# Maximizing University Enrollment Using Institutional-Based Aid Scholarship

Ahmad Slim*†, Don Hush†, Tushar Ojha†, Georges El-Howayek‡, Chaouki Abdallah§ and Terry Babbitt¶
*School of Arts & Sciences – Computer Science & Math
Lebanese American University, Beirut, Lebanon
†Institute of Design and Innovation
University of New Mexico, Albuquerque, USA
‡College of Engineering
Valparaiso University, Valparaiso, USA
§School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, USA
¶Office of the President
University of New Mexico, Albuquerque, USA
Email: ahmad.slim@lau.edu.lb

***Abstract*—The overall student enrollment has declined over the past eight years in US. The administrators in enrollment management are applying ever more sophisticated analytical techniques in efforts to determine the characteristics of students that are most probable to enroll upon offering them monetary incentives (i.e., financial aid and/or scholarship). In this paper we present a framework that reveals the characteristics of such students using different machine learning models. Particularly, we use a genetic algorithm optimization method along with three different classification models: logistic regression (LR), support vector machines (SVMs) and bayesian networks (BNs). The results show that students with particular gender, ethnicity, socioeconomic and academic backgrounds are most probable to enroll upon offering them institutional money. We validated our results using data of actual students from a large public Tier 1 research university.**

***Index Terms*—Enrollment management, improving enrollment, genetic algorithm, bayesian network, logistic regression, support vector machines, scholarship**

## I. Introduction

In the last decades of this century, enrollment management emerged as a new concept and entity in colleges and universities with the goal of exercising more structured control over the constituent student population - in terms of number, characteristics and persistence post enrollment [1]. The rise of such entity can be attributed to the change in public-policy in reducing subsidies and incentives towards higher educational institutions and post-secondary students, which resulted in pushing higher educational institutions towards self generation and management of budget through tuition revenue [2]. Enrollment management offices can be seen as a cooperative organizational entity at higher educational institutions that brings together many administrative functions with the aim of optimizing student enrollment [1]. This is exemplified through the close relationship between financial and enrollment management executives in many institutions to strategize financial plans and decisions. Thus, the financial aid office turned to be considered a cornerstone, among other key offices under the purview of enrollment management[3]. Their primary purpose is to help students with financial barriers to enroll in colleges and universities by fairly allocating institutional-, state- and federal-based aid scholarship. Many efficacious papers have assessed the influence of such scholarships on students decisions toward their enrollment in college [4], [5], [6], [7], [8], [9]. For example, Dynarski in [4], using Current Population Survey (CPS) data, draws a before and after comparison between enrollment rates in Georgia to other Southern state studying the impact of Georgia Helping Outstanding Students Educationally (HOPE) Scholarship program; a merit based program. Her findings suggest that the HOPE program had an unexpectedly high effect on the college enrollment rate for students coming from middle- and high-income class. The quantitative results showed an increase in college attendance rate in Georgia by a half percentage point (from 3.7 to 4.2 percentage points) for every $1000 in aid (in 1998 dollars). Cornwell, Mustard and Sridhar carried out another study on Georgia Hope using the Integrated Post-secondary Education Data System (IPEDS) data [10]. Their results showed a 6.9 percentage point increase in the overall enrollment rate of freshman students; notably four-year institutions. Thus the money allocated for students through the scholarship programs becomes a major factor to increase the overall enrollment in colleges and universities. However enrollment management are facing a critical problem in finding the best strategy to optimally allocate the scholarship money. Particularly, people in enrollment management are struggling to find the characteristics of students whose probability of enrollment can significantly increase upon offering them scholarships. It is not hidden that some students, for many reasons, would still rather not enroll even if they are offered a scholarship. For example, some students would prefer to attend a better ranked university. Others might decide to work and earn some money before enrolling in a college. Thus, in this case and in order to maximize their overall enrollment, it would be

978-1-6654-4337-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMLA52953.2021.00196

reasonable for enrollment management to shift the scholarship money from students with very low odds of enrollment to those with higher odds. The question is how to effectively identify the characteristics of such students? In this paper we present a framework that approaches this question from a data-driven approach. In particular, we first use machine learning models to reveal the underlying function that correlates student characteristics to their enrollment decision. We then apply optimization techniques utilizing the computed underlying function in order to determine the students who are most likely to enroll upon receiving institutional aid. The machine learning models used in this work are: logistic regression (LR), support vector machines (SVMs) and Bayesian networks (BNs). As for the optimization model, we use a genetic algorithm (GA) method that fits the non-linear mixed discrete nature of the optimization problem. The remainder of this paper is organized as follows. In Section 2, we formulate mathematically the enrollment problem we introduced in this section. In Section 3, we provide a detailed discussion on student and institutional characteristics that are nationally significant and commonly cited as being highly influential in predicting enrollment. In Section 4, we go over the implementations of our proposed models. In Section 5, we present our experimental results and in Section 6, we provide some concluding remarks.

## II. Problem Formulation

As mentioned earlier, the main objective in this paper is first to define the underlying function characterizing the relationship between students' characteristics and their respective enrollment decisions. The next step is to utilize this function to find regions of the covariate space where enrollment is the most sensitive to institutional aid. In particular, we are interested in regions where increasing institutional aid from \$0 to approximately \$1500 would have the largest effect on enrollment. Here's a mathematical statement of the problem. Let $X$ be the covariate space representing student and institutional characteristics and let the covariate $x \in X$ be decomposed as follows:

$$x = (x_1; x_{inst}) \tag{1}$$

where $x_{inst}$ represents the institutional aid and $x_1$ represents all the other covariates. Define $p_e(.)$ as follows:

$$p_e(x) = p_e(x_1; x_{inst}) \tag{2}$$

the probability that students with covariate $x$ will enroll. The goal is to find the covariate value that solves the following optimization problem:

$$x_1^* \in argmax_{x_1}(p_e(x_1, 1500) - p_e(x_1, 0)) \tag{3}$$

More generally we would like to find the set of covariate values where $(p_e(x_1, 1500) - p_e(x_1, 0))$ is largest.

In the next section we present the features that define our covariate space $X$. These features capture a varied range of information on student and institutional characteristics.

## III. Data and Descriptive Statistics

Studies proved the existence of a correlation between student characteristics and college enrollment [11]. These characteristics are usually not hard to reveal with some rational analysis. For example, it is not surprising to assume that the socioeconomic situation of the parents has a direct influence on the student's decision. Studies showed that parents with high socioeconomic status would prefer their children to enroll in prestigious colleges whose tuition is often relatively expensive [12]. Another convenient characteristic that has a direct influence on choosing a college is student merit/competency-level. For example students with very high SAT/ACT scores and high school GPA would rather go to highly selective colleges [13]. These factors and many others (i.e., gender, ethnicity, etc.) are the core of our analysis in this work. On the other side, the characteristics of college play an important role in the decision of the student choosing the college to attend. Some of these characteristics are fixed and basically the college has no control over them. For example the location of the college is fixed and constant. Students usually prefer to attend colleges that are close to the places where they live [11]. Therefore the location does indeed influence the student's decision. The location has even a bigger influence on the decision of students who live outside the state. The tuition for out-of-state students is relatively higher. In this case the students, especially with low socioeconomic status, would rather choose an in-state college. This brings up another influential and important college characteristic which is the cost. The tuition & fees and other college expenses are one of the most significant factors in deciding which college to choose [14]. To compensate for that, colleges usually offer financial aid and scholarship awards which are other influencing college characteristics. The availability of certain programs, the ranking, athletic success and many others are additional college characteristics that play a major role in deciding which college to attend [15], [16].

In this work, we select a total of 15 features that encompass information characterizing students as well as institutions. The following is a list of these features along-with description:

- Gender: This is a binary variable denotes the sex of students (males or females).
- Ethnicity: It is a categorical variable; identifies the ethnicity of the student as per IPEDS reporting categories
- GPA: It is a continuous variable; provides the high school GPA of the student and is within the range 0 to 4.
- First_generation: It is a binary variable; 1 if at least one of the parent has college degree.
- Income_Level: It is a categorical variable. The income level of the parents was quantized into three categories(low-, middle- and high-income).
- Student_Income: It is a continuous variable. It denotes student income and implicitly denotes the job status of the student.
- Residency_State: It is a categorical variable having four categories. It denotes student residency status relative to

the university's physical location.

- Institutional_Money: It is a continuous variable. It denotes the financial aid provided to the student by the institution at various levels of university administration (at university level, college level, department level etc.).
- Federal_Money: It is a continuous variable. It denotes the federal financial aid allocated to eligible students. It primarily includes the Pell and Work Study money awarded to students.
- State_Money: It is a continuous variable. It denotes the financial aid awards to eligible students funded by the state.
- Appl_Decision_Diff: It is a discrete variable. It is a measure of the time it takes, in days, between receipt of application and notification of admission decision to the student.
- Apply_Afeb: It is a binary variable. It denotes the month of application - 1 if the application was submitted after February of the prior academic year, else 0.
- Fafsa_ADeadline: It is a binary variable. It denotes the timeline of submission of Free Application for Federal Student Aid (FAFSA) forms by students with respect to the insitutional FAFSA deadline.
- Father_Educ_Level/Mother_Educ_Level: It is a categorical variable. It denotes the highest education level of student's father/mother (parent). It has three labels: 0 indicating middle school level; 1 indicating high school degree; 2 indicating a college (or beyond) degree.

## IV. Models

As mentioned earlier, in this work, we approach the problem of maximizing enrollment from an optimization perspective. Particularly, we use a GA technique to solve the optimization function introduced in (3). This is driven by the fact that (3) is a non-linear mixed-discrete optimization problem and hence, this makes GA a suitable model to solve it. On the other side, we implement the probability function in (3), $p_e(.)$, using the machine learning classification models mentioned earlier: LR, SVMs and BNs. That is we approach the conditional probability of enrollment, $p_e(x)$, from a classification standpoint: our objective is to estimate the probability of enrollment for a cohort of students given their characteristics. We train these models, compare their performance and use them to compute $p_e(x)$.

### A. LR

The relationship between a binary dependent variable $y$ and an independent variable(s) $x$, can be determined using a two class LR (a fundamental classification model) [17]. The independent variable(s) can be continuous or categorical. Success and fail are the two classes for the binary LR dependent variable $y$. Often, 1 is used for indicating success and 0 for indicating failure. We denote student enrollment as 1, and 0 otherwise. The LR model can be written as

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}} \tag{4}$$

where noitemsep,topsep=0pt

$\pi$ is logit function denoting the probability of being in class 1.

$\beta_i$ is the regression coefficient of the $i$-th predictor.

In LR models, the output is interpreted as a probability ratio. In our case, it is the probability of a student to enroll given a student characteristics . However, logistic regression suffers from the drawback that it is not capable of capturing possible nonlinear structures in the data.

### B. SVMs

SVMs provide the ability to use more complex kernel functions, like the widely used RBF kernel, in place of the linear kernel. Optimization of the kernel parameter ($\sigma$) is required by an RBF kernel. This parameter ($\sigma$) regulates the decision boundary's curvature. Margin maximization classification criterion is employed by SVM. The estimated error in SVM is based on the sign of the following linear model:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{5}$$

Given this hyperplane, the SVM model aims at minimizing the estimated classification error while simultaneously maximizing the region between the hyperplanes $\mathbf{w}^\top \mathbf{x} + b = \pm 1$.

### C. BN

Unlike LR and SVMs, BNs facilitate learning about causal relationships between variables. This allows us to visualize the interactions between such variables. Technically, a BN is defined as a graphical representation of the joint probability distribution for a number of random variables. It is made up of two main parts: a directed acyclic graph (DAG) representing the causal relationship among these variables and a set of conditional probability tables (CPTs) quantifying this relationship [18]. The nodes in the DAG represent the random variables and the directed edges represent the causal relationships. However, such networks are well known to be limited when it comes to applications with continuous variables. So, generally discretizing these variables is a common practice in BN models. Given these characteristics of a BN, we decided to implement this model in this paper not only to compute the conditional probability of enrollment, $p_e(x)$, but also to visualize how different student and college characteristics influence the enrollment decision of students.

## V. Experimental results

We leveraged actual university data for our experimentation to empirically validate our proposed framework's performance. We created a dataset comprising all the First Time Full Time (FTFT) students admitted to the university between the years 2011 to 2018. The size of this dataset is 54,678 rows. We represented the various features as per our experimentation needs. Thereon, we trained our models, tested their performance and solved the optimization problem introduced in (3) using this dataset.

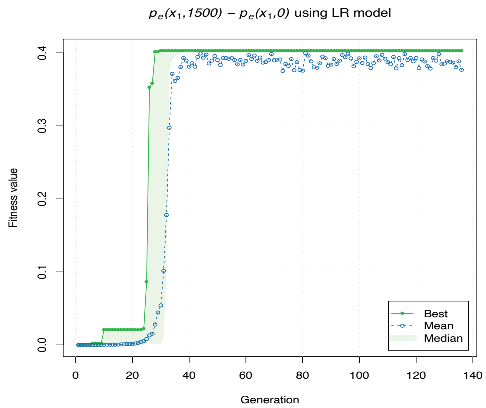


Fig. 1. The median, mean and best fitted values for each of the 200 generation generated by the GA to solve the optimization problem of (3) using the LR model.

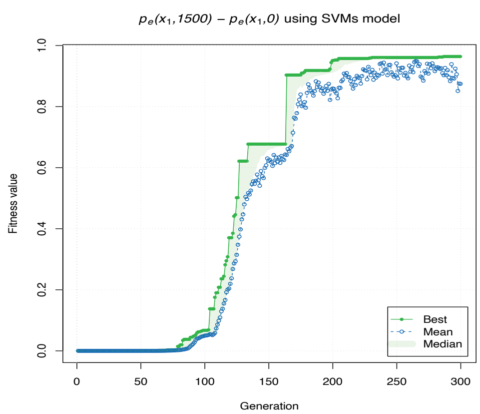


Fig. 2. The median, mean and best fitted values for each of the 300 generation generated by the GA to solve the optimization problem of (3) using the SVMs model.

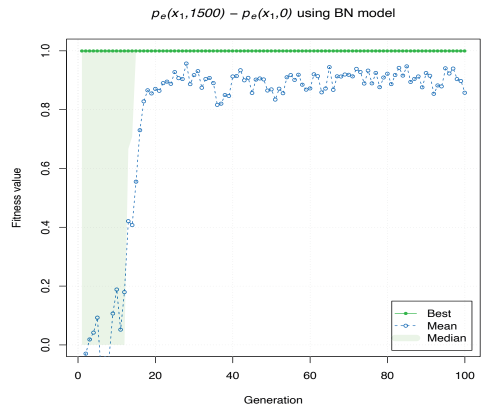


Fig. 3. The median, mean and best fitted values for each of the 100 generation generated by the GA to solve the optimization problem of (3) using the BN model.

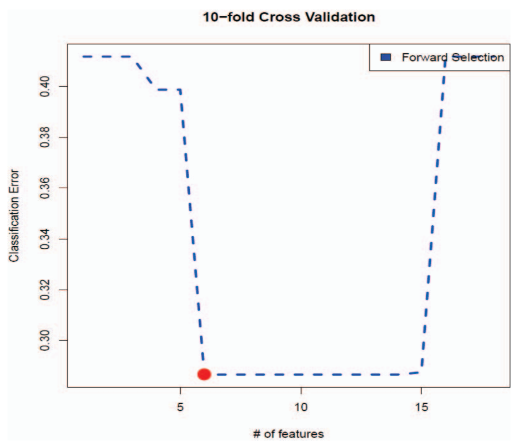


Fig. 4. The classification error of the forward variable selection method implemented using the BN model to predict enrollment.

### A. Data-preprocessing

Data preprocessing is an important step for feature representation, standardization, and normalization of the data so as to ascertain and improve the consistency of results. We carried out multiple standard preprocessing techniques widely used in machine learning experiments. We came across missing values in our dataset, i.e. many students did not have any data for multiple features. This was because of, among other reasons, the fact that many features utilize information provided by students in their admission or financial aid applications which are not always mandatorily required (for e.g. parents income is only required by financial aid applications, and thus if a student did not fill out that application, we would have a blank for this feature). Discarding the rows and/or columns containing missing values, or imputation (replacing blanks with mean, median, or mode values for the corresponding column) are two of the widely adopted techniques to resolve the missing data issue. The former option may result in discarding information, unless the size of the dataset is big enough that it does not impact the performance of models. In our work, we chose to discard the rows having missing values. The size of the resulting dataset was 37,137 rows. Each row corresponds to a unique student applicant. Thereon, we carried out standardization procedures for the discrete and continuous features in our dataset. We achieved this by subtracting the mean and dividing by standard deviation for each value of these features. This ensures that the features are on the same scale, which in turn improves the performance of the models. We employed one-hot encoding technique to transform categorical features to binary representation. This technique transforms a categorical feature having $m$ possible values into $m$ possible features, and only one feature is active for a given row.

### B. Numerical results: ***performance accuracy***

As mentioned previously, we implemented the LR, SVMs and BN models using 15 features. We used K-fold cross validation (with k=10) to asses the performance of each of the models and to compare their performances to each other . It is important to mention here that for the BN model, particularly, we used only a subset of these features. It turns out that using all the features together introduces overfitting. So to overcome this problem, we implemented a feature selection model, particularly, forward selection model. This step aims to enhance prediction results by removing irrelevant or redundant variables. Fig. 4 illustrates the results. The figure shows the classification error of the BN model for forward selection method using 10-fold cross validation. It shows that only a subset of six features (Table I) out of the 15 features used in this work is capable of achieving the best performance - a 28.7% classification error rate. The performance accuracy of the rest of the models are provided in Table II.

TABLE I
THE SIX FEATURES THAT ACHIEVE THE BEST PERFORMANCE ACCURACY OF THE BN MODEL COMPUTED USING THE FORWARD SELECTION METHOD

| Features | | |
|---|---|---|
| ETHNICITY | APPLY_AFEB | FAFSA_ADEADLINE |
| GENDER | STATE_MONEY | INSTITUTIONAL MONEY |

TABLE II
THE PERFORMANCE ACCURACY OF THE LR, SVMS AND BN MODELS USING 10-FOLD CROSS VALIDATION

| | LR | SVMs | BN |
|---|---|---|---|
| Performance Accuracy (%) | 92.1 | 98.6 | 71.22 |

Another important aspect we noticed in this work is the biasness of the dataset. The number of observations are not equal for the two classes in our dataset. The percentage of students who enrolled is 58% compared to 42% for those who did not. Thus, the prediction accuracy measure, solely, can be misleading in this case. A complementary metric that can be used to better validate the performance of the classifiers is confusion matrix. A confusion matrix is a more comprehensive technique that gives more detailed information about the type of errors made. Table III, Table IV and Table V show the confusion matrices for the LR, SVMs and BN models respectively. Using these matrices, we computed the $F-$scores of the LR, SVMs and BN models to be 0.93, 0.99 and 0.78 respectively. These results show that indeed the SVMs classification model outperforms the rest of the models.

TABLE III
THE CONFUSION MATRIX OF THE LR MODEL

| | | Actual | | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| **Predicted** | Positive | 20312 | 1484 | 21796 |
| | Negative | 1450 | 13891 | 15341 |
| | Total | 21762 | 15375 | |

TABLE IV
THE CONFUSION MATRIX OF THE SVMS MODEL

| | | Actual | | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| **Predicted** | Positive | 21593 | 203 | 21796 |
| | Negative | 304 | 15037 | 15341 |
| | Total | 21897 | 15240 | |

TABLE V
THE CONFUSION MATRIX OF THE BN MODEL

| | | Actual | | |
|---|---|---|---|---|
| | | Positive | Negative | Total |
| **Predicted** | Positive | 18685 | 7579 | 26264 |
| | Negative | 3110 | 7763 | 10873 |
| | Total | 21795 | 15342 | |

### C. *Numerical results:* ***GA solution***

The function introduced in (3) is a non-linear mixed-discrete optimization problem. So, as mentioned earlier, this makes GA a good candidate to solve it. We run the GA three separate times. In each time, we replaced the conditional probability of enrollment, $p_e(x)$, in (3) with one of the functions computed using the three classification models: LR, SVMs and BN. For each of these models, we chose the initial population of the GA to be 100 and the maximum number of generations before the GA halts to be 300. We chose the crossover and mutation probabilities to be 0.8 and 0.1 respectively. The performance of GA for the LR, SVMs and BN models is illustrated in Fig. 1, Fig. 2 and Fig. 3 respectively. These figures show that the values of the maximum difference in the probability of enrollment, upon increasing the institutional aid from \$0 to \$1500 (i.e., $p_e(x_1, 1500)-p_e(x_1, 0)$) are 0.403, 0.964 and 1.0 for the LR, SVMs and BN models respectively. The solution of the GA for the values of the covariate $x_1$ that maximizes this difference in probability are shown in

TABLE VI
THE GA SOLUTION OF THE LR, SVMS AND BN MODELS

| Feature | LR | SVMs | BN |
|---|---|---|---|
| ETHNICITY | “american indian” | “white” | “american indian” |
| GENDER | “female” | “female” | “female” |
| APPLY_AFEB | “before Feb” | “before Feb” | “before Feb” |
| FAFSA_ADEADLINE | “before deadline” | “before deadline” | “before deadline” |
| STATE_MONEY | “7199” | “306” | “0-500” |
| APPL_DECISION_DIFF | “242” | “111” | |
| FEDERAL_MONEY | “4978” | “373” | |
| STUDENT_INCOME | “17103” | “21230” | |
| FATHER_EDUC_LEVEL | “college or beyond” | “college or beyond” | |
| MOTHER_EDUC_LEVEL | “high school” | “college or beyond” | |
| GPA | “2.85” | “3.4” | |
| FIRST_GENERATION | “first generation” | “not first generation” | |
| RESIDENCY_STATE | “non-resident” | “resident” | |
| INCOME_LEVEL | “middle income” | “high income” | |

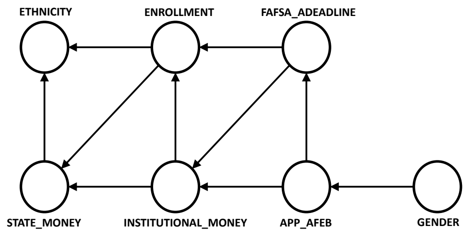


Fig. 5. The constructed BN.

Table VI. The table shows that the three models coincide on three features to maximize the probability of enrollment upon increasing the institutional aid from $0 to $1500: the student is a "female" who had applied to college "after February" and submitted the FAFSA application before deadline. We constructed a BN to better understand and visualize how these three features, besides other student and college characteristics, are related to one another. This network is shown in Fig. 5. One of the interesting observations revealed by this network is the correlation between STATE_MONEY and ETHNICITY. It is evident that some students who belong to a certain ethnic group get more state money than others. This can be explained by the different financial needs that differentiate these groups. It is also interesting to see how STATE_MONEY is related to INSTITUTIONAL_MONEY. This is reasonable as well because usually institutions adjust their institutional money offer according to the state money offer. This would guarantee, to an extent, that money is fairly distributed among students. One final observation we need to mention here is how all these variables together influence, directly and indirectly, ENROLLMENT at an institution. For example, this network shows that the enrollment is indirectly influenced by the gender of the student through the amount of the institutional money offered. It is important to mention that we automate the process of learning the structure of the BN using a score-based learning algorithm [19]. Particularly, we use the hill climbing (hc) greedy search that explores the space of the DAGs by single-arc addition, removal and reversals.

## VI. Conclusion

In this paper we present a framework that reveals the characteristics of students who are most probable to enroll upon offering them institutional money. The main goal of this work is to provide the enrollment management office at the administrators' level with a tool that help them improve their overall enrollment. We approach this problem from an optimization perspective. In particular, we use a genetic algorithm (GA) model to find the cohort of students that maximizes enrollment. In this work the GA is implemented using three different classification methods: logistic regression (LR), support vector machines (SVMs) and Bayesian networks (BN). The results show that the SVMs method achieved the best accuracy performance with a classification error of 1.2%. The GA solutions for these three classification models were slightly different. However, the three of them match up on three student characteristics that maximize their probability of enrollment upon offering them institutional aid: the student should be a "female" who had applied to college "after February" and submitted the FAFSA application "before the deadline".

## References

[1] "Enrollment management in higher education - defining enrollment management, key offices and tasks in enrollment management, organizational models," Education Encyclopedia - StateUniversity.com, 2013. [Online]. Available: http://education.stateuniversity.com/pages/1959/Enrollment-Management-in-Higher-Education.html

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

[3] D. Hossler, "The role of financial aid in enrollment management," 2000.

[4] S. Dynarski, "Hope for whom? financial aid for the middle class and its impact on college attendance," National Bureau of Economic Research, Working Paper 7756, June 2000. [Online]. Available: http://www.nber.org/papers/w7756

[5] M. Hurwitz, "The impact of institutional grant aid on college choice," *Educational Evaluation and Policy Analysis*, vol. 34, no. 3, pp. 344–363, 2012.

[6] J. Kim, "Exploring the relationship between state financial aid policy and postsecondary enrollment choices: A focus on income and race differences." *Res High Educ*, vol. 53, pp. 123–151, 2012.

[7] T. J. Kane, "A quasi-experimental estimate of the impact of financial aid on college-going," National Bureau of Economic Research, Working Paper 9703, May 2003. [Online]. Available: http://www.nber.org/papers/w9703

[8] L. L. Leslie and P. T. Brinkman, "Student price response in higher education: The student demand studies," *The Journal of Higher Education*, vol. 58, no. 2, pp. 181–204, 1987. [Online]. Available: http://www.jstor.org/stable/1981241

[9] A. Slim, D. Hush, T. Ojha, and T. Babbitt, "Predicting student enrollment based on student and college characteristics," in *Proceedings of the 11th International Conference on Educational Data Mining*. Buffalo, NY: ERIC, 2019.

[10] C. Cornwell, D. B. Mustard, and D. J. Sridhar, "The enrollment effects of merit-based financial aid: Evidence from georgia's hope program," *Journal of Labor Economics*, vol. 24, no. 4, pp. 761–786, 2006. [Online]. Available: http://www.jstor.org/stable/10.1086/506485

[11] D. W. Chapman, "A model of student college choice," *Higher Education*, vol. 52, no. 5, pp. 490–505, 1981.

[12] M. Loveheim and C. L. Reynolds, "The effects of housing wealth on college choice: Evidence from the housing boom," *Human Resources*, vol. 48, no. 1, pp. 1–35, 2012.

[13] K. H. kim and M. F. Hull, "Effects of motivation, act/sat, gpa, and ses on college choice for academically advanced students and other students," *World Journal of Educational Research*, vol. 2, no. 2, pp. 140–167, 2015.

[14] J. kim, S. L. DesJardins, and B. P. McCall, "Exploring the effects of student expectations about financial aid on postsecondary choice: A focus on income and racial/ethnic differences," *Research in Higher Education*, vol. 50, no. 8, pp. 741–774, 2009.

[15] A. Griffith and K. Rask, "The influence of the u.s. news and world report collegiate rankings on the matriculation decision of high-ability students: 1995-2004," *Cornell Higher Education Research Institute*, pp. 1–24, 2005.

[16] D. G. Pope and J. C. Pope, "Understanding college application decisions: Why college sports success matters," *Sports Economics*, vol. 15, no. 2, pp. 107–131, 2014.

[17] H.-A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *Korean Society of Nursing Science*, vol. 43, no. 2, 2013.

[18] M. Horny, "Bayesian networks," Boston University School of Public Health, Tech. Rep., 2014.

[19] D. Heckerman, "Learning in graphical models," M. I. Jordan, Ed. Cambridge, MA, USA: MIT Press, 1999, ch. A Tutorial on Learning with Bayesian Networks, pp. 301–354.