Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

CSE4108: Artificial Intelligence Lab

Fall 2020

Project Report

US Graduate School's Admission Prediction Based on GRE Score, TOEFL Score, University Ranking, SOP, LOR, CGPA and Research

Lab Section: A1

Submitted To

Md. Siam Ansary
Department of CSE, AUST

Mr. Ashek Seum
Department of CSE, AUST

Submitted By

Samiul Islam Niloy

Student ID: 170204016

September 8, 2021

1 Introduction

The United States of America (USA) hosts the most number of international students in the world. Quality education, unique curriculum, multicultural environment, and abundant opportunities are just some of the reasons why many International students want to study in the US. The US boasts of some of the finest universities, a lot of which consistently rank in the world university rankings. American institutions are also known to have high academic standards, follow rigorous practices to maintain quality and are well-supported to be able to offer excellent education to its students. The US is a melting pot of different cultures, races and ethnicities. Its diverse environment ensures that there is acceptance among all communities and there is no room for any sort of discrimination. One will be learning with students from different regions of the world thereby making it a rich and stimulating education experience.

In Bangladesh, it is many student's dream to higher study in the US. But to be able to get the chance of admission, a student has to take some extra exams and procedures other than his university degree, such as GRE, TOEFL, SOP etc. So, having only high CGPA will not ensure his admission in US university. Also, not every parameter can be measured by numerical value such as human factors. In this project, only numerical values are considered for doing the prediction of getting chance of admit in the US graduate school.

2 A Brief Description of the Dataset

At a Glance Overview

Name of the Dataset	US Graduate Schools Admission Parameters Dataset
File Format of the Dataset	.csv
Dimension of the Dataset	500 x 9
Number of Total Columns	9
Number of Total Rows	500
Number of Feature Columns	7
Name of Feature Columns	GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research Experince
Number of Target Column(s)	1
Name of Target Columns	Chance of Admit

Description

The dataset has 9 columns and 500 rows. Of 9 columns, 7 columns are feature columns and they are GRE Score, TOEFL Score, University Ranking, SOP, LOR, CGPA and Research Experience. The last column is the target column which we are going to predict the value of and the name of that column is Chance of Admit. Below is a brief description of each columns:

Name of the Feature : GRE Score
Unit : Integer
Range : 280 - 340
Description : The Graduate Record Examinations (GRE) is a standardized test that is an admissions requirement for many graduate schools in the United States and Canada and few in other countries. The GRE exam is a two part examination. Second part is out of 340 and minimum score is 280. Whatever the score is, it will always be an integer number and never be a fraction.

Name of the Feature : TOEFL Score
Unit : Integer
Range : 90 - 120
Description : Test of English as a Foreign Language (TOEFL) is a standardized test to measure the English language ability of non-native speakers wishing to enroll in English-speaking universities. Lowest possible mark for TOEFL exam is 90 and highest possible is 120. Like GRE, the score of TOEFL exam will always

be an integer.

Name of the Feature : University Rating
Unit : Integer
Range : 1 - 5
Description : Universities are ranked from 1 to 5 where 1 being the best university and 5 being the worst (comparing to 1). For same score in GRE/TOEFL/SOP etc, The probability of getting the chance of admission will be higher in rank 5 university than that of rank 1 university.

Name of the Feature : SOP
Unit : Float
Range : 1 - 5
Description : The SOP is one's Statement of Purpose. It is a document in which one write about himself and make the first impressions to the admissions committee. It is the first part of GRE examination and is out of 5. Lowest possible score is 1 and highest is 5 out of 5. The score can be fraction but will always be divisible by 0.5.

Name of the Feature : LOR
Unit : Float
Range : 1 - 5
Description : The LOR is a Letter of Recommendation from reputable teachers and professors. A letter of recommendation for USA universities is a mandatory requirement for admission at all study levels, from bachelor to doctoral courses. Students are asked to submit two to three academic LORs or professional LORs, depending upon their choice of program. It is sometimes scored out of 5. Lowest possible score is 1 and highest is 5 out of 5. The score can be fraction but will always be divisible by 0.5 just like LOR.

Name of the Feature : CGPA
Unit : Float
Range : 2.20 - 4.00
Description : Cumulative Grade Point Average or CGPA along with GRE scores are the two most important scores for the admission in US Graduate School. Different countries measure CGPA in different methods and the range differs from one another. In Bangladesh, it is calculated out of 4.00. The minimum CGPA score to get an undergraduate degree in any university in Bangladesh is 2.20. So, the range of CGPA will be minimum 2.20 and maximum 4.00 and will be a floating number.

Name of the Feature : Research Experience
Unit : Integer
Range : 0 - 1
Description : By research experience column, it is referring to one's having the experience of working in research before. If the answer is yes, then the value will be 1 and 0 otherwise.

Name of the Target Column : Chance of Admit
Unit : Float
Range : 0 - 1
Description : After training and testing using different regression models, the final output will be this column or the target column. The values of this column are probability of getting the chance of admit in US graduate school. So, the value will be in between 0 to 1.

3 Description of the Models used in this Project

There are total 6 regression models used in this project. They are:

1. **Linear Regression**
   First model used in this project is Linear Regression Model. 70% data was used for training and 30% data was used for testing in this model.

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
   Mathematically, we can represent a linear regression as:

   $$y = a_0 + a_1 x + \epsilon$$

   Here,
   y = Dependent Variable (Target Variable)
   x = Independent Variable (predictor Variable)
   a_0 = intercept of the line (Gives an additional degree of freedom)
   a_1 = Linear regression coefficient (scale factor to each input value)
   ϵ= random error

2. **Decision Tree Regression**
   Second model used in this project is Decision Tree Regression Model. 70% data was used for training and 30% data was used for testing in this model.

   Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
Step-3: Divide the S into subsets that contains possible values for the best attributes.
Step-4: Generate the decision tree node, which contains the best attribute.
Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

3. **K-Nearest Neighbor Regression**
Third model used in this project is K-Nearest Neighbor Regression Model. 70% data was used for training and 30% data was used for testing in this model.

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
The K-NN working can be explained on the basis of the below algorithm:
Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors.
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.

4. **Support Vector Regression**
Fourth model used in this project is Support Vector Regression Model. 70% data was used for training and 30% data was used for testing in this model.

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

Linear SVR

$$y = \sum_{i=1}^{N} (a_i - a_i^*).(x_i, x) + b$$

Non-linear SVR
The kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation.

$$y = \sum_{i=1}^{N} (a_i - a_i^*).(\phi(x_i), \phi(x)) + b$$

$$y = \sum_{i=1}^{N} (a_i - a_i^*).K(x_i, x) + b$$

Kernel Functions
Polynomial

$$k(x_i, x_j) = (x_i, x_j)^d$$

Gaussian Radial Basis Function

$$k(x_i, x_j) = exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2})$$

5. **Random Forest Regression**
   Fifth model used in this project is Random Forest Regression Model. 70% data was used for training and 30% data was used for testing in this model.

   Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
   Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. The Working process can be explained in the below steps:
   Step-1: Select random K data points from the training set.
   Step-2: Build the decision trees associated with the selected data points (Subsets).
   Step-3: Choose the number N for decision trees that you want to build.
   Step-4: Repeat Step 1 & 2.
   Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

6. **Ridge Regression**
   Sixth model used in this project is Ridge Regression Model. 70% data was used for training and 30% data was used for testing in this model.

   Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the

actual values.
The cost function for ridge regression:

$$Min(\left\|Y - X(\theta)\right\|^2 + \lambda\|\theta\|^2)$$

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.
For any type of regression machine learning models, the usual regression equation forms the base which is written as:

$$Y = XB + e$$

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.
Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

4 Performance Scores of Each Model

Regression Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R^2 Score
Linear Regression	0.0414	0.0033	0.05769	81.85%
Decision Tree Regression	0.0587	0.0065	0.08085	64.35%
K-Nearest Neighbor Regression	0.0539	0.0049	0.06999	73.28%
Support Vector Regression	0.0615	0.0055	0.07406	70.09%
Random Forest Regression	0.0456	0.0040	0.06359	77.95%
Ridge Regression	0.0414	0.0033	0.05769	81.85%

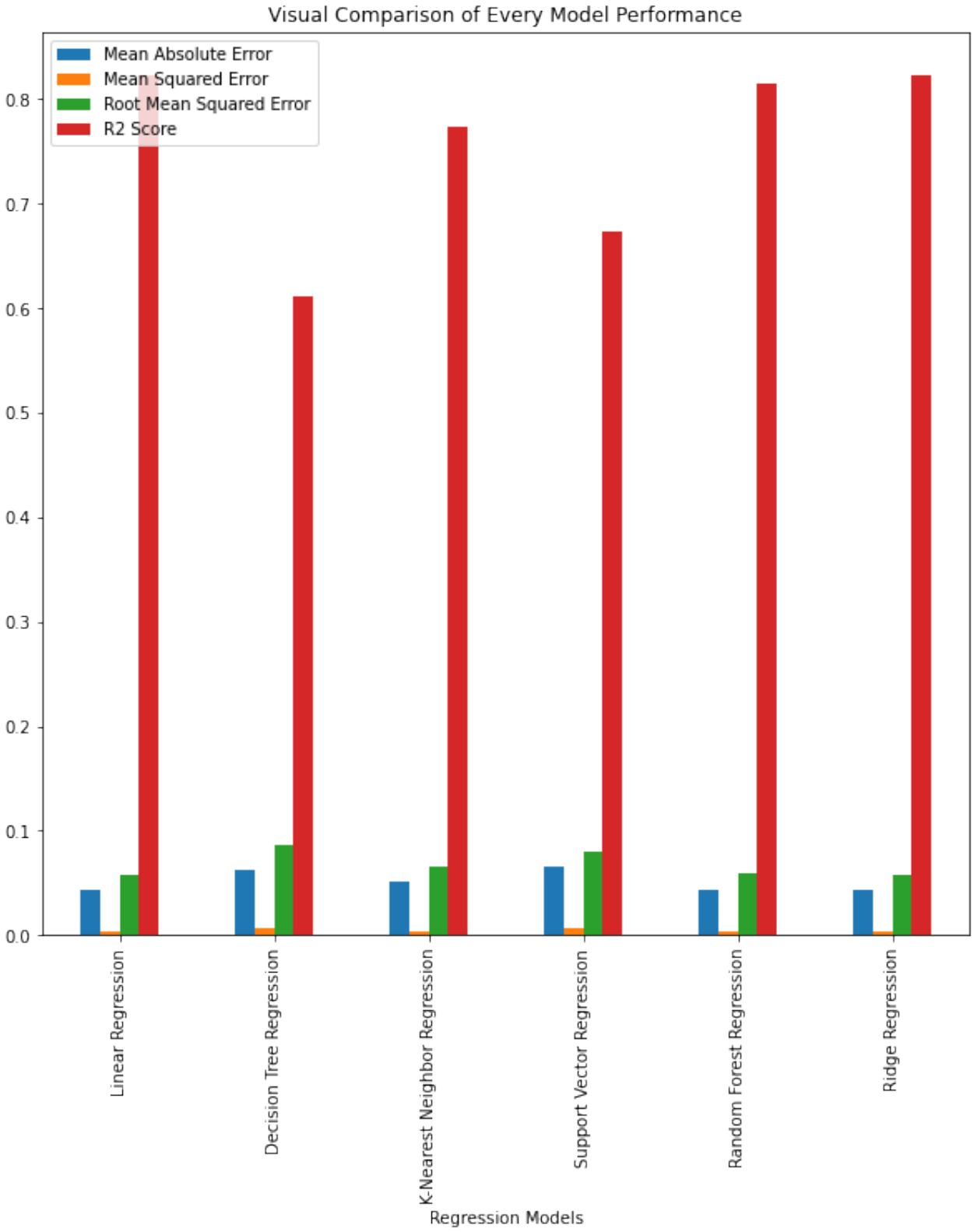

Figure 1: Visual Comparison of Every Model's Performance

5 Conclusion

Before drawing conclusion, we have to check the facts for four performance score. What kind of score makes a model better than the other? For Mean Absolute Error, a perfect MAE value is 0.0, which means that all predictions matched the expected values exactly. For Mean Squared Error, a perfect MSE value is 0.0, which means that all predictions matched the expected values exactly. For Root Mean Squared Error, a perfect RMSE value is 0.0, which means that all predictions matched the expected values exactly. Lastly, the most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

So, the lower the score of MAE, MSE, RMSE is and the higher the score of R^2 is, the better the model will be.

Now, we can come to the conclusion that, with the given performance scores for the six models, we can say both Linear Regression and Ridge Regression models are most suitable for this dataset and Random Forrest Regression is the least suitable.