University Recommendation System

Akash Prabhudesai (51)
Zain Abbas (52)

Under the guidance of
Professor Nasim Shah

Agenda

- Problem Statement
- Approach
- Data Characteristics
- Data Preprocessing
- Solution Implementation
- Test Evaluation
- Summary

Problem Statement

- For an aspiring student who wants to apply for higher studies in other countries, university selection process is a challenging task.
- Lot of different criteria need to considered during application process based on individual's requirement.
- This problem can be addressed by modeling a recommender system based on various classification algorithms.
- In this project based on the Graduate and Undergraduate student dataset and user profile, a list of 10 best universities will be suggested such that it maximizes the chances of a student getting admission into those universities.

Approach

As we have large set of data and User profile, I planned to use Knowledge based recommendation techniques using two different models. For Graduate recommendations it is Case based knowledge recommendation and for Under graduate recommendations its constraint based recommendation.

- K-Nearest Neighbors
- Feature weighted algorithms.

K Nearest Neighbors

In KNN, the trained data is compared with test data and distances are calculated using Euclidean distance. It then classifies an instance by finding its nearest neighbors and recommend the top n nearest neighbor universities. Algorithm is stated as below.

- Initialize the value of k
- For getting recommendation, iterate from 1 to number of trained data.
- Calculate distance between test data and each row
- Sort the distances in ascending order
- Get top k rows and recommend to the user

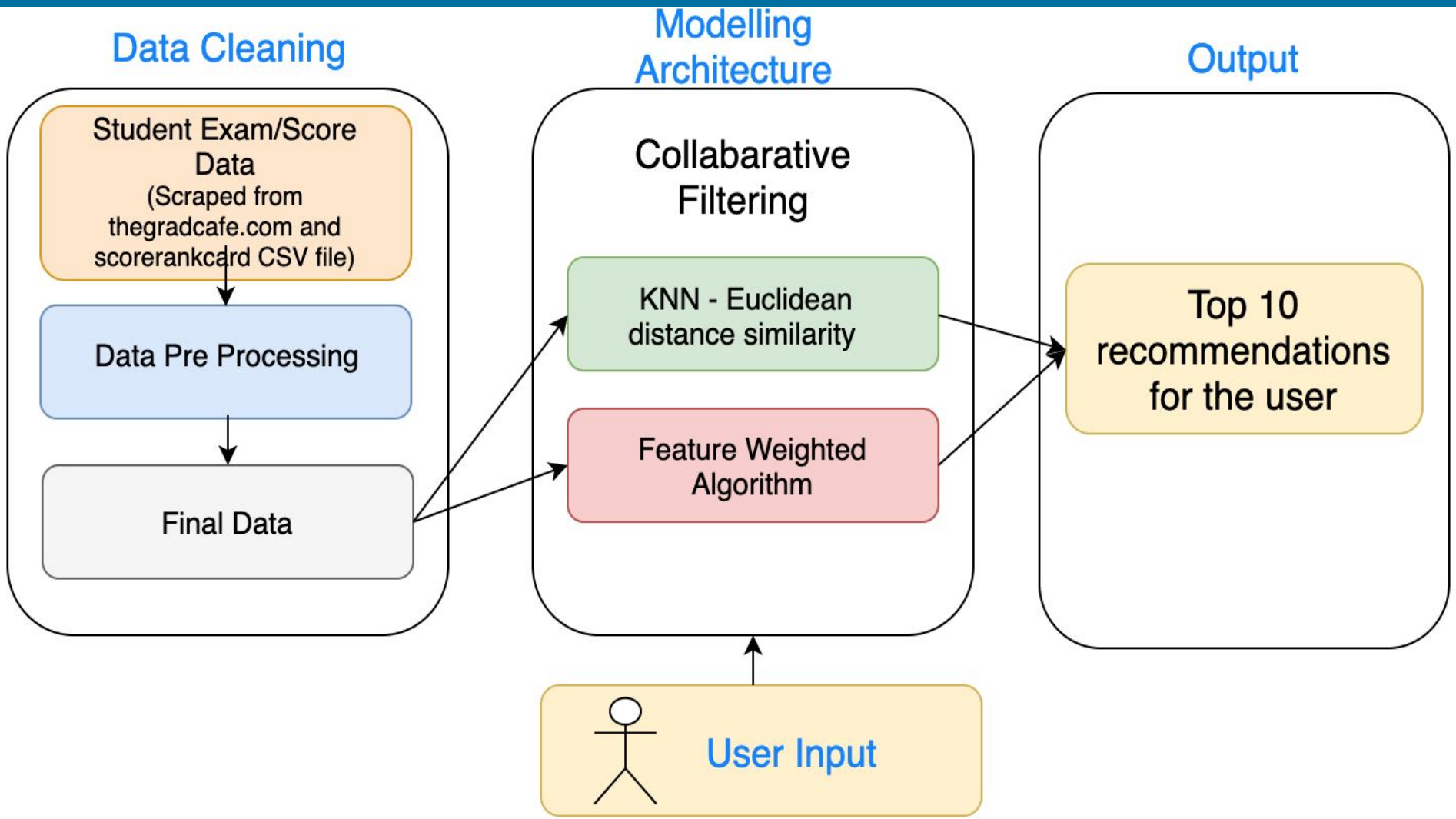
Feature Weighted Algorithm

The Weightage of all the features are taken and find the similarity score. Based on the similarity score, The universities with highest similarity is recommended.

Suppose w1 and w2 are the weights and f1 and f2 are the features the similarity is calculated by formula

$$\text{Similarity Score} = w1*f1+w2*(1-f1)$$

Block Diagram



Dataset

The first step in building any recommendation system is the identification of the data set.

- Graduate student data was scraped from www.thegradcafe.com and the Undergraduate student data was scraped from https://collegescorecard.ed.gov/data/.
- About 271807 rows of raw student data was obtained as a result of web scraping, which is being processed to use as final dataset.

Dataset Characteristics

- The Graduate Student data has the columns related to their GRE and TOEFL scores and University name they got admitted/rejected into.
- The Undergraduate Student data contains the columns related to SAT scores and in which University they got admitted/Rejected into.

Data Pre Processing

In order to use the obtained data for our analysis, preprocessing and cleansing of dataset is required. The following are the data processes we did in our project

- Data Scraped from website are loaded into different files. Merged all data.
- Data from the admitted student rows are taken and rejected student rows are deleted.
- Column names are set to the dataset.
- The new/old GRE scores were also cleansed

Data Pre Processing….ctnd

- Null values are deleted or filled with appropriate values
- GPA scores available were based on different point systems, so all the GPA scores were uniformly scaled to 4 point scale by using normalize functions.
- All the unnecessary columns are dropped.
- Changed the order of columns making train and test dataset for algorithm

Data Pre Processing….

Initial dataset after web scraping the data and merging the csv files.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
0	University Of Waterloo	Systems Design Engineering	MS	NaN	Accepted	Website	(1, 7, 2019)	1.561964e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	(10, 7, 2019)	1562742000	
1	Northeastern University	Electrical Engineering	PhD	F19	Rejected	Website	(8, 7, 2019)	1.562569e+09	NaN	NaN	NaN	NaN	NaN	NaN	NaN	(10, 7, 2019)	1562742000	Gc
2	The University Of Auckland	Electrical And Electronic Engineering	MS	NaN	Accepted	Website	(19, 6, 2019)	1.560928e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	(9, 7, 2019)	1562655600	Student\ı 7.5\nCoı Super
3	Radford University	Counseling Psychology PsyD.	Other	F19	Accepted	Phone	(4, 3, 2019)	1.551686e+09	NaN	NaN	NaN	NaN	NaN	NaN	American	(9, 7, 2019)	1562655600	
4	University Of Chittagong	Computer Science	MS	NaN	NaN	Other	(9, 7, 2019)	1.562656e+09	3.2	163.0	168.0	4.0	True	NaN	International	(9, 7, 2019)	1562655600	

Data Pre Processing….

Dataset after setting the names of the columns

```
data.columns = ['univName', 'major', 'program', 'season', 'decision', 'Method', 'decdate', 'decdate_ts', 'cgpa', 'greV'
           'greA', 'is_new_gre', 'gre_subject','status', 'post_data', 'post_timestamp', 'comments']
data.head()
```

	univName	major	program	season	decision	Method	decdate	decdate_ts	cgpa	greV	greQ	greA	is_new_gre	gre_subject	status	post
0	University Of Waterloo	Systems Design Engineering	MS	NaN	Accepted	Website	(1, 7, 2019)	1.561964e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	
1	Northeastern University	Electrical Engineering	PhD	F19	Rejected	Website	(8, 7, 2019)	1.562569e+09	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	The University Of Auckland	Electrical And Electronic Engineering	MS	NaN	Accepted	Website	(19, 6, 2019)	1.560928e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	
3	Radford University	Counseling Psychology PsyD.	Other	F19	Accepted	Phone	(4, 3, 2019)	1.551686e+09	NaN	NaN	NaN	NaN	NaN	NaN	American	
4	University Of Chittagong	Computer Science	MS	NaN	NaN	Other	(9, 7, 2019)	1.562656e+09	3.2	163.0	168.0	4.0	True	NaN	International	

Data Pre Processing….

Description of the numerical data

```
data.describe()
```

	decdate_ts	cgpa	greV	greQ	greA	gre_subject	post_timestamp
count	6.145400e+04	55589.000000	61474.000000	61474.000000	61474.000000	7175.000000	6.147400e+04
mean	1.431551e+09	3.715970	231.556333	248.826447	4.144757	796.411150	1.431763e+09
std	9.540728e+07	0.506153	174.575147	208.551820	1.111126	122.305977	8.079993e+07
min	-1.000000e+00	0.400000	130.000000	130.000000	0.000000	310.000000	1.263283e+09
25%	1.363244e+09	3.520000	155.000000	157.000000	3.500000	710.000000	1.363417e+09
50%	1.426662e+09	3.750000	161.000000	164.000000	4.000000	800.000000	1.427094e+09
75%	1.490771e+09	3.900000	167.000000	170.000000	5.000000	890.000000	1.491030e+09
max	1.360120e+10	9.990000	800.000000	800.000000	6.000000	990.000000	1.562569e+09

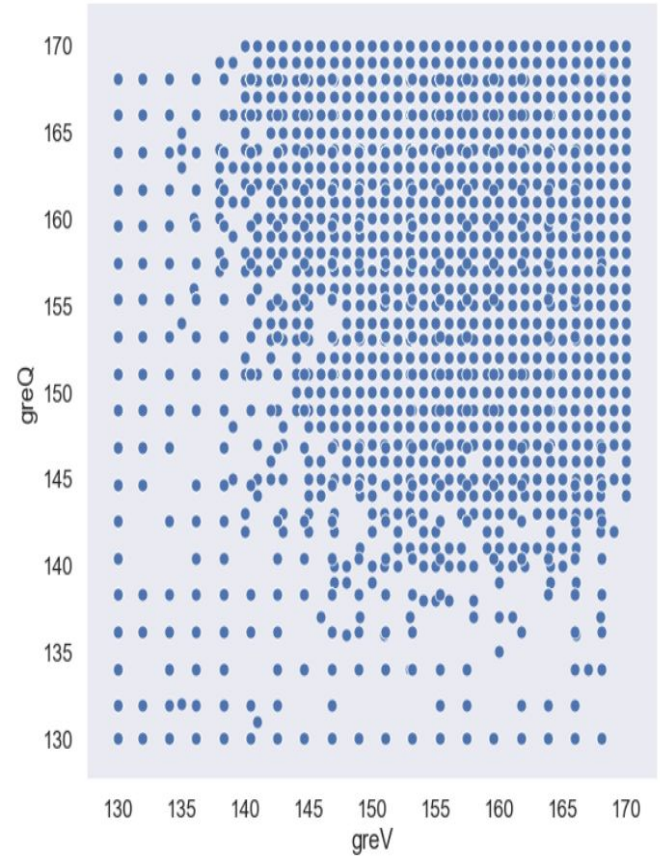
Data Pre Processing....

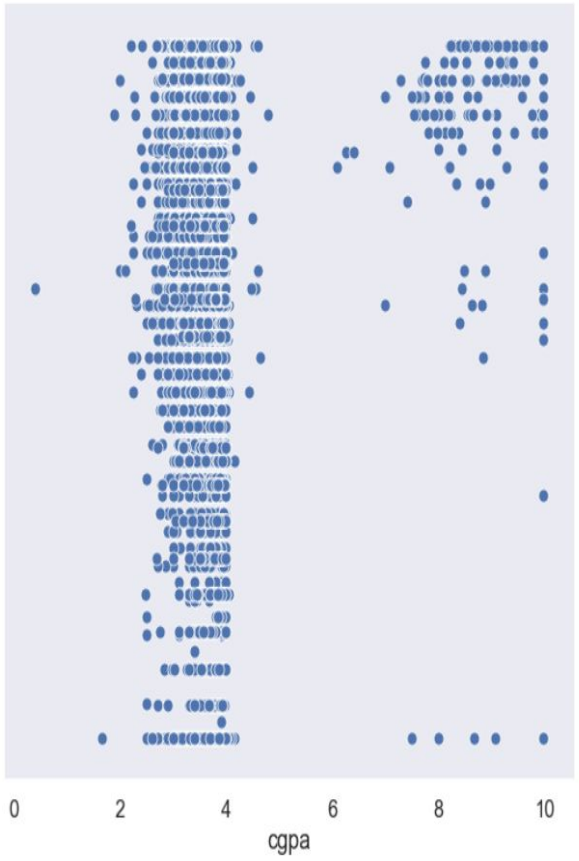
Converting the old GRE scores to new GRE scores

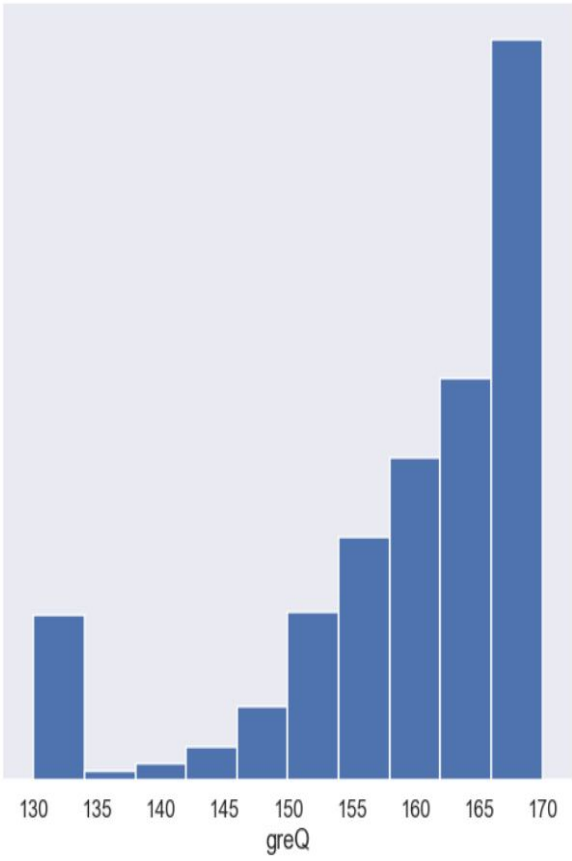
```
def convert_quant_score(quant_score):
    quant_list = []
    quant_score = quant_score.tolist()
    for old_quant in quant_score:
        if old_quant <= 170:
            quant_list.append(old_quant)
            continue
        else:
            old_quant = old_quant/4.7
            if old_quant <=130:
                quant_list.append(130)
            else:
                quant_list.append(old_quant)
    return quant_list

def convert_verbal_score(verbal_score):
    verbal_list = []
    verbal_score = verbal_score.tolist()
    for old_verbal in verbal_score:
        if old_verbal <= 170:
            verbal_list.append(old_verbal)
            continue
        else:
            old_verbal = old_verbal/4.7
            if old_verbal <=130:
                verbal_list.append(130)
            else:
                verbal_list.append(old_verbal)
    return verbal_list
```

Data Pre Processing….

Plots of the greV and greQ data.







Data Pre Processing….

Processed data which has only University name, gre scores of the user.

	univName	cgpa	greV	greQ	greA
14	Ohio State University	4.00	150.0	166.0	3.0
17	Texas A&M University	3.57	157.0	151.0	5.5
46	University Of California, Irvine	3.66	155.0	167.0	4.0
64	Boston University	3.10	161.0	157.0	4.0
203	Oregon State University	3.38	154.0	170.0	4.0

Solution Implementation

- In KNN, trained and Test data will be sent to algorithm to find the Euclidean distance and the top 5 nearest neighbors are taken into consideration.

- The features like GREA, GREV, GREQ, CGPA of the user as test data are taken as features and provide weightage to them to find the similarity score. and provide top 5 recommendations to the user.

- Developed a data analysis notebook where user can easily provide his/her score details to the application and get recommendations.

Test Evaluation

- As there is analysis application to the recommendation system, any user can test the website any time.
- Planning to develop the feedback page where users can come back and provide feedback to the website and provide data in which University he/she got the admit. (if time permits)
- By capturing admission information I can train my algorithm to increase accuracy.

Summary

- In short, This project will help students in decision making of which University to choose for their higher education in other countries like USA.
- with the User friendly web Interface which takes in test scores, helps students apply to the Universities in which there is a high chance of getting the admit.
- Data Mining techniques like KNN and Feature weighted algorithms are used

References

- https://www.semanticscholar.org/paper/Recommender-System-for-Graduate-Studies-in-USA-Suresh/22924fda3f293f80a3f62f32799c08d0b81a9b20
- http://://ieeexplore.ieee.org/document/7760053

Thank You