Statistical Modelling in Enrolment Management: A Higher Education Case Study

Humphrey Brydon, Sarel Steel, Dineo Mahlangu, Jessica Maloma, and Renette Blignaut

Abstract

Enrolment management is important to institutions of higher learning. Administrators at these institutions are annually faced by the question: how many offers for a given academic programme should be made to applicants to meet the registration target set by the authorities?

Data on past and new applicants are available at most institutions. In this paper, data from the Faculty of Natural Sciences at the University of the Western Cape are used to develop a statistical model that provides estimates of the likelihood of new applicants accepting registration offers from the Faculty. The paper therefore contributes to the important field of strategic enrolment management. The paper shows how a statistical model estimated from historical data can assist administrators to determine the number of offers that should be extended to applicants to reach a given registration target.

Key words: admission points, decision tree, higher education, logistic regression, registration targets

ABOUT THE AUTHORS: HUMPHREY BRYDON (EMAIL: HBRYDON@UWC.AC.ZA), SAREL STEEL, DINEO MAHLANGU, JESSICA MALOMA, AND RENETTE BLIGNAUT: UNIVERSITY OF THE WESTERN CAPE, SOUTH AFRICA

Brydon, H., Steel, S., Mahlangu, D., Maloma, J., & Blignaut, R. (2025). Statistical Modelling in Enrolment Management: A Higher Education Case Study. *International Journal of African Higher Education, 11*(2), 1-18. https://doi.org/10.6017/ijahe.v11i2.17815

Résumé: La gestion des inscriptions est importante pour les établissements d'enseignement supérieur. Chaque année, les administrateurs de ces établissements sont confrontés à la question suivante : combien d'offres pour un programme académique donné doivent être proposées aux candidats pour atteindre l'objectif d'inscription fixé par les autorités ? La plupart des établissements disposent de données historiques sur les candidats et de données relatives aux nouveaux candidats. Dans cet article, les données de la Faculté des sciences naturelles de l'Université du Cap Occidental sont utilisées pour développer un modèle statistique qui fournit des estimations de la probabilité que les nouveaux candidats acceptent les offres d'inscription de la Faculté. Cet article contribue donc au domaine important de la gestion stratégique des inscriptions. Il montre comment un modèle statistique estimé à partir de données historiques peut aider les administrateurs à déterminer le nombre d'offres à faire aux candidats pour atteindre un objectif d'inscription donné.

Mots clés: points d'admission, arbre de décision, enseignement supérieur, régression logistique, objectifs d'inscription

Introduction

Universities and other institutions of higher learning annually face the challenge of identifying prospective students who should be offered registration opportunities. The available facilities place an upper limit on the number of first-time entry students who can be admitted. Financial considerations on the other hand encourage the admission of larger numbers of students. The dilemma is exacerbated by throughput considerations, requiring admitted students to be successful and graduate in the shortest possible time.

There are several sources of uncertainty in this scenario which should be taken into account in an admission/enrolment strategy: applicants who are offered the opportunity to register may decline the invitation; those who do accept an invitation may fail to register; those who do register may not graduate in the minimum time, or at all. Accurate and reliable data on current applicants, and historical data on the profiles of successful students, can be used to minimise the uncertainty by using advanced data analysis methods. This paper describes such an approach, developed at the University of the Western Cape for applicants to the Faculty of Natural Sciences. More specifically, historical data is used to develop a statistical model that can be used to identify first-time applicants with a high likelihood of accepting a registration offer. These form the subset

of applicants who should be approached first by the Faculty with offers of registration. Using such a model streamlines the admission process, which is important given the tight timelines at the start of an academic year. A conclusion from the research is that the output from judicious application of statistical (machine-learning based) algorithms to relevant data can be a valuable aid to decision makers in enrolment management. "Aid" is a key word in this statement: the intention must not be to replace the human enrolment manager with an artificial agent.

Several research questions are addressed in the paper. Firstly, what are the important variables when historical data available at the Faculty of Natural Sciences of the University of the Western Cape are used to develop a statistical model for predicting the response of new applicants to registration offers? Following on this, the second question concerns the specific statistical model that should be used for this purpose, keeping in mind the conflicting requirements of model accuracy and interpretability. Finally, the third question deals with the value that can be derived from including the predictions from the model into the enrolment strategy at the institution.

Section 2 of the paper contains a survey of selected papers on data-based support in the enrolment process. This is followed in Section 3 by a summary of the data that were analysed in this paper, and a description of the steps required to clean and prepare the data for analysis. A non-technical description of the method that was used is provided in Section 4, and the results are presented and discussed in Section 5. The paper closes with conclusions and suggestions for further research in Section

Literature Survey

Papers on the application of statistical algorithms in the enrolment problem have been regularly published. The focus in Basu et al. (2019) is similar to the focus in this paper: which qualifying applicants should be offered an opportunity to register? This question is framed as a binary classification problem: an applicant offered a registration opportunity can either accept or decline the offer, where the latter response includes cases where an applicant receiving an offer does not respond. The authors analyse a dataset consisting of 11001 cases, each having 35 variables, and compare the performance of seven techniques for solving the classification problem: logistic regression, a naïve Bayes approach, decision trees, support vector machines, nearest neighbours, random forests and gradient boosting. The metrics used in comparing the different methods are accuracy, precision,

recall, score, area under the receiver operator characteristic (ROC) curve and the Matthews correlation coefficient (MCC). According to Basu et al. (2019), these metrics address the imbalance property of the data: only a relatively small proportion of applicants offered a registration opportunity accept such an offer (this proportion is known as the yield rate). Definitions and a discussion of the suitability of these and other metrics appear in Section 4.

Basu et al. (2019) find that, overall, logistic regression performs best, albeit by small margins. Although not mentioned by these authors, logistic regression, in addition to its good performance in terms of the evaluation metrics, has the added advantage of producing output that is easy to interpret. This is important from the perspective of explaining to an applicant who did not receive a registration offer the reasons for such a decision. The authors do, however, discuss the importance of the input variables and conclude that the five most important ones are GPA score, campus visit indicator, high school class size, reader academic rating, and gender.

Although the focus in Lofti and Maki (2018) is on graduate applications and admissions, it contains several insights that are valuable in the present context. The importance of an enrolment strategy incorporating an effective predictive model is emphasised, and it can be argued that this is even more important in first-time entry enrolment strategies. In their thorough literature survey, Lofti and Maki (2018) highlight several earlier contributions. Thomas et al. (2001) investigate the problem of identifying the applicants that will be most impacted by recruitment efforts. They develop a logistic regression model for predicting enrolment based on four groups of variables, viz. demographic, academic, geographic and behavioural. It should be noted that their study did not include any financial aid variables. Thomas et al. (2001) find that targeting students with predicted enrolment probabilities between 0.3 and 0.6 leads to an increase in the yield rate.

Lofti and Maki (2018) focus on developing a model that can be used to predict registration of applicants after admission to a post-graduate study programme. They use data on approximately 40 variables to develop a decision tree model for this purpose. The authors argue that decision trees hold several advantages over logistic regression: a decision tree can model non-linear relationships, is easy to interpret, and is able to deal with categorical inputs with large numbers of categories. However, Hastie et al. (2009, p. 310), cast doubt on the last point. Lofti and Maki

(2018) also find that the most important predictor variables are financial aid related variables, geographic location, academic variables and age. A further interesting significant variable turned out to be the number of days following application until admission.

Langston et al. (2016) discuss strategic enrolment management at tertiary institutions. They emphasise the importance of accurate enrolment prediction in this process, stating that efficient enrolment management is a combination of accurate enrolment forecasts and informed judgment on the part of enrolment managers. Regarding specific modelling approaches, they present an extensive discussion of trend analysis, with logistic regression also being mentioned. One of their findings is that prediction per subpopulation (academic program) yields more accurate results, provided the sample sizes are not too small.

Goenner and Pauls (2006) consider enrolment forecasting based on inquiry data, i.e., data obtained from inquiries made to the institution by possible applicants. This is more challenging than using applicant data. The authors propose a Bayesian model averaging approach, in which a posterior weighted average of different candidate logistic regression models is determined. One of the interesting points arising from their investigation is that distance from the institution seems to have a non-linear effect on the likelihood of a student enrolling.

A slightly different angle is explored by Mountford-Zimdars and Moore (2020), who discuss the use of contextual data in enrolment forecasting. The authors discuss the use of contextual data in an admission strategy with the aim of redress: contextual data place the school marks of a student within the context of social environment. Although this is a relevant aspect, a disadvantage is that the study is qualitative, based on results from interviews conducted with relevant individuals.

Basu et al. (2022) use random forests to address two versions of the enrolment forecasting problem. The first version considers a three-class classification problem: a student receiving an offer from the institution can either accept the offer and register, or accept the offer but not register, or reject the offer. In the second version, a hierarchical approach is employed. In the first stage one of two categories is predicted for an applicant: accept or reject an offer that was made. In the second stage only those who accepted the offer are considered and split into two sub-categories, viz. register or do not register. The authors refer to these latter cases as applicants who "melt away".

Soltys et al. (2021) present a detailed proposal for enrolment forecasting that implements XGBoost on an Amazon Web Services (AWS) platform. Snippets of the actual Python code that was used are provided, making the paper valuable to someone intending to duplicate the analysis in a different context. Another contribution of the present paper is the fact that an applicant can be placed in one of three categories: those who will most probably register (and therefore do not need recruitment effort), those who will most probably not register (and for whom recruitment effort will most probably be wasted), and those falling in-between, who are the ones who should be targeted during recruitment.

The focus in this paper is specifically on predicting how an applicant qualifying for a programme at a university will react to a registration offer. Consequently, the model proposed in the paper operates at an individual applicant level. There are many papers dealing with the broader problem of predicting enrolment numbers from historical data, and the related problem of predicting student success. Two examples in an African context are Satope (2014), dealing with university enrolment in Nigerian universities, and Nyenya and Rupande (2015), discussing Zimbabwean institutions. However, these and other similar contributions fall outside the scope of the present paper.

Data Summary

Historical data are available on students applying for admission to an academic program in the Faculty of Natural Sciences at the University of the Western Cape. The data go back to 2022 and were used to develop the statistical model. Details on the data and the steps followed to clean it are now provided. Ethical clearance for this project was granted by the University of the Western Cape (HS23/4/17).

The data used in the development of the statistical model contained 84,424 observations (prior to data cleaning and preparation) in total. Data for the years 2022 (20,969 observations) and 2023 (28,890 observations) were used as training data and data for 2024 (34,565 observations) were used as validation data. Additionally, the number of applicants included in the study was reduced in consideration of certain inclusion criteria as discussed below.

The variables included in an initial decision tree that was fit to the data were APS (Admission Points System) score, age, school quintile, race,

gender and area (urban or rural). A binary variable indicating whether an applicant accepted an offer or not was used as the target.

For the purposes of this study, only those applicants with available National Senior Certificate (NSC) results were included in the study. The reason for this lies in the admission criteria of the University which involves the conversion of NSC results, using a university points-based system, to an APS score. This APS score, obtained from final matric results, is then used to determine whether an applicant meets the minimum requirements for different programmes.

It is worth mentioning at this point that applicants do not apply for the extended curriculum programmes (ECP) within the Faculty of Natural Sciences. Applicants apply for one of the 10 offered programmes and are cascaded into the respective ECP programmes (Programme 1 does not offer an extended programme) during the application review process. In total there are 19 programmes offered within the Faculty of Natural Sciences. The proportions of applicants for the 10 mainstream programmes, for the years 2022 - 2024, are given in Table 1.

Table 1: Distribution of Applications per Mainstream Programme

Programme	Proportion of applicants (%)
1	10.09
2	6.38
3	5.59
4	10.78
5	10.02
6	17.84
7	6.65
8	12.82
9	12.68
10	7.14

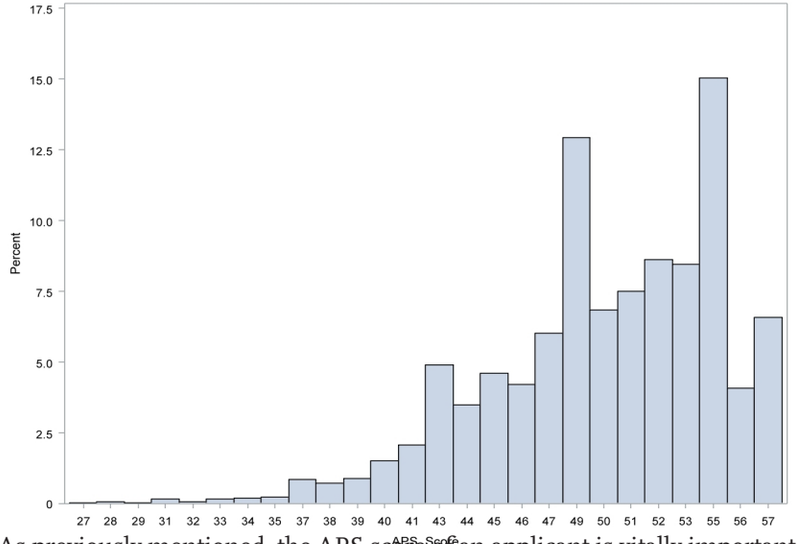
As shown in Table 1. four of the 10 programmes offered in the Faculty receive less than 10% of applications, with Programme 6 receiving a substantial proportion of applications. As discussed later in Section 5 in conjunction with registrations targets, these proportions of applications received for each programme further emphasise the need for a statistical

model to determine which applicants should receive registrations offers. The proportions of applicant race groups and gender are presented in Table 2. It can be clearly seen that African applicants make up the vast majority of applications to the University of the Western Cape, followed by coloured applicants. From Table 2, it is evident that there are slightly more female applicants to the Faculty of Natural Sciences than male applicants.

Table 2: Race and Gender Groups

Race/Gender	Female	Male	Other
African	43.12%	39.75%	0.10%
Asian	0.16%	0.13%	0%
Coloured	5.00%	4.19%	0%
Indian	2.20%	1.86%	0%
White	1.99%	1.25%	0.02%
Other	0.08%	0.11%	0%

Figure 1: APS Distribution for Training Data



As previously mentioned, the APS score of an applicant is vitally important in the decision as to whether or not to extend an offer to the applicant. As shown in Figure 1, the APS scores (with mean and median of 49.760 and 51, respectively) are slightly negatively skewed (skewness of -0.759) with a kurtosis value of 0.446. Another important point to note here is

that applicants with an APS score below 27 were excluded from the model development as these scores do not meet minimum requirements for degree admission to the University.

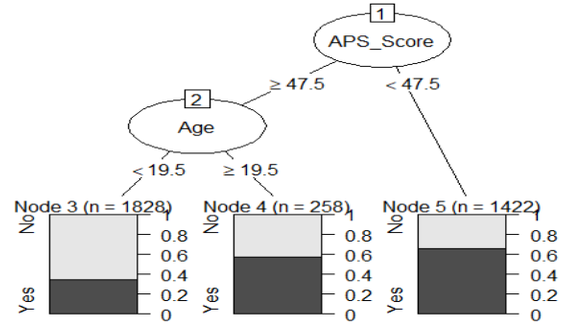
The target variable for this study was a binary variable, where the values 0 and 1 reflected whether an applicant declined or accepted an offer, respectively. Of importance in this study is the attachment of a probability to these binary indicators, thereby assisting those involved in enrolment in the decision of whether to extend an offer to an applicant.

Method

The first research question stated in Section 1 of the paper requires using the data described in the previous section to develop a model for predicting the likelihood that an applicant to the Faculty will accept a registration opportunity. From a statistical perspective, this can be viewed as a binary classification problem. It is customary in binary classification problems to denote the response variable by Y, with $Y = 0$ signifying that an applicant who received a registration offer declines the offer, and $Y = 1$ signifying acceptance of the offer. A statistical model for such a scenario can be used to estimate the probability of the event $Y = 1$ for a given applicant.

During a first exploratory step in the analysis a classification tree was fit to all the data, resulting in the tree shown in Figure 2.

Figure 2: Decision Tree fit to the Data



It is seen that the first split is on the APS score of a student, with $APS < 47.5$ and APS ≥ 47.5 describing the two branches. For the APS ≥ 47.5 branch a further split occurred at and . The tree shows that for those students with APS score below 48, approximately 65% accepted an offer that was extended by the University. For those students with , the older students

(aged 20 or more) tended to accept an offer more readily than younger ones (19 or younger). Almost 60% of these APS score students older than 19 accepted, while only approximately 35% of those 19 or younger did so. The numbers of students in the different nodes are shown in Figure 2.

The above results suggest that APS score and Age are the two most important input variables when predicting whether an offer extended to a student is accepted or not. As a second step in the analysis, a logistic regression model was fit to the reduced dataset containing only the response, APS score and Age. For this second step of the analysis, a separate logistic regression model was developed for each of the 10 mainstream programmes. As an example, the fitted model for Programme 3 is given below (the model results for other programmes are included in the Appendix to this paper, including the training and validation observation counts):

$$f(x) = -5.41 - 0.01 * APS + 0.59 * Age - 0.01 * (APS * Age)$$

Interestingly, the model contains a negative coefficient for APS. It therefore seems that the likelihood of accepting an offer decreases with an increase in the APS score. This result was also observed in other fitted models and is not surprising: students with higher APS scores would likely apply to many institutions and only accept an offer from their more preferred institution.

The proposed logistic regression model can be used in different ways to estimate the number of offers to be made to a cohort of N qualifying applicants. One of these is now described. Denote the number of vacancies (i.e., registration target) that must be filled by M (a known quantity), and the estimated number of offers that must be extended by T (to be determined). For each applicant in the cohort an estimate is obtained regarding the probability that an offer which was extended to the applicant will be accepted. These estimated probabilities are denoted by:

$$pi = P(\text{student } i \text{ accepts the offer}) \text{ for } i = 1,2, \ldots\ldots, N.$$

Now consider the small artificial example in Table 3, where $M = 5$ vacancies have to be filled. Scenario A shows 10 applicants with their probabilities of accepting a registration offer in the second row. These probabilities are obtained by applying the logistic regression model to the attributes of the applicants. In Scenario A, the probabilities are all assumed to be equal to 0.5. This implies that for every two extended offers one can, on average, expect one applicant accepting. The third row in Scenario A contains the

cumulative probabilities. This shows that we can expect to fill the $M = 5$ vacancies if we extend $T = 10$ offers (the first point at which the cumulative probabilities reach the target of 5).

It is highly unlikely that all the applicants will have the same predicted probability of accepting a registration offer. Scenario B shows a more realistic case. The same argument as in Scenario A shows that now only $T = 9$ offers need to be extended, since the cumulative probability reaches the value 5 at the ninth applicant.

Scenario C considers the same probabilities as in Scenario B, but now the applicants have been reordered in descending order of probability. If offers are extended to the applicants in this order, the cumulative probabilities show that we can expect offers to be sufficient to reach the target.

Table 3: Determining the Number of Offers from Predicted Probabilities

Scenario A										
Applicant	1	2	3	4	5	6	7	8	9	10
P(accept offer)	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Accumulated	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Scenario B										
Applicant	1	2	3	4	5	6	7	8	9	10
P(accept offer)	0.2	0.5	0.4	0.9	0.8	0.5	0.6	0.3	0.8	0.9
Accumulated	0.2	0.7	1.1	2.0	2.8	3.3	3.9	4.2	5.0	5.9
Scenario C										
Applicant	4	10	5	9	7	2	6	3	8	1
P(accept offer)	0.9	0.9	0.8	0.8	0.6	0.5	0.5	0.4	0.3	0.2
Accumulated	0.9	1.8	2.6	3.4	4.0	4.5	5.0	5.4	5.7	5.9

Formulated generally, the predicted probabilities are accumulated until the first point at which their sum exceeds the target, i.e., we take T = minimum$\{n: \sum_{i=1}^{n} p_i \geq M\}$.

In practice, if we determine T in this way and it turns out that $\sum_{i=1}^{n} p_i < M$, then the cohort seems to be too small to fill all the vacancies and an offer is extended to every student, i.e., we take $T = N$.
A decision must also be made as to the order in which the students should be approached with offers. Table 3 illustrates two options. In Scenarios A and B the applicants are offered registration opportunities without taking

the predicted acceptance probabilities into account (i.e., offers are made based on applicant ordering). In Scenario B, this leads to nine offers having to be extended. Scenario C shows another possibility: approach the students in decreasing order of the predicted probabilities. This makes sense since the applicants who are, according to the model, more likely to accept an offer are approached first. In Scenario C, this leads to a reduction in the predicted number of offers that have to be extended, with only seven instead of nine offers required. The disadvantage of this approach is that it does not necessarily prioritise the students who have the highest APS scores, i.e., those most likely to be successful in their studies. This is a consequence of the negative relationship between APS score and likelihood of accepting an offer. A further option is therefore to approach the students in order of decreasing APS score. This will, however, most likely lead to a larger value of T than that produced by the previous method.

Suppose a given classifier is applied to the cases in a dataset. What metrics should be used to evaluate its performance? The answer to this question should take into account the imbalance in enrolment data: usually only a small proportion of students accept an invitation to enrol. We therefore focus on evaluation metrics suitable for imbalanced data. It is well known that accuracy is an unsuitable metric in imbalanced cases, and that recall (the proportion of students accepting an offer who are predicted to do so), precision (the proportion of students predicted to accept who actually do accept) and the F_1 - score (the harmonic mean of precision and recall) are better options. In addition, the Matthews correlation coefficient has desirable properties as a classification metric, also in imbalanced scenarios (see for example Chicco et al., 2021 for a definition and a discussion of its properties).

Result

Model Result

In this section we discuss the logistic regression model results for each of the mainstream programmes. Table 4 shows the values of six classification metrics obtained from the validation data for each of the ten programmes.

Although there is considerable variation in the values of the metrics amongst the programmes, the overall performance is somewhat disappointing.

Table 4: Final Logistic Regression Model Metrics

Programme	Precision	Recall	F1 Score	MCC	AUE	Misclassification Rate
Programme 1	0.314	0.745	0.442	0.004	0.524	0.458
Programme 2	0.397	0.568	0.467	0.003	0.668	0.332
Programme 3	0.250	0.618	0.356	0.004	0.686	0.295
Programme 4	0.783	0.714	0.747	0.006	0.671	0.401
Programme 5	0.272	0.524	0.358	0.002	0.697	0.218
Programme 6	0.546	0.717	0.620	0.003	0.738	0.207
Programme 7	0.489	0.579	0.530	0.002	0.694	0.374
Programme 8	0.099	0.571	0.168	0.004	0.679	0.206
Programme 9	0.412	0.615	0.494	0.002	0.663	0.246
Programme 10	0.105	0.300	0.156	-0.001	0.731	0.227

An important reason for this is the imbalanced nature of the training data, with the number of positive responses being a small minority. A direction for future research is to implement an approach for addressing this imbalance, thereby potentially improving the performance of the model – see for example Chen et al. (2024) for a recent discussion.

For the purposes of this study, it is worth noting the precision metric (i.e., probability of correctly predicting an acceptance). All models, apart from that for Programme 4, produced quite low values for this metric. The opposite is true for the recall metric, where models appeared to be able to predict non-acceptance more accurately than acceptance (except for the 4th programme listed where prediction appeared to remain stable).

The ability of the models to produce better predictions for the non-acceptances is somewhat confirmed by the F1-Score as all models produced values that are fairly high (i.e., indicative of better false positive/negative predictions). This trend is further confirmed with the misclassification rate, where no model produced a rate less than 20%, with Programme 1 producing a high rate of 45.8%.

The MCC and AUC metrics appear to suggest the same performance from the models (i.e., a "coin toss decision"). All MCC values are quite close to zero, suggesting that the decision to extend an offer to an applicant and that the applicant will accept the offer, is a coin toss. In terms of the AUC metric, all models, besides two, produced values less than 0.700, with

Programme 1 producing a value of 0.524 (with 0.500 considered to be a "coin toss decision").

Offer Estimates

Table 5 contains a summary of the estimated number of offers to make using the approach described in Section 4. The table shows additional information as well that would be available to the enrolment manager at the time of deciding how many offers to extend.

Table 5: Final Offer Estimates for 2024

Programme	Total Applicants	Registration Target	Number of Offers to Make
Programme 1	360	80	279
Programme 2	241	60	234
Programme 3	410	70	311
Programme 4	207	60	All applicants
Programme 5	377	55	275
Programme 6	716	85	536
Programme 7	211	50	199
Programme 8	504	40	200
Programme 9	564	70	304
Programme 10	282	40	192
Programme 2 - Extended	N/A	20	All remaining
Programme 3 – Extended	N/A	30	80
Programme 4 – Extended	N/A	20	N/A - all applicants made mainstream offer
Programme 5 – Extended	N/A	20	All remaining
Programme 6 – Extended	N/A	50	155
Programme 7 – Extended	N/A	20	All remaining
Programme 8 – Extended	N/A	15	90
Programme 9 – Extended	N/A	20	69
Programme 10 - Extended	N/A	20	76

The second and third columns in Table 5 contain the number of applicants and the registration target in 2024 for each programme, respectively.

As was previously mentioned, applicants do not apply for an extended programme but are cascaded into the equivalent extended programme. For the methods that have been put forward, if an applicant was not made an offer for the mainstream programme, they were included in the applicant list for the extended programme and subsequently either made or not made an offer for the extended programme.

The estimated number of offers to be extended to applicants appear to be in line with the previous year's offers, except for programmes 2, 4, 5 and 7.. For these programmes, either all remaining applicants needed to be made an offer (with no guarantee that these were the correct number of applicants needed for the counting methods), or there were no applicants remaining in the pool for the extended programme (i.e., all applicants were made an offer for the mainstream programme).

The above does occur in practice and it is in these types of scenarios that enrolment managers rely on ad hoc measures of collating additional applicants (e.g., in agreement with an applicant, they could be moved to a programme where applications are needed). A more common occurrence is where late applications for programmes would be accepted.

Conclusions and Recommendations

The discussion in this section is structured around the research questions stated in the introduction.

What are the important variables determining the response of an applicant to a registration offer? The analysis of the Faculty of Natural Sciences data at the University of the Western Cape showed that two variables are particularly important: the APS score of the applicant, which gives an indication of academic ability, and age. Other variables in the data, such as school quintile, race, gender and area (urban or rural), did not distinguish well between positive and negative responses to registration offers. An important conclusion is that there are significant interaction between APS score and age: in the high APS score group the older applicants more readily accepted registration offers than the younger ones.

The second research question relates to the specific statistical model that should be used for predicting the response of new applicants to registration offers. Logistic regression was used in this study. The main advantage of logistic regression is its easy interpretability. This is important, since it

should be possible to explain to applicants not receiving registration offers why this is the case. Also, similar published studies confirm that the

performance of logistic regression compares well with that of other more complex machine learning approaches – see for example Basu et al. (2019).

The third research question refers to the value added by using the proposed model as part of an enrolment strategy. This question can only be answered fully after the model has actually been incorporated by the Faculty. The results from applying the model to historical data suggest that using the model will serve to improve the enrolment process.

This paper deals with an important and complex problem, and the following is a selection from the many possibilities for further research:

(i) How can the performance of the logistic regression model be improved by using a suitable method to address the imbalanced nature of the training data?

(ii) The proposed logistic regression model provides point predictions of the responses of applicants to registration offers. How can measures of confidence in the predictions, for example confidence intervals, be computed?

(iii) The example in Section 4 explained how the predicted probabilities obtained from the model can be used to estimate the number of applicants who should be approached with registration offers. Are there other more effective approaches?

(iv) On a more practical level, a conclusion from the study is that applicants with high APS scores are less likely to accept registration offers, reflecting the availability of offers from other institutions. How can this group be targeted to increase their registration rate?

References

Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive models of student college commitment decisions using machine learning. *Data, 4(2), 65.*

Basu, T., Buckmire, R., & Tweneboah, O. (2022). An application of machine learning to college admissions: The summer melt problem.

Journal of Machine Learning for Modelling and Computing, 3(4), 93 - 117.

Chen, W., Yang, K., Yu, Z., Shi, Y. & Chen, C.L.P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10759-6, 57 - 137.*

Chicco, D., Warrens, M., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access, 3, 3083 - 3084.*

Goenner, C., & Pauls, K. (2006). A predictive model of inquiry to enrolment. *Research in Higher Education, 47, 935 - 956.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Springer.*

Langston, R., Wyant, R., & Scheid, J. (2016). Strategic enrolment management for chief enrolment officers: Practical use of statistical and mathematical data in forecasting first year and transfer college enrolment. *Strategic Enrolment Management Quarterly, 4, 74 - 89.*

Lotfi, V., & Maki, B. (2018). A predictive model for graduate application to enrolment. *Open Access Library Journal, 5.*

Mountford-Zimdars, A., & Moore, J. (2020). Identifying merit and potential beyond grades: Opportunities and challenges in using contextual data in undergraduate admissions at nine highly selective English universities. *Oxford Review of Education, 46, 752 - 769.*

Nyenya, T. & Rupande, G. (2015). Enrollment management in open and distance learning institutions: A case study of the Zimbabwe Open University. *International Journal of Research in Humanities and Social Studies,* 2(2), 26 - 36.

Satope, B.F. (2014). Determinants of enrolment in Nigerian universities. *Economics World,* 2(4), 238 - 251.

Soltys, M., Dang, H., Reilly, G., & Soltys, K. (2021). Enrolment predictions with machine learning. *Strategic Enrolment Management Quarterly, 9(2), 9 - 18.*

Thomas, E.H., Reznik, G.L., & Dawes, W. (1999). Using predictive modeling to target student recruitment: Theory and practice. AIR 1999 Annual Forum Paper.

Appendix

Table 6: Fitted Models for Mainstream Programms

Programme	Training (N)	Validation (N)	Coefficient values			
			Intercept	APS	Age	APS*Age
Programme 1	263	360	11.6460	-0.3569	-0.6402	0.0191
Programme 2	153	241	-58.5793	1.1956	3.4754	-0.0718
Programme 3	256	410	-5.4088	-0.0145	0.5943	-0.0065
Programme 4	138	207	30.8757	-0.7138	-1.1637	0.0277
Programme 5	242	377	79.1506	-1.7948	-3.8356	0.0867
Programme 6	386	716	-17.9661	0.2236	1.3001	-0.0199
Programme 7	200	211	-9.9018	0.1928	0.8103	-0.0168
Programme 8	288	504	-13.7073	0.1999	0.8490	-0.0142
Programme 9	219	564	3.1810	-0.1050	0.0475	0.0001
Programme 10	159	282	-45.2553	0.9725	2.6528	-0.0576