Knoblock, C.; Turchetta, R.; & Yales, A. (2003). To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price. ACM. [www.isi.edu/integration/papers/etzioni03-kdd.pdf](http://www.isi.edu/integration/papers/etzioni03-kdd.pdf)
- [3] Fennell, G.; & Allenby, G. (2004). Market definition, market segmentation, and brand positioning create a powerful combination. [fisher.osu.edu/~allenby\\_1/2004%20Integrated%20Approach.pdf](http://fisher.osu.edu/~allenby_1/2004%20Integrated%20Approach.pdf)
- [4] Lee, D. (1999). CRM Definitions. CRM.Talk #054. [www.crmguru.com/content/crmtalk/2000a/crmt054.htm#1](http://www.crmguru.com/content/crmtalk/2000a/crmt054.htm#1)
- [5] Pritscher, L.; & Feyen, H.; (2001). Data Mining and Strategic Marketing in the Airline Industry. Atraxis AG, Swissair Group, Data Mining and Analysis, CKCB. [www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/pritscher1.pdf](http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/pritscher1.pdf)
- [6] Ramachandran, P. (2001). Mining for Gold. WIPRO Technologies. [www.wipro.com/whitepapers/services/businessintelligence/dataminingmininggold.htm](http://www.wipro.com/whitepapers/services/businessintelligence/dataminingmininggold.htm)