Graduate Admissions using Advanced Regression Techniques

Authors Name: Mr. I.K.Pradeep Sir, Mr. K.Bhargav Kiran Sir, M. Madhu Kiran Varma :

CSE dept., Vishnu Institute of Technology, Vishnupur, Bhimavaram.

***Abstract*— This is a research paper on the scenario that In India Pursuing a Masters Degree in abroad is a dream for every individual who aspires to advance the vision in his domain immediately after the Under-Graduation.Hence an analysis is made by Mohan S. Acharya, which was posted in Kaggle.By the sheer exploration of the data it is possible to develop certain rating based on the related characteristics or features which can be used to help students in shortlisting the universities and the rating gives an insight to the individual about a clear idea about their scope or chances for an admission in a specific university of their desire.Hence, this project is aimed at developing a dynamic model that can predict the individual's chances of getting an admission into an abroad university through a score 'Chance Of Admit', based on his/her performance in the prerequisite examinations like GRE , TOEFL and undergraduate score , CGPA and additionally their research experience and ratings of factors that aid the admissions like Letter Of Recommendation(LOR) and Statement Of Purpose(SOP) and the rating of the university that the individual is willing to pursue his education.**

Keywords—Predictive Modeling, Feature Selection, Recursive Feature Elimination, Supervised Learning, Features,Python.

I. INTRODUCTION

Our project 'GRADUATE ADMISSIONS' is a standard Machine Learning approach for predicting the chance of admission for an individual aspiring for a Masters Degree specifically in the United States.Machine Learning is an application of Artificial Intelligence(AI) that provides the capability to a system to think and make decisions just like a human mind without being explicitly programmed.Machine Learning algorithms and methodologies are specifically designed to understand the data provided and apply it to learn on it's own.The people who are aspiring for a Masters Degree will make sure that they are applying for a good university with limitless opportunities for both professional and career orientation.This project aims at providing the individual's an insight of their admission chances by taking in the data of their GRE, TOEFL, CGPA scores and the recommendations like LOR and SOP eventually drawing inferences by providing a rating out of 1(for higher chances) and 0 (for critical chances).

The goal is to develop a model that can predict the score 'Chance Of Admit', that determines an individual's chances of getting an admission in the university he desires to pursue his/her education**.**

The tasks involved are :-

- Download and preprocess the Graduate Admissions data.
- Train a BenchMark Model and record it's performance.
- Then three supervised learning models were trained using the training data and a comparison is done based on

the performance metric and decided which among the three is the best model.

- The best model thus selected is optimized using GridSearchCV technique.
- The Optimized model is then compared with the Benchmark model and deciding which is the best for the given data.
- Then the best model is validated against unseen data and documenting the intuition.

The final model can be applied to determine the Chance's for an individual of getting an admission into a specific university of his desire.

II. TOOLS AND TECHNOLOGIES USED

The entire project is built upon Python programming language. The Algorithms and methodologies used in the respective model building are facilitated by the implementation of “SCIKIT- LEARN” framework specifically built for applying Machine Learning to build practical systems that have the capability to make decisions without any external human interference. The statistical part of the project is implemented by the use of “NUMPY” module in Python. The Data Munging part is handled by the implementation of “PANDAS” module and the Data Visualization part is implemented by the aid of “MATPLOTLIB” and “SEABORN ” libraries available in Python. All the algorithms and techniques used in the project for building the model including training,testing,analyzing it’s performance, pre-processing,evaluation and validation are performed by the methodologies available in the “SCIKIT-LEARN” framework of Python.

III. DATA ACQUISITION

Data Acquisition is one of the key steps in Machine Learning.Since, it is more prerogative to understand the data or just to feed the data as an input to the specific Machine Learning Model is achievable only in a 'Tabular Sense' i.e., if the input dataset is converted to a tabular data, then it is more flexible for the model to build accurate predictions out of the perfected data format.

IV. DATA EXPLORATION

Data Exploration is a crucial step in the process of Machine Learning.It helps us to understand the patterns and available features in a data set from which we can determine the sort of actions that we can perform for further analysis.

Data Exploration gives an intuition that a cursory investigation of the data-set is necessary for familiarizing Yourself with the data through an explorative process and is a fundamental practice to help you better understand and justify your results.

Since, the main goal of this project is construct a working model that has the capability of predicting the 'ChanceOfAdmit' scores, we will need to separate the dataset into features and the target.

V. DATA VISUALIZATION

Exploratory Visualization can be defined as an approach for analyzing data sets to summarize their important characteristics, often by the application of visual methods.The Primary theme of Exploratory Visualization is for observing what the data can give us an intuition far beyond the conventional modeling or hypothesis testing tasks.

The Following session of Data Visualization will provide the intuition of how each feature is 'Correlated' with the target variable 'Chance of Admit'.

1. Correlation of ‘GRE Score’ with respect to ‘Chance of Admit’

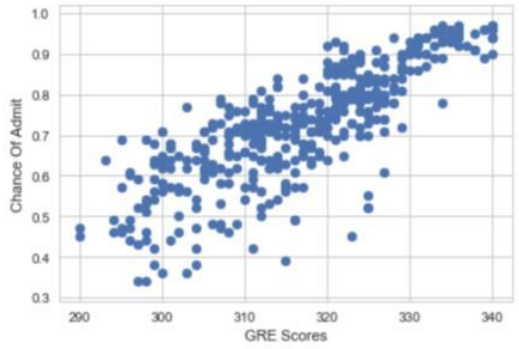

Fig.1. Scatterplot between ‘GRE Score’ and ‘Chance of Admit’

2. Correlation of ‘TOEFL Score’ with respect to ‘Chance of Admit

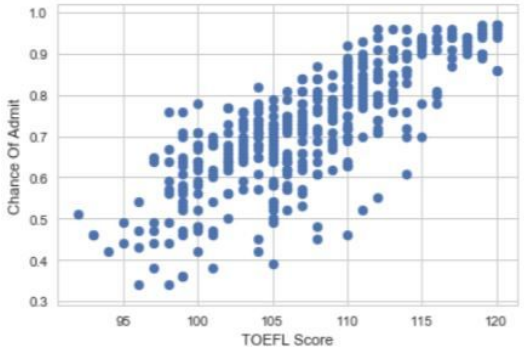

Fig.2. Scatterplot between ‘TOEFL Score’ and ‘Chance of Admit’

3. Correlation of ‘University Rating’ with respect to ‘Chance of Admit’

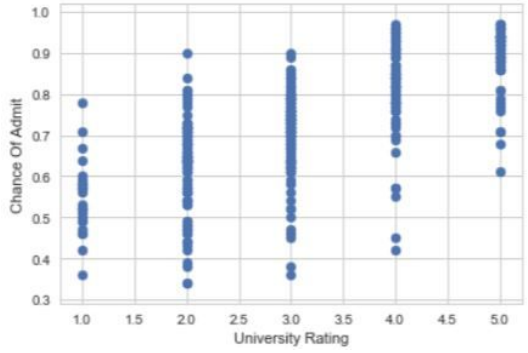

Fig.3. Scatterplot between ‘University Rating’ and ‘Chance of Admit’

4. Correlation of ‘Statement of Purpose’ with respect to ‘Chance of Admit’

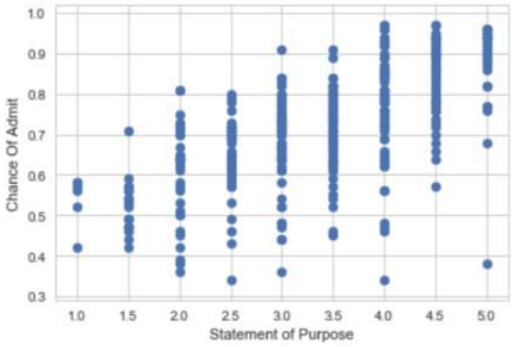

Fig.4. Scatterplot between ‘Statement of Purpose’’ and ‘Chance of Admit’

5. Correlation of ‘Letter of Recommendation’ with respect to ‘Chance of Admit’

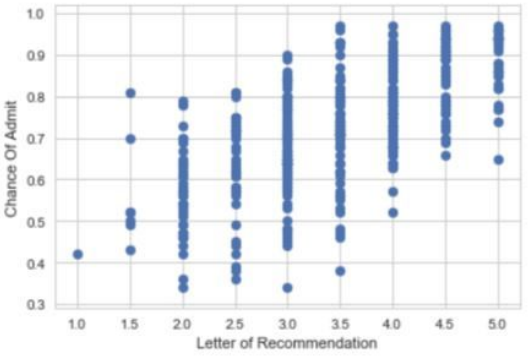

Fig.5. Scatterplot between ‘Letter of Recommendation’’ and ‘Chance of Admit’

6. Correlation of ‘CGPA’ with respect to ‘Chance of Admit’

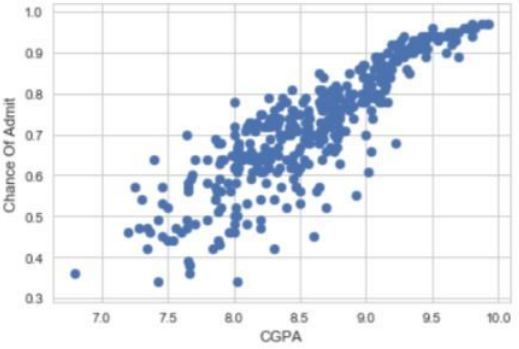

Fig.6. Scatterplot between ‘CGPA’ and ‘Chance of Admit’

VIII. DATA PREPROCESSING

- Data Preprocessing is one of the key procedures in Machine Learning that involves transforming the raw input data into an understandable format.
- This step is quite crucial because the Real-World data is often incomplete and inconsistent, or lacking in certain behaviours or trends, and is most probably consists of many errors.
- Data Preprocessing is a proven procedure for solving this problems and it prepares raw data for further processing.
- In the current problem an irrelevant feature is present that is 'Serial No',which actually plays no role in the determination of the 'Chance Of Admit' rate.Hence, it is quite valid to remove it from the dataset.

Feature Selection

- Feature Selection is an important concept in the Data Preprocessing phase which involves the removal of irrelevant attributes and retaining only those features that have show better performance.
- I implemented Feature Selection using a Linear Regressor, the benchmark model that i used to prune the features that actually have a lower rank of performance.
- Recursive Feature Elimination(RFE) uses a model (here Linear Regressor) to select either the best or worst performing feature, and then simples prunes the feature. After this the entire process is iterated until all the features in the data set are used up(limit).
- Sklearn has an inbuilt RFE function i.e., sklearn.feature_selection and I'm using this along with my Linear Regressor Model.

IX. ALGORITHMS AND TECHNIQUES

By observing the problem, it is quite evident that it is a 'Regression' Problem. It is important to understand the intuition behind the consideration of specific model, since it has to generate an optimal possibility of results that can improve the model's performance on a sample of new data. Hence, taking the performance of the model into consideration, I chose three Supervised Machine Learning Algorithms that can be better compatible for the data being available.

They are :-

- Decision Trees
- Ensemble Methods - Random Forests
- Support Vector Machines (SVM)

Decision Trees

- Decision Trees are very flexible,easy to understand and easy to debug.They usually work on Classification and Regression Problems i.e.,for categorical problems having [green,red,blue..etc.] and continuous inputs like [2.9,3.8..etc].They usually cover both the perspectives.
- It divides the 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. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.
- As the given data is comprised of continuous features,Decision Trees perform well in regression tasks.

Random Forest Regressor

Random forests are an ensemble learning method for classification, regression and other tasks, that

operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes in case of the classification tasks or mean prediction for the regression tasks of the individual trees.

- The Random forests have a stroke of brilliance when a performance optimization happens to enhance precision of the model, or vice versa.Tuning down the fraction of features that is considered at any given node can let you easily work on datasets with thousands of features.
- Since Random Forests perform well on almost every machine learning problem and they also show less overfit behavior when compared to Decision Trees. Since our problem is composed of a lot of continuous features for which Random Forests serve a better choice.

Support Vector Machines

- SVM's are simple,accurate and perform well on smaller and cleaner datasets.It can be more efficient as it uses subset of training points.
- The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. Initially as the output is a real number and continuous it becomes very difficult to predict the information at hand, which has infinite possibilities.
- In the case of regression, the factor margin of tolerance (epsilon) is set in approximation to the SVM.
- The main theme for SVM is always to minimize error, particularize the hyperplane which maximizes the margin, keeping in mind that part of the error is convinced.
- Since For the current regression problem consists of a lot of continuous features the application of SVM's can serve a better purpose.

IX. METRICS

- The current problem is a Regression task, since it takes certain features as inputs and attempts to find a score that helps an individual to get an idea about the chances of getting an admission in a specific university.
- Hence, Coefficient Of Determination is considered as the performance metric that can be applied to compare the performances of the scores obtained from the BenchMark and the Optimal Model considered.
- The CoEfficient Of Determination(R^2) is the key output of the Regression Analysis.It can be defined as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- It's values ranges from 0 to 1, and the results are given intuition by, if:-
    1. The value of R^2 -> 0 , indicates that the model is a worst fit to the given data.
    2. The value of R^2 -> 1 , indicates that the model is the best fit to the given data.
    3. The value of R^2 in between 0 and 1 -> indicates that the respective variability exhibited by

the target variable.

- The formula for CoEfficient Of Determination(R^2) is given by :-

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

X. BENCHMARK MODEL

- A Benchmark Model can be defined as a standard model that already shows a better performance on a given data.The factors on which our results or the solution is tested, are mostly going to be the amount of training/testing data, and then we compare your solution with that of the benchmarked solution obviously based on a performance metric(here r2_score).
- The main theme here is to understand which model works delivers the best solution than their existing solution.So, it can be achieved by sheer analysis, implementing standard algorithms and observance and coming to the conclusion that the model shows good solutions or results than the benchmark model's solution.
- Since, the problem is a 'Regression' task , I'm implementing a 'Linear Regression' model as my BenchMark Model.

XI. IMPLEMENTATION

In the further section of the project, I'll intuitively select the best out of the three models that I considered for the current problem by using the performance metric(r2_score), based on the results generated, I'll decide best of the three models, which is optimal for the given problem.

INITIAL MODEL EVALUATION

In this section, a clear description of the coding implementation of the three supervised learning models is given,

- Import the necessary libraries and initialize the models and store them in respective variables.
- And finally comparing the r2_scores of the three learning models and decide which one is the best.

CHOOSING THE BEST MODEL

Since, it is obvious that the model which has best high r2_score when compared to the other models can be termed as the best optimal model for the current problem, as the fact that if :-

- r2_score is 0 -> it indicates that the model is a worst fit to the given data.
- r2_score is 1 -> it indicates that the model is the best fit to the given data.
- r2_score in between 0 and 1 -> indicates that the respective variability exhibited by the target variable.

The Values are tabulated and documented for further exploration.

XII REFINEMENT

In this section of the project, the model('Random Forest Regressor Model') is optimized by the application of 'GridSearchCV' technique for fine tuning the parameters for the final model thus chosen and later calculating the performance metric(r2_score) of the optimized model.

MODEL TUNING

Here, we will find the implementation of GridSearchCV by initially importing the libraries sklearn.grid_search.GridSearchCV and sklearn.metrics.make_scorer.

1. Initialize the regressor('Random Forest Regressor Model') and store it in the variable 'lgr_grid'.

2. Creating a dictionary of parameters, in the variable parameters.

3. Using make_scorer to create a r2_score scoring object. -> scorer

4. Perform Grid Search on the Regressor lgr_grid using the 'scorer', and store it in grid_obj.

5. Fit the Grid Search Object to the training data (X_train, y_train) and store it in the grid_fit.

FINAL MODEL EVALUATION

In this part of the project a clear demonstration of the comparison of the performances between the BenchMark Model('Linear Regressor') and the Optimal Model('Random Tree Regressor') based on their performance metrics(r2_score) in a tabular form.

MODEL VALIDATION

In this part of the project, a demonstration of the performance of the Best Model for the given regression task i.e., The Optimal Model against unseen data.

XIII. CONCLUSION

By observing the validation results above, it is quite evident that the model is performing well on the given data.

When compared to the BenchMark Model, the Optimal Model('Random Forest Regressor Model') indeed shows a striking performance as shown in the Table-4 of the 'Final Model Evaluation' section.

- Although the performance scores are so close and similar, it is likely to understand that the Optimal Model('Random Forest Regressor Model') holds a tight control in generalizing the input data.
- The performance of the BenchMark Model is 75.7 % and that of the Optimal Model is 76.1%.

It is also important to note that the individual's chances of getting an admission in their dream university are predominantly dependent on his/her performance in the prerequisite tests and will to succeed in their respective careers.

The model can be applied or is useful for students who are aspiring for a certain university of their desire to pursue their education and build their career.

In summary the application of the model is quite useful only in the domain of 'Education' and that the data should be very consistent that if any discrepancies in the features may lead to bad predictions i.e., a good student may get a bad prediction and a student with ill IQ can gain advantage of it.

XIV IMPROVEMENT

Potentially the 'Data Preprocessing' phase is the crucial part of any Machine Learning Problem, Since during this phase we can potentially identify the flaws in the data set that could actually mess the results and performance of the model thus considered.

- Hence, removing irrelevant data during this phase can predominantly increase the model's performance and can benefit in generalized results.
- But I personally feel that the feature selection that I used is good but there are other techniques that are used for the application of Feature Selection.
- Since, this is a regression task, the Wrapper Method implementation i.e., RFE may not be the best one.
- There is also room for trying Embedded method such as LASSO and Elastic Net and Ridge Regression that don't need the external

implementation of the feature selection techniques , since these methods have embedded feature selection and regularization built in. It's worth a trial, since there is always scope for improving the model performance agaInst the given dataset.

- Furthermore, some ensemble methods such as 'XGBOOST' should also take care of larger data dimensions and these methods can themselves be used for feature selection.

There are lot of other possibilitIes that can make the feature selection more intuitive and literally I am new to this and the 'Feature Selection' implementation gave me a hard core challenging of applying it in real time.

REFERENCES

[1]https://www.kaggle.com/mohansacharya/graduate-admissions

[2]https://en.wikipedia.org/wiki/Residual_sum_of_squares

.

[3] https://en.wikipedia.org/wiki/Total_sum_of_squares

[4]http://scholarstrategy.com/importance-of-gpa-and-gre-scores-in-your-ms-applications/.

[5] https://studyabroad.careers360.com/what-toefl-test

.

[6]https://en.wikipedia.org/wiki/College_and_university_rankings

[7]https://studyabroad.shiksha.com/sop-statement-of-purpose-aplycontent1701

[8] https://en.wikipedia.org/wiki/Letter_of_recommendation

[9]https://www.phdstudent.com/Getting-Involved-in-Research/the-importance-of-research-to-grad-school-admission

[10]https://scikit-learn.org/stable/

[11] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[12] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection

[13]https://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination

[14] https://en.wikipedia.org/wiki/Linear_regression

[15] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[16]https://en.wikipedia.org/wiki/Random_forest

[17]https://en.wikipedia.org/wiki/Residual_sum_of_squares

.

[18]https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[19]http://scholarstrategy.com/importance-of-gpa-and-gre-scores-in-your-ms-applications/.

[20]https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

.

[21]https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[22]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[23]https://scikit-learn.org/stable/modules/svm.html

[24]https://scikit-learn.org/stable/modules/grid_search.html

[25]https://stats.stackexchange.com/questions/153131/gridsearchcv-regression-vs-linear-regression-vs-stats-model-ols

[26] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[27] https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[28]https://stackoverflow.com/questions/43590489/gridsearchcv-random-forest-regressor-tuning-best-params

[29]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

[30] https://stackoverflow.com/questions/23309073/how-is-the-r2-value-in-scikit-learn-calculated