Mining Airline Data for CRM Strategies

LENA MAALOUF, NASHAT MANSOUR
Department of Computer Science and Mathematics, Lebanese American University,
Mme Curie St., Kreitem, Beirut, LEBANON
E-mail: nmansour@lau.edu.lb

***Abstract**:* In today's competitive climate, customer relationship management (CRM) has become an essential component in airline business strategies. CRM in the airline industry would be based on analyzing customer data in order to understand preferences and behavior. In this paper, we apply data mining techniques to real airline frequent flyer data in order to derive 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 our customers allow our marketing specialist to improve campaign strategy, increase response and manage campaign costs by using targeting procedures, and facilitate cross-selling, and up-selling.

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

1 Introduction

The airline industry faces real challenges based on changes in customer behavior, competition and technology. Thus airlines need to identify, develop, and implement better business strategies [1]. The variety of offers and availability of communication technologies provide customers the power to access information on competitors, products, availability, and prices. Due to these factors, business has to become customer centric. With new challenges and competition, companies need to

understand customers, and to quickly respond to their preferences and needs. 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 [5] 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 relationships must be promoted for airlines to retain aggressive benefit and success in the long term. 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 [4]. A segmentation example in the airline business is one that defines business travelers versus leisure travelers for the purpose of developing schedules and pricing policies.

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 [7]; (b) classifying trip purposes into leisure, business, etc… [6]; (c) addressing airline ticket prices behavior over time [3].

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; using our CRM recommendations we can give the airline a competitive advantage; (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.

This paper is organized as follows. Section 2 describes the problem and the data. Section 3 discusses the solution strategy. Section 4 presents the experimental results which are discussed in Section 5. Section 6 presents a summary of our CRM recommendations.

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 allows getting a better understanding of customer types and behaviors. The key program features are mileage accumulation (members can earn miles for air travel, but also for activities like hotel stays, and credit card usage) and mileage redemption (members can spend miles for air travel). The currency of such a program is miles. The program intends to identify high value customers and provide them with special services and benefits such as upgrades.

In our 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 4 categories of members: Basic member (used for new customer or customer with qualified miles less than 20,000); Prestige member (used for customer with 20,000 qualified miles); Elite member (used for customer with 40,000 qualified miles) and Prestige Club member (used for special customer identified by top management independent from any criteria).

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. The variables chosen for this study include: mileage, passenger membership, number of services used over the passenger membership period, and number of services used in the recent year, and other business variables. Other variables are calculated by merging transaction data to each passenger record.

2.2 Problem Description

The market experts concerns are the key business processes for customer management. Key business processes typically include customer value management, customer retention, customer growth, customer acquisition, customer communication and multi-channel optimization. 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 this data to divide customers into clusters based on variables such as current customer profitability, some measure of risk, a measure of the lifetime value of a customer, and retention probability. Creating clusters based on such variables highlights marketing opportunities. Cross-selling (selling new products) and up-selling (selling more of what customers currently buy) are the marketing initiatives of choice.

Data understanding

Our source of data is the frequent flyer program database. The target population contains only the members of this program.

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, which is mandatory for the clustering algorithms. The queries presented below illustrate the manipulation done on each transaction data. It includes pivoting, aggregating, and inserting into each passenger record.

The first query (Q1) is based on "Individuals" and "Activities" tables. It groups customer data with activities types; identifies the "Financial", "Flight", and "Hotel" activities; gives the total mileage of each activities per customer; and prepares the calculation of membership time per month. It includes 174,900 records.

The second query (Q2) is based on Q1. It groups customer data; calculates the "Financial", "Flight", and "Hotel" services used by the customer during his lifetime; gives the total mileage per customer; and finalizes the membership time per month. It includes 79,782 records (Record for each customer).

The third query (Q3) is based on "Individuals" and "Activities" tables. It groups customer data with activities types; and identifies the "Financial", "Flight" and "Hotel" activities done during the last year (2005). It includes 200,243 records.

The fourth query (Q4) based on Q3. It groups customer data; and calculates the "Financial", "Flight", and "Hotel" services used by the customer during the last year. It includes 79,782 records (Record for each customer).

The fifth query (Behavioral Activities) based on Q2 and Q4. It includes the Customer ID, First Name and Last Name; calculates the "Financial", "Flight", and "Hotel" services used by the customer during his lifetime; computes the "Financial", "Flight", and "Hotel" services used by the customer during the last year; calculates the revenue mileage, membership period, Revenue mileage

per membership period, and “Financial”, “Flight”, “Hotel” services used by the customer during his lifetime per membership. It includes 79,782 records (Record for each customer).

We discard customer records with missing values. The records remaining are 50,830 records.

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 association rules

The result generated by the clustering provides customer segmentation with respect to important dimensions of customers’ needs and value. One of this segment identified Frequent Flyer best customers. Two different approaches have been used for Association Rules application. Each approach is based on different data. Below we describe the data used in both approaches:

Approach Based on Original Activities

As mentioned in the clustering process; the “Flight”, “Financial”, and “Hotel” activities are used as services purchased by customers. A query (Q5) based on Q1. It includes the Customer ID, and the “Financial”, “Flight”, and “Hotel” services used by the customer during his lifetime. It groups all the Frequent Flyer Customer information. It includes 52,338 records.

Query (Q6) based on a selected Cluster and Q5. It includes the Customer ID, and the “Financial”, “Flight”, and “Hotel” services used only by the selected Cluster customers. It groups best customers information. It includes 3,788 records.

Using the pivot table function on Q6, we can rotate its rows and columns to see different summaries of the source data (Original Activities of selected Cluster). It includes 1,886 records (Record for each one from our best customers). In one record, we can found the customer ID, and for each activities (Flight, Financial, or Hotel); if it is used then a “1” will be associated to the field; otherwise it will be “0”.

Approach Based on Flight Activities Only

In the second approach, we consider from our best customer (Selected Cluster) only the Flight activity studying and analyzing the sector used taking into account that the original have to be one of our main Hubs. A query (Q7) based on "Activities" table. It 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. It includes 139,708 records.

Query (Q8) 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. It includes 10,828 records.

Using the pivot table function on Q8, we can rotate its rows and columns to see different summaries of the source data (Activities Selected Cluster). It includes 1,886 records (Record for each one from our best customers). In one record, we can found the customer ID, and for each sector; if it is used then a "1" will be associated to the field; otherwise it will be "0". The Cluster 16 customers have used 145 sectors.

Model building and evaluation

Using a data mining tool, we apply clustering and association rules techniques [2] in order to generate new 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. It is then possible to target those customers to show the desired behavior (such as buying a service) based on a predictive model.

The clustering techniques separate the data into subgroups or classes that share common characteristics. We are interested in learning more about loyal customers. Then, the company might infer what could be done to keep customers loyal. The data mining shall perform hierarchical clustering using an enhanced version of the k-means algorithm and O-Cluster. A cluster is

characterized by its centroid, attribute histograms, and place in clustering model hierarchical tree. The cluster centroid is the vector that encodes, for each attribute, either the mean (if the attribute is numerical) or the mode (if the attribute is categorical) of the cases in the build data assigned to a cluster. Clusters discovered by k-means or O-cluster algorithms are used to create rules that capture main characteristics of data assigned to each cluster. Clusters are also used to generate probability model, which is used during scoring for assigning data points to clusters.

The enhanced K-Means is based on the standard k-means algorithm. It relies on a distance metric (function) to measure the similarity ("closeness") between data points. The selected distance metric is Cosine. A hierarchical version of enhanced k-means algorithm is implemented. Unbalanced trees are built. The tree can grow one node at a time. The node with the largest distortion (Sum of distance to the node's centroid) is split to increase the size of the tree until the desired number of clusters is reached.

Association rules capture the frequent co-occurrence of items or events in large volumes of customer transaction data. Finding such rules is invaluable for cross-marketing and mail order promotion. We use the Apriori algorithm for association rules mining [2]. Association rules are mined within selected clusters in order to derive some CRM suggestions.

4 Experimental Results

4.1 Clustering

4.1.1 Empirical Procedure

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 scenario has been tested. For brevity, we present only a limited set of results. The algorithms are applied on "Behavioral Clustering" query including 50,830 records.

4.1.2 Input Variables

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).

4.1.3 K-Means Algorithm Results

The basic parameters available for k-means clustering include:

- Maximum number of clusters. We specify the maximum number of clusters allowed; the algorithm may come up with fewer. The default value is 4.
- Maximum iterations or Maximum number of passes through the data. This parameter indicates the maximum number of times the algorithm will read the data. The longer the algorithm will run, and the more accurate the result will be. This parameter is a stopping criterion for the algorithm. It must be between 2 (slow build) and 30 (fast build). The default is 6.
- Minimum Error Tolerance. It must be between 0.001 (slow build) and 0.1 (fast build). The default value is 0.005. 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. In Table 1, the following information is displayed about the cluster: Cluster ID, Cluster Level, Record Count (the number of records or cases in the cluster), the attributes in the cluster, and the Centroid Value.

Table 1: Clusters Details of K-Means Algorithm

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
9	4	9,066	ACTLASTYEAR	1.02-1.05	MILEAGE	3740.78-12616.56
			ACTLIFE	1.02-1.05	RAM	0.1273-0.134
			MEMBERSHIP	9.2-9.92	RMM	1210.9099-1671.2133
11	4	6,177	ACTLASTYEAR	0.99-1.02	MILEAGE	21492.34-30368.12
			ACTLIFE	0.99-1.02	RAM	0.0201-0.0268
			MEMBERSHIP	43.04-43.76	RMM	290.3033-750.6066
12	4	3,981	ACTLASTYEAR	0.0-0.03	MILEAGE	3740.78-12616.56
			ACTLIFE	0.93-0.96	RAM	0.0201-0.0268
			MEMBERSHIP	38.72-39.44	RMM	-170-290.3033
13	4	3,676	ACTLASTYEAR	0.0-0.03	MILEAGE	3740.78-12616.56
			ACTLIFE	0.99-1.02	RAM	0.0402-0.0469
			MEMBERSHIP	21.44-22.16	RMM	290.3033-750.6066
14	4	5,538	ACTLASTYEAR	0.0-0.03	MILEAGE	12616.56-21492.34
			ACTLIFE	0.93-0.96	RAM	0.0067-0.0134
			MEMBERSHIP	68.96-69.68	RMM	-170-290.3033
15	4	2,965	ACTLASTYEAR	0.0-0.03	MILEAGE	12616.56-21492.34
			ACTLIFE	0.99-1.02	RAM	0.0134-0.0201
			MEMBERSHIP	53.84-54.56	RMM	-460.3033
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
17	5	8,774	ACTLASTYEAR	0.99-1.02	MILEAGE	39243.9-48119.68
			ACTLIFE	0.99-1.02	RAM	0.0067-0.0134
			MEMBERSHIP	67.52-68.24	RMM	290.3033-750.6066

4.2 Association Rules

4.2.1 Empirical Procedure

In ODM, we use an SQL-based implementation of the Apriori algorithm. The candidate generation and support counting steps are implemented using SQL queries. The result generated by k-means

clustering best scenario will be 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. The Support of a rule is a measure of how frequently the items involved in it occur together. The Confidence of a rule is the conditional probability of consequent given the antecedent. Support and confidence can be used to rank the rules and hence the predictions.

4.2.2 Input Variables

K-means algorithm scenario divides the customers into 9 clusters. For the association rules study, 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”). A sample of the results, for some customers, is given in Table 2. The “Activities Cluster 16” query includes Customer ID and 145 fields including the Name of Sectors used by customers and originated from the main airline Hubs.

Table 2: Sample of "Original Activities Cluster 16" Query

Cust. ID	Financial	Flight	Hotel
206	1	1	1
486	1	1	0
906	0	1	1
1352	0	1	1

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 manipulate only the sectors that have a percentage of use greater than 20%.

4.2.3 Apriori Algorithm Results

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 default algorithm settings are as follows:

- Minimum support: Support of a rule is a measure of how frequently the items involved in it occur together; we set it to 0.1.
- Minimum confidence: Confidence of a rule is the conditional probability of a consequent given the antecedent; we set it to 0.5.
- Number of attributes in each rule: is number between 2 and 100 that specifies the maximum number of attributes in each rule; we set it to 3.

4.2.3.1 Scenario 1

The run is based on "Original Activities Cluster 16" query. Table 3 displays the rules with support and confidence for each rule.

4.2.3.2 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. This scenario leads to 2,082 rules. Table 4 displays some significant rules with support and confidence for each rule.

Table 3: Association Rules for Best Customers Activities (Scenario 1)

Rule Id	If (condition)	Then (association)	Confidence	Support
4	FINANCIAL=1	FLIGHT=1	1	0.92099684
3	FLIGHT=1 and HOTEL=0	FINANCIAL=1	1	0.91251326
2	FINANCIAL=1 and HOTEL=0	FLIGHT=1	1	0.91251326
8	HOTEL=0	FINANCIAL=1	1	0.91251326
9	HOTEL=0	FLIGHT=1	1	0.91251326
1	FINANCIAL=1 and FLIGHT=1	HOTEL=0	0.9907887	0.91251326
5	FINANCIAL=1	HOTEL=0	0.9907887	0.91251326
6	FLIGHT=1	FINANCIAL=1	0.92099684	0.92099684
7	FLIGHT=1	HOTEL=0	0.91251326	0.91251326

Table 4: 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
495	BEYCAI=1 and BEYCDG=1	BEYAMM=1	0.5307517	0.12479914
494	BEYAMM=1 and BEYCDG=1	BEYCAI=1	0.6005155	0.12479914
497	BEYAMM=1 and BEYDXB=1	BEYCAI=1	0.58469945	0.11462239
96	BEYAMM=1	BEYCAI=1	0.5473888	0.15158008
1303	BEYDXB=1 and BEYRUH=1	BEYCDG=1	0.84347826	0.103910014
1297	BEYDXB=1 and BEYJED=1	BEYCDG=1	0.82520324	0.108730584
493	BEYAMM=1 and BEYCAI=1	BEYCDG=1	0.8233216	0.12479914
138	BEYFCO=1	BEYCDG=1	0.82287824	0.119442955
1228	BEYCAI=1 and BEYDXB=1	BEYCDG=1	0.8157182	0.1612212
1419	BEYDXB=1 and BEYLHR=1	BEYCDG=1	0.8051118	0.1349759
647	BEYAMM=1 and BEYDXB=1	BEYCDG=1	0.7978142	0.15640064
60	BEYGVA=1	BEYCDG=1	0.78039217	0.10658811
141	BEYJED=1	BEYCDG=1	0.7751323	0.15693626
63	BEYIST=1	BEYCDG=1	0.77260274	0.15104446
68	BEYKWI=1	BEYCDG=1	0.7615894	0.12319229
16	BEYAMM=1	BEYCDG=1	0.7504836	0.20782003
131	BEYCAI=1	BEYCDG=1	0.7428088	0.23513658
142	BEYLCA=1	BEYCDG=1	0.7237569	0.14033209
57	BEYDXB=1	BEYCDG=1	0.71935856	0.3363685
72	BEYRUH=1	BEYCDG=1	0.71780825	0.14033209
71	BEYLHR=1	BEYCDG=1	0.7161172	0.2094269
646	BEYAMM=1 and BEYCDG=1	BEYDXB=1	0.7525773	0.15640064
1302	BEYCDG=1 and BEYRUH=1	BEYDXB=1	0.740458	0.103910014
161	BEYKWI=1	BEYDXB=1	0.7086093	0.11462239
97	BEYAMM=1	BEYDXB=1	0.7079304	0.19603643
1296	BEYCDG=1 and BEYJED=1	BEYDXB=1	0.69283277	0.108730584
1227	BEYCAI=1 and BEYCDG=1	BEYDXB=1	0.6856492	0.1612212
160	BEYJED=1	BEYDXB=1	0.6507937	0.13176219
1418	BEYCDG=1 and BEYLHR=1	BEYDXB=1	0.64450127	0.1349759
163	BEYRUH=1	BEYDXB=1	0.63013697	0.12319229

5 Discussion Of Results

5.1 K-Means Clustering

Table 5 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 5, 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 5. 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.

Table 5: Clustering Analysis for K-Means Algorithm

(Average Number of Services used = Sum of Activities used over lifetime / Number of Customers)

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

Best Customers:	Clusters 11 - 16 -17 (Higher Mileage per Customer)

Strategies:
Retention strategy for best customers.
Cross-Sell strategy for clusters 8 by contrasting with cluster 16 (Service Index is close. By comparing which services is used by the best customers)
Cross-Sell strategy for clusters 15 by contrasting with cluster 11 (Service Index is close. By comparing which services is used by the best customers)
Cross-Sell strategy for clusters 14 - 13 by contrasting with cluster 17 (Service Index is close. By comparing which services is used by the best customers)
Cluster 9 to wait (New Customers)
Cluster 12 (The Worst Cluster)

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. Table 6 shows some of the best routes (and some percentages) deduced from the results of the k-means algorithm.

Table 6: Best Route Originated from CDG (with K-Means Algorithm)

CDGBES	0.0829		CDGAMM	0.3314		CDGCPH	1.242751
CDGBKK			CDGBUD			CDGSFO	
CDGBOG			CDGGOT			CDGVIE	
CDGBZV			CDGHAJ				
CDGCCS			CDGZYR			CDGLYS	1.325601
CDGCFE							
CDGDLA			CDGABJ	0.41425		CDGORD	1.574151
CDGDTW			CDGCKY				
CDGEZE			CDGLIN			CDGDXB	1.657001
CDGFIH			CDGOSL			CDGIAH	
CDGFNI			CDGSXB			CDGMXP	
CDGFRL			CDGTRN				
CDGGIG						CDGATL	1.739851
CDGHKG			CDGATH	0.4971		CDGFCO	
CDGKWI			CDGCGN			CDGLIS	
CDGLBV			CDGCVG				
CDGLED			CDGMPL			CDGBOS	1.822701
CDGNAP			CDGPHL				
CDGNKC						CDGFRA	1.905551
CDGNSI			CDGTXL	1.159901			
CDGOUA			CDGYYZ			CDGAMS	2.071251
CDGPHC			CDGZRH				

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. This result has been confirmed. 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. The first one dedicated to Flight/Financial customers, recommending "hotel" special offer. The second dedicated to Flight/Hotel customers, recommending "financial" special offer.

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

Our objective is to manage the customer information. The results obtained reveal different customer groups. The best scenario for clustering using k-means algorithm generates 9 different clusters with specific profile for each one. These clusters allow the airline to generate revenue from customer's business. 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 with the higher revenue mileage per customer. A retention strategy should be applied to these customers. It is possible, for example, to recognize an individual in those clusters who usually travels once a month and has not showed up for three months. The sales specialist could contact this customer to check the reason for his behavior change and try to rectify the situation in order to retain such valuable customer. 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. Cross-selling is applied to increase the number of services purchased. With little effort, we might convert customers from mid-range clusters to best customers clusters.

The third type of clusters identified in this study is new customer cluster. The recommendation is to observe these customers to determine their behavior. The marketing of available services to this group will be useful 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 any marketing campaign 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. It personalizes the passenger's interaction with services.

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. We wish to thank the anonymous reviewer whose comments improved the presentation of the paper.

References

[1] Alaska Airlines soars in Meeting the Needs of More than 17 Million Customers Annually. (2005). Siebel Systems, Inc. http://www.siebel.com/downloads/case_studies/alaska_air.pdf

[2] Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics.* Prentice Hall.

[3] Etzioni, O.; Knoblock, C.; Tuchinda, R.; & Yales, A. (2003). To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price. http://www.isi.edu/integration/papers/etzioni03-kdd.pdf

[4] Fennell, G.; & Allenby, G. (2004). Market definition, market segmentation, and brand positioning create a powerful combination. http://fisher.osu.edu/~allenby_1/2004%20Integrated%20Approach.pdf

[5] Lee, D. (1999). CRM Definitions. CRM.Talk #054. http://www.crmguru.com/content/crmtalk/2000a/crmt054.htm#1

[6] Pritscher, L.; & Feyen, H.; (2001). Data Mining and Strategic Marketing in the Airline Industry. Atraxis AG, Swissair Group, Data Mining and Analysis, CKCB. http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/pritscher1.pdf

[7] Ramachandran, P. (2001). Mining for Gold. WIPRO Technologies. http://www.wipro.com/whitepapers/services/businessintelligence/dataminingmininggold.htm