Prediction of Domain Reputation Score

Payodhi Mandloi Geetha Kotagiri Mohnisha Mandwani
University of Maryland, Baltimore County
{payodhi.m, geethak2, mmandwa1}@umbc.edu

December 13, 2021

1 Introduction

Universities around the globe are striving to maintain an overall good reputation and trying to make a better place for the prospective students which instigates the creativity,entrepreneurship and intellectually of the students. All this effort is reflected on their overall reputation and most of the prospective students get carried away by the overall reputation instead of their specific department background which plays a major role in curving a path for their future plans.The data set provided is giving a clear perspective on the constraints that every student takes into consideration while selecting the Universities. Attributes like SAT score, tuition fees, Acceptance rate,Average GPA give a clear picture of the reputation of the University. Although it is projecting a good analysis on the overall ranking, it is vague considering domain specific requirements. We are trying to bridge this gap between the domains by designing an algorithm that predicts the university reputation score.

Initially, we proposed to solve a classification problem that segregated universities into Good/Average/Bad classes based on their domain. However, due to data-set restrictions we modified the problem statement to predict the domain based reputation score.

2 Problem Statement

There have been many prior works on assessing the University Reputation and its impact on the students.In this research project, we attempt to design an algorithm that predicts reputation score of the university based on the domain. This would help students to choose the best fit as per their competency. The scope of this project is limited to predicting the law reputation score. The similar idea can be extended to other domains as well.

3 Data sets

To predict the domain reputation score, we have used **"University Statistics"** data-set that was sourced from "Kaggle" along with **"Law School Data-set"** sourced from "publiclegal". University Statistics Data-set has 23 attributes out of which only 4 attributes were having an impact on the Engineering Reputation Score: Acceptance rate, Avg-sat score, Tuition fees and Avg GPA. **"Law school data-set"** consisted of 3 sub files consisting of - Law school ranking, Tuition and Salary. We have post-processed these files to generate a single data frame having only relevant values. The attributes includes Acceptance Rate, SAT Score, Avg-GPA and Tuition that had substantial amount of impact on the reputation score which is to be predicted. We have changed the data-set from **"US Law School Disclosures to the ABA"** to a **"Law School Data-set"** as the proposed data-set had multiple NULL values and was difficult to pre-process.

4 Exploratory Data Analysis

```
university_stats_dataset.describe().T
```

	count	mean	std	min	25%	50%	75%	max
act-avg	291.0	23.024055	4.159257	15.0	20.00	22.0	26.00	34.0
sat-avg	291.0	1044.027491	157.701571	715.0	930.00	1010.0	1130.00	1510.0
enrollment	300.0	14895.256667	10660.572830	133.0	6428.00	12104.5	21661.75	55776.0
acceptance-rate	302.0	60.390728	22.546806	5.0	47.25	65.0	76.00	100.0
percent-receiving-aid	143.0	35.279720	17.163426	5.0	21.00	35.0	47.00	81.0
cost-after-aid	143.0	33920.867133	7625.493383	13186.0	28613.00	34621.0	38936.00	51810.0
avg-gpa	244.0	3.543443	0.252066	2.8	3.40	3.5	3.70	4.0
businessRepScore	234.0	2.832479	0.590240	2.0	2.40	2.7	3.20	4.8
tuition	311.0	31121.340836	11995.242460	5460.0	21949.00	28500.0	41255.00	57208.0
engineeringRepScore	206.0	2.803398	0.645916	2.0	2.30	2.6	3.10	4.9
overallRank	311.0	83.623794	75.788351	-2.0	-1.00	75.0	151.00	223.0

Figure 1: Descriptive Statistics: University Statistics Dataset

Checking the number of NULL value in each column - University Statistics

```
#checking the null values
print(university_stats_dataset.isnull().sum())
```

```
act-avg                  20
sat-avg                  20
enrollment               11
acceptance-rate           9
percent-receiving-aid   168
cost-after-aid          168
state                     0
avg-gpa                  67
rankingDisplayRank        0
businessRepScore         77
tuition                   0
engineeringRepScore     105
UniName                   0
schoolType                0
overallRank               0
institutionalControl      0
dtype: int64
```

Figure 2: Checking the NULL values

Replacing the NULL values with Mean

```
university_stats_dataset.fillna((university_stats_dataset.mean()), inplace= True)
university_stats_dataset.head(10)
```

	act-avg	sat-avg	enrollment	acceptance-rate	percent-receiving-aid	cost-after-aid	state	avg-gpa	rankingDisplayRank	businessRepScore	tuition	engineeringRepScore	UniName
0	32.0	1400.0	5400.0	7.0	60.0	16793.0	NJ	3.900000	#1	2.832479	47140	4.100000	Princeton University
1	32.0	1430.0	6710.0	5.0	55.0	16338.0	MA	4.000000	#2	2.832479	48949	3.600000	Harvard University
2	32.0	1450.0	5941.0	8.0	42.0	27767.0	IL	4.000000	#3	2.832479	54825	2.803398	University of Chicago
3	32.0	1420.0	5472.0	6.0	50.0	18385.0	CT	3.543443	#3	2.832479	51400	3.400000	Yale University
4	32.0	1430.0	6113.0	6.0	48.0	21041.0	NY	3.543443	#5	2.832479	57208	3.800000	Columbia University
5	33.0	1460.0	4524.0	8.0	58.0	20331.0	MA	3.543443	#5	4.600000	49892	4.900000	Massachusetts Institute of Technology

Figure 3: Replacing the NULL values to mean

Figure 4: Graphical Representation: SAT score vs. Acceptance Rate

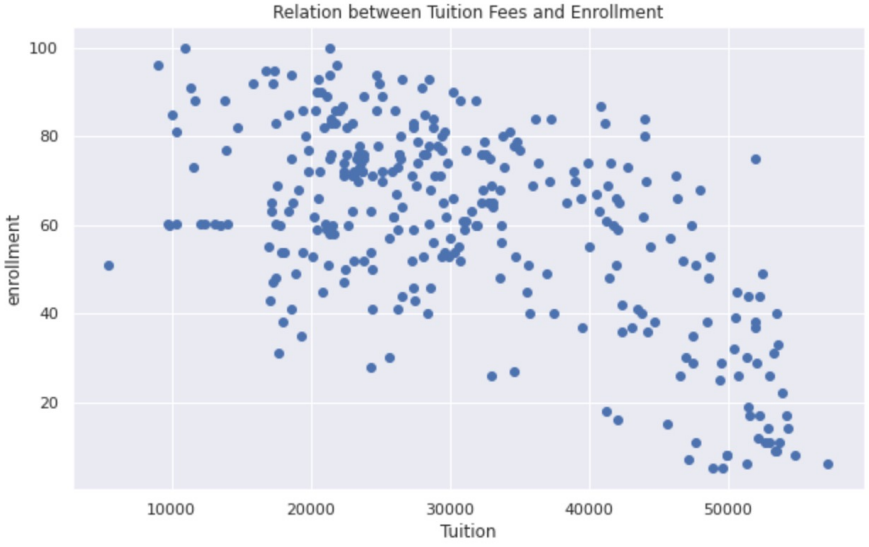

Figure 5: Graphical Representation: Tuition vs. Enrollment

```
combined_dataset_upd.describe().T
```

	count	mean	std	min	25%	50%	75%	max
GPA Median	196.0	3.442653	0.244236	2.80000	3.28000	3.45000	3.61000	3.93000
Accept	196.0	0.455969	0.149756	0.06900	0.34725	0.48100	0.56100	0.86100
S/F Ratio	192.0	7.178646	1.933857	3.50000	5.97500	6.90000	8.02500	17.00000
Employment at Grad	168.0	0.509863	0.189154	0.08600	0.37300	0.48600	0.61025	0.94900
Employment at 10Mos	196.0	0.748883	0.125892	0.27600	0.68100	0.76900	0.83500	0.96900
Median Salary Private	185.0	87446.286486	39691.587822	45500.00000	62400.00000	72500.00000	95000.00000	180000.00000
Median Salary Public	184.0	54461.510870	7092.035884	40000.00000	50000.00000	54791.50000	60000.00000	90518.00000
Tuition	197.0	42143.647208	11939.794275	13060.00000	32950.50000	42190.00000	50700.00000	69916.00000
SAT Median	196.0	156.323383	6.711503	142.01726	151.76726	155.01726	161.01726	173.01726

Figure 6: Descriptive Statistics: Law School Dataset

Checking the presence of NULL values in dataset - Law School

```
[42] combined_dataset_upd.isna().sum()
```

```
GPA Median                4
Accept                    4
S/F Ratio                 8
Employment at Grad       32
Employment at 10Mos       4
State                     0
Median Salary Private    15
Median Salary Public     16
University Name           0
Tuition                   3
SAT Median                4
dtype: int64
```

Figure 7: Checking the NULL values

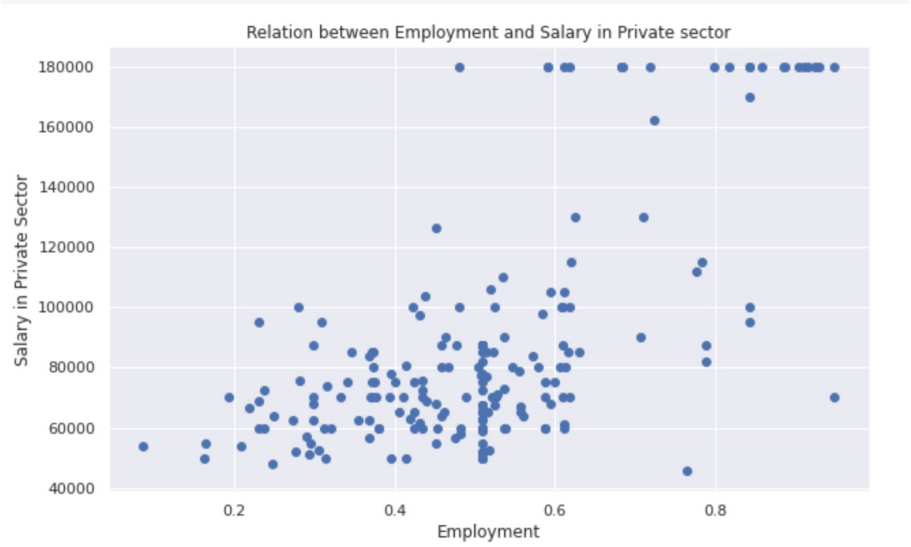

Figure 8: Graphical Representation: Salary vs. Employment Private

Figure 9: Graphical Representation: Salary vs. Employment Public

5 Modelling and results

We divided University Statistics data-set in ratio of 70:30 to train and test the Recurrent Neural Network model. To maintain the uniformity in the data-set we are using MinMaxScaler model from Sklearn, so that data-set values lie in the range 0 and 1. The model consist of 1 hidden layer with RELU as the activation function, ADAM as optimizer, and Mean Squared Error as the loss function. The model doesn't have any dropout layer.

After evaluation of Recurrent Neural Network on 30% validation set, we generated the law reputation score on the common university list consisting for 35 universities.

Results: The results generated for the law prediction score were as expected, because the Recurrent Neural Network model gave us values of the test data set which predicted the reputation score of the Engineering schools. We have visualized the reputation score data points of both Engineering and Law data-sets in scatter plots and it gives a clear idea of the projections done by the Recurrent Neural Network model.

Figure 10: Scatter Plot for Engineering Reputation score vs. Law Reputation Score

6 Future Scope

To progress with this model, if we find a data-set which contains attributes dependent on the their respective domains, we will be able to train the model and predict the data to classify the Universities into 3 classes: Good, Average and Bad Universities. Example: In University Statistics data-set, assuming the data has a unique attribute which defines the engineering attribute such as

"engineering knowledge" and law data-set has an attribute like "court points". Then it would possible for us to segregate/classify the Universities into different classes based on these attributes.

References

[1] 1. Bowers, T. A., Pugh, R. C. (1973). Factors underlying college choice by students and parents. Journal of College Student Personnel, 14(3), 220–224.

[2] 2. Angliss, K. An Alternative Approach to Measuring University Reputation. Corp Reputation Rev (2021). https://doi.org/10.1057/s41299-021-00110-y

Dataset:

[3] 1.https://www.kaggle.com/orrisis/us-law- schools?select=ABA Admissions.csv

[4] 2. https://www.kaggle.com/JobsPikrHQ/usa-based-job- data-set-from-300-companies

[5] 3.https://www.kaggle.com/jessemostipak/college- tuition-diversity-and-pay?select=salary potential.csv

7 Description of the code

We are submitting the code in a separate file with the report.