University of Illinois Springfield

Machine Learning: Graduate Admission

Vamshi Bairagoni

Dec, 2022.

# Abstract

Prospective graduate students are usually at a loss when it comes to picking which colleges to go to for master's degrees. While there are many predictors and consultancies that might help a student, they aren't always trustworthy because decisions are made based on selective historical admissions. Given the student's profile, we provide a Machine Learning-based technique for comparing several regression methods, such as Linear Regression, Support Vector Regression, Decision Trees, and Random Forest. The error functions for the various models are then computed and compared in order to identify the best performing model. The results reveal if the chosen university is ambitious or conservative.

# Problem Definition and Goals

This project sought to investigate and analyze Graduate Admission data in order to extract relevant information and apply machine learning algorithms to forecast the 'chance of admit' based on several criteria. The data given provides the essential information to learn more about the chance of admits, and other metrics required to create forecasts and draw conclusions. Originally, the variables in the data were as follows:

MOHAN S ACHARYA contributed the data, which was collected through Kaggle.com.

The dataset is linked in the R notebook for this project. It consists of 500 observations and 9 variables. The following are the variables:

* Serial.No. - Unique Identifier ( Not a significant variable)
* GRE.Score - Graduate Record Examinations
* TOEFL.Score - Test of English as Foreign Language.
* University.Rating - Ranking of the university on a scale of 5.
* SOP - Statement of Purpose on a scale of 5.
* LOR - Letter of Recommendation on a scale of 5.
* CGPA - Cummulative Grade Point Average of Undergrad.
* Research - Research Experience
* Chance.of.Admit - Chance of admit range from 0 to 1.

# Related Work

Because this is a public dataset posted on Kaggle, a website famous for datasets that attract a lot of data enthusiasts, and because this dataset was posted in 2019, there has been a lot of work done on it, work that is not necessarily related to machine learning modeling, but also exploratory data analysis and visualizations. However, for this activity, the associated work was not referred to prior to the conclusion of the project in order to avoid duplicating ideas and to examine the challenge from a fresh perspective.

Citation:-

**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**

Data Exploration & Preprocessing

The dataset consists a total of 500 rows and 9 columns. A basic summary tells us that this data doesn’t have missing values. The identifier columns ‘Serial No’is removed since it would not be used for modeling purposes.

Calendar

Description automatically generated with medium confidence

Before we proceed, we must first ensure that our data is clean and valuable. The following steps will be taken: 1. Converting categorical data to factor data - Because variable Research is either 0 or 1, we will convert this column to factor data. 2. Determine whether there are any missing values in the dataset. 3. Remove an unnecessary column from our dataset. The identifier columns ‘Serial No’is removed since it would not be used for modeling purposes.

Table

Description automatically generated with low confidence

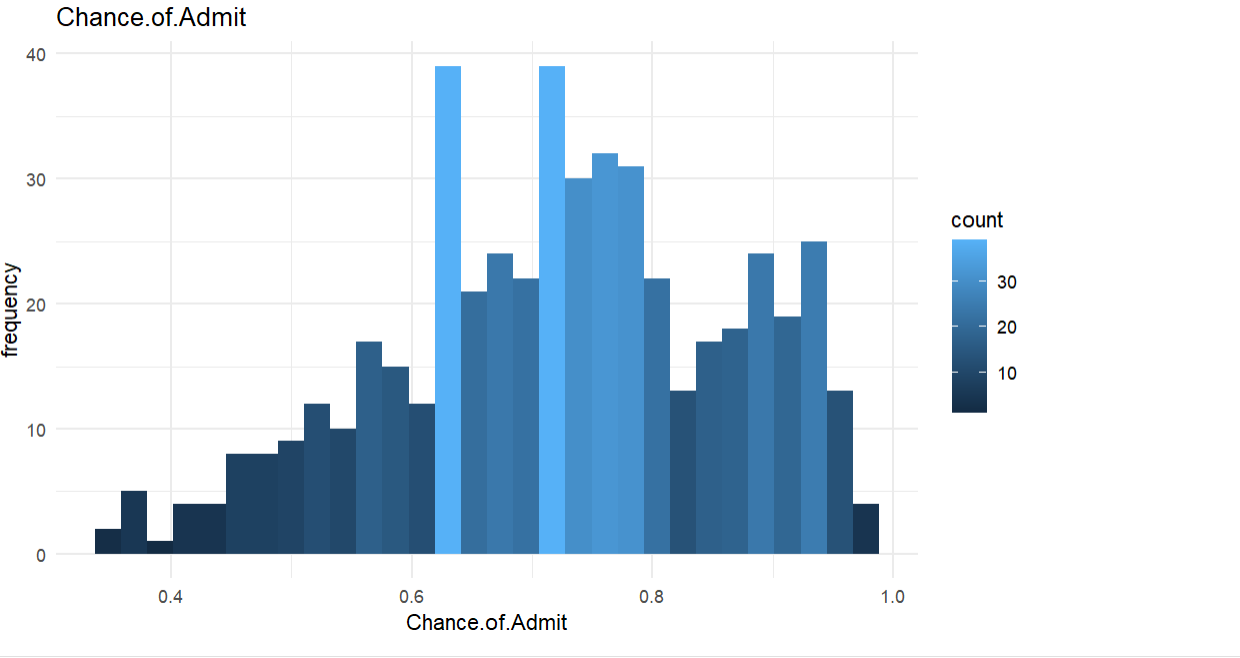
Chart, bubble chart

Description automatically generated

Correlation plot of the dataset.

Above correlation plot clearly states that all the variables has strongest correlation with the target variable chance.of.Admit.

**Histogram of target variable**:



Highest frequency count of 40 has between 0.6 and 0.8 percentile chance.

**Plots:**

**Chart, box and whisker chart

Description automatically generated**

# Data Analysis and Results

The following models were used:

* Linear Regression

-Lasso

-Ridge

-Elastic net

* Random Forest
* Gradient Boosting
* Support Vector Machine

-Linear Kernel

-Radial Kernel

* Neural Network

Every model except the neural network was tuned using the caret package and 10-fold cross-validation.

RandomForest was tuned using a tune grid of 3 values between 2 and the number of features in the dataset for mtry

The neural network was tuned manually by trying different values for the mini-batch

The results

|  |  |
| --- | --- |
| MODELS | RMSE |
| Multiple Linear Regression | 0.06374082 |
| Stepwise Backward Selection | 0.06357101 |
| Regression Tree | 0.0761262 |
| Regression with Lasso | 0.06365759 |
| Regression with Ridge | 0.06335264 |
| Regression with Elastic Net | 0.06355819 |
| Random Forest | 0.06240536 |
| Gradient Boosting | 0.06506031 |
| SVM Linear Method | 0.06444613 |
| SVM Ridge Method | 0.06292008 |
| Neural Network | 0.06585196 |

All the models behave similarly. On average, the models RMSE values are very close to each other.

Following are the summaries and observations from the above models :

Multiple Linear Regression:

Text, letter

Description automatically generated

Linear regression is a model with a great interpretability, thus we’ll interpret this simple linear regression model immediately. - Intercept-based:

Based on the coefficient or slope: When CGPA increases by one value, the Chance.of.Admit increases by around 0.12.

According to the P-value, CGPA is a significant predictor with a linear impact.

Based on R-squared values of 0.8244, the predictor chosen is enough to explain the target variable.

Regression Tree:

**Diagram

Description automatically generated**

This states the variables Grescore and Cgpa is significant for chance of admit.

Random Forest:

**Text, letter

Description automatically generated**

Here method = rf is used , with 10 fold cross validation. Pre proc function is also used for imputation and numerical data normalization.

**mtry** is the parameter that is tuned here to get best model. Mtry states how many features are randomly selected at each split.

grid\_rf <- expand.grid(mtry = c(2,4,8))

Neural Network model:

ANN Model:

A feed forward Deep neural network is built here.

Keras and Tensorflow packages are used here to build ANN model.

For building Neural Network for Regression:

1.Target variables is continuous.

2. All categorical variable should be one hot coded with binary values(or nay other method but Neural networks only takes numerical inputs) i.e Research variable.

3. All numerical variables should be scaled and normalized for best model

4. We can tune hyper parameters for that we need to split our training data in to train and validation data with 90 and 10 % split.

Training the model on neural network using 128 unites, activation function = “relu”.

And the last layer with 1 unit/neuron with a dropout rate of 0.2%

Following output has 17,793 params.

Table

Description automatically generated

Model built here has two hidden layers with first layer has 128 neurons with relu activation. Second has 128 with relu. And output layer is 1.

Optimizer used for compiling is mse and loss “Stochastic Gradient Descent”.

Model trained with batch size of 100 and epochs of 30.

**Tuning parameters:**

Multiple hyper parameters are tuned here to get the best model. Nodes, learning rate, batch size, epochs, activation, drop outs are tuned and are run to get the best model.

Graphical user interface, application

Description automatically generated

Best run has 32 Neurons for first hidden layer, 50 for second hidden layer with relu activation and tanh respectively trained for 50 epochs, with learning rate of 0.01.

By Comparing various models**, Random forest** last lowest RMSE.

NOTE : Backward Selection, Support vector machine models using linear and radial kernels are explained in the source code file. There are several works done on similar models. This project is intended to extend with **Random Forest** & **Neural Network** Models.

# Conclusion

Graduate school requirements are largely the same for many programs, with certain schools emphasizing more on one component than another, and which component that may be varies from school to school. The main topic we intended to investigate was how much these five characteristics (academics, statement, letters, research experience, and institution rating) influence chances of admission to grad school.

According to an article published on usnews.com, graduate schools want to see that you have a demonstrated track record of accomplishment in your profession, as the aim of graduate school is to build competence in a certain academic topic. GPA and standardized exams are popular admissions variables used by graduate schools to gauge an applicant's academic potential. The GRE is the most often used standardized test for admissions, and how relevant standardized test scores are to a graduate school largely depends on the topic of study. I also learnt that potential graduate schools want to know if you are a good fit for their program, thus your letters of reference may be the most crucial aspect in your application, aside from your academic qualities. Because graduate schools are frequently hyper-focused on a certain academic area, one easy approach to communicate interest is through your personal statement, both of these sites cited passion for your profession as an apparent component . My linear regression model confirmed that university ranking was not regarded essential. So it appears that the specific qualities that grad schools look for in potential students are factors such as good academic standing, relevant work experience, a strong personal statement, and strong letters of recommendation, as they can all demonstrate a deep and long-standing interest in your field of study.