E-BANKING: HOW DATA LEAD TO ACTION

Vasilis Aggelis
aggelisv@winbank.gr

WINBANK, PIRAEUSBANK SA,
Syggrou Ave., 87,
Athens 11745, Greece

Abstract: The role of data in a modern organization is most important in our days. Apparently data are not only for statistical purposes and reporting. Nowadays data management is the motive force for actions. In this sense, information and knowledge exercise significant influence in establishment and development of a company. A relatively new approach is data mining. In Greek banking sector, a remarkable number of organizations have adopted data mining methods. The introduction of such methods in the banking area due to the nature and sensitivity of bank data can already be considered of great assistance to banks as to prediction, forecasting, decision support and action. Object of this paper is to demonstrate that keeping track of customer groups according to their relations with e-banking services and products, is of major importance firstly for all customers' interest and secondly for e-Banking Services improvement and redesign.

1. INTRODUCTION. In previous works we introduced e-Trans association rules for Greek internet banking, concerning patterns in e-banking transactions through internet, Aggelis (1). Those rules were established taking in account the Greek e-banking market.

Determination of association rules concerning internet banking data is a demanding task since:

- The volume of data is large.
- The goal must be clearly set from the beginning.
- Good knowledge and great experience of the data is a prerequisite not only for the data miner but also for the final analyst.
- Not all rules are of interest.

The challenge of the whole process rises from the fact that some of established relations are not easily observed without use of data mining methods.

The increase of electronic transactions during the last two years is quite rapid in Greek banking sector, Aggelis (2), Aggelis (3), Aggelis and Christodoulakis (4). E-banking nowadays offers a complete set of products and services facilitating not only the individual customers but mainly corporate customers to conduct their transactions easily and securely. Discovery of rules between different electronic services is therefore of great significance to a bank.

Identification of such rules offers advantages as the following:

- Good knowledge of the relationships between different types of electronic transactions.
- Description and establishment of most popular internet transactions
- The electronic services become more easily familiar to the public since specific groups of customers are approached, that uses specific payment manners.
- Customer approach is well designed with higher possibility of successful engagement.
- The improvement of already offered bank services is classified as to those used more frequently.
- Redesign internet transaction structures for those which used rarely
- Reconsidering of the usefulness of products exhibiting little or no contribution to the rules.

The most important established e-Trans rules (1) are:

If someone pays Greek Telecommunication Organization bill *then* also pays Public Power Corporation bill(confidence=67%). (Rule 1)

Additionally it can be observed that this rule is valid for the reverse case too, namely:
If someone pays Public Power Corporation bill *then* also pays Greek Telecommunication Organization bill (confidence= 62%). (Rule 2)

Two other rules obtained exhibiting confidence > 45% are the following:
If someone sends Social Insurance Institute Payment Order *then* also sends VAT Payment Order (confidence=64%), (Rule 3)
valid also reversely.
If someone sends VAT Payment Order *then* also sends Social Insurance Institute Payment Order (confidence=46%), (Rule 4)

The next step in our research is the identification of customers' attributes who satisfy the most important e-Trans rules. In order to achieve the above scope, we apply clustering algorithms for these customers who satisfy Rule 1. The software used is SPSS Clementine 9.0 Desktop, SPSS (13), SPSS (14). Section 2 contains clustering basics and algorithms, while in section 3 the process of application for clustering is described. Experimental results can be found in section 4 and finally section 5 contains the main conclusions of this work accompanied with actions taken from bank.

2 CLUSTERING BASICS

2.1 Clustering Techniques. Clustering techniques, Collier et al (7), Hand et al (10), fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. There is no target variable to be predicted, thus no distinction is being made between independent and dependent variables.
Clustering techniques are used for combining observed examples into clusters (groups), Chen et al (6), COMPAQ (8), DataPlus (9), Hilderman and Hamilton (11), Im and Park (12), Toivonen (15), that satisfy two main criteria:

1. each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
2. each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

Depending on the clustering technique, clusters can be expressed in different ways:

- identified clusters may be exclusive, so that any example belongs to only one cluster.
- they may be overlapping; an example may belong to several clusters.
- they may be probabilistic, whereby an example belongs to each cluster with a certain probability.
- clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

2.2 K-means algorithm. K-means, Bradley and Fayyad (5), (7), (10), Zha et al (16), is the simplest clustering algorithm. This algorithm uses as input a predefined number of clusters that is the *k* from its name. Mean stands for an average, an average location of all the members of a particular cluster. When dealing with clustering techniques, a notion of a high dimensional space must be adopted, or space in which orthogonal dimensions are all attributes from the table of analysed data. The value of each attribute of an example represents a distance of the example from the origin along the attribute axes. Of course, in order to use this geometry efficiently, the values in the data set must all be numeric and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space.

K-means algorithm is a simple, iterative procedure, in which a crucial concept is the one of centroid. Centroid is an artificial point in the space of records that represents an average location of the particular cluster. The coordinates of this point are averages of attribute values of all examples that belong to the cluster. The steps of the K-means algorithm are given below.

Select randomly k points (it can be also examples) to be the seeds for the centroids of k clusters.
Assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples.
Calculate new centroids of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (centroid).
Check if the cluster centroids have changed their "coordinates".
If yes, start again form the step 2).
If not, cluster detection is finished and all examples have their cluster memberships defined.

Usually this iterative procedure of redefining centroids and reassigning the examples to clusters needs only a few iterations to converge.

3 DATA AND PREPARATION. From customers who satisfy e-Trans Rule 1, we select only the individuals. For each customer we take into consideration the following attributes:

- Gender
- Age
- Home area
- Enrollment years

A data sample is shown in Table 1.

Table 1. Data Sample

Cust -id	*Gender*	*Home-area*	*Age*	*Enrollment Years*
…	…	…	…	…
4303970	M	Attica A	25	4
4270227	M	Salonica B	26	0
4476694	M	Peloponissos	26	3
…	…	…	…	…

Home Areas are in total 13 distinct geographical areas of Greece. Values of Enrollment Years are rounded. Value 0 (zero) indicates that customer has less than 6 months enrolled in e-banking services.

4 EXPERIMENTAL RESULTS. After K-Means algorithm application, the following age histogram produced (Fig. 1). An easily assumption is that distribution is high between 30 and 40 years of age. This fact is absolute normal for Greek banking market.

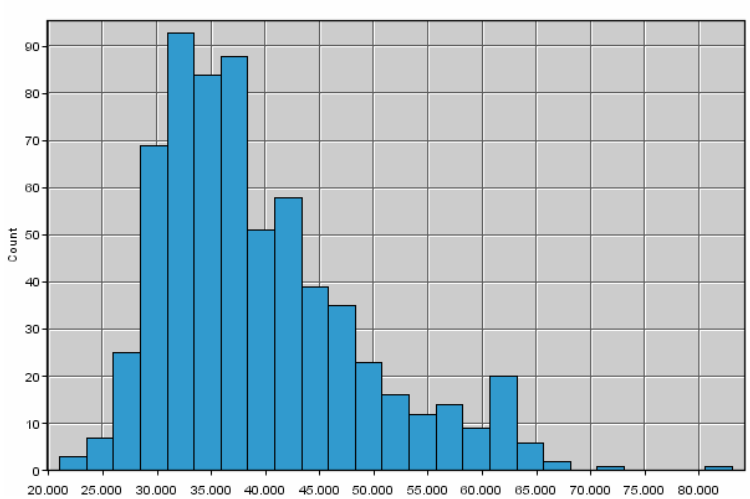


Fig. 1 – Age Histogram

However, the basic result is the five distinct customer clusters, which are shown in Fig. 2. Those customers pays Greek Telecommunication Organization bill and also pays Public Power Corporation bill, with great confidence (67%).

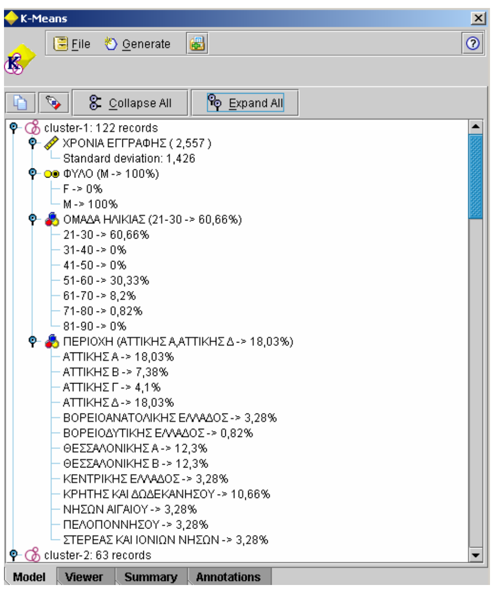


Fig.2 – Customer Clusters

Analytically, the five clusters are:

Cluster 1 (122 customers)

Enrollment Years = 2.557

Gender = M (Male)

Age Group

21-30 → 60.66%, 51-60 → 30.33%, 61-70 → 8.2%, 71-80 → 0.82%

Home Area

Attica A → 18.03%, Attica B → 7.38%, Attica C → 4.1%, Attica D → 18.03%, North-East Greece → 3.28%, North-West Greece → 0.82%, Salonica A → 12.3%, Salonica B → 12.3%, Central Greece → 3.28%, Crete and Dodekanissa → 10.66%, Aegean Islands → 3.28%, Peloponissos → 3.28%, Sterea and Ionion Islands → 3.28%

Cluster 2 (63 customers)

Enrollment Years = 2.619

Gender = F (Female)
Age Group
31-40 → 100%
Home Area
Attica A → 28.57%, Attica B → 7.94%, Attica C → 6.35%, Attica D → 11.11%, North-West Greece → 1.59%, Salonica A → 15.87%, Salonica B → 12.7%, Central Greece → 1.59%, Crete and Dodekanissa → 9.52%, Aegean Islands → 3.17%, Sterea and Ionion Islands → 1.59%

__**Cluster 3 (73 customers)**__
Enrollment Years = 2.315
Gender = F (Female)
Age Group
21-30 → 41.1%, 41-50 → 35.62%, 51-60 → 19.18%, 61-70 → 2.74%, 81-90 → 1.37%
Home Area
Attica A → 24.66%, Attica B → 13.7%, Attica C → 12.33%, Attica D → 12.33%, North-East Greece → 1.37%, North-West Greece → 1.37%, Salonica A → 9.59%, Salonica B → 9.59%, Crete and Dodekanissa → 10.96%, Aegean Islands → 1.37% Peloponissos → 2.74%

__**Cluster 4 (145 customers)**__
Enrollment Years = 3.207
Gender = M (Male)
Age Group
41-50 → 88.97%, 61-70 → 11.03%
Home Area
Attica A → 11.03%, Attica B → 5.52%, Attica C → 6.21%, Attica D → 16.55%, North-East Greece → 2.76%, North-West Greece → 1.38%, Salonica A → 22.07%, Salonica B → 7.59%, Central Greece → 2.07%, Crete and Dodekanissa → 12.41%, Aegean Islands → 3.45%, Peloponissos → 4.83%, Sterea and Ionion Islands → 4.14%

__**Cluster 5 (253 customers)**__
Enrollment Years = 2.625
Gender = M (Male)
Age Group
31-40 → 100%
Home Area
Attica A → 13.83%, Attica B → 7.11%, Attica C → 6.72%, Attica D → 19.76%, North-East Greece → 5.14%, North-West Greece → 1.98%, Salonica A → 14.23%, Salonica B → 8.3%, Central Greece → 1.58%, Crete and Dodekanissa → 9.09%, Aegean Islands → 3.16%, Peloponissos → 4.74%, Sterea and Ionion Islands → 4.35%

Each attribute which is involved in clustering has an importance value. Those attributes with value near 1 are the most important. Fig. 3 shows that group of age, gender and enrollment years are the most important factors for customer clustering. Importance of home area is smaller and this is normal because of the high distribution of the 13 distinct values. If we have bounded areas in less distinct values, the importance would be higher.

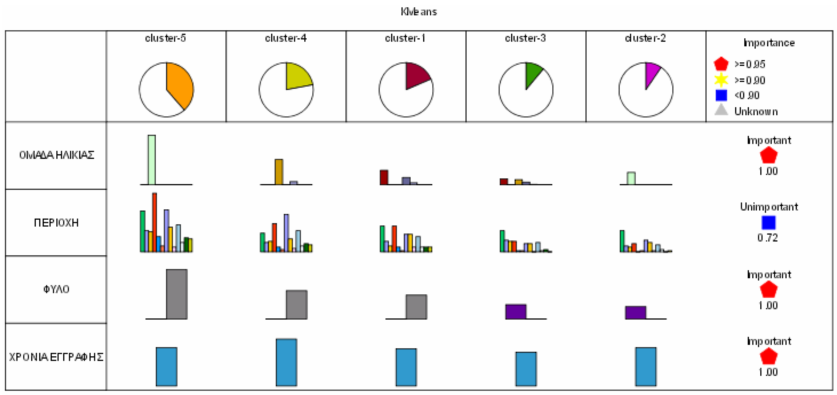


Fig. 3 – Importance of attributes

In Fig. 4 we see the customer distribution in each cluster, taking into consideration gender. Each group has one distinct value.

In Table 2 we see the same distribution. The difference is that values are age average. In Table 3 cell values are enrollment years' average.

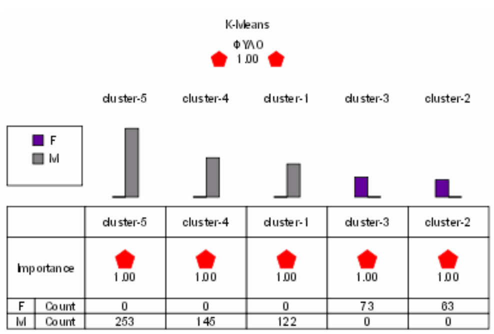


Fig. 4 – Customer distribution according gender

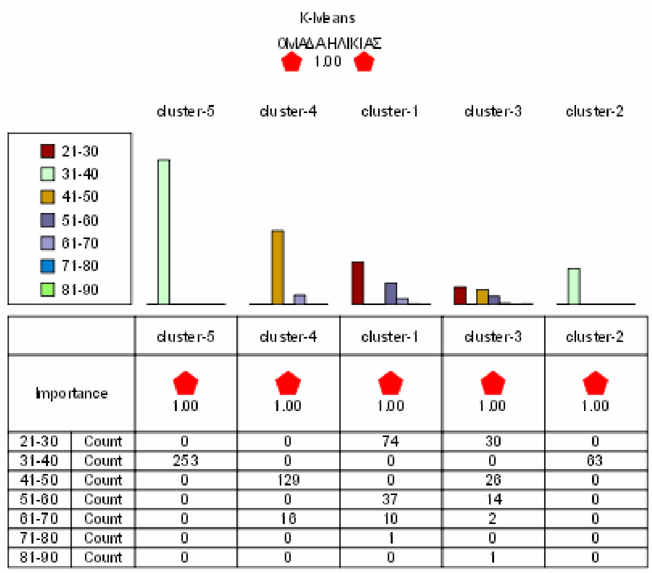
Table 2 – Customer distribution according gender (with age averages)

Gender	*Cluster5*	*Cluster4*	*Cluster1*	*Cluster3*	*Cluster4*
F	0	0	0	41.5	34.28
M	35.42	46.62	39.59	0	0

Table 3 – Customer distribution according gender (with enrollment years averages)

Gender	*Cluster5*	*Cluster4*	*Cluster1*	*Cluster3*	*Cluster4*
F	0	0	0	2.3	2.619
M	2.618	3.217	2.557	0	0

In Fig. 5 we see the customer distribution in each cluster, taking into consideration group of age. In Table 4 we see the same distribution. The difference is that values are enrollment years average. In Table 5 cell values are age average.



		cluster-5	cluster-4	cluster-1	cluster-3	cluster-2
Importance		1.00	1.00	1.00	1.00	1.00
21-30	Count	0	0	74	30	0
31-40	Count	253	0	0	0	63
41-50	Count	0	129	0	28	0
51-60	Count	0	0	37	14	0
61-70	Count	0	16	10	2	0
71-80	Count	0	0	1	0	0
81-90	Count	0	0	0	1	0

Fig. 5 – Customer distribution according group of age

Table 4 – Customer distribution according group of age (with enrollment years averages)

Age Group	*cluster1*	*cluster2*	*cluster3*	*cluster4*	*cluster5*
21-30	2.43	0	2.20	0	0
31-40	0	2.62	0	0	2.62
41-50	0	0	2.38	2.98	0
51-60	2.84	0	2.46	0	0
61-70	2.40	0	1,00	5.00	0
71-80	3.00	0	0	0	0
81-90	0	0	3.00	0	0

Table 5 – Customer distribution according group of age (with age averages)

Age Group	*Cluster1*	*Cluster2*	*Cluster3*	*Cluster4*	*Cluster5*
21-30	28.33	0	28.43	0	0
31-40	0	34.30	0	0	35.43
41-50	0	0	45.26	44.72	0
51-60	54.81	0	56.14	0	0
61-70	63.50	0	65.50	62.00	0
71-80	71.00	0	0	0	0
81-90	0	0	83.00	0	0

5 CONCLUSIONS AND ACTIONS. Basic conclusions from this work summarized below.

The 5 distinct groups of customers who pays Greek Telecommunication Organization (GTO) and Power Public Corporation (PPC) bills are:

1. Males between 31-40 years old, living in big cities and having 2.625 enrollment years.
2. Males between 41-50 years old, living in big cities and having 3.207 enrollment years
3. Males between 21-30 years old, living in big cities and having 2.5577 enrollment years
4. Females between 21-30 years old, living in big cities and having 2.315 enrollment years
5. Females between 31-40 years old, living in big cities and having 2.619 enrollment years

A valuable remark is that the enrolment years average is more than 2 years. This fact indicates that "old" users are more familiar with electronic payments.

The majority of customers live in big cities. Especially for women this percentage is higher. This result confirms the small internet penetration in Greek countryside.

Also the majority of customers are young people, between 21-40 years old. Those customers are more familiar with internet, e-banking and technology.

The following actions are under consideration from e-banking division, in order to ensure customer loyalty in a first phase.

- **Consolidation of the two payments types (GTO, PPC) in one screen**. The main advantage for customer would be less time spending in those payments. Hidden advantage for both sides would be that customers have the sense of care from their bank.
- **On line Cross selling campaign**. When a user ends such transaction, a message would appear, informing him, about the other payment and vice versa.
- **On line knowledge returning to the customer**. After ending such transaction, a customer would have the ability to see on line the most popular electronic payments for those who conduct GTO and PPC payments.

In a second phase, e-banking division could approach bank's customers who are making these payments from other channels (cashier, ATM, call center) in order to increase its customer portfolio firstly and secondly to contribute bank's cost saving.

6. REFERENCES

1. Aggelis V., "Application and Diachronic Strength of e-Trans Association Rules", 5th IBIMA Conference 2005, Cairo, Egypt.
2. Aggelis V., "Data Mining for Decision Support in e-banking area (II)", 1st International Conference on Knowledge Engineering and Decision Support 2004, Porto, Portugal.
3. Aggelis V., "e-Commerce Models for Banks' Profitability", Transactions on Information and Communication Technologies Journal, ISSN 1743-3517, 2005.
4. Aggelis V., and Christodoulakis D., "Customer Clustering using RFM analysis" , 9th WSEAS International Conference on Computers (Special Session Data Mining: Techniques and Application) 2005, Vouliagmeni, Greece.
5. Bradley P., and Fayyad U., "Refining Initial Points for K-Means Clustering", 15th International Conference on Machine Learning, 1998.
6. Chen M., Han J. and Yu P., "Data Mining: An Overview from Database Perspective", IEEE Transactions On Knowledge And Data Engineering, 1997.
7. Collier, K., Carey B., Grusy E., Marjaniemi C., and Sautter D., "A Perspective on Data Mining", Northern Arizona University, 1998
8. COMPAQ "Retain Customers and reduce risk", White Paper, 2001.
9. DataPlus Millenium, "Data-Driven Analysis Tools and Techniques", White Paper, 2000
10. Hand D., Mannila H., and Smyth P., "Principles of Data Mining". The MIT Press, 2001
11. Hilderman R., and Hamilton H., "Knowledge Discovery and Interestingness Measures: A Survey", Technical Report CS 99-04, Department of Computer Science, University of Regina ,1999
12. Im K., and Park S., "A Study on Analyzing Characteristics of Target Customers from Refined Sales Data", APIEMS, 1999.
13. SPSS, "Clementine 9.0 Users's Guide", Integral solutions Limited, 2005.
14. SPSS, "Clementine Application Template for Customer Relationship Management 6.5", Integral solutions Limited, 2005.
15. Toivonen H., "Discovery of Frequent Patterns in Large Data Collections", Technical Report A-1996-5, Department of Computer Science, University of Helsinki, 1996.
16. Zha, H., Ding, C., Gu, M., He, X. and Simon, H., "Spectral Relaxation for K-means Clustering", Neural Info. Processing Systems, 2001.