Predicting Student Enrollment Based on Student and College Characteristics

Ahmad Slim
University of New Mexico
Albuquerque, NM 87131, USA
ahslim@unm.edu

Don Hush
University of New Mexico
Albuquerque, NM 87131, USA
dhush@unm.edu

Tushar Ojah
University of New Mexico
Albuquerque, NM 87131, USA
tushar3309@unm.edu

Terry Babbitt*
University of New Mexico
Albuquerque, NM 87131, USA
tbabbitt@unm.edu

ABSTRACT

Colleges are increasingly interested in identifying the factors that maximize their enrollment. These factors allow enrollment management administrators to identify the applicants who have higher tendency to enroll at their institutions and accordingly to better allocate their money rewards (i.e., scholarship and financial aid). In this paper we identify factors that affect the likelihood of enrolling. We use machine learning methods to statistically analyze the enrollment predictability of such factors. In particular, we use logistic regression (LR), support vector machines (SVMs) and semi-supervised probability methods. The LR and the SVMs methods predict the enrollment of applicants at an individual level whereas the semi-supervised probability method does that at a cohort level. We validate our methods using real data for applicants admitted to the university of New Mexico (UNM). The results show that a small set of factors related to student and college characteristics are highly correlated to the applicant decision of enrollment. This outcome is supported by the relatively high prediction accuracy of the proposed methods.

Keywords

Student enrollment, student characteristics, college characteristics, classification, logistic regression, support vector machines, time series analysis, variable selection

1. INTRODUCTION

In the past years enrollment management emerged as an important structure in academic institutions [3]. Its direct influence on the performance of such institutions made it a cornerstone. Don Hossler, John P. Bean, and colleagues defined enrollment management as "an organizational concept and a systematic set of activities designed to enable educational institutions to exert more influence over their student enrollments. Organized by strategic planning and supported by institutional research, enrollment management activities concern student college choice, transition to college, student attrition and retention, and student outcomes. These processes are studied to guide institutional practices in the areas of new student recruitment and financial aid, student support services, curriculum development, and other academic areas that affect enrollments, student persistence, and student outcomes from college" [5].

A direct consequence of this process is the major involvement of enrollment management in budgeting and financial aid planning. This requires that administrators of the enrollment management communicate with administrators of the financial aid office to better allocate scholarship and financial aid rewards in order to maximize enrollment. Considering the large expenditure on the scholarship and the financial aid awards, this research explores different factors that presumably influence the enrollment decision of applicants. The intention of this work is to provide decision makers in the enrollment management administration a better understanding of the factors that are highly correlated to the enrollment process. These factors might better identify the applicants who have higher tendency to enroll at an institution relative to others. This allows enrollment management to assign money rewards efficiently and thus not only maximize enrollment but also save a big portion of institutional money. These factors basically include a wide range of features related to student characteristics and institutional characteristics.

For this purpose, we use real data for applicants admitted to the university of New Mexico (UNM) as a case study. UNM represents a variety of regional comprehensive universities and thus the results of this work could be widely applicable to other universities. UNM is a public research university in Albuquerque, New Mexico. It is the largest post-secondary institution in the state in total enrollment across all campuses and one of the state's largest employers. The acceptance rate at UNM is 45% with an average enrollment of 3,500 new beginning student per year [2].

*Associate Vice President of Enrollment Management

The results in this work are presented using machine learning methods and data mining techniques. We use logistic regression (LR) and support vector machines (SVMs) models in addition to time series analysis and probability approaches.

The remainder of this paper is organized as follows. Section 2 presents a descriptive definition of nationally effective student and institutional characteristics in addition to others that are introduced for the first time in the literature. Section 3 introduces our proposed models. Section 4 shows some experimental results. Finally, Section 5 presents some concluding remarks.

2. FEATURE DESCRIPTION

The data set used in this work contains more than fifty features for admitted students at UNM. These features describe some of the student and the college characteristics. These features are represented by a set of binary, categorical, discrete and continuous variables. The section below lists a description for each of these features and explains the intuition guiding us to include them in our analysis.

- GENDER: a binary variable indicating the sex of the applicant (i.e., male or female). This feature might be a good enrollment predictor if the population at UNM tends to lean towards one sex more than the other one.
- ETHNICITY: a categorical variable indicating the ethnicity of the applicant (i.e., black, white, latino and others). This feature might be a good predictor in case applicants of certain ethnicity has a tendency to enroll at UNM more than others.
- ACT_SCORE, SAT_SCORE: discrete variables reflecting the competence level of the applicant. These variables are represented by the ACT and/or the SAT scores. This feature might be a good predictor since students usually would rather enroll in colleges whose student population has a similar competence level.
- GPA: a continuous variable representing the high school GPA of the applicant. Its value ranges between 0 and 5. Similar to the ACT_SCORE and SAT_SCORE variables, GPA reflects the competence level of the applicant.
- FIRST_GENERATION: a binary variable indicating the education level of the parents. The label is 1 if at least one of the parents went to college and 0 otherwise. Usually parents provide their children with an advice on deciding which college to attend. Thus this variable might be a good predictor.
- PARENT_INCOME: a continuous variable indicating the total income of the parents. As mentioned earlier, this feature reflects the socioeconomic status of the parents and should have a major influence on the student's decision choosing which institution to attend.
- STUDENT_INCOME: a continuous variable indicating the income of the student in case he or she has a job. Similar to the PARENT_INCOME variable, STUDENT_INCOME has an influence on the applicant's decision.
- RESIDENCY_STATE: a categorical variable indicating the residency status of the applicant. This variable is a relative measure of the distance from the applicant's residency to UNM campus. It has four labels: 0 indicating that the applicant resides in New Mexico and thus considered as in-state student; 1 indicating that the applicant resides in either Texas, California, Arizona or Colorado. Applicants in those states are eligible to the Amigo scholarship which allows them to pay in-state tuition if they meet certain criteria; 2 indicating that the applicant is non-resident; 3 indicating that the applicant is international. This variable might be a good predictor since it reveals the type of relation between the distance from the campus (implicitly the cost) and the applicant's decision.
- INSTITUTIONAL_MONEY: a continuous variable indicating the amount of the financial aid assigned to the applicant by the institution (i.e., UNM).
- BRIDGE, SUCCESS: binary variables indicating the type of the financial aid offered by UNM. BRIDGE is a reward given for freshman students in their first semester. It is exclusively given to applicants with certain aptitude levels; SUCCESS is a reward given for freshman students in their first semester. It is eligible to applicants with financial needs.
- FEDERAL_MONEY: a continuous variable indicating the amount of the financial aid assigned to the applicant by the federal government.
- APPL_DECISION_DIFF: a discrete variable indicating the total number of days between the time of the application submission and the time of the admission decision. The gap between these two events might be a good predictor. For example, if the admission decision was taken shortly after the application submission, this might provoke the applicant's tendency to enroll at UNM.
- APPLY_AFEB: a binary variable indicating the month during which the application is submitted. The label is 1 if the application is submitted after February and 0 otherwise.

There are many other features that are used in this work. However we did not mention them all because their characteristics are similar-to an extent- to the above listed features. We believe that the features described above give realistic examples of factors that have the potential to influence the applicant's decision of enrollment.

3. MODELS

In this work we approach the enrollment prediction question from a classification perspective. That is we have a pool of applicants and we want to identify or classify those that are most likely to enroll at UNM. We further divide the classification problem into two main approaches: classification at individual level and classification at a cohort level. The individual level approach predicts the enrollment of an applicant based on a given set of features. Then it determines the total number of enrollment by simply counting the applicants who

are predicted to enroll. For this purpose, we use two common machine learning classifiers with two class responses: LR and SVMs. The LR with two class responses is one of the basic classification models that aim to find the relationship between a binary response y and a predictor variable(s) x, which can be in a categorical or numerical scale [6]. The response variable y of the binary logistic regression consists of two categories i.e. success and fail. In some cases, the categories are denoted as 1 for success and 0 for fail. In our case the label is 1 if the applicant is predicted to enroll and 0 otherwise. However, a disadvantage of logistic regression is that the technique is not able to identify possible nonlinear structures in the data. A good alternative in this case would be SVMs. On the other side the cohort approach predicts the enrollment of a cohort of applicants based on a given set of features. Using this approach we directly determine the portion of the applicants' pool that would enroll without identifying them individually. The sections below explain in more detail the difference between these approaches and the models used to implement them.

Variable selection

In this work we use a total of 60 features to predict enrollment. It is always possible that some of these features are redundant or irrelevant. Thus they can be removed from the prediction models without incurring much loss of information. One approach to encounter such features is variable selection. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [4]. The variable selection methods are typically presented in three classes: filter method, wrapper method and embedded method. In this work we implement the wrapper methodology which come in two flavors: forward selection and backward elimination. In forward selection, variables are continuously added into bigger and bigger subsets, whereas in backward elimination we start with a set containing all the variables and iteratively remove the variable with the least predictability. Both methods yield nested subsets of variables.

3.1 Classification at cohort level

The LR and SMVs models used in this work predict enrollment at individual level. That is, given a set of features for an applicant a_i, the LR and SVMs models predict the probability of enrollment of a_i (LR) or alternatively give a 0/1 flag indicating if a_i will enroll or not (SVMs). In this work we present a new approach in which we predict enrollment at cohort level. That is, given a set of features for a cohort of applicants c_i, we predict the portion of c_i that would enroll. The main concepts underlying this approach are probability and time series analysis. In the probabilistic approach we define a probability distribution over the features of c_i and accordingly compute the respective portion that would enroll. The results shown by this approach prove to be promising.

The probabilistic model is based on a semi-supervised learning method. It is defined as following:

$$p_X(x) = P_0.p_{X|0}(x) + P_1.p_{X|1}(x) \quad (1)$$

where

$p_{X|0}$ is the distribution over the features of applicants that do not enroll. It is estimated from the training data.

$p_{X|1}$ is the distribution over the features of applicants that do enroll. It is estimated from the training data.

$P_1 = 1 - P_0$ is the fraction of applicants that do enroll.

p_X is the distribution over the features of unlabeled applicant data.

So if the total number of applicants is n then the predicted number that would enroll is $n_{enroll} = nP_1$. In this case all what we need to do is to estimate P_1. Solving (1) for P_1 (using $P_1 = 1 - P_0$) gives:

$$P_1 = \frac{p_X(x) - p_{X|0}(x)}{p_{X|1}(x) - p_{X|0}(x)} \quad (2)$$

true for all x.

The second approach used in this work to predict enrollment at cohort level is based on time series analysis. A time series is a sequence of observations collected over time. Usually these observations are taken at constant intervals (i.e., daily, monthly, annually, etc.). The main object of time series analysis is to reveal the model underlying the process generating the series data. Such a model is used to describe the patterns in the series (i.e., trend , seasonality), explain how past observations influence future ones, and accordingly forecast future values of the series [1]. In this work we use a seasonal autoregressive integrated moving average ($ARIMA$) model to forecast the number of applicants that would enroll at UNM. A seasonal $ARIMA$ model is defined as an $ARIMA(p,d,q)\text{x}(P,D,Q)_m$ model, where

- p is the number of autoregressive terms
- d is the number of differences
- q is the number of moving average terms
- P is the number of seasonal autoregressive terms
- D is the number of seasonal differences
- Q is the number of seasonal moving average terms
- m is the number of periods per season

4. EXPERIMENTAL RESULTS

In an attempt to empirically validate the performance of our proposed models, we analyzed actual university data. For this purpose we used the data of 54,692 First Time Full Time (FTFT) students who were admittted to UNM between years 2009 and 2016. We used this data set in our work in order to layout the needed features, train our models and test their performance.

4.1 Data-preprocessing

In order to get more consistent and discipline results it is essential to preprocess the data set. For this purpose, we implemented a number of common preprocessing techniques used in machine learning.

For various reasons, the data set used in this work contains missing values. That is for some admitted students, UNM does not have all the required information (ex. parents income). This leaves the values of some features in our data set blank. A basic strategy to overcome this problem is to implement imputation methods such as the mean, median or mode of the row or column in which the missing values are located. Another strategy would be simply to discard or remove the rows and/or columns containing missing values. This might come at the price of losing information. However, this might not be the case if the training data set is big enough in which removing some rows will not impact the model performance. In this work we simply discarded the rows with missing values. Consequently, we were left with a data set of 37,500 student which is enough to train our models. Next we standardized the continuous and discrete features of the data set. We removed the mean value of those features and scale them by their respective standard deviation values. Standardization improve the performance of the models by adjusting the features to the same scale. We also converted categorical features to binary features using one-hot encoding. This estimator transforms each categorical feature with m possible values into m binary features, with only one active.

4.2 Numerical results

We used 60 features to train the LR and the SVMs models. We used 10-fold cross validation to examine the performance of these models and compare their results as well. The performance accuracy of both models is presented in Table 1.

	Performance Accuracy (%)
LR	89.41
SVMs	91.25

Table 1: The performance accuracy of the LR and SVMs models using 10-fold cross validation.

It is important to mention that in our training data set the number of observations in each class is not equal. The number of applicants who enroll at UNM is relatively higher than those who do not enroll. In this case the performance accuracy of the classifier can be misleading. A better metric to test the performance of a classifier is a confusion matrix. It is a technique for summarizing the performance of a classification algorithm. A confusion matrix gives a better idea of what the classification model is getting right and what types of errors it is making. Table 2 and Table 3 show the confusion matrices for the LR model and the SVMs model.

The precision and recall values for the LR and the SVMs models are shown in Fig. 1. Both precision and recall are good measures to examine the relevance of the predicted instances to the actual ones. They are calculated using the confusion matrices and hence they are reliable measures to summarize such matrices.

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	2060	267	2327
	n′	130	1293	1423
	total	2190	1560	

Table 2: The confusion matrix of the LR model.

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	2063	201	2265
	n′	127	1359	1485
	total	2190	1560	

Table 3: The confusion matrix of the SVMs model.

The performance accuracy, the confusion matrices and the precision and recall scores of the LR and the SVMs models are very similar. Thus the advantage of the SVMs model over the LR model in identifying nonlinear structures is not utilized here. This suggests that the LR model is sufficient to achieve enrollment prediction with a reliable performance. In this context we applied the LR model again to find a subset of the features that attain a similar performance accuracy without losing much information. So we implemented the forward and the backward variable selection models to remove possible redundant and irrelevant features. The results are shown in Fig. 2. The figure presents the classification error of the LR model using 5-fold cross validation for both backward and forward methods. It shows that the forward method has slightly better performance over the backward method. In fact using a subset of 14 features only (the red circle) can achieve a performance accuracy of 89%. This result is almost equal to the performance accuracy of the LR model when using all the features (Table 1). In other words, we can use these 14 features only to predict the enrollment for any future pool of applicants without the need to include the rest of the features in our prediction models. We provide a description for 10 of these features. The description is provided below. These features are sorted according to the predictive importance criterion proposed by the forward selection method. They are listed in a descending order:

- STATE_AWARD_ORIGINAL: This is a continuous variable. It is the amount of the scholarship offered to the

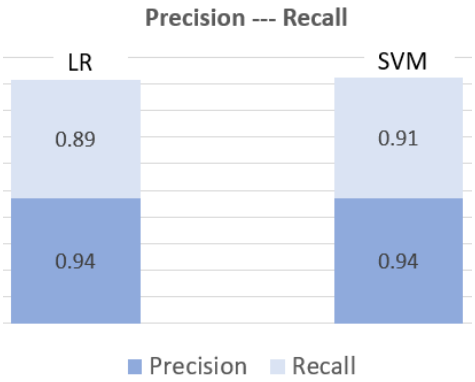

Figure 1: The precision and recall for the LR and SVMs models.

applicants by the state of New Mexico. Perhaps the most important among others is the lottery scholarship. The results presented by our LR model show that on average applicants with state awards tend to enroll more at UNM.

- FIRST_DECISION_AFEB: This is a binary variable. It represents the time of the admission decision. The label is 1 if the admission decision is taken after February (i.e., March, April, May, June and July) and 0 otherwise. The results show that applicants have more tendency to enroll at UNM if the admission decision is taken after February.

- SUCCESS: This is a binary variable. It is the reward given for applicants in their first semester. It is eligible to those with financial needs. The total amount of the reward is 1,000 $. The LR model shows that applicants with SUCCESS rewards are more likely to enroll at UNM.

- GPA: This is a continuous variable. It represents the high school GPA of the applicants. The results show that applicants with high school GPA between 3.0 and 3.5 tend to enroll more at UNM compared to other applicants.

- RESIDENCY_STATE: This is a categorical variable indicating the residency status of the applicant. The results show that applicants who resides in NM are more likely to enroll at UNM (not surprising!).

- FAFSA_BDEADLINE: This is a binary variable. It indicates if the applicants submit the Free Application for Federal Student Aid (FAFSA) before the deadline set by UNM. The results show that applicants who submit the FAFSA before the deadline tend to enroll more at UNM compared to those who submit after the deadline.

- LOW_INCOME: This is a binary variable. It reflects the socioeconomic status of the parents. The results show that applicants whose parents have a low income are more likely to enroll at UNM.

- BRIDGE: This is a binary variable. It is the reward given for freshman students in their first semester. It is exclusively given to applicants with certain aptitude levels. The total amount of the reward is 1,500 $. The LR model shows that applicants with BRIDGE rewards are more likely to enroll at UNM.

- APP_AFEB: This is a binary variable. It represents the time when the applicants submit their applications. The label is 1 if the submission is done after February (i.e., March, April, May) and 0 otherwise. The results show that applicants have more tendency to enroll at UNM if the submission is done after February.

- FED_AWARD_ORIGINAL: This is a continuous variable. It is the amount of the financial aid offered to the applicants by the federal state. Perhaps the most important is the Pell grant. The results presented by our LR model show that on overage applicants with federal awards tend to enroll more at UNM.

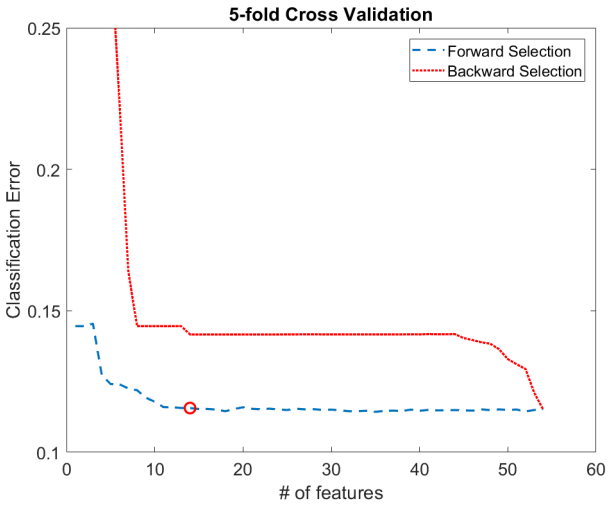

Figure 2: The classification error of the backward and forward variable selection methods implemented using the LR model to predict enrollment.

Cohort prediction

As mentioned earlier the LR and the SVMs models predict enrollment at individual level. In this work we propose alternative approaches where we predict enrollment at cohort level. The first approach is probabilistic in which the total enrollment is computed using (2). We implemented this approach following these set of steps:

- Use previous year data to estimate $p_{X|0}$ and $p_{X|1}$ (labeled data).
- Use current year data to estimate p_X (unlabeled data).
- Use (2) to estimate P_1 for current year.
- Predict the total enrollment: $n_{enroll} = nP_1$

To empirically validate our proposed model we used actual university data for UNM students admitted in years 2015

and 2016. We used the 2015 cohort as a training data set to estimate $p_{X|0}$ and $p_{X|1}$. Then we used the 2016 cohort to estimate p_X and accordingly compute P_1 using (2). The actual and the predicted total enrollment for the 2016 cohort are shown in Table 4.

	Total enrollment (2016)
Actual	3402
Predicted	3478

Table 4: The total enrollment of 2016 cohort at UNM.

We repeated the same procedure, however this time using the 2016 cohort as a training data set to predict the enrollment of the 2015 cohort. The results are shown in Table 5.

	Total enrollment (2015)
Actual	3320
Predicted	3239

Table 5: The total enrollment of 2015 cohort at UNM.

It is essential to mention that we estimated p_X, $p_{X|0}$ and $p_{X|1}$ using only one feature. We evaluated these densities using the histogram method. Then we used (2) to estimate P_1 at multiple x values (histogram bins) and averaged these results to obtain a final P_1 estimate. A remarkable observation using this approach is the accurate predictions using just one feature. This is reasonable. The feature does not have to provide good discrimination because we are not trying to predict individual enrollment; instead we just need to estimate P_1.

Time series analysis is another approach to forecast student enrollment. Unlike the other classification models used in this work, a time series model does not require features to carry out predictions. It only requires a sequence of observations collected over time. This sequence enables us to reveal the model underlying the process generating the series data. In this context we collected the number of students enrolled at UNM for spring, summer and fall semesters of each year. The study contains students enrolled at UNM between years 2003 and 2016. The time series data for this study is shown in Fig. 3 (black color). Note that the number of periods, m, in the series is 3 referring to spring, summer and fall semesters. In this work we used the Akaike Information Critera (AIC) as a statistical measure to choose the $ARIMA$ model that best fits the series. AIC is a widely used measure in statistics. It reflects the robustness of the fitted model in a single value. When comparing two $ARIMA$ models, the one with the lower AIC is generally "better". The parameters of the $ARIMA$ model that best fit the time series of the UNM enrollment data are $p = 0$, $d = 0$, $q = 0$, $P = 1$, $D = 0$ and $Q = 3$ (i.e., $ARIMA(0,0,0)\text{x}(1,0,3)_3$). The fitted model is represented by the red curve in Fig. 3. This model has the lowest AIC and we used it to predict the enrollment at UNM for spring, summer and fall semesters of the 2017 cohort. The predicted numbers are represented by the blue curve of Fig. 3. Table 6 shows the actual versus the predicted enrollment numbers at UNM for the 2017 cohort with 80% confidence interval.

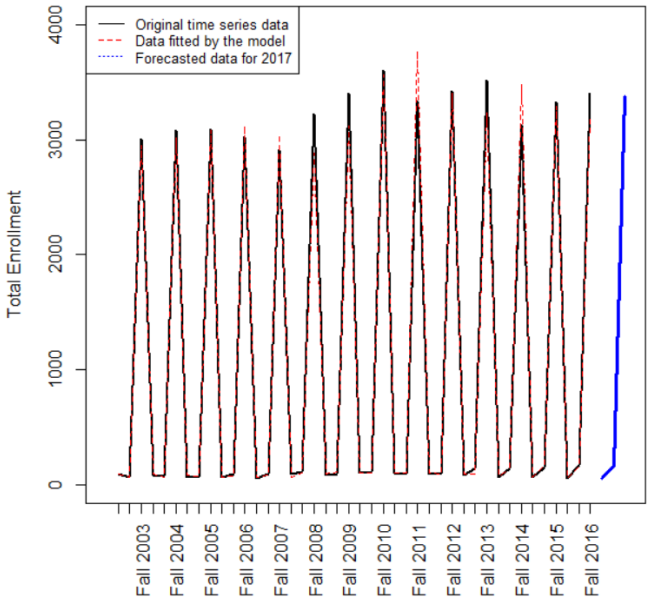

Figure 3: The $ARIMA(0,0,0)\text{x}(1,0,3)_3$ model for the enrollment data at UNM.

	Actual total enrollment (2017)	Predicted total enrollment (2017)	Lower bound (80%)	Upper bound (80%)
Spring	70	57	48	68
Summer	132	153	129	181
Fall	3219	3380	2850	4010

Table 6: The actual and the predicted enrollment of the 2017 cohort at UNM.

5. CONCLUSION

In this paper we shed the light on factors that influence the enrollment decision of applicants. We use machine learning methods to measure the level of correlation between enrollment and such factors. In particular we approach the enrollment prediction question from a classification perspective where we need to identify the likelihood of enrollment for a pool of applicants. We further divide the classification problem into two main approaches: classification at individual level and classification at a cohort level. The individual level approach predicts the enrollment of an applicant based on a given set of features. Then it determines the total number of enrollment by simply counting the applicants who are predicted to enroll. For this approach we implemented a LR model and an SVM model. On the other side the cohort approach predicts the enrollment of a cohort of applicants based on a given set of features. For this approach we implement a semi-supervised probability model and a time series model. Using this approach we directly determine the portion of the applicants' pool that would enroll without identifying them individually. The results show that our proposed models can predict enrollment with reliable accuracy using only a small set of features related to student and college characteristics.

6. REFERENCES

[1] Applied time series analysis. Penn State Eberly College of Science. Available at `https://onlinecourses.science.psu.edu/stat510/node/47`.

[2] Office of institutional analytics. Available at `http://oia.unm.edu/facts-and-figures/freshman-cohort-tracking-reports.html`. Accessed: 1-2-2018.

[3] Enrollment management in higher education - defining enrollment management, key offices and tasks in enrollment management, organizational models. Education Encyclopedia - StateUniversity.com, 2013.

[4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[5] D. Hossler and . Bean, John P. *The strategic management of college enrollments*. San Francisco, Calif. : Jossey-Bass, 1st ed edition, 1990. Includes bibliographical references (p. 303-318) and index.

[6] H.-A. Park. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Korean Society of Nursing Science*, 43(2), 2013.