A Comparison of Regression Models for Prediction of Graduate Admissions

Mohan S Acharya
Dept. of ECE
National Institute of Engineering
Mysuru, India
mohansacharya@ieee.org

Asfia Armaan
Dept. of ECE
National Institute of Engineering
Mysuru, India
asfiaarmaan@ieee.org

Aneeta S Antony
Dept. of ECE
National Institute of Engineering
Mysuru, India
aneetaantony_ece@nie.ac.in

***Abstract*— Prospective graduate students always face a dilemma deciding universities of their choice while applying to master's programs. While there are a good number of predictors and consultancies that guide a student, they aren't always reliable since decision is made on the basis of select past admissions. In this paper, we present a Machine Learning based method where we compare different regression algorithms, such as Linear Regression, Support Vector Regression, Decision Trees and Random Forest, given the profile of the student. We then compute error functions for the different models and compare their performance to select the best performing model. Results then indicate if the university of choice is an ambitious or a safe one.**

Keywords-Linear Regression; Support Vector Regression; Decision Trees; Random Forest; Mean Squared Error

I. INTRODUCTION

The Graduate Program is an exhaustive task that requires thorough preparations, both in terms of building a noteworthy profile and choosing universities that offer relevant programs. A majority of students applying to master's programs face difficulty in shortlisting universities either because they are not aware of university rankings or would have been misinformed by seniors and fellow applicants. This often results in students missing out on admissions and leads to a complete wastage of resources. Here, we present a Machine Learning based approach where the data is trained on a range of values, from stellar profiles to mediocre ones. After training the data, new values are fed to the system to determine the outcome. A sample profile is tested against all the four models defined earlier in order to understand the performance of each model, both theoretically and visually. We aim to bring students closer to their university of choice through a robust evaluation of their profiles. A good number of predictors and consultancy services fail in understanding the admission procedure and either suggest extremely ambitious schools or lower ranked ones. In this paper, we have included parameters that are all relevant for graduate admissions. Barring a few exceptional cases in which a student may unexpectedly fetch an admit in a top school, most of the results are as expected and give a fair idea about the selection criteria. In further sections, we explore the different models and try to understand their functioning.

II. DATASET

The dataset is available at [1]. At the time of writing this paper, the dataset has over 400 downloads and more than 2000 views. This dataset contains parameters that are considered carefully by the admissions committee. First section contains scores including GRE, TOEFL and Undergraduate GPA. Statement of Purpose and Letter of Recommendation are two other important entities. Research Experience is highlighted in binary form. All the parameters are normalized before training to ensure that values lie between the specified range. A few profiles in the dataset contain values that have been previously obtained by students. A unique feature of this dataset is that it contains equal number of categorical and numerical features. The data has been collected and prepared typically from an Indian student's perspective. However, it can also be used by other grading systems with minor modifications. A second version of the dataset will be released which will have an additional two hundred entries.

Fig. 1. First five entries of the dataset with all the parameters.

III. RELATED WORK

There are a number of predictors that evaluate a profile based on past admissions. With the scheme of evaluation changing every year and with stricter guidelines, requirements vary considerably. One significant work observed in the same direction is [2]. The classification

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

algorithms used in [2] uses data from an old format of UCLA graduate dataset. The test scores and other parameters are more suited for US student applications. In our work, we use Regression models that give a definite value between 0 and 1 which is useful in understanding a student's profile. It also helps in analysing how important a particular parameter is for the admission and greatly affects the output value when one parameter is changed. Our dataset was created for the defined problem and is original in the true sense.

IV. METHODOLOGY

The dataset is evaluated using four regression methods namely Linear Regression, Support Vector Regression, Decision Tree and Random Forest. Errors in each case are calculated and then tabulated. The model with the best performance is then chosen.

A. Linear Regression

Linear Regression is one of the simplest regression models used for prediction of results. It models the relationship between the independent variable and the dependent variable. In the case of simple linear regression, one independent variable and one dependent variable are involved. In our dataset, however, we use Multiple Linear Regression since there are several independent variables and a single dependent variable. Multiple Linear Regression [3] attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to

observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

$$y = a1x1 + a2x2 + a3x3 + \cdots + anxn \quad (1)$$

The equation represents a Multiple Linear Regression where y is the dependent variable and $x1$, $x2$,..xn are independent variables.

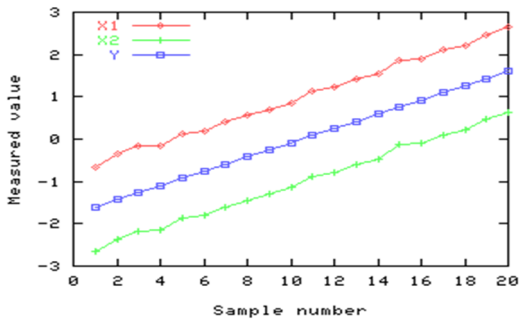

Fig. 2. Multiple Linear Regression with two variables.

We evaluate our dataset using the Linear Regression Model with a train test split of 0.25. The evaluation metrics obtained are as follows.

TABLE I. EVALUATION METRIC FOR LINEAR REGRESSION

Evaluation Metric	Value
Mean Squared Error (MSE)	0.00480149
Root Mean Squared Error	0.06929284
Mean Squared Log Error	0.00181057
R2 Score	0.72486310

R-Squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient determination, or in our case, coefficient of multiple determination since we deal with Multiple Linear Regression. A higher R-squared value generally indicates that the model fits the data better.

B. Support Vector Regression

Support Vector Machines support Linear and Non-Linear Regression commonly referred to as SVR [4]. Instead of trying to fit the largest possible street between two classes while limiting margin violations, SVR [5] tries to fit as many instances on the street while limiting margin violations. The width of the street is controlled by the hyperparameter epsilon. Here, we use the kernel trick to predict or data. The kernels used [6] in our case are RBF/Gaussian kernel and Polynomial kernel. We evaluate the metrics on the two kernels and select the one with least MSE and higher value of R2 score.

TABLE II. EVALUATION METRIC FOR RBF KERNEL

Evaluation Metric	Value
Mean Squared Error (MSE)	0.00724206
Root Mean Squared Error	0.08510033
Mean Squared Log Error	0.00267933
R2 Score	0.58501301

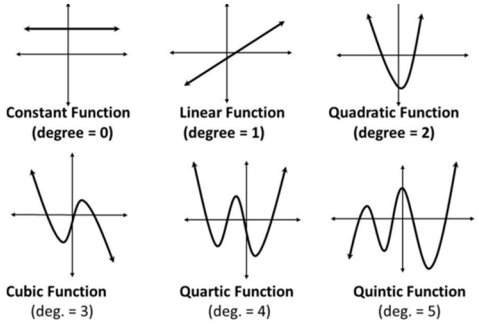

Fig. 3. Polynomial functions of degrees starting from 0 through 5

The figure above shows the graphs for polynomial functions of varying degrees. It is evident from the figure

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

that the model generally fits the curve better with increasing degree of the polynomial. Lower degree polynomial functions tend to underfit the model and therefore do not provide satisfactory results. With increasing degree of the polynomial, the curve is fit better. However, if the degree increases beyond a certain limit, the model tends to overfit and results in lack of generalization. Based on the results obtained from the table below, we demonstrate that higher degree polynomials overfit and have a larger MSE.

TABLE III. POLYNOMIAL DEGREE FUNCTIONS v/s MSE

Degree of Polynomial	MSE
Three	0.00624807
Four	0.00754321
. Five	0.00983652
Six	0.01906544
Seven	0.03423610

The MSE increases exponentially for degrees higher than eight. R-Squared score is negative for polynomials with degree greater than five. For our problem, a polynomial of degree three is chosen. The R-Squared score obtained is 0.64

C. *Decision Tree Regression*

Decision Tree is an example of CART (Classification and Regression Trees) algorithm. Decision Tree follows a top-down greedy layer approach. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision Trees can handle both numerical and categorical data.

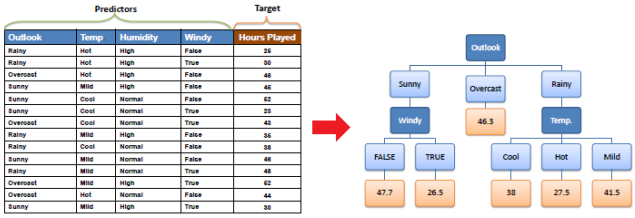

Fig. 4 Example of Decision Tree Regression

The dataset is evaluated using Decision Trees with Mean-Squared Error (MSE) being the criterion. We first evaluate the Decision Tree with maximum number of features which is the default.

TABLE IV. EVALUATION METRICS FOR DECISION TREES

Evaluation Metric	Value
Mean Squared Error (MSE)	0.01108793
Root Mean Squared Error	0.10529954
Mean Squared Log Error	0.00424281
R2 Score	0.36463222

We also considered changing the number of features to check the performance of the model. We evaluated MSE for two additional values, square root and log2.

TABLE V. NUMBER OF FEATURES v/s MSE

Number of Features	MSE
Max_Features	0.01108793
Sqrt (Max_Features)	0.00874299
Log2 (Max_Features)	0.00984356

MSE is least for a Decision Tree having only square root number of maximum features. The MSE obtained for log2 is also lower compared to a model that has maximum number of features. R2 Score for square root is 0.5 which is considerably higher than the default number of features. This is the best case scenario for Decision Tree Regression.

D. *Random Forest Regression*

The Random Forest Regression [7] is a type of additive model that makes predictions by combining decision from a sequence of base models. Each base model is a Decision Tree and the result of the Random Forest model is the cumulative output of the Decision Trees. This technique of using multiple models to obtain better predictive performance is called model ensembling. In Random Forests, all the base models are constructed independently using a different subsample of the data. The random forest model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, random forests are able to capture non-linear interaction between the features and the target.

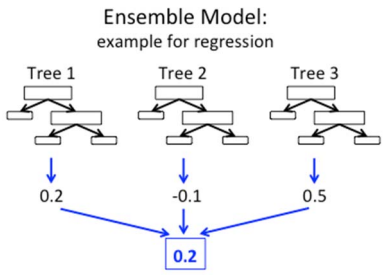

Fig. 5 Random Forest as a combination of multiple Decision Trees

We evaluated our dataset on the model for different values of estimators(Decision Trees) with depth of the tree as default value.

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

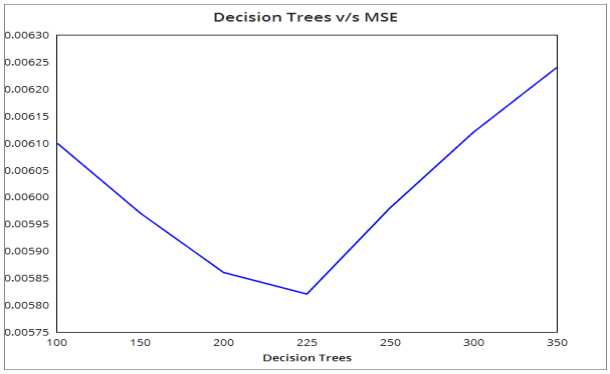

Fig. 6 Plot of number of Decision Trees v/s MSE

Generally, higher the number of estimators, better is the predictive performance of the model. However, it is evident from the plot that MSE is lowest for 225 estimators. MSE steadily increases afterwards. Therefore, we consider 225 estimators for our model. We then evaluate other metrics for this value of n.

TABLE VI. EVALUATION METRIC FOR RANDOM FOREST

Evaluation Metric	Value
Mean Squared Error (MSE)	0.00582112
Root Mean Squared Error	0.07629623
Mean Squared Log Error	0.00229122
R2 Score	0.66017010

V. CONCLUSION

After evaluating all four models on the dataset, we compare the performances to find out which model predicts better. MSE and R2 Scores are tabulated for all the models.

TABLE VII. PERFORMANCE ANALYSIS

Regression Models	MSE	R2 Score
Linear Regression	0.00480149	0.72486310
Support Vector Regression	0.00724206	0.64401301
Decision Tree Regression	0.00874299	0.50134421
Random Forest Regression	0.00582112	0.66017010

It is clear that Linear Regression performs the best on our dataset, with a low MSE and high R2 score, closely followed by Random Forest. This can be attributed to the linear dependencies of features in the dataset. Higher values of test scores, GPA and other factors generally result in greater chances of admission. The inclusion of a few outliers has influenced the Linear model to some extent.

VI. SCOPE AND FUTURE WORK

We evaluated additional profiles (unseen data) using Linear Regression, our best performing model to see how well it performs. The profile of an applicant with the test scores, starting from GRE and TOEFL, College Ranking, Statement of Purpose, Letter of Recommendation, GPA and Research Experience is displayed as follows:

[335,117,5,5,5,9.7,1] -> 0.93
[324,110,4,4,5,9.04,1] -> 0.82
[296,95,2,1.5,2,7.1,0] -> 0.43

The results obtained closely resemble the actual chances of admit. The performance of the model is clearly indicative of the fact that our algorithm also works well on unseen data.

We developed a plausible solution for the problem keeping in mind the various factors that affect the chances of admission. Although the entire admission process is subjective, we have been successful in developing an original solution that gives satisfactory results for the dataset used. As indicated, we aim to expand our dataset and increase the number of profiles with some variations. The number of outliers (profiles that do not seem impressive but had a high chance of admission) would be significantly increased to reduce the linear dependency of features. We will also use Deep Neural Networks as another plausible model to understand the subjective nature of admission.

978-1-5386-9471-8/19/$31.00 ©2019 IEEE

REFERENCES

[1] Mohan S Acharya, "Graduate Admissions", Predicting admission from important paramaters, Kaggle, April 2018. https://www.kaggle.com/mohansacharya/datasets

[2] Naman Doshi, "Predicting MS Admission", https://medium.com/data-science-weekly-dsw/predicting-ms-admission-afbad9c5c599 February, 2018.

[3] A Tutorial on Multiple Linear Regression, http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[4] Alex J.Smola and Bernard Scholkopf , "A Tutorial on Support Vector Regression" ,September 2003.

[5] A Tutorial on Support Vector Machine – Regression, https://www.saedsayad.com/support_vector_machine_reg.htm

[6] Paul Paisitkriangkrai, "Linear Regression and Support Vector Regression modules", https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf , The University of Adelaide, October 2012.

[7] Random Forest Regression,Turi Machine Learning Platform, https://turi.com/learn/userguide/supervised-learning/random_forest_regression.html

978-1-5386-9471-8/19/$31.00 ©2019 IEEE