Applying BI Techniques To Improve Decision Making And Provide Knowledge Based Management

Alexandra Maria Ioana FLOREA
Bucharest University of Economic Studies
alexandra.florea@ie.ase.ro

The paper focuses on BI techniques and especially data mining algorithms that can support and improve the decision making process, with applications within the financial sector. We consider the data mining techniques to be more efficient and thus we applied several techniques, supervised and unsupervised learning algorithms The case study in which these algorithms have been implemented regards the activity of a banking institution, with focus on the management of lending activities.

Keywords: *business intelligence, data mining, Naïve Bayse, Support Vector Machine,*

1 Introduction

Business Intelligence refers to information systems for identifying, extracting and analyzing data available in enterprise systems whose purpose is to provide real support for business decisions [1].

[2] identifies a various number of BI techniques such as: predictive modelling through which we can predict value for a specific data item attribute; characterization and descriptive data mining used for data distribution, dispersion and exception; classification, used to determine to which class a data item belongs; clustering and outlier analysis which partitions a set into classes, whereby items with similar characteristics are grouped together; OLAP (OnLine Analytical Processing) with tools that enable users to analyze different dimensions of multidimensional data (for example, it provides time series and trend analysis views). Other techniques we can mention are Association, correlation, causality analysis (Link Analysis), Temporal and sequential patterns analysis, Model Visualization, Exploratory Data Analysis (EDA).

Machine learning can offer a set of tools that could be useful in summarizing various types of unliniar connections between data. [3]

A new concept that is quickly making its way in the knowledge management efforts is the use of Big Data. As mentioned in [4] Big Data found its way quickly in online shopping. For example we can identify the behavior of each customer, even by correlating his logins with IP addresses for tracking views when he is not authenticated. With the help of such analyses we can identify the products or the class of products that even though are being viewed and/or added to the cart they aren't eventually bought as much as other products.

There are almost unlimited options for using BI techniques to support the decision process in any type of business, from energy power plants to financial institutions.

In our research, the knowledge discovered by applying data mining techniques will enable representatives of a financial institution to asses the lending activity of the institution. The solution we are presenting in this paper represents a case study whose aim is to develop an integrated solution which automates the sales processes of a banking institution, with focus on the management of lending activities.

2. Algorithms used for determining the likelihood of contracting the loan

In order to implement the functionalities regarding the decisions to grant or reject loan applications it is necessary to develope performant algorithms to determine the values of the pre-scoring and scoring processes, to estimate the probability of granting the loan and to determine the maximum threshold for the amount awarded.

The objectives are:

- creating a scoring model, which contains characteristics that are relevant and have impact on lending decision as well as financial information, the relationship with the bank, areas of interest;
- creating a predictive model which is based on data mining techniques such as logistic regression, the Naïve Bayes classifier and the Support Vector Machines algorithm;
- data analysis through clustering methods in order to obtain a profile of the customer.

The algorithms will take into account the financial indicators and socio-demographic data that characterize a customer and include the following:

Table 1. Financial and socio-demographic indicators related to customers

Attribute name	Variable type	Explanation
CNP	Varchar2	Identity number used to identify customers
ID_CERERE	Varchar2	Loan application ID
ID_PRODUS	Varchar2	Loan product ID
NUME_CLIENT	Varchar2	Client's name and surname
PROFESIA	Varchar2	Client's profession
SEX	Varchar2	Genul persoanei (feminin/masculin)
MONEDA	Varchar2	Currency in which the loan is requested. (RON, EUR, USD)
GARANTIEVALOARE	Number	Total value of collaterals offered by the client.
VALOARE_ESTIMATA_BUNURI	Number	Total value of assests owned by the client
VENIT_ANUAL_RON	Number	Total annual income
SUMA_INDATORARE_RON	Number	Total amount of debt
DATA_CERERE	Date	Date of the lending application.
VARSTA	Number	Client's age
CATEGORIE	Varchar2	Type of credit.
DESCRIERE	Varchar2	Description of the requested loan.
PRESCORING/SCORING	Number	Value of calculated scor after applying the scoring algorithm (maximum 100)
SUMA_DEPOZIT	Number	Total amount of deposits owned by the client.
FIDELITATE	Number	Customer history relationship with the bank (0 - client without historical relationship with the bank, 1 - client with other products such as current account and / or other loans 2 - client with products including deposit)
STARE_CIVILA	Varchar2	Customer status (unmarried , married , divorced, widow)
SUMA_SOLICITATA	Number	The loan amount requested by the client

The algorithm is based on data mining techniques with supervised learning mechanism (classification algorithms using Bayes classifier, regression algorithms, significant attributes identification) and unsupervised learning mechanism (clustering). The following algorithms have been used

- *Naïve Bayes classification algorithm* is a supervised learning technique which enables the relationship between each independent variable and the dependent variable, by calculating a conditional probability for each of these relationships [5]. Thus the prediction is achieved by determining the effects of independent variables on the dependent variable

- The *SVM method* allows binary classification and is based on the structural risk minimization principle. The algorithm involves the following steps [6] separating classes using linear programming to obtain linear and nonlinear patterns of discrimination between data points; determining overlapping classes; application of kernel techniques to eliminate nonlinearity; determining the optimal solution.
- *Regression* is a method for determining the relationship between a dependent variable Y (response type variable) and one or more independent variables X1, ..., Xn (predictors or explanatory variables). Regression allows the determination of Y variation when the independent variables Xi values change. A value or range of values of the dependent variable for certain values of the independent variables can be estimated
- Determining significant attributes (Attribute Importance) is a method that is based on the "Minimum Description Length" (MDL) algorithm to classify a set of attributes depending on their usefulness in making predictions. This method significantly reduces the time and resources necessary in the calculation of prediction models by selecting a subset of meaningful attributes through the elimination of redundant, irrelevant or informal attributes and identification og those attributes useful in making predictions.
- *Clustering* is a method of determining the similarities and dissimilarities between elements of a set, in order to group them into distinct and homogeneous classes. The method involves a classification with unsupervised learning, in which it is not a priori known either the number of possible classes or the inclusion of objects in certain classes

The Naïve Bayes and SVM classification algorithms will be applied to determin the likelihood of contracting a credit, regression to determine the maximum amount that can be awarded based on the financial situation of clients and clustering will be used for grouping customers into clusters representing financial and socio-demographic profile thereof.

3. The analysis and reporting module

In the initial phase we identified two specific reporting requirements: an operational level of reporting that regards current lending applications and a tactical reporting level, which refers to the analysis of activity on a higher level in various areas, territories, sales agents, and product categories, types and groups of customers

In this regard two reporting modules are required. A module within the operating system, which will be built along with the process management application and a multidimensional analysis module will be based on a Data Mart that could be integrated with existing decision support system within the organization.

For designing the data mart it is necessary to identify objects of type: sizes, hierarchies, tables facts, mappings, this being modeled using UML stereotypes presented in Table 2.

Table 2. Stereotypes defined for multidimensional modeling

Object	Stereotype	Meaning
Class	<<Dimensiune>>	Dimension class
Class	<<Tabela_Fapte>>	Facts class
Attribute	<<O_ID>>	Identifying attribute
Attribute	<<O_DESC>>	Descriptive attribute
Attribute	<<ID_Parinte>>	Parent attribute (dimension class)
Attribute	<<Atrib_Dim>>	Attribute (dimension class)
Attribute	<<Masura_Baza>>	Measure attribute (facts class)
Attribute	<<Masura_Derivată >>	Calculated attribute (facts class)

For multidimensional analysis the following entities are designed:

- *Clienti* dimension with information regarding name, status, sex, adress, county, group and client type;
- *Zone* dimension with information regarding the unit, region and area;
- *Produse* dimension, with information on the name, category and type of product, associated fees and amount limits;
- *Utilizatori* dimension, with information on salespeople, managers and their roles;
- *Cereri* facts table, in which the data in the loand application is summarized;
- *Punctaj_scoring* facts table, in which the scorecards obtained on the basis of loan applications are mantained.

The class diagram for the multidimensionl objects is shown in Figure 1.

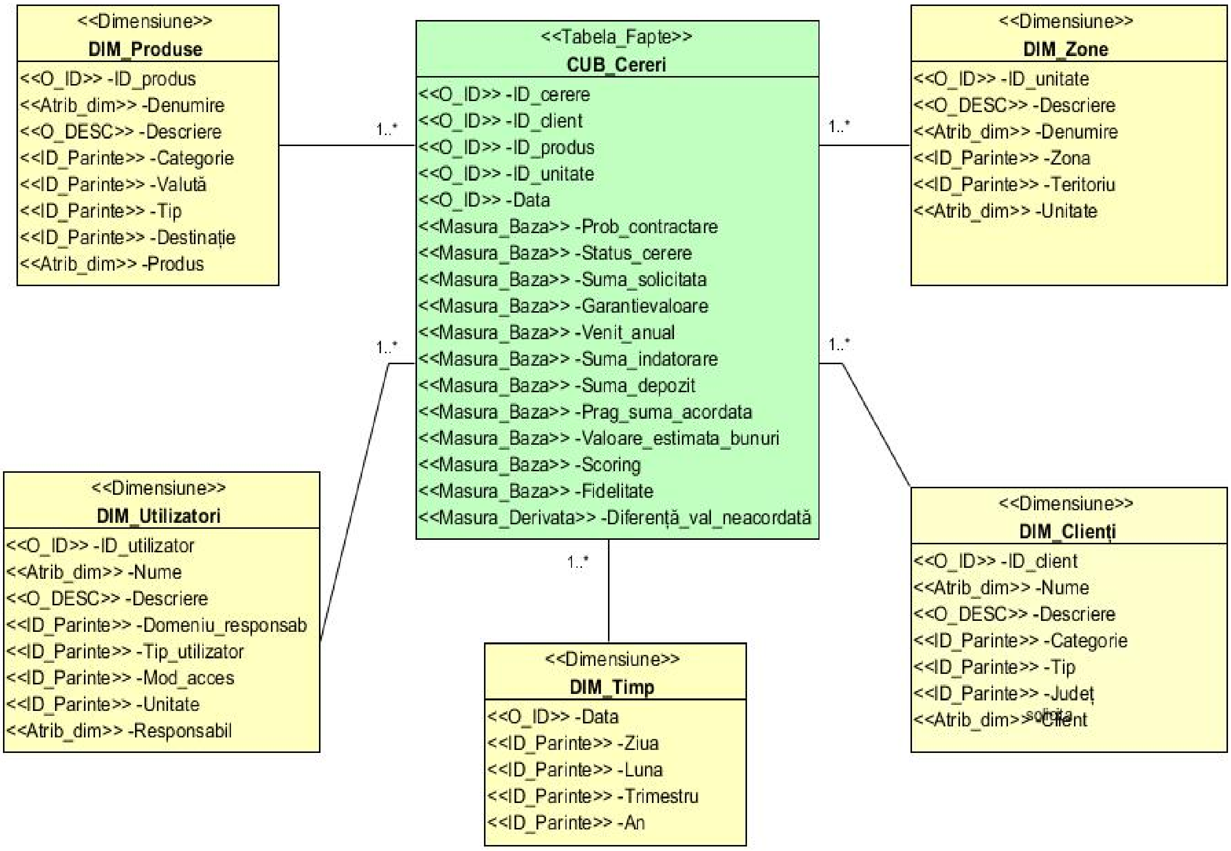


Fig. 1. Class Diagram for the Data Mart

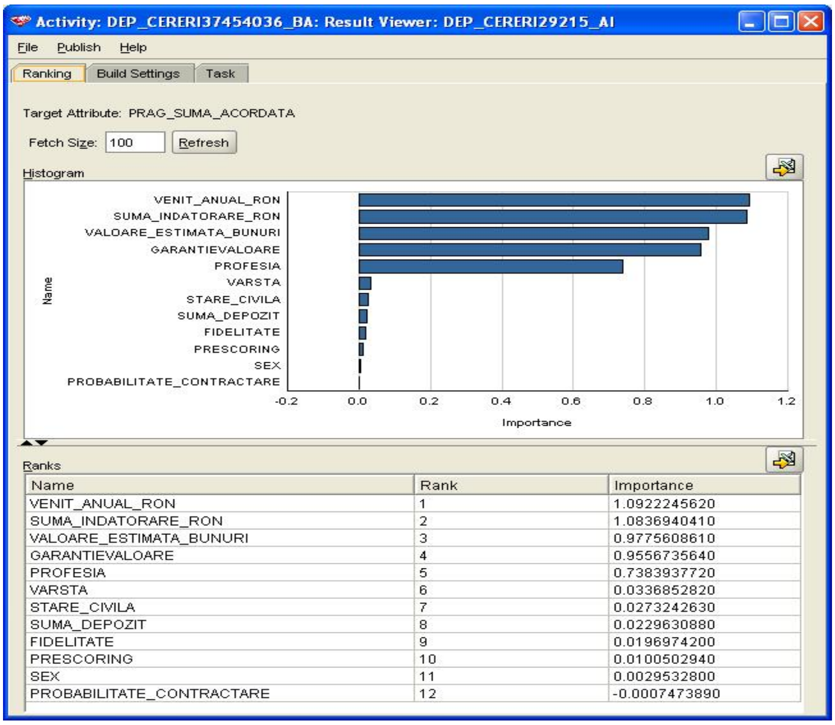
The objects of the multidimensional model can interconnect with the objects of the organizational data warehouse through XSD schemas.

4 Implementing the Data Mining algorithms

The full implementation of the system involves the realization of the prescoring / scoring algorithm used to calculate scores for each loan applications, implementing data mining algorithms to determine the likelihood of contracting a loan, the maximum amount that can be granted and grouping customers in clusters based on their financial and socio-demografic profile and building a data mart that can be used for multivariate analysis of the lending activity.

In the next section of the paper we will present how we implemented the Data Mining algorithms to determine the likelihood of contracting a loan.

We applied the algorithm for determining the important attributes and determine the maximum amount that can be awarded as shown in Figure 2.



Target Attribute: PRAG_SUMA_ACORDATA

Fetch Size: 100

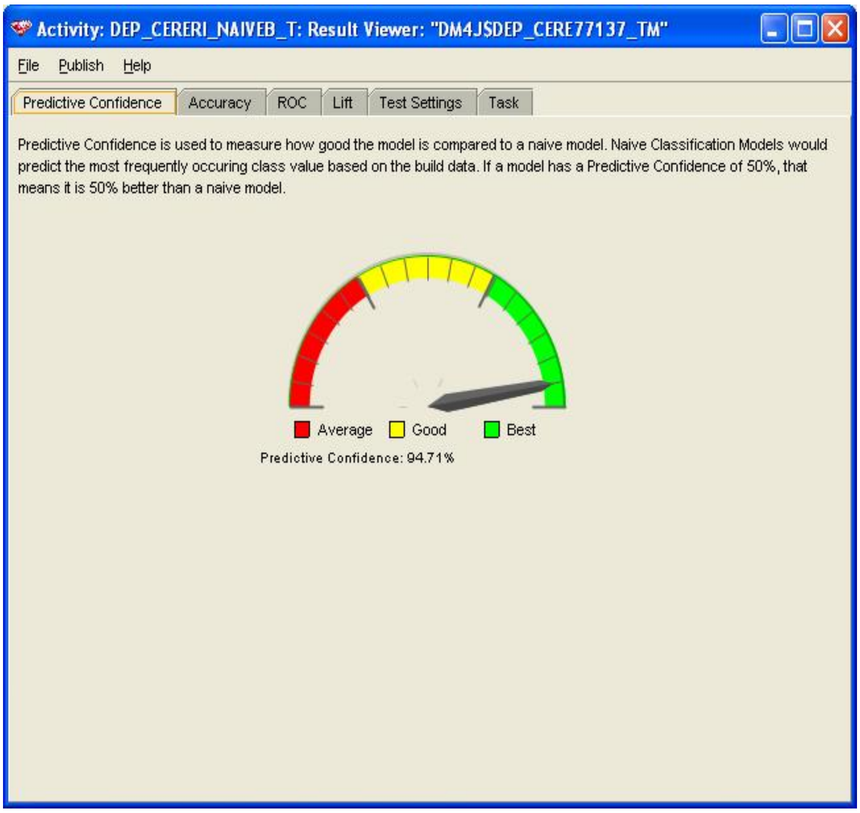
Name	Rank	Importance
VENIT_ANUAL_RON	1	1.0922245620
SUMA_INDATORARE_RON	2	1.0836940410
VALOARE_ESTIMATA_BUNURI	3	0.9775608610
GARANTIEVALOARE	4	0.9556735640
PROFESIA	5	0.7383937720
VARSTA	6	0.0336852820
STARE_CIVILA	7	0.0273242630
SUMA_DEPOZIT	8	0.0229630880
FIDELITATE	9	0.0196974200
PRESCORING	10	0.0100502940
SEX	11	0.0029532800
PROBABILITATE_CONTRACTARE	12	-0.0007473890

Fig. 2. The determination of significant attributes for the maximum amount that can be awarded to a client

We can observe from the analysis results that to determine the maximum amount that can be awarded to a particular customer, the attributes with the highest degree of relevance are: annual income, the total amount of debt, the total value of goods held by the client, the guarantees provided for the loan, profession, age, marital status, amount in the customer's deposits, if they exists, loyalty to the bank, prescoring/scoring value. These attributes will be used in the regression model that will determine the maximum allowable amount for lending.

5. Determining the lending likelihood

Next we determined the likelihood of granting the loan through two classification methods: Naive Bays and SVM. The target attribute is the probability of granting the loan with values 0/1, where 0 - loan is given 1 - no loan is given. After selecting the relevant attributes that are used in the algorithms and implementing the corresponding steps it is observed that models have a high accuracy of 94.71% and 94.34% for Naive Bays for SVM (Figure 3).



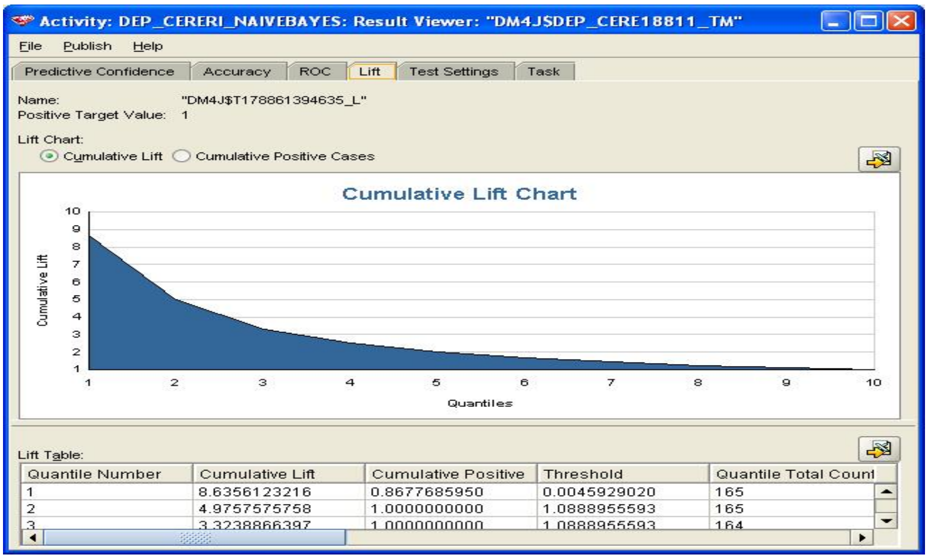
Predictive Confidence is used to measure how good the model is compared to a naive model. Naive Classification Models would predict the most frequently occuring class value based on the build data. If a model has a Predictive Confidence of 50%, that means it is 50% better than a naive model.

Predictive Confidence: 94.71%

Fig. 3. Naïve Bayes algorithm accuracy

To analyze the accuracy of the algorithm we analyzed a series of values that characterize the determination of the dependent attributes. Such a set of values is represented by the LIFT matrix which represents the learning rate of the model.

From the graph in Figure 4 we can observe a high rate of learning the model in the first three quintiles. Basically, the matrix is the ratio between the percentage of correct classification carried out and the percentage of real positive model classifications.



Name: "DM4J$T178861394635_L"
Positive Target Value: 1

Lift Chart:
● Cumulative Lift ○ Cumulative Positive Cases

Lift Table:

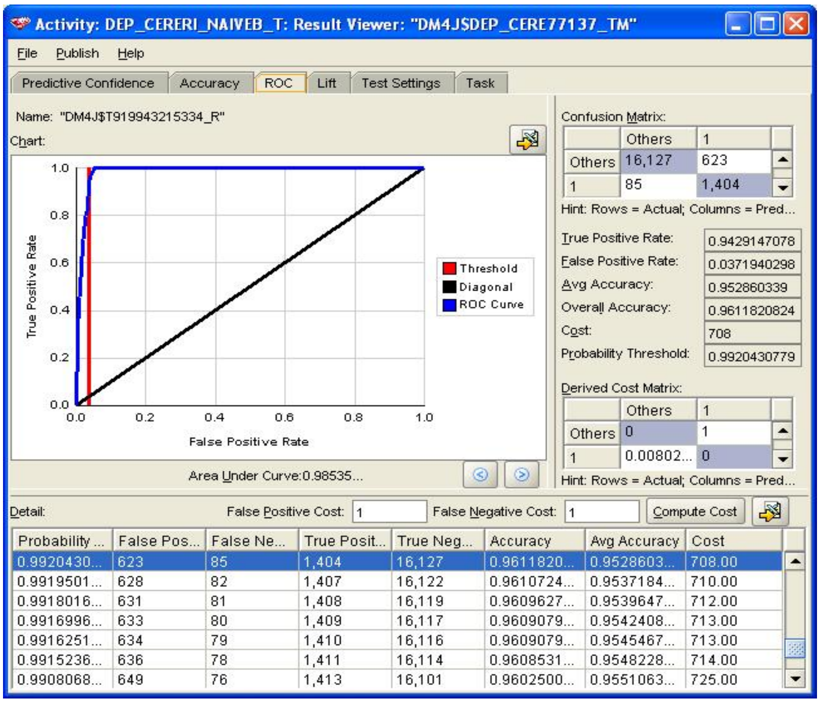
Quantile Number	Cumulative Lift	Cumulative Positive	Threshold	Quantile Total Count
1	8.6356123216	0.8677685950	0.0045929020	165
2	4.9757575758	1.0000000000	1.0888955593	165
3	3.3238866397	1.0000000000	1.0888955593	164

Fig. 4. LIFT Matrix for the Naïve Bayes algorithm

For the Naïve Bayes algorithm we can observe the ROC matrix (Figure 5) which represents a metric for comparison of existing (real) values with those predicted by the built model.

.

The ROC matrix, as well as the LIFT matrix, applies classification models and can be used to gain insight into the ability of the model to determine the values.



Name: "DM4J$T919943215334_R"

Area Under Curve:0.98535...

Confusion Matrix:

	Others	1
Others	16,127	623
1	85	1,404

Hint: Rows = Actual; Columns = Pred...

True Positive Rate: 0.9429147078
False Positive Rate: 0.0371940298
Avg Accuracy: 0.952860339
Overall Accuracy: 0.9611820824
Cost: 708
Probability Threshold: 0.9920430779

Derived Cost Matrix:

	Others	1
Others	0	1
1	0.00802...	0

Hint: Rows = Actual; Columns = Pred...

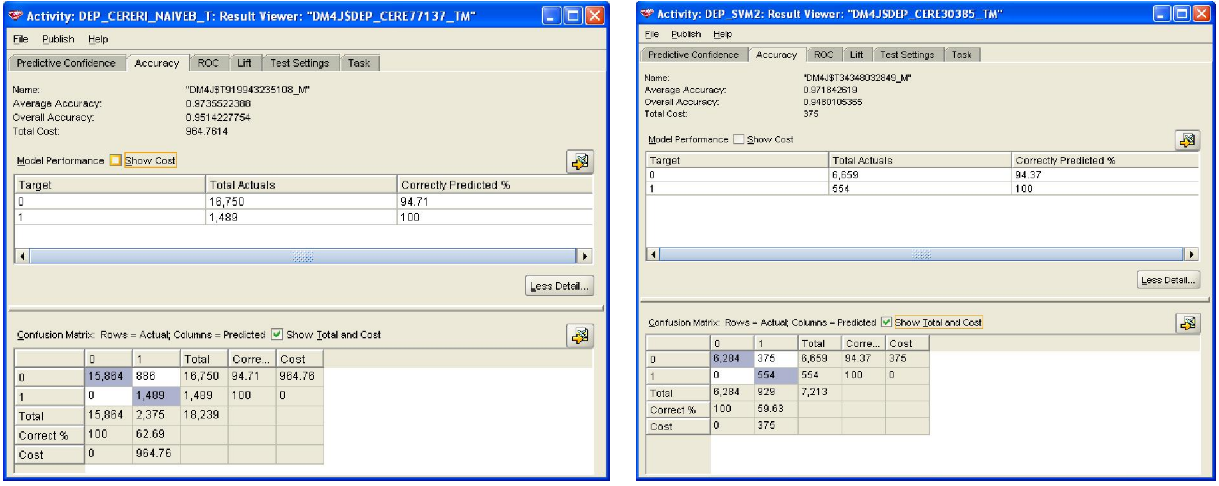
Detail: False Positive Cost: 1 False Negative Cost: 1

Probability ...	False Pos...	False Ne...	True Posit...	True Neg...	Accuracy	Avg Accuracy	Cost
0.9920430...	623	85	1,404	16,127	0.9611820...	0.9528603...	708.00
0.9919501...	628	82	1,407	16,122	0.9610724...	0.9537184...	710.00
0.9918016...	631	81	1,408	16,119	0.9609627...	0.9539647...	712.00
0.9916996...	633	80	1,409	16,117	0.9609079...	0.9542408...	713.00
0.9916251...	634	79	1,410	16,116	0.9609079...	0.9545467...	713.00
0.9915236...	636	78	1,411	16,114	0.9608531...	0.9548228...	714.00
0.9908068...	649	76	1,413	16,101	0.9602500...	0.9551063...	725.00

Fig.5. ROC Matrix for the Naïve Bayes algorithm

We analyzed the cost matrix for both classification models to compare the results and accuracy obtained.

For the Naive Bayes model there are 16750 instances associated with the value 0 (loan can be granted) of which 94.71% were correctly predicted and 1489 cases of default, corresponding to the value 1, from which the model has correctly predicted a percentage of 100 %



Name: "DM4J$T919943235108_M"
Average Accuracy: 0.9735522388
Overall Accuracy: 0.9514227754
Total Cost: 964.7614

Model Performance □ Show Cost

Target	Total Actuals	Correctly Predicted %
0	16,750	94.71
1	1,489	100

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

	0	1	Total	Corre...	Cost
0	15,864	886	16,750	94.71	964.76
1	0	1,489	1,489	100	0
Total	15,864	2,375	18,239		
Correct %	100	62.69			
Cost	0	964.76			

Name: "DM4J$T34348032849_M"
Average Accuracy: 0.971842619
Overall Accuracy: 0.9480105365
Total Cost: 375

Model Performance □ Show Cost

Target	Total Actuals	Correctly Predicted %
0	6,659	94.37
1	554	100

Confusion Matrix: Rows = Actual; Columns = Predicted ☑ Show Total and Cost

	0	1	Total	Corre...	Cost
0	6,284	375	6,659	94.37	375
1	0	554	554	100	0
Total	6,284	929	7,213		
Correct %	100	59.63			
Cost	0	375			

Fig. 6 –Cost matrix for :

a. Naïve Bayes algorithm

b. SVM algorithm

The total cost is 964.76 u.m. The confusion matrix indicates 0 false-negative predictions when the real value is 1 and 886 false pozitive predictions predictions when the real value is 0.

For the SVM model we used a smaller training set of 6659 instances related to the value 0, out of which 94.37% were correctly predicted and 554 cases related to the value 1 from which the model correctly predicted 100% . The total cost is 375 u.m. The confusion matrix indicates 0 false-negative predictions when the real value is 1 and 375 false-pozitive predictions when the real value is 0.

6. Evaluating the system

To evaluate the system there must be organized a series of working session with the users involved in order to validate the system's functionalities, identified its main evaluation criteria and developed a report on the degree of fulfillment of these criteria. We will present the identified criteria and theri degree of completion below.

1. *Integrating data from multiple sources - the degree of fulfillment: High.* Due to the implementation of the CRM system, data sources are multiple and diverse, being integrated into two functional modules namely the CRM system and data mart. For data integration various techniques and methods are used such as JDBC and bridge connectors, XML/XSD schemas and data warehouses. No data is taken from outside the bank and as such it does not influence the performance of the solution.
2. *Integration Service – the degree of fulfillment: High.* Service integration is done in the Microsoft CRM and SharePoint via plug-in components and Web services
3. *Portal integration – the degree of fulfillment: Average* – Portal integration is only performed for processing document of two types: client and collateral documents. There is no portal structure provided for integrating applications within the bank.
4. *Business process interoperability – the degree of fulfillment: High.* In the system we ensure the implementation of all business processes identified in the analysis phase through software components described previously. The processes are fully automated and run fluently and seamlessly and the interconnection of heterogeneous platforms.
5. *Flexibility – degree of fulfillment: High.* From both a technical and functional perspective the solution is flexible and can be easily adapted by adding new features such as integration with an ERP system or a full organizational portal. It is also The interconnection with objects in the Data Mart is also provided through an organizational data warehouse.
6. *Scalability – degree of fulfillment: High.* Thanks to the CRM, SharePoint portal and Data Mart system the resizing of solution can be achieved according to the changes in the organization without affecting overall system performance.
7. *Maintenance - degree of fulfillment: Average.* The solution was developed using five homogeneous servers in terms of the operating system but heterogeneous in terms of platforms and installed software products. Therefore maintaining the solution may cause problems if not properly monitored and specialists in database administration and management of operating systems are not involved.
8. *Decision support - degree of fulfillment: Average.* The system provides reporting tools for the current operations in the CRM system and business intelligence tools to realize analytical dashboards or reports dedicated to senior managers for analysis of the lending activity on different time periods, different types of products and depending on the customer profile.

9. *Performance – High.* Due to advanced technologies and next-generation database management systems used the solution offers a lower response time. Even during the multidimensional analyzes we can obtain a high performance in developing analytical reports. Also, the data processing is carried out consistently and rapidly, being used a single database to manage current activities.
10. *Friendly interface - High* - From the perspective of the end user using the solution through the use of mobile devices such as laptop, PDA, tablet or smartphone is a welcome feature in the easy conduct of activities. The interfaces present the information both graphically and in spreadsheet form via videoformats and exports can be made to spreadsheets or PDF documents. To use the system users do not need advanced knowledge in IT.

The analysis of these criteria by the managers involved in decision making can conclud that the system's functional and technical requirements are met and the solution is suitable for broad deployment and in other financial and banking institutions.

7. Conclusions

In the case study discussed in the paper we developed three data miming models in order to establish the likelihood of granting a loan, the maximum amount that can be granted and the grouping of customers based on their profile, presenting in detail the model of determining the maximum amount that can be granted. The values so determined will be associated to the loan applications to develop the final decision on granting or not granting the loan.

We have also built a Data Mart based in the entities of the CRM system in order to enable integration with an organizational data warehouse and also multidimensional analysis of current activities. The Data Mart is used for analytical reports which are summarized information on the volume of loans granted in different time intervals, by category, by product and by customer typology.

Finally, we evaluated the proposed solution on a series of criteria such as the degree of integration, interoperability, flexibility, performance, maintenance, scalability and decision support offered.

Acknowledgment: This paper was co-financed from the European Social Fund, through the Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/138907 "Excellence in scientific interdisciplinary research, doctoral and postdoctoral, in the economic, social and medical fields - EXCELIS", coordinator The Bucharest University of Economic Studies

References

[1] S. Chaudhuri, U. Dayal, V. Narasayya, An Overview Of Business Intelligence Technology, *Communications of the ACM* 54 (8): 88–98, 2011.

[2] J. Ranjan - Business intelligence: concepts, components, techniques and benefits, *Journal of Theoretical and Applied Information Technology*, Vol 9, No 1, 2009.

[3] Vlad Diaconita, Procesarea volumelor mari de date folosind HADOOP Yarn, *"Studii si Cercetari de Calcul Economic si Cibernetica Economica",* Nr. Special 1-2, pg. 43-51, ISSN:1843-0112

[4] Vlad Diaconita, Big Data and Machine Learning for Knowledge Management, *Proceedings of the 9th International Conference On Business Excellence*, Economica Printing Press, Bucharest 2014, pp 244-248, ISBN 978-973-709-738-5

[5] O.L Mangasarian. - Linear and non-linear separations of patterns by linear programming, Operations Research, Volume 13, Issue 3, 1965, pg. 444-452

[6] N. Friedman, N, D. Geiger, D, M.Goldszmidt - Bayesian Network Classifiers, *Machine Learning*, Volume 29 (2-3), p. 131.

Alexandra Maria Ioana FLOREA has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 2007 and also from the Faculty of Marketing in 2008. She has a PhD in Economic informatics and since February 2014 she is a lecturer. She teaches Databases, DBMS and Software Packages seminars and courses at the Economic Cybernetics, Statistics and Informatics Faculty. She is co-author of 3 books, has 8 articles published in prestigious journals included in international recognized databases (SCOPUS, Elsevier, EBSCO, ProQuest, or DOAJ) and also 22 papers in the volumes of national and international scientific manifestations, of which 5 are indexed Thomson ISI Web of Science and her fields of interest include integrated information systems, information system analysis and design methodologies and database management systems.