

Rt
416
c.1

MINING AIRLINE DATA FOR CRM STRATEGIES

by

LENA MAALOUF

Maitrise, Computer Science, Lebanese University, 1995

Thesis submitted in partial fulfillment of the requirements for the Degree of Master of
Science in Computer Science

Division of Computer Science and Mathematics

LEBANESE AMERICAN UNIVERSITY

JUNE 2006



Thesis approval Form

Student Name : Lena Maalouf I.D.: 200202846


Thesis Title : Mining Airline Data for CRM Strategies


Program : M.S. in Computer Science


Division/Dept : Computer Science and Mathematics

School : Arts and Sciences - Beirut

Approved by :


Nashat Mansour, Ph.D. (Advisor)
Professor of Computer Science, LAU


Faisal Abu Khzam, Ph.D.
Assistant Professor of Computer Science, LAU


Toufic Mezher, Ph.D.
Professor of Engineering Management, AUB

Date : June 27, 2006

Plagiarism Policy Compliance Statement

I certify that I have read and understood LAU's Plagiarism Policy. I understand that failure to comply with this Policy can lead to academic and disciplinary actions against me.

This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that by acknowledging its sources.

Name: Lena Maalouf

Signature:



Date: 29-Jun.-2006

I grant to the LEBANESE AMERICAN UNIVERSITY the right to use this work, irrespective of any copyright, for the University's own purpose without cost to the University or its students and employees. I further agree that the University may reproduce and provide single copies of the work to the public for the cost of reproduction.

The work is dedicated to my family, who teach me to love hard work to achieve valued ambitions. Thank you for your love, inspiration, and encouragement.

Acknowledgment

I would like to thank my advisor Dr. Nashat Mansour for his guidance throughout my M.S. studies.

Middle East Airlines is the solo owner of the data used in this study. I would like to express my sincere thanks to all who have help me, in particular my colleagues at IT Department – Middle East Airlines who provided me with full support. Especially, Adib Charif – Head of IT Department for his enthusiasm, patience, and support.

Finally, I would like to thank my friends and family for their long support.

ABSTRACT

In today's competitive climate, Customer Relationship Management (CRM) has become an essential component in the airline business strategies. Building CRM in the airline industry requires a comprehensive view of customer behavior. This view has to be based on analyzing customer data in order to understand customer preferences and learn from his/her behavior.

In this thesis, 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. Furthermore, we explore the characteristics of data mining algorithms for this application and uncover relative merits of the algorithm employed.

CONTENTS

<i>List of Figures</i>	<i>xi</i>
<i>List of Tables</i>	<i>xii</i>
<i>Abbreviations</i>	<i>xv</i>
1. Introduction	1
2. Background	5
2.1 Cross-Industry Standard Process for Data Mining (CRISP-DM).....	5
2.1.1 Phase One: Business Understanding.....	6
2.1.2 Phase Two: Data Understanding	6
2.1.3 Phase Three: Data Preparation.....	7
2.1.4 Phase Four: Modeling	7
2.1.5 Phase Five: Evaluation.....	7
2.1.6 Phase Six: Deployment.....	7
2.2 Data Mining and Business Intelligence	8
2.3 Related Work.....	9
3. Problem and Data Description	11
3.1 Customer Segmentation Study – Data Description.....	11
3.1.1 Frequent Flyer Concept.....	11
3.1.2 Cedar Miles Services.....	11
3.1.3 Membership Categories	12
3.1.4 Customer Activities	12
3.1.5 Data Model.....	12
3.2 Customer Segmentation Study – Problem Description.....	15
3.2.1 Customer Value Measurement	15
3.2.2 Customer Retention	15
3.2.3 Customer Growth.....	15
3.2.4 Customer Acquisition	16
4 – Solution Strategy	17
4.1 Overview.....	17
4.2 CRISP Implementation.....	17
4.2.1 Business Goals.....	17
4.2.2 Data Mining Goals	17
4.2.3 Data Understanding	18
4.2.4 Data Transformation and Aggregation for Clustering	18
4.2.5 Data Preparation	22
4.2.6 Data Transformation and Aggregation for Association Rules.....	23
4.2.6.1 Based on Original Activities.....	23
4.2.6.2 Based on Flight Activities Only.....	24
4.2.7 Model Building and Evaluation.....	24

4.3	Data Mining Techniques – Algorithms	25
4.3.1	Clustering.....	25
4.3.1.1	Enhanced K-Means	26
4.3.1.2	O-Cluster	27
4.3.1.3	Expectation-Maximization (EM)	30
4.3.1.4	COBWEB Algorithm.....	31
4.3.2	Association Rules	32
4.3.2.1	APRORI Algorithm	32
4.3.2.2	PREDICTIVEAPRIORI Algorithm	33
5	– Experimental Results	36
5.1	Overview.....	36
5.2	Hardware and Software Platform.....	36
5.3	Clustering	37
5.3.1	Empirical Procedure.....	37
5.3.2	Input Variables	38
5.3.3	K-Means Algorithm Results	38
5.3.3.1	Algorithm Parameters	38
5.3.3.2	Scenario 1	39
5.3.3.3	Scenario 2	42
5.3.3.4	Scenario 3	45
5.3.3.5	Scenario 4	48
5.3.3.6	Scenario 5	51
5.3.3.7	Scenario 6	52
5.3.4	O-Cluster Algorithm Results	56
5.3.4.1	Algorithm Parameters	56
5.3.4.2	Scenario.....	56
5.3.5	EM Algorithm Results	60
5.3.5.1	Algorithm Parameters	60
5.3.5.2	Scenario.....	61
5.3.6	COBWEB Algorithm Results	63
5.3.6.1	Algorithm Parameters	63
5.3.6.2	Scenario.....	63
5.4	Association Rules.....	63
5.4.1	Procedure.....	63
5.4.2	Scoring (Applying Models).....	64
5.4.3	Input Variables	66
5.4.3.1	“Original Activities Cluster 16” Query	66
5.4.3.2	“Activities Cluster 16” Query	68
5.4.4	Apriori Algorithm Results	68
5.4.4.1	Algorithm Parameters	68
5.4.4.2	Scenario 1	69
5.4.4.3	Scenario 2	70
5.4.4.4	Scenario 3	74
5.4.5	Predictive Apriori Algorithm Results.....	76
5.4.5.1	Algorithm Parameters	76
5.4.5.2	Scenario 1	76
6	– Discussion of Results	78

6.1	Overview	78
6.2	Discussion of K-Means Clustering Results	78
6.2.1	Scenario 1 Result	79
6.2.2	Scenario 2 Result	81
6.2.3	Scenario 3 Result	83
6.2.4	Scenario 4 Result	86
6.2.5	Comparison between Different Scenarios	88
6.3	Discussion of O-Cluster Clustering Results	91
6.3.1	Scenario Result	91
6.3.2	Scoring (Apply Result)	94
6.3.3	Comparison O-Cluster to K-Means Scenario 3	94
6.3.4	Best Route from CDG	97
6.4	Discussion of EM Clustering Results	100
6.4.1	Scenario Result	100
6.5	Comparison of Clustering Algorithm	103
6.6	Discussion of Association Rules Results	105
6.6.1	Scenario 1	105
6.6.2	Scenario 2	106
6.6.3	Scenario 3	106
6.7	Comparison of Association Rules Algorithm	107
6.8	Summary of CRM Recommendations	108
7	– Conclusion and Further Work	110
7.1	Conclusion	110
7.2	Future Work	112
	References	114
	Appendix	116

LIST OF FIGURES

Figure 2.1 Phases of the CRISP-DM Reference Model	6
Figure 3.1 Frequent Flyer Case Study Data Model	14
Figure 4.1 K-Means Algorithm	26
Figure 4.2 O-Cluster Algorithm Block Diagram	30
Figure 4.3 EM Algorithm	31
Figure 4.4 PREDICTIVEAPRIORI Algorithm	34
Figure 4.5 Generation of Rules Procedure	35
Figure 5.1 WEKA Preprocess View	61

LIST OF TABLES

Table 4.1 Sample of “Activities” Table	19
Table 4.2 Sample of “Individuals” Table	20
Table 4.3 Sample of “Behavioral Activities” Query	22
Table 5.1 K-Means Algorithm Rules (Scenario 1)	40
Table 5.2 Clusters Details of K-Means Algorithm (Scenario 1)	41
Table 5.3 K-Means Algorithm Rules (Scenario 2)	43
Table 5.4 Clusters Details of K-Means Algorithm (Scenario 2)	44
Table 5.5 K-Means Algorithm Rules (Scenario 3)	46
Table 5.6 Clusters Details of K-Means Algorithm (Scenario 3)	47
Table 5.7 K-Means Algorithm Rules (Scenario 4)	49
Table 5.8 Clusters Details of K-Means Algorithm (Scenario 4)	50
Table 5.9 Parameters Change Results	52
Table 5.10 Comparison between Different Values of Minimum Error Tolerance	52
Table 5.11 K-Means Algorithm Rules for Partitioned Data	54
Table 5.12 Comparison between Scenario 3 and Scenario 6	55
Table 5.13 O-Cluster Algorithm Rules	57
Table 5.14 Clusters Details of O-Cluster Algorithm	58

Table 5.15 Centroid Value of O-Cluster Algorithm	59
Table 5.16 EM Algorithm Result	62
Table 5.17 K-Means Algorithm Scoring Sample (Scenario 3)	65
Table 5.18 Sample of "Original Activities Cluster 16" Query	67
Table 5.19 Association Rules for Best Customers Activities (Scenario 1)	70
Table 5.20 Association Rules for Best Customers Activities (Scenario 2)	71
Table 5.21 Association Rules for Best Customers Activities (Scenario 2)	73
Table 5.22 Association Rules for Best Customers Activities (Scenario 3)	75
Table 5.23 Association Rules for Best Customers Activities (Scenario 3)	75
Table 5.24 Predictive Apriori Result (Scenario 1)	77
Table 6.1 Clustering Analyst for K-Means Algorithm (Scenario 1)	80
Table 6.2 Clustering Analyst for K-Means Algorithm (Scenario 2)	82
Table 6.3 Clustering Analyst for K-Means Algorithm (Scenario 3)	85
Table 6.4 Clustering Analyst for K-Means Algorithm (Scenario 4)	87
Table 6.5 Cluster 16 Sample	89
Table 6.6 Cluster 12 Sample	90
Table 6.7 Clustering Analyst for O-Cluster Algorithm	93
Table 6.8 Sample of Applying O-Cluster Algorithm	95
Table 6.9 Comparison between K-Means (Scenario 3) and O-Cluster Result	96

Table 6.10 Best Route Originated from CDG with K-Means Algorithm (Scenario 3)	98
Table 6.11 Best Route Originated from CDG with O-Cluster Algorithm	99
Table 6.12 Best Route Comparison between K-Means (Scenario 3) and O-Cluster Result	100
Table 6.13 Clustering Analyst for EM Algorithm	102

ABBREVIATIONS

ACTLASTYEAR	Number of Services (“Financial”, “Flight”, or “Hotel”) the customer used in the last 12 months
ACTLIFE	Number of Services (“Financial”, “Flight”, or “Hotel”) the customer used over lifetime
B	Basic Member
BI	Business Intelligence
C	Prestige Club Member
CRM	Customer Relationship Management
CRISP-DM	Cross-Industry Standard Process for Data Mining
E	Elite Member
EM	Expectation-Maximization
ETL	Extraction, Transformation, and Loading
MEA	Middle East Airlines
ODM	Oracle Data Miner
OLAP	Online Analytical Processing
P	Prestige Member
RAM	Number of Services over Lifetime / Membership Period
RMM	Revenue Mileage / Membership Period

SPA Special Prorate Arrangements

WEKA Waikato Environment for Knowledge Analysis

Chapter 1

INTRODUCTION

The airline industry has been exposed to many challenges due to the changes in customer behavior, competition and technology. By considering these factors, airline can identify, develop, and implement business strategies.

The changes in customer behavior are due to demography and customer expectations. Demography covers the distribution of income and ages. The huge amounts of information and the availability of communication technologies provides customer the power to access information on competitors, products, availability, and prices made possible. Switch between competitors is easy thru Internet.

Due to those effects, business has 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 analyze their markets to identify the most valuable customers and the appropriate strategies to use in developing relationships with these customers. Such strategies include the developing of one-to-one relationship with customers using market segmentation and Customer Relationship Management (CRM).

In a recent study (Boland, Morrison, and O'Neill, 2002) see that airline industry stands on a crossroad. Economic crash and result of September 11th attacks impacted airline economics. The focus was on reducing costs, but we cannot ignore the customer. Customer relationships must be promoted for airlines to retain aggressive benefit and success in the long term.

We have two different categories of CRM; Operational CRM and Analytical CRM. Operational conducts the top line and analytical conducts the bottom line of business. Operational CRM alone is not enough; it must be based on business analytics capable of increasing profitability and impact business model. The airline should apply assessment,

acquisition, and customer management analytics. Assessment measures customer's value. Customer acquisition deals with profiling, segmentation, and ranking of customers based on tendency to buy, order frequency, and purchasing behavior. Customer management determines the impacts of order fulfillment, returns, and call center activity on actual sale performance.

To implement CRM, we have to analyze the customer behavior. Based on the result, new customer-centric strategies are implemented. The Airline data used consists of Frequent Flyer database. Any decisions require manually processing of huge data. So often, airlines use methods based on human expertise.

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 (Fennell and Allenby, 2004). A segmentation example in the airline business focuses on defining business travelers versus leisure travelers for the purpose of developing schedules and pricing policies.

Lee (1999) 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, and raising significantly an airline's efficiency and effectiveness.

Previous works in this field are minimal. Those works focus on proposing technique and describing the results briefly. The technique used is only clustering. All of them have used mainly the frequent flyer data. Details are presented in Chapter 2. The objective of previous works that have used data mining for frequent flyer airline data have been:

- a.** Categorizing customer into groups based on sectors most frequently flown, class flown, period of year, hometown compared to sector flown (Ramachandran, 2001).
- b.** Classifying trip purposes into leisure, business, etc... (Pritscher and Feyen, 2002).

Our objective is to explore the Frequent Flyer database using data mining methods in order to prepare for CRM implementation. To enable CRM, the first task is to identify market segments containing customers with high profit potential. We apply clustering and association rules data mining methods. During our work we have used the Cross-Industry Standard Process for Data Mining (CRISP-DM) process cycle.

Our contribution in this thesis 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. For example, one major objective is to classify customers based on services used by passenger over lifetime, services used by customer in the last year, passenger's mileage over lifetime, passenger membership period in months, mileage / membership period, and number of services over lifetime / membership period. The services considered are "Financial" (Credit Card), "Flight", and "Hotel". Based on the application of these selected DM techniques, we derived some CRM strategies; one example of these strategies is to adopt refrain strategy for the group with low spending tendency.
- b.** We have compared a few data mining algorithms and drew conclusions about the quality of the solutions produced.
- c.** We proposed preprocessing technique for processing the huge amount of data for a feasible application of DM techniques.
- d.** We used real data from MEA and conducted experimental work for validating our techniques. In these experiments, we included design decisions to optimize the operation of the Data Mining algorithms. Also we validated the clustering results by reapplying clustering again on a part of the data and using the remaining part for evaluation.

Chapter 2 discusses the background covering CRISP-DM, Business Intelligence and Data Mining. Chapter 3 provides an overview of the problem and data description. Chapter 4 discusses the solution strategy. Chapter 5 offers the experimental result. However, the result

will be discussed in Chapter 6. Finally, we have in Chapter 7 the conclusion and propose the future work.

Chapter 2

BACKGROUND

2.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

In 2000, CRISP_DM version 1.0 was introduced by industry leaders reflecting significant progress in the development of a standardized data processing model (Chapman et al., 2000).

CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, & deployment. These phases help organizations understand the data mining process & provide a road map to follow while planning a data mining project.

Chapman et al., shows Figure 2.1 with data mining process phases. The arrows indicate the dependencies between the phases, while the outer circle symbolizes the cyclical nature of data mining.

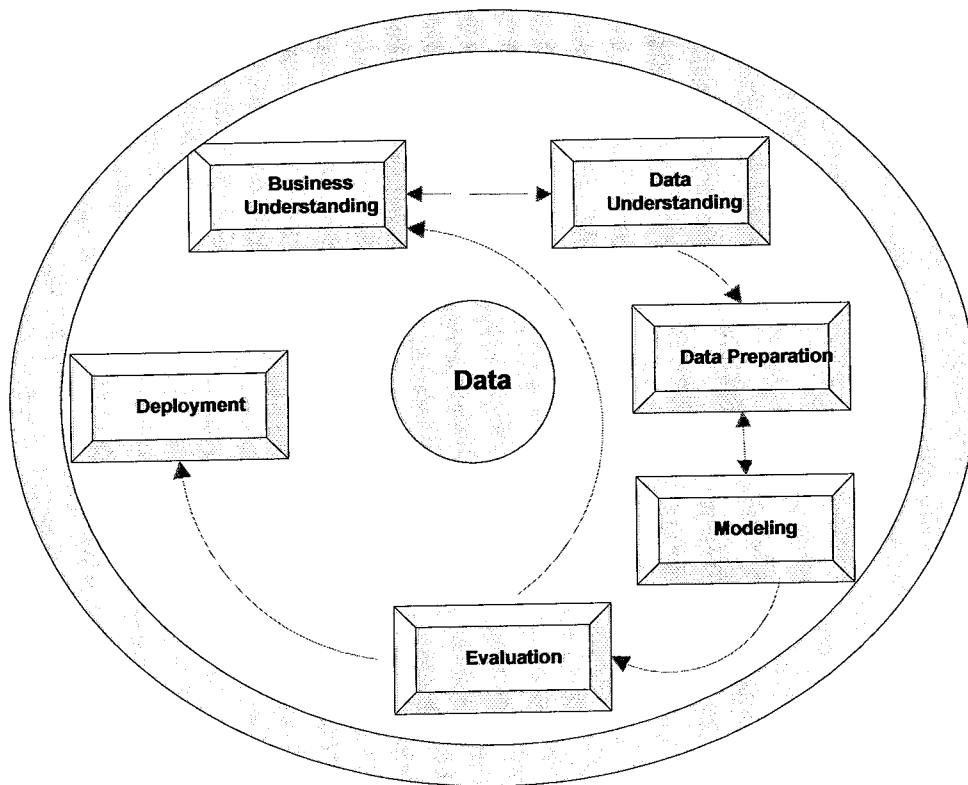


Figure 2.1 - Phases of the CRISP-DM Reference Model

2.1.1 Phase One: Business Understanding

The business understanding phase focuses on understanding the project objectives from a business perspective. This knowledge is converted into a data mining problem definition. Then a preliminary plan is developed. This plan is designed to achieve the objectives. It involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan.

2.1.2 Phase Two: Data Understanding

The data understanding phase starts with data collection. This phase permits to increase familiarity with the data, to identify data quality problems, or to detect interesting

subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality.

2.1.3 Phase Three: Data Preparation

The data preparation phase covers all activities to build the final data set or the data that will be fed into the modeling tool from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data.

2.1.4 Phase Four: Modeling

In this phase, different modeling techniques are selected and applied and their parameters are adjusted to best values. Several techniques exist for the same data mining problem type. Therefore, returning to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models.

2.1.5 Phase Five: Evaluation

Before proceeding to final deployment of the model built, the model is evaluated. The model's construction is reviewed to be certain it achieves the business objectives. At the end of this phase, the analyst should decide how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps.

2.1.6 Phase Six: Deployment

The knowledge gained must be organized and presented in a usable way. As final result, the analyst has the actions that must be taken in order to make use of the created models. The key steps here are plan deployment, plan monitoring and maintenance, the production of the final report, and review of the project.

2.2 Data Mining and Business Intelligence

Business Intelligence is very critical; Adelman, Miss, and Abai (2005) states the importance of collecting facts correctly otherwise it will conduct us in a wrong direction. We can find different definitions for Business Intelligence (BI). BI provides a 360-degree view of business, enabling the organizational decision makers to make faster and more reliable decisions.

BI applications allow uncovering abnormalities in the business. The benefits realized from BI include increase in revenue and profits; fraud and abuse detection; competitive advantage due to accurate information; and better relationships with customers. CRM software has to utilize BI tools on a variety of issues as customer segmentation. The basic components of BI are Data Warehouse, Enterprise Data Warehouse, Operational Data Store, Data Mart, Metadata Repository, Data Transformation and Cleansing, Online Analytical Processing (OLAP) and Analytics, and Data Presentation and Visualization. In addition to the basic components of BI, there are many other important tools such as Data Mining, or Knowledge Discovery.

The key is the data. Mining without data is only opinions. Data mining is one of the key supporting technologies for Business Intelligence and CRM. Data mining is the process of using computer models and algorithms against internal and external organizational data to find hidden patterns in organization data. Data Mining is designed to either predict or discover undetected behavior. In predictive data mining, the algorithm predicts the behavior of an entity, person, or object based on given parameters and previously recorded data. Discovery algorithms, allow organizations to identify patterns, exceptions and deviations in data that are not evident to business analysts.

Linoff (2004) describes the data mining usage. Data mining can help spot sales trends, develop smarter marketing campaigns, and predict customer loyalty. Specific uses of data mining include market segmentation (Identify the common characteristics of customers who buy the same products from our company); customer churn (Predict which customers are likely to leave our company and go to a competitor); fraud detection (Identify which transactions are most likely to be fraudulent); direct marketing (Identify which prospects

should be included in a mailing list to obtain the highest response rate); interactive marketing (Predict what each individual accessing a Web site is most likely interested in seeing); market basket analysis (Understand what products or services are commonly purchased together); and trend analysis (Reveal the difference between a typical customer this month and last).

There are a series of steps involved in data mining: getting the data organized, determining the desired outcomes, selecting tools, carrying out the mining, reducing the results so that only the useful ones are considered further, taking actions, and evaluating the actions to determine benefits. Some of the tools used for data mining are Artificial Neural Networks, Decision Trees, Rule Induction, Genetic Algorithms, and Nearest Neighbor.

2.3 Related Work

Minimal work has been reported on applying data mining in the Airline Industry. Some of these papers are summarized in this section.

Ramachandran (2001)'s white paper presented some results of data mining on frequent flyer data. The objectives were to identify the characteristics of the customers and to find the relationship among the sectors based on the customer behavior. These characteristics were sectors most frequently flown, class flown, period of year, hometown compared to sector flown. The result categorizes data types into category, booking, and sector. The tool used is Clementine, Business Miner and Intelligent Miner. Clementine permits to use the association rules and factor analysis techniques for modeling discovery analysis; and rule induction and decision trees techniques for predictive modeling.

The customer success story of Alaska Airlines from Siebel was reported in "Alaska Airlines soars in Meeting the Needs of More than 17 Million Customers Annually" (2005). Based on customer loyalty program over several years, a great deal of customer data was acquired. Alaska Airlines chose Siebel Business Analytics to tie together customer data from numerous sources, and to design marketing programs to derive customer loyalty. Alaska Airlines is now able to better understand, respond to, and anticipate customer needs. We conclude from this paper some customer's implementation advice such as understanding data and infrastructure, defining the goal and the business needs before choosing a solution, and

implementation accelerates dramatically as user familiarity increases. The tool used is Siebel Business Analytics to group customer data into one entity, giving the capability of launching marketing programs improving customer loyalty. This paper describes another experience without any concrete application.

With Etzioni, Knoblock, Tuchinda, and Yates (2003) the behavior of airline ticket prices over time is addressed. This work presents data mining methods capable to detect patterns in price data, and how Web price tracking coupled with data mining save consumers money in practice. The data mining methods applied are Ripper, Q-learning, and time series. The main result of Mining prices is to give recommendations to customers to rush or delay purchases based on price change prediction.

A recent related work is on Data Mining and Strategic Marketing in the Airline Industry (Pritscher and Feyen, 2001). A "Trip Builder" algorithm rebuilds the trip and categorizes flight by local and connected. Three major factors have been considered as input variables; the segment distribution, segment order, and number of countries visited with return or no return. With flight categories and revenue information, Pritsher and Feyen (2004) begin the next step. Using k-means clustering algorithm, each customer is assigned to a group. The clustering result is verified by Kohonen methods. Six segments have been discovered permitting to define the trip purpose such as few weekend flights, and few long stay returns.

Chapter 3

PROBLEM AND DATA DESCRIPTION

3.1 Customer Segmentation Study – Data Description

The goal of the study is to extract business and CRM strategies. We extract data from the Frequent Flyer Program and loaded it onto our data mining platform in order to build, test, and apply a predictive model to score customer base, producing a list of customers to increase our profits. The data we describe in this section is real data from the Middle East Airlines.

3.1.1 Frequent Flyer Concept

Due to frequent flyer programs there is a rich data available, which 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 is also used to identify high value customers and provide them with special services and benefits such as lounge access and upgrades.

3.1.2 Cedar Miles Services

We are conducting a customer segmentation study for the frequent flyer customers (Cedar Miles Program) in Middle East Airlines (MEA), the Lebanese national carrier. The Cedar Miles Program is an air miles reward program for an association of more than 79,782 customers including in addition to the flight services, a financial service of a credit card and a hotel service. Each time a passenger uses his dedicated credit card for any transaction or has a stay in the dedicated hotel, he will win additional miles in the reward program. Due to agreements with bank and hotel, MEA 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. The Air Miles Reward Program – Cedar Miles – is a frequent flyer program. The passenger accumulates air travel miles by making transactions on his special credit card or having a reservation in the contractual hotel. Passengers exchange the travel miles collected for rewards.

3.1.3 Membership Categories

Actually our customer are divided into 4 categories of members: Basic member (used for new customer or customer with qualified miles less than 20,000 or customer qualified segment less than 15,000); Prestige member (used for customer with 20,000 qualified miles or for customer with 15,000 qualified segment); Elite member (used for customer with 40,000 qualified miles or for customer with 30,000 qualified segment) and Prestige Club member (used for special customer identified by top management independent from any criteria). To mention that the qualified segment miles are dedicated from purely MEA flights; even with share code flights, the segment must have the MEA flight number. Those categories vary depending on the collected mileage by each passenger.

3.1.4 Customer Activities

All association partners capture passenger transactions and transmit them to MEA Frequent Flyer Group. Then the passengers are stored and the data is used for database marketing initiatives. MEA Frequent Flyer Group data currently contains more than 1,322,409 activities transaction records.

3.1.5 Data Model

The primary source of data for this study is shown in Figure 3.1 data model. The data consisted of approximately 79,000 passengers and their relative transactions for a

period of 6 years. The variables chosen for this study included mileage, passenger membership, number of services used over the passenger membership, and number of services used the last year, in addition to specific business variables.

The "Activities" table in the data model of Figure 3.1 contains revenue data. Each transaction record includes the mileage status used to estimate profitability. Other variables are calculated by merging transaction data to each passenger record.

The Frequent Flyer Program identified the customers by passenger number in order to protect the confidentiality.

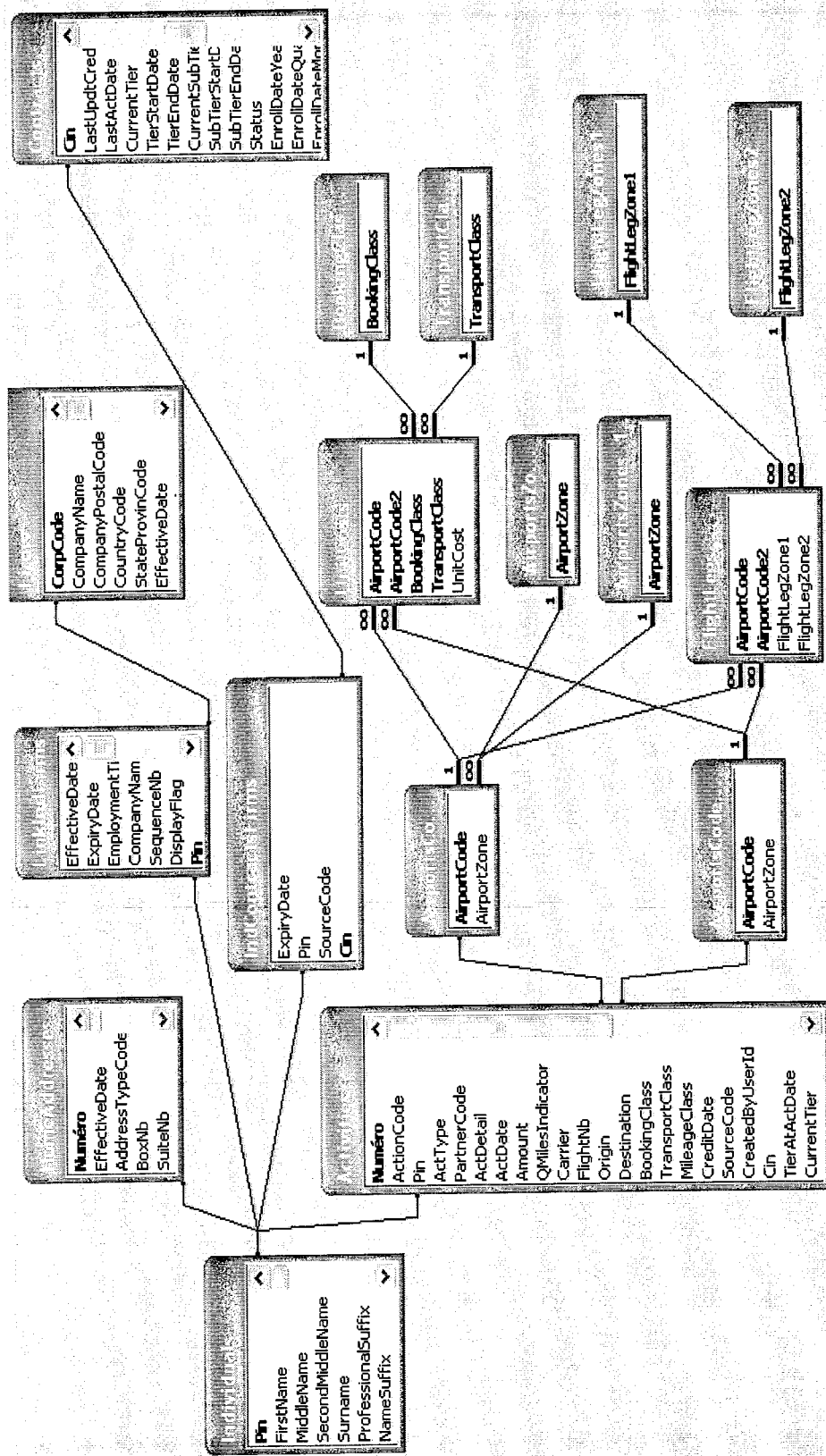


Figure 3.1 - Frequent Flyer case study data model

3.2 Customer Segmentation Study – Problem Description

The market experts concerns are the key business processes for customer management. The key business process typically includes customer value management, customer retention, customer growth, customer acquisition, customer communication and multi-channel optimization.

The objective of this thesis 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 presented in the following subsections:

3.2.1 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 unprofitable or low profit customers to a position of improved profitability?

What is the predicted lifetime value by customer segment?

3.2.2 Customer Retention

Define best market segment

3.2.3 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

3.2.4 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?

Chapter 4

SOLUTION STRATEGY

4.1 Overview

The study is conducted based on the frequent travel data. The data preparation task includes data cleansing and preprocessing. The resultant data will be the input for the data mining process. Clustering and Association Rules are two of the most important data mining methodologies used in marketing and CRM. For clustering, we have used two algorithms: k-means and o-cluster. APRIORI algorithm is used for Association Rules.

4.2 CRISP Implementation

4.2.1 Business Goals

Anyone in the sales field will imagine that the secret to success is to identify and retain top customers. Our targets customers should be not only who spend too much, but they should be valuable candidates for cross-selling. Building CRM environment allows serving airlines customers individually. The main concern was to understand each customer in order to implement new strategic customer segments. The results will be used for marketing issues such as promotions and targeted campaigns, and improving customer service such as information availability for call centers. The questions defined in paragraph 3.2; give a clear idea about our business goals.

4.2.2 Data Mining Goals

In our current situation, the individual mileage determines the customer value. The mileage is a unit measure for customer profitability. The distribution of data into several sources is the primary data problem. Our goal is to merge data sources and to develop a model that generates passenger revenue value, based on the booking history.

4.2.3 Data Understanding

Given the data mining goals, we have to explore which data are available and might be useful for achieving the goals. The data must contain information, be relevant, and be in a format which data mining can use effectively. Our primary source of data is the frequent flyer program database. The target population doesn't contain all customers, only the members of the frequent flyer program. No missing values in the data covering flight activities.

4.2.4 Data Transformation and Aggregation for Clustering

Before we have applied the transformation and aggregation on the available data, we have considered several other approaches. We have considered the use of sample data but since the tool used is very powerful with very low execution time, the complete data gives more powerful and accurate result. In this case, we will have an accurate and complete view by using the complete data. We have also considered the use of two clustering layers. The behavioral clustering has been realized without problem since the data is available. The second clustering layers have to be demographic segmentation. This segmentation will help in choosing suitable advertising, marketing channels, and campaigns to suit strategic behavior segmentation. But the data is not mature; we will consider this feature in our future work.

Several queries have been built to merge the "Activities" transaction data to the "Individuals" passenger file. Those queries create the clustering input record which is mandatory for the clustering algorithms we used. Table 4.1 and Table 4.2 are samples of "Activities" and "Individuals" tables respectively. We display some fields of both tables for privacy reason. The queries presented below illustrate the manipulation done on each transaction data. It includes pivoting, aggregating, and inserting into each passenger record.

Table 4.1 - Sample of "Activities" Table

Numéro	ActionCode	Pin	ActType	PartnerCode	ActDetail	ActDate	Amount
2340852	C	760000000011	FLIGHT	MEA	MEA 201 BEY LHR JJ	12/4/1999	2161
2340853	C	760000000011	FLIGHT	MEA	MEA 572 ABJ BEY YL	12/4/1999	3201
2340854	C	760000000011	PROGRAM	MEA	CSBO MEA 201 JJ	12/4/1999	2161
2340855	C	760000000011	FLIGHT	MEA	MEA 202 LHR BEY JJ	12/5/1999	2161
2340856	C	760000000011	FLIGHT	MEA	MEA 205 BEY CDG JJ	12/5/1999	1979
2340857	C	760000000011	PROGRAM	MEA	CSBO MEA 202 JJ	12/5/1999	2161
2340858	C	760000000011	PROGRAM	MEA	CSBO MEA 205 JJ	12/5/1999	1979
2340859	C	760000000011	FLIGHT	MEA	MEA 201 BEY LHR JT	12/7/1999	2161
2340860	C	760000000011	FLIGHT	MEA	MEA 202 LHR BEY YD	12/7/1999	2161
2340861	C	760000000011	FLIGHT	MEA	MEA 217 BEY FRA JJ	12/7/1999	1762
2340862	C	760000000011	PROGRAM	MEA	CSBO MEA 202 YD	12/7/1999	2161
2340863	C	760000000011	PROGRAM	MEA	CSBO MEA 217 JJ	12/7/1999	1762
2340864	C	760000000011	PROGRAM	MEA	WBME MEA 201 JT	12/7/1999	3000
2340865	C	760000000011	MISCELLANEOUS	ME	CMPL	12/7/1999	18500
2340866	C	760000000011	FLIGHT	MEA	MEA 213 BEY GVA JJ	12/9/1999	1759
2340867	C	760000000011	PROGRAM	MEA	CSBO MEA 213 JJ	12/9/1999	1759
2340868	C	760000000011	FLIGHT	MEA	MEA 218 FRA BEY JJ	12/10/1999	1762
2340869	C	760000000011	PROGRAM	MEA	CSBO MEA 218 JJ	12/10/1999	1762
2340870	C	760000000011	REWARD CLAIM	N/A	CMBMX 000000001	12/10/1999	-15000
2340871	C	760000000011	FLIGHT	MEA	MEA 214 GVA BEY JJ	12/11/1999	1759
2340872	C	760000000011	PROGRAM	MEA	CSBO MEA 214 JJ	12/11/1999	1759
2340873	C	760000000011	FLIGHT	MEA	MEA 205 BEY CDG JJ	12/16/1999	1979
2340874	C	760000000011	PROGRAM	MEA	CSBO MEA 205 JJ	12/16/1999	1979
2340875	C	760000000011	FLIGHT	AFR	AFR511 ORY CMF YY	12/20/1999	500
2340876	C	760000000011	FLIGHT	AFR	AFR760 CDG MRS PP	12/20/1999	406
2340877	C	760000000011	PROGRAM	AFR	FCAF AFR760 PP	12/20/1999	812
2340878	C	760000000011	REWARD CLAIM	N/A	CMBMX 000000001	12/22/1999	15000

Table 4.2 - Sample of "Individuals" Table

Pin	ProfessionalSuffix	BirthDate	Gender	NbOfChildren	PrivacyIndicator	CreatedDate	StatusDate
76000000184		1/1/1972	M		Y	12/7/1999	12/7/1999
76000000195		1/1/1968	M		Y	12/7/1999	12/7/1999
76000000206		4/1/1962	M	02	Y	12/7/1999	12/7/1999
76000000210		1/1/1965	M		Y	12/7/1999	12/7/1999
76000000221		1/1/1979	M		Y	12/7/1999	12/7/1999
76000000232		1/1/1986	M		Y	12/7/1999	12/7/1999
76000000243		1/1/1961	M		Y	12/7/1999	12/7/1999
76000000254		1/1/1972	M		Y	12/7/1999	12/7/1999
76000000265		1/1/1968	M		Y	12/7/1999	12/7/1999
76000000276		1/1/1978	M		Y	12/7/1999	12/7/1999
76000000280		1/1/1987	M		Y	12/7/1999	12/7/1999
76000000291		1/1/1954	M		Y	12/7/1999	12/7/1999
76000000302		1/1/1979	M		Y	12/7/1999	12/7/1999
76000000011		10/23/1968	F		Y	12/3/1999	12/3/1999
76000000022		1/1/1965	F		Y	12/3/1999	12/3/1999
76000000033		1/1/1979	F		Y	12/3/1999	12/3/1999
76000000044		1/1/1986	F		Y	12/3/1999	12/3/1999
76000000055		1/1/1961	M		Y	12/7/1999	12/7/1999
76000000066		1/1/1972	M		Y	12/7/1999	12/7/1999
76000000070		1/1/1968	M		Y	12/7/1999	12/7/1999
76000000081		7/15/1965	M		Y	12/7/1999	12/7/1999
76000000092		8/1/1944	M		Y	12/7/1999	12/7/1999
76000000103		1/1/1965	M		Y	12/7/1999	12/7/1999
76000000114		1/1/1979	M		Y	12/7/1999	12/7/1999
76000000125		1/1/1986	M		Y	12/7/1999	12/7/1999
76000000136		12/3/1945	M		Y	12/7/1999	12/7/1999
76000000140		1/1/1965	M		Y	12/7/1999	12/7/1999

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). The “Behavioral Activities” query is shown in Table 4.3 without the first name and the last name for privacy reason.

After all the data variables were created on each customer record, the missing values needed to be treated. The missing values capable of changing the distribution and statistics of the field are “Financial”, “Flight”, and “Hotel” services used by the customer during his lifetime, and Revenue mileage. We have to discard the customers records with those values missed. The records remaining are 50,830. The “Behavioral

Activities” query is used as input data all clustering algorithms K-Means, O-Cluster, Expectation-Maximization (EM), and COBWEB.

Table 4.3 - Sample of Behavioral Activities Query

custid	actlastyear	actlife	membership	mileage	ram	rmm
11	1	1	74	201960	0.01	2729.19
22	0	1	74	15759	0.01	212.96
33	0	1	74	3905	0.01	52.77
44	0	1	74	6068	0.01	82
66	0	1	74	9303	0.01	125.72
81	0	1	74	53558	0.01	723.76
92	1	1	74	3093	0.01	41.8
103	1	1	74	35549	0.01	480.39
125	1	1	74	28915	0.01	390.74
136	0	1	74	3056	0.01	41.3
140	0	1	74	9709	0.01	131.2
151	0	1	74	1329	0.01	17.96
173	1	1	74	24272	0.01	328
184	0	1	74	24864	0.01	336
206	3	3	74	74132	0.04	1001.78
221	0	1	74	13290	0.01	179.59
232	1	1	74	97337	0.01	1315.36
243	0	1	74	2658	0.01	35.92
254	0	0	74	26240	0	354.59
265	1	1	74	15154	0.01	204.78
276	1	1	74	44796	0.01	605.35
280	1	1	74	28177	0.01	380.77
291	1	1	74	71127	0.01	961.18
313	1	1	74	39804	0.01	537.89
335	1	1	74	6462	0.01	87.32
350	0	1	74	3979	0.01	53.77
383	0	1	74	49152	0.01	664.22
405	1	1	74	44764	0.01	604.92
416	0	1	74	1329	0.01	17.96
431	0	1	74	2658	0.01	35.92
464	0	1	74	8427	0.01	113.88

4.2.5 Data Preparation

Data is prepared using Normalization. Normalizing converts individual attribute values in such a way that all attributes values lie in the same range. Normalization involves scaling continuous values down to specific range such that $x_{new} = (x_{old} - shift)/scale$. It

applies only to numerical attributes. Our study is based on Z-Score Normalization. The normalization definition for each attribute is computed based on the values for mean and standard deviation that are computed from the data. The values for shift and scale are computed to be shift = mean, and scale = standard deviation respectively.

4.2.6 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 MEA 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 approach:

4.2.6.1 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”.

4.2.6.2 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 Hub; the Rafic Harriri Airport (BEY) or Charles-De-Gaulles Airport (CDG).

A query (Q7) based on “Activities” table. It includes the Customer ID, Sector (concatenation of Origin and Destination), Origin (must be “CDG” and “BEY” 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.

4.2.7 Model building and Evaluation

Using data mining tool, the clustering and association rules techniques was applied. For clustering, two different algorithms are used. The result will be analyzed to generate new business rules. The details are described in chapters 5 and 6.

4.3 Data Mining Techniques – Algorithms

Clustering and association rules are two of the most important data mining methodologies used in marketing and customer relationship management. They use customer purchase 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.

4.3.1 Clustering

Behavioral clustering help derive strategic marketing initiatives using the variables that determine customer shareholder 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) by creating predictive model.

The clustering techniques resolve the segmentation data mining problem. It separates the data into subgroups or classes that share common characteristics.

Concept description aims at a description of concepts or classes. The purpose is to gain insights. Our company is interested in learning more about our loyal customers. From a description of these concepts, the company might infer what could be done to keep customers loyal. Typically, segmentation is performed before concept description. Data mining tool performs segmentation and concept description at the same time. Data mining tool performs hierarchical clustering using an enhanced version of the k-means algorithm and O-Cluster. In a data mining tool, 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. Rules represent the bounding boxes. Clusters are also used to generate probability model which is used during scoring for assigning data points to clusters.

4.3.1.1 Enhanced K-Means

The enhanced K-Means is based on the standard k-means algorithm. As per Dunham (2003); the k-means algorithm is given in Figure 4.1:

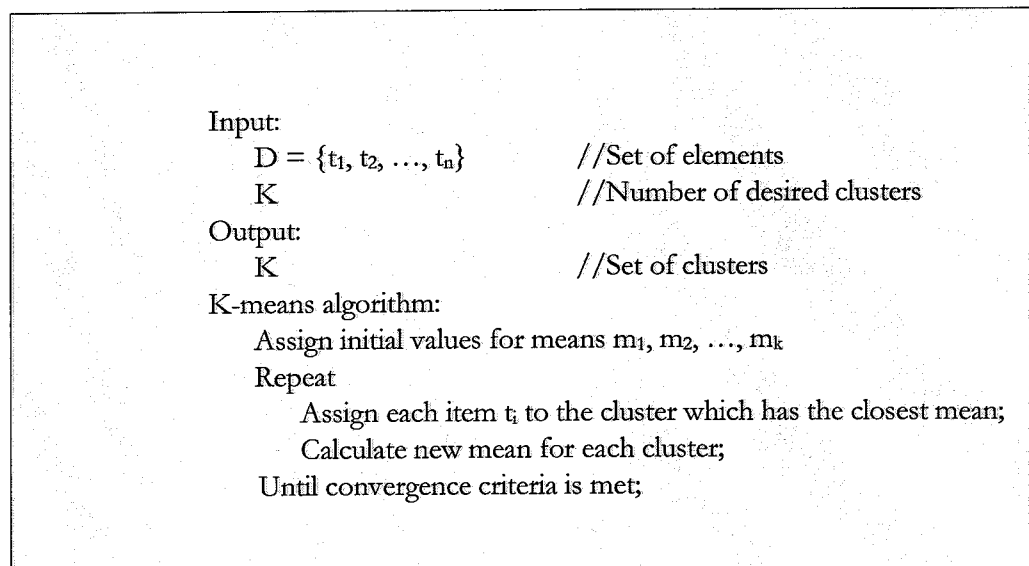


Figure 4.1 – K-Means Algorithm

The time complexity of k-means algorithm for n objects is $O(tkn)$ where t is the number of iterations, and k is the number of clusters specified by the user.

The enhanced k-means as distance-based algorithm rely on a distance metric (function) to measure the similarity (“closeness”) between data points. The selected distance metric is Cosine (Ali, Bagherjeiran, and Chen, 2004). A hierarchical version of enhanced k-means algorithm is implemented. Only unbalanced trees are built. The tree can grown 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. Unbalanced trees give better results with smaller overall distortion while the balanced approach (tree grow one level at a time) is faster.

The enhanced k-means algorithm works best with a moderate number of attributes (at most 100); however, there is no upper limit on the number of attributes. It uses at most one database scan. For each table that doesn't fit in memory, it employs a smart summarization approach that creates a summary of the data table that can be stored in memory. It handles small dataset with numeric mining attributes only.

4.3.1.2 O-Cluster

Ali, Bagherjeiran, and Chen (2004) view the O-Cluster as combination of active sampling technique with an axis-parallel partitioning strategy. It identifies continuous areas of high density in the input space. O-Cluster is an Oracle proprietary algorithm. It is a hierarchical, grid-based clustering. It handles large dataset with all mining attribute types. This algorithm makes two major contributions:

- a.** It proposes the use of a statistical test to validate the quality of a cutting plane.
- b.** It can operate on a small buffer containing a random sample from the original data set.

O-Cluster operates recursively. It evaluates possible splitting points for all projection in a partition, selects the "best" one, and splits the data into two new partitions. The algorithm proceeds by searching for good cutting planes inside the newly created partitions. O-Cluster creates a hierarchical tree structure that groups the input space into rectangular regions. Figure 4.2 provides an outline of O-Cluster algorithm.

The main processing stages are as follows:

- i. Load Data Buffer:** If the entire data set does not fit in the buffer, a random sample is used.
- ii. Compute Histograms for Active Partitions:** The goal is to determine a set of projections for the active partitions and compute histograms along these projections. Any partition that represents a leaf in the clustering hierarchy and is not explicitly marked ambiguous or 'frozen' is considered active. The process whereby an active partition becomes ambiguous or 'frozen' is explained in Step 4.
- iii. Find 'best' Splitting Points for Active Partitions:** For each histogram, O-Cluster attempts to find the 'best' valid cutting plane, if such exist. A valid cutting plane passes through a point of low density (a valley) in the histogram. The point of low density should be surrounded on both sides by points of high density (peaks). O-Cluster attempts to find a pair of peaks with a valley between them where the difference between the peak and the valley histogram counts is statistically significant. Statistical significance is tested using a standard X^2 test:

$$X^2 = \frac{2(\text{observed} - \text{expected})^2}{\text{Expected}} \geq X^2_{\alpha,1}$$

Where the observed value is equal to the histogram count of the valley and the expected value is the average of the histogram counts of the valley and the lower peak. Since multiple splitting points can be found to be valid separators per partition according to this test, O-Cluster choose the one where the valley has the lowest histogram count as the 'best' splitting point.

- iv. **Flag Ambiguous and 'Frozen' Partitions:** If no valid splitting points are found, O-Cluster checks whether the X^2 test would find a valid splitting point at a lower confidence level. If that is the case, the current partition can be considered ambiguous. More data points are needed to establish the quality of the splitting point. If no splitting points were found and there is no ambiguity, the partition can be marked as 'frozen' and the records associated with it marked for deletion from the active buffer.
- v. **Split Active Partitions:** If a valid separator exists, the data points are split along the cutting plane and two new active partitions are created from the original partition. For each new partition, the processing begins recursively from step ii.
- vi. **Reload Buffer:** This step can take place after all recursive partitioning on the current buffer has completed. If all existing partitions are marked as 'frozen' and/or there are no more data points available, the algorithm exits. Otherwise, if some partitions are marked as ambiguous and additional unseen data records exist, O-Cluster proceeds with reloading the data buffer. The new data replace records belonging to 'frozen' partitions. When new records are read in, only data points that fall inside ambiguous partitions are placed in the active buffer. New records falling within a 'frozen' partition are not loaded into the buffer. Loading of new records continues until either: 1) the active buffer is filled again; 2) the end of the data set is reached; or 3) a reasonable number of records have been read, even if active buffer is not full and there are more data. Once the buffer reload is completed, the algorithm proceeds from Step 2.

The O-Cluster algorithm complexity is $O(N \times d)$ where N is the number of data points in the buffer and d is the number of dimensions.

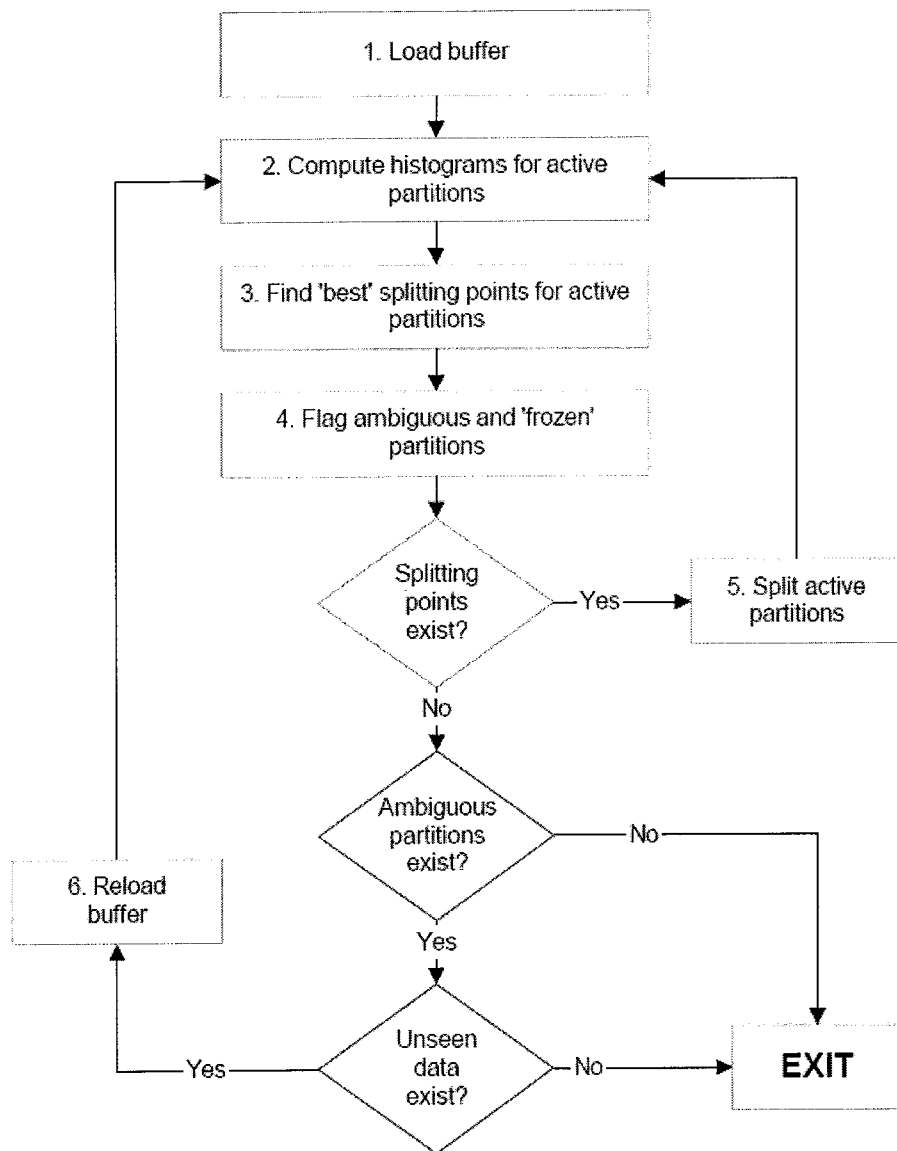


Figure 4.2 - O-Cluster Algorithm Block Diagram

4.3.1.3 Expectation-Maximization (EM)

Dellaert (2002) describes the EM algorithm as unsupervised algorithm. Using the Gaussian mixtures model, EM performs the statistical model function. Such K-Means algorithm, the EM algorithm re-computes a set of parameters until a desired convergence value is achieved. It assumes that the attributes are

independent random variables. A set of N probabilities distributions is called mixture. Each distribution represents a cluster. Figure 4.2 shows the EM algorithm general procedure:

1. Guess initial values for five parameters
2. Compute the cluster probability for each instance using the probability density function. With single independent variable, mean μ , and standard deviation σ , the formula is display below:
3. Use the probability scores to re-estimate the five parameters.
4. Return to Step 2

The algorithm terminates when a formula that measures cluster quality no longer shows significant increases.

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Figure 4.3 – EM Algorithm

The time complexity of the EM algorithm is linear with total length of the data. Each iteration of the EM algorithm takes $O(n^2)$ time and a few iterations are typically sufficient.

4.3.1.4 COBWEB Algorithm

As per Fisher (1987), the COBWEB algorithm is an incremental algorithm; whereas the K-means is an iterative distance-based algorithm. Covering the whole dataset, the K-means algorithm runs until convergence in the clusters is reached. COBWEB works incrementally. It updates the algorithm cluster instance by instance. COBWEB creates a tree. The leaves represent instance in the tree, the root node represents the entire dataset, and the branches represent clusters and sub-clusters.

COBWEB starts with a tree including the root node only. Instances are added one by one. This algorithm is easier to understand than K-Means but instances order can impact the clustering. Two instances very similar can appear as the first input instances at opposite ends of the tree.

COBWEB complexity for n objects is $O(tn)$, depends non-linearly on tree characteristics packed into a constant t .

4.3.2 Association Rules

Dependency analysis finds a model that describes significant dependency (or associations) between data items or events. Dependencies can be used to predict the value of a data item, given information on other data items. Associations are a special case of dependencies. It describes data items or events that frequently occur together.

Association Rules function is often associated with “market basket analysis”, which is used to discover relationships or correlations among a set of items. It is used in data analysis for direct marketing. Association rules capture the co-occurrence of items or events in large volumes of customer transaction data. Finding such rules is valuable for cross-marketing and mail-order promotions.

4.3.2.1 APRIORI Algorithm

Algorithms that calculate association rules work in two phases. In the first phase, all combinations of items that have the required minimum support (called the “frequent item sets”) are discovered. In the second phase, rules of the form $X \rightarrow Y$ with the specified minimum confidence are generated from the frequent item sets.

We use the Apriori algorithm for Association Rules mining problem (Dunham, 2003). The Apriori algorithm for finding frequent itemsets makes multiple passes over the data. In the K th pass, it finds all itemsets having k items, called the k -itemsets. Each pass consists of two phases. Let F_k

represent the set of frequent k -itemsets, and C_k the set of candidate k -itemsets. First, is the candidate generation phase where the set of all frequent $(k-1)$ itemsets, F_{k-1} , found in the $(k-1)$ th pass, is used to generate the candidate itemsets C_k . The candidate generation procedure ensures that C_k is a superset of the set of all frequent k -itemsets. A specialized in-memory hash-tree data structure is used to store C_k . Then, data is scanned in the support counting phase. For each transaction, the candidates in C_k contained in the transaction are determined using the hash-tree data structure and their support count is incremented. At the end of the pass, C_k is examined to determine which of the candidates frequent, yielding F_k are. The algorithm terminates when F_k or C_{k+1} becomes empty.

The main advantage of APRIORI algorithm is a low number of database passes made when searching the hypothesis space. Its disadvantage is its time complexity ($O(n^2)$) with respect to the number of attributes, which becomes difficult when analyzing data with several hundreds of items.

4.3.2.2 PREDICTIVEAPRIORI Algorithm

The PREDICTIVEAPRIORI Algorithm (Sceffer, 2004) doesn't have fixed confidence and support. Its objective is to find the n best rules. The algorithm is displayed in Figure 4.4; the generation of rules procedure is presented in Figure 4.5. Time complexity $O(n^2)$ with respect to the number of attributes.

1. **Input:** n (desired number of association rules), database with items a_1, \dots, a_k .
2. Let $\tau = 1$.
3. **For** $i = 1 \dots k$ **Do:** Draw a number of association rules $[x \Rightarrow y]$ with i items at random. Measure their confidence (provided $s(x) > 0$). Let $\pi_i(c)$ be the distribution of confidences.
4. **For all** c , Let $\pi(c) = \frac{\sum_{i=1}^k \pi_i(c) \binom{k}{i} (2^i - 1)}{\sum_{i=1}^k \binom{k}{i} (2^i - 1)}$.
5. Let $X_0 = \{\emptyset\}$; Let $X_1 = \{\{a_1\}, \dots, \{a_k\}\}$ be all item sets with one single element.
6. **For** $i = 1 \dots k - 1$ **While** ($i = 1$ or $X_{i-1} \neq \emptyset$).
 - (a) **If** $i > 1$ **Then** determine the set of candidate item sets of length i as $X_i = \{x \cup x' \mid x, x' \in X_{i-1}, |x \cup x'| = i\}$. Generation of X_i can be optimized by considering only item sets x and $x' \in X_{i-1}$ that differ only in the element with highest item index. Eliminate double occurrences of item sets in X_i .
 - (b) Run a database pass and determine the support of the generated item sets. Eliminate item sets with support less than τ from X_i .
 - (c) **For all** $x \in X_i$ **Call** RuleGen(x).
 - (d) **If** *best* has been changed, **Then Increase** τ to be the smallest number such that $E(c|1, \tau) > E(c(\text{best}[n]) | \hat{c}(\text{best}[n]), s(\text{best}[n]))$ (refer to Equation 6). **If** $\tau >$ database size, **Then Exit**.
 - (e) **If** τ has been increased in the last step, **Then eliminate** all item sets from X_i which have support below τ .
7. **Output** $\text{best}[1] \dots \text{best}[n]$, the list of the n best association rules.

Figure 4.4 – Predictive APRIORI Algorithm

Algorithm RuleGen(x) (find the best rules with body x efficiently)

10. Let γ be the smallest number such that $E(c|\gamma/s(x), s(x)) > E(c(best[n])|\hat{c}(best[n]), s(best[n]))$.
11. For $j = 1 \dots k - |x|$ (number of items not in x)
 - (a) If $j = 1$ Then Let $Y_1 = \{a_1, \dots, a_k\} \setminus x$.
 - (b) Else Let $Y_j = \{y \cup y' | y, y' \in Y_{j-1}, |y \cup y'| = j\}$ analogous to the generation of candidates in step 6a.
 - (c) For all $y \in Y_j$ Do
 - i. Measure the support $s(x \cup y)$. If $s(x \cup y) \leq \gamma$, Then eliminate y from Y_j and Continue the for loop with the next y .
 - ii. Calculate predictive accuracy $E(c([x \Rightarrow y])|s(x \cup y)/s(x), s(x))$ according to Equation 6.
 - iii. If the predictive accuracy is among the n best found so far (recorded in $best$), Then update $best$, remove rules in $best$ that are subsumed by other, at least equally accurate rules (utilize Theorem 1 and test for $x \subseteq x' \wedge y \supseteq y'$), and Increase γ to be the smallest number such that $E(c|\gamma/s(x), s(x)) \geq E(c(best[n])|\hat{c}(best[n]), s(best[n]))$.
12. If any subsumed rule has been erased in 11(c)iii, Then recur from step 10.

Figure 4.5 – Generation of Rules Procedure

Chapter 5

EXPERIMENTAL RESULTS

5.1 Overview

In this chapter, we describe the data mining results based on the clustering and association rules application is exposed. For clustering, we have used the k-means algorithm with four different scenarios and the O-cluster algorithm. For association rules, we have used the Apriori algorithm.

5.2 Hardware and Software Platform

The computer used for our study is DELL Computer Corporation; DELL LATITUDE D505 – Intel ® Pentium ® M – Processor 1.70 GHz – 592 MHz, 512 MB of RAM.

In our study, we have used the Oracle Data Mining Tool called Oracle Data Miner – the User Interface to the Data Mining option of Oracle 10g. “Oracle is unique, because it integrates the processing-intensive components needed for business intelligence – Extraction, Transformation, and Loading (ETL); Online Analytical Processing (OLAP); and data mining – directly into the database”.

Oracle Data Miner 10.1.0.2 is a user interface to Oracle Data Mining (ODM) 10.1. In addition to Oracle Data Miner, there is the ODM Java Code Generator, which is an Oracle JDeveloper extension.

The Oracle Data Miner 10.1.0.2 targets data analysts directly. In addition, Oracle Data Miner is designed to increase the analyst’s success rate in properly utilizing ODM algorithms. These two goals are addressed in several ways. First, users need more assistance in applying a methodology that addresses both data preparation and algorithm selection. Oracle Data Miner meets this need by providing a Data Mining Activity to step users through the proper methodology. Next, Oracle Data Miner includes improved and expanded heuristics in

the model building and transformation wizards to reduce the chance of error in specifying model and transformation settings. Finally, Oracle Data Miner has added additional transformation wizards to round out the data preparation features. ODM has Regression and Feature Extraction Models, an additional classification algorithm, and the ability to mine text data. To ensure ODM's strong support for model deployment, users can continue to generate Java code for existing models and mining results using the ODM Java Code Generator.

Another tool has been also used; the purpose was to validate and evaluate our Oracle Data Mining result. The Waikato Environment for Knowledge Analysis (WEKA) has been written by Mark Hall (2002 – 2005) at the University of Waikato, New Zealand. We have applied the EM and COBWEB clustering algorithm; in addition to Predictive Apriori association rules algorithm.

5.3 Clustering

5.3.1 Empirical Procedure

The first steps in the clustering process involve selecting the data set and the algorithm we want to use. The next step in the process is to choose the basic run parameters for the algorithm. K-means is the first algorithm applied, based on two different scenarios. The difference between the scenarios in the first approach is the number of clusters. Scenario 1 has 4 clusters, scenario 2 has 6 clusters, scenario 3 has 9 clusters, and scenario 4 has 15 clusters. The second approach for k-means algorithm begins by choosing the best scenario already developed. The best scenario is scenario 3. With scenario 3, we will try to work on the remaining parameters; the maximum iterations, and the minimum error tolerance. The second clustering algorithm is O-cluster with 9 clusters. The algorithms are applied on "Behavioral Clustering" query including 50, 830 records after processing 79,782 records from "Individuals" table, and 1,322,409 records from "Activities" table. 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.

The second approach of clustering used k-means algorithm but divides the “Behavioral Clustering” query into two different parts. The first grouping 80% of the data is used to build the model. The second grouping 20% of the data is used to apply the model. This approach permits to validate the result.

5.3.2 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).

5.3.3 K-Means Algorithm Results

5.3.3.1 Algorithm Parameters

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. ODM used an enhanced k-means algorithm.

5.3.3.2 Scenario 1

For our first clustering run, we choose a maximum of four clusters, a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. Many trials of clustering run has been done based on changes of clustering parameters in order to choose the best clustering to generate association rules and understand our customer behavior. The execution time is less than 1 second (Start time: 2:05 PM – End time: 2:05 PM). Table 5.1 displays the rules with support and confidentiality for each rule.

In Table 5.2, the following general information is displayed about the cluster: Cluster ID, Cluster Level, Record Count (the number of records or cases in the cluster), and the attributes in the cluster centroid. For each cluster centroid attribute, the Attribute name and Centroid Value are displayed.

Table 5.1 – K-Means Algorithm Rules (Scenario 1)

Cluster ID	Act. Life		Act. Last Year		Mileage		Membership		Mileage / Membership		Activities / Membership		Confidence	Support
	0.99	1.02	0.99	1.02	-5,135.00	136,877.48	40.88	74	-170	2,131.52	0.0067	0.0201		
4	0.99	1.02	0.99	1.02	-5,135.00	136,877.48	40.88	74	-170	2,131.52	0.0067	0.0201	0.9054595	0.235904
5	0.99	1.02	0	0.03	-5,135.00	48,119.68	45.92	74	-170	750.61	0.0067	0.0201	0.9374785	0.161066
6	0.99	1.02	0.99	1.02	-5,135.00	56,995.46	4.88	40.16	-170	3,052.12	0.0134	0.0201	0.9485229	0.399843
7	0.99	1.02	0	0.03	-5,135.00	30,368.12	12.8	45.2	-170	1210.91	0.0134	0.0603	0.9145011	0.133622

Table 5.2 - Clusters Details of K-Means Algorithm (Scenario 1)

Cluster ID	Cluster Level	Record Count	Attribute	Centroid Value	Attribute	Centroid Value
4	3	13,243	ACTLASTYEAR	1.05-1.08	MILEAGE	30368.12-39243.9
			ACTLIFE	1.05-1.08	RAM	0.0134-0.0201
			MEMBERSHIP	61.76-62.48	RMM	290.3033-750.6066
5	3	8,733	ACTLASTYEAR	0.0-0.03	MILEAGE	12616.56-21492.34
			ACTLIFE	0.96-0.99	RAM	0.0067-0.0134
			MEMBERSHIP	63.2-63.92	RMM	-460.3033
6	3	21,427	ACTLASTYEAR	1.05-1.08	MILEAGE	12616.56-21492.34
			ACTLIFE	1.05-1.08	RAM	0.0737-0.0804
			MEMBERSHIP	19.28-20.00	RMM	750.6066-1210.9099
7	3	7,427	ACTLASTYEAR	0.0-0.03	MILEAGE	3740.78-12616.56
			ACTLIFE	0.96-0.99	RAM	0.0335-0.0402
			MEMBERSHIP	30.08-30.80	RMM	290.3033-750.6066

5.3.3.3 Scenario 2

For our second clustering run, we choose a maximum of six clusters, a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. The execution time is less than 1 second (Start time: 9:49 PM – End time: 9:49 PM). The Table 5.3 displays the rules with support and confidentiality for each rule.

In Table 5.4, the following general information is displayed about the cluster: Cluster ID, Cluster Level, Record Count (the number of records or cases in the cluster), and the attributes in the cluster centroid. For each cluster centroid attribute, the Attribute name and Centroid Value are displayed.

Table 5.3 – K-Means Algorithm Rules (Scenario 2)

Cluster ID	Act. Life		Act. Last Year		Mileage		Membership		Ratio Mileage to Membership		Ratio Activities to Membership		Confidence	Support
	Act. Life	Act. Last Year	Mileage	Act. Last Year	Membership	Ratio Mileage to Membership	Membership	Ratio Activities to Membership	Confidence	Support				
5	0.99	1.02	0.00	0.03	-5,135.00	48,119.68	48.80	74.00	-170	750.6066	0.0067	0.0201	0.92799807	0.147826
7	0.99	1.02	0.00	0.03	-5,135.00	30,368.12	12.8	48.08	-170	1,210.9099	0.0134	0.0603	0.92161727	0.146193
8	0.99	1.02	0.99	1.02	-5,135.00	65,871.24	18.56	39.44	-170	2,591.8198	0.0268	0.0536	0.9233863	0.176175
9	0.99	1.02	0.99	1.02	-5,135.00	39,243.90	2.72	18.56	-170	3,512.4265	0.0536	0.335	0.96259356	0.205036
10	0.99	1.02	0.99	1.02	-5,135.00	136,877.48	59.6	74	-170	2,131.5166	0.0067	0.0201	0.90492755	0.153551
11	0.99	1.02	0.99	1.02	-5,135.00	92,498.58	39.44	59.6	-170	2,131.5166	0.0134	0.0201	0.9467391	0.102813

Table 5.4 - Clusters Details of K-Means Algorithm (Scenario 2)

Cluster ID	Cluster Level	Record Count	Attribute	Centroid Value
5	3	8,097	ACTLASTYEAR	0.0-0.03
			ACTLIFE	0.96-0.99
			MEMBERSHIP	64.64-65.36
			MILEAGE	12616.56-21492.34
			RAM	0.0067-0.0134
			RMM	-170-290.3033
7	3	8,063	ACTLASTYEAR	0.0-0.03
			ACTLIFE	0.96-0.99
			MEMBERSHIP	31.52-32.24
			MILEAGE	3740.78-12616.56
			RAM	0.0268-0.0335
			RMM	290.3033-750.6066
8	4	9,698	ACTLASTYEAR	1.05-1.08
			ACTLIFE	1.08-1.11
			MEMBERSHIP	27.2-27.92
			MILEAGE	12616.56-21492.34
			RAM	0.0402-0.0469
			RMM	750.6066-1210.9099
9	4	10,827	ACTLASTYEAR	1.02-1.05
			ACTLIFE	1.02-1.05
			MEMBERSHIP	10.64-11.36
			MILEAGE	3740.78-12616.56
			RAM	0.1139-0.1206
			RMM	750.6066-1210.9099
10	4	8,625	ACTLASTYEAR	1.08-1.11
			ACTLIFE	1.08-1.11
			MEMBERSHIP	68.96-69.68
			MILEAGE	39243.9-48119.68
			RAM	0.0134-0.0201
			RMM	290.3033-750.6066
11	4	5,520	ACTLASTYEAR	1.02-1.05
			ACTLIFE	1.02-1.05
			MEMBERSHIP	48.08-48.8
			MILEAGE	30368.12-39243.9
			RAM	0.0134-0.0201
			RMM	290.3033-750.6066

5.3.3.4 Scenario 3

For clustering run, we choose a maximum of nine clusters, a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. The execution time is 1 second (Start time: 10:22 PM – End time: 10:23 PM). Table 5.5 displays the rules with support and confidentiality for each rule.

In Table 5.6, the following general information is displayed about the cluster: Cluster ID, Cluster Level, Record Count (the number of records or cases in the cluster), and the attributes in the cluster centroid. For each cluster centroid attribute, the Attribute name and Centroid Value are displayed.

Table 5.5 – K-Means Algorithm Rules (Scenario 3)

Cluster ID	Act. Life		Act. Last Year		Mileage		Membership		Mileage / Membership		Activities / Membership		Confidence		Support	
8	0.99	1.02	0.99	1.02	-5,135.00	56,995.46	15.68	34.4	-170	2,591.82	0.0268	0.0603	0.9511366	0.17615582		
	0.99	1.02	0.99	1.02	-5,135.00	30,368.12	2.72	15.68	-170	3,512.43	0.067	0.335	0.951246381	0.169663578		
11	0.99	1.02	0.99	1.02	-5,135.00	83,622.80	34.4	55.28	-170	2,131.52	0.0134	0.0335	0.946414113	0.11501082		
	0.99	1.02	0	0.03	-5,135.00	30,368.12	30.8	46.64	-170	750.61	0.0134	0.0335	0.943230331	0.073873699		
13	0.99	1.02	0	0.03	-5,135.00	21,492.34	12.8	30.8	-170	1,210.91	0.0268	0.0737	0.954298198	0.069014363		
	0.99	1.02	0	0.03	-5,135.00	48,119.68	61.76	74	-170	750.61	0.0067	0.0201	0.951968193	0.103718273		
15	0.99	1.02	0	0.03	-5,135.00	39,243.90	46.64	61.04	-170	750.61	0.0134	0.0201	0.964924097	0.056285657		
	1.98	2.01	1.98	2.01	-5,135.00	163,504.81	25.76	74	-170	3,512.43	0.0268	0.0737	0.934624612	0.022781821		
17	0.99	1.02	0.99	1.02	-5,135.00	136,877.48	56.72	74	-170	2,131.52	0.0067	0.0201	0.925803483	0.159807205		

Table 5.6 - Clusters Details of K-Means Algorithm (Scenario 3)

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

5.3.3.5 Scenario 4

For this clustering run, we choose a maximum of fifteen clusters, a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. The execution time is 1 second (Start time: 10:22 PM – End time: 10:23 PM). Table 5.7 displays the rules with support and confidentiality for each rule. Support and confidence can be used to rank the rules and hence the predictions.

In Table 5.8, the following general information is displayed about the cluster: Cluster ID, Cluster Level, Record Count (the number of records or cases in the cluster), and the attributes in the cluster centroid. For each cluster centroid attribute, the Attribute name and Centroid Value are displayed.

Table 5.7 – K-Means Algorithm Rules (Scenario 4)

Cluster ID	Act. Life		Act. Last Year		Mileage		Membership		Mileage / Membership		Activities / Membership		Confidence	Support
	0.99	1.02	0	0.03	-5,135.00	39,243.90	46.64	61.04	-170	750.61	0.0134	0.0201		
16	1.98	2.01	1.98	2.01	-5,135.00	216,759.50	62.48	74	-170	3,052.12	0.0268	0.0335	0.8544726	0.012591
17	0.99	1.02	0.99	1.02	-5,135.00	136,877.48	58.88	74	-170	2,131.52	0.0067	0.0201	0.9250275	0.148554
18	1.98	2.01	1.98	2.01	-5,135.00	119,125.92	30.8	61.04	-170	3512.427	0.0268	0.0603	0.8967495	0.009227
19	0.99	1.02	0.99	1.02	-5,135.00	83,622.80	38	58.16	-170	2,131.52	0.0134	0.0201	0.9333962	0.097324
20	0.99	1.02	0.99	1.02	-5,135.00	21,492.34	2.72	7.04	-170	3972.73	0.134	0.335	0.9493343	0.050502
21	0.99	1.02	0.99	1.02	-5,135.00	39,243.90	7.76	19.28	-170	2591.82	0.0536	0.1206	0.9444379	0.15483
22	1.98	2.01	1.98	2.01	-5,135.00	110,250.14	3.44	29.36	-170	5,353.64	0.067	0.335	0.9138627	0.013358
23	0.99	1.02	0.99	1.02	-5135	65871.24	20	37.28	-170	2591.82	0.0268	0.0536	0.9561006	0.153394
24	0.99	1.02	0	0.03	-5135	30368.12	30.8	46.64	-170	750.6066	0.0134	0.0335	0.9486156	0.070096
25	0	0.03	0	0.03	-5135	3740.78	43.76	45.2	-170	290.3033	0	0.0067	0.9035087	0.004053
26	0.99	1.02	0	0.03	-5135	48119.68	62.48	74	-170	750.6066	0.0067	0.0134	0.7926783	0.082215
27	0	0.03	0	0.03	-5135	12616.56	66.8	74	-170	290.3033	0	0.0067	0.9285715	0.004859
28	0.99	1.02	0	0.03	-5135	21492.34	13.52	30.08	-170	1210.91	0.0268	0.0737	0.9538547	0.067913
29	0	0.03	0	0.03	-5135	12616.56	2.72	22.16	-170	750.6066	0	0.0067	0.9122807	0.001023

Table 5.8 - Clusters Details of K-Means Algorithm (Scenario 4)

Cluster ID	Cluster Level	Record Count	Attribute	Centroid Value	Attribute	Centroid Value
15	4	2,962	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	-170-290.3033
16	5	749	ACTLASTYEAR	1.86-1.89	MILEAGE	56995.46-65871.24
			ACTLIFE	1.98-2.01	RAM	0.0268-0.0335
			MEMBERSHIP	70.4-71.12	RMM	750.6066-1210.9099
17	5	8,163	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	290.3033-750.6066
18	5	523	ACTLASTYEAR	1.89-1.92	MILEAGE	39243.9-48119.68
			ACTLIFE	1.98-2.01	RAM	0.0402-0.0469
			MEMBERSHIP	43.76-44.48	RMM	750.6066-1210.9099
19	5	5,300	ACTLASTYEAR	0.99-1.02	MILEAGE	21492.34-30368.12
			ACTLIFE	0.99-1.02	RAM	0.0134-0.0201
			MEMBERSHIP	46.64-47.36	RMM	290.3033-750.6066
20	5	2,704	ACTLASTYEAR	0.99-1.02	MILEAGE	3740.78-12616.56
			ACTLIFE	0.99-1.02	RAM	0.1943-0.201
			MEMBERSHIP	4.88-5.6	RMM	1210.9099-1671.2133
21	5	8,333	ACTLASTYEAR	0.99-1.02	MILEAGE	12616.56-21492.34
			ACTLIFE	0.99-1.02	RAM	0.0804-0.0871
			MEMBERSHIP	12.8-13.52	RMM	750.6066-1210.9099
22	5	743	ACTLASTYEAR	1.92-1.95	MILEAGE	21492.34-30368.12
			ACTLIFE	1.98-2.01	RAM	0.1407-0.1474
			MEMBERSHIP	17.12-17.84	RMM	1210.9099-1671.2133
23	5	8,155	ACTLASTYEAR	0.99-1.02	MILEAGE	21492.34-30368.12
			ACTLIFE	0.99-1.02	RAM	0.0335-0.0402
			MEMBERSHIP	27.2-27.92	RMM	750.6066-1210.9099
24	5	3,756	ACTLASTYEAR	0.0-0.03	MILEAGE	3740.78-12616.56
			ACTLIFE	0.99-1.02	RAM	0.0201-0.0268
			MEMBERSHIP	38.0-38.72	RMM	-460.3033

25	5	228	ACTLASTYEAR	0.0-0.03	MILEAGE	-5135-3740.78
			ACTLIFE	0.0-0.03	RAM	0.0-0.0067
			MEMBERSHIP	44.48-45.2	RMM	-170-290.3033
26	5	5,272	ACTLASTYEAR	0.0-0.03	MILEAGE	12616.56-21492.34
			ACTLIFE	0.99-1.02	RAM	0.0067-0.0134
			MEMBERSHIP	68.96-69.68	RMM	-170-290.3033
27	5	266	ACTLASTYEAR	0.0-0.03	MILEAGE	-5135-3740.78
			ACTLIFE	0.0-0.03	RAM	0.0-0.0067
			MEMBERSHIP	71.12-71.84	RMM	-170-290.3033
28	5	3,619	ACTLASTYEAR	0.0-0.03	MILEAGE	3740.78-12616.56
			ACTLIFE	0.99-1.02	RAM	0.0402-0.0469
			MEMBERSHIP	22.16-22.88	RMM	290.3033-750.6066
29	5	57	ACTLASTYEAR	0.0-0.03	MILEAGE	-5135-3740.78
			ACTLIFE	0.0-0.03	RAM	0.0-0.0067
			MEMBERSHIP	8.48-9.2	RMM	290.3033-750.6066

5.3.3.6 Scenario 5

As discussed in chapter 6, we found scenario 3 the most suitable for our study. In this section, we will discuss the remaining parameters: the maximum iterations and the minimum error tolerance. We try the scenario 3 with maximum iteration of 30. The execution time increased a little bit. The result doesn't change as shown in the table 5.9. We can see that the result obtained in the scenario 3 is correct since with more iteration we have the same result.

The Minimum Error Tolerance parameter must be between 0.001 (slow build) and 0.1 (fast build). We have tried two different values 0.001 and 0.01. The default value is 0.005. Increasing minimum error tolerance builds models faster, but with lower accuracy. The result is shown in Table 5.10. The comparison between cases for 0.001 and 0.01 are very close. In order to get an average result we apply the default value.

Table 5.9 - Parameters Change Results

Cluster ID	Cases
8	9414
9	9066
11	6177
12	3981
13	3676
14	5538
15	2965
16	1239
17	8774

Table 5.10 - Comparison between different value of Minimum Error Tolerance

Cluster ID	0.01Cases	Default Cases (0.005)	0.001 Cases
9	6093	9414	6394
10	8543	9066	9184
11	10471	6177	9117
12	5533	3981	4990
13	3419	3676	4148
14	3532	5538	6753
15	3676	2965	269
16	1170	1239	1212
17	8393	8774	8763

5.3.3.7 Scenario 6

We use this scenario to validate our result of scenario 3. In this scenario, we divide the data into two different parts. The first part groups 80% of the data is used to build the model. The second part groups 20% of the data is used to apply the model. In order to split the "Behavioral Clustering" query data, including 50,830 records, we divide the data sequentially into groups of 100 records each. The first 80 records are added to the first part and the remaining 20 are added to the second part. The last part grouping only 30 records; the first 24 records are added to the first part. It represents 80% of the remaining data. The last 6 records are added to the second part. It contains 20% of the data. The first part includes 40,664 records

and the second part includes 10,166 records. For clustering model build, we choose a maximum of nine clusters, a maximum of 6 passes through the data, and a minimum error tolerance of 0.005. The first part of the data is used to build the model. Table 5.11 displays the rules with support and confidentiality for each rule.

We apply the result on the second part of the data. This part groups 20% of the data. We found that the result 100% similar to the result of scenario 3 as shown in Table 5.12. The table lists just an example of cluster 14 data. This scenario proves the validity of the clustering result.

Table 5.11 - K-Means Algorithm Rules for Partitioned Data

Cluster ID	ActLife		ActLastYear		Mileage		Membership		Ratio Mileage to Membership		Ratio Activities to Membership		Confidence	Support
8	0.99	1.02	0.99	1.02	5,135.00	65,871.24	20.75	41.34	-93.8	2,232.14	0.0134	0.0536	0.95044476	0.183946
9	0.99	1.02	0.99	1.02	5,135.00	39,243.90	3.71	20.04	-93.8	3,265.89	0.0469	0.2549	0.9410086	0.202808
11	0.99	1.02	0.99	1.02	5,135.00	92,498.58	41.34	61.22	-93.8	1,715.27	0.0134	0.0201	0.9461847	0.086907
12	0.99	1.02	0	0.03	5,135.00	30,368.12	33.53	47.02	-93.8	681.51	0.0134	0.0335	0.93404007	0.078005
13	0.99	1.02	0	0.03	5,135.00	21,492.34	12.94	30.69	-93.8	939.95	0.0268	0.0737	0.94707972	0.072177
14	0.99	1.02	0	0.03	5,135.00	48,119.68	64.77	74	-93.8	681.51	0.0067	0.0134	0.95196819	0.103718
15	0.99	1.02	0	0.03	5,135.00	39,243.90	51.99	64.06	-93.8	681.51	0.0134	0.0201	0.9553796	0.065291
16	1.98	2.01	1.98	2.01	5,135.00	207,883.72	34.95	74	-93.8	3,007.46	0.0268	0.0536	0.9511158	0.02201
17	0.99	1.02	0.99	1.02	5,135.00	154,629.05	61.93	74	-93.8	1,973.70	0.0067	0.0201	0.885159	0.135525

Table 5-12 – Comparison between Scenario 3 and Scenario 6

Cluster ID	CUSTID with Clustering	
	Complete Data	Partitioned Data
14	107796	107796
14	107800	107800
14	107881	107881
14	107892	107892
14	107936	107936
14	108006	108006
14	108032	108032
14	108065	108065
14	108080	108080
14	108091	108091
14	108102	108102
14	108113	108113
14	108150	108150
14	108183	108183
14	108194	108194
14	108205	108205
14	108216	108216
14	108253	108253
14	108345	108345
14	108371	108371
14	108415	108415
14	108430	108430
14	108463	108463
14	108474	108474
14	108533	108533
14	108544	108544
14	108581	108581
14	108625	108625
14	108636	108636
14	108673	108673
14	108706	108706
14	108743	108743
14	108776	108776
14	108813	108813
14	108835	108835
14	108883	108883
14	108905	108905
14	108916	108916
14	108931	108931
14	108975	108975
14	109023	109023
14	109045	109045

5.3.4 O-Cluster Algorithm Results

For comparison purposes, the same data set used for the K-means algorithm scenario 3 is used for the O-Cluster Algorithm.

5.3.4.1 Algorithm Parameters

The next step in the process is to choose the basic run parameters for the algorithm. The basic parameters available for O-Cluster 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 10.
- Sensitivity. By increasing the sensitivity value, the number of clusters created may be increased, but the model will take longer to build. A higher sensitivity value makes the algorithm detect smaller density variation as clusters. The range value is between 0 (fewer clusters) to 1 (more clusters). The default value is 0.5.

5.3.4.2 Scenario

For O-Cluster clustering run, we choose a maximum of nine clusters, and a sensitivity value of 0.5. The O-Cluster algorithm is used to verify the k-means algorithm results for 9 clusters. The execution time is less than 1 second (Start time: 8:15 PM – End time: 8:15 PM).

The Table 5.13 displays the rules with support and confidentiality for each rule. Support and confidence can be used to rank the rules and hence the predictions.

We have tried to modify the sensitivity parameters with 0 and 1. This will allow overwriting the minimum number of clusters chosen. Since we are using the o-cluster algorithm to verify the result of k-means algorithm scenario 3, we will keep the default value.

Table 5.13 – O-Cluster Algorithm Rules

Cluster ID	ActLife	ActLastYear	Mileage		Membership		Ratio Mileage to Membership		Ratio Activities to Membership		Confidence	Support
3	1	1	-5,135.00	21,492.34	2	4.4	-170	4,893.34	0.2479	0.335	0.857442319	0.00818
5	1	1	-5,135.00	30,368.12	4.4	9.2	290.3033	3,972.73	0.134	0.201	0.883638501	0.029920001
7	1	1	-5,135.00	39,243.90	6.8	16.4	-170	3,052.12	0.067	0.1206	0.888447344	0.120739996
11	1	0	-5,135.00	39,243.90	38	74	-170	750.61	0.0067	0.0201	0.948440909	0.200139999
13	1	(0, 1)	-5,135.00	48,119.68	14	23.6	-170	2,591.82	0.0469	0.0603	0.950794995	0.113619998
14	1	1	-5,135.00	92,498.58	38	52.4	-170	2,131.52	0.0134	0.0201	0.957131326	0.071000002
15	1	1	-5,135.00	136,877.48	54.8	74	-170	2,131.52	0.0067	0.0201	0.86458886	0.16256
16	1	1	-5,135.00	65,871.24	21.2	40.4	-170	2,131.52	0.0268	0.0402	0.950081468	0.128279999
17	(1, 2)	(0, 2)	-5,135.00	74,747.02	21.2	74	-170	1,210.91	0.0268	0.0402	9251291156	0.0860000003

In Table 5.14, the following general information is displayed about the cluster: Cluster ID, Cluster Level, and Record Count (the number of records or cases in the cluster).

Table 5.14 - Clusters Details of O-Cluster Algorithm

Cluster ID	Cluster Level	Record Count
3	2	477
5	3	1,693
7	4	6,795
11	6	10,551
13	6	5,975
14	7	3,709
15	7	9,401
16	7	6,751
17	7	4,648

The Table 5.15 shows the attributes in the cluster centroid. For each cluster centroid attribute, the Attribute name and Centroid.

Table 5.15 - Centroid Value of O-Cluster Algorithm

Cluster ID	ActLife	ActLastYear	Mileage		Membership		Ratio Mileage to Membership		Ratio Activities to Membership	
3	1	1	3,740.78	12,616.56	2	4.4	2,131.5166	2,591.8198	0.2814	0.2881
5	1	1	3,740.78	12,616.56	6.8	9.2	1,210.9099	1,671.2133	0.1608	0.1675
7	1	1	12,616.56	21,492.34	11.6	14	750.6066	1210.9099	0.0871	0.0938
11	1	1	12,616.56	21,492.34	59.6	62	-170	290.3033	0.0134	0.0201
13	1	1	12,616.56	21,492.34	18.8	21.2	750.6066	1210.9099	0.0469	0.0536
14	1	1	30,368.12	39,243.90	42.8	45.2	290.3033	750.6066	0.0134	0.0201
15	1	1	39,243.90	48,119.68	64.4	66.8	290.3033	750.6066	0.0067	0.0134
16	1	1	21,492.34	30,368.12	30.8	33.2	750.6066	1210.9099	0.0268	0.0335
17	1	1	21,492.34	30,368.12	35.6	38	290.3033	750.6066	0.0268	0.0335

5.3.5 EM Algorithm Results

5.3.5.1 Algorithm Parameters

The basic parameters available for EM clustering include:

- 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 default is 100.
- Maximum number of clusters. We specify the maximum number of clusters allowed. The default value is -1; it permits to select the number of clusters automatically by cross-validation.
- Minimum Standard Deviation. It set the minimum allowable standard deviation. The default value is 1.0 E-6.

The model stops after either the change in standard deviation between two consecutive iterations is less than minimum standard deviation or the maximum number of iterations is greater than maximum iterations. The EM algorithm is applied using WEKA tool. Figure 5.1 shows the WEKA preprocess view. It permits to examine the data preparation.

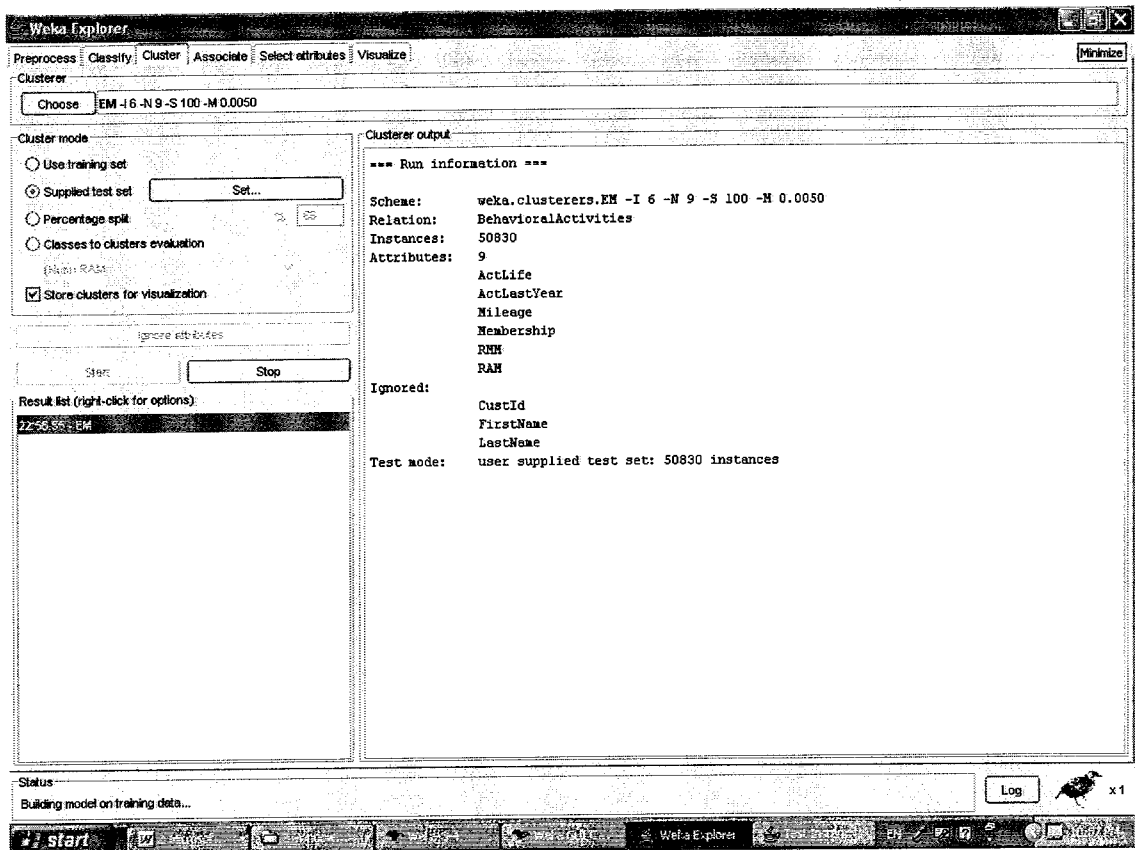


Figure 5.1 - WEKA Preprocess View

5.3.5.2 Scenario

For EM algorithm clustering run, we choose a maximum of nine clusters, a maximum of 6 passes through the data, and a minimum standard deviation of 0.005. The execution time is about 7 minutes. Table 5.16 displays the clusters characteristics with percentage and probability for each cluster.

Table 5.16 – EM Algorithm Result

Cluster ID	Record Count	%	Prob.	ActLife		ActLastYear		Mileage		Membership		RMM		RAM	
				Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
0	9,466	19%	0.19	0.98	0.14	0.41	0.49	15,197.17	11,193.67	68.19	4.61	222.95	162.80	0.01	0.05
1	968	2%	0.02	1.99	0.09	1.90	0.30	37,716.49	34,044.35	21.69	10.33	1,995.87	2,259.02	0.12	0.09
2	8,451	17%	0.16	0.99	0.09	0.98	0.12	9,909.00	7,706.68	8.53	2.97	1,278.66	1,083.52	0.14	0.06
3	3,660	7%	0.07	0.98	0.13	0.93	0.26	79,673.36	61,395.94	68.87	4.93	1,166.37	929.76	0.01	0.05
4	1,085	2%	0.02	2.02	0.15	1.90	0.36	68,122.54	64,955.61	64.06	10.27	1,080.46	1,022.17	0.03	0.01
5	4,506	9%	0.10	1.00	0.23	0.15	0.35	8,628.30	4,169.13	25.23	6.56	356.78	171.58	0.04	0.01
6	8,635	17%	0.17	1.00	0.23	0.89	0.31	18,200.35	15,989.79	22.14	5.74	855.22	789.56	0.05	0.01
7	10,009	20%	0.17	0.96	0.20	0.19	0.40	9,538.49	6,545.06	45.60	7.07	211.78	143.71	0.02	0.01
8	4,050	8%	0.10	1.00	0.23	0.91	0.29	33,052.71	27,241.58	44.41	8.20	761.54	669.88	0.02	0.05

5.3.6 COBWEB Algorithm Results

5.3.6.1 Algorithm Parameters

The basic parameters available for COBWEB clustering include:

- Acuity. Set the minimum standard deviation for numeric attributes. The default value is 1.0.
- Cutoff. Set the category utility threshold by which to prune nodes. The default is 0.0028.
- Save Instance Data. Save instance information for visualization purpose.

5.3.6.2 Scenario

For COBWEB algorithm clustering run, we choose an acuity of 0.005, a cutoff of 0.09, and a save instance data sets to "True". The execution time has been very long. After 5 hours, we got the following message: "Not enough memory. Please load a smaller dataset". Several attempts have been repeated by changing the parameters value (Acuity, and cutoff), but the result has been always the same due the volume of the database which consists of 50,830 records. Due to those reasons, we discard the result of this algorithm.

5.4 Association Rules

5.4.1 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. We do not use any specialized in-memory data structures. The SQL queries are fine-tuned to run efficiently in the database server by using various hints.

The result generated by k-means clustering Scenario 3 will be used as a basis for the Association Rules algorithm. The first steps in the association rules process involve selecting the data set. The algorithm used is Apriori Algorithm. The next step in the process is to choose the basic run parameters for Apriori algorithm. Three different scenarios have been applied. The first two scenarios are based on the flight activities; especially the sectors, with 1,867 records. The third scenario is based on “Financial”, “Flight”, and “Hotel” activities with 1,896 records. In chapter 6, we discuss the results from business side.

5.4.2 Scoring (Applying Models)

Clusters discovered are used to generate a profitability model that is used during scoring (applying models) for assigning data points to clusters.

We apply the k-means clustering model Scenario 3 on the data using the following parameters:

Values of highest cluster Ids.

The number of top cluster Ids is 1.

Table 5.17 shows a sample of the scoring based on K-Means algorithm – Scenario 3.

Table 5.17 – K-Means Algorithm Scoring Sample (Scenario 3)

CLUSTER ID	PROBABILITY	ACTLASTYEAR	ACTLIFE	CUSTID	MEMBERSHIP	MILEAGE	RAM	RMM
14	0.9999967	0	1	33	74	3905	0.01	52.77
14	0.9999967	0	1	44	74	6068	0.01	82
14	0.9999967	0	1	140	74	9709	0.01	131.2
14	0.9999975	0	1	151	74	1329	0.01	17.96
16	0.9997239	3	3	206	74	74132	0.04	1001.78
14	0.9999975	0	1	221	74	13290	0.01	179.59
17	0.9999986	1	1	232	74	97337	0.01	1315.36
14	0.9999975	0	1	243	74	2658	0.01	35.92
17	0.9999927	1	1	313	74	39804	0.01	537.89
17	0.9999927	1	1	405	74	44764	0.01	604.92
14	0.9999967	0	1	464	74	8427	0.01	113.88
17	0.9999877	1	1	490	74	16715	0.01	225.88
14	0.9999975	0	1	501	74	14824	0.01	200.32
14	0.9999967	0	1	523	74	10466	0.01	141.43
14	0.9999975	0	1	814	74	575	0.01	7.77
17	0.99998957	1	1	840	74	1329	0.01	17.96
14	0.9999967	0	1	943	74	4658	0.01	62.95
17	0.9999879	1	1	991	74	11251	0.01	152.04
14	0.9999967	0	1	1024	74	7974	0.01	107.76
17	0.9999877	1	1	1050	74	15290	0.01	206.62
17	0.9999879	1	1	1061	74	9974	0.01	134.78
17	0.9999927	1	1	1072	74	39411	0.01	532.58
14	0.9999967	0	1	1105	74	3769	0.01	50.93
14	0.9999975	0	1	1120	74	2754	0.01	37.22
14	0.9999975	0	1	1186	74	19225	0.01	259.8
17	0.9999986	1	1	1282	74	96304	0.01	1301.41
17	0.99998957	1	1	92	74	3093	0.01	41.8
17	0.99999154	1	1	103	74	35549	0.01	480.39
17	0.99999064	1	1	125	74	28915	0.01	390.74
14	0.9999975	0	1	136	74	3056	0.01	41.3

5.4.3 Input Variables

K-means algorithm scenario 3 divides the customers into 9 groups or clusters. Using the Scoring techniques, we assign to each customer a cluster ID. To proceed to Association Rules study, we choose our best customers cluster which forms the Cluster 16. Cluster 16 has 1,886 records or customers. The input variables are divided into two different scenarios depending to the cases studied with the Association Rules. The first case is based on Original Activities using the Query "Original Activities Cluster 16". The second case is based on flight activities only using the query "Activities Cluster 16".

5.4.3.1 "Original Activities Cluster 16" Query

The 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 5.18 - 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
1746	1	1	0
2343	1	1	0
2984	1	1	0
3485	1	1	0
3964	1	1	0
4012	1	1	0
4023	1	1	0
4130	1	1	0
4454	1	1	0
4480	1	1	0
4664	1	1	0
5110	1	1	0
5250	1	1	0
5386	1	1	0
5460	0	1	1
5493	1	1	0
5983	1	1	0
6451	1	1	0
6462	1	1	0
6554	1	1	0
6753	1	1	0
6786	1	1	0
7081	1	1	0
7486	1	1	0
8061	1	1	0
8072	1	1	0
8083	1	1	0
8116	1	1	0
8186	1	1	0
8256	1	1	0
8400	1	1	0
8411	1	1	0
8422	0	1	1
8514	1	1	0

5.4.3.2 “Activities Cluster 16” Query

The query includes:

Customer ID.

145 fields including the Name of Sectors used by customers and originated from BEY or CDG.

We apply for Association Rules two different approaches. The first one will keep only the sectors which have a percentage of use greater than 10%. The second approach will manipulate only the sectors which have a percentage of use greater than 20%.

5.4.4 Apriori Algorithm Results

5.4.4.1 Algorithm Parameters

We implement the Apriori algorithm of ODM to build association models. The algorithm settings in the Apriori algorithm depend deeply of the marketing professional decision. They can decide which rules will be shown. The minimum support controls the rules displayed depending of the application percentage of this rule on existing data. Otherwise minimum confidence controls the displays of rules depending on the probability of having this rule in the future data. We will apply different parameters values on the data to show the difference that will be on commercial issues only. The default algorithm settings (Dunham, 2003) are as follows:

- **Minimum support:** Support of a rule is a measure of how frequently the items involved in it occur together. Using probability notation, support ($A \rightarrow B$) = $P(A, B)$. A real number between 0 and 1; smaller numbers result in faster builds. The default is 0.1.

- **Minimum confidence:** Confidence of a rule is the conditional probability of B given A; confidence $(A \rightarrow B) = P(B|A)$, which is equal to $P(A, B)$ or $P(A)$. Confidence of a rule is the conditional probability of the consequent given the antecedent. Confidence in the rules; a number between 0 and 1; high confidence results in a slower build. The default is 0.5.
- **Limit Number of Attributes in Each Rule:** If we specify this value, it is number between 2 and 100 that specify the maximum number of attributes in each rule; the default is to specify the limit as 3. If we don't want to specify any limit, we click the checkbox. If we increase the minimum support and confidence, we will limit the number of rules generated.

5.4.4.2 Scenario 1

For our first association rules run, we use the default values which prove to be the best. We choose a minimum support of 0.1, a minimum confidence of 0.5, and a limit number of attributes in each rule of 3. The first run is run based on “Original Activities Cluster 16” query. The execution time is less than 1 second (Start time: 10:55 PM – End time: 10:55 PM).

These statistical measures can be used to rank the rules and enhance the prediction.

The Table 5.19 displays the rules with support and confidentiality for each rule.

Table 5.19 - 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

5.4.4.3 Scenario 2

For our second association rules run, we use the default values. We choose a minimum support of 0.1, a minimum confidence of 0.5, and a limit number of attributes in each rule of 3. The second scenario is based on “Activities Cluster 16” query. In this round, 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. The execution time is less than 1 second (Start time: 11:00 PM – End time: 11:00 PM).

This scenario has 2,082 rules. The Table 5.20 displays the significant rules with support and confidentiality for each rule.

Table 5.20 - 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
132	BEYCAI=1	BEYDXB=1	0.6243655	0.19764328
257	BEYLCA=1	BEYDXB=1	0.6132597	0.11890734
158	BEYIST=1	BEYDXB=1	0.59178084	0.11569363
162	BEYLHR=1	BEYDXB=1	0.57326007	0.16764863
56	BEYCDG=1	BEYDXB=1	0.52029824	0.3363685

Another run on the same data will be applied by changing the default values. We choose a minimum support of 0.05, a minimum confidence of 0.25, and a limit number of attributes in each rule of 3. The execution time is less than 1 second. This scenario has 3,897 rules. We can note that the rules number has increased. The Table 5.21 displays some significant rules with support and confidentiality for each rule.

Table 5.21 - Association Rules for Best Customers Activities (Scenario 2)

Rule Id	If (condition)	Then (association)	Confidence	Support
1394	BEYJED=1 and BEYLHR=1	BEYCDG=1	0.90225565	0.06427424
1102	BEYCAI=1 and BEYFCO=1	BEYCDG=1	0.8947368	0.0546331
1409	BEYKWI=1 and BEYRUH=1	BEYCDG=1	0.88785046	0.05088377
3066	BEYCAI=1 and BEYJED=1	BEYCDG=1	0.8797814	0.0862346
551	BEYAMM=1 and BEYJED=1	BEYCDG=1	0.87586206	0.06802357
1283	BEYDXB=1 and BEYFCO=1	BEYCDG=1	0.8757764	0.07552223
3303	BEYJED=1 and BEYKWI=1	BEYCDG=1	0.8703704	0.05034815
1423	BEYLHR=1 and BEYRUH=1	BEYCDG=1	0.870229	0.061060525
1123	BEYCAI=1 and BEYRUH=1	BEYCDG=1	0.85882354	0.07820032
554	BEYAMM=1 and BEYKWI=1	BEYCDG=1	0.8552632	0.06963042
3063	BEYCAI=1 and BEYIST=1	BEYCDG=1	0.8545455	0.07552223
3075	BEYCAI=1 and BEYLHR=1	BEYCDG=1	0.84951454	0.09373326
2505	BEYAMM=1 and BEYLCA=1	BEYDXB=1	0.8476821	0.068559185
1294	BEYDXB=1 and BEYIST=1	BEYCDG=1	0.8472222	0.098018214
2464	BEYAMM=1 and BEYRUH=1	BEYCDG=1	0.84615386	0.07070166
1288	BEYDXB=1 and BEYGVA=1	BEYCDG=1	0.8445946	0.06695233
548	BEYAMM=1 and BEYIST=1	BEYCDG=1	0.8435374	0.06641671
1315	BEYDXB=1 and BEYRUH=1	BEYCDG=1	0.84347826	0.103910014
560	BEYAMM=1 and BEYLHR=1	BEYCDG=1	0.84313726	0.09212641
1303	BEYDXB=1 and BEYKWI=1	BEYCDG=1	0.8411215	0.096411355
3328	BEYLCA=1 and BEYLHR=1	BEYCDG=1	0.8362069	0.051955007
3308	BEYJED=1 and BEYRUH=1	BEYCDG=1	0.8333333	0.08302089
1300	BEYDXB=1 and BEYJED=1	BEYCDG=1	0.82520324	0.108730584
508	BEYAMM=1 and BEYCAI=1	BEYCDG=1	0.8233216	0.12479914
1114	BEYCAI=1 and BEYKWI=1	BEYCDG=1	0.8231707	0.07230852
2054	BEYFCO=1	BEYCDG=1	0.82287824	0.119442955
581	BEYAMM=1 and BEYKWI=1	BEYDXB=1	0.82236844	0.06695233
2499	BEYAMM=1 and BEYJED=1	BEYDXB=1	0.8206897	0.06373862
898	BEYAUH=1 and BEYCDG=1	BEYDXB=1	0.8161765	0.05945367
3054	BEYCAI=1 and BEYDXB=1	BEYCDG=1	0.8157182	0.1612212
3099	BEYCAI=1 and BEYKWI=1	BEYDXB=1	0.8109756	0.07123728
3217	BEYDXB=1 and BEYLHR=1	BEYCDG=1	0.8051118	0.1349759
536	BEYAMM=1 and BEYDXB=1	BEYCDG=1	0.7978142	0.15640064
1117	BEYCAI=1 and BEYLCA=1	BEYCDG=1	0.794702	0.06427424
1161	BEYCAI=1 and BEYLCA=1	BEYDXB=1	0.794702	0.06427424
593	BEYAMM=1 and BEYRUH=1	BEYDXB=1	0.78846157	0.065881096
1302	BEYCDG=1 and BEYKWI=1	BEYDXB=1	0.7826087	0.096411355
93	BEYGVA=1	BEYCDG=1	0.78039217	0.10658811
2057	BEYJED=1	BEYCDG=1	0.7751323	0.15693626
96	BEYIST=1	BEYCDG=1	0.77260274	0.15104446
1170	BEYCAI=1 and BEYRUH=1	BEYDXB=1	0.7705882	0.070166044

5.4.4.4 Scenario 3

For our third association rules run, we use the default values which prove to be the best. We choose a minimum support of 0.1, a minimum confidence of 0.5, and a limit number of attributes in each rule of 3. The third scenario is based on “Activities Cluster 16” query. In this round, we keep from the “Activities Cluster 16” query the sectors used by the customer with a percentage greater than 20%. The remaining number of sector field is 14 sectors. The execution time is less than 1 second (Start time: 10:48 PM – End time: 10:48 PM).

This scenario has 217 rules. The Table 5.22 displays the significant rules with support and confidentiality for each rule.

Another run on the same data will be applied by changing the default values. We choose a minimum support of 0.20, a minimum confidence of 0.75, and a limit number of attributes in each rule of 3. The execution time is less than 1 second. This scenario has 61 rules. We can note that the rules number has decreased. The Table 5.23 displays some significant rules with support and confidentiality for each rule.

Table 5.22 - Association Rules for Best Customers Activities (Scenario 3)

Rule Id	If (condition)	Then (association)	Confidence	Support
171	BEYCAI=1 and BEYDXB=1	BEYAMM=1	0.5799458	0.11462239
168	BEYCAI=1 and BEYCDG=1	BEYAMM=1	0.5307517	0.12479914
167	BEYAMM=1 and BEYCDG=1	BEYCAI=1	0.6005155	0.12479914
170	BEYAMM=1 and BEYDXB=1	BEYCAI=1	0.58469945	0.11462239
97	BEYDXB=1 and BEYJED=1	BEYCDG=1	0.82520324	0.108730584
166	BEYAMM=1 and BEYCAI=1	BEYCDG=1	0.8233216	0.12479914
192	BEYCAI=1 and BEYDXB=1	BEYCDG=1	0.8157182	0.1612212
208	BEYDXB=1 and BEYLHR=1	BEYCDG=1	0.8051118	0.1349759
175	BEYAMM=1 and BEYDXB=1	BEYCDG=1	0.7978142	0.15640064
20	BEYJED=1	BEYCDG=1	0.7751323	0.15693626
9	BEYAMM=1	BEYCDG=1	0.7504836	0.20782003
125	BEYCAI=1	BEYCDG=1	0.7428088	0.23513658
128	BEYDXB=1	BEYCDG=1	0.71935856	0.3363685
131	BEYLHR=1	BEYCDG=1	0.7161172	0.2094269
62	BEYAMM=1 and BEYLHR=0	BEYCDG=1	0.69009584	0.11569363
169	BEYAMM=1 and BEYCAI=1	BEYDXB=1	0.75618374	0.11462239
174	BEYAMM=1 and BEYCDG=1	BEYDXB=1	0.7525773	0.15640064
10	BEYAMM=1	BEYDXB=1	0.7079304	0.19603643
96	BEYCDG=1 and BEYJED=1	BEYDXB=1	0.69283277	0.108730584
191	BEYCAI=1 and BEYCDG=1	BEYDXB=1	0.6856492	0.1612212
138	BEYJED=1	BEYDXB=1	0.6507937	0.13176219
17	BEYCAI=1	BEYDXB=1	0.6243655	0.19764328
26	BEYLHR=1	BEYDXB=1	0.57326007	0.16764863
127	BEYCDG=1	BEYDXB=1	0.52029824	0.3363685

Table 5.23 - Association Rules for Best Customers Activities (Scenario 3)

Rule Id	If (condition)	Then (association)	Confidence	Support
5	BEYAMM=1	BEYCDG=1	0.7504836	0.20782003

5.4.5 Predictive Apriori Algorithm Results

5.4.5.1 Algorithm Parameters

We implement the Predictive Apriori algorithm of WEKA to build association models. It permits to found association rules stored by predictive accuracy. The default algorithm settings are as follows:

- **Num Rules:** Number of rules to find. The default is 100.

5.4.5.2 Scenario1

For association rules run, we use the default values. The run is based on “Activities Cluster 16” query. In this round, 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 with 1,209 records. The execution time is 45 minutes. This scenario has 100 rules as per default value. The Table 5.24 displays the best rules with the accuracy for each rule.

Table 5.24 – Predictive Apriori Result (Scenario 1)

1.	BEYDXB=BEYDXB BEYGVA=BEYGVA 15 ==> BEYCDG=BEYCDG 15	acc:(0.99488)
2.	BEYGVA=BEYGVA BEYLHR=BEYLHR 12 ==> BEYCDG=BEYCDG 12	acc:(0.99476)
3.	BEYLCA=BEYLCA 10 ==> BEYCDG=BEYCDG 10	acc:(0.99456)
4.	BEYCAI=BEYCAI BEYGVA=BEYGVA 9 ==> BEYCDG=BEYCDG 9	acc:(0.99435)
5.	BEYDXB=BEYDXB BEYRUH=BEYRUH 9 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 9	acc:(0.99435)
6.	BEYAMM=BEYAMM BEYCDG=BEYCDG BEYRUH=BEYRUH 8 ==> BEYDXB=BEYDXB 8	acc:(0.99435)
7.	BEYAUH=BEYAUH BEYGVA=BEYGVA 8 ==> BEYCDG=BEYCDG 8	acc:(0.994)
8.	BEYGVA=BEYGVA BEYIST=BEYIST 8 ==> BEYCDG=BEYCDG 8	acc:(0.994)
9.	BEYLHR=BEYLHR BEYRUH=BEYRUH 8 ==> BEYCDG=BEYCDG 8	acc:(0.994)
10.	BEYAUH=BEYAUH BEYDXB=BEYDXB BEYLHR=BEYLHR 8 ==> BEYCDG=BEYCDG 8	acc:(0.994)
11.	BEYAMM=BEYAMM BEYFCO=BEYFCO 7 ==> BEYCDG=BEYCDG 7	acc:(0.99338)
12.	BEYAMM=BEYAMM BEYJED=BEYJED 7 ==> BEYCDG=BEYCDG 7	acc:(0.99338)
13.	BEYAUH=BEYAUH BEYFCO=BEYFCO 7 ==> BEYDXB=BEYDXB 7	acc:(0.99338)
14.	BEYCAI=BEYCAI BEYIST=BEYIST 7 ==> BEYAMM=BEYAMM 7	acc:(0.99338)
15.	BEYFCO=BEYFCO BEYLHR=BEYLHR 7 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 7	acc:(0.99338)
16.	BEYGVA=BEYGVA BEYRUH=BEYRUH 7 ==> BEYCDG=BEYCDG 7	acc:(0.99338)
17.	BEYCAI=BEYCAI BEYDXB=BEYDXB BEYGVA=BEYGVA 7 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 7	acc:(0.99338)
18.	BEYCAI=BEYCAI BEYGVA=BEYGVA BEYLHR=BEYLHR 7 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 7	acc:(0.99338)
19.	BEYAMM=BEYAMM BEYAUH=BEYAUH 6 ==> BEYDXB=BEYDXB 6	acc:(0.99225)
20.	BEYAUH=BEYAUH BEYCAI=BEYCAI 6 ==> BEYCDG=BEYCDG 6	acc:(0.99225)
21.	BEYDXB=BEYDXB BEYKWI=BEYKWI 6 ==> BEYCDG=BEYCDG 6	acc:(0.99225)
22.	BEYFCO=BEYFCO BEYGVA=BEYGVA 6 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.99225)
23.	BEYAMM=BEYAMM BEYGVA=BEYGVA BEYRUH=BEYRUH 6 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.99225)
24.	BEYAMM=BEYAMM BEYIST=BEYIST BEYLHR=BEYLHR 6 ==> BEYCAI=BEYCAI 6	acc:(0.99225)
25.	BEYAMM=BEYAMM BEYLHR=BEYLHR BEYRUH=BEYRUH 6 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.99225)
26.	BEYAMM=BEYAMM BEYKWI=BEYKWI 5 ==> BEYCAI=BEYCAI BEYRUH=BEYRUH 5	acc:(0.99012)
27.	BEYAMM=BEYAMM BEYLCA=BEYLCA 5 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.99012)
28.	BEYAUH=BEYAUH BEYIST=BEYIST 5 ==> BEYDXB=BEYDXB 5	acc:(0.99012)
29.	BEYCAI=BEYCAI BEYFCO=BEYFCO 5 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.99012)
30.	BEYCAI=BEYCAI BEYKWI=BEYKWI 5 ==> BEYAMM=BEYAMM BEYRUH=BEYRUH 5	acc:(0.99012)
31.	BEYCAI=BEYCAI BEYLCA=BEYLCA 5 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.99012)
32.	BEYFCO=BEYFCO BEYKWI=BEYKWI 5 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.99012)
33.	BEYGVA=BEYGVA BEYJED=BEYJED 5 ==> BEYCDG=BEYCDG 5	acc:(0.99012)
34.	BEYLCA=BEYLCA BEYLHR=BEYLHR 5 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.99012)
35.	BEYCAI=BEYCAI BEYGVA=BEYGVA BEYIST=BEYIST 5 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 5	acc:(0.99012)
36.	BEYCAI=BEYCAI BEYGVA=BEYGVA BEYIST=BEYIST 5 ==> BEYAMM=BEYAMM BEYLHR=BEYLHR 5	acc:(0.99012)
37.	BEYCAI=BEYCAI BEYGVA=BEYGVA BEYRUH=BEYRUH 5 ==> BEYAMM=BEYAMM BEYDXB=BEYDXB 5	acc:(0.99012)
38.	BEYCAI=BEYCAI BEYGVA=BEYGVA BEYRUH=BEYRUH 5 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 5	acc:(0.99012)
39.	BEYAMM=BEYAMM BEYCAI=BEYCAI BEYGVA=BEYGVA BEYIST=BEYIST 5 ==> BEYCDG=BEYCDG BEYLHR=BEYLHR 5	acc:(0.99012)
40.	BEYDXB=BEYDXB BEYFCO=BEYFCO 14 ==> BEYCDG=BEYCDG 13	acc:(0.98769)
41.	BEYAUH=BEYAUH BEYJED=BEYJED 4 ==> BEYCDG=BEYCDG 4	acc:(0.986)
42.	BEYAUH=BEYAUH BEYRUH=BEYRUH 4 ==> BEYCDG=BEYCDG BEYGVA=BEYGVA 4	acc:(0.986)
43.	BEYCAI=BEYCAI BEYJED=BEYJED 4 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 4	acc:(0.986)
44.	BEYFCO=BEYFCO BEYJED=BEYJED 4 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 4	acc:(0.986)
45.	BEYFCO=BEYFCO BEYRUH=BEYRUH 4 ==> BEYAMM=BEYAMM BEYCAI=BEYCAI 4	acc:(0.986)
46.	BEYFCO=BEYFCO BEYRUH=BEYRUH 4 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 4	acc:(0.986)
47.	BEYGVA=BEYGVA BEYKWI=BEYKWI 4 ==> BEYCDG=BEYCDG 4	acc:(0.986)
48.	BEYGVA=BEYGVA BEYLCA=BEYLCA 4 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 4	acc:(0.986)
49.	BEYIST=BEYIST BEYKWI=BEYKWI 4 ==> BEYCDG=BEYCDG 4	acc:(0.986)
50.	BEYIST=BEYIST BEYRUH=BEYRUH 4 ==> BEYAMM=BEYAMM BEYCAI=BEYCAI 4	acc:(0.986)
51.	BEYIST=BEYIST BEYRUH=BEYRUH 4 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 4	acc:(0.986)
52.	BEYJED=BEYJED BEYLCA=BEYLCA 4 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 4	acc:(0.986)
53.	BEYJED=BEYJED BEYRUH=BEYRUH 4 ==> BEYCDG=BEYCDG 4	acc:(0.986)
54.	BEYKWI=BEYKWI BEYLHR=BEYLHR 4 ==> BEYCDG=BEYCDG 4	acc:(0.986)
55.	BEYAMM=BEYAMM BEYAUH=BEYAUH BEYGVA=BEYGVA 4 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 4	acc:(0.986)
56.	BEYAMM=BEYAMM BEYCAI=BEYCAI BEYFCO=BEYFCO 4 ==> BEYCDG=BEYCDG BEYRUH=BEYRUH 4	acc:(0.986)
57.	BEYAMM=BEYAMM BEYCAI=BEYCAI BEYFCO=BEYFCO 4 ==> BEYDXB=BEYDXB BEYRUH=BEYRUH 4	acc:(0.986)
58.	BEYAMM=BEYAMM BEYCDG=BEYCDG BEYKWI=BEYKWI 4 ==> BEYCAI=BEYCAI BEYDXB=BEYDXB 4	acc:(0.986)
59.	BEYAMM=BEYAMM BEYCDG=BEYCDG BEYKWI=BEYKWI 4 ==> BEYDXB=BEYDXB BEYRUH=BEYRUH 4	acc:(0.986)
60.	BEYAMM=BEYAMM BEYGVA=BEYGVA 13 ==> BEYCDG=BEYCDG 12	acc:(0.98483)
61.	BEYAMM=BEYAMM BEYLHR=BEYLHR 13 ==> BEYCDG=BEYCDG 12	acc:(0.98483)
62.	BEYAMM=BEYAMM BEYLHR=BEYLHR 13 ==> BEYDXB=BEYDXB 12	acc:(0.98483)
63.	BEYAMM=BEYAMM BEYCDG=BEYCDG BEYGVA=BEYGVA 12 ==> BEYDXB=BEYDXB 11	acc:(0.97989)
64.	BEYAUH=BEYAUH BEYLCA=BEYLCA 3 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 3	acc:(0.97775)
65.	BEYJED=BEYJED 11 ==> BEYCDG=BEYCDG 10	acc:(0.97176)
66.	BEYCAI=BEYCAI BEYLHR=BEYLHR 11 ==> BEYCDG=BEYCDG 10	acc:(0.97176)
67.	BEYAMM=BEYAMM BEYRUH=BEYRUH 10 ==> BEYCAI=BEYCAI 9	acc:(0.95823)
68.	BEYAMM=BEYAMM BEYRUH=BEYRUH 10 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 9	acc:(0.95823)
69.	BEYAUH=BEYAUH BEYLHR=BEYLHR 10 ==> BEYCDG=BEYCDG 9	acc:(0.95823)
70.	BEYCAI=BEYCAI BEYRUH=BEYRUH 10 ==> BEYAMM=BEYAMM 9	acc:(0.95823)
71.	BEYCAI=BEYCAI BEYRUH=BEYRUH 10 ==> BEYCDG=BEYCDG 9	acc:(0.95823)
72.	BEYGVA=BEYGVA 20 ==> BEYCDG=BEYCDG 18	acc:(0.94466)
73.	BEYDXB=BEYDXB BEYLHR=BEYLHR 20 ==> BEYCDG=BEYCDG 18	acc:(0.94466)
74.	BEYCAI=BEYCAI BEYGVA=BEYGVA 9 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 8	acc:(0.93589)
75.	BEYDXB=BEYDXB BEYRUH=BEYRUH 9 ==> BEYAMM=BEYAMM BEYCAI=BEYCAI 8	acc:(0.93589)
76.	BEYDXB=BEYDXB BEYRUH=BEYRUH 9 ==> BEYCAI=BEYCAI BEYCDG=BEYCDG 8	acc:(0.93589)
77.	BEYIST=BEYIST BEYLHR=BEYLHR 9 ==> BEYCDG=BEYCDG 8	acc:(0.93589)
78.	BEYIST=BEYIST BEYLHR=BEYLHR 9 ==> BEYDXB=BEYDXB 8	acc:(0.93589)
79.	BEYFCO=BEYFCO 19 ==> BEYCDG=BEYCDG 17	acc:(0.93178)
80.	BEYAMM=BEYAMM BEYIST=BEYIST 8 ==> BEYCAI=BEYCAI 7	acc:(0.90025)
81.	BEYAMM=BEYAMM BEYIST=BEYIST 8 ==> BEYCDG=BEYCDG 7	acc:(0.90025)
82.	BEYGVA=BEYGVA BEYIST=BEYIST 8 ==> BEYCDG=BEYCDG BEYLHR=BEYLHR 7	acc:(0.90025)
83.	BEYGVA=BEYGVA BEYIST=BEYIST 8 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 7	acc:(0.90025)
84.	BEYAMM=BEYAMM BEYFCO=BEYFCO 7 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.84723)
85.	BEYAMM=BEYAMM BEYJED=BEYJED 7 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.84723)
86.	BEYAUH=BEYAUH BEYFCO=BEYFCO 7 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 6	acc:(0.84723)
87.	BEYCAI=BEYCAI BEYIST=BEYIST 7 ==> BEYAMM=BEYAMM BEYDXB=BEYDXB 6	acc:(0.84723)
88.	BEYCAI=BEYCAI BEYIST=BEYIST 7 ==> BEYAMM=BEYAMM BEYCDG=BEYCDG 6	acc:(0.84723)
89.	BEYLHR=BEYLHR 24 ==> BEYCDG=BEYCDG 21	acc:(0.82923)
90.	BEYDXB=BEYDXB 38 ==> BEYCDG=BEYCDG 32	acc:(0.79727)
91.	BEYRUH=BEYRUH 14 ==> BEYCDG=BEYCDG 12	acc:(0.78606)
92.	BEYAUH=BEYAUH BEYCDG=BEYCDG 14 ==> BEYDXB=BEYDXB 12	acc:(0.78606)
93.	BEYCAI=BEYCAI BEYDXB=BEYDXB 14 ==> BEYAMM=BEYAMM 12	acc:(0.78606)
94.	BEYCDG=BEYCDG BEYLHR=BEYLHR 21 ==> BEYDXB=BEYDXB 18	acc:(0.77687)
95.	BEYAMM=BEYAMM BEYAUH=BEYAUH 6 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.77678)
96.	BEYAUH=BEYAUH BEYCAI=BEYCAI 6 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 5	acc:(0.77678)
97.	BEYLHR=BEYLHR 24 ==> BEYDXB=BEYDXB 20	acc:(0.74469)
98.	BEYAMM=BEYAMM BEYGVA=BEYGVA 13 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 11	acc:(0.73938)
99.	BEYAMM=BEYAMM BEYLHR=BEYLHR 13 ==> BEYCDG=BEYCDG BEYDXB=BEYDXB 11	acc:(0.73938)
100.	BEYCAI=BEYCAI 19 ==> BEYCDG=BEYCDG 16	acc:(0.73278)

Chapter 6

DISCUSSION OF RESULTS

6.1 Overview

After running the cluster and association rules algorithms, the next step is to characterize the clusters and rules qualitatively. The characterization of result qualitatively is done by profiling. The purpose of profiling is to assess the potential business value of each cluster and rules quantitatively by profiling the aggregate values of the variables by cluster and rules.

6.2 Discussion of K-Means Clustering Results

The result of scenarios is presented in different tables. Each table provides a review of the profile of revenue mileage, number of services used, and customer membership period. The Weight column is a fraction of revenue mileage per customer. The service index is the fraction of the average number of services used by the customers in the cluster divided by the average number of services purchased overall. In summary, we have used the following parameters for evaluation:

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

6.2.1 Scenario 1 Result

Table 6.1 shows the result of k-means clustering scenario 1. The most profitable cluster is cluster 4. It groups about 53.12 percent of the mileage with only 28.04 percent of the passengers. This cluster has the highest weight fraction.

A valuable business opportunity is shown in the cluster profile. This opportunity is identified by increasing the number of services used by passengers.

It is obvious that cluster 4 contains the best customers. These passengers have a higher mileage per passenger than other clusters, as shown by the weight column. Some possible strategies include:

- A retention strategy for best customers (those in clusters 4).
- A cross-sell strategy for cluster 6 by contrasting with cluster 4. Cluster 6 has a service index close to this of cluster 4. Cluster 4 has the highest number of services used. The effort needed to convert passengers from cluster 6 to cluster 4 should be minimal, since both clusters are close in number of services used. The comparison of services bought by the best passengers to those purchased to that in cluster 6; we can predefine services that are candidates for cross-selling.
- The same cross-selling strategies are applied between 7 and 5 because they are close in services value.

Table 6.1 - Clustering Analyst for K-Means Algorithm (Scenario 1)

Cluster ID	Mileage %	Customers %	Avg. Services per Cluster	Service Index	Weight	Membership (Sum Membership/ Customer)
4	53.12	28.04		1.04	1.89	60.84
6	30.82	40.38		1.02	0.76	18.69
5	10.42	18.06		0.94	0.58	62.51
7	5.64	13.52		0.93	0.42	29.26

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

1.03

Best Customers: Clusters 4 (Higher Weight)
Strategies:
Retention strategy for best customers.
Cross-Sell strategy for clusters 6 by contrasting with cluster 4
(Service Index is close. By comparing which services is used by the best customers)
Cross-Sell strategy for clusters 7 by contrasting with cluster 5
(Service Index is close. By comparing which services is used by the best customers)
New Customers and worst cluster are not shown

6.2.2 Scenario 2 Result

Table 6.2 shows the result of the second scenario. It provides the same review as the Scenario 2. The most profitable cluster is cluster 10. It groups about 35.89 percent of the mileage with only 19.09 percent of the passengers. This cluster has the highest weight fraction.

It is obvious that clusters 10 and 11 contain the best customers. These passengers have a higher mileage per passenger than other clusters, as shown by the weight column. Some possible strategies include:

- A retention strategy for best customers (those in clusters 10, and 11).
- A cross-sell strategy for cluster 8 and 9 by contrasting with cluster 10. Cluster 8 and 9 has a service index close to this of cluster 10. Cluster 10 has the highest number of services used. The effort needed to convert passengers from cluster 8 and 9 to cluster 10 should be minimal, since both clusters are close in number of services used. The comparison of services bought by the best passengers to those purchased to that in cluster 8 and 9; we can predefine services that are candidates for cross-selling.
- The same cross-selling strategies are applied between 5, 7 and 11 because they are close in services value.
- The new customers and the worst cluster are not shown.

Table 6.2 - Clustering Analyst for K-Means Algorithm (Scenario 2)

Cluster ID	Mileage %	Customers %	Avg. Services per Customer	Service Index	Weight	Membership (Sum Membership/Nb. Customer)
10	35.89	16.09	1.074	2.23	69.40	
8	21.43	19.64	1.052	1.09	27.91	
11	16.01	11.56	0.971	1.38	49.50	
9	10.76	21.18	1.004	0.51	10.69	
5	9.55	16.28	0.940	0.59	64.33	
7	6.37	15.24	0.937	0.42	31.12	

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

Best Customers:	Clusters 10 - 11 (Higher Mileage per Customer)
Strategies:	Retention strategy for best customers.
	Cross-Sell strategy for clusters 8 - 9 by contrasting with cluster 10 (Service Index is close. By comparing which services is used by the best customers)
	Cross-Sell strategy for clusters 5 - 7 by contrasting with cluster 11 (Service Index is close. By comparing which services is used by the best customers)

6.2.3 Scenario 3 Result

Table 6.3 shows the result of k-means clustering scenario 3. The most profitable cluster is cluster 16. It groups about 8.88 percent of the mileage with only 3.71 percent of the passengers. This cluster has the highest weight fraction.

A valuable business opportunity is shown in the cluster profile. This opportunity is identified by 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. Some possible strategies include:

- A retention strategy for best customers (those in clusters 11, 16, and 17).
- A cross-sell strategy for cluster 8 by contrasting with cluster 16. Cluster 8 has a service index close to this 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 to those purchased to that in cluster 8, we can predefine services that are candidates for cross-selling.
- The same cross-selling strategies are applied between 15 and 11; 13 and 17 because they are close in services value.
- The cluster 9 has to be observed very closely during some period of time. It defines a group of new passengers. We have to collect sufficient data to determine the behaviors 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 to accelerate profitability.
- The clusters 12 are the worst cluster, since they have a very low mileage percentage. These passengers use very few services even though they have been with the

company for 37 months. The strategy will be to minimize spending on any significant marketing on them.

Table 6.3 - Clustering Analyst for K-Means Algorithm (Scenario 3)

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

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

1.03

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)	

6.2.4 Scenario 4 Result

Table 6.4 shows the result of k-means clustering scenario 4. The most profitable cluster is cluster 16. It groups about 4.74 percent of the mileage with only 1.52 percent of the passengers. This cluster has the highest weight fraction.

A valuable business opportunity is shown in the cluster profile. This opportunity is identified by increasing the number of services used by passengers.

It is obvious that clusters 16, 17, and 18 contain the best customers. These passengers have a higher mileage per passenger than other clusters, as shown by the weight column. Some possible strategies include:

- A retention strategy for best customers (those in clusters 16, 17, and 18).
- A cross-sell strategy for cluster 22 by contrasting with cluster 18. Cluster 22 has a service index close to this of cluster 18. Cluster 18 has the highest number of services used. The effort needed to convert passengers from cluster 22 to cluster 18 should be minimal, since both clusters are close in number of services used. The comparison of services bought by the best passengers to those purchased to that in cluster 22, we can predefine services that are candidates for cross-selling.
- The same cross-selling strategies are applied between 15, 19, 20, 21, 23, 24, 26, 28 and 16 because they are close in services value.
- The clusters 25, 27, and 29 are the worst cluster, since they have a very low mileage percentage. These passengers use very few services even though they have been with the company for long time. The strategy will be to minimize spending on any significant marketing on them.

Table 6.4 - Clustering Analyst for K-Means Algorithm (Scenario 4)

Cluster ID	Mileage %	Customers %	Avg. Services per Customer	Service Index	Weight	Membership (Sum Membership/ NB. Customer)
17	32.25	15.22		0.97	2.12	69.02
23	16.59	16.04		0.97	1.03	28.58
19	14.30	10.72		0.97	1.33	48.66
21	9.61	17.16		0.97	0.66	13.26
26	6.05	9.43		0.97	0.64	69.82
16	4.74	1.52		1.94	3.12	69.42
15	3.67	7.28		0.97	0.68	54.53
24	3.52	7.63		0.97	0.46	36.81
28	2.46	6.16		0.97	0.40	24.23
18	2.33	0.99	2.33	1.97	2.35	44.14
22	2.13	1.45	2.01	1.95	1.47	17.26
20	2.00	5.33		0.97	0.38	5.34
27	0.11	0.51		0	0.22	71.02
29	0.03	0.12		0	0.25	12.98
25	0.02	0.44		0	0.05	44.89

Best Customers:	Clusters 16 -17-18 (Higher Mileage per Customer)
Strategies:	
Retention strategy for best customers.	
Cross-Sell strategy for clusters 22 by contrasting with cluster 18 (Service Index is close. By comparing which services is used by the best customers)	
Cross-Sell strategy for clusters 15-19-20-21-23-24-26-28 by contrasting with cluster 16 (Service Index is close. By comparing which services is used by the best customers)	
Cluster 25-27-29 (The Worst Cluster)	

6.2.5 Comparison between Different Scenario

As we can conclude the third scenario lets view a clearer picture with more details concerning new and worst customers. The first two scenarios are not showing the new and worst customers. In this case, we will miss the opportunities with new valuable customers and we will spend marketing on customers without any future. The fourth scenario has many details and many clusters with same services index. The market professional will be confused. The real customer sales value is missed. To continue our study for parameters preference and association rules, we will use the scenario 3 results.

We will begin by evaluating the result of the most valuable scenario, scenario 3. We inspect some clusters in detail and detect similarity properties. We review the cluster 16 including the best customers. We compare the result to the company grouping and we can see that we have to enhance the company policies. We can see in table 6.5 a sample of data from cluster 16. The cluster 16 includes 1,886 customers. We found that the most important characteristics of this cluster are the high number of activities done during the lifetime of the customer and the activities done during the last year.

We review also another cluster, the cluster 12 including the new customers. We compare the result to the company grouping and we can see that all customers are in the Basic group which is true for new customers. We can see in table 6.6 a sample of data in the cluster 12. The cluster 12 includes 3,101 customers.

Another method to test our result is by applying the O-Cluster, EM, and COBWEB algorithms and compares their results with k-means result (Scenario 3). This will be discussed in the following sections.

Table 6.5 - Cluster 16 Sample

Cust. ID	probability	Act. Last Year	Act. Life	membership	mileage	ram	rmm	Current Tier
206	0.999723911	3	3	74	74132	0.04	1001.78	P
486	1	2	2	74	136617	0.03	1846.18	E
906	1	2	2	74	117644	0.03	1589.78	E
1352	1	2	2	74	35727	0.03	482.8	E
1746	0.988215029	1	2	74	71791	0.03	970.15	B
2343	1	2	2	74	154950	0.03	2093.92	P
2984	1	2	2	74	53488	0.03	722.81	P
3485	1	2	2	74	33940	0.03	458.65	B
3964	1	2	2	74	90764	0.03	1226.54	V
4012	1	2	2	74	190609	0.03	2575.8	E
4023	1	2	2	74	87483	0.03	1182.2	V
4130	1	2	2	74	42455	0.03	573.72	B
4454	1	2	2	74	50498	0.03	682.41	V
4480	1	2	2	74	66461	0.03	898.12	V
4664	0.968755186	1	2	74	8880	0.03	120	B
5110	1	2	2	74	17263	0.03	233.28	B
5250	1	2	2	74	110741	0.03	1496.5	P
5386	1	2	2	74	149421	0.03	2019.2	V
5460	1	2	2	74	146205	0.03	1975.74	E
5493	1	2	2	74	136713	0.03	1847.47	P
5983	1	2	2	74	115609	0.03	1562.28	E
6451	1	2	2	74	41134	0.03	555.86	E
6462	1	2	2	74	63533	0.03	858.55	P
6554	1	2	2	74	26981	0.03	364.61	B
6753	1	2	2	74	31559	0.03	426.47	P
6786	1	2	2	74	96332	0.03	1301.78	B
7081	1	2	2	74	30563	0.03	413.01	B
7486	1	2	2	74	34539	0.03	466.74	B
8061	1	2	2	74	215133	0.03	2907.2	V
8072	1	2	2	74	21155	0.03	285.88	V
8083	1	2	2	74	25489	0.03	344.45	B
8116	1	2	2	74	122069	0.03	1649.58	P
8186	1	2	2	74	270493	0.03	3655.31	V
8256	1	2	2	74	172704	0.03	2333.84	V
8400	1	2	2	74	57591	0.03	778.26	V
8411	1	2	2	74	154519	0.03	2088.09	V
8422	1	2	2	74	38570	0.03	521.22	V
8514	1	2	2	74	49580	0.03	670	V
8536	1	2	2	74	93590	0.03	1264.73	V

Table 6.6 - Cluster 12 Sample

custid	probability	actlastyear	actlife	membership	mileage	ram	rmm	CurrentTier
236946	0.990084469	0	0	54	13116	0	242.89	B
241555	0.998773396	0	0	53	3577	0	67.49	B
255194	0.999504924	0	0	51	1000	0	19.61	B
262942	0.999016282	0	2	49	17132	0.04	349.63	B
286473	0.999957919	0	0	46	1000	0	21.74	B
288886	0.999957919	0	0	46	3518	0	76.48	B
289100	0.999957919	0	0	46	1000	0	21.74	B
289166	0.999957919	0	0	46	1000	0	21.74	B
289192	0.999957919	0	0	46	1000	0	21.74	B
289251	0.999957919	0	0	46	1000	0	21.74	B
292795	0.999973655	0	0	45	1000	0	22.22	B
292821	0.999973655	0	0	45	1000	0	22.22	B
292854	0.999973655	0	0	45	1000	0	22.22	B
292865	0.999973655	0	0	45	1000	0	22.22	B
292880	0.999973655	0	0	45	1000	0	22.22	B
292946	0.999973655	0	0	45	1000	0	22.22	B
293005	0.999973655	0	0	45	1000	0	22.22	B
293016	0.999973655	0	0	45	1000	0	22.22	B
293031	0.999973655	0	0	45	1000	0	22.22	B
293145	0.999973655	0	0	45	1000	0	22.22	B
293171	0.999973655	0	0	45	1000	0	22.22	B
293182	0.999973655	0	0	45	1000	0	22.22	B
293193	0.999973655	0	0	45	1000	0	22.22	B
293215	0.999973655	0	0	45	1000	0	22.22	B
293226	0.999973655	0	0	45	1000	0	22.22	B
293230	0.999973655	0	0	45	1000	0	22.22	B
293252	0.999973655	0	0	45	1000	0	22.22	B
293285	0.999973655	0	0	45	1000	0	22.22	B
293300	0.999973655	0	0	45	1000	0	22.22	B
293333	0.999973655	0	0	45	1000	0	22.22	B
293355	0.999973655	0	0	45	1000	0	22.22	B
293414	0.999973655	0	0	45	1000	0	22.22	B
293425	0.999973655	0	0	45	1000	0	22.22	B
293451	0.999973655	0	0	45	1000	0	22.22	B
293462	0.999973655	0	0	45	1000	0	22.22	B
293473	0.999973655	0	0	45	1000	0	22.22	B
293484	0.999973655	0	0	45	1000	0	22.22	B
293495	0.999973655	0	0	45	1000	0	22.22	B
293506	0.999973655	0	0	45	1000	0	22.22	B
293510	0.999973655	0	0	45	1000	0	22.22	B
293554	0.999973655	0	0	45	1000	0	22.22	B
293580	0.999973655	0	0	45	1000	0	22.22	B

6.3 Discussion of O-Cluster Clustering Results

As discussed in the previous section, the best scenario for k-means clustering is the scenario 3. For this reason, we apply the O-Cluster algorithm with 9 clusters. The result of O-Cluster algorithm is presented in table 6.7. It provides a profile review including mileage, number of services used, and passenger membership period. The Weight column is a fraction of mileage per passenger. The service index is the fraction of the average number of services used by the passengers in the cluster divided by the average number of services used overall.

6.3.1 Scenario Result

Table 6.7 shows the result of o-cluster clustering scenario. The most profitable cluster is cluster 15. It groups about 36.37 percent of the mileage with only 18.13 percent of the passengers. This cluster has the highest weight fraction.

A valuable business opportunity is shown in the cluster profile. This opportunity is identified by increasing the number of services used by passengers.

It is obvious that clusters 14, 15, 16, and 17 contain the best customers. These passengers have a higher mileage per passenger than other clusters, as shown by the weight column. Some possible strategies include:

- A retention strategy for best customers (those in clusters 14, 15, 16, and 17).
- A cross-sell strategy for cluster 11 by contrasting with cluster 14. Cluster 11 has a service index close to this of cluster 14. Cluster 14 has the highest number of services used. The effort needed to convert passengers from cluster 11 to cluster 14 should be minimal, since both clusters are close in number of services used. The comparison of services bought by the best passengers to those purchased to that in cluster 11, we can predefine services that are candidates for cross-selling.
- The same cross-selling strategies are applied between 13 and 16; 7 and 17 because they are close in services value.

- Same strategy will be adopted for cluster 3 and 5. Those clusters have to be observed very closely during some period of time. They define a group of new passengers. We have to collect sufficient data to determine the behaviors of those new passengers. We have to adopt some marketing efforts to inform cluster 3 and 5 passengers of the Frequent Flyer program's products and services to accelerate profitability.

Table 6.7 – Clustering Analyst for O-Cluster Algorithm

Cluster ID	Mileage %	Customers %	Avg. Services per Customer	Service Index	Weight	Membership (Sum Membership/Nb. Customer)	K- Means Cluster ID
15	36.37	18.13	1.08	0.974	2.01	67.33	17
16	13.54	13.09	0.976	0.976	1.03	30.28	
17	12.24	11.10	0.974	1.297	1.10	35.60	16 - 13
14	11.33	8.26	0.974	0.974	1.37	45.90	11
11	10.35	20.21	0.974	0.920	0.51	59.74	15
13	7.88	11.78	0.974	0.974	0.67	18.84	14 - 8
7	6.21	11.96	0.981	0.981	0.52	11.33	9
5	1.52	3.74	0.972	0.972	0.40	6.17	
3	0.57	1.68	0.970	0.970	0.34	3.68	

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

1.03

Total Customers	50,830
Best Customers:	Clusters 14 - 15 - 16 - 17 (Higher Mileage per Customer)
Strategies:	
Retention strategy for best customers.	
Cross-Sell strategy for clusters 11 by contrasting with cluster 14	
(Service Index is close. By comparing which services is used by the best customers)	
Cross-Sell strategy for clusters 13 by contrasting with cluster 16	
(Service Index is close. By comparing which services is used by the best customers)	
Cross-Sell strategy for clusters 7 by contrasting with cluster 17	
(Service Index is close. By comparing which services is used by the best customers)	
Cluster 5 - 3 to wait (New Customers)	

6.3.2 Scoring (Apply Result)

Clusters discovered are used to generate a profitability model that is used during scoring (model apply) for assigning data points to clusters.

We apply the o-cluster clustering model Scenario on the data using the following parameters:

Values of highest cluster Ids.

The number of top cluster Ids is 1.

Sample of the result is shown in table 6.8.

6.3.3 Comparison O-Cluster to K-Means Scenario 3

Using the scoring technique, in k-means clustering (Scenario3) and O-cluster Algorithm, data points are assigned to clusters. Each customer will be assigned two different cluster numbers. One is due to k-means algorithm (scenario 3) and the second is due to o-cluster algorithm. With two cluster numbers, we can identify which customers with the two techniques (k-means and o-cluster) still in the same group or cluster. In Table 6.9, we have considered the k-means result as the base. And in each cluster, we have calculated the highest percentage of the o-cluster cluster in this group. Finally, we have added this percentage and divided them by the number of cluster (9). We concluded the average percentage of the similarity between k-means (Scenario 3) and O-Cluster. We have a similarity at 75.44 %. This percentage gives us a good evaluation of the result generated by the k-means algorithm (Scenario 3).

Table 6.8 – Sample of Applying O-Cluster Algorithm

Cluster_ID	Probability	ActLastYear	ActLife	CustID	Membership	Mileage	RAM	RMM
3	0.99998742	1	1	768,795	4	11,830	0.25	2,957.50
3	0.99996144	1	1	768,810	4	6,987	0.25	1,746.75
3	0.99996144	1	1	768,913	4	6,936	0.25	1,734
3	0.99992853	1	1	769,005	4	1,500	0.25	375
3	0.99994463	1	1	769,860	4	5,161	0.25	1,290.25
3	0.99994463	1	1	769,963	4	5,000	0.25	1,250
3	0.9999941	1	1	770,022	4	12,370	0.25	3,092.50
3	0.99377209	1	1	770,103	4	35,967	0.25	8,991.75
3	0.99996144	1	1	770,243	4	6,762	0.25	1,690.50
3	0.99993044	1	1	770,265	4	4,357	0.25	1,089.25
3	0.99993044	1	1	770,350	4	3,958	0.25	989.5
3	0.99992853	1	1	770,383	4	2,479	0.25	619.75
3	0.99377209	1	1	770,431	4	34,070	0.25	8,517.50
3	0.99997646	1	1	770,512	4	9,322	0.25	2,330.50
3	0.99993044	1	1	770,523	4	4,000	0.25	1,000
3	0.99998742	1	1	770,545	4	11,916	0.25	2,979
3	0.99993163	1	1	770,630	4	500	0.25	125
3	0.99993044	1	1	770,652	4	4,528	0.25	1,132
3	0.99997646	1	1	770,696	4	8,958	0.25	2,239.50
3	0.99994463	1	1	770,766	4	6,210	0.25	1,552.50
3	0.99998742	1	1	771,024	4	10,916	0.25	2,729
3	0.99992853	1	1	771,072	4	1,230	0.25	307.5
3	0.99920106	1	1	771,142	4	29,050	0.25	7,262.50
3	0.99994463	1	1	771,396	4	5,299	0.25	1,324.75
3	0.99994463	1	1	771,411	4	5,125	0.25	1,281.25
3	0.99993163	1	1	771,562	4	800	0.25	200
3	0.99993163	1	1	771,573	4	615	0.25	153.75
3	0.99993044	1	1	771,761	4	4,125	0.25	1,031.25
3	0.99994463	1	1	771,956	4	5,036	0.25	1,259
3	0.99992853	1	1	772,026	4	1,605	0.25	401.25
3	0.99996144	1	1	772,144	4	8,121	0.25	2,030.25
3	0.99998742	1	1	772,236	4	10,936	0.25	2,734
3	0.99992597	1	1	772,284	4	23,380	0.25	5,845
3	0.99998742	1	1	772,295	4	11,238	0.25	2,809.50
3	0.99996144	1	1	772,306	4	8,458	0.25	2,114.50
3	0.99993163	1	1	772,310	4	1,018	0.25	254.5
3	0.99999976	1	1	772,365	4	21,040	0.25	5,260
3	0.99992853	1	1	772,402	4	2,036	0.25	509

Table 6.9 - Comparison between K-Means (Scenario3) and O-Cluster			
cluster_KMeans	cluster_OCluster	CountOfcluster_OCluster	Percentage
8	7	267	3.65303051
8	13	4248	58.12012587
8	14	1	0.013681762
8	16	2777	37.99425366
8	17	16	0.218908195
		7309	
9	3	856	10.10029499
9	5	1901	22.43067847
9	7	5715	67.43362832
9	11	1	0.01179941
9	13	2	0.02359882
		8475	
11	14	4113	48.57682768
11	15	419	4.94862407
11	16	3878	45.8013464
11	17	57	0.673201842
		8467	
12	11	1176	37.92325056
12	14	7	0.225733634
12	17	1918	61.8510158
		3101	
13	7	98	2.660152009
13	11	34	0.922909881
13	13	1736	47.12269273
13	17	1816	49.29424539
		3684	
14	11	4326	97.19164233
14	15	125	2.808357672
		4451	
15	11	4729	98.41831426
15	14	45	0.936524454
15	15	31	0.64516129
		4805	
16	5	1	0.053022269
16	13	2	0.106044539
16	14	7	0.371155885
16	15	19	1.007423118
16	17	1857	98.46235419
		1886	
17	11	6	0.069348128
17	14	26	0.300508553
17	15	8620	99.63014332
		8652	
Same grouping at:		75.44203302	%

6.3.4 Best Route from CDG

The result of clustering was used to prepare data for Association Rules. The clustering result of k-means scenario 3 was used as foundation to continue our study. As shown before, based on our best customers (Cluster 16) we have prepared the query “Activities Cluster 16”. The “Activities Cluster 16” 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 our second hub (CDG). This approach will be applied on the result given by the k-means algorithms K-means scenario 3. Table 6.10 shows the best routes dedicated from the k-means algorithm scenario 3.

On other side, the application of o-cluster algorithm shows that our best customers are in the cluster 15. Based on cluster 15, a query called “Activities Cluster 15” will be build. The query includes:

Customer ID.

148 fields including the Name of Sectors used by customers and originated from BEY or CDG.

In parallel, Table 6.11 shows the best routes derived from the o-cluster algorithm.

By comparing the result derived from table 6.10 (k-means – scenario 3) and table 6.11 (o-cluster), we found that the preferable routing in both table are the same. The only difference is the percentage of use.

Table 6.12 shows the percentage comparison between the two algorithms. The result demonstrates that the best route in both methods is almost the same. The difference can be shown in the percentage.

Table 6.10 - Best Route Originated from CDG with K-Means Algorithm (Scenario 3)

CDGBES	0.0829		CDGAMM	0.3314		CDGTXL	1.159901
CDGBKK			CDGBUD			CDGYYZ	
CDGBOG			CDGGOT			CDGZRH	
CDGBZV			CDGHAJ				
CDGCCS			CDGZYR			CDGCPH	1.242751
CDGCFE						CDGSFO	
CDGDLA			CDGABJ	0.41425		CDGVIE	
CDGDTW			CDGCKY				
CDGEZE			CDGLIN			CDGLYS	1.325601
CDGFIH			CDGOSL				
CDGFNI			CDGSXB			CDGORD	1.574151
CDGFRL			CDGTRN				
CDGGIG						CDGDXB	1.657001
CDGHKG			CDGATH	0.4971		CDGIAH	
CDGKWI			CDGCGN			CDGMXP	
CDGLBV			CDGCVG				
CDGLED			CDGMPL			CDGATL	1.739851
CDGNAP			CDGPHL			CDGFCO	
CDGNKC			CDGRBA			CDGLIS	
CDGNSI			CDGTLS				
CDGOUA						CDGBOS	1.822701
CDGPHC			CDGBOD	0.57995			
CDGSCL			CDGHAV			CDGFRA	1.905551
CDGVRN							
CDGXDB			CDGBRU	0.6628		CDGAMS	2.071251
CDGBIQ	0.1657		CDGDKR				
CDGCAN			CDGLOS			CDGMUC	2.154101
CDGHEL			CDGVCE				
CDGLFW			CDGVLC			CDGMAD	2.651201
CDGMAN			CDGWAW				
CDGMLH			CDGMEX	0.74565		CDGAGP	2.734051
CDGPEK			CDGSTR			CDGDUS	
CDGRUH			CDGBLQ	0.8285		CDGTUN	2.816901
CDGSOF			CDGBSL			CDGCMN	2.899751
CDGCOO	CDGPSA	0.24855	CDGCAI			CDGIAD	2.982601
CDGDFD	CDGPUF		CDGARN	0.91135		CDGLAX	3.231152
CDGFLR	CDGPVG		CDGGRU			CDGBCN	3.728252
CDGICN	CDGSVO		CDGPRG			CDGLHR	4.722452
CDGIST	CDGOPO		CDGHAM	0.9942		CDGYUL	
CDGJED			CDGMIA			CDGGVA	5.799503
CDGJNB			CDGEWR	1.077051		CDGJFK	
CDGNRT			CDGMRS			CDGNCE	10.35626

Table 6.11 - Best Route Originated from CDG with O-Cluster Algorithm

CDGBOM	0.02	CDGBES	0.12	CDGSXB	0.36	CDGEWR	1.00
CDGCMF		CDGBIQ		CDGAMM	0.38	CDGVCE	
CDGFMO		CDGBZV		CDGBOD		CDGSFO	1.02
CDGKRK		CDGFIH		CDGGIG			
CDGLCY		CDGFLR				CDGARN	1.18
CDGLJU		CDGHAJ		CDGBSL	0.40	CDGATL	
CDGNIM		CDGLBV				CDGLYS	
CDGNSI		CDGDTW	0.14	CDGVLC	0.44		
CDGNUE		CDGNTE		CDGWAW		CDGCAI	1.20
CDGPTP		CDGPSA				CDGRUH	
CDGSOF		CDGPUF		CDGATH	0.46		
CDGXDB		CDGPVG		CDGBHX		CDGZRH	1.30
CDGBGF	0.04			CDGJNB			
CDGDAM		CDGNKC	0.16			CDGCPH	1.34
CDGDFW				CDGCVG	0.52		
CDGDOH		CDGDLA	0.18	CDGGRU		CDGBLQ	1.36
CDGMLH		CDGGOT		CDGMAN		CDGLAH	
CDGMNL		CDGICN				CDGLIS	
CDGMRU		CDGNCL		CDGZXR	0.54		
CDGRAK		CDGPHC				CDGORD	1.44
CDGSHA		CDGVRN		CDGPRG	0.56	CDGFCO	1.48
CDGSIN						CDGTXL	
CDGSXM		CDGCGN	0.20	CDGHAV	0.58	CDGMIA	1.60
CDGTHR		CDGMPL				CDGVIE	1.76
				CDGMEX	0.62	CDGMXP	1.98
CDGBOG	0.06	CDGHEL	0.22			CDGFRA	2.24
CDGCAN		CDGOUA		CDGBRU	0.68	CDGBOS	2.56
CDGFNI						CDGAGP	2.60
CDGLRT		CDGAUH	0.24	CDGPHL	0.70	CDGAMS	2.74
CDGNDJ		CDGIST				CDGTUN	2.78
CDGPPT		CDGNRT		CDGDKR	0.72	CDGDXB	2.93
CDGSDQ				CDGTLS		CDGLAX	
CDGSSG		CDGBUD	0.26			CDGMUC	3.15
		CDGCOO		CDGLOS	0.82	CDGDUS	3.51
CDGCFE	0.08	CDGLAD				CDGYYZ	3.59
CDGEZE		CDGNAP		CDGLIN	0.84	CDGBCN	3.89
CDGLED						CDGIAD	3.99
CDGLFW		CDGCCS	0.28	CDGSTR	0.86	CDGMAD	4.01
CDGBKO	0.10	CDGPEK				CDGJFK	4.65
CDGKBP		CDGCKY	0.30	CDGABJ	0.88	CDGCMN	5.65
CDGOTP		CDGKWI	0.32			CDGLHR	6.73
CDGSCL		CDGOPO		CDGJED	0.94	CDGGVA	7.87
		CDGOSL		CDGRBA	0.96	CDGYUL	9.04
		CDGSVO	0.34	CDGHAM	0.98	CDGNCE	10.98
		CDGTRN		CDGMRS		CDGNCE	10.98

Sector	K-Means %	O-Cluster %
CDGLAX	3.23	2.93
CDGBCN	3.73	3.89
CDGLHR	4.72	6.73
CDGYUL	4.72	9.04
CDGGVA	5.80	7.87
CDGJFK	5.80	4.65
CDGNCE	10.36	10.98

**Table 6.12 –
Best Route Comparison
between
K-Means (Scenario 3)
and
O-Cluster Result**

6.4 Discussion of EM Clustering Results

In the previous part of our work, we have used the ODM tool for clustering. In the section 6.4 and 6.5; we will use another analytical tool called WEKA. The result of scenario is presented in a table. The same parameters applied in the evaluation of k-means clustering result are used.

6.4.1 Scenario Result

Table 6.13 shows the result of EM clustering scenario. The most profitable cluster is cluster 3. It groups about 28.45 percent of the mileage with only 7 percent of the passengers. This cluster has the highest weight fraction.

The business opportunity is shown in the cluster profile. This opportunity is identified by increasing the number of services used by passengers. But in contrast with the k-means clustering with ODM tool, where the cross-selling opportunities occurs between a mid-range cluster and high-profit cluster, the sales increase chances occurs between the member of same cluster categories.

The EM algorithm clustering with WEKA tool generates four categories of clusters; the high-profit customers (Cluster 1 – 3 – 4), the Mid-Range customers (Cluster 5 – 6 – 7 – 8), low-profit customers (Cluster 0), and new customers (Cluster 2). As mentioned before the cross-selling strategies are limited to the same category of clusters. Then the business opportunity in this order is not important, there will not generate a big enhancement in the customer situation. The result of k-means (ODM) gives more significant information and conclusion with less execution time.

Table 6.13 Clustering Analyst for EM Algorithm

Cluster ID	Mileage %	Customers %	Avg. Services per Cluster	Service Index	Weight	Membership (Sum Membership/ NB. Customer)
3	28.45	7	0.98	0.96	4.06	68.87
4	24.33	2	2.02	1.96	12.16	64.06
1	13.47	2	1.99	1.93	6.73	21.69
8	11.80	8	1.00	0.97	1.48	44.41
6	6.50	17	1.00	0.97	0.98	22.14
0	5.43	19	0.98	0.95	0.29	68.19
2	3.54	17	0.99	0.96	0.21	8.53
7	3.41	2	0.96	0.93	1.70	45.60
5	3.08	9	1.00	0.97	0.34	25.23

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

Best Customers: Clusters 1 - 3 - 4 (Higher Weight)
Strategies:
Retention strategy for best customers.
Cross-Sell strategy for clusters 8 by contrasting with cluster 7
(Service Index is close. By comparing which services is used by the best customers)
Cross-Sell strategy for clusters 5 by contrasting with cluster 6
(Service Index is close. By comparing which services is used by the best customers)
Cluster 2 to wait (New Customers)
Cluster 0 (The Worst Cluster)

6.5 Comparison of Clustering Algorithm

We have used for clustering four different algorithms; k-means, O-Cluster, EM, COBWEB. The first two belongs to ODM tool, and the last two belongs to WEKA tool. The ODM tool algorithms have an execution time much better than the algorithm of WEKA tool. All the methods are informative. But, it is data and cluster dependent, to determine which algorithm is the best to find the true clusters.

Several K-Means scenarios have been applied. All of them have a valuable result, dividing the 50,830 into marketing groups. The scenario 3 with 9 clusters covers all the commercial aspects without exaggeration and without using an over partitioning process. The first two scenarios give two different clustering categories; the best customers and the mid-range customers. The strategies to be adopted is to retain the best customers, and to apply cross-selling to mid-range clusters in order to ameliorate the customers sells and change their status in order to be in the best customer clusters. But two important clusters categories are absents; the new customers, and the bad customers. Instead, scenario 3 generates four categories of clusters; best customers, mid-range, worst customers, and new customers. All commercial aspects are available. Some possible strategies are the retention and up-selling for best customers; cross-selling of mid-range customers with best customers in order to change the status of those customers from mid-range to best customers' categories; minimize the marketing campaign addressed to the worst customers in order to reduce expensive; and observe the situation of the new customers and adopt some marketing effort to inform them about available products and services. The scenario 4 doesn't show the new customers categories and the three remaining categories have a lot of unnecessary details. The worst customers are divided into three clusters.

In another hand, we have build a K-Means model based on 80% of the data using the same parameters of K-Means scenario 3. We apply the model on the remaining 20% of the data. By comparing the result to the K-means Scenario 3 results, we found that they match 100%.

The K-Means has been fast and easily implemented but lots of experiments are needed and even then we can't be sure that we have found the best answer.

The O-Cluster results are mainly used to verify the result of K-means algorithm. Since the K-Means scenario 3 is considered as the best scenario covering all the commercial aspects, we will use the scenario 3 parameters in the O-Cluster algorithm. The O-Cluster generates three clusters categories; best customers, mid-range, and new customers. Same strategies used before for those three categories can be applied here. But the O-Cluster algorithm doesn't show the worst customers.

The clustering methodologies for enhanced k-means and O-Cluster algorithms are distance-based and grid-based respectively. Both algorithms use the hierarchical clustering method. They assign scoring data to clusters probabilistically. The O-Cluster has the advantage of less model build time.

The EM results generate four categories of customers; best customers, mid-range, bad customers, and new customers. The best customers' strategy is the retention and the enhancement of internal up-selling of this category. New customers have to be observed, and we have to minimize our marketing budget for bad customers. But the problem is shown in the mid-range category. The cross-selling strategies are limited to the same category of clusters. The business opportunities are limited, since all work that can be done, will not enhance the customer situation. We cannot move a customer from one category to another. The EM algorithm cluster boundaries are not so strict and results are quite consistent but predictions are not solid and vary over experiments.

The COMWEB algorithm has tree architecture. The results are translated into procedure. But the results depend on sample order. Due to that, we have problem in the execution. In addition, the experimenting process with the cutoff value to find acceptable results is mandatory.

6.6 Discussion of Association Rules Results

The association rules evaluation is based on the scenarios discussed before. We analyze the rules for each scenario. We have listed those rules into tables with two measurement parameters confidence and support. Those parameters give an analyst view on the rules and how it will be used in the business context. The support defines the percentage of the rule on the current database. The confidence gives the probability percentage to enhance work based on this rule. Our future plan has to be founded on the confidence. This plan can be a marketing campaign, or special offers. We launch in this section an analyst review of the three scenarios of the Association Rules algorithm application.

6.6.1 Scenario 1

The scenario 1 is based on the original activities of cluster 16 – best customer cluster. The original activities are “Flight”, “Financial”, and “Hotel”. We conclude from the result 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. The result has been confirmed. To enhance business, we have to divide customers into two different categories; Flight/Financial customers and Flight/Hotel customers. Two different marketing campaigns have to be launched. The first one dedicated to Flight/Financial customer, recommending “hotel” special offer. The second dedicated to Flight/Hotel customer, recommending “financial” special offer.

6.6.2 Scenarios 2

The scenario 2 is based on the activities of cluster 16 – best customer cluster. The activities are mainly the sectors flown by the customers. The sectors are restricted to the sectors originated from the main two hubs; CDG and BEY. The second association rules scenario treats the sectors with flown percentage over 10%. In this paragraph, we will expose the more interesting rules.

- $BEYDXB = 1$ and $BEYRUH=1 \rightarrow BEYCDG = 1$; with Support = 0.10 and Confidence = 0.84. Then 10% of the best customers are traveling to Beirut/Dubai, Beirut/Riyadh, and Beirut/Charles-De-Gaulle. We have also a probability of 84% to enhance our business for customers traveling on the sectors Beirut/Dubai, and Beirut/Riyadh. This can be done by marketing campaign or special offer for those customers for Beirut/Charles-De-Gaulle sector.
- $BEYDXB = 1$ and $BEYJED=1 \rightarrow BEYCDG = 1$; with Support = 0.11 and Confidence = 0.83. Then 11% of the best customers are traveling to Beirut/Dubai, Beirut/Jeddah, and Beirut/Charles-De-Gaulle. We have also a probability of 83% to enhance our business for customers traveling on the sectors Beirut/Dubai, and Beirut/Jeddah. This can be done by marketing campaign or special offer for those customers for Beirut/Charles-De-Gaulle sector.
- $BEYAMM = 1$ and $BEYCAI=1 \rightarrow BEYCDG = 1$; with Support = 0.12 and Confidence = 0.82. Then 12% of the best customers are traveling to Beirut/Amman, Beirut/Cairo, and Beirut/Charles-De-Gaulle. We have also a probability of 82% to enhance our business for customers traveling on the sectors Beirut/Amman, and Beirut/Cairo. This can be done by marketing campaign or special offer for those customers for Beirut/Charles-De-Gaulle sector.

6.6.3 Scenarios 3

The scenario 3 is founded on the activities of cluster 16 – best customer cluster. With the same characteristics mentioned in the scenario 2, the third association rules scenario

only difference is the treatment of the sectors with flown percentage over 20%. In this paragraph, we will explore some interesting rules.

- $BEYDXB = 1$ and $BEYJED=1 \rightarrow BEYCDG = 1$; with Support = 0.11 and Confidence = 0.83. $BEYAMM = 1$ and $BEYCAI=1 \rightarrow BEYCDG = 1$; with Support = 0.11 and Confidence = 0.82. We can see that at the scenario 3, we have the same rule as scenario 2. This is another way to verify that our result is similar and correct.
- $BEYCAI = 1$ and $BEYDXB=1 \rightarrow BEYCDG = 1$; with Support = 0.13 and Confidence = 0.81. Then 13% of the best customers are traveling to Beirut/Cairo, Beirut/Dubai, and Beirut/Charles-De-Gaulle. We have also a probability of 81% to enhance our business for customers traveling on the sectors Beirut/Dubai, and Beirut/Cairo. This can be done by marketing campaign or special offer for those customers for Beirut/Charles-De-Gaulle sector.
- The same logic can be used in the following rules: $BEYDXB = 1$ and $BEYLHR=1 \rightarrow BEYCDG = 1$; with Support = 0.13 and Confidence = 0.81. $BEYAMM = 1$ and $BEYDXB=1 \rightarrow BEYCDG = 1$; with Support = 0.16 and Confidence = 0.80.

6.7 Comparison of Association Rules Algorithms

We have used two different algorithms; the Apriori from ODM tool, and the Predictive Apriori from WEKA tool. Both algorithms generate valuable results from a marketing and commercial view. The difference between them is purely technical. In the Apriori algorithm, the analyst can control the result and define its needs. He can predefine the support, the confidentiality, and the number of attributes in each rule. Then the rules found are controlled by the analyst but they can include some insignificant results.

The Predictive Apriori algorithm is not controlled by the analyst. The only parameter used is the number of rules. It shows the best rules. The advantage of these rules is the possibility of retrieving some good rules not shown in Apriori algorithm due to the

restriction applied by the analyst. PredictiveApriori finds the best n rules maximizing the predictive accuracy, which combines confidence and support in one measure.

6.8 Summary of CRM Recommendations

Our objective is to managing the customer knowledge. The market analyst has now data on the frequent flyer customer's value. Information provides several customer groups. The best scenario for clustering using k-means algorithm is scenario 3. It generates 9 different clusters with specific profile for each one. These clusters allow MEA to generate revenue from customer's business. Such information is valuable in determining the resources MEA should commit to gain and retain a customer in the event he or she should defect. The cluster profile shows the business opportunity in increasing the number of services purchased by customers.

We track and monitor high-value customers or groups of customers. The result shows three clusters as best customers with the higher revenue mileage per customer. A retention strategy is applied for best customers. It is possible, for example, to recognize an individual in those clusters who usually travels once a month and has not revealed up for three months. The sales specialist can contact this customer to check the reason for his behavior change and try to rectify the situation in order to retain this valuable customer. Another result in these clusters is improving opportunity of identification. It provides opportunities for MEA to produce more revenue from a customer based on the information available on the customer. Based on the existing situation, MEA, for example, may try to apply up-selling strategy by selling a higher fare seat.

The second type of clusters defined in this study is the mid-range clusters. The analyst of the best customer behavior permits to propose an enhance strategy for those clusters in order to increase services usage and revenue mileage per passenger. This strategy defines services candidates for cross-selling. A cross-selling is applied to enlarge the number of services purchased. With little effort, we can convert customers from mid-range clusters to best customers' clusters.

The third type of clusters identified in this study is new customer cluster. The strategy is to observe those customers to determine their behavior. The marketing of available services to this group will be useful in order to make them profitable quickly.

The fourth type of clusters groups the bad customers with very low revenue mileage per passenger. The strategy is to retain any marketing campaign for those customers.

MEA has two main hubs; BEY and CDG. We have found the best route from CDG 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 or market it.

The association rules algorithm based on the best customer cluster illustrates more results. By analyzing the services used, we characterize the service integration. It enables MEA 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 approach of association rules algorithm is used to explore routes. It permits to propose customers additional route flight tailored to the needs, behavior, and values of MEA's most profitable customers.

Using this CRM strategic result, we can give MEA a competitive differentiation among airlines by making such customer-specific services available. It is a new way of doing business by providing the correct information and enables the appropriate service to each valuable customer at each touch point.

Chapter 7

CONCLUSION AND FURTHER WORK

7.1 Conclusion

The role of data mining is to insert intelligence back into the customer relationship. We conclude from this thesis:

- a. In our study, we have conducted a clustering and association rules for data mining purpose. They are applied to Frequent Flyer airline data. For clustering purpose, we have used different algorithms in order to evaluate our result. The algorithms used are K-means, O-cluster, EM, and COBWEB. In addition, different scenarios were applied in order to found the best result. One of the k-means scenario results has been used as input for association rules. The cluster grouping the best customers has been the base for the remaining part of the study. The study of association rules has been approach with two different views. The first one has analyzed different services used by the best customers. The other one has examined the sector flown by the best customers.

From the clustering, we have mainly two different results. We have categorized our customers. The customers have been grouped into different groups such as best customers, new customers, or not important customers. In addition, we have been able to compare clusters to each other. This comparison enables the up-selling and cross-selling procedure.

We can use the study result to optimize marketing campaigns. The targeting of the message is only part of a much larger process. This process includes determining the budget for the campaign, planning the offer, preparing marketing scripts or e-mail messages, and delivering the service to the customer. An important part of the CRM strategy is to understand and manage customer value. The goal of customer value

measurement should be to increase average customer value for all customers, rather than focus on only the most valued ones.

- b.** We have compared data mining algorithms and drew conclusions about the quality of the solutions produced. We have used for clustering four different algorithms; k-means, O-Cluster, EM, COBWEB. All the methods are informative. But, it is data and cluster dependent, to determine which algorithm is the best to find the true clusters.

Several K-Means scenarios have been applied. All of them have a valuable result, dividing the 50,830 into marketing groups. The best scenario is scenario 3 with 9 clusters covering all the commercial aspects without exaggeration and without using an over partitioning process. It generates four categories of clusters; best customers, mid-range, worst customers, and new customers. All commercial aspects are available. Some possible strategies are the retention and up-selling for best customers; cross-selling of mid-range customers with best customers in order to change the status of those customers from mid-range to best customers' categories; minimize the marketing campaign addressed to the worst customers in order to reduce expensive; and observe the situation of the new customers and adopt some marketing effort to inform them about available products and services.

In another hand, we have build a K-Means model based on 80% of the data using the same parameters of K-Means scenario 3. We apply the model on the remaining 20% of the data. By comparing the result to the K-means Scenario 3 results, we found that they match 100%. The K-Means has been fast and easily implemented but lots of experiments are needed and even then we can't be sure that we have found the best answer.

The O-Cluster results are mainly used to verify the result of K-means algorithm. It generates three clusters categories; best customers, mid-range, and new customers. But the O-Cluster algorithm doesn't show the worst customers.

The clustering methodologies for enhanced k-means and O-Cluster algorithms are distance-based and grid-based respectively. Both algorithms use the

hierarchical clustering method. They assign scoring data to clusters probabilistically. The O-Cluster has the advantage of less model build time.

The EM results generate four categories of customers; best customers, mid-range, bad customers, and new customers. Same strategies discussed before can be applied. But the problem is shown in the mid-range category. The cross-selling strategies are limited, since all work that can be done, will not enhance the customer situation.

The COMWEB algorithm has tree architecture. But the results depend on sample order. Due to that, we have problem in the execution.

- c. In this study, we have defined our most valuable customers with the activities that contribute to their value. The behavior of a good customer has been discovered. The behavior includes the attributes and characteristics. We have determined the customers that are most promising for a defined campaign. We have defined marketing rules to transform unprofitable or low profit customers to a position of improved profitability. The predicted lifetime value by customer segment has been identified. We have defined the best market segment. We have identified the customer segment that has a potential to purchase additional travel segment by Identifying up-selling and cross-selling opportunities in order to design packages or grouping of services. Our main achievement is to match new customers to the right services.

7.2 Future Work

By using the frequent flyer data source only part of the current customers are considered, since there are also many customers of an airline, which are non-members of the frequent flyer program. Our future plan is to have the ability to view and analyze the Passenger Name Record (PNR). The PNR contains details of a reservation/booking for a passenger which can be divided into categories. We mention here below some of this categories:

- Flight Bookings Information, including when the booking was made, the status of the bookings, revenue versus non-revenue, seat requests, special meal requests, identifying and capturing all changes to the bookings.

- Sales Information, including channel, travel agency, campaign, promotions, and sales commission.
- Ticket Information, including ticket number, fare basis, and taxes.
- Passenger Information, including name, salutation, age, occupation, loyalty card and contact details.
- Itinerary Information, including flight segments, legs, O&D (Origin & Destination), connection point, and journey.
- Non-Flight Bookings Information, including hotel, car rental, and tour details.
- Miscellaneous Information, including person who makes the booking, and organization name.

The business benefits will be to enhance the analysis based on journey and booking information; to capture and understand customer behavior with the integration of seat inventory, flight schedules, departure control, and tickets; to improve operations such as overbooking profile, airport operations, cargo, investigations and security; and to assure revenue.

By analyzing a journey, an airline can optimize its complete network. PNR information helps to determine alliance, code share, or special prorate arrangements (SPA). The data can also help differentiate between leisure and business travel that will be reflected on the segmentation process. It can be used also to measure the effectiveness of marketing and sales programs. In addition, segmenting the passengers on a flight based on their PNR data can influence overbooking practices.

REFERENCES

- 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
- Ali, A.; Bagherjeiran, A.; & Chen, C. (2004). Scalable Clustering Algorithms.
<http://www2.cs.uh.edu/~ayaz/Scalable%20Clustering.pdf>
- Boland, D.; Morrison, D.; & O'Neill, S. (2002). The Future of CRM in the Airline Industry: A New Paradigm for Customer Management. IBM Institute for Business Value.
http://www-5.ibm.com/e-business/fi/pdf/highlights/integration/crm_airline.pdf
- Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. NCR Systems Engineering Copenhagen (USA & Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) & OHRA Verzekeringen en Bank Groep B.V (The Netherlands).
- Deck, S. (1999). Data Mining. ComputerWorld.
<http://www.computerworld.com/hardwaretopics/hardware/de.../0.10801,43509.00.htm>
- Dellaert, F. (2002). The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology. Technical Report number GIT-GVU-02-20.
<http://www-static.cc.gatech.edu/~dellaert/em-paper.pdf>
- Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Prentice Hall.
- Etzioni, O.; Knoblock, C.; Tuchinda, R.; & Yales, A. (2003). To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price. ACM.
<http://www.isi.edu/integration/papers/etzioni03-kdd.pdf>
- 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

- Fisher, D. (1987). Knowledge Acquisition Via Incremental Conceptual Clustering.
<http://kiew.cs.uni-dortmund.de:8001/mlnet/instances/81d91eaa-db6c2e1493>
- Lee, D. (1999). CRM Definitions. CRM.Talk #054.
<http://www.crmguru.com/content/crmtalk/2000a/crmt054.htm#1>
- Linoff, G. (2004). Survival Data Mining for Customer Insight.
<http://www.intelligententerprise.com/showArticle.jhtml?articleID=26100528>
- Moss, L.; Adelman, S.; & Abai M. (2005). *Data Strategy*. Addison Wesley Professional.
- 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>
- Ramachandran, P. (2001). Mining for Gold. WIPRO Technologies.
<http://www.wipro.com/whitepapers/services/businessintelligence/dataminingmininggold.htm>

Appendices

Airport Code

Code	Airport Name	Airport Location	State	Country
ABJ	Abidjan	Abidjan		Cote D'Ivoire
AGP	Malaga	Malaga		Spain
AMM	Amman	Amman		Jordan
AMS	Amsterdam	Amsterdam		Netherlands
ARN	Arlanda	Stockholm		Sweden
ATH	Athens	Athens		Greece
ATL	Atlanta	Atlanta	Georgia	USA
AUH	Abi Dhabi	Abi Dhabi		United Arab Emirates
BCN	Barcelona	Barcelona		Spain
BES	Brest	Brest		France
BEY	Beirut	Beirut		Lebanon
BGF	Bangui	Bangui		Central African Republic
BHX	Birmingham	Birmingham		United Kingdom
BIQ	Biarritz	Biarritz		France
BKK	Bangkok	Bangkok		Thailand
BKO	Bamako	Bamako		Mali
BLQ	Bologna	Bologna		Italy
BOD	Bordeaux	Bordeaux		France
BOG	Bogota	Bogota		Colombia
BOM	Bombay	Bombay		India
BOS	Boston	Boston	Massachusetts	USA
BRU	Brussels	Brussels		Belgium
BSL	Basle	Basle		Switzerland
BUD	Budapest	Budapest		Hungary
BZV	Brazzaville	Brazzaville		Congo
CAI	Cairo	Cairo		Egypt
CAN	Guangzhou	Guangzhou		China
CCS	Caracas	Caracas		Venezuela
CDG	Aeorporte Charles De Gaulle	Paris		France
CGN	Cologne	Cologne		Germany
CKY	Conakry	Conakry		Guinea
CMN	Mohamed V	Casablanca		Morocco
COO	Cotonou	Cotonou		Benin
CPH	Copenhagen	Copenhagen		Denmark
CVG	Cincinnati	Cincinnati	Ohio	USA
DAM	Damascus	Damascus		Syrian Arab Republic

Code	Airport Name	Airport Location	State	Country
DFW	Dallas/Ft. Worth	Dallas/Ft. Worth	Texas	USA
DKR	Dakar	Dakar		Senegal
DLA	Douala	Douala		Cameroun
DOH	Doha	Doha		Qatar
DTW	Detroit	Detroit	Michigan	USA
DUS	Dusseldorf	Dusseldorf		Germany
DXB	Dubai	Dubai		United Arab Emirates
EWR	Newark	New York	New York	USA
EZE	Ministro Pistarini	Buenos Aires		Argentina
FCO	Fiumicino	Rome		Italy
FDF	Fort De France	Fort De France		Martinique
FIH	Kinshasa	Kinshasa		Zaire
FLR	Florence	Florence		Italy
FMO	Muenster	Munster		Germany
FNI	Nimes	Nimes		France
FRA	Frankfurt	Frankfurt		Germany
GIG	Internacional	Rio De Janeiro		Brazil
GOT	Gothenburg	Gothenburg		Sweden
GRU	Guarulhos International	Sao Paulo		Brazil
GVA	Geneva	Geneva		Switzerland
HAJ	Hanover	Hanover		Germany
HAM	Hamburg	Hamburg		Germany
HAV	Havana	Havana		Cuba
HEL	Helsinki	Helsinki		Finland
HKG	Hong Kong	Hong Kong		
IAD	Dulles International	Washington	District of Colombia	USA
IAH	Intercontinental	Houston	Texas	USA
IST	Istanbul	Istanbul		Turkey
JED	Jeddah	Jeddah		Saudi Arabia
JFK	John F. Kennedy Int'l Airport	New York	New York	USA
JNB	Johannesburg	Johannesburg		South Africa
KBP	Borispol	Kiev		Ukraine
KWI	Kuwait	Kuwait		Kuwait
LAD	Luanda	Luanda		Angola
LAX	Los Angeles	Los Angeles	California	USA
LBV	Libreville	Libreville		Gabon
LCR	Larnaca	Larnaca		Cyprus
LCY	London City Airport	London		United Kingdom
LED	St-Petersburg	St-Petersburg		CIS
LFW	Lome	Lome		Togo

Code	Airport Name	Airport Location	State	Country
LHR	Heathrow Airport	London		United Kingdom
LIN	Linate	Milan		Italy
LIS	Lisbon	Lisbon		Portugal
LJU	Ljubljana	Ljubljana		Yugoslavia
LOS	Lagos	Lagos		Nigeria
LYS	Lyon	Lyon		France
MAD	Madrid	Madrid		Spain
MAN	Manchester	Manchester		United Kingdom
MEX	Mexico	Mexico		Mexico
MIA	Miami	Miami	Florida	USA
MLH	Mulhouse	Mulhouse		France
MNL	Manila	Manila		Philippines
MPL	Montpellier	Montpellier		France
MRS	Marseille	Marseille		France
MRU	Mauritius	Mauritius		Mauritius
MUC	Munich	Munich		Germany
MPX	Malpensa	Milan		Italy
NAP	Naples	Naples		Italy
NCE	Nice	Nice		France
NCL	Newcastle	Newcastle		United Kingdom
NDJ	N'Djamena	N'Djamena		Chad
NIM	Niamey	Niamey		Niger
NKC	Nouakchott	Nouakchott		Mauritania
NRT	Narita	Tokyo		Japan
NTE	Nantes	Nantes		France
NUE	Nuremberg	Nuremberg		Germany
OPO	Porto	Porto		Portugal
ORD	O'Hare International Airport	Chicago	Illinois	USA
OSL	Oslo	Oslo		Norway
OTP	Otepeni	Bucharest		Romania
OUA	Ouagadougou	Ouagadougou		Burkina Faso
PEK	Beijing	Beijing		China
PHC	Port Harcourt	Port Harcourt		Nigeria
PHL	Philadelphia	Philadelphia	Pennsylvania	USA
PPT	Papeete	Papeete		French Polynesia
PRG	Prague	Prague		Czechoslovakia
PSA	Pisa	Pisa		Italy
PTP	Point A Pitre	Point A Pitre		Guadeloupe
PUF	Pau	Pau		France
RAK	Marrakech	Marrakech		Morocco

Code	Airport Name	Airport Location	State	Country
RBA	Rabat	Rabat		Morocco
RUH	Riyadh	Riyadh		Saudi Arabia
SCL	Santiago	Santiago		Chile
SDQ	Santo Domingo	Santo Domingo		Dominican Republic
SFO	San Francisco	San Francisco	California	USA
SHA	Shanghai	Shanghai		China
SIN	Singapore	Singapore		Singapore
SOF	Sofia	Sofia		Bulgaria
STR	Stuttgart	Stuttgart		Germany
SVO	Sheremetyevo	Moscow	Moscow	CIS
SXB	Strasbourg	Strasbourg		France
SXM	St. Marten	St. Marten		Netherlands Antilles
THR	Tehran	Tehran		Iran
TLS	Toulouse	Toulouse		France
TRN	Turin	Turin		Italy
TUN	Tunis	Tunis		Tunisia
TXL	Tegel	Berlin		Germany
VCE	Venice	Venice		Italy
VIE	Vienna	Vienna		Austria
VLC	Valencia	Valencia		Spain
VRN	Verona	Verona		France
WAW	Warsaw	Warsaw		Poland
YUL	Dorval International	Montreal	Quebec	Canada
YYZ	Toronto	Ontario	Ontario	Canada
ZRH	Zurich	Zurich		Switzerland