Predicting Enrolments in University Units

George Gemayel
Department of Computing
Macquarie University
Sydney, Australia
30629438@students.mq.edu.au

Abstract

In this paper we present a prediction model capable of forecasting undergraduate unit enrolments at the Faculty of Science, Macquarie University. With limited data, we were able to define a model for 86% of 200 and 300 level units in this faculty. For semester 1 2009, our model was able to predict 43% of these defined units within 4 enrolment counts, and 57% of these defined units within 10 enrolment counts. Through the analysis of related work, we examine the many external and internal factors that may affect unit enrolments, and their feasibility in our model. We discuss how we were able to simplify the model and restrict its variables to internal institutional data only, allowing it to be effective and self correcting.

1 Introduction

The need for flexible programs of study at Universities has complicated the dependencies and hierarchies between units. What motivates a student to enrol into a unit may depend on a number of factors, predominantly:

- If it is a core unit
- If it is a pre-requisite unit
- If it is a popular unit

The first and the second examples are relatively simple to analyse since we can track where students are coming from and where they will go to next. The last example however is dependent on many factors both internal and external to the university.

Study patterns also have an effect on unit enrolments. For example, some students choose to study full time while others study part time. Some students may choose night time classes while the majority prefer daytime classes. Some core units are also offered in both semesters of an academic year for those who have failed or those who have an unorthodox program of study. These students are harder to track.

To add to the complexity of unit dependencies, students are not limited to units in the department they are enrolled in. For example, a student enrolled in the Division of Science with an IT undergraduate degree can legitimately enrol into 100 level law units.

Postgraduate units at Macquarie University are less complex in their structure. There are no unit pre-requisites and so unit enrolments are purely dependent on the following factors:

- It is a core unit
- It is a popular unit
- The units offering does not clash

Forecasting unit enrolments is therefore a complicated exercise. Ideally, the model would take into consideration external factors such as the economy, the unemployment rate, tertiary study trends and internal movements of students within degrees. However the aim of any good prediction model is to keep it simple. Therefore in this paper, we aim to address notable factors of unit enrolments and explore their purpose, if any, in our prediction model.

The rest of the paper is structured as follows. In Section 2 we will discuss the need to predict unit enrolments and the requirements of Macquarie University. Section 3 will provide a summary of related work in the area of prediction, which will be used to determine the necessity and importance of each predictor. Section 4 will report on the data used in a form of a data audit. Section 5 will explore generic forecasting techniques and how best to approach a prediction model. Section 6 will begin to describe in detail

the model, which includes clustering the units, the use of dependency maps and comparison techniques. In Section 7 we report the results of the model, and finally in Section 8 we draw some conclusions from these results and outline any further work.

2 The Problem

Each year, Divisions at Macquarie University are required to predict unit enrolments for that current year. The predictions are required to prepare timetables, to allocate lecture rooms, and to allocate other resources. It can also be used to determine expected gross income for that division that year.

Currently there are no concrete methods to predict unit enrolments. Generally the result is a best guess based on the number of enrolments in that unit the previous year, which doesn't take into effect drops in university enrolments, failure rate of a unit, or students who have completed pre-requisites.

The goal of this project is therefore three parts. Firstly it must be determined wether a prediction model is feasible, given the complexity of the unit system and the amount of data that can be obtained. Secondly, a best fit model must be applied, which is flexible enough to account for the various factors and not too loose that it is not reliable. Thirdly, if a model can be proven to work, a system should then eventuate that implements this model and can use current data to forecast unit enrolments across the Department of Science and possibly the rest of Macquarie University.

3 Related Work

There are no published works on predicting university unit enrolments; however there are a few papers on forecasting university enrolment figures. These projects are significant as they expose the various external factors that may determine changes in university course enrolments and hence changes in university unit enrolments. Other interesting aspects of these models are their approach and technique towards forecasting and the final outcome.

3.1 Data, Parameters and Variables

Many of the models assessed defined similar dependent variables. These included the population count of the city, the strength of the economy and the value of education. Demographic variables of students were also commonly used such as age, gender and financial situation.

O'Heron (1997) performed an in-depth study on forecasting undergraduate enrolment figures. His paper focused on two distinct parts of enrolment. The first analysis tackled the factors that influenced university enrolments, by looking at demands for a higher education, primarily for full time students at the typical age of enrolment (18 to 21 years old). The second analysis looked at the institution and movement of students between different parts of the post secondary system. He found that predictions based solely on population figures are usually wrong; hence his analysis included demographic trends, population rates, economic cycles, social values, institutional fees and policy changes

3.1.1 Participation Rates

O'Heron (1997) had found it difficult to establish any solid relationship between population and enrolment.

Anderson (2006) describes and compares traditional enrolment forecasting models. His data sets included existing university population by retention rates; graduating and newly enrolled students by ethnicity; and ratios of non-resident and graduate students to undergraduate students. He also compared two trends; one where the population rate steadily increased (aggressive analysis) and the other the population rate remained constant (trend analysis).

Anderson (2006) found that the trend analysis yielded 50% more accurate results. This result complements O'Heron's findings that changes in population rates are not strong predictors of enrolment figures, and better results are achieved when the population rates is considered a constant.

3.1.2 Economic Cycles

According to O'Heron (1997), growth in the economy generally had a negative impact on university enrolments. The reverse is also true; enrolment growth was observed when the economy was at a stall. This can be attributed the wider range of choices a prospecting student may have when the economy is improving. These observations are conclusive; however including economic evolution in a prediction model is not feasible given the nature of the data that can be obtained.

3.1.3 Financial Standing

Lim et al. (2008) developed a predictive model of student enrolments which had a heavy reliance on Siefert and Galloways (2006) model of students' financial 'tipping point'. The aim was to determine how likely a prospecting student was to enrol based on the amount of awards. To calculate a student's financial 'tipping point', the model required financial data and admission data of the institution. Using logistic regression, the model was at best able to predict 20.2% of enrolled students based purely on financial controls. The low figure was attributed to the quality of the data available and not to the low significance financial standing may have in enrolment projections.

O'Heron (1997) also theorised that the cost of attending university seemed to have little impact on participation.

3.1.4 Institutional Data

The University Of Central Florida (UCF) developed a prediction model for course enrolments which did not include any external factors. The approach they employed was to limit their data to internal university figures only. That is, no external sources were considered. The model builds student headcount by starting with the returning students, based on the previous two years. Also the undergraduates are estimated using cohort retention from previous years.

By limiting the control variables to just the last two years for returning students, and the last decade for new students, the model was found to almost self-adjust when external factors take effect.

This model requires revision every few years, however at the moment; it is successfully predicting head count accurately within 0.5% for a one year prediction, and 2% for a five year prediction.

4 Data Audit

Enrolment figures from 2001 through to 2009 were provided for the Division of Science, which encompasses units in computing, mathematics, electronics, physics and information systems. The files provided were in varying formats, with unit counts having to be joined on unit name and semester offered.

Various units are no longer offered and so these were ignored. Other units, specifically those that began in or after 2006, are too recent to have significant historical data and these too were dropped. Finally, other units had too small a number to extrapolate any significant findings. These units had less than 100 enrolments in total for the previous six years. These units may be revisited after the results have been validated; however the numbers are too small to train the model.

The trends of unit enrolments across the 8 years showed a strong decrease, with most computing units having 2008 enrolment figures a quarter of what they were in 2003. This was found to be general trend in the sector coupled with the many degree changes involving these units. It was also found that actuarial studies had deferred 100 level computing units to the second year, which made their enrolment trends inconsistent with other 100 level units.

A handful of other prominent and core computing units were also found to be approaching zero. Comp225 for example had only 49 enrolments in 2008, down from 410 in 2003. This was attributed to the change in programs of study which now offers alternatives to Comp225. Also this drop was attributed to the extra 200 level units introduced with the ISYS stream, which, as expected, cannibalised some of these students.

Other inconsistent trends that required closer examination belonged to the 100 level ISYS units. Defying the general decreasing trend, enrolment figures for these units increased in 2005 and 2006, which is widely due to the number of Business students taking up these units.

5 Prediction Methods

Forecasting is the ability to predict the future by examining trends and dependencies. Prediction models should be kept simple, and only a handful of significant variables should be used. The most common types of prediction models fall into the following categories:

- Regression
- Decision Tree
- Neural Network

Regression analysis is the ability to extrapolate numerical information based on response variables, and so this was used to predict unit enrolments.

With all types of prediction models, there are three data sets that should be used. The *training data* is used to calculate any correlations and build the model. The *validation data* is used to validate the model to ensure the correct variables

and constants were used. Lastly, the *test data* which is used to test the success rate of the model. The test data is not compulsory since the model is already validated using the *validation data*.

Given the limited amount of data available to build this model, all historical figures, up to and including 2008, were assigned as the training data, with 2009 semester 1 figures being defined as the validation data.

6 The Solution

6.1 Comparison techniques

Unit enrolment figures varied considerably and consistent trends were difficult to determine. It was found that there is no direct correlation between unit figures of any unit and its pre-requisites. For example, Comp330 may have Comp225 as a pre-requisite; however its enrolment figures do not emulate that of Comp225. This finding conveys the difficulty in following student study trends from one unit to the next.

It was found however that by looking at the change in unit enrolments as a percentage, solid patterns could be found. As an example, Comp226 enrolment figures have decreased since 2003, however the rate of change was not constant; in 2005 it decreased by 33% and in 2006 it decreased by 40%. Comp229 enrolment figures are slightly higher than that of Comp226, however in 2005 it decreased by 22% and in 2006 it decreased by 25%. A clear correlation was found when the changes in enrolments were graphed, where Comp226 trends following that of Comp249 by a factor of 1.6. That is, if Comp249 figures decreases by 20%, as it did in 2007, then Comp226 figures are expected to drop by 41%.

Let us define x_{09} as the enrolment figure for unit x in year 2009. Let us also define x_d, the change in enrolment figure for unit x as:

$$x_d = (x_{09} - x_{08})/x_{08}$$

Let us also introduce unit y, where y_d is dependent on x_d by a factor of k (in our Comp249 and Comp226 case, k was equal to 1.6). Therefore we have:

$$y_d = k.x_d$$
$$(y_{09} - y_{08})/y_{08} = k.(x_{09}-x_{08})/x_{08}$$
$$y_{09}/y_{08} = k.(x_{09} - x_{08})/x_{08} + 1$$
$$y_{09} = k.y_{08}.(x_{09} - x_{08})/x_{08} + y_{08}$$

The final equation above allows us to predict the final enrolment figures of unit y in 2009 as a function of figure x_{08}, x_{09}, y_{08} and of course the constant k. Enrolment figures for x_{08} and y_{08} are known. The figure for x_{09} is to be determined from other units using the same technique or by using course enrolment figures for 100 level units. This is explained in section 6.2. Hence the role of this model is to define dependencies between units and calculate the constant k.

To simplify this task, each unit was compared to units at the same level (e.g. 300 level) to find a core subject that dictates the rise and drop in all the other units. This technique worked for the majority of units, with other unmatched units having too many outliers, or too few data points to extrapolate anything relevant.

What we have at this point is what we aimed for; a model dependent solely on internal institutional figures. This approach reduces the models susceptibility to external factors; for example, a rise in Comp225 will predict a rise in Comp248, without having to take into consideration why the rise happened.

Another benefit of comparing deltas (x_d) of enrolment figures is that the accuracy of the model is constantly being corrected. i.e., an error in prediction for one year is not carried through to the prediction the following year, as the new deltas are applied to the actual figures of the preceding year.

6.2 Clustering of Units

Our modelling techniques can be applied to the majority of units, however not each unit should be represented the same. Three segments were created in this model, each representing a different method and approach for prediction.

- 100 level unit
- Core units
- Elective units.

100 level units cannot be treated the same way as 200 or 300 level units when determining dependencies and the constant k. This is because most 100 level units don't rely on pre-requisites. 100 level units have the most diverse students in regards to which university division the students are enrolled in. Also they are more dependent on enrolments into programs of study, and so these units are more susceptible to external factors.

However this does not mean that trends of some 100 level units do not correlate with other

100 level units; for example, Math135 semester 1 is a predictor for Math136 semester 2. For those 100 level units that have no dependencies, these must be predicted using course enrolment figures. An example is Comp 125 enrolments may increase or decrease as enrolments in BCompSc increase or decrease. Unfortunately, no data was able to be sourced to follow through with this method, and so predictions for 100 level units remain unresolved.

As for core units, these were preferred as predictor units wherever possible, mainly due to the high number of enrolments figures which made calculating x_d more precise and significant.

Elective units tend to follow the trends of core units; however changes in enrolments of these units are more susceptible to popularity and offering of the unit. Hence, these units were found to have some data points completely against the norm for a random year, while the rest of the years followed the general trends of the core units. These were singled out as there predictions are more likely to deviate from the actuals.

6.3 Dependency map

A dependency map was created to demonstrate the dependencies between the units, as well as illustrate the order that predictions should be made. For example, Comp226 is predicted by Comp249, and in turn Comp247 is predicted by Comp226. The dependency map for undergraduate computing units is described in figure 6-1.

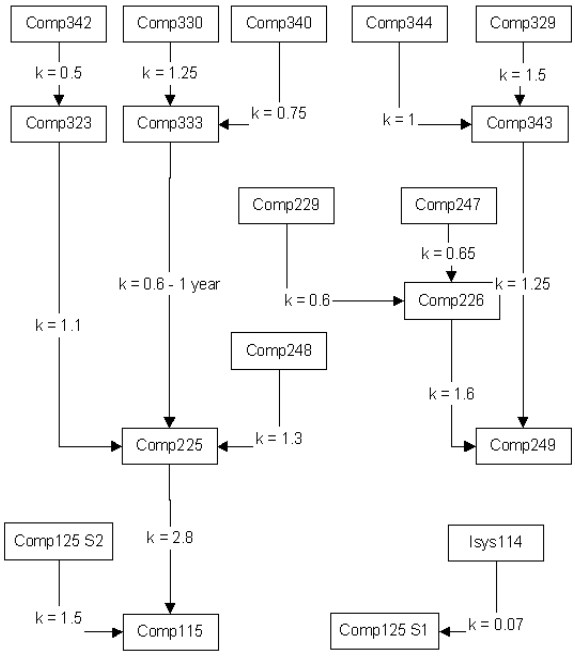

Figure 6-1: Dependency Map for COMP

This map also gives the opportunity to roll up the dependencies, with the aim of simplifying the number of connections. The result is that all units in a division are dependent on the same predictor In our example, if Comp247 is dependent on Comp226, and Comp226 is dependent on Comp249, then Comp247 is dependent on Comp229.

From our equation above, where y is dependent on x, we introduce z, which is dependent on y with a constant j. Therefore we have the two prediction equations:

$$y_{09} = k.y_{08}.(x_{09} - x_{08})/x_{08} + y_{08}$$
$$z_{09} = j.z_{08}.(y_{09} - y_{08})/y_{08} + z_{08}$$

We use the following substitution of y_d to remove the dependency between z and y:

$$(y_{09} - y_{08})/y_{08} = k.(x_{09}-x_{08})/x_{08}$$

Therefore we now have:

$$z_{09} = j.k.z_{08}.(x_{09}-x_{08})/x_{08} + z_{08}$$

Making z completely dependent on x with a constant ($j.k$).

It must be noted that a unit in the second half of the year can depend on units offered in the first half of the year. Likewise, a unit in the first half of the year can depend on a pre-requisite which is offered in the second half of the previous year. For these cases, delta calculations must be shifted back one year. For example, Phys301 semester 1 is predicted by Phys202 semester 2 of the previous year. Therefore we have:

$$y_d = k.x_{d-1}$$

7 Results

Predictions were made across computing, physics and mathematic units using 2009 enrolment figures as at the 23rd of March 2009. Since course enrolment figures were not available to predict 100 level units, the actual figures of the base units were used to kick off the prediction process. The aim was to verify that our prediction method is feasible for those units we do predict.

Predictions, actuals and deviations are shown in table 7-1 for all computing subjects where a dependent unit and a constant k were identified.

Unit	Predicted	Actual	Deviation
COMP125	187 (57%)	216 (82%)	-29
COMP225	101 (106%)	108 (120%)	-7
COMP229	83 (24%)	81 (21%)	2
COMP247	170 (26%)	168 (24%)	2
COMP226	35 (40%)	38 (52%)	-3
COMP323	53 (61%)	15 (-55%)	38
COMP330	26 (-7%)	23 (-18%)	3
COMP342	17 (31%)	13 (0%)	4
COMP343	43 (30%)	24 (-27%)	19
COMP340	5 (0%)	13 (160%)	-8
COMP329	33 (43%)	19 (-17%)	14
COMP333	21 (-5%)	23 (5%)	-2
COMP344	65 (30%)	41 (-18%)	24
ISYS114	241 (3%)	280 (20%)	-39

Table 7-1: Predictions vs. Actuals 2009

As table 7-1 illustrates, there are several units that were predicted correctly within an acceptable degree of error; which in this case is a handful of students.

Comp229 is an excellent example of how our prediction model is not dependent on external factors that affect enrolment figures. Since 2005 Comp229 enrolment figures have dropped consistently, i.e. 20% in 2005, 24% in 2006, 28% in 2007 and 22% in 2008. If our model was dependent on time trends, then it would have forecasted another drop in enrolment. Instead, our model has forecasted an increase of 24% in 2009 to raise the enrolment count to 83 which was only 2 counts away from actual (which was 81).

The decrease and subsequent increase in Comp229 figures could be attributed to many factors, including a general trend in university enrolments; however this does not concern us. What is important to the model is that we have successfully linked Comp229 enrolment changes to that of Comp226, which rose 52%, and in doing so was able to not only accurately depict if the figure would rise or fall, but also by how much.

Another interesting unit is Comp333 since it depends on Comp225 figures of the previous year. The prediction was off by two counts.

Other predictions, specifically for units Comp125, Comp323 and Comp343 were not so accurate. For Comp125, it was only offered in the evenings in 2009, which may have affected the results. For Comp323 and Comp343, the dependencies and constant k appear to be chosen incorrectly. The training data consisted of declining figures throughout the past five years, and now that enrolments are on the rise, it is expected that dependencies on predictors may vary.

8 Conclusion

We have presented a model capable of defining dependencies between undergraduate units, and with these dependencies, we were able to show how trends in one unit can predict enrolment figures for other units in the same faculty. We defined the model free from external sources, and only relied on institutional data, hence simplifying the model, the variables and the data required.

The model is limited by its data, so that we were not able to predict most 100 level units based on course enrolment figures, however, we were able to define dependent variables for 86% of 200 and 300 level units.

In the department of Computing, we were able to predict 43% of units within an error of 4 students and 57% of units within an error of 10 students.

As for the other inaccurate predictions, the model is expected to improve pre semester as more data is collected. Hence further work is necessary in two areas. Firstly more work is required to better define those units with inaccurate predictions. Secondly course enrolments must be taken into consideration for 100 level units, which in turn, will improve the model definitions for all undergraduate units. These two issues must be addressed before any system that implements the model be attempted.

References

Anderson, D., 2006. Enrolment prediction techniques. [PowerPoint slides]. Arizona Board of Regents.

Lim, H., Davies, D., Jackson, S., 2008. 'Hark who goes there?': Developing a predictive model of student enrolment. [PowerPoint slides]. Southampton Solent University.

O'Heron, H., 1997. Undergraduate enrolment forecasts: A tricky science. Research File, 2(1), pp.1-15

University Of Central Florida,. Detailed enrolment prediction modelling method. [Online]. Available at: http://www.uaps.ucf.edu/enrollment/methods_detailed.html [Accessed 21 March 2009]