Machine Learning Engineer Nano degree

Capstone Project Proposal

COLLEGE ACCEPETANCE RATE PREDICTION

MODUGAMUDI.FREDRICK ALEX

February 16th, 2019

Proposal :-College acceptance rate prediction

Domain Background

The main reason of this project(code) is to get an better accuracy of predicting the accepetence of student in an university from the Indian point of view. This project is very helpful for the students to know whether they get admission in a particular university or not.

If you consider applying for a graduate school, you are interested in useful information such as chances for admission and minimum requirements because you can make a powerful strategy and prepare it efficiently by the news. In fact, there are lots of communities for graduate admission.

A college's overall quality and prestige is determined in part by its selectivity, of which acceptance rate is a main indicator. The acceptance rate figures are provided by the IPEDS database, a tool offered by the U.S. Department of Education.

The school with the lowest acceptance rate in 2017 was Stanford University, located in Stanford, California.

Many students aim for admission to a prestigious college or university, but the supply of open seats often does not meet the demand from applicants. Earning admission to the schools on this list can be especially difficult.

Acceptance rates at top schools are continually dropping. For example, a recent Huffington Post article examining the dropped acceptance rate cited that back in 1991, University of Pennsylvania had a 47% acceptance rate, where by contrast 2016's rate came in at 12.3%.

So my goal is to apply machine learning techniques and predict the outcome(whether he may be accepted or not).

Link of the above data: https://oedb.org/rankings/acceptance-rate/

INSPIRATION:

This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

PERSONAL MOTIVATION:-

As from the starting of this Nano degree program I very much interested in the topics like Neural Networks , Linear Regression, Random Forest algorithm .I want to know the best among them. But as we already know these algorithm results changes from problem to problem .Even though I choose this problem so I can apply number of algorithms on this dataset and choose the best among them(related to this problem).

Problem Statement

The aim of this project is to predict the acceptance of a student based on a minimum of data collected of the student(Predicting the admission of a student at a particular university can help shorten the time required for

managing agencies to provide support and plan targeted maintenance(i.e money) for the students career. Here we use supervised learning and get the output in a continuous order between 0 and 1.

Datasets and Inputs: The dataset is downloaded from the below link

https://www.kaggle.com/mohansacharya/graduate-admissions

Content.

The UCLA Graduate Dataset inspires this dataset. The test scores and GPA are in the older format. Mohan S Acharya owns the dataset.

The data set contains 500 student records.

The dataset contains several parameters, which are considered important during the application for Masters Programs. The parameters included are :

1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1)

All the columns i.e:-(9 attributes)

#Serial No .#GRE Score #TOEFL Score #University Rating #SOP #LOR #CGPA #Research #Chance of Admit are all numerical data, meaning they are quantifiable measurements so, we can perform mathematical operations.

In this dataset we neglect “Serial No.” coloum as it does not play role in the prediction.

Here we take 400 records. So we use 240 as training data and 160 as testing data.

In simple **linear regression** a single independent variable **is** used to predict the value of a dependent variable. In multiple **linear regression** two or more independent variables are used to predict the value of a dependent variable. The difference between the two **is** the number of independent variables.

So we use “CGPA” attribute as an central variable and compare it with other features and find the dependencies.The above data is taken from:

https://www.google.com/search?q=data+predicting+using+linear+regression&oq=data+predicting+using+linear+regression&aqs=chrome..69i57j0.34758j1j7&sourceid=chrome&ie=UTF-8

Solution Statement

By using sklearn and linear regression we train up to 240 to 250 epochs and preprocessing,and data cleaning and will finally predict the students acceptance accuracy. I will also try with decision trees and random forest and neural networks.

But neural network works good for larger data sets not small.

Benchmark Model

Here I consider decision trees and random forest as my bench mark because ther are already some predictions that are done with the above algorithm but could not meet perfect accuracy. Random forest gave 94% and I would like to increase the accuracy using linear regression.

Ill train and test the data so to give a better prediction

Link for previous project: https://www.kaggle.com/vai1995/graduate-admission-prediction-with-94-accuracy

Evaluation Metrics

The linear regression equation

[Rˆ2/RMSE]

Wher R^2= l gives some information about the goodness of fit of a model.

RMSE=ROOT MEAN SQUARE ERROR

But in sklearn we already have built-in functions for linear regression.

So by using the regression we can predict the training and testing set of the given data .At some point you want to know the classification accuracy. Cutting to the chase,the as this is a small data set we can go with traditional regression(linear).

Project Design

The project is composed of different steps as follows:-

Preprocessing:

We first performed a feature screening and decided to use only 8 of the 9 features. Our screening process excluded 1feature for the following reasons:

Irrelevance: some of the features were deemed irrelevant to our project and we decided to exclude them to reduce the computational cost of our algorithm.

Redundancy: some of the numerical features had exact or almost exact duplicates and we decided to only keep one out of the two or three identical features. In these cases, we kept themost granular feature. In particular, this reduced the number of serial no

- There are null values in our features which are needed to beupdated for better training of our model.

steps involved in project design

step 1:we load the data in to the code.

Step2: We haven't looked too closely at the data .so to get involved in features engineering, but we can already start with some items that will make our life easier. Right off the bat, this is what we can do:

- Get rid of the 'Serial No.' Column, as it only serves the purpose of identifying entries and would not contribute to data exploration/visualization/predictions.

Step 3:so, I’ll consider the bench mark model and find its accuracy.

Step 4: now I’ll work on linear regression an see whethwr it can beat the accuracy of the random forest.

Step 5:implement the best technique for getting better accuracy.