**University Enrollment Case Study**

**Scenario:**

In the fall of 2004, the administration of a large private university requested that the Office of Enrollment Management and the Office of Institutional Research work together to identify prospective students who would most likely enroll as new freshmen in the Fall 2005 semester. The administration would particularly like to understand what factors are most closely related to a student’s decision to enroll in the university so that they can implement outreach plans to increase enrollment.

Historically, inquiries numbered about 90,000+ students, and the university enrolled from 2400 to 2800 new freshmen each Fall semester.

The Office of Institutional Research maintains a data warehouse that contains historical information about enrollment. It was decided that inquiries for Fall 2004 would be used to build the model to help shape the Fall 2005 freshman class. The data set **Inq05\_samp** was built over a period of a several months in consultation with Enrollment Management. The data set includes variables that could be classified as demographic, financial, number of correspondences, student interests, and campus visits. Many variables were created using historical data and trends. For example, high school code was replaced by the percentage of inquirers from that high school over the past five years who enrolled (**HSCRAT**). The resulting data set included over 90,000 observations and over 50 variables. For this case study, the number of variables was reduced and a sample was taken. The variables are described in the table below. Some of the variables were automatically rejected based on the number of missing values.

The nominal variables **ACADEMIC\_INTEREST\_1**, **ACADEMIC\_INTEREST\_2**, and **IRSCHOOL** were rejected because they were replaced by the interval variables **INT1RAT**, **INT2RAT**, and **HSCRAT**, respectively. For example, academic interest codes 1 and 2 were replaced by the percentage of inquirers over the past five years who indicated those interest codes and then enrolled. The variable **IRSCHOOL** is the high school code of the student, and it was replaced by the percentage of inquirers from that high school over the last five years who enrolled. The variables **ETHNICITY** and **SEX** were rejected because they cannot be used in admission decisions. Several variables count the various types of contacts the university has with the students.

The suggested roles and measurement levels are shown below as well. You may use any of the variables in the data set for exploratory analysis to gain better understanding of your data. However, you should NOT use variables with a status of REJECTED in your model. As always, you may use any combination of software tools that you choose to analyze the data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Role** | **Measurement Level** |
| ACADEMIC\_INTEREST\_1 | Primary academic interest code | Rejected | NOMINAL |
| ACADEMIC\_INTEREST\_2 | Secondary academic interest code | Rejected | NOMINAL |
| CAMPUS\_VISIT | Campus visit code | Input | NOMINAL |
| CONTACT\_CODE1 | First contact code | Rejected | NOMINAL |
| CONTACT\_DATE1 | First contact date | Rejected | DATE |
| ETHNICITY | Ethnicity | Rejected | NOMINAL |
| ENROLL | 1 = Enrolled F2004, 0 = not enrolled F2004 | Target | NOMINAL (binary) |
| IRSCHOOL | High school code | Rejected | NOMINAL |
| INSTATE | 1 = In-state, 0 = Out of state | Input | NOMINAL |
| LEVEL\_YEAR | Student academic level | Rejected | NOMINAL |
| REFERRAL\_CNTCTS | Referral contact count | Input | INTERVAL |
| SELF\_INIT\_CNTCTS | Self-initiated contact count | Input | INTERVAL |
| SOLICITED\_CNTCTS | Solicited contact count | Input | INTERVAL |
| TERRIOTRY | Recruitment area | Input | NOMINAL |
| TOTAL\_CONTACTS | Total contact count | Input | INTERVAL |
| TRAVEL\_INIT\_CNTCTS | Travel initiated contact count | Input | INTERVAL |
| AVG\_INCOME | Commercial household income estimate | Input | INTERVAL |
| DISTANCE | Distance from university | Input | INTERVAL |
| HSCRAT | 5-year enrollment rate from student’s high school | Input | INTERVAL |
| INIT\_SPAN | Time from first contact to enrollment date | Input | INTERVAL |
| INT1RAT | 5-year primary interest code rate | Input | INTERVAL |
| INT2RAT | 5-year secondary interest code rate | Input | INTERVAL |
| INTEREST | Number of indicated extracurricular interests | Input | INTERVAL |
| MAILQ | Mail qualifying score (1 = very interested) | Input | ORDINAL |
| PREMIERE | 1 =Attended campus recruitment event, 0 = Did not | Input | NOMINAL |
| SATSCORE | SAT (original score) | Rejected | INTERVAL |
| SEX | Gender | Rejected | NOMINAL |
| STUEMAIL | 1 = Has e-mail address, 0 = No e-mail address | Input | NOMINAL |
| TELECQ | Telecounciling qualifying score (1 = very interested) | Rejected | INTERVAL |

**Your Task:**

Use these data to create a model for predicting whether or not a student will enroll in the university. You will need to address missing values. You may use any of the classification modeling methods that we have discussed in class. You should use a validation data set for selecting the final model.

**Your Deliverable:**

Develop an annotated power point slide deck. The slide deck will have two parts:

* **Executive Summary:** The audience for this section is university administration. These are business people who understand the problem at hand but who do not have any formal training in analytics. Your slides should focus on “telling the story” of the data. The slides should include:
  + Introduction to the business problem
  + Overview of the data / data quality issues
  + Exploratory analysis of the data
  + Presentation of the final model
  + Conclusion / recommendations for the university (should include “action items” that the university should consider in order to increase enrollment in the coming years)

This section of the slide deck should highlight the factors that seem to be influencing a student’s decision to enroll in the university and should give a description of the student who is most likely to attend.

Further, the university administration believes that a student’s residency status (in state student vs out of state student) will have a significant impact on their decision to attend the university or not. Your discussion should include a couple of statements regarding the impact of state of residency on the decision to enroll in the university (or not).

**NOTE:** The slides should be clean and easy to read. They should NOT be wordy. Instead, you should include a short paragraph in the notes section of each slide explaining the information that is included on the slide. You can think of this as what you would say if you were presenting the slide to an audience.

* **Technical Appendix:** The audience for this section is technical (e.g., your professor or your peers who are familiar with predictive modeling). The slides should include:
  + Outline of your modeling process
  + Complete output / technical details for the final model that you presented in the Executive Summary section

These slides should provide enough detail (on the slides or in the notes section) for someone to be able to completely replicate your analysis. That said, you don’t need to include all of the details about every model that you create. The detailed output should relate primarily to how you selected the final model and any relevant details associated with the final model.

**Grading:** Your grade will be based on

* Appropriate application of the predictive modeling process and “goodness” of the final model from both a business and statistical perspective (60%)
* Clear, concise slides appropriate for a business audience and accurate discussion of the results (in the notes section of the slides) (20%)
* Clear explanation of how the results can be used to address the business problem at hand (e.g., what recommendations would you make to the university regarding student enrollment) (15%)
* Other factors relating to the overall goodness of your model and appropriate presentation of results. (5%)