Admission prediction

Pavankalyan kanakam

Feb,2019

I.Definition

Project overview

To apply for a master's degree is a very expensive and intensive work. With this idea, students will guess their capacities and they will decide whether to apply for a master's degree or not.

This is probably a question that every aspiring MS aspirant wants to know. Is my profile good enough to get a good college? Being an aspirant myself even I also have so many doubts, whether my CGPA is good enough, how should I write a solid SOP, etc .

The main goal of this problem is to predict the 'Chance of Admit' of a student in a particular university given various parameters.

As the dataset contains many entries and the required output is continues it is better to use the regression ML Algorithms.(Regression is a technique from statistics that is used to predict values of a desired target quantity when the target quantity is continuous.) and also there is no academic research has been published on this topic.

Problem Statement

• The main aim of this problem is to predict the 'Chance of Admit' with high accuracy by applying various ML Algorithms and then comparing their scores.

• Regression is a technique from statistics that is used to predict values of a desired target quantity when the target quantity is continuous.

•Here we have to predict the ‘chance of admit ‘ by using Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR ,CGPA, Research.

• Compare different models of regression to check for best model depending on r_squared score

(R-square tells the percentage of variance in dependent variable that can be explained by the independent variable.)

Datasets and Inputs

This dataset is created for prediction of graduate admissions and the dataset link is below: • https://www.kaggle.com/mohansacharya/graduate-admissions

Features in the dataset:

• GRE Scores (290 to 340)

• TOEFL Scores (92 to 120)

• University Rating (1 to 5)

- Statement of Purpose (1 to 5)
- Letter of Recommendation Strength (1 to 5)
- Undergraduate CGPA (6.8 to 9.92)
- Research Experience (0 or 1)
- Chance of Admit (0.34 to 0.97)

Solution Statement

The main aim of this problem is to predict the 'Chance of Admit' with high accuracy by applying various ML Algorithms and then comparing their scores..

In this problem we use both classification and regression models

Algorithms Considered:

- Linear Regression
- Decision Tree Regressor

Evaluation Metrics

As we are using regression models in order to find and compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term.

In this problem I will compare different models to check for best model depending on rsquared score. R-square tells the percentage of variance in dependent variable that can be explained by the independent variable.

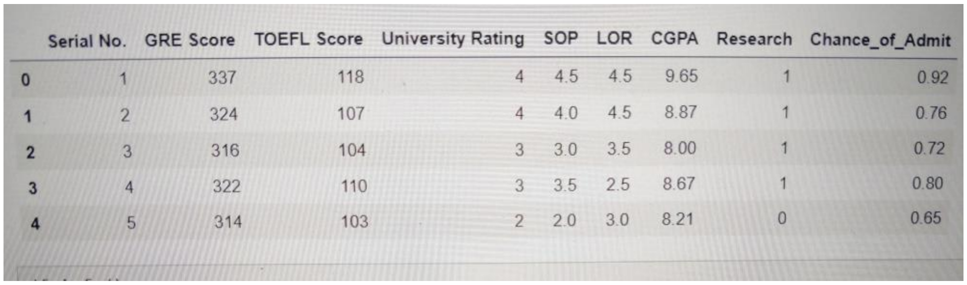
II.Analysis

• There are 9 columns: Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR , CGPA, Research, Chance of Admit

• There are no null records. It's good.

• There are 400 samples in total.

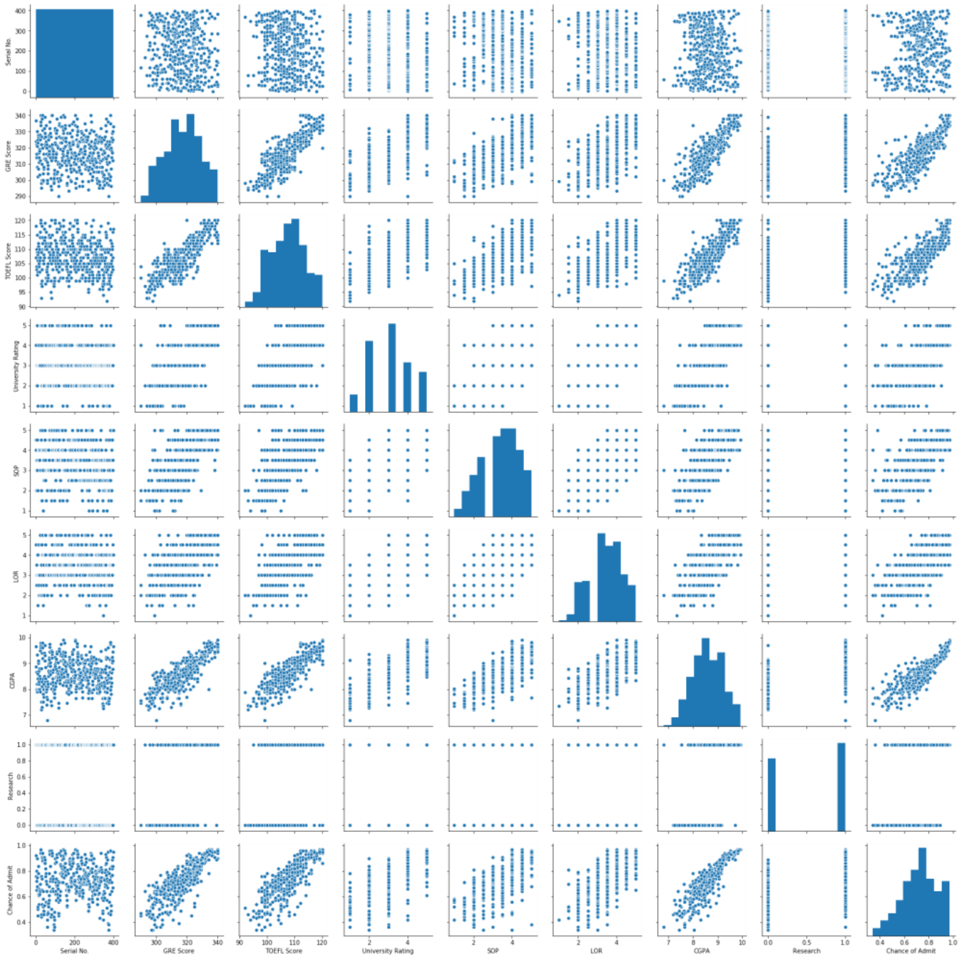
The structure of the dataset is like as below:

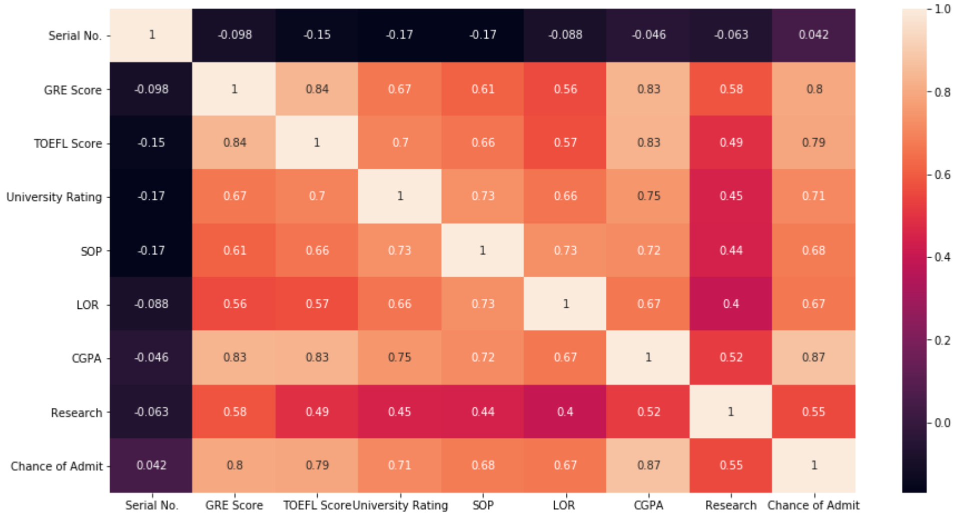


	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance_of_Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Checking for any Linear Relationship between the given Parameters:

- By constructing Pairplot
- By constructing Correlation heatmap





	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
Serial No.	1	-0.098	-0.15	-0.17	-0.17	-0.088	-0.046	-0.063	0.042
GRE Score	-0.098	1	0.84	0.67	0.61	0.56	0.83	0.58	0.8
TOEFL Score	-0.15	0.84	1	0.7	0.66	0.57	0.83	0.49	0.79
University Rating	-0.17	0.67	0.7	1	0.73	0.66	0.75	0.45	0.71
SOP	-0.17	0.61	0.66	0.73	1	0.73	0.72	0.44	0.68
LOR	-0.088	0.56	0.57	0.66	0.73	1	0.67	0.4	0.67
CGPA	-0.046	0.83	0.83	0.75	0.72	0.67	1	0.52	0.87
Research	-0.063	0.58	0.49	0.45	0.44	0.4	0.52	1	0.55
Chance of Admit	0.042	0.8	0.79	0.71	0.68	0.67	0.87	0.55	1

From above the Parameters with High Correlation against 'Chance of Admit' are:

- GRE Score

- TOEFL Score
- CGPA

Algorithms and techniques

The main aim of this problem is to predict the 'Chance of Admit' with high accuracy by applying various ML Algorithms and then comparing their scores..

In this problem we use both classification and regression models

Algorithms Considered:

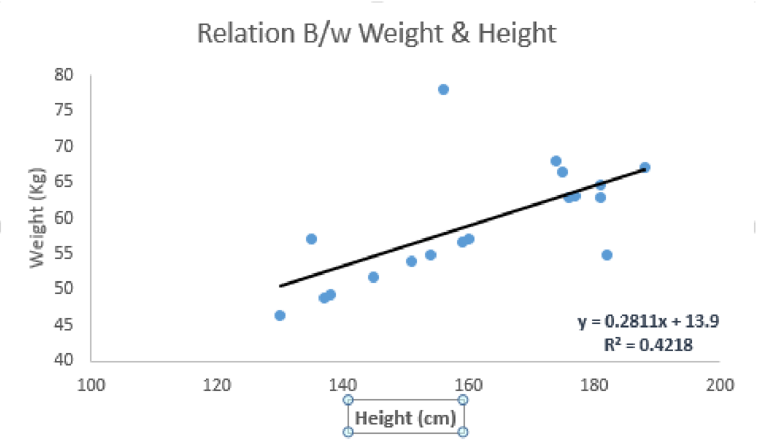
- Linear Regression
- Decision Tree Regressor

Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

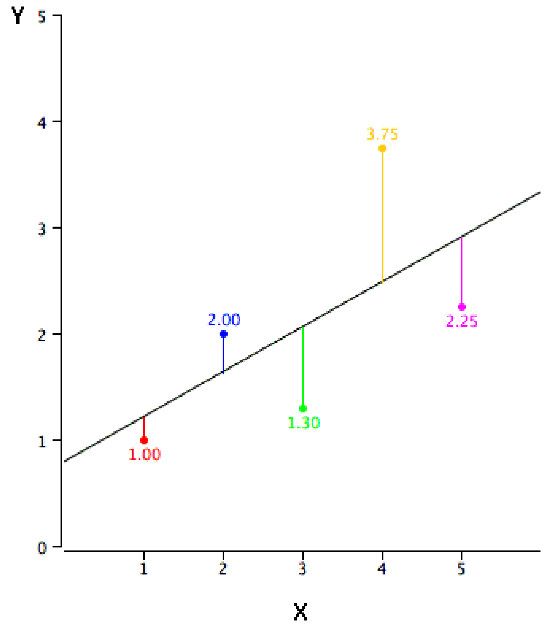


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is “How do we obtain best fit line?”.

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

$$\underset{w}{min}\,||Xw - y||_2{}^2$$



We can evaluate the model performance using the metric **R-square**.

Benchmark Model

The benchmark model I will choose is Decission tree regressor which gives r-squared score of 65.81%

In order to beat this I will choose anothor regression model called linear regression

While I'm using linear regression I successfully got 82.14% which is far better than the decision tree regressor.

III.Methodology

Data preprocessing:

In data preprocessing ,I will load the data and check for whether dataset need cleaning or not.

In this dataset there are no null values,duplicate.so,the dataset is already cleaned so normalization of data is required.

Implementation:

finding the Parameters with High Correlation against 'Chance of Admit'

Here I will Check for any Linear Relationship between the given Parameters

- By constructing Pairplot
- By constructing Correlation heatmap

the parameters which effects the final value of chance of admit

- GRE Score
- TOEFL Score
- CGPA

While , I'm splitting the data tnto train and test dets I set the size of test set as 0.25.

Next I applied the linear regression algorithm to my problem

```
from sklearn.linear_model import
LinearRegressionlr = LinearRegression()
```

After that I applied fitting function to my training data set.

Later I calculated the r-score and comared with the benchmark model.

Refinement:

In this problem I considered regression models linear regression and decision tree regressor,while im considered decision tree regressor as benchmark model I got r-square score around 65%.

Linear regression is a linear model, which means it works really nicely when the data has a linear shape. in order to improve the r-square score I use linear regression which gives me r-squared score as 82%.so I considered linear regression as best one for this problem.

IV. Results

Model Evaluation and Validation

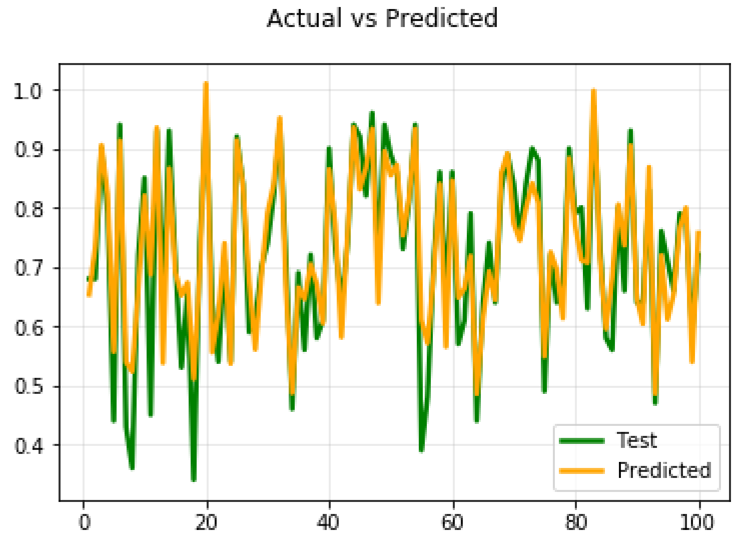
In statistics, the **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

im considered decision tree regressor as benchmark model I got r-square score around 65%.in order to improve the r-square score I use linear regression which gives me r-squared score as 82%.so I considered linear regression as best one for this problem.

The ability to perform well on unseen data is called **generalization**, and is the desirable characteristic we want in a model.

When a model performs well on training data (the data on which the algorithm was trained) but does not perform well on test data (new or unseen data), we say that it has overfit the training data or that the model is overfitting. This happens because the model learns the noise present in the training data as if it was a reliable pattern.in this problem linear regression performs well as u can see the graph between atual vs predicted.so this model performs well for unseen data.



Sometimes outliers are bad data, and should be excluded, such as typos. Sometimes they are Wayne Gretzky or Michael Jordan, and should be kept.

"This regression line fits pretty well for most of the data. 1% of the time a value will come along that doesn't fit this trend, but hey, it's a crazy world, no system is perfect"

So, finally we trust this model for this problem we can also get better results by using other ML algorithms.

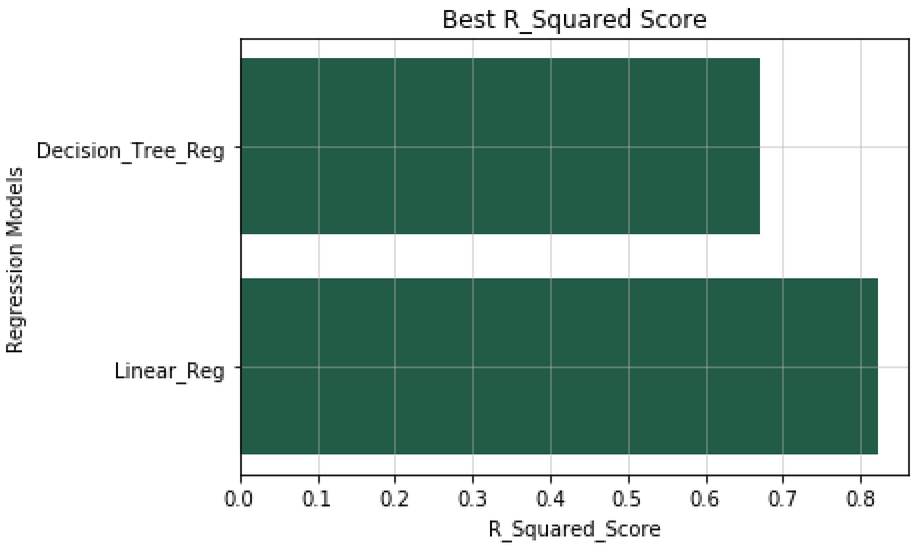
Justification

i considered decision tree regressor as benchmark model I got r-square score around 65%.in order to improve the r-square score I use linear regression which gives me r-squared score as 82%.so I considered linear regression as best one for this problem.

V. Conclusion

Free-Form Visualization

Final comparision between r-scores of benchmark model and linear regression:



Reflection

1)First I have gone through some of the problems in Kaggle, but this admission prediction seems to be very interesting and thought this can be helpful for students

2) Next I gained knowledge about the procedure of allocation of seats and how it is effected

3) Afterwards I started downloading dataset and plot some graphs which gives me correlation of attributes

4) As I mentioned , There are 9 columns: Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR , CGPA, Research, Chance of Admit.

5)first I went with decision tree regressor but this one gives me r-square score around 0.66.so I considered decision tree regressor as benchmark model.

6)in order to compete with the above r-score I choose linear regression.

7)the linear regression gives me r-score as 0.82 which is far better than previous one.

Improvement

- For this problem we can improve a better r_score by using the ensembling methods called AdaBoost, XGBoost.
- In future we can also use grid search techniques for parameter optimization.
- As the dataset contains only 400 rows .it is better to add some more data in order get the best results.