PHILIPPINE ENGINEERING JOURNAL
PEJ 2022; Vol. 43, No. 1: 47-64

Received: 8 July 2021
Accepted: 1 February 2022

Identifying Factors Influencing Engineering Undergraduate Student Graduation in UP Diliman

Darvy P. Ong and Jhoanna Rhodette I. Pedrasa
Electrical and Electronics Engineering Institute, College of Engineering,
University of the Philippines Diliman, Quezon City, Philippines

Abstract — *Understanding how different factors affect the performance of a student in the university setting is important in policy making and providing a better environment for learning. Existing studies on student graduation rates typically employ the use of statistical analyses to correlate a student's profile and their chances of graduation. Building on the success of these methods for Western institutions, we used Logistic Regression together with other statistical metrics such as Wald's $\chi 2$ Test and Odds Ratios to evaluate the contributing factors that may affect student graduation chances. The results show that the main factors affecting graduation include enrollment in the preferred degree, Math scores in the college admission test, high school academic performance, proximity to the university, and economic background. Finally, we also found that including post-matriculation factors in the model increased model performance significantly.*

Keywords — logistic regression, engineering education, factor analysis, odds ratio, graduation

I. INTRODUCTION

In any academic institution, it is of utmost importance that educators are able to understand the needs of their students. Understanding how different factors may be affecting the experience of students inside our institutions is important specially since it may help administrators gain better insight on their students and be able to formulate new and revise existing policies according to their students' needs.

The objective of this study is to identify factors affecting student graduation in the College, with the aim of using this information to make more informed decisions on how to improve graduation and retention in the program. A quantitative approach was originally proposed focusing on pre- matriculation factors, but over the course of the study, post-matriculation factors were also used to gain more insight. Furthermore, while there are studies tackling this subject matter, there is a lack of studies from Asian universities, the Philippines, and the College of Engineering. It is therefore important that we be able to utilize these techniques, and apply them to a localized student body, thereby providing a benchmark for studies in the Asian and local setting.

Data was collected on freshmen students from 2009 to 2013 and students who graduated between 2013 to 2018 to generate statistics on graduation rates. From this, it was found that among those who enter the COE through the UP College Admission Test (UPCAT), only 66.69% graduate within the college. Furthermore, only 58.06% of the freshmen graduate in their original degree programs, and only 39.53% of the freshmen graduate on-time. In contrast,

the findings of a study conducted in 2016, [1], found that among the 7163 who entered the UPD through the UPCAT from 2010 to 2014, only 61% have graduated at the time of study, with 3428 (47.86%) students graduating from their original degree program and 922 (12.87%) students graduating from a different degree program. From these numbers it can be seen that the conversion rate of the university from freshmen to graduates is not promising, especially when compared to graduation rates from similar degree programs from well-known universities overseas which have graduation rates between 85% to 90% [2][3]. It is our hope that through this study, we may be able to gain insight on how different factors affect the graduation chances of students in our college.

The following are the research contributions of this work. Our work provides an extensive insight into Engineering students, with a total of 26 pre- and post-matriculation factors, included in the study. The study also tracks the students admitted over a five-year period covering a total of 4,809 students. Furthermore, the model was broken down to graduation with original degree program and within the College of Engineering, allowing us insights in how shifting between programs within the college affects graduation. Finally, this study also documents graduation statistics pre-K-12 transition to serve as benchmarks for future studies. Archiving such statistics is important as it is difficult to track progress or determine the effects of any policy change without such data. To the best of our knowledge, this study is also the first such study conducted for a Southeast Asian University.

II. RELATED WORK

Existing studies on student retention have employed different statistical methods to determine the causes of student success. This section will discuss studies that used statistical tools for identifying significant factors to student retention in engineering.

A study, [4], used digital signal processing to create a simple model capable of predicting student graduation state. They found that using this technique yielded varied results in accuracy, listed as follows: 55% (for 1990 students), 75% (1992 students), 80% (for 1992 students), and 94% (for 1993 students). The study also found that the subjects taken in the first two years of their college stay are critical in determining whether or not the student will graduate. This study is specially of interest to us since it was conducted on data from students of the then EEE Department of the COE.

In the study, [5], exploratory factor analysis was used in order to determine the relationship between engineering students performance in their first year and the following factors: SAT Math Score, SAT Verbal Score, HS Physics Regent Exam Score, HS Chemistry Regent Exam Score, HS Physics Grade, HS Chemistry Grade, Gender, Interest in Science, Math, or Physics. They found that the factors affecting student performance were the SAT Math score and taking extra classes in math and physics.

On the other hand, another study, [6], used logistic regression with backwards elimination to determine which among the following factors have a significant effect on the graduation of students: Sex, Race, Residency, High School Ranking, High School GPA, ACT Science, ACT

Math, ACT English, and ACT Reading. They found that for engineering students, HS GPA, and ACT Math, English, and Science were significant factors to graduation.

Another study [7] used statistical analysis with logistic regression to create a model that calculates the probation probability of a student. To test their model, they employed a goodness of fit test, presenting results in the form of odds ratios. They found that SAT score, high school rank, gender, and summer bridge programs had a significant effect on student probation probability.

Finally, multiple logistic regression was used in [8] to create a model for graduation vs factors being considered. Wald Chi-square statistics were used to test for significance of factors, and it was found that high school GPA, gender, ethnicity, SAT Quantitative and Verbal scores, and citizenship are significant factors to graduation. Furthermore, it was determined using the coefficient of determination that the factors included explained only 12.6% of the variance in the graduation.

The studies previously mentioned used statistical methods to identify important factors. It is prevalent among these studies that the method of logistic regression widely used for analyzing student retention and graduation due to its capability of modelling given categorical and continuous data into a binary output of yes or no. It should be of note that while these studies used a variety of student factors in their analyses, none of them have included post-matriculation factors. Our work includes both pre-matriculation and post-matriculation factors to be able to paint a more complete picture of what factors really influence student success.

III. DATA PREPARATION

3.1 Data Collection

For this study, data was collected and collated from three main sources: (1) the Computerized Registration System, (2) the College Secretary's office, and (3) the University's Office of Admissions.

The Computerized Registration System (CRS) hosts the university's registration process system and can generate different standard reports on student profile. The COE College Secretary's office on the other hand, provides additional information such as official lists of freshmen, transferees, shiftees, and second degree takers. Finally, the Office of Admissions (OA) is the initial entry point of all students in the university. The OA holds all information submitted by the students during their college applications; at the same time, the OA is also responsible for determining a student's eligibility to enroll in the university.

Based on existing studies, the pre-matriculation data or factors found in Table 1 were selected as an initial starting point for this study. Some of these factors have been localized, for example, instead of ACT scores, we will be using the university's equivalent University of the Philippines College Admission Test (UPCAT) scores. Table 1 also indicates the type of data for each factor, may it be categorical or continuous. The scope of the data collected for

this study spans the COE freshmen coming from five (5) academic years, starting from 2009 and ending with 2013, with a total of 4,809 students.

3.2 Data Processing

To prepare the collected data for use with the second part of this study, the data was further processed with the following considerations: first, Factors made up of NaN or No Data were removed since their inclusion wouldn't be beneficial to the analysis; second, rows or data points with empty cells (NaN, No Data, or blanks) were removed to simplify the analysis; third, factors with non-changing values were removed. Doing the above simplifies the study so that we can focus on the analysis of the dataset.

Table 1. Collated Student Factors

CONTINUOUS FACTORS	CATEGORICAL FACTORS	
Age of Entry to UP	Degree Program Upon Entry	School Type
High School GWA	Year of First Enrolment	HS Province
UPCAT Science T-Score	Sex	Enrolled in 1st or 2nd choice program
UPCAT Math T-Score	Country of Citizenship	Dependent on UP Faculty or Employee
UPCAT Reading Comprehension T-Score	STFAP Bracket Upon Entry (Socialized Tuition System)	Had Leave of Absence (LOA) or Absence without Leave (AWOL)*
UPCAT Language Proficiency T-Score	Religion	Had a Scholarship*
UPCAT Average T-Score	Address Location (Permanent)	Lived in a Residence Hall*
UP Predicted Grade	Civil Status	Shifted within Engineering*
UP Math Predicted Grade	Minority Group Member	

**Post-Matriculation Factors*

3.2.1 Handling Categorical Factors – One-Hot Encoding

To ensure that the usability of the dataset with the selected modelling method, all categorical factors were converted into a numerical format using one-hot encoding. One-hot encoding converts each category value in each factor into a new column and assigns a one or zero based on the original factors' value [9]. This method of encoding was chosen because it converts categorical values into continuous values without introducing weights to its values, which is especially important for nominal factors.

An example of this is shown in Table 2; the table shows a categorical factor, color, with three distinct values: Red, Blue, and Yellow. Each of these values are converted into a new column, with its zero as its values unless the original data in the color column includes the columns namesake value, in which case a one is indicated.

Table 2. One-Hot Encoding Example

Color	Color_Red	Color_Blue	Color_Yellow
Red	1	0	0
Blue	0	1	0
Yellow	0	0	1
Red	1	0	0

Furthermore, after using One-hot encoding to convert the categorical factors, one column from each categorical factors' columns is dropped to avoid statistical redundancy. For example, in Table 2, at least one of the three dummy Color columns should be removed to avoid creating a multicollinear set of columns [10].

3.2.2 Identifying Factors with High Association/Correlation

To handle multicollinearity amongst the factors used for modelling, the correlation of each factor against all other factors from the same type was computed. This information is presented in the form of a correlation matrix, with each cell showing the correlation coefficient between two different factors indicated in the row and column headers [11].

To compute for the strength of the relationship between continuous factors, Pearson's correlation coefficient was utilized. The Pearson's Correlation (r) is commonly used in regression to determine the linear relationship between two sets of data [12]. The correlation between two variables x and y can be calculated by using (1), where n is the number of data points in the dataset.

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

Using (1) returns a value between negative and positive one. A negative one indicates a strong negative relationship, meaning that x and y are inversely proportional to each other; on the other hand, a positive one indicates a strong positive relationship, meaning that x and y are directly proportional to each other. Finally, a value of zero indicates that there is no correlation between x and y.

Once the correlation coefficients have been calculated, factors which have correlations higher than or equal to a magnitude of 0.5 will be identified and marked as factors with strong relationships. Finally, factors will be selected and removed to lessen the strength of correlation between factors in the data set.

On the other hand, for categorical factors, Cramer's V Coefficient (V) was used to identify high associations between factors [13]. Cramer's V Coefficient is calculated using (2) and (3), where nij is the number of observations in the intersection of category i and j, with i coming from the first factor and j from the second factor.

$$\chi^2 = \sum_{ij} \frac{\left(n_{ij} - n_{i.}\frac{n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \tag{2}$$

$$V = \sqrt{\frac{\frac{\chi^2}{n}}{\min\left(categories_a - 1,\ categories_b - 1\right)}} \tag{3}$$

Using (3) yields a value between zero and one, where a value closer to one means the factors have high association and a value closer to zero means the factors have minimal association. Once the Cramer's V coefficient is computed, factors who have a coefficient value that is higher than or equal to 0.5 as highly associated are marked and the factor with more categories is removed to further minimize the variance of the dataset.

IV. MODELS AND STATISTICAL TESTS

4.1 Logistic Regression

To identify the factors that influence the graduation rate of engineering students, the relationship of each of the factors to the graduation status of students will be measured using Logistic Regression. Using this, a model of the relationship between the independent categorical and continuous factors to the binary output factor, Graduation Status, is made.

$$\pi(X) = \frac{e^{\beta_o+\beta_1X_1+\beta_2X_2+\cdots+\beta_NX_N}}{1 + e^{\beta_o+\beta_1X_1+\beta_2X_2+\cdots+\beta_NX_N}} \tag{4}$$

The logistic regression equation is defined by Eq. 4, where the X's are the factors being considered and the β's are the weights of the factors [14][15]. The goal of the equation is to find the best fit slopes for the factors using maximum likelihood estimation such that the standard logistic function shown in Fig. 1 has a minimized or nearly vertical slope [16].

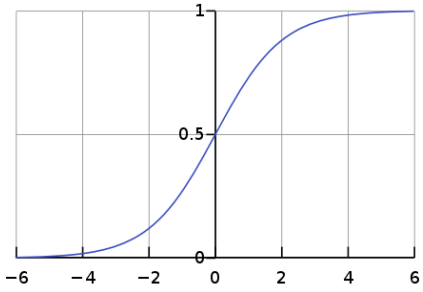

Figure 1. Standard Logistic Function [17]

Once the logistic regression model has been created, it can be used to determine whether a student will graduate given a specific set of values as their factors. As can be seen in Fig. 1, the expression does not result in a binary output, instead, it provides a probability value between zero and one. The decision boundary is typically set at 0.5, which means that inputs that result in a probability value greater than or equal to 0.5 will be mapped as graduated, while all values less than 0.5 will be mapped as not graduated.

4.2 Likelihood Ratio Test

Once the model has been created, the likelihood ratio test is used to determine whether the input factors have a significant effect on Graduation Status. The global null hypothesis, H_o, is set as the case wherein including the input factors in the model does not have a significant effect on the Graduation Status of students. To test the null hypothesis, the log-likelihood of the model is computed using (5) [18].

$$L(\theta|x) = \log\left(\prod_i \pi_\theta(x_i)^{y_i}(1-\pi_\theta)^{1-y^i}\right) \tag{5}$$

Equation (5) produces two likelihood values: $L(q\,|x)_{model}$ which is the likelihood value for when the model includes all input factors, and $L(q\,|x)_{null}$ the likelihood value for when the model doesn't have any input factors. Using these two values, the likelihood ratio of the model can be computed using (6).

$$LLR = 2[L(\theta|x)_{model} - L(\theta|x)_{null}] \tag{6}$$

The log-likelihood ratio (LLR) can then be used together with the χ^2 distribution to determine the p-value of the model: $p_{value} = P[\chi^2(df) > LLR]$. In this study, the significance level is set at a = *0.05*, which means that for all p-values less than 0.05, it can be said that there is strong evidence that the input factors have an effect on Graduation Status. On the other hand, when the p-value is greater than or equal to 0.05, there is not enough evidence to say that the input factors influence the outcome.

4.3 Coefficient of Determination (McFadden's Pseudo R^2)

While the likelihood ratio test helps determine the significance of a model, it is also important to know how the model performs with respect to the outcome variable. Using the Coefficient of Determination, the goodness of fit of the input factors to Graduation Status can be determined [19].

To determine the goodness of fit of logistic regression models, Mc-Fadden's Pseudo R^2 is used. The pseudo R^2 can be computed using (7) and yields a value between zero and one.

$$R^2_{McFadden} = 1 - \frac{L(\theta|x)_{model}}{L(\theta|x)_{null}} \tag{7}$$

It uses the log-likelihood values of the full model and null model to determine goodness of fit. To interpret the values calculated from (7), it is important to understand that the equation provides insight on the proportional reduction in error variance of a model. As such, while the

value can range from zero to one, the range for good models is not linearly distributed. In fact, according to McFadden, R^2 values in the range of 0.2 to 0.4 are considered as good fit models [20].

Additionally, it should be noted that pseudo R^2's in general only have meaning when compared to another pseudo R^2 value obtained from a model using the same data and predicting the same outcome. This means that while comparisons can be made between different models using the same data, the pseudo R^2 values cannot be used to make comparisons between the same model on different datasets.

4.4 Wald's χ^2 Statistic

After evaluating the whole model, analysis on the individual input factors is then performed. The significance of each input factor to the model is determined using Wald's Test, which can be computed using (8), where b is the coefficient of the factor in the logistic regression model and $\beta_{std.err.}$ is the standard error of the factor's coefficient [21].

$$\chi^2 = \frac{\beta}{\beta_{staderr}} \tag{8}$$

Equation (8) gives a statistic capable of identifying factor significance regardless of the type of factor. For continuous and bi-categorical factors, it yields one value, but for multi-categorical values, it provides one overall value, with $n-1$ values representing the n categories of the factor. To interpret the statistic, the χ^2 distribution is used to determine the p-value of the input factor: $p_{value} = P[z > \chi^2]$.

In this study, the confidence level for factor significance is set to $a = 0.05$. This means that all factors that have a p-value less than 0.05 is considered as significant to Graduation Status, while those that have p-values greater than or equal to 0.05 does not have a significant effect on Graduation Status. Furthermore, for multi-categorical values, the $n-1$ additional values can be used to determine specific categories in the factor that affects Graduation Status.

4.5 Odds Ratio

Once a factor is determined as significant to Graduation Status, the relationship between the input factor and Graduation Status is further explored. To do this, Odds Ratio is used to measure the effect of changes in the input factor to the model output.

$$Odds = \frac{p}{1-p} \tag{9}$$

To understand odds ratios, it is important to first understand what odds are, as defined in (9). Odds is the ratio between the probability that an event occurs and the probability that the event does not occur. In this sense, it can be said that the odds ratio, which is defined by (10), is just a comparison between the odds of one factor against the odds of another factor [22].

$$OR = e^{\beta} \tag{10}$$

For continuous factors, the odds ratio can be interpreted as the increase or decrease in odds of an outcome given a unit increase or decrease in the input factor. For categorical factors on the other hand, the odds ratio can only be computed between two categories in a factor since it provides a ratio of odds. Therefore, for a categorical factor with n categories, a minimum of *n – 1* odds ratio values will be obtained.

Furthermore, interpreting odds ratios is as simple as looking at the values: for ratios equal to one, it means that exposure to the input does not affect the odds of the outcome; on the other hand, ratios greater than one translate to higher odds of the outcome when exposed to the input, and ratios less than one translate to lower odds of the outcome. Finally, to properly interpret odds ratios, Wald confidence intervals is used, computed using (11).

$$OR: \left[e^{\beta - \beta_{stderr}}, e^{\beta + \beta_{stderr}}\right] \tag{11}$$

This is the 95% confidence interval for the odds ratio, and it is useful for determining the precision of the odds ratio. A confidence interval with a small range of values translates to a high confidence on the relationship of the input to the outcome, while a large range means that while there is an idea on how the input translates to the outcome, there is not enough information to determine the exact relationship of the two. Additionally, the confidence interval can be used to determine the validity of the odds ratio. A good odds ratio interval should not contain one in its range, since an odds ratio of one is the boundary between positive and negative relationship of the input to the output. Therefore, whenever a confidence interval contains one, it cannot be said with certainty that the input variable has a positive or negative effect on the outcome even if the range of values is very small.

V. RESULTS AND ANALYSIS

5.1 Factor Association and Correlation

The results of using Pearson's correlation and Cramer's V to identify correlated and associated factors is discussed in this section. Figure 2 shows the correlation matrix for all continuous factors in the dataset. From this it can be said that there are three factors with high correlations to the other factors: (1) UPCAT Average T-Score, (2) UP Predicted Grade, and (3) UP Math Predicted Grade. These factors are used by the university's Office of Admissions to determine ranking during freshman student application, and are known to be computed using a combination of UPCAT related factors and student background, as such, it is to be expected that these factors will have a high correlation to other student factors and therefore will be removed from further analysis.

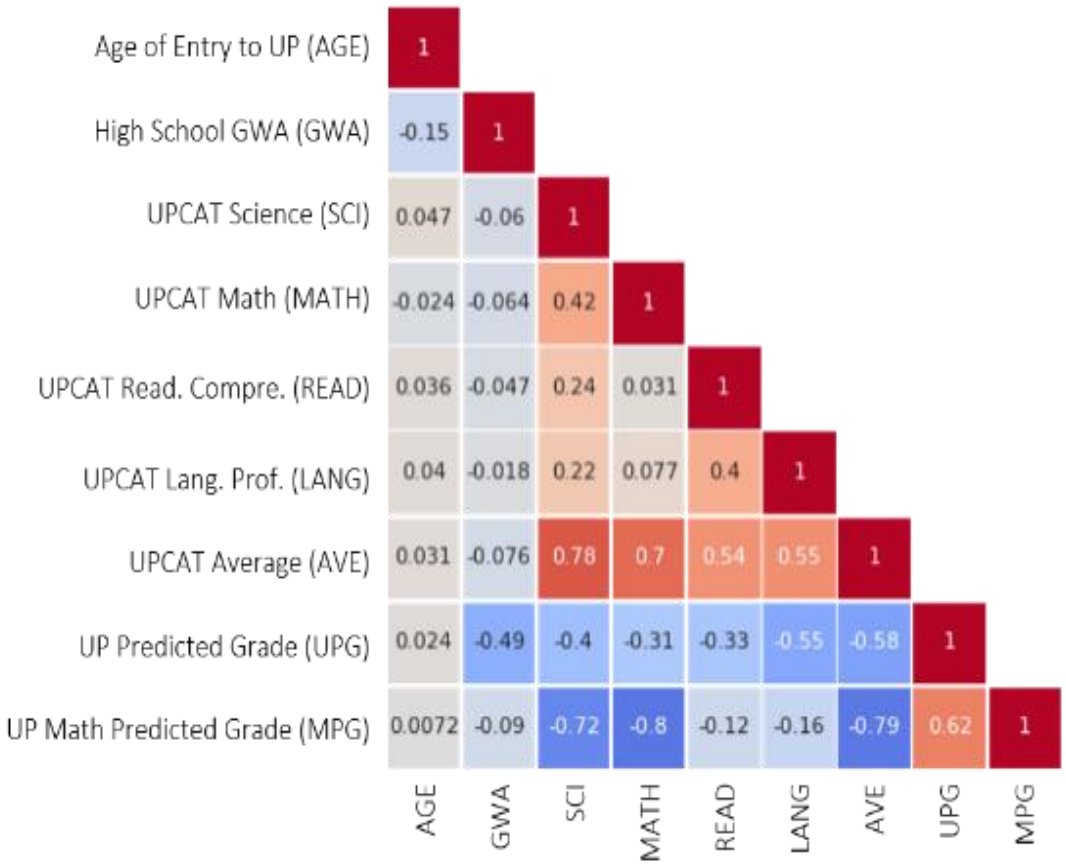

Figure 2. Pearson's Correlation Matrix

For categorical factors, the results of using Cramer's V is presented in Fig. 3. From this it can be said that High School Province and School Type are highly associated, with V=0.53. A high association score between school type and high school province is not unexpected since many students come from science and private high schools, a majority of which are situated in Luzon. Because of this, the factor with more categories, School Type, is removed to avoid multicollinearity in further analysis.

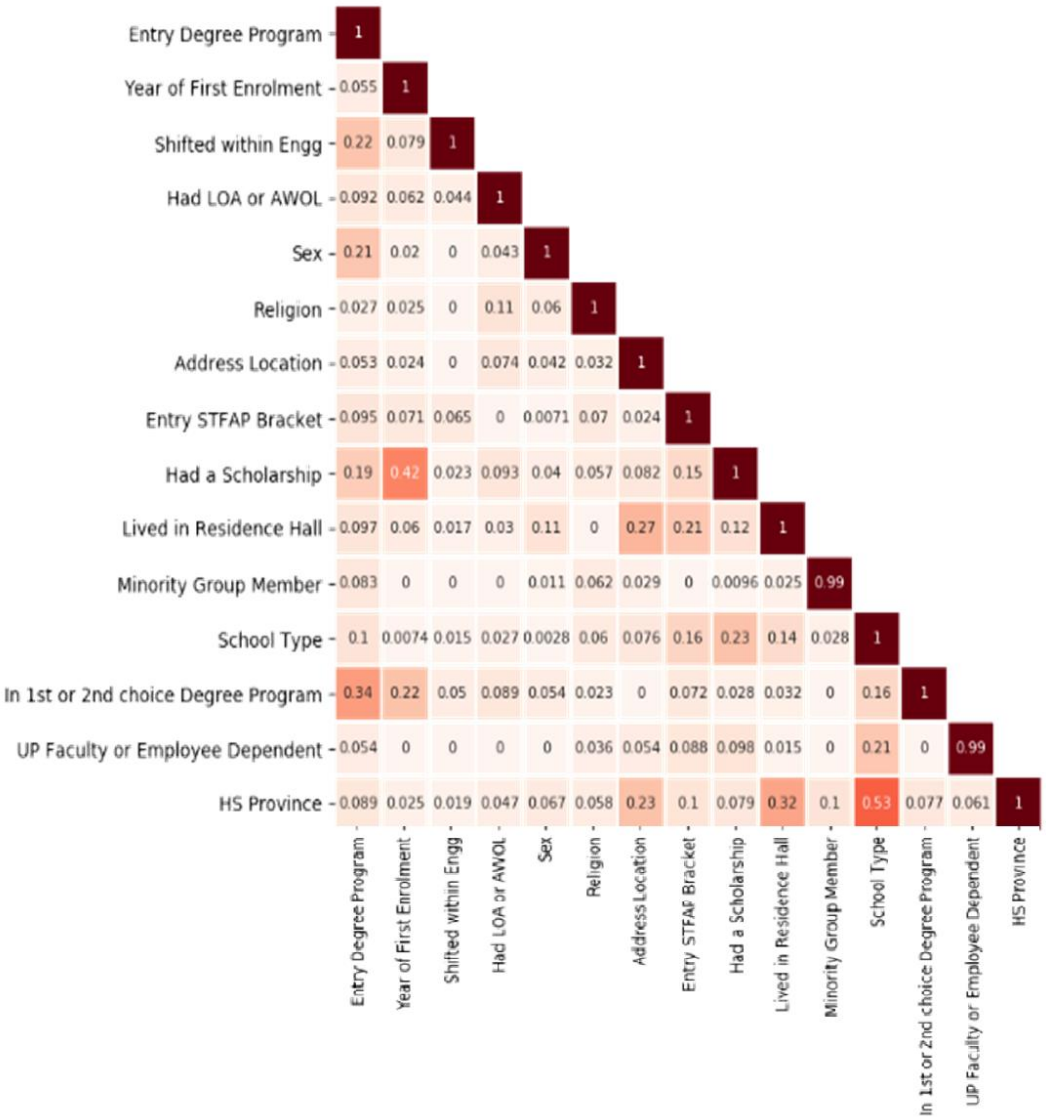

Figure 3. Cramer's V Matrix

In addition to this, from looking at simple distributions of each factor against graduation status, it was identified that some categorical factors in the dataset had minimal or no variance. In particular, the Country of Citizenship and Civil Status, both of which are bi-categorical factors, are affected. For Country of Citizenship, the distribution of data is as follows: 4779 for Philippines and 9 for Others. For Civil Status on the other hand, all data points were labelled as Single. Distributions like this that have more than 99% of their total data points in one category results in minimal variance for the factor and may lead to biased conclusions when used in analysis. Because of this, both factors are removed from the analysis.

Finally, standard statistical checks for correlation and distribution were performed during data pre-processing, and results were found to be inline with the assumptions of the abovementioned methods. After performing these, a total of 20 factors remained, and together with the 4809 data points, these comprise the final dataset used in analysis for this study.

5.2 Logistic Regression Model

To understand the effect of including post-matriculation factors in the model, two separate models were created. Table 3 presents the results from these models. The first column presents the results of the model which does not include any post-matriculation factors, while the second presents the results for the model which includes all available factors. From this, it can be said that while both models are significant compared to the null model (model with no factors) since the p-values are less than 0.05, the difference of including four post-matriculation factors is significant. The model with no post-matriculation factors has a pseudo R^2 value of 0.1065, which is below the threshold for good fit models, 0.2. On the other hand, the model with post-matriculation factors has a pseudo R^2 value of 0.3346, which means that by adding even just four post-matriculation factors has significantly increased the model fit of the dataset. A more detailed discussion of the logistic regression model can be found in [23].

Table 3. Model Information

	No Post-Matriculation Factors	**All Factors**
Number of Factors	16	20
Likelihood Ratio (LLR)	457.1193	1435.7896
LLR p-value	<0.0001	<0.0001
McFadden's Pseudo R^2	0.106541	0.334641

5.3 Wald's χ^2 Statistic and Odds Ratios

Table 4 presents the Wald's χ^2 statistics and odd ratios for the continuous and bi-categorical factors in the dataset. Out of the six continuous factors, it can be said that only *High School General Weighted Average, UPCAT Math,* and *UPCAT Reading Comprehension T-Score* were found to be significant. Furthermore, looking at the odds ratios, it should be noted that higher *High School GWA* and *UPCAT Math T-Scores* have a positive effect on graduation. On the other hand, while the point estimate of the odds ratio for *UPCAT Reading T-Score* is less than one, it can be seen that the 95% confidence interval for this factor includes the value 1.0, which means that while there is an observed negative effect on graduation status in the dataset, there is not enough data to conclude that there is a definite negative effect at the 95% confidence level.

For the bi-categorical factors, Table 4 shows that out of the 8 factors, only *Enrolled in 1st or 2nd choice degree program, Shifted within Engineering, Had Leave of Absence (LOA) or Absence without Leave (WOL),* and *Had a Scholarship* have a significant effect on graduation. It should be noted that out of the four, only *Enrolled in 1st or 2nd choice Program* is a pre-matriculation factor. The factor *Enrolled in 1st or 2nd choice Program* has an odds ratio of 3.124, which means that those who enrolled in their preferred degree programs experience an increase by 3.124 times in their graduation odds as compared to those who did not.

Furthermore, looking at the odds ratios of the significant post-matriculation factors, it can be said that students who do not undergo LOA or AWOL are 27.1 times more likely to graduate than those who do; students who have scholarships are 2.79 (1/0.3589) times more likely to

graduate than those who do not; and finally, students who shift within engineering are 7.81 (1/0.128) times more likely to graduate than those who stayed in their original degree program.

It should be noted that out of the four post-matriculation factors, three were found to be significant, with *Lived in a Residence Hall* being the only post-matriculation factor which was found to be not significant. This further reinforces the idea that including post-matriculation factors in creating models pertaining to student graduation and retention is important.

Table 4. Wald Chi Test and Odds Ratios

Continuous and Bi-Categorical Factors		
Factor	***Wald Chi (p-value)***	***Odds Ratio***
Age of Entry to UP	---	
High School GWA	6.1266 (<0.0001)	1.1331 [1.089, 1.179]
UPCAT Science T-Score	---	
UPCAT Math T-Score	3.8967 (<0.0001)	1.0214 [1.011, 1.032]
UPCAT Reading Comprehension T-Score	-2.0288 (0.0425)	0.9833 [0.967, 0.999]
UPCAT Language Proficiency T-Score	---	
Sex (Male vs Female)	---	
Ethnic Group Member (Yes vs No)	---	
Enrolled in 1st or 2nd choice Program (Yes vs No)	10.7196 (<0.0001)	3.124 [2.537, 3.847]
Dependent on UP Faculty or Employee (Yes vs No)	---	
Shifted within Engineering (No vs Yes)	-11.2708 (<0.0001)	0.128 [0.089, 0.183]
Had LOA or AWOL (No vs Yes)	25.515 (<0.0001)	27.1425 [21.06, 34.98]
Had a Scholarship (No vs Yes)	-10.3202 (<0.0001)	0.3589 [0.296, 0.436]
Lived in a Residence Hall (Yes vs No)	---	

--- Indicates Factor was not found to be significant

Table 5 presents the Wald's χ^2 statistics for the multi-categorical factors in the dataset. It can be seen from this table that all the multi-categorical factors were found to be significant, with p-values less than 0.05.

Table 5. Wald Chi Test: Multi-Categorical Factors

Factor	Wald Chi (p-value)
Year of First Enrolment	69.2976 (<0.0001)
STFAP Bracket Upon Entry	10.4175 (0.034)*
High School Province	25.5086 (<0.0001)
Permanent Address Location	28.1262 (<0.0001)
Entry Degree Program	196.3716 (<0.0001)
Religion	28.2243 (0.003)*

* Indicates Factor that does not have significant categories

It should also be noted that while the *STFAP Bracket Upon Entry* and *Religion* were found to be significant as factors in the dataset, further analysis on their individual categories did not yield significant results (p-values of all categories in the two factors were greater than 0.05), therefore, further discussion on these two factors will not be included in this study. Aside from these two factors, the rest of the multi-categorical factors were found to have significant categories, the result of which are presented below. Note that for multi-categorical factors, the factor with the least number of datapoints was set as the reference category.

Table 6 presents the Wald's χ^2 statistics and odds ratios of the categories of the *Year of First Enrollment* factor. From the table, freshman students from batch *2010* and *2013* are significant and have lower odds of graduation compared to batch *2009* with odds ratios of 0.736 and 0.3878, respectively. On the other hand, results for batch *2011* and *2012* are inconclusive with p-values greater than 0.05.

Table 6. Year of First Enrollment

Categories	Wald Chi (p-value)	Odds Ratio
2009	*Reference Category*	
2010	**-2.3081 (0.021)**	**0.7326 [0.5624, 0.9541]**
2011	0.1708 (0.8644)	1.0782 [0.4547, 2.5564]
2012	-0.3943 (0.6933)	0.8406 [0.3545, 1.9929]
2013	**-2.1139 (0.0345)**	**0.3878 [0.1611, 0.9333]**

Table 7 presents the Wald's χ^2 statistics and odds ratios of the categories of the *Entry Degree Program* factor. Out of the twelve (12) degree programs, only 5 programs were found to have a significant effect on graduation status, with p-values less than 0.05. The 5 programs are as follows: *BS Computer Science, BS Chemical Engineering, BS Computer Engineering, BS Electronics and Communications Engineering,* and *BS Electrical Engineering*, with odds ratios of 0.3982, 0.4218, 0.1769, 0.29, and 0.33, respectively. From these odds ratios, it can be said that compared to students from *BS Mining Engineering*, students from these degree programs have significantly lower odds of graduating in the College.

Table 7. Entry Degree Program

Categories	Wald Chi (p-value)	Odds Ratio
BS Civil Engineering	-1.0354 (0.3005)	0.7264 [0.3966, 1.3303]
BS Computer Science	**-3.0488 (0.0023)**	**0.3982 [0.2203, 0.7197]**
BS Chemical Engineering	**-2.8116 (0.0049)**	**0.4218 [0.2311, 0.7699]**
BS Computer Engineering	**-5.6975 (<0.0001)**	**0.1769 [0.0975, 0.3211]**
BS Electronics and Communications Engineering	**-4.0632 (<0.0001)**	**0.29 [0.1596, 0.5269]**
BS Electrical Engineering	**-3.1985 (0.0014)**	**0.3333 [0.17, 0.6535]**
BS Geodetic Engineering	-1.2472 (0.2123)	0.6621 [0.3463, 1.2657]
BS Industrial Engineering	0.4892 (0.6247)	1.1679 [0.627, 2.1755]
BS Mechanical Engineering	-0.1069 (0.9149)	0.9659 [0.511, 1.8257]
BS Materials Engineering	-0.5203 (0.6029)	0.847 [0.453, 1.5834]
BS Metallurgical Engineering	-1.184 (0.2364)	0.6556 [0.3259, 1.3188]
BS Mining Engineering	*Reference Category*	

Table 8 presents the Wald's χ^2 statistics and odds ratios of the categories of the *Permanent Address Location* factor. It can be seen from the table that the significant categories: *Quezon City*, *National Capital Region* (NCR), and *Luzon* all have odds ratios larger than 1.0, with 15.47, 15.44, and 13.63, respectively. From this it can be said that students with permanent addresses near the university campus, which is located within *Quezon City*, have higher odds of graduating in the college compared to those coming from Visayas. Furthermore, it should be noted that the nearer the address is to the campus, the higher the odds of graduation are.

Table 8. Permanent Address Location

Categories	Wald Chi (p-value)	Odds Ratio
Quezon City	**5.0222 (<0.0001)**	**15.4655 [5.3114, 45.034]**
National Capital Region	**4.9602 (<0.0001)**	**15.438 [5.2352, 45.525]**
Luzon	**4.6765 (<0.0001)**	**13.6308 [4.5607, 40.739]**
Visayas	*Reference Category*	
Mindanao	1.414 (0.1574)	4.5286 [0.5581, 36.743]

Table 9 presents the Wald's χ^2 statistics and odds ratios of the categories of the *High School Province* factor. Out of the five categories, only *National Capital Region* (NCR) was found to be significant, with a p-value of 0.0489. The NCR is the largest metropolitan area in the Philippines. This category yielded an odds ratio of 2.4374, meaning that students who studied in high schools located in the NCR have an increase in odds by 2.44 times of graduating in the college as compared to those coming from overseas.

Table 9. High School Province Location

Categories	Wald Chi (p-value)	Odds Ratio
National Capital Region	**1.9698 (0.0489)**	**2.4374 [1.0044, 5.9145]**
Luzon	1.9177 (0.0552)	2.3814 [0.981, 5.7806]
Visayas	0.4216 (0.6733)	1.2198 [0.4843, 3.0728]
Mindanao	0.7597 (0.4474)	1.442 [0.5609, 3.7074]
Overseas	*Reference Category*	

VI. CONCLUSION

This study was able to analyze graduation status using both pre-matriculation and post-matriculation factors. In analyzing the factors affecting student graduation the following were observed: Among the pre-matriculation factors, the following had a positive effect on graduation: First, being *Enrolled in 1st or 2nd choice Degree Program* has the most significant effect on graduation status, with an increase of 3.12 times in graduation odds. Additionally, it was also observed that students who have *Permanent Address Locations* near the campus are at least 13.6 times more likely to graduate than those from *Visayas*, and those who graduated from high schools in the *National Capital Region* are 2.44 times more likely to graduate than those from *Overseas*. Furthermore, it was observed that students with higher *UPCAT Math T-Scores* and *High school GWA* have higher chances of graduation. On the other hand, it was found that students who enroll in *BS Computer Science, BS Chemical Engineering, BS Computer Engineering, BS Electronics and Communications Engineering,* and *BS Electrical Engineering* are 50% less likely to graduate than students from *BS Mining Engineering*.

For post-matriculation factors, we saw that students who *Shifted within Engineering* are 7.81 times more likely to graduate than those who did not. Additionally, students who *Had a Scholarship* are 2.79 times as likely to graduate as those who did not, while students who did not undergo *Had LOA or AWOL* are 27.1 times more likely to graduate than those who did.

By performing statistical analyses on the student data, previously unseen correlations between student factors and the students' chances of graduation were identified. Further proving the importance of studies on student success for policy making. This study shall serve as a precursor for studies on student success in schools in the Philippines.

VII. RECOMMENDATIONS

This section will discuss possible changes or additions for future studies in the same field. For the dataset, it might be a good idea to include analysis of students who exited from the college outside of graduation (i.e. drop out, shifted out, transferred to other school, etc.). Some metrics that may be useful are semesters from first entry to exit and drop out or exit rates for the college. In addition to this, it may be beneficial to study year-on-year persistence of students and the effects of qualitative factors if access to the necessary data, such as semestral GWA, student organization affiliation, and other post-matriculation factors, is received. For

factor analysis, additional tests on the data, such as ANOVA, may be used to gain better insight on the relationship between input factors and the output variable, graduation status.

Finally, while this study focuses on analyzing the relationship of the individual factors to the outcome, it may be useful if a predictive model is created using the dataset, which can be used to gain insight on specific students given a subset of data. These predictive models may be trained to determine metrics such as probability of graduation, on-time graduation, and number of semesters to graduation, which can be used to help gain a better understanding on the state of the College of Engineering.

VIII. ACKNOWLEDGEMENTS

We would like to thank the University of the Philippine's Office of Admissions and the University of the Philippines College of Engineering College Secretary's Office for providing access to the data needed to conduct this study.

This study was funded by the University of the Philippines Diliman Office of the Vice Chancellor for Research and Development (OVCRD) Source of Solutions (SOS) Grant.

REFERENCES

[1] Office of Advancement of Teaching. 2016. Retrieved from https://oat.upd.edu.ph/wp-content/uploads/2016/09/API_balangkasan_091516_FINAL.pdf on 15 May 2020.

[2] Engineering Berkeley. 2020. Retrieved from https://engineering.berkeley.edu/admissions/ undergrad/faqs on 15 May 2020.

[3] College Factual. 2020. Retrieved from https://www.collegefactual.com/colleges/ massachusetts-institute-of-technology/academic-life/graduation-and-retention/ on 15 May 2020.

[4] Denoga G. 1999. EEE student mortality-performance analysis. Philippine Engineering Journal. 20(1):1-8.

[5] Issapour M, Kelly A. 2015. How student gender, SAT score, and interest in science relates to performance in introductory engineering technology coursework. 5th IEEE Integrated STEM Education Conference; Nagoya, Japan. IEEE. p. 221-224. doi:10.1109/ISECon.2015.7119928.

[6] Hahler S, Orr MK. 2015. Background and demographic factors that influence graduation: A comparison of six different types of majors. Frontiers in Education Conference; El Paso, Texas. IEEE. doi:10.1109/FIE.2015.7344306.

[7] Scalise A, Besterfield-Sacre M, Shuman L, Wolfe H. 2000. First term probation: models for identifying high risk students. 30th Annual Frontiers in Education Conference; Kansas, Missouri. IEEE. doi:10.1109/FIE.2000.897696.

[8] Zhang G, Anderson T, Ohland M, Thorndyke B. 2004. Identifying factors influencing engineering student graduation: a longitudinal and cross-institutional study. Journal of Engineering Education. p. 313–320. doi:10.1002/j.2168-9830.2004.tb00820.x.

[9] Towards Data Science. 2020. Retrieved from https://towardsdatascience.com/categorical-encoding-techniques-93ebd18e1f24 on 15 May 2020.

[10] The Analysis Factor. 2013. Retrieved from https://www.theanalysisfactor.com/strategies-dummy-coding/ on 15 May 2020.

[11] DisplayR. 2018. Retrieved from https://www.displayr.com/what-is-a-correlation-matrix/ on 15 May 2020.

[12] Statistics How To. 2013. Retrieved from https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/ on 15 May 2020.

[13] Towards Data Science. 2018. Retrieved from https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9 on 15 May 2020.

[14] Sperandei S. 2014. Understanding logistic regression analysis. Biochemia Medica. 24(1):12-18.
[15] Hosmer D, Lemeshow S. 2000. Applied Logistic Regression. 2nd Edition. Canada: John Wiley & Sons Inc.
[16] Variables and Observation. Retrieved from https://czep.net/stat/mlelr.pdf on 15 May 2020.
[17] Wikipedia. 2008. Retrieved from https://en.wikipedia.org/wiki/Logistic_function#/ media/File:Logistic-curve.svg on 15 May 2020.
[18] Nowling Lab. 2017. Retrieved from http://rnowling.github.io/machine/learning/ 2017/10/07/likelihood-ratio-test.html on 15 May 2020.
[19] The Stats Geek. 2014. Retrieved from https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/ on 15 May 2020.
[20] McFadden D. 1973. Chapter 4 - Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics. 1st edition. New York: Academic Press. p. 105-142.
[21] Forthofer R., Hernandez M., Lee E. 2007. Chapter 14 - logistic and proportional hazards regression. Biostatistics: A Guide to Design, Analysis and Discoverys. 2nd edition. Amsterdam, Boston: Academic Press. p. 387-419.
[22] National Center for Biotechnology Information. 2010. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/ on 15 May 2020.
[23] Ong D, Pedrasa JR. 2021. Student risk assessment: predicting undergraduate student graduation probability using logistic regression, SVM, and ANN. 2021 IEEE Region 10 Conference.