

A Machine Learning Approach for Graduate Admission Prediction

Amal AlGhamdi
Abdulaziz University
Information Systems Dept.
Jeddah
aalghamdi12094@stu.kau.edu.sa

Hanadi AlMshjary
Abdulaziz University
Information Systems Dept.
Jeddah
halmshjary@stu.kau.edu.sa

Amal Barsheed
Abdulaziz University
Information Systems Dept.
Jeddah
abarsheed0002@stu.kau.edu.sa

Hanan AlGhamdi
Abdulaziz University
Information Systems Dept.
Jeddah
hsaalghamdi@kau.edu.sa

ABSTRACT

With the increase in the number of graduates who wish to pursue their education, it becomes more challenging to get admission to the students' dream university. Newly graduate students usually are not knowledgeable of the requirements and the procedures of the postgraduate admission and might spent a considerable amount of money to get advice from consultancy organizations to help them identify their admission chances. However, giving the limited number of universities that can be considered by a human consultant, this approach might be bias and inaccurate. Thus, in this paper, a machine learning approach is developed to automatically predict the possibility of postgraduate admission to help graduates recognizing and targeting the universities which are best suitable for their profile. This paper evaluates three learning strategies of regression to predict the university rate given the students' profile; namely, linear regression, decision tree, and logistic regression model. This paper evaluates, these models to select the best model in terms of the highest accuracy rate and the least error. Logistic Regression model shows the most accurate prediction in our experiments, and hence, we suggest employing this model to predict the future applicant's university chance of admission.

CCS Concepts

• Information systems → Information systems → Decision support systems → Data analytics

Keywords

University rating; chance of admission; machine learning; predictive analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
IVSP 2020, March 20–22, 2020, Singapore, Singapore
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7695-2/20/03...\$15.00
DOI: <https://doi.org/10.1145/3388818.3393716>

1. INTRODUCTION

Prospective graduate students usually face several challenges identifying universities while applying to master's programs. Most students have difficulty in the graduate admission program as fellow applicants might misinform them or because they are not fully aware of the universities' requirements. Besides, shortlisting all possible universities and their requirements would take much time and might result in the student's missing out the admission deadline. This paper discusses the application of three machine learning algorithms as a method of predicting potential university ratings to admit the students by using data from Kaggle. We compare the performance between the models; Linear Regression Model, Decision Tree Model, and Logistic Regression Model in terms of the least error and suggest some direction of further investigations. This paper is structured as follows. Section II discusses the problem of students who want to study a master's and how we can help them through using machine learning to solve the problem and presents some related works. Section III discusses the methods used to predict data and explores three different machine learning algorithms, including the Linear Regression Model, Decision Tree Model, and Logistic Regression Model. Section IV discusses the experimental setup in terms of the dataset, features, correlations and data visualization. Section V outlines the implementation and evaluation approaches, while section VII presents and discusses the results of the evaluated models. Section VIII concludes the paper and suggests some directions for future works.

2. LITERATURE SURVEY

Many aspiring graduate students want to complete their studies, prepare for the next stage, which is a master's degree. Many of them may wonder about the basic requirements for admission to universities, and about the universities where they can be admitted based on their requirement [1].

The literature contains several studies that perform statistical analyses on admissions decisions. For example authors in [2], presents an expert system, called PASS, in which Logistic Regression is used to predict the potential of high school students in Greece to pass the national exam for entering higher education institutes. The authors in [3] used predictive modeling to assess admission policies and standards based on features like GPA score, ACT score, residency race, etc. Limitations of this research include not taking into consideration other important factors such as past work experience, technical papers of the students, etc.

These researchers' authors in [4] have used data mining and ML techniques to analyze the current scenario of admission by predicting the enrolment behavior of students. They have used the Apriori technique to analyze the behavior of students who are seeking admission to a particular college. They have also used the Naïve Bayes algorithm which will help students to choose the course and help them in the admission procedure. In their project, they were conducting a test for students who were seeking admissions and then based on their performance, they were suggesting students a course branch using Naïve Bayes Algorithm. But human intervention was required to make the final decision on the status.

3. METHODS

3.1 Linear Regression Model

Linear Regression is a supervised learning machine learning algorithm and one of the most well-known algorithms in machine learning and statistics. Linear Regression is an attractive model for researchers because its representation is simple, and it works well for many problems. Learning algorithms are used to estimate the coefficients of the LR model [5].

The objective of Linear regression model is to figure out the relationship between two variables by fitting a linear model to the training data. It is a predictive algorithm that provides a Linear relationship between Prediction (Call it 'Y') and Input (Call it 'X'). The simplest form of the regression model with one feature and one target variable is defined by the formula:

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (\sigma(\theta^T \cdot \mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

$$y = c + b * x,$$

Where y is the target variable value, c is a y-intercept, b is the slope, and x is the value of the feature variable [5]. To train a Linear Regression model, you need to find the value of θ that minimizes the RMSE by the Equation MSE cost function for a Linear Regression model:

$$\text{MSE}(\mathbf{X}, \mathbf{h}_\theta) = \frac{1}{m} \sum_{i=1}^m (\theta^T \cdot \mathbf{x}^{(i)} - y^{(i)})^2$$

Where h_θ is the hypothesis function using the model parameters θ , m = number of samples in dataset, θ^T is the transpose of θ , \mathbf{x} is the instance's feature vector, $\theta^T \cdot \mathbf{x}^{(i)}$ is the dot product of θ^T and $\mathbf{x}^{(i)}$, and y = expected value [6].

3.2 Decision Tree Model

A decision tree is a machine learning model that uses a tree-like graph of decisions. Decision trees are widely used algorithms in data mining and machine learning because of their simplicity and the ease of their interpretation. Decision Tree algorithms can model non-linear relationships [7]. The cost function that the algorithm tries to minimize is given by the following formula:

$$J(K, L_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

Where

$\begin{cases} G_{\text{left/right}} \text{ measures the impurity of the left/right subset.} \\ m_{\text{left/right}} \text{ is the number of instances in the left/right subset.} \end{cases}$

The idea is really quite simple: the algorithm first splits the training set in two subsets using a single feature k and a threshold t_k (e.g., petal length ≤ 2.45 cm). It searches for the pair (k, t_k) that produces the purest subsets (weighted by their size) [6].

To train a Linear Regression model, you need to find the value of θ that minimizes the RMSE by the Equation MSE cost function for a Linear Regression model:

$$\text{MSE}(\mathbf{X}, \mathbf{h}_\theta) = \frac{1}{m} \sum_{i=1}^m (\theta^T \cdot \mathbf{x}^{(i)} - y^{(i)})^2$$

Where h_θ is the hypothesis function using the model parameters θ , m = number of samples in dataset, θ^T is the transpose of θ , \mathbf{x} is the instance's feature vector, $\theta^T \cdot \mathbf{x}^{(i)}$ is the dot product of θ^T and $\mathbf{x}^{(i)}$, and y = expected value [6].

3.3 Logistic Regression Model

Logistic Regression (also called Logit Regression) is another technique originates from the field of statistics. It is widely used for binary classification problems (problems with two class values) and commonly used to estimate the probability that a data sample belongs to a particular class. Below is an example logistic regression formula:

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$

Logistic regression is often used with regularization techniques to prevent overfitting. To use MSE with the partial derivatives of the cost function with regards to the j^{th} model parameter θ_j is given by Equation [6]:

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (\sigma(\theta^T \cdot \mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

4. EXPERIMENTAL SETUP

The data was taken from a sample dataset on Kaggle (GRADUATE ADMISSION 2) [8]. The dataset contains several parameters that are considered important during the application for Masters Programs. Here are two class labels— Chance of Admit labeled out of 5. The dataset had 401 data points, each labeled Chance of Admit either 0 or 1. The dataset included various important features (see Table 1).

Fortunately, the dataset contains no missing value and no categorical values, therefore no preprocessing has been performed. Table 2 shows the correlation between the dataset attributes while table 3 shows the histograms of each attributes and figure 1 presents the frequency values for each attribute.

Table 1. Table of description data

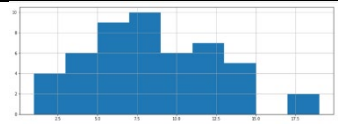
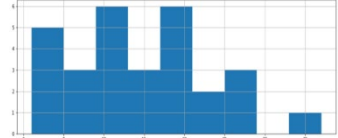
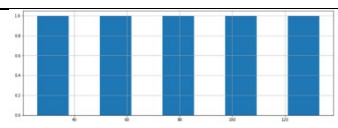
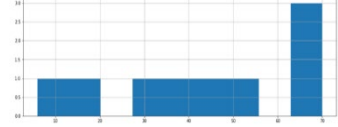
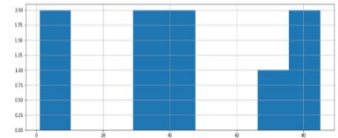
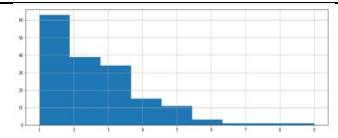
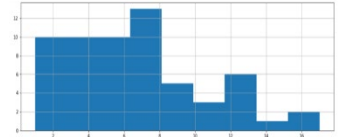
Column	description
GRE Score (out of 340)	The graduate record examination (GRE) is a standardized exam for measuring student's aptitude for abstract thinking
TOEFL Score (out of 120)	Test of English as a Foreign Language (TOEFL) score
University Rating (out of 5)	is the rankings of the higher education institute
SOP (out of 5)	A Statement of Purpose is an essay of purpose of applying to a specific course in a particular university.
LOR (out of 5)	A Letter of Recommendation (LOR) from a professional who has taught a student.
CGPA (out of 10)	CGPA is the average grades obtained by a student in all the semesters
Research (either 0 or 1)	"Research Experience" means any academic research activity.
Chance of Admit (ranging from 0 to 1)	Is the likelihood of a student's chance to be admitted to a university. 1 means great possibility, 0 means no chance of admission.

❖ There was a total of nine features, seven were numeric.

Table 2. Correlation coefficients between the dataset attributes

GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
GRE Score	-0.097526	1.000000	0.835977	0.668976	0.612831	0.557555	0.833060
TOEFL Score	-0.147932	0.835977	1.000000	0.695590	0.657981	0.567721	0.828417
University Rating	-0.169948	0.668976	0.695590	1.000000	0.734523	0.660123	0.746479
SOP	-0.166932	0.612831	0.657981	0.734523	1.000000	0.729593	0.718144
LOR	-0.088221	0.557555	0.567721	0.660123	0.729593	1.000000	0.670211
CGPA	-0.045608	0.833060	0.828417	0.746479	0.718144	0.670211	1.000000
Research	-0.063138	0.580391	0.489858	0.447783	0.444029	0.396859	0.521654
Chance of Admit	0.042336	0.802610	0.791594	0.711250	0.675732	0.669889	0.873289

Table 3. Table of frequency for each attribute

Attribute	Histogram
GRE Score (out of 340)	
TOEFL Score (out of 120)	
University Rating (out of 5)	
SOP (out of 5)	
LOR (out of 5)	
CGPA (out of 10)	
Research (either 0 or 1)	

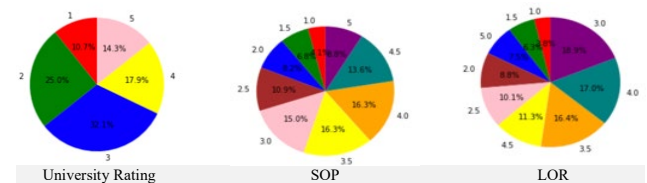


Figure 1. Feature frequency chart pie

In data GRADUATE ADMISSION, we used correlation to show how the features correspond with the output, and which one set of data may correspond to another set (see Table2). GRADUATE ADMISSION's Correlation Coefficient indicates the relationship between two quantities. It gives the measure of the strength of the association between two variables. The value of the GRADUATE ADMISSION's Correlation Coefficient range between -1 to +1, 0 for no correlation, -1 means that there is a negative correlation [9] and +1 means that there is a positive correlation. For example (see Table2), First their positive correlation between University Rating and GRE Score because it is 1, Second there almost negative correlation between TOEFL Score and University rating because of it -0.169, the small value in all correlation. Finally, note there all

correlation with TOEFL Score attribute it is almost negative except the chance of admission.

It has been also used some visualizations of the dataset to get an insight into what the data looks like. Including a pie chart and a histogram. A pie chart is a good way of presenting data. In data GRADUATE ADMISSION, we used a pie chart (see Fig 1) to display frequency numbers for each attribute in data as a simple and easy-to-understand picture. A histogram in table 3 is used to summarize frequency for all attributes. In other words, it provides a visual interpretation of numerical data by showing the number of frequency data points that fall within a specified range of values. For example, (as seen in Fig 1) the University Rating chart indicates that most of the students have chosen the third and second universities where that representing (32.1% and 25.0%) respectively. Other examples are for the SOP and LOR graphs where in both graphs small portion of students have one SOP and LOR.

5. MODEL EVALUATION

In this work, we split the original dataset into training (80 %) and test (20 %) and then train three machine learning models, Logistic Regression, Linear Regression and Decision Tree model to fit the training data. Then we use the trained models to predicts the Chance of Admit. The performance of the models was measured through the MSE. All models were run on Anaconda specific (jupyter) to run code and train the three models. After finish training data and used to predict the chance of admission, then evaluation to choose the perfect model that has less error rate, so to do this evaluation, we used RMSE to choose the best. Logistic regression model mostly does not match MSE because the main reason not to use the MSE as the cost function for logistic regression is that we don't want cost function to be non-convex. If the cost function is not convex, then it isn't straightforward for the function to converge optimally [10]. After evaluation, we compare the three models Logistic Regression, Decision Tree Model and Linear Regression based on the results of RMSE. Finally, we find that the best model when use RMSE is Logistic Regression that has (0.072) is minimal rates of error than other models.

6. RESULTS AND DISCUSSIONS

We are interested in the accurate predictions, so we are using cross-validation to determine the root-mean-square error for each model. The logistic regression model has the smallest RMSE (0.072) as shown in table 4, so the logistic regression model is the best and that giving the best predictions. The logistic regression gives us a details outcome and a broadly used technique since it is beneficial, does not require as well many computational resources, it doesn't require inputs to be scaled, and it outputs well-predicted probabilities [11].

Table 4. RMSE of the evaluated algorithms

Algorithm	RMSE
Logistic Regression	0.072
Decision Tree Model	0.11
Linear Regression	0.076

7. CONCLUSION AND FUTURE WORKS

This paper outlines the possibilities of creating an algorithm that can apply to student Graduate Admission. It appears there that relationship is between all attributes (Requirements for study the postgraduate) and one attribute (chance of admit) experienced

during learning. To verify this, we implemented the algorithm uses a Linear Regression Model, Decision Tree Model, and Logistic Regression Model to see how to uses all Requirements for study the postgraduate to predict the chance of admission in different values. To classify different machine learning algorithms, the Logistic Regression Model was the algorithm that achieved the best classification prediction and the most accurate to predict the chance of admission: As it contains the smallest number of errors (7.2% RMSE) than other algorithms. So, the goal of this paper is to create software by using machine learning, especially using Logistic Regression in the future to help students can know the how the possibility of postgraduate admission in universities to help graduates in recognizing and targeting universities that have suited their requirements.

8. REFERENCES

- [1] "Admissions Requirements," Berkeley Graduate Division, [Online]. Available: <https://grad.berkeley.edu/admissions/requirements/>. [Accessed 09 november 2019].
- [2] I. Hatzilygeroudis, "PASS: An Expert System with Certainty Factors for Predicting Student Success," ResearchGate, 20 September 2004. [Online]. Available: https://www.researchgate.net/publication/221017874_PASS_An_Expert_System_with_Certainty_Factors_for_Predicting_Student_Success.
- [3] E. Roberts, "using machine learning and predictive modeling to assess admission policies and standards," 2013. [Online]. Available: http://sites.tntech.edu/weberle/wp-content/uploads/sites/87/2018/06/NSSR_2013-1.pdf.
- [4] M. M. P. K. P. S. J. Heena Sabnani, "Prediction of Student Enrolment Using Data Mining Techniques," April 2018. [Online]. Available: <https://www.irjet.net/archives/V5/i4/IRJET-V5I4408.pdf>.
- [5] J. Brownlee, "Linear Regression for Machine Learning," [Online]. Available: <https://machinelearningmastery.com/linear-regression-for-machine-learning/>. [Accessed 25 March 2016].
- [6] A. Géron, Hands on Machine Learning with Scikit Learn and Tensorflow, O'Reilly Media, 2017.
- [7] R. S. Brid, "Decision Trees," Greyatom, 26 Oct 2018. [Online]. Available: <https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb>.
- [8] mohansacharya, "Graduate Admissions 2," kaggle, 28 12 2018. [Online]. Available: <https://www.kaggle.com/mohansacharya/graduate-admissions>.
- [9] "Data Correlation can make or break your Machine Learning Project," Medium, 14 Oct 2018. [Online]. Available: <https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9>.
- [10] A. V. Kumar, "What are the main reasons not to use MSE as a cost function for Logistic-Regression," Quora, 22 Feb 2016. [Online]. Available: <https://www.quora.com/What-are-the-main-reasons-not-to-use-MSE-as-a-cost-function-for-Logistic-Regression>.
- [11] "The Logistic Regression Algorithm," machinelearning, 23 April 2018. [Online]. Available: <https://machinelearning-blog.com/2018/04/23/logistic-regression-101/>.