MACHINE LEARNING ENGINEER NANODEGREE

June 12, 2018

CAPSTONE PROPOSAL
MADHU KIRAN VARMA
June 12th , 2018
Predicting The Chance Of Admit For The Students of India Aspiring For Masters Program abroad
DOMAIN BACKGROUND :-

- In India Pursuing a Masters Degree 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.
- Link for the data set in Kaggle : https://www.kaggle.com/mohansacharya/graduate-admissions
- By the sheer exploration of the data it is possible to develop certain rating based on the related charactersitics or features which can be used to help students in short-listing 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.
- Using the data effectively can potentially help an individual to determine the chance of being admitted in a specific university of their interest.
- The major motivation for exploring this project is to develop the skill and understanding how to work out on real time datasets and how Machine-Learning is potentially making the lives of people easier.

PROBLEM STATEMENT :-

- By accurately predicting the chances of an individual for getting an admission in a specific university.
- By using the data-set on Graduate Admissions, the task is to predict the 'ChanceofAdmit' score that explains the chances for an individual for getting an admission in a specific university.The model utilizes the important characteristics of the data to develop models that can predict the scores.
- I decided to implement Machine-Learning techniques to predict the 'ChanceofAdmit' score by which an individual decides his chance of admission in a specific university of his desire.

- This Project constitutes Data Exploration and Visualizatios,Data Preprocessing and finally testing various algorithms and techniques.

DATASETS AND INPUTS :-

- The data-set constitutes a single CSV files which is obtained from Mohan S.Acharya from Kaggle.
- Link :- https://www.kaggle.com/mohansacharya/graduate-admissions
- The data has 9 columns i.e., they are Serial No.,GRE Score,TOEFL Score,University Rating,SOP,LOR,CGPA,Research, and the target variable being 'Chance of Admit' and has 400 rows.
- We are comprised with a good amount of features which are potentially utilized to estimate a good score.

SOLUTION STATEMENT :-

- The brighter solution to this problem, is drawn from the scores of the individuals from different features and the individuals who claim to secure highest scores in all the respective domains can get a better chance of admission in a specific university of their interest.
- This potential data can be helpful for testing the supervised machine learning models to predict or estimate the scores of the new individuals.
- I will eventually work out on diverse models such as SVM's,Decision Trees and Random Forests etc., along with GridSearchCV for Model optimization.

BENCHMARK MODEL :-

- Since the given problem expects to predict a continuous output,it is more obvious to determine a metric value(R2_score) that will potentially help us to establish a comparison between the performances of the Bench Mark Model and the Optimal Model thus selected.
- The BenchMark model is a base model that potentially has no intelligence, hence we start with finding the r2_score that provides an indication of the goodness of fit of a set of predictions to the actual values.

EVALUATION METRICS :-

- The current problem is a regression problem, since it takes certain features as inputs and tries to figure out a score that will be helpful for an individual to determine his chances for an admission in a specific university he desires to go to.
- Hence, I decided to use 'Co-Efficient Of Determination' as the performance metric that could be applied to chech the performance of the scores obtained from the Bench Mark Model and the Optimal Model considered.

PROJECT DESIGN :-
The project will be designed in correspondence to the following steps :-

- Data Acquisition :- Indeed the first step of the project is acquiring data which I potentially collected from the Kaggle.
- Data Exploration and Visualization :- Exploring and visualizing the dataset, cleaning and observing the relevance of every feature.
- Data Pre-Processing :- Data Preprocessing is the key factor in Machine-Learning, in my current project I used this process to clean and remove any irrelevant data.
- Model Evaluation and Validation :- Model is tested for the performance in it's BenchMark State and the Optimized State respectively.
- Optimization :- The naive model is optimized by the application of GRIDSEARCHCV which helps in tuning the parameters optimal for the model.