Academic Analytics in Customer Relationship Management Perspective using Data Mining

Reshma Desai
Assistant Professor
Computer Science Department
Thakur College of Science and
Commerce

ABSTRACT
Customer relationship management (CRM) comprises a set of processes and enabling systems supporting a business strategy to build long term, profitable relationships with specific customers. Customer data and information technology (IT) tools form the foundation upon which any successful CRM strategy is built. In addition, the rapid growth of the Internet and its associated technologies has greatly increased the opportunities for marketing and has transformed the way relationships between companies and their customers are managed. Many organizations have collected and stored a wealth of data about their current customers, potential customers, suppliers and business partners. Data mining tools could help these organizations to discover the hidden knowledge in the enormous amount of data. The emerging fields of academic analytics and educational data mining are rapidly producing new possibilities for gathering, analyzing, and presenting student data. Faculty might soon be able to use these new data sources as guides for analyzing student dropout rate, student retention, course redesign and as evidence for implementing new assessments and lines of communication between instructors and students. This paper uses the college admission data with data mining techniques to objectively and methodically comment on the retention percentage in the college.

General Terms
Customer Relationship Management (CRM), Data Mining, Customer Retention.

Keywords
Student Retention, Classification, Categorization.

1. INTRODUCTION
The emergence of Information Technology and use of computer in every field of activities has created a new buzz in the field of marketing and that is the concept of Customer Relationship Management (CRM). The concept of CRM is defined as "the process of acquiring, retaining and growing profitable customer which requires a clear focus on service attributes that represent value to the customer and creates loyalty".

Since customer relationship is neither a concept nor a project, instead a business strategy that aims to understand, anticipate and manage the needs of an organization's current and potential customer it is required to integrate both the software i.e. Data Mining tool along with the analytics of CRM.

Data mining techniques can help to accomplish extracting or detecting hidden customer characteristics and behaviors from large databases. The generative aspect of data mining consists of the building of a model from data (Carrier & Povel, 2003).

Each data mining technique can perform one or more of the following types of data modeling:

- Association
- Classification
- Clustering
- Forecasting
- Regression
- Sequence Discovery
- Visualization

Today there are various data mining tools available in the market. These tools can be broadly placed in following three categories

- General purpose tools
- Integrated DSS / OLAP / DM tools and
- Application specific tools

2. Overview of Data Mining Applications for CRM
From the CRM point of view, the data mining applications include but not limited to the following:

- **Customer Retention**: Sophisticated customer-retention programs begin with modeling those customers who have defected to identify patterns that led to their defection. These models are then applied to the current customers to identify likely defectors so that preventive actions can be initiated.

- **Sales and Customer Services:** In today's highly competitive environment, superior customer service creates the sales leaders. When information is properly aggregated and delivered to front-line sales and service professionals, customer service is greatly enhanced. If customer information is **available**, rule based software can be employed to automatically recommend products. The programs like market-basket analysis have already shown phenomenal gains in cross-selling ratios, floor and shelf layout and product placement improvements and better layout of catalog and web pages.

- **Marketing:** Marketing depends heavily on accurate information to execute retention campaigns, lifetime value analysis, trending targeted promotions, etc. Only by having a complete customer profile can promotions be targeted and targeting dramatically increase response rates and thus decreases campaign cost.

- **Risk Assessment & Fraud Detection:** An accessible customer base significantly reduces the risk of entering into undo risk. For example, a bank can identify fiscally related companies that may be in financial jeopardy before extending a loan to them.

2.1 Applications of Data Mining in Educational Industry

Identify risk factors that predict results:

One critical question in any educational institution is "What are the risk factors or variables that are important for predicting the results (pass/fail) of students?" Although many risk factors that affect results are obvious, subtle and non-intuitive relationships can exist among variable that are difficult, for not impossible to identify without applying more sophisticated analysis. Modern data mining models such as decision trees can more accurately predict risk than current models, educational institutions can predict the results more accurately, which in turn can result in quality education.

Student Level Analysis: Successfully training the student requires analyzing the data at the student level. Using the associated discovery data mining technique, educational institutions can more accurately select the kind of training to offer to different kinds of students. With the help of this technique, educational institutions can.

Segment the student database to create student profiles: Conduct analysis on a single student segment for a single factor. For example: the institution can perform in-depth analysis of the relationship between attendance and academic achievement.

Analyze the student segments for multiple factors using group processing and multiple target variables. For example, ―What are the characters shared by students who drop out from colleges?

Developing new strategies: Teachers can increase the pass percentage by identifying the most lucrative student segments and organize the training sessions accordingly. The results may be affected, if teachers do not offer the right kind of training to the right student segment at the right time. With data mining operations such as segmentation or association analysis, institutions can now utilize all of their available information for betterment of students.

2.2 Defining Data Mining and CRM Framework for the chosen field

The emerging fields of academic analytics and educational data mining are rapidly producing new possibilities for gathering, analyzing, and presenting student data. Faculty might soon be able to use these new data sources as guides for course redesign and as evidence for implementing new assessments and lines of communication between instructors and students.

In order to understand how and why data mining works, it's important to understand a few fundamental concepts. First, data mining relies on four essential methods: Classification, categorization, estimation, and visualization.

Classification identifies associations and clusters, and separates subjects under study.

Categorization uses rule induction algorithms to handle categorical outcomes, such as "persist" or "dropout," and "transfer" or "retention."

Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts.

Higher education institutions can use classification, for example, for a comprehensive analysis of student characteristics, or use estimation to predict the likelihood of a variety of outcomes, such as transferability, persistence, retention, and course success.

Data mining is already fundamental to the private sector. Many of the data mining techniques used in the corporate world, however, are transferable to higher education. The table below shows the higher education equivalents of critical business questions answered by data mining.

Table 1: Private sector questions and their Higher education equivalents [7]

Industry oriented questions	Higher education Oriented questions
Who are my most profitable customers?	Which students are taking the most credit hours?
Who are my repeat Web site visitors?	Which students are most likely to return for more classes?
Who are my loyal customers?	Who are the "persisters" at my university/college?
Who is likely to increase his/her purchases?	Which alumni are likely to make larger donations?
Which customers are likely to defect to competitors?	What types of courses will attract more students?

3. Review of related literature

One of the biggest challenges that higher education faces is to improve student retention. In general, more students remaining in the university means better academic programs and higher revenue.

To understand the factors influencing university student retention, questionnaires are often used to collect data including personal history of the student, implication of student behaviour, perceptions of the student, for example in (Superby et al., 2006) the authors applied different approaches such as decision tree, random forests, neural networks, and linear discriminate analysis to their questionnaires. However, possibly because of the small sample size, the prediction accuracy is not very good. Herzog(2006) collected data from institutional student information system, the American College Test's Student Profile, National Student Clearinghouse, SPSS software are chosen to estimate student retention and degree-completion time. Nearly 50 features including demographics, campus experience, academic experience, and financial aid are applied to predict student retention. The research shows that decision tree and neural networks performed better when larger data sets are available.

Student fee income also relates closely to student retention. For a medium size college who enrolls about 2000 new students each year. If 5% first year students drop out, the fee loss will be increased. Furthermore, dropped out students have a recruitment cost upfront and new students have to be recruited in order to keep college students number steady. The most widely accepted model in the student retention literature is Tinto's (Tinto, 1995). It examines factors contributing to a student's decision about whether to continue their higher education. It claims that the decision to persist or drop out is quite strongly predicted by their degree of academic integration, and social integration. Tinto argues that from an

academic perspective, performance, personal development, academic self-esteem, enjoyment of subjects, identification with academic norms, and one's role as a student all contribute to a student's overall sense of integration into the university (Tinto, 1995). Students who are highly integrated academically are more likely to persist and complete their degrees. The same is true from a social perspective. Student who have more friends at their university, have more personal contact with academics, enjoy being at the college, are likely to make the decision to persist. Poor retention is normally caused by unclear career goals, uncertainty about the course, lack of academic challenge, transition or adjustment problems, limited or unrealistic expectations, lack of engagement, and a low level of integration. According to Tinto, students are most likely to stay on the course if there are close links between their own academic objectives, and the academic and social characteristics of the college. If students find the particular course can combine education and their chosen subject, and greatly help them achieve their goals, their chances of completing would increase dramatically.

4. The Research objective

This research crystallizes the broad objective by using student admission data of a College in Mumbai for a period from 2007 to 2010 and using data mining techniques to objectively and methodically comment on the retention percentage in the college. These methods could be used across various colleges to find out their respective retention percentages.

5 .Data Cleaning and Preparation

The data was collected in excel and hence data cleaning and preparation of the data as required for Data Mining tools and techniques had to be meticulously done. During data exploration many data anomalies were discovered, like missing values, data type mismatch and so on.

The business understanding phase begins with setting goals for the data mining project. The scope this research in terms of data used is limited by the data available in the given system through the enrolment form used for collecting data from newly enrolled students. It is important to have a full understanding of the nature of the data and how it was collected and entered before proceeding further.

Data preparation is the most important and the most time consuming phase in data mining. In this phase the data are put into a form suitable for the modeling phase. If required some selected variables are combined, transformed or used to create new variables.
Data are cleaned for any duplication of records.

Some data mining and multivariate statistical methods are not able to deal with categorical variables measured on a nominal scale, but require a numerical variable. Therefore, for the categorical variables dummy variables were created, each with two possible values: 1 and 0.

Some of the attributes are completely irrelevant to the scope of studies but have been retained for further use if any.

6. Methodology and Analysis of data

Tools and Techniques used for Analysis of data

In the modeling phase different models were on the training data set. Then it was decided whether a suitable model for the data set was found was acceptable from both an analytical and a managerial standpoint.

1)For Classification-The Naïve Bayes algorithm was found most suitable for the given limited scope of data. Bayesian classifier try to model probabilistic relationships between the attributes and the class variable. It uses the well-known Bayes theorem to combine a priori information with evidence from data. Let X denote the attribute set and Y denote the class variable. The classifier learns the posterior probability P(Y jX) for every combination of X and Y, so a new record can be classified such that the posterior probability is maximal.

2)For Clustering- K-Means algorithm was found suitable. It created two cluster for the given two data sets gave out the difference in distance for each cluster.

A non-hierarchical approach to forming good clusters is to specify a desired number of clusters, say, k, then assign each case (object) to one of k clusters so as to minimize a measure of dispersion within the clusters. A very common measure is the sum of distances or sum of squared Euclidean distances from the mean of each cluster. The problem can be set up as an integer programming problem but because solving integer programs with a large number of variables is time consuming, clusters are often computed using a fast, heuristic method that generally produces good (but not necessarily optimal) solutions. The k-means algorithm is one such method.

K-Means Training starts with a single cluster with its center as the mean of the data. This cluster is split into two and the means of the new clusters are iteratively trained. These two clusters are again split and the process continues until the specified number of clusters is obtained. If the specified number of clusters is not a power of two, then the nearest power of two above the number specified is chosen and then the least important clusters are removed and the remaining clusters are again iteratively trained to get the final clusters. When the user specifies random start the algorithm generates the k cluster centers randomly and goes ahead by fitting the data points in those clusters. This process is repeated for as many random starts as the user specifies and the Best value of start is found. The outputs based on this value are displayed. The drawback of standard clustering methods is that they ignore measurement errors, or uncertainty, associated with the data. If these errors are available, they can play a significant role in improving the clustering decision. This approach to clustering is called Error based clustering. Error based clustering explicitly incorporates errors associated with data into the clustering algorithm.

3)For Prediction-Multiple linear regression algorithm was used which has shown minimum average error. This procedure performs linear regression on the selected dataset. This fits a linear model of the form

$$Y= b0 + b1X1 + b2X2+ .... + bkXk+ e$$

where Y is the dependent variable (response) and X1, X2,.. .,Xk are the independent variables (predictors) and e is random error. b0 , b1, b2, ... bk are known as the regression coefficients, which have to be estimated from the data. The multiple linear regression algorithm in XLMiner™ chooses

regression coefficients so as to minimize the difference between predicted values and actual values.

Linear regression is performed either to predict the response variable based on the predictor variables, or to study the relationship between the response variable and predictor variables.

6.1 Analysis of data

The data sets was divided into two parts

1) Records of Bachelor of science stream for the year 2007-2009 was considered for first dataset and 2010-2012 was considered for second dataset to find out the retention of students.

2) The students are termed retained by researcher for this project in the following ways.

- If they take admission into the next year in next class or
- next year in same class in case of failures, or
- If the student takes a break (a year or two) and comes back in the same college to further continue studies.
- The student can be retained for two year or three year if the student is consistently passing
- The student can be termed retained if the student completes the three year bachelor degree from this college.

6.2 Analysis for the first data set

Data partition of the data was done using Standard partition . All the attributes were considered for the data partitioning.

1)For Classification-The Naïve Bayes algorithm was found most suitable for the given limited scope of data

2)For Clustering- K-Means algorithm was found suitable. It created two cluster for the given two data sets gave out the difference in distance for each cluster.

3)For Prediction-Multiple linear regression algorithm was used which has shown minimum average error.

6.3 For the second dataset

1) The second data set was prepared after getting the admission data of the current year.

For Classification-The Naïve Bayes algorithm was found most suitable for the given limited scope of data.

Same step as above were repeated for the second data set.

7. Findings of study and Conclusion

In this project, Naïve Bayes classification algorithm is used for Student's retention in the college system , to predict the relevancy of the incoming student data from a new data set to the already existing data sets. The Naïve Bayes model shows 93.0% retention for the first data set, 95.5% retention for the second model the result shown may vary from the actual results, as XL Miner chooses only limited set of data for its working.

The empirical results show that we can produce short but accurate prediction of attributes for the student retention purpose by applying the Naïve Bayes classification model to the records of incoming new students.

Lift charts are visual aids for measuring model performance. They consist of a lift curve and a baseline. The greater the area between the lift curve and the baseline, the better the model.

Lift Curve and the Declie Lift curve obtained in both the case of data set used show the model performance on the given data as fair.

For Multiple Linear Regression for first data set total sum of squared errors summaries for the training and validation data sets are calculated. The total sum of squared errors is the sum of the squared errors (deviations between predicted and actual values) and the root mean square error (square root of the average squared error). Average error is typically very small, because positive prediction errors tend to be counterbalanced by negative ones.

For Clustering the model has created two clusters for both the data sets .the difference in distance between the clusters is shown in the above diagram

Finally, this study shows that data mining techniques can also have their use on less rich datasets, provided the data preparatory steps are carried out carefully. The basic analysis presented here shows an accuracy of 75% to 80% based on data collected during the enrollment of the student, and shows several ways of possible improvement without having to collect additional data. As a supplement to the study it was found out that a survey be conducted from the students who have been retained in the college for more than one year to find out more social environment which can also be the factors affecting the retention rate of the college

A sample survey of 125 B.Sc. second year and third year students was conducted to understand the social and other factors which affect the retention. Most of the students in the sample survey were found to have joined the college in 2010 for their first year. Some students had directly joined in second year or third year .The parent background survey showed that the mothers are less educated where as fathers were at least graduate. The students on an average came from middle class or upper middle class families.

Almost 77% of the students lived in the vicinity of the college where the distance is defined in the range of 10-15 km around the institution. This contributes to understanding that students whose who have stayed back chose nearby college.

Overwhelming number of students have been retained as they were able to study the stream they preferred. Almost 85.95%

are happy with the quality of teacher and rated it as satisfactory, good and even excellent.

Laboratory and Classroom infrastructure has played important role for the student retention as overwhelming

83.3% are stratified and rated it as satisfactory, good or excellent. Library facility seems to make students be stratified with it as it has large number of books and a book bank facility for students.

Social environment developed through extra circular and co-curricular activities contributed 85% as a retention factor and 93.39 % of the students are satisfied with good overall environment for their education.

8. Limitations of the study

The software used to conduct the research was XLMiner DEMO version. The version is freely downloadable but with limitation of 600 row set for standard partition. Hence the result obtained from the analysis may not be the actual representation of the overall data. Nevertheless the researcher in further studies could like to use other versions of the software's or any other software found suitable, which can deal with voluminous data.

Though large no of students join the first year some of them leave due to their admission in professional college whose results are declared late after first year admission is over. Hence to get a fair idea about the student admitted in the college for the first year the data has been tallied with the list of students appearing the first year examination.

9. Suggestions for further research

It was found out that even though the admission data can statistically contribute in finding the retention of the students further more attributes also can be used for data mining process. The survey conducted has found out many factors which have contributed to better understanding of the results, hence my suggestions is that admission data, results data, Library usage data, participation of students in different activities of the college along teachers feedback and in combination with other attributes which reveal more information about the students also can contribute of find the retention rate and find factors contributing to it.

ACKNOWLEDGMENTS

I extend my gratitude to Thakur College of Science and Commerce, Mumbai for sharing invaluable information for the conducted of the research project. This minor research project was funded by University of Mumbai 2011-2012.

REFERENCES

[1] Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques Mr. M. N. Quadri1 Dr. N.V. Kalyankar2 Global Journal of Computer Science and Technology

[2] Mining Educational Data Using Classification to Decrease Dropout Rate of Students Dr. Saurabh Pal;Department of Computer Applications, VBS Purvanchal University, Jaunpur – 222001 (U.P.), India; International Journal Of Multidisciplinary Sciences And Engineering, Vol. 3, No. 5, May 2012

[3] Md Rashid Farooqi, and Khalid Raza. "A Comprehensive Study of CRM through Data Mining Techniques." Proceedings of the National Conference; NCCIST-2011, pp. 61-65, 2011

[4] Use Data Mining To Improve Student Retention In Higher Education – A Case Study Ying Zhang, Samia Oussena Thames Valley University, London,UK ying.zhang@tvu.ac.uk, samina.oussena@tvu.ac.uk Tony Clark, Hyeonsook Kim Middlesex University, London,UK t.n.clark@mdx.ac.uk, hyeonsook.kim@tvu.ac.uk

[5] Prediction Of Student Academic Performance By An Application Of Data Mining Techniques Sajadin Sembiring, Faculty of Computer System & Software Engineering,

[6] Universiti Malaysia Pahang, Malaysia Teknik Informatika STT Harapan Medan, Indonesia

[7] Data Mining Applications in Higher Education Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories Executive report

[8] Building profitable customer relationshipswith data mining Herb Edelstein, President Two Crows Corporation

[9] Application of data mining techniques in customer relationship management:A literature review and classification E.W.T. Ngai a,*, Li Xiu b, D.C.K. Chau a a Department of Management and Marketing, The Hong Kong Polytechnic University, Hong Kong, PR China b Department of Automation, Tsinghua University, Beijing, PR China

[10] Enrollment Prediction Models Using Data Mining Ashutosh Nandeshwar Subodh Chaudhari April 22, 2009

[11] Predicting students drop out: a case study Gerben W. Dekker g.w.dekker@student.tue.nl Department of Electrical Engineering, Eindhoven University of Technology April 10, 2009

[12] A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell and Charles Kaprolet Arizona State University