PREDICTIVE MODELING FOR GRADUATE ADMISSIONS USING MACHINE LEARNING TECHNIQUES

[1]I. SHANTHI, [2]Dr. K. VENKATA RAO
[1]M. TECH SCHOLAR, [2]PROFESSOR
Department Of Computer Science & Systems Engineering,
Andhra University College of Engineering (A), Visakhapatnam, India.

Abstract: Predictive modeling for graduate admissions using machine learning is used to help graduates for getting admission into the top universities in which they are desired to admit and also the university administration to give admission to the graduates based on their performance. In this paper, machine learning techniques are applied to predict the chance of admission and number candidates the university is admitting based on the graduate admission dataset. Three machine learning methods are explored including logistic regression (LR), support vector machine classifiers and linear regression regressor. These predictive models are constructed from graduate admission dataset to predict the chance of admission of a graduate in the university

IndexTerms **- Predictive modeling, Classification, Regression, Graduate Admissions dataset, Machine learning.**

I. INTRODUCTION

Currently, applying master's degree is a very expensive and intensive work. Since, there are many graduates who are applying for master's degree in other countries whether there is an admission chance in a particular university or the graduate can apply to another university. For the university administration to filter the graduates who applied for the university. So, predictive modeling using machine learning techniques are applied to the dataset in order to predict. Predictive modeling is the problem of developing a model using historical data to make a prediction on a new data where we do not have any answer. Predictive modeling is a mathematical problem of approximation where a mapping function (f) from input variables (x) to output variable(y).

$$y=f(x) \qquad \text{Eq.1}$$

The modeling algorithm finds the best mapping function that can give time and resources available [1]. Based on the mean errors we can predict the model but in this paper, classification predictive modeling metrics and regression predictive modeling metrics are used on the dataset for prediction. Classification is the problem of predicting a discrete value whereas regression is used to predict the continuous value. An algorithm that is capable of learning a classification predictive model is known as classification algorithm whereas an algorithm that is capable of learning regression predictive model is called regression algorithm.

In Classification predictive modeling, based on classification accuracy given by the confusion matrix for the model is calculated for making a prediction. In Regression modeling based on regression metrics such as adjusted R square, mean square error, root mean squared error etc., are used for prediction. The main objective of the study is to explore the feasibility of applying machine learning algorithms to graduate admissions dataset and developing predictive models that will help in predicting the chance of admission. The objective of the study are: (1) to prepare a dataset for training the predictive models; (2) to develop different predictive models using machine learning algorithms such as linear regression, logistic regression, support vector machine in R ; (3) to evaluate and select most appropriate predictive model through their accuracies and some of the regression metrics.

In this paper, Section II describes the related work; Section III describes the methodology; Section IV describes about the dataset and performance measurements used and Section V describes the results obtained.

II. RELATED WORK:

The predictive modeling is used to develop a model using historical data to find an answer [1]. In order to predict a model we need to train the data and then algorithms are applied to it so that the model builds the relationships between independent and dependent variables [2]. The models are developed so that we can make predictions [3]. The models are compared based on the performance metrics so that the best model was chosen [4][10].

III. METHODOLOGY:

The methodology consists of machine learning techniques used for predicting and the flow of the system.

3.1 MACHINE LEARNING TECHNIQUES

The machine learning techniques used are linear regression, logistic regression, support vector machine these are supervised learning algorithms which are used to model the data and relationship between the dependent and independent variables occur which are used to model the data and predict the results [9].

3.1.1 LINEAR REGRESSION:

Linear regression is one of the most commonly used predictive modeling techniques. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variables i.e., independent variables so that you can use this regression model to predict the Y when only the X is known.

$$Y=\beta_0 + \beta_1 X \qquad \text{Eq.2}$$

3.1.2 LOGISTIC REGRESSION:

Logistic regression is a classification model in which the prediction n is based on discrete value. It computes the distribution between the independent variable(s) X and dependent variable Y of a boolean class by P (X|Y). Logistic regression classifies boolean class label Y as follows [2]:

$$P(Y=1|X) = \frac{1}{1+\exp(W_0 + \sum_{i=1}^{n} W_i X_i)}$$

$$P(Y=0|X) = \frac{\exp(W_0 + \sum_{i=1}^{n} W_i X_i)}{1+\exp(W_0 + \sum_{i=1}^{n} W_i X_i)}$$

3.1.3 SUPPORT VECTOR MACHINE:

Support vector machine is a supervised learning algorithm which is used for classification to predict the category of data. The main idea of svm is to create a hyperplane between two classes in training data .The function to calculate the hyper plane is [2][6][7]

$$\max W(\alpha) = \max \alpha - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \langle x_i | x_j \rangle + \sum_{k=1}^{l}\alpha_k$$

The decision function is given by the formula to calculate the output as

$$f(x) = \text{sign}\left[\sum_{i=1}^{l}\alpha_i d_i K(x, x_i) + b\right]$$

The process of predictive modeling is described by the following flow chart.

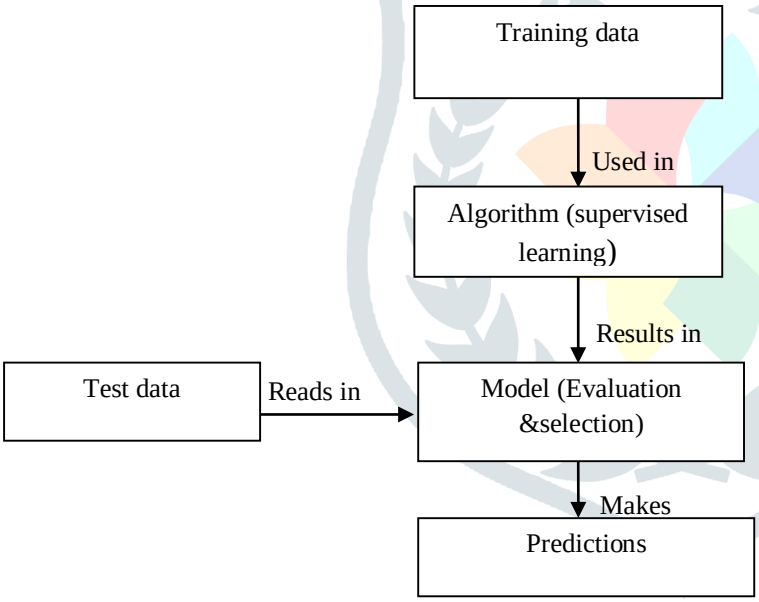

Figure 1: Process Flow

In this paper, mainly three techniques are used to predict the chance of admission. The proposed method compares the classification performance of logistic regression, support vector machine and linear regression algorithm performance [1]. The proposed process of constructing the predictive models is shown in figure 1. In the first step, the dataset is trained, the data is spit in the ratio 70:30 i.e., 70% is training data and rest as testing data. The model learns from training data only. Supervised algorithms are applied for building the model. In classification, the output has defined labels i.e., binary attributes whereas in case of regression we get a continuous value. The model learns from training data only. The model contains the learned relationships and the model builds some logic itself. Since the model contains relationships we can give the test data to it in order to make predictions and can calculate accuracy of the model.

IV. DESCRIPTION OF DATASET AND PERFORMANCE MEASUREMENTS:

4.1 Dataset:

This dataset was built from UCLA graduate dataset with the purpose to help students in shortlisting universities with their profiles. The output which is predicted gives an idea for their admission chance in particular university. The Graduate Admissions dataset consists of 400 instances and 9 attributes collected from kaggle platform. The dataset consists of the attributes like Serial No, GRE Score, TOEFL Score, CGPA (Undergraduate Cumulative GPA), SOP (Statement of Purpose), LOR (Letter of

Recommendation), University Rating, Research, Chance.of.Admit [5]. The data set is divided into two groups, one for training and another for testing. The ratio of training and testing is 70% and 30% respectively. These are built on R Studio.

4.2 Performance Measurement:

In this paper, the performance of the proposed method is measured by accuracy, sensitivity, specificity, MSE, RMSE, MAE, R^2 and MSE [8] [10].

Accuracy (ACC) is the overall success rate of the classifier defined as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FT+FN}$$

Sensitivity is the proportion of actual true positive that got predicted as positive.

$$\text{Sensitivity/Recall}=\frac{TP}{TP+FN}$$

Specificity is the proportion of actual negatives which got predicted as negative.

$$\text{Specificity}=\frac{TN}{TN+FP}$$

Regression Metrics such as [10]

Mean Square Error measures average squared error of our predictions. It calculates square difference between predictions and the target variable and average of those values.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(yi - yi`)^2$$

Root Mean Square Error is the root of the MSE. It is used to make scale of errors to be same as the scale of dependent/target variable.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error is used to calculate the average of absolute differences between the target values and predictions. It calculates all the individual differences are equally weighed.

$$\text{MAE}=A = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2$$

R^2 (R Squared) is called coefficient of determination is used to evaluate how good the model is. The value of R^2 lies between –α to 1. It gives the ration between how the predicted model is and naïve mean model

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

The MSE (Mean Square Error) baseline is the naïve model i.e., the simplest possibility model we can get.

$$\text{MSE (baseline)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The MAPE measures the percentage of error that measure to absolute error

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{|y_i|} * 100$$

V. RESULTS AND DISCUSSION

The results of the dataset and the predictive models which are obtained after applying machine learning techniques are discussed.

5.1 Correlation Matrix of the Dataset:

A correlation matrix is a table which consists of correlation coefficients for a set of variables which is used to determine the relationship between the variables. The number measures the percentage of variable fluctuation in one variable is explained by other. A correlation of 1 means the variable is in perfect unison whereas -1 indicates opposite relationship and 0 indicates no relationship between the variables.

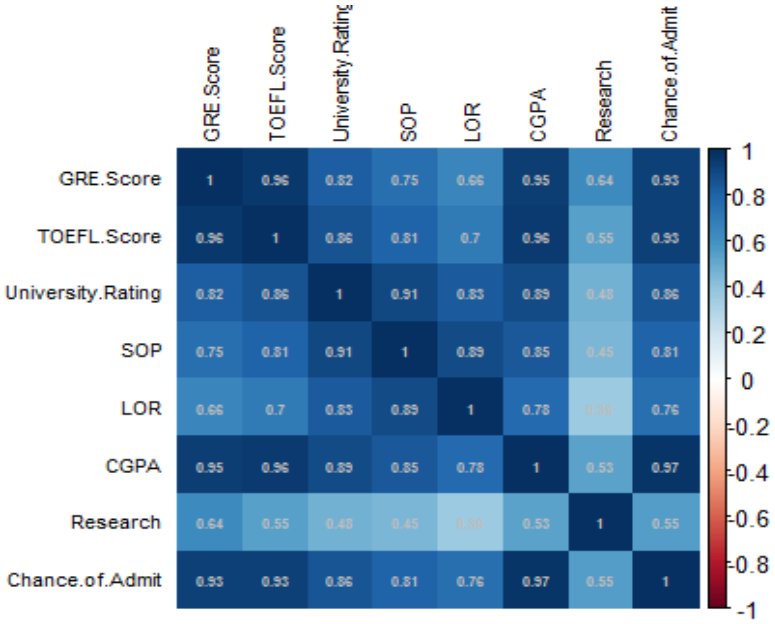

Figure 2: Correlation matrix of the dataset

5.2 Descriptive statistics:

Table 1: Descriptive statistics of the dataset

Variable	Minimum	Maximum	Std. Deviation	Mean
GRE.Score	290	340	11.47365	316.8
TOEFL.Score	92	120	6.069514	107.4
University.Rating	1	5	1.143728	3.087
SOP	1	5	0.8984775	3.4
LOR	1	5	0.8984775	3.453
CGPA	6.8	9.92	0.5963171	8.5
Research	0	1	0.498362	0.5475
Chance.of.Admit	0.34	0.97	0.1426093	0.7244

The descriptive statistics measures the minimum, maximum, mean, standard deviation values of the variables from which we can know range of the variables.

Linear Model:

The performance metrics of the linear model such as mean error, root mean square error, mean absolute error, mean precision error, mean accurate and precision error obtained are below in the table

Table 2: Regression Performance Metrics

ME	-0.720243
RMSE	**0.7320644**
MAE	0.720243
MPE	-2201.599
MAPE	8448.466
R^2	**0.8013**

Table 3: Comparison of different classification algorithms

Algorithm	Accuracy	Sensitivity	Specificity
Logistic Regression	0.9339	0.66667	0.95536
Support Vector Machine	0.9833	1.00000	0.98182

V1. CONCLUSIONS

The above results are obtained by applying different machine learning algorithms. Three machine learning algorithms such as linear regression, logistic regression, support vector machine are applied to the dataset to predict the which model is better and the decision tree classification algorithm gives the best result compared to other algorithms implemented on the dataset under same conditions. The linear regression gives adjusted R^2 as 80.1 % whereas the classification algorithms such as logistic regression, support vector machine classifiers gives confusion matrix accuracies as 93.3%, 98.3% respectively.

REFERENCES:

[1] https://machinelearningmastery.com/how-machine-learning-algorithms-work

[2] Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques, Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, The 2016 Management and Innovation Technology International Conference MITiCON,2016

[3] Machine learning predictive models for improved Acoustic Disdrometer, Ma. Madecheen S. Pangaliman, IEEE, 2018.

[4] An Artificial Immune Recognition System-based Approach to Software Engineering Management: with Software Metrics Selection Xin Jin 1 , Rongfang Bie 1* , and X. Z. Gao 2, Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06),IEEE,2006.

[5] Mohan S Acharya, Asfia Armaan, Aneeta S Antony: A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

[6] Schölkopf, B. and Smola, A. J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond-, MIT Press,2002.

[7] Vapnik, V., S. Golowich, and A. Smola . Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, Cambridge, MA, pp. 281–287. MIT Press,1997.

[8] https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

[9] Y. Li, C. Szepesvri, and D. Schuurmans. Learning exercise policies for American options. Journal of Machine Learning, Research - Proceedings Track, 5:352–359, 01 2009.

[10] https://towardsdatascience.com/how-to-select right evaluation metrics.