CASE STUDY C: Identifying at-risk students at New York Institute of Technology

As with many other universities, New York Institute of Technology (NYIT) has a problem with retention and wished to intervene early with at-risk students. Developing their own model and dashboard with the help of the counselling staff who would be using it to support students, NYIT has been able to identify at-risk students with a high degree of accuracy.

Key takeaway points

- The expertise of counselling staff who support students was deployed to help define the model for identifying at-risk students
- Data on previous students was used to train the model using four different mathematical approaches
- Key risk factors included grades, the major subject and the student's certainty in their choice of major subject, and financial data such as parental contribution to fees
- Dashboards were developed for support staff showing whether each student was predicted to return to their studies the following year, the percentage confidence in that prediction from the model and the reasons for the prediction – this provided a basis for discussion with the student
- Recall of the model is 74%; in other words, approximately three out of every four students who do not return to their studies the following year had been predicted as at-risk by the model. This high recall factor is due to the choice of model as well as the inclusion of a wider range of data than other similar models. Financial and student survey data were included in the model as well as pre-enrolment data.

Rationale

The aim was to increase retention of students in the first year of their studies by creating an at-risk model to identify students most in need of support and to provide information about each student's situation that would assist support counsellors in their work. The model, built using educational data mining, was designed using an 'end-to-end' approach. This involved the entire process from mining the data, through running the analytics and producing the output in a format that was helpful to the counselling staff. There were two reasons for this approach: however powerful the predictive model, it would be useless unless the counselling staff were willing and able to incorporate it into their day-to-day work; and the model needed to work quickly and automatically, without time-consuming manual intervention.

The initial project

The problem definition originated from the users: the counselling staff responsible for supporting students. An external IT solution provider worked with NYIT staff to identify the relevant data. The IT provider gathered and prepared the data; and deployed and evaluated the model. The design process was an iterative cycle, with NYIT counselling staff involved at each stage. The model was built in house at NYIT, so the database, prediction models and front end used by the counsellors were all on the same platform, Microsoft SQL Server.

Data sources and indicators of engagement

There were two versions of the model. Version 1.0 was built entirely in-house at NYIT, simply by gathering data on each student from various sources and combining it in an Excel spreadsheet. The three data sources were: admission application data, registration/placement test data and a survey taken by every student when they did the placement exam. Risk factors were identified by staff based upon their experience and the retention literature. Each risk factor was assigned a score of '1' (increased risk of drop out), or '0' (not a risk), and the score was added up for each student. Risk factors included: incoming grades, major subject and the student's certainty in their choice of major subject.

This version was simplistic in its approach, with equal weighting for all factors. Also, the risk factors were derived from published literature about student behaviour at other institutions, rather than actual student behaviour at NYIT. Furthermore, the Excel spreadsheet was compiled by hand, so it required considerable staff time.

Version 2.0 was designed to address these issues. Technically, there were two important steps: the dataset was built automatically in the NYIT data warehouse (with a new profile created as soon as a new student registered); also the risk classification factors were identified by machine learning models which were trained on NYIT student data. This enabled a more authentic match between the classification of risk factors and actual NYIT student behaviour, and a more nuanced risk analysis with weightings, rather than a simple '1' or '0'.

This version used the same three data sources as Version 1.0, plus financial data (fees required to complete the qualification, whether the student has bursary or other support, whether they have a source of income, parental contribution to fees etc.) The financial data was included because it was known to influence student completion rates, although the relationship between the various financial influences and the risk of attrition was acknowledged to be complicated.

The model was trained on a set of records from previous NYIT students. Attributes included the risk classification for each student. The model was then tested on a different set of records, to see how accurately it could predict risk classifications. A range of mathematical approaches were compared, to see which performed the best in modelling the risk as a function of the other attributes (incoming grades, finance, etc.) Each mathematical approach was tested in different variations. In total, 372 variant models based upon four mathematical approaches were used.

Dashboards and interventions

The dashboard was designed with and for the student support staff. It is a simple table, with one line for each student, showing whether they are predicted to return to their studies the following year, the percentage confidence in that prediction from the model and, importantly, the reasons for the prediction. The reasons may include: a disparity between fees for the rest of the qualification and the student's funds, the student is uncertain about their career goal, or they are working a large number of hours per week as well as studying. The counsellor then has the basis for a conversation with each student about their situation and future plans.

Findings and outcomes

The models were tested for precision and *recall*, which measures the match between the actual and predicted student behaviour:

- The higher the recall percentage, the better the model is at identifying students who are at risk of leaving. If the recall percentage is low, that means that there was a large proportion of students who actually left, but were *not* identified as at risk by the model.
- The higher the precision percentage, the lower the proportion of 'false alarm' predictions made by the model. If the precision percentage is low, then the model is predicting a lot of students to leave, when in fact they continue their studies.

The best version of the model, which was an ensemble combining several mathematical approaches, had 74% recall and 55% precision. This compares very favourably with a similar model developed independently at Western Kentucky University, which only had a 30% recall. The WKU model was based only upon pre-enrolment data. The enhanced recall of the NYIT model is due to the inclusion of financial and student survey data and the type of model used. There are multiple factors in the model, although the following are some of the factors that impact upon the 'at risk' classification: full-time or part-time student, number of working hours per week, where the student has a completion plan for their qualification.

In practice, recall matters because the purpose of the model is to identify students at risk so support staff can intervene. With a recall of 74%, for every four students not returning to study the following year, three of those students will have been predicted as at risk by the model correctly. Precision also matters, because 'false alarms' may impact upon staff resources.

References

Agnihotri, L., Ott, A. (2014). Building a Student At-Risk Model: An End-to-End Perspective From User to Data Scientist. *Proceedings of the 7th International Conference on Educational Data Mining.*