ENGINEERING STUDENTS DATASET

SAI DURGA KARTHIK NANDIRAJU

Abstract

I took a student dataset from an engineering college located in India. The dataset contains information about the students high school scores, undergraduate scores and other personal details.

I explored the relation between undergraduate scores and high school scores of the student data set to answer the research questions. I used regression and python matplotlib while exploring the data.

I found that there is a fairly positive correlation between undergraduate scores and high school scores(12 grade).

Motivation

Engineering students play a major role in creating a better world. The scores obtained by a student during an undergraduate education in engineering determine how well a student understood the concepts so that he can apply and contribute to the real world. Therefore the poor performance of students can put the reputation of an academic institution to which they belong at stake and might have a negative effect on the society. All the students may not be suitable for engineering or for a particular engineering major, therefore identifying the potential students for a major is necessary.

The insight gained by exploration of the student data set would be valuable to the management or faculty of an engineering college especially during admission process.

Student Dataset

I took the student data set from an engineering college affiliated to Osmania University, Hyderabad, India. The student data set contains information about the students high school(10 grade and12 grade) scores(percentage), undergraduate scores(percentage), Major and other personal details like name, email id, address etc. I provided the meaning of some of the columns in my student data set below.

GPAX- 10 grade overall score in percentage BioTech- Biotechnology

GPAXII- 12 grade overall score in percentage Specialization- Major or Department

UGGPA- Undergraduate overall score in percentage

CSE- Computer Science and Engineering

Data Preparation and Cleaning

I used the pandas read_csv function to extract the data from a csv file into the data frame. I dropped all the null values using dropna function from the GPAX,GPAXII and UGGPA columns whose meaning was provided in the previous slide. I removed some unnecessary features from the original data set.

I faced a problem when some of the data in GPAX,GPAXII and UGGPA columns was in string format instead of number format. I finally converted everything to number format.

My data set contained 865 rows and 20 columns.

Data Preparation and Cleaning Code

```
data.columns
```

```
Index(['SNO', 'CID', 'firstname', 'lastname', 'fullname', 'Gender', 'EMAILID',
       'MobileNo', 'DOB', 'National', 'Colname', 'University', 'GPAX', 'XYEAR',
       'GPAXII', 'XIIYEAR', 'diployear', 'Current', 'Degree', 'Specialization',
       'UGGPA', 'UG Year', 'Unnamed: 23'],
      dtype='object')
```

```
features= [ 'SNO', 'CID', 'firstname', 'lastname', 'fullname', 'Gender', 'EMAILID',
       'MobileNo', 'DOB', 'National', 'Colname', 'University', 'GPAX', 'XYEAR',
       'GPAXII', 'XIIYEAR', 'Current', 'Degree',
       'Specialization', 'UGGPA']
```

```
pd= data[features]
```

```
pd= pd.dropna()
```

```
pd=pd[pd.GPAX != "."]
```

```
pd['GPAX'].astype('float')
```

Data Preparation and Cleaning Code

```
In [117]: pd.shape
Out[117]: (865, 20)
```

```
In [12]: pd
```

	SNO	CID	firstname	lastname	fullname	Gender	EMAILID	MobileNo	DOB	National	Colname	Uni
0	1	8070769	KUMMARA	YOGITHA	KUMMARA YOGITHA	Female	yogitha.shyni.k@gmail.com	9550355342	2/5/1995	Indian	CBIT	OS UNIVE
3	4	8070502	Chitra	Yerra	Chitra Naidu Yerra	Female	chitrayerra94@gmail.com	8106289298	12/12/1994	Indian	CBIT	OS UNIVE
4	5	8070824	bhavana	yendamuri	bhavana yendamuri	female	bhavanayendamuri.2012@gmail.com	9494227456	7/24/1994	Indian	CBIT	OS UNIVE
5	6	8070835	madhuri	yellaturi	yellaturi madhuri	female	madhu 3495@gmail.com	9885728740	6/13/1995	Indian	CBIT	OS UNIVE
6	7	8071080	TENNYSON	YELLAMATI	TENNYSON YELLAMATI	Male	y.tennyson9999@gmail.com	8985898203	5/20/1995	Indian	CBIT	OS UNIVE
7	8	8070533	Vamshi Krishna	Yedire	YEDIRE VAMSHI KRISHNA	Male	krishna.vamshi224@gmail.com	9676790231	5/2/1994	Indian	CBIT	OS UNIVE
8			balisetty		balisetty							

nal	Colname	University	GPAX	XYEAR	GPAXII	XIIYEAR	Current	Degree	Specialization	UGGPA
ian	CBIT	OSMANIA UNIVERSITY	91.6	2010	96.0	2012.0	UG	B.E	EEE	70.53
ian	CBIT	OSMANIA UNIVERSITY	90	2010	85.2	2012.0	UG	B.E	CSE	72.00
ian	CBIT	OSMANIA UNIVERSITY	93	2010	92.8	2012.0	UG	B.Tech	BIOTECH	87.70
ian	CBIT	OSMANIA UNIVERSITY	83	2010	84.2	2012.0	UG	B.Tech	BIOTECH	68.80
ian	CBIT	OSMANIA UNIVERSITY	86	2010	79.3	2012.0	UG	B.E	MECHANICAL	69.73
ian	CBIT	OSMANIA UNIVERSITY	92	2009	96.3	2011.0	UG	B.E	CSE	81.51

Research Question(s)

1.Does a student with a good overall high school score has a good overall undergraduate score?

2.Computer Science is generally a challenging course which requires good programming and mathematical skills whereas biotechnology does not require programming skills. Do students who got into computer science and performing well(80 and above score in computer science) have a better high school score than Biotechnology majors(biotech)?

3.What would be the estimate of an undergraduate score for a student given his/her high school scores?

Methods

I used regression to estimate the undergraduate scores given the high school scores of the students. I utilized python packages matplotlib for plotting the graphs, sklearn for regression, numpy and pandas for analysis.

```
import pandas as pd
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
import numpy as np
```

Findings(Research Question #3)

I used regression to estimate the undergraduate score given high school(10, 12 grade) scores which would be helpful in identifying potential engineering students.

```
x_features= ['GPAXII','GPAX']
```

```
x= pd[x_features].astype(float)
```

```
y_features= ['UGGPA']
```

```
y= pd[y_features].astype(float)
```

```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=324)
```

```
X_train.head()
```

	GPAXII	GPAX
558	72.1	71.50
901	80.0	90.00
578	87.7	75.70
378	94.6	90.33
821	94.0	92.00

```
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

```
y_prediction = regressor.predict(X_train)
y_prediction
```

```
In [20]: y_prediction = regressor.predict(X_train)
         y_prediction
```

```
Out[20]: array([[ 63.50035466],
                [ 68.77599206],
                [ 72.47534135],
                [ 77.01498279],
                [ 76.75151932],
                [ 78.10536679],
                [ 78.38812394],
                [ 78.19269887],
                [ 78.47603081],
                [ 78.54128594],
                [ 78.43797289],
                [ 77.8534599 ],
                [ 78.02068941],
                [ 61.59792687],
                [ 78.59113489],
                [ 77.82626268],
                [ 77.65803483],
                [ 78.10345332],
                [ 65.33032712],
                [ 74.60880244],
```

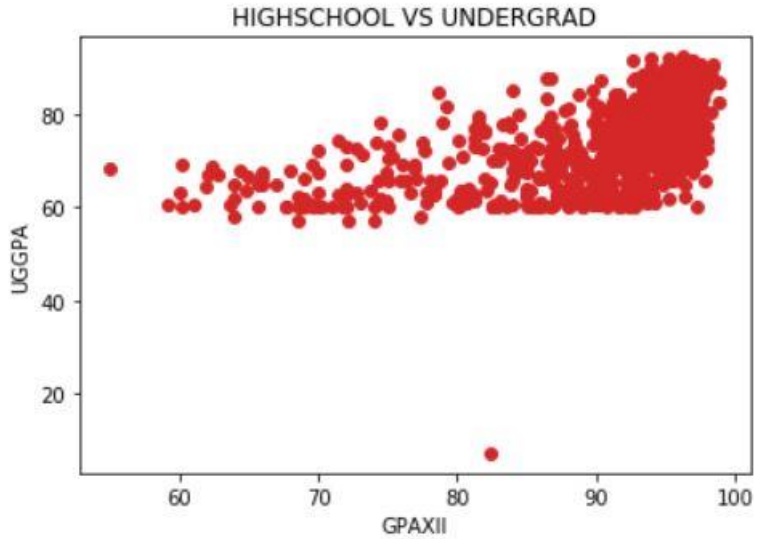
Findings(Research Question #1)

There is a fair possibility for a student with a good 12 grade score(GPAXII) to get a good undergraduate score. I found a correlation value of 0.563 between undergraduate score(percentage) and high school score(12 grade percentage). There is a low possibility for a student with better 10 grade score(GPAX) in percentage to get a better undergraduate score. The correlation value between 10 grade score and undergraduate score is 0.275.

```
: pd['UGGPA'].corr(pd['GPAXII'])
: 0.56322291599501195

: pd['UGGPA'].corr(pd['GPAX'].astype(float))
: 0.27519276026125
```

Findings(Research Question #1)

```
plt.scatter(pd['GPAXII'].values, pd['UGGPA'].values)
plt.xlabel('GPAXII')
plt.ylabel('UGGPA')
plt.title('HIGHSCHOOL VS UNDERGRAD')
plt.show()
```

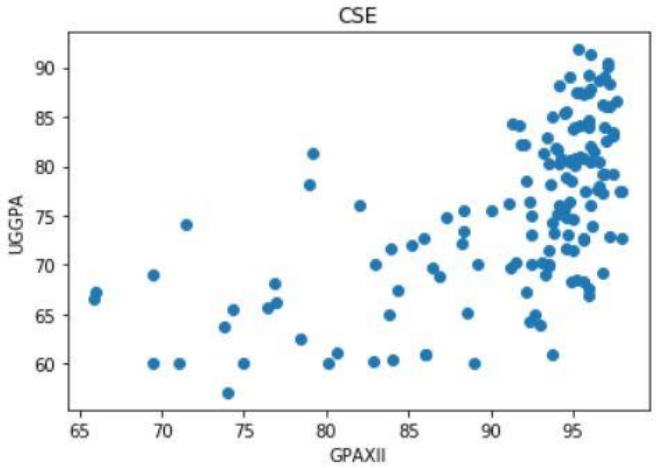


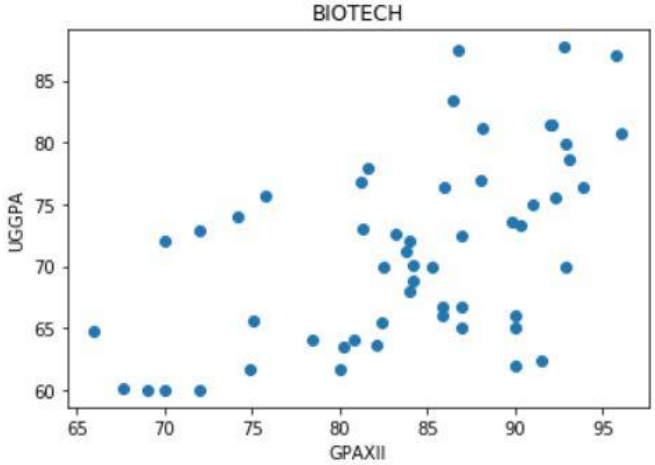
Findings(Research Question #2)

Students with Computer Science course as major got a better 12 grade score on an average(91.165) than those with Biotechnology as a major(84.010)

```
HIGHSCOOL AVERAGE PERCENTAGE OF CSE STUDENTS

In [28]: dp2.mean()
Out[28]: 91.16500000000002

UNDERGRAD AVERAGE PERCENTAGE OF CSE STUDENTS

In [29]: dp1.mean()
Out[29]: 75.76534246575346

In [30]: dp3= pd['UGGPA'].where(pd['Specialization'] == 'BIOTECH').dropna()

UNDERGRAD AVERAGE PERCENTAGE OF BIOTECH STUDENTS

In [31]: dp3.mean()
Out[31]: 71.22781818181818

In [48]: dp4= pd['GPAXII'].where(pd['Specialization'] == 'BIOTECH').dropna()

HIGHSCOOL AVERAGE PERCENTAGE OF BIOTECH STUDENTS

In [49]: dp4.mean()
Out[49]: 84.0101818181818
```

Findings Research Question #2





Limitations

These findings are true for an engineering college in India. These results may not be applicable to all the engineering colleges but can be same for some of them.

Conclusions

1. There is fair possibilty for a student with a good high school score(12 grade) to get a good undergraduate score.

2. The management can use high school score as a factor to determine the potential students for particular majors like Biotechnology and Computer Science. For a student, 80 and above score in computer science course might require a 90 and above 12 grade score whereas course like biotechnology might not require as high 12 grade score as computer science does for a good score.

3. Regression can be used by taking high school scores as independent variables and undergraduate score as dependent variable to filter future better undergraduate academic performers.

Acknowledgements

I had no one to give me feedback.

References

www.google.com

www.edx.org