Predicting the Uptake of a University's Offers of Places

Stephen Peplow, PhD, Kwantlen Polytechnic University

Author's Contact Information

Stephen Peplow, PhD,
Kwantlen Polytechnic University
12666 – 72nd Avenue, Surrey, BC, V3W 2M8
Email: Stephen.Peplow@kpu.ca

Abstract:

Institutions such as Universities spend considerable resources in recruiting and following up on applicants. Unfortunately, much wastage results from the sending out of follow-up letters to students who never arrive, and who perhaps had applied only as a form of insurance; and also in hiring instructors and scheduling classes when the attendance is uncertain. In a competitive market, a predictive model of the uptake of offers made might well be helpful.

Key Words:

Recruiting, offers of place, predictive model, students, faculty.

Introduction

This paper sets out to develop a predictive model for one particular institution, but the techniques and the modeling process could be applied elsewhere with adaptations. In fact, the techniques could be used in any environment in which customers are free to make multiple applications.

As well as reducing the more immediate wastage problem, a predictive model based on historical information would be helpful in diverting attention towards the most plausible reasons for the rejection of an offer, highly useful information at a time when the values that students place on higher education are changing. As Maringe points out in an interesting but limited study of student motivations in the United Kingdom, students are becoming increasingly 'consumerist' in their approach (Maringe, 2006).

This paper provides an example of the construction of a predictive and analytical model. Free open-source software is used in both the GIS (QGIS) and statistical analysis ('R') parts of the paper (R Core Team, 2013) . There are therefore no software costs in implementation. The accuracy of the model is reasonably high at nearly eighty per cent.

Data and methods

The data has been kindly provided by Kwantlen Polytechnic University in British Columbia, Canada. The dataset identifies students who were accepted for the academic year 2011 and also whether or not they took up the offer. Each applicant's record is georeferenced by postcode. There are 8,899 unique postcodes in the dataset, the majority of which (seventy-three percent) contain information on only one applicant. The total number of student records is 12,968. The acceptance or refusal of the offer provides the binary dependent variable to be used in the analyses. The dataset also provides information such as age; faculty applied for; ratecode (international or domestic) and other pertinent details. Of the nearly thirteen thousand students in the dataset, just over twenty-two per cent declined their offers.

The dataset provides the names of the Faculties to which the student applied and also the age-group of the student.

Faculty Name	Code
Academic and Career Advancement	1
Arts	2
Business	3
Community and Health	4
Design	5
Non-credential	6
Science and horticulture	7
Trade and Technology	8

Table 1. Coding of Faculty names

Agegroups have been coded as follows:

Agegroup	Code
18	1
19-22	2
23-28	3
29-32	4
33-38	5
39-44	6
45-50	7
51-55	8
56-60	9
60 +	10
Below 18	11

Table 2. Coding of agegroups

Statistical methods

The task is to 'explain' the binary dependent variable in terms of the other independent variables. I have approached the task in two ways; by using the classification tree, and by logistic regression. These methods have their own strengths and weaknesses, and a combination of the two yields deeper insights (Long, 1997). This is not an instructional document, and so I have not explained the statistical theory behind the approaches in depth. Instead I have provided suitable references and would be pleased to enter into correspondence.

The classification tree

The classification tree method dates from the 1980s (Breiman, 1993). An algorithm partitions the data using splits or nodes. At each possible split, the algorithm decides whether the node contributes useful information about the dependent variable. If it does then it is defined as a split. The method has been used in a wide range of disciplines, for example in mental health care to calculate suicide risks , while in oncology, Camp and Slattery use the classification tree to identify types of cancer (Camp & Slattery, 2002). In remote sensing, the tree has been used to assess ground cover (Davranche, Lefebvre, & Poulin, 2010). One advantage of the classification tree approach is that the nodes are ordered in decreasing statistical significance. This means that the node which contributes the most information comes first.

Logistic regression

Logistic regression is a well-established method of calculating the odds ratio of the occurrence of one of the two alternatives in a binary dependent variable. An odds ratio is defined as the probability of an event (p) divided by (1-p). An odds ratio of 1 thus means that the event is as likely to occur as not to occur. In betting parlance, this is 'evens'. It is common practice to present results from the analysis as the natural logarithm of the odds ratio. This is to obviate problems with zeroes and negative numbers. Logistic regression is used in a wide range of disciplines, but naturally is especially popular in disciplines in which binary dependent variables are common. This includes marketing and in finance. For example, Yeung and Yee use the tool to predict customer propensity to purchase (Yeung & Yee, 2011). Restaurant bankruptcies have been predicted by logistic regression (Youn & Gu, 2010).

The tests

From the university's perspective, interesting questions might be:

1. Does the acceptance rate change between faculties? Are some faculties more successful than others in retaining the students who have applied to them?
2. Is the age of the applicant related to the acceptance rate?
3. Are international students more or less likely than domestic students to accept their offers?

4. Does the home location of the applicant affect his or her acceptance rate? This is tied to a supplementary question regarding applications to multiple institutions as an 'insurance policy'.

Tests 1 - 3 can be answered by the classification tree and logistic regression. Test 4 will use data gleaned from GIS.

Results of Tests 1 -3

Using classification tree

The plot below provides a classification tree using the 'rpart' algorithm with the splits in order of statistical significance.

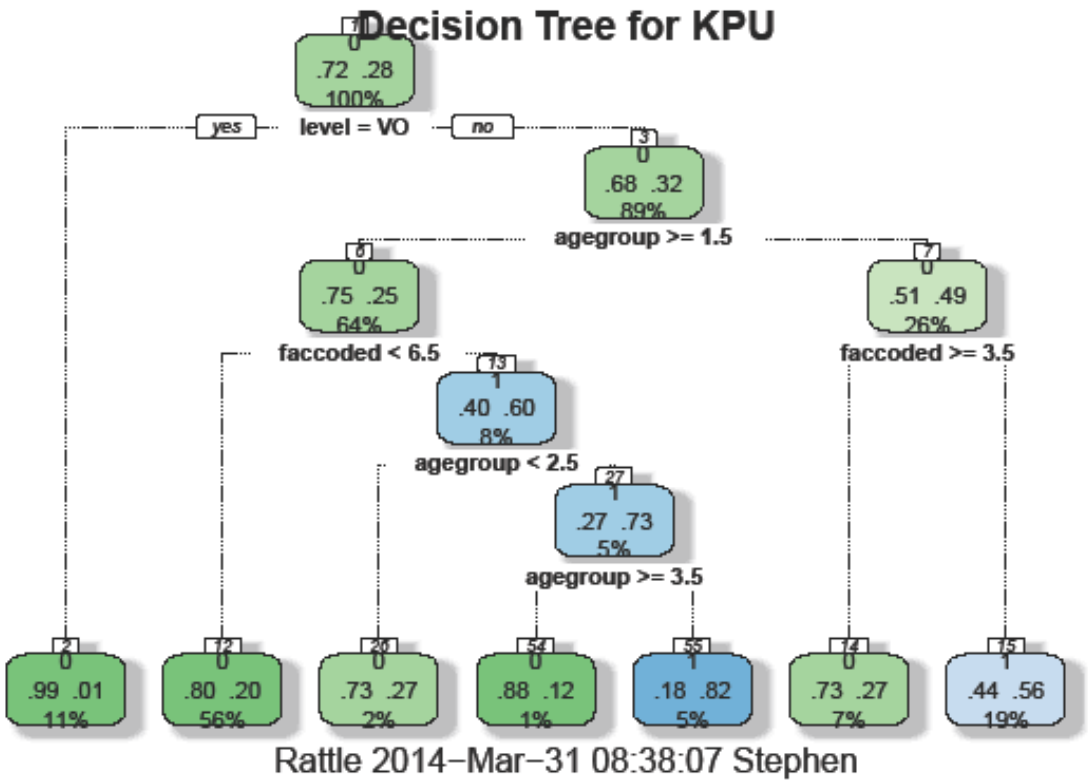

Figure 1. Classification tree output

The data has been split into training and testing sets. The most important split is at node 1, concerning whether the applicant was seeking vocational or undergraduate study. If YES, then the probability of accepting the offer (0) was 99%. If he or she was seeking undergraduate training, then the next most important node concerning age group. The algorithm has identified a split at age group >= 1.5, which, referring back to the table, meaning that the applicant was aged over 18. The rest of the tree is straightforward. The overall error is 23%, meaning that 77% of the classification were correct.

The model may be used for prediction in bulk. A large dataset of applicants detailed could be applied to the model, and the predicted probability of offer acceptance for each applicant produced. This output could then be ranked and appropriate action taken.

Using logistic regression

I have repeated the analysis using the same variables. Given the coding of zero for accept and 1 for reject, the meaning of the coefficients is this: the more positive (or less negative) the coefficient, the more likely the student is to reject the offer and vice-versa.

	Refuse
Arts	-0.094
Business	-0.520**
Community and Health Studies	-2.749***
Design	-14.677
Non-credential students (Academic)	0.065
Science and Horticulture	0.005
Trades and Technology	-3.297***
INTERNATIONAL	-0.367**
age18	0.666**
age19 - 22	-0.073
age23 - 28	-0.294
age29 - 32	-0.330
age33 - 38	-0.220
age39 - 44	0.010
age45 - 50	0.273
age51 - 55	-0.004
age56 - 60	0.322
age60+	-0.347
Constant	-0.954**
N	12,968
Log Likelihood	-6,289.654
AIC	12,617.310
*p < .05; **p < .01; ***p < .001	

Table 3. Logistic regression output

The asterisks after the coefficients indicate the statistical significance of the variable, as shown in the table below the results. There are only seven faculties listed above, while Table 1 provides eight. This is because we need one faculty to be the reference level. The missing faculty is academic and career advancement. The coefficients for the seven displayed faculties should be considered in reference to Academic and Career Advancement, which had a forty-one per cent drop rate.

It is immediately apparent that the probability of a student failing to take up an offer from the Faculty of Design is extremely small, and in fact no student dropped an offer.

LevelVO refers to Vocational or Undergraduate, with Undergraduate being the reference level. We already know that Vocational students rarely drop offers, and so the negative sign is expected. The probability of a vocational student rejecting an offer is

much lower than that of the reference level, undergraduate students. The age variable uses age < 18 as the reference level. Age (18) shows high statistical significance and also a positive coefficient. This means that eighteen year olds are the group most at risk of failing to take up offers. The size of the negative coefficient decreases with age, perhaps reflecting greater stability and decision-making maturity. The exception is the coefficient for the age 23-28 group. Perhaps students of this age are active in the job market and turn down the offer because they have obtained a job? I have no other plausible explanation for this change, but it might be worthy of further research. A section below explores this issue a little further.

Ratecode is a dummy variable, splitting the applicants into domestic and international students, with the international students paying more and perhaps having more choice. The reference level for this variable is domestic, simply because there were many more of them in the dataset. The negative sign shows that international students are less at risk for rejecting offers compared to domestic students. In fact, the difference is quite stark: twenty-three per cent of domestic students failed to take up offers made, while the figure for international students was fourteen per cent.

GIS

The statistical analysis above can be complemented by insights from geographical information systems (GIS). We can use GIS in three ways.

1. to gain a visual impression of the geographical distribution of the offer take-up

2. to gauge the effect of competition. KPU students almost certainly apply for other institutions apart from KPU as a form of insurance. It is interesting to gauge the effect of competing offers. The dataset does not provide data on alternative offers of course, but we can estimate it by constructing 'buffers' around both KPU and competing institutions and observing whether the takeup rate differs. We can also estimate the effect of having to cross a bridge or travel great distances.

Visualisation

It may be helpful to visualise the geographical distribution of the acceptance of offers. As with the statistical analyses above, I have assigned a zero to a student who accepted an offer, and a one to a student who did not take up the offer. I have selected only those postcodes which contained one applicant. Figure 1 below shows the distribution. Yellow marks those who took up offers, red those who received an offer but who did not take it up.

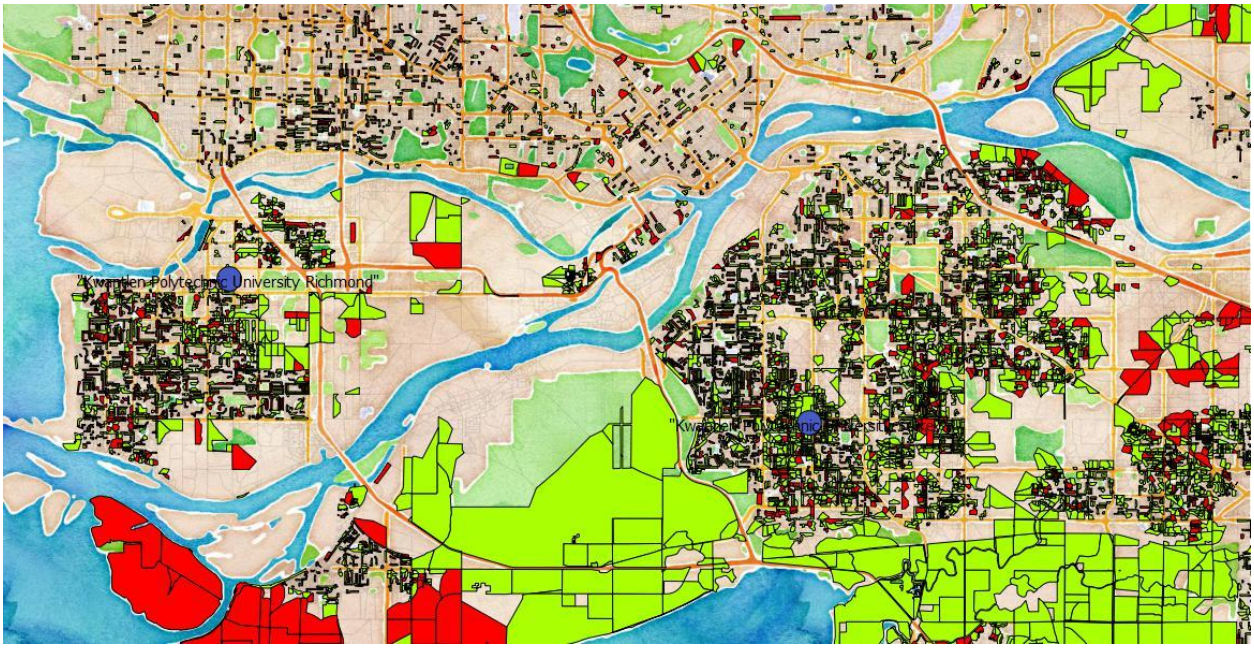

Figure 2. Map of offer takeup (light green) and refusal (red).
KPU campuses indicated with blue circles.

It is noticeable that acceptances (the light green colour) are clustered around the campuses of KPU.

Buffers

We can examine this further by placing a buffer around KPU campuses and also those of institutions which might be considered 'competing'. I chose a 2 km radius buffer, but this was an arbitrary choice. Table 3 below shows the numbers and also the odds of a student rejecting an offer. The column headings are: the names of post-secondary institutions likely to be attractive to KPU applicants. 'Drop' and 'total offers' are the number of applicants who did not take up offers and the total KPU offers made within a 2 km radius of the institution. Pdrop is the probability of a drop, shown as an odds ratio in the next column.

Institution	Drop	Total_Offers	Pdrop	2km odds
BCIT	16	27	0.593	1.455
Langara	31	108	0.287	0.403
Langley	14	99	0.141	0.165
Richmond	53	245	0.216	0.276
SFU	0	3	0.000	0.000
Surrey	13	142	0.092	0.101
TWU	4	12	0.333	0.500
UBC	1	11	0.091	0.100

Table 3. Odds within 2km buffers of competing institutions

I also calculated the odds using a 1 km buffer but only for KPU's three campuses: Richmond, Surrey and Langley. Data was insufficient at the other campuses. From the

2 km figures, it appears that BCIT is KPU's greatest competitor, which is hardly surprising since both institutions are similar. The overall rejection rate is 22 per cent, or an odds of rejection of 0.28. Since the odds of rejection are lower within the 2 km buffer, it is possible that proximity to a campus increases the take up rate. Again this is not surprising; students who live near a campus are perhaps more likely to apply to only that institution. It is interesting that the 1 km figures confirm the proximity finding; the odds are lower. Langley is particularly small, perhaps because students there have few nearby alternatives.

Effect of age

Above I noted that the sign for age in the logistic regression changed. The plot below in Figure 3 uses the age group as the explanatory variable, with the probability of rejection of KPU's offer on the vertical y axis.

This is interesting because the plot begins and ends with a highly defined probability, because the line is tightly focussed. However the focus diminishes in the middle age ranges, reinforcing the logistic regression result.

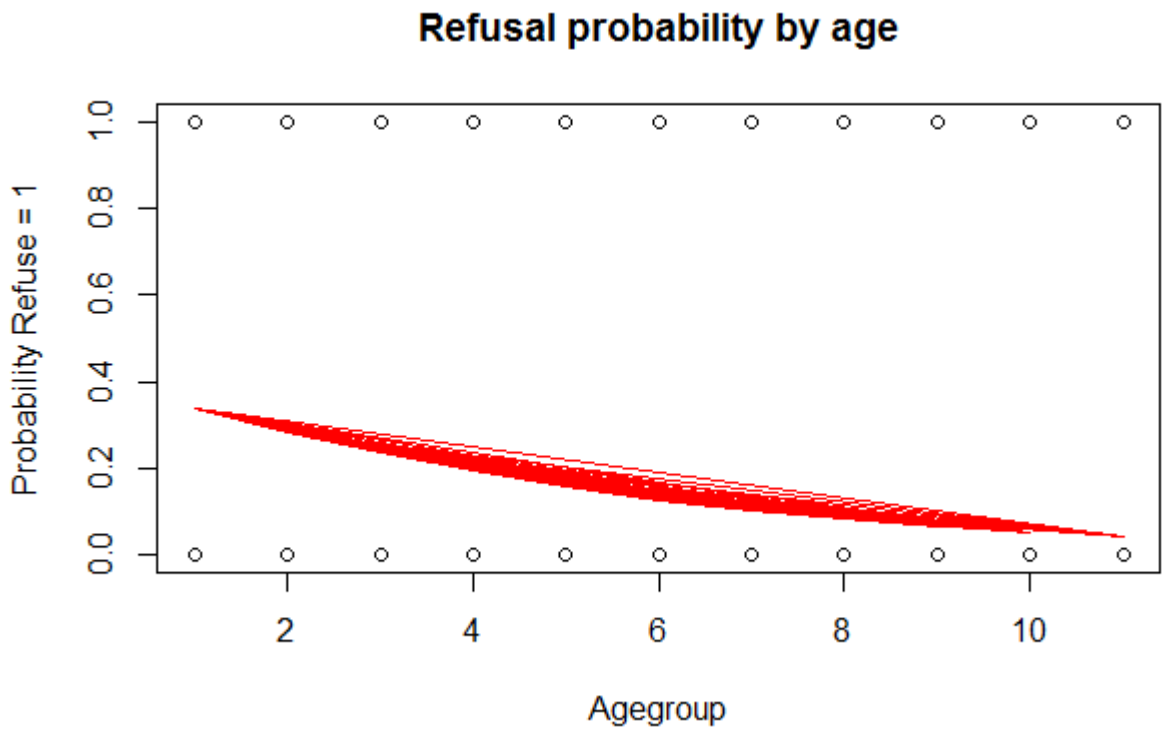

Figure 3. Refusal probability by agegroup

Prediction example

The logistic regression model above yields insights into the areas of concern in recruitment. It is also possible to use the model to predict the probability of refusal or acceptance either for an individual or a whole group. As an example, I predict the probability of refusal for this applicant:

Age = 18; level = UG; ratecode = DOMESTIC; faculty = Arts. The response is 0.4009791, meaning that there is a forty per cent change of refusal. It is possible to feed in a large dataset of applicants and receive probabilities for each one. The applicants could then be ranked by probability and appropriate action taken.

Operationalization

The information discussed and presented here is a beginning only. However, the following insights drawn from the analysis could be acted upon.

Triage system. KPU's Office of International Analysis and Planning prepared an internal report in March 2013, the Acceptance/Declined Survey. In the survey, five per cent of respondents claimed that other institutions communicated more quickly than KPU, and this was an important reason for their decision not to take up the offer. Perhaps the frontline staff in the registrar's office could use the full classification tree to prioritize work. Of course, all students should be attended to promptly, and no doubt are, but students whose applications fall into 'risky' nodes might warrant extra care.

The logistic regression has highlighted particular areas of concern. By faculty, Academic and Career Advancement, Non-credential, and Science and Horticulture students deserve attention at the institutional level; why are so many students rejecting offers from these two faculties? Age groups also provide interesting questions and opportunities. If scholarships are to be offered, they could be targeted to the youngest age-group; the effect of a scholarship on mature students is likely to be negligible and would therefore be money unwisely spent.

Further work

I would like to combine raw responses to the KPU Acceptance/Decline Survey with geospatial and observed offer take-up. From this, we might be able to observe patterns which match students' responses to the Survey. For example, did those who claimed that lack of public transport was a major factor live inconveniently distant from a KPU campus? Correspondence and principal component analysis would also be possible. In addition, investigating more recent data might yield some interesting intertemporal insights.

Conclusion

The analyses which I have performed above are very elementary, and yet have brought out some interesting insights. In addition, the dataset is limited and in particular lacks intertemporal data. We cannot therefore examine trends over time, perhaps the most interesting feature of dynamic student enrolments. Data analyses such as mine are the bread and butter of modern business, and organisations which fail to build data capture and data analysis into their regular routines are likely to fall behind. In contrast, as some interesting recent studies have found, even modest data applications can propel organisations forward

References

Breiman, L. (1993). *Classification and regression trees*. New York: Chapman & Hall.

Camp, N. J., & Slattery, M. L. (2002). Classification tree analysis: a statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes & Control*, *13*(9), 813–823

Davranche, A., Lefebvre, G., & Poulin, B. (2010). Wetland monitoring using classification trees and SPOT-5 seasonal time series. *Remote Sensing of Environment*, *114*(3), 552–562. doi:10.1016/j.rse.2009.10.009

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE.

Maringe, F. (2006). University and course choice: Implications for positioning, recruitment and marketing. *International Journal of Educational Management*, *20*(6), 466–479.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Yeung, R. M. W., & Yee, W. M. S. (2011). Logistic Regression: An advancement of predicting consumer purchase propensity. *Marketing Review*, *11*(1), 71–81.

Youn, H., & Gu, Z. (2010). Predict US restaurant firm failures: The artificial neural network model versus logistic regression model. *Tourism & Hospitality Research*, *10*(3), 171–187.