CHAPTER 16

Use of data mining to examine an outreach call center’s effectiveness and build a predictive model for classifying future marketing targets

J. Luan, C. Summa & M. Wieland
Cabrillo College, USA.

Abstract

Data mining is demonstrated through the evaluation of an enrollment management call center at a community college in the US with predictive modeling of data mining being the dominant tool for the key research questions. In responding to the three research questions, the study found a statistically significant 5% higher registration rate in a four way post facto control and experimental analysis. Through examining the applicants by the manner in which they were reached by the call center, the study found that those who spoke directly with the call center staff had at least a 10% increase in registering for classes compared to those who had no contact with the call center. After removing the applicants who interacted with the call center directly, data mining analysis identified six rules for applicants who would register for classes and another six for those who would not. Further, the data mining analysis built a predictive model with an 85% accuracy to predict those who would be less likely to register in a future semester. In conclusion, there is enough evidence to suggest that data mining has a place in enrollment management.

1 Background

A call center is a medium through which organizations communicate with their members and with which service providers keep in touch with their customers, mostly through the use of telephones staffed by trained personnel. In higher education in America, call centers have been used to perform similar functions, with actual tasks varying from fund raising, to enrollment follow-up, to survey research,
such as those at Piedmont and Sinclair Community Colleges. Sometimes call centers are named ‘phone banks’. If done well, call centers may provide a positive impact on college enrollment.

Personal touch in an era of a fast-paced, impersonal lifestyle can leave a person reached by the call center with a lasting impression, not the least of which is the power of persuasion inherently present in a call from a college representative compared to receiving a piece of mail. To reinforce this concept, in the March 2005 issue of *Campus Technologies* [1], an article entitled ‘Getting Personal’ discussed strategies being implemented at various institutions to boost their enrollment. Ferris State, for example, attributed the increase of 2327 students up from 9495 a couple of years ago to customized recruiting. After providing many examples, the article stated that ‘Campuses … to interact with potential students have reported success in meeting their enrollment goals’. In the *University Business* [2] magazine published in the same month, the article ‘A New Definition of Marketing’ discussed the concept of engaging students through organizational as well as departmental marketing efforts.

Students enroll and leave college without graduation for many reasons [3]. Hossler in 1984 [4] systematically identified several factors influencing students’ enrollment. College reputation, cost, financial aid, all played certain roles. Active outreach, or customer relationship management concepts borrowing a modern term, did not receive significant mention in Hossler’s work. Neither was information driven analytical approaches, such as data mining. Rounds [5] in 1984 discussed several promising practices in attrition and retention of community college students, which helped shed light on interpreting some of the behaviors of students, such as peer influence and inertia. Luan and several scholars [6, 7] in 2004 and 2005 presented case studies to demonstrate the use of data mining in managing applicants’ yield, outreach, and predictive modeling for both registration and attrition. Their work helped with designing the data mining approach.

Spring 2005 marked another round of California community college fee increases enacted by the California legislature. Past fee increases had proven to negatively impact enrollment. For example, one study conducted by Borden [8] showed that for every fee increase, there had been a corresponding drop in enrollment. A $13 increase would effect a 6% drop in headcounts, which translates into hundreds of students not enrolling. In confirmation of this, in 2004 the Chancellor’s Office for the 109 California’s community colleges estimated that system-wide annually a total of 142,500 students would be ‘lost’ due to such an increase [9].

In anticipation of a potential enrollment dip, Cabrillo College’s Marketing & Communications department, with assistance from several areas of the college, helped form the enrollment fee task force with the goal to maintain, if not increase, the spring 2005 enrollment. Among many areas identified as worthy of improvement, the issue of low registration rate from a large pool of applicants rose to the top. Although a majority of the applicants would register, thousands may never proceed beyond turning in their applications. Back in spring 2004, a total of 1836 applicants out of 7137 did not register for classes. Therefore, one of the strategies identified by
the taskforce was to direct call center outreach activities to those who have applied, but have not registered in spring 2005. As of January 18, 2005, roughly a month before the spring census day, a total of 2649 applicants were identified. Between January 10 and January 30, 2005, Cabrillo College, with the generous donation of time from 50 volunteers, made calls to a list of applicants at the beginning of the spring 2005 semester.

The purpose of the call center was to maintain or increase enrollment for spring 2005. The specific objectives of the call center effort were the following:

- Primary: to remind and encourage students to register.
- Secondary: to help students resolve problems stopping them from registering.
- Tertiary: to gather data about registration problems and identify any trends.

2 Three key questions addressed

This study addresses three key questions in evaluating the spring 2005 call center effectiveness. The general term of effectiveness used here means to include yield rate and productivity. Yield refers to the number of applicants who have become registrants as a result of the call center’s efforts. Productivity refers to the average units taken by these registrants. An additional question for designing a predictive model by pre-classifying future applicants into groups scaled according to their likelihood to register is also explored by the study. A predictive model would reduce the cost of call center by identifying those who are less or least likely to register, so that calls are more focused and targeted.

Specifically, this study addresses three questions:

- Question one (yield): How many of the applied-but-not-registered applicants became registrants as a result of being reached by the call center?
- Question two (productivity): What are the average units taken by the registrants as compared to other registrants who were not directly reached by the call center?
- Question three (predictive modeling): How many applicants can be predicted to be less likely to register so that the call center can concentrate on these applicants?

For the sake of saving space, answers to the productivity question are not stated in this chapter.

3 Data sources

Cabrillo College’s Computing Resources (CR) department provided lists of applicants who applied, but had not registered, for select dates based on request from the call center volunteers. CR also provided summary counts of applicants for both the current semester and the historical spring 2004 semester. The call center provided feedback data in the form of notes taken by the call center volunteers. The Planning and Research Office (PRO) conducted data matching where possible prior to conducting statistical analyses.
4 Design and method

In order to answer all three questions, this study employed a variety of methods and tools. The study adopted a post facto control and experimental design for seeking answers to question one. Chi-square statistics was used for question one. Regression equations, neural networks, and classification and regression tree (C&RT) were used for algorithm bias analysis for question three. Also, data were split into training and test sets for empirical accuracy validation. Data warehousing and SQL (structured query language) technologies directly supported the datasets merging and querying tasks.

The results of the call center were hand coded into an Excel worksheet that was imported into Brio Query, a business intelligence (BI) tool for the purpose of querying and pivoting variables (building various reports). Most of the answers to question one and question two are provided by Brio Query, assisted by Excel and SPSS (another statistical analysis tool). For question three of predictive modeling, the study utilized data mining and a tool called Clementine, a leading industrial strength business analytics (BA) application.

The study chose Clementine as the data mining tool because of its ability to directly interface with static or live relational databases, to calculate new fields using GUI (graphical user interface) guided nodes, to convert transactional data files into analytical data files, and to allow infinite number of scenarios to be built and examined using its 16 modeling algorithms. All analyses are conducted inside one data stream, which makes it much easier for cross-validation, interpretation, replication and documentation. The screenshot in Fig. 1 illustrates the ‘data stream’ built within Clementine for the entire study, including the nodes used for calculating new fields (variables).

![Screenshot of Clementine data stream on design canvas.](image_url)
Since Brio Query queried the datasets and produced a data cube containing most of the needed data elements that lent themselves readily as input variables for the data mining tasks, Clementine directly imported a tab delimited text files from Brio Query as its data source.

5 Findings

5.1 Yield

The answer to question one was obtained through two separate steps. The first step examined the differences in registration rates between a control group and an experimental group. The experimental group would be the group that had the presence of a call center and the control group had not. To compute specific yield rates, which was the second step, required those in the experimental group who were identified to be those who applied but had not yet registered for classes for the call center to contact. Not all applicants could be contacted by the call center.

5.1.1 Step one for answering question one: overall effect of the presence of call center

As the first step in addressing question one, the study made refinement to the original pair of a control group and an experimental group by splitting them further. The rationale is as follows. As mentioned earlier, all those who turned in their applications as of spring 2004 semester census date (February 23, 2004) became a pseudo-control group because no call center activities took place in that semester. All those who had their applications on file as of spring 2005 semester census date (February 22, 2005) were the experimental group. The call center only functioned for a brief period of time, a month before the start of the spring 2005 semester and the college continued to receive applicants since the lists of applicants were extracted for the calls. This has provided a good opportunity to examine the registration rates with and without the call center in the same semester. Therefore, applicants in both groups were then split by a specific date. For spring 2005, the date of January 19 was chosen because none of the applicants who turned in their applications after 19 January were contacted. This group is called ‘Pool A’. January 19, 2005, was 31 days before the census day of spring 2005. For spring 2004, the date of January 20, 2004, was chosen (31 days before spring 2004 census date). This group is called ‘Pool B’. Hypothetically speaking, if call center had no effect, then the rates for Pools A and Pools B in their respective semesters should be very similar.

The following table presents the rates of registration for the control (Pools A & B) and experimental groups (Pools A & B).

Table 1 shows that those spring 2005 students in Pool A had a higher registration rate than Pool B. The difference is 5%. Both the equivalent pools of students in spring 2004 showed no change in their registration rate.

Is the observed 5% difference statistically significant? The study turned to chi-square analysis for answers. The following output from chi-square analysis showed a high level of significance.
Table 1: Registration rates by treatment and control groups.

<table>
<thead>
<tr>
<th>Pool</th>
<th>Spring 2005</th>
<th>Spring 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Call Center Available 71%</td>
<td>No Call Center 74%</td>
</tr>
<tr>
<td>B</td>
<td>No Call Center 66%</td>
<td>No Call Center 74%</td>
</tr>
</tbody>
</table>

Difference = 5%  Difference = 0%

Note:
1. Pool A contain all applicants who turned in their applications as of spring census date. For Pool A from spring 2005, a sub-group of the applicants, namely those who applied but had not registered as of the date of January 19, 2005, was the group contacted by the call center and consequently discussed in detail in the study.
2. Online application was made available for the first time in spring 2005. A great number of students utilized it to apply online. This is a key difference other than a call center available to both Pool A and Pool B in spring 2005.

\[ \chi^2 = \sum \frac{(O-E)^2}{E} = 24.52 \]

degrees of freedom = 1

\[ p = 0.000001 \]

Figure 2: Chi-square output.

The chi-square analysis indicates that the observed 5% difference was statistically a significant event for the registration rate for those who applied but had not registered in spring 2005. The occurrence of a difference of 5% purely by chance is deemed to be one in 100,000, or very unlikely.

Although the 5% difference is considered statistically significant, in order to completely answer the question on the yield rate the next step is to look at the actual number of yields.

5.1.2 Step two for answering question one: computing specific yield rates

The applicants under study have been categorized into several distinct groups. Those who spoke with the call center staff directly and said they were going to register were in the Promised group. Those who received a voice mail message from the call center volunteers were in the Left Msg. group. Those whose phones never answered were in the Not Accessible group. Those who spoke Spanish only were in the Spanish Spk group. Those who provided a wrong number on their applications were in the Wrong No. group. Those whose phone numbers had an area code other than 831 were in the Out of Area group. Table 2 contains detailed information on these groups.

In Table 2, of the applicants who had not registered by January 19, 2005, 370 of them spoke with the call center staff directly and said they would register. Eventually, 194 of them were found to have registered as of spring 2005 semester census day, thus producing a yield rate of 52% for the Promised group. Across all
categories of applicants in Table 2, this is by far the highest yield rate. The next
group that had the highest yield rate (48%) was those who received a voice mail
message from the callers.

The study removed all cases from all categories if they were found to have a
registration date prior to January 19, 2005. This will help with making sure that
the subjects under study have not been included in error. Secondly, the researchers
went through the actual survey forms filled out by the call center volunteers and
paid particular attention to those in the Promised group. The purpose of examining
the actual feedback from the applicants was to get a sense of the reasons behind
those 194 direct yields. Many of them stated reasons such as ‘not clear on what to
do next’, ‘have not gotten the time’, or ‘procrastinating’. Many were thankful that
they got the call. It was clear that they indeed may not have registered if they had
not receive the calls.

Apart from those who received a voice mail message from the callers, those who
registered without speaking directly with the call center volunteers (or without
being reached by the callers) can be regarded as those who enrolled in classes of
their own volition; therefore, subjects with no contact from call center (without
receiving any treatment). They were the groups of Not Accessible, Wrong No. and
Out of Area. After collapsing the above six application types (Table 3) into only
three, the yield rate of the applicants for the Promised group is ranked 11% higher
than the All Other group and is still 3% higher than all others after it is combined
with the Left Msg. group (Promised & Msg) (Fig. 3).

5.2 Predictive modeling

Question three: How would predictive modeling help with identifying among the
future applicants those who are less likely to register to help better focus on calls
made by the call center?
Table 3: Yield rates by collapsed applicants types.

<table>
<thead>
<tr>
<th>Applicants*</th>
<th>Yield</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promised</td>
<td>370</td>
<td>194</td>
</tr>
<tr>
<td>Promised &amp; Msg</td>
<td>1212</td>
<td>594</td>
</tr>
<tr>
<td>All Other</td>
<td>633</td>
<td>261</td>
</tr>
</tbody>
</table>

Promised denotes those who spoke to the callers directly and promised to register. Promised & Msg denotes those of the Promised group as well as those who received a voice mail message. All Other denotes those who were not reachable, wrong number, and out of the area (who are likely distant or non-residents).

*Applicants column has been revised by removing those who had already registered by January 19, 2005.

To answer this question, the study first used a Venn diagram to group the 2949 applicants based on their presence in the college’s MIS data warehouse and the spring 2005 census database. A Venn diagram, developed by John Venn, a logician, is an effective way to visually identify subgroups (also called sets in algebra) of any populations, particularly when there is a large amount of overlapping. Venn diagrams have found use in many fields, such as database design, combinatorics [10], and assessment of learning [11].

The following Venn diagram (Fig. 4) indicates the overlapping of the population of 2949 applicants with the colleges’ MIS historical data warehouse and the spring 2005 census database. Four distinct groups are therefore clearly visible. They are:
Group 1: Those who applied for spring 2005, had enrolled at Cabrillo before, but did not enroll in spring 2005 ($n = 521$);
Group 2: Those who applied, enrolled, and had enrolled at Cabrillo before ($n = 1094$);
Group 3: Those who applied, enrolled in spring 2005, but had no prior enrollment ($n = 737$); and
Group 4: Those who applied for spring 2005, did not enroll, and had no prior history ($n = 597$).

These groups helped the rest of the analysis by making it possible to focus on each of them while drilling down to its population details. They guided the rest of the study and are frequently referenced.

Since Venn diagrams do not display data in proportion to their distributions in a dataset, a pie chart below will correct that by making the distributions adjusted to their appropriate scale. First, for those who eventually registered, the majority (62%, Groups 2 and 3) of the 2949 applicants eventually registered as of spring 2005 census time. At least a quarter (25%) of the applicants was new to Cabrillo College because no prior academic records existed for them in the college MIS historical data warehouse going back 15 years.

For those who never registered, regardless of being reached by the call center or not, over a third (38%, Groups 1 and 4) of the 2,949 applicants on the call list did not eventually register, but half of them (Group 1) had attended Cabrillo College before. The other half, or 20% of the 2949 applicants, had never been to Cabrillo College.
Figure 5: Registration percent distribution of applicants to be called \((n = 2949)\).

Overall, 45\% (Groups 3 and 4, \(n = 1334\)) of the 2949 applicants had never been to Cabrillo College before. It is unique to have so many of the potential ‘new’ students among the 2949 applicants to be called by the call center. Was there a reason for a disproportionate number of applicants who had never been to Cabrillo College to be slow in registering for classes? Spring 2005 enrollment statistics showed that a total of 2741 students enrolled were new students. Hypothetically, the number of new students could have been 3338 \((2741 + 597 \text{ of Group 4})\). In other words, 17.8\% \((597/3338)\) were missing from the new students pool.

The pie chart (Fig. 5) also seems to indicate that the 2949 applicants almost had an equal 25\% chance in falling into any of the four groups. Overall, having been to Cabrillo College seemed to increase the chance of registering for classes (37\%, Group 2). For those who had never been to Cabrillo, their chances of registration were about 50/50. This means that the outcomes of the applicants are really a set of four: those who had been to Cabrillo but did not register and those who did register; those who were new to Cabrillo and registered; and those who did not register and their prior background information is unknown.

The following five charts (figures) and tables display background characteristics of the groups identified in the Venn diagram. However, Group 4 is not in any of the analysis due to lack of data.

Figure 6 above shows the distribution of age ranges across the three Venn diagram groups.

Overall, the age of students in Group 1 (those applicants who had been at Cabrillo before but did not eventually register) was higher than the other two groups. Compared to Groups 2 and 3, Group 1 had fewer students younger than 20. The reverse is true for Groups 2 and 3.

Students ages 17 and below or 19 and below as shown in Fig. 5 are likely concurrently enrolled students. For Group 2 (the group of applicants who had taken classes at Cabrillo College and registered), there were fewer students in the age range of a recent high school graduate (18–19) compared to Group 3 (new applicants who had never been to Cabrillo). Comparing Group 2 to Group 3, fewer
students in Group 2 were from the age range of 18 to 19. The missing ones may have been recent high school graduates who had decided to move on following their study at Cabrillo College.

Figure 7 above shows the distribution of gender across the three Venn diagram groups. Across the three groups, gender seemed to have an opposing trend. More females were in Group 1, less in Group 2 and much less in Group 3, but the reverse is true for males. New applicants (Group 3) tended to be male. Applicants who had been to Cabrillo College and had not registered tended to be female.

Figure 8 above shows the distribution of ethnicities across the three Venn diagram groups. There is no major difference across major ethnic minorities among all three groups of applicants. There appeared to be fewer white students in Group 3 (new applicants without Cabrillo College experience) while there is an increase in the Unknown category in Group 3. Research has shown that most of the students in ‘unknown’ or ‘unreported’ categories tend to be White students.

Figure 9 above shows the distribution of education background across the three Venn diagram groups. There were fewer concurrently enrolled students in Group 1 (those applicants who had been to Cabrillo College, but never registered) compared
to other groups. This is a validation of the observations made about their age. More of them in Group 1, on the other hand, were high school graduates, but not necessarily recent high school graduates.

Figure 10 above shows the distribution of enrollment status across the three Venn diagram groups.

The largest portion of Group 1 was those who were continuing students when they applied. The largest portion of applicants in Group 2 was those who were new. The largest portion of applicants for Group 3 was those who were new, too.

Distributions of enrollment status for the three groups of applicants were generally very diverse. Very few in Group 1 were concurrently enrolled students. Very few in Group 2 were continuing students when they applied.
5.2.1 Data mining rationale and discoveries

The above visual analysis by five select background variables of demographics and academic status help develop an impression of the different characteristics of the applicants in Groups 1, 2 and 3. However, the impression is at best a fuzzy one, not accurate or evidential to help classify individual future applicants into respective groups. Plus, should one decide to cross tabulate the groups by two or more variables, the examination of these background variables using the method above can go on forever and there can be infinite number of tables and charts.

A 3-dimensional environment with multiple variables intermingled in a myriad of ways is impossible for the human eyes to quickly spot trends or monitor changes. It is precisely such a spatial sphere in which hidden patterns exist that can lead to new information on how the applicants become a registrant. Conventional tools may start to show inadequacies in handling the infinite number of coordinates in such a spatial sphere. Even the traditional regression analysis, which computes the residual statistics of multiple variables around a mean to determine their contribution to explain the variance of a phenomenon, is not entirely adequate.

For example, there may be a dozen variables going into a regression analysis to identify the key ones that would determine whether or not an applicant would register. The regression analysis may find that among these variables gender, ethnicity, age, location, and GPA would be significant. It then provides a specific value (odds ratio) associated with each variable to determine the likelihood of an applicant’s registration status. The equation typically functions as a polynomial model: if the value of a variable changes by one unit, such as one year of age increase for an applicant, the likelihood of registration would change by an X amount. Plugging in an applicant whose age is 18, gender is male, and ethnicity is Asian, out comes the likelihood of his registering for classes.

This likelihood is essentially a quadratic equation following a Sigmoid curve (regression line) with its values ranging from 0 to 1. Every new case is fit into
this regression equation. The sizes of the confidence intervals drawn for individual cases are much greater or wider than those drawn for the averages predicted by the regression lines [12]. For example, not necessarily every 18 year old, male, Asian applicant would be likely to register. Some 18-year-old male Asian applicants may register because of variables A, B, and C, and others because of variables C, D and F.

Another regression test, the linear regression has long been considered inadequate to address most social science problems and has remained as an introduction to modern regression statistical analysis [13]. Approaches to nonlinear data do exist, such as Nonlinear Regression (NLR), but it requires both input and output data to be quantitative. Many data mining algorithms do not have such requirement. In addition, the Neural Net algorithm, for example, does not make assumptions about the data, nor does it require data to be coded any particular way [14].

Jesus Mena in 1998 [15] observed the differences between statistics and data mining by classifying traditional statistics as a top down approach and data mining bottom up. Mena stated that traditional statistics require the researcher to first formulate a theory in the form of a hypothesis and then query the data to prove or disprove the theory. Data mining on the other hand relies on data to self-organize and may not ever rise up to the level of formulating a theory [16]. Mena also discussed scaling both the analytical prowess and data file to speed and size that is particularly important in a fast moving business environment. For example, according to Mena, a typical database may contain 100 attributes about customers, which would mean, for any customer, there would be $100 \times 99 \times 98$ possible combinations. A small task such as classifying customers into high, medium or low groups, would mean $970,200$ ($100 \times 99 \times 98$) possible combinations.

The rest of the study moved beyond data visualization and employed predictive modeling data mining technique. Specifically, the study used Neural Net and C&RT nodes. Both of them are predictive modeling nodes based on artificial intelligence and machine learning. The predicative modeling nodes, to state simply, has the ability to study known cases in order to make predictions onto unknown cases. The unknown cases would be incoming applicants for a future semester. Along the way, data mining also provides a number of sophisticated ways of examining and describing the data.

Not all cases have been used for the predictive modeling. First, hundreds of applicants spoke directly with the call center staff, which could have contaminated the subjects because of the extra intervention they received. The direct interaction with the call center staff would artificially make the likelihood of registering for classes higher, particularly such an interaction has been proven to be effective in generating enrollments. Therefore, the study removed these 370 applicants. Secondly, the study removed the entire Group 4, because there was no background data for this group of applicants who never registered. Thirdly, Group 3 (those who had no prior records at Cabrillo but eventually registered) presented another dilemma. The most significant marker for this group is that they had no prior records, which would trick the algorithm to put too much weight on this fact that could overwhelm all other variables. Therefore, the entire group was removed as well. The remaining 1413 applicants came from Group 1 ($n = 463$) and Group 2 ($n = 950$).
Between applicants registering for classes and not registering for classes, of interest to the study is those who did not register. The reason is simple. If data mining algorithms can predictively identify those who are less likely to register, then the call center staff can concentrate on these applicants, which is a far effective use of time and resources. Defined as between group accuracy, all the applicants who did not register (Actual) should be 'predicated' as not having registered (Predicted). The idea is for the data mining algorithms to first examine all the variables associated with each applicant’s registration outcome to learn the rules and then apply the rules to similar cases to predict their registration outcome. Since the focus of the predictive modeling is on those who would not register, the level of Between Group Accuracy should be higher than 85%, which means out of all the applicants predicted to be less likely to register, it would get 9 out 10 of them correct. The outcome variable, S05Regd, is binary: registered ‘S05’ and not registered ‘NoS05’.

The Neural Net node was run first with the sensitivity statistics shown in Table 4. The sensitivity statistics from the Neural Net node helps the analysts understand the level of importance of input variables that determine an outcome. In this case, the outcome is registered or not registered. The node indicates (Table 4) that the enrollment status is the most important variable and gender is the least important variable.

Yet, the predication accuracy by the Neural Net node is not impressive. The between group predictive accuracy of Neural Net node for those who did not register is low (55.9%), evidenced by the following matrix (Table 5). This represents a slightly better chance than a coin toss. Neural Net node was not used after this discovery.

The C&R Tree node was used next without any adjustments inside the node to accommodate for misclassification. Misclassification is the equivalent of false positives. It is often used when the analysts choose to err on the side of predicting more cases favoring one outcome over the other, such as not registering for classes, so as to increase the chances of getting more cases of one side (type) correctly predicted at the expense of the other.

Table 4: Relative importance of inputs of Neural Net node.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>System variable names</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrollment Status</td>
<td>Group of EnrlStatAll</td>
<td>0.40344</td>
</tr>
<tr>
<td>High School Origin</td>
<td>HighSchAll</td>
<td>0.33039</td>
</tr>
<tr>
<td>Age (in ranges)</td>
<td>AgeRange2005</td>
<td>0.307335</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Group of EthnicityAll</td>
<td>0.30281</td>
</tr>
<tr>
<td>Education Background</td>
<td>Group of EdStatusAll</td>
<td>0.249625</td>
</tr>
<tr>
<td>Number of Terms Previously Enrolled</td>
<td>CntDTL</td>
<td>0.176079</td>
</tr>
<tr>
<td>Gender</td>
<td>GenderAll</td>
<td>0.115245</td>
</tr>
</tbody>
</table>

Note: Neural Net node: 71.2% accuracy; 59 neurons, 1 hidden layer; Quick, seed = true.
Table 5: Between group accuracy matrix by Neural Net node.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted Not Reg’d</th>
<th>Registered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Reg’d</td>
<td>Count</td>
<td>259</td>
</tr>
<tr>
<td>Row %</td>
<td>55.9</td>
<td>44.1</td>
</tr>
<tr>
<td>Registered</td>
<td>Count</td>
<td>160</td>
</tr>
<tr>
<td>Row %</td>
<td>16.8</td>
<td>83.2</td>
</tr>
</tbody>
</table>

Table 6: Between group accuracy matrix by C&RT node.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted Not Reg’d</th>
<th>Registered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Reg’d</td>
<td>Count</td>
<td>393</td>
</tr>
<tr>
<td>Row %</td>
<td>84.9</td>
<td>15.1</td>
</tr>
<tr>
<td>Registered</td>
<td>Count</td>
<td>344</td>
</tr>
<tr>
<td>Row %</td>
<td>36.2</td>
<td>63.8</td>
</tr>
</tbody>
</table>

In Table 6, the between group predictive accuracy produced by the C&RT node for those who did not register is much better (85.7%) than those who eventually registered (70.4%).

The rest of the analysis utilized a decision tree graph and rule sets from the C&RT node.

Figure 11 is a partial screenshot of a binary decision tree produced by the C&RT algorithm. The tree is six branches deep from the first node of registration status ($R$-$S05Regd$), but only the first four top branches can fit on a landscape page in this report. The top branches typically contain variables more influential to the outcome than the rest toward the bottom.

In the decision tree graph (Fig. 11), applicants' latest Enrollment Status seemed to be most important, therefore, it became the first split. Those who were Concurrently Enrolled, New or Returning applicants ($n = 905$) were split from those who were Continuing or with unknown status ($n = 508$). Of those 905 applicants, 21.3% ($n = 193$) did not register and 78.7% of them did. For the sake of understanding the tree graph, let us focus on these 193 applicants who did not register. In the next split that occurred on the variable of Education Background, 183 were left. They were those who had already obtained a Bachelors degree, or a Foreign High School Diploma, or a General Education Degree, and those who either graduated or did not graduate from high school. The next split took place on the variable of number of terms enrolled at Cabrillo College prior to applying for college (CntDTL), 109 applicants who had less than 1.5 total terms enrolled were left. They reside in the terminal node, which means no further split occurred. Therefore, tracing the above splits, it can be reasonably stated that the 109 applicants who did not register were those who had a very short or no prior attendance at the college, who had a Bachelors degree, or a high school diploma, regardless of their high school
To identify these patterns by studying the decision tree diagram can be cumbersome. Fortunately the data mining algorithm has already combed through the tree diagram and produced a list of the patterns, called rules. There are a total of six rules for those who did not register and another set of 6 rules for those who did. If more variables were introduced to the algorithm, the number of rules would likely increase. The first rule reads as follows:

Rules for No Registration - contains 6 rule(s)
Rule 1 for No Registration (262.0, 0.584)
if Group of Enrollment Status in ["Concurrent" "New" "Returning" ]
and Group of Education Background in ["BA+" "FOREIGN HS" "GED" "HS GRAD" "NO HS"]
and Number of Terms Ever Enrolled <= 1.500
then No Registration
Rule 2 for No Registration (36, 0.528)
if Group of Enrollment Status in ["Concurrent" "New" "Returning" ]
and Group of Education Background in ["BA+" "FOREIGN HS" "GED" "HS GRAD" "NO HS"]
and Number of Terms Ever Enrolled CntDTL > 1.500
and Group of Enrollment Status in ["Concurrent" "Returning" ]
and City Locations in ["Aptos" "Campbell" "Capitola" "Moss Landing" "Mount Hermon" "Santa Cruz" "Soquel" "Watsonville" ]
and Age Range in ["25 - 29" "35 - 39"]
then No Registration
The first rule corresponds to the observations made early for the 109 applicants. Because 153 of those who actually registered were also classified into the terminal node with the 109 applicants, therefore, the rule states at the beginning that there were a total of 262 cases and the confidence is only 58%. It is not unusual that the accuracy for individual rules may not be always high. The fourth rule had 40 cases in it and it had a confidence level of 100%. The rules are listed without a hierarchical order. The first rule is as important as the last rule. Let us take a look at the first rule for those who would register.

Rules for Registration- contains 6 rule(s)
  Rule 1 for Registration (248, 0.867)
    if Group of Enrollment Status in [ "Concurrent" "New" "Returning" ]
      and Group of Education Background in
      [ "BA+" "FOREIGN HS" "GED" "HS GRAD" "NO HS" ]
      and Number of Terms Ever Enrolled > 1.500
      and Group of Enrollment Status in [ "New" ]
      then Registered.

There were a total of 248 cases belonging to this rule and its confidence level was 87%.

It is clear through just viewing a few of the rules that there are many different profiles that could determine the final outcome of registering for classes. This proves that there is no one equation that can adequately illustrate the highly various nature behind the registration status of all the applicants. Case in point, only about 40 students fit the fourth rule for those who did not register. Another 248 of the applicants fit the first rule for those who registered.

The 2949 applicants, a subset of all the applicants for spring 2005, were those who for some reason had not registered just days before the semester was to start, the analysis of their behaviors and background information should bear in mind that real reasons and the motivation factor were unknown. Some of them may be straddlers or stragglers, others may have a legitimate reason to be slow. As the notes taken by the call center staff showed, some applicants were confused by the registration process, some were waiting for their appointment dates, some took a backseat when they realized they had to go through assessment. Quite a few applicants reached by the call center said they were moving out of the area, had financial aid issues, or downright unmotivated and ‘lazy’ quoting their own words.

At the time of this study, academic history data is only available for Groups 1 and 2. When background data becomes available for Group 4, the predictive modeling may be further enhanced.

6 Discussion

In responding to the three research questions, the study uncovered the following findings. In terms of call center effectiveness as defined by yields and significance of the yields (question one), the study found a statistically significant 5% higher
registration rate in a four way post facto control and experimental analysis. But this significance should also take into consideration a few college initiated changes, although no specific impact to any of the four way analyses was immediately perceivable. For example, the college implemented online application in spring 2005, which increased the total applicants by 12% as compared to the spring a year earlier.

The study found a noticeable difference in registration rate large enough to attribute the increase in registration to the call center after examining the different categories of applicants coded for callers. Those who spoke directly with the call center staff had at least a 10% increase in registering for classes compared to those who had no contact with the call center. Yet in answers to question three on productivity (not discussed in this chapter), the increase was found at best to have helped maintain the level of FTES, a productivity measure. However, it can be viewed that without the call center, there could have been a perceivable drop in both headcounts and FTES in spring 2005. From a summative evaluation perspective, the call center receives a B+.

After removing the applicants who interacted with the call center directly, data mining analysis identified six rules for applicants who would register for classes and another six for those who would not. Further, the data mining analysis built a predictive model with an 85% accuracy to predict those who would be less likely to register in a future semester.

References

298  DATA MINING IN E-LEARNING


