

# High School Scoring System Based on TOPSIS

UConn Team 2

Haoxian Ruan, Botao Wang, Shubham Rajaram

## Abstract

The report introduces the analysis process and the technical method used in the project. We first perform exploratory data analysis and then select the features we need, then we process the data so that we get the 5 factors to be used to score. We finally score the high scores by TOPSIS method with the weights calculated by analytic hierarchy process.

## Objective

We utilize data provided by the UConn admissions and analytics teams to create a novel scoring mechanism for domestic high schools. The score will be utilized to deploy marketing and outreach resources in the most efficient way possible in out-of-state markets.

In this project, there are 5 factors to be considered: brand power (yield rate), applicant's academic level (APPLICATION\_ASSESSMENT), applicant's distance to UConn (distance), college going rate (a\_hs\_college\_access) and financial background (a\_hs\_medfaminc\_ptile). And our scoring mechanism is based on these 5 factors.

## Exploratory Data Analysis

After merging all the dataset together on key [PUBLIC\_ID], there are 225 columns. We don't need to use all the columns since some of them are not useful for us to achieve our goal and some of them have too much missing values, so we just select the columns we need.

And the applicants mainly consist of three terms: 1218 (2021 fall), 1228 (2022 fall), 1238 (2023 fall). And we will calculate the score of each term and take the average of all the term as the final score.

## Technical Method Introduction

### 1. Comprehensive Assessment

AHP (Analytic Hierarchy Process) is a decision-making framework that uses a multi-level hierarchical structure to model a decision problem and then applies pairwise comparisons and mathematical calculations to derive priority scales.

It is commonly used in decision making, where decisions are made on merit by calculating which option has the highest score. We scored different high schools in this project by treating them as different options.

And in this project, we use the idea of AHP to get the weight of each factor and then we can use the weight to calculate score by TOPSIS method.

### (1) Weighting Calculation

We begin by subjectively ranking the importance of five factors. We think that when the admission team consider the importance of factors to evaluate high school, the rank is:

yield rate > APPLICATION\_ASSESSMENT > distance > a\_hs\_college\_access > a\_hs\_medfaminc\_ptile

And we can construct a pairwise comparison matrix to represent the relationship of importance. Each value in the matrix represents relative

importance between two factors. We can use  $a_{ij}$  represents the value of the  $i^{th}$  row  $j^{th}$  column.

If  $a_{ij} = 1$ : factors i and j have the same importance.

If  $a_{ij} > 1$ : factor i is more important than factor j. For example, if  $a_{ij} = 3$ , it means that factor i is 3 times more important than factor j.

If  $a_{ij} < 1$ : factor i is less important than factor j.

	yield rate	APPLICATION_ASSESSMENT	distance	a_hs_college_access	a_hs_medfaminc_ptile
yield rate	1	3	5	7	9
APPLICATION_ASSESSMENT	1/3	1	1.6	2.5	3
distance	1/5	1/1.6	1	1.5	2
a_hs_college_access	1/7	1/2.5	1/1.5	1	1.5
a_hs_medfaminc_ptile	1/9	1/3	1/2	1/1.5	1

We subjectively rate the importance relationships between factors and construct the matrix above.

Based on the pairwise comparison matrix, we can get the weights of each factor by eigenvector method. And we get the weights:

yield rate: 0.56, APPLICATION\_ASSESSMENT: 0.19, distance: 0.12, a\_hs\_college\_access: 0.08, a\_hs\_medfaminc\_ptile: 0.06.

And then we need the consistency check to ensure the reliability of the judgment matrix, thereby increasing the credibility of the comprehensive evaluation results.

We need to calculate the consistency ratio (CR), when  $CR < 0.1$ , the consistency of the pairwise comparison matrix is acceptable. when  $CR \geq 0.1$ , the consistency is poor, requiring adjustment of matrix.

The consistency ratio of the pairwise comparison matrix we construct is 0.001, so the pairwise comparisons are consistent and logically coherent. And we can use the weights for further calculation.

### (2) Score Calculation

We can directly calculate the score by multiplying weight and value for each factor, but we think that TOPSIS is a reasonable and interpretable way to get the score.

The basic process of TOPSIS is to use the cosine method to identify the optimal and worst alternatives among the finite set of options, and then calculate the distances between each evaluation object and the optimal and worst alternatives. This results in the relative closeness of each evaluation object to the optimal solution, which serves as the basis for evaluating their performance. The optimal and worst alternatives are ideal, the closer to the optimal alternatives, the higher the score, the closer to the worst alternatives, the lower the score. For example:

$D_i^+$ : the distance of  $i^{th}$  high school to the optimal alternatives

$D_i^-$ : the distance of  $i^{th}$  high school to the worst alternatives

The score of the  $i^{th}$  high school:  $\frac{D_i^-}{D_i^+ + D_i^-}$

And the optimal high school we want is the one with:

- ① lowest yield rate: we tend to give high score to the high schools with low yield rate. Since UConn admissions team need to go to these high school to set up a brand power instead of going to the high schools with high yield rate. And they can also go into the reasons why the student would not choose UConn even if they are admitted by then.
- ② lowest APPLICATION\_ASSESSMENT: the lower APPLICATION\_ASSESSMENT, the higher academical level the applicants. And we think that UConn admissions team tend to admit applicants with high academical level.
- ③ shortest distance: the distance factor considers the average distance from each student's home to UConn within a high school. Students with short distance are more likely to visit the campus, participate in university events, build their familiarity and comfort with UConn. Moreover, shorter distances lower travel costs and time for both the admissions team and prospective students, making it easier to establish and maintain strong relationships.
- ④ highest a\_hs\_college\_access: the high schools with high a\_hs\_college\_access have more students go to 4-year college. And we think that UConn admissions team tend to let these students choose UConn as their future college.
- ⑤ highest a\_hs\_medfaminc\_ptile: the tuition are high for out-of-state students. The higher a\_hs\_medfaminc\_ptile means that students are more likely to afford out-of-state rates.

## 2. Distance Calculation

We first want to consider the distance from UConn to each high school, but the given data does not tell us the location of each high school, so we are not able to

calculate the distance. And the column geo in the Landscape dataset tells us the Geographic ID of each applicant, so we can use it to calculate the distance from the applicants' home to UConn, and we select the column APPLICATION\_FIRST\_CHOICE\_CAMPUS as the destination for distance calculation for each applicant.

The geo id here is presented as 11-digit, and we extracted the first 5 digits to locate. By this way, we can get the address down to the county. These addresses are then batch processed to generate the corresponding coordinates, and we can get the distance by calculating the distance between coordinates.

### **3. Yield Rate Calculation**

To calculate the yield rate, we need label applicants if they are admitted and enrolled. We checked if applicants are admitted and enrolled by combining several datasets with multiple columns for consideration.

#### **(1) Check Admitted**

We check the columns from [Rank 1 Released Code] to [Rank 5 Released Code], if AT (admit) appears in any Rank Released Code columns, we label the row (applicant) as admitted.

#### **(2) Check Enrolled**

We check enrolled by several situations.

First, if DEIN (Deposit Enrolled) appears in any Rank Released Code columns, we label the row (applicant) as enrolled.

However, some applicants may withdrawal their enrollment, so those applicants should not be label as enrolled. If WADM (Withdrawal - Administrative) or WAPPDD (Withdrawal - Declined Decision) or WAPP (Withdrawal - Applicant) appears in any Rank Released Code columns, we label the row (applicant) as not enrolled.

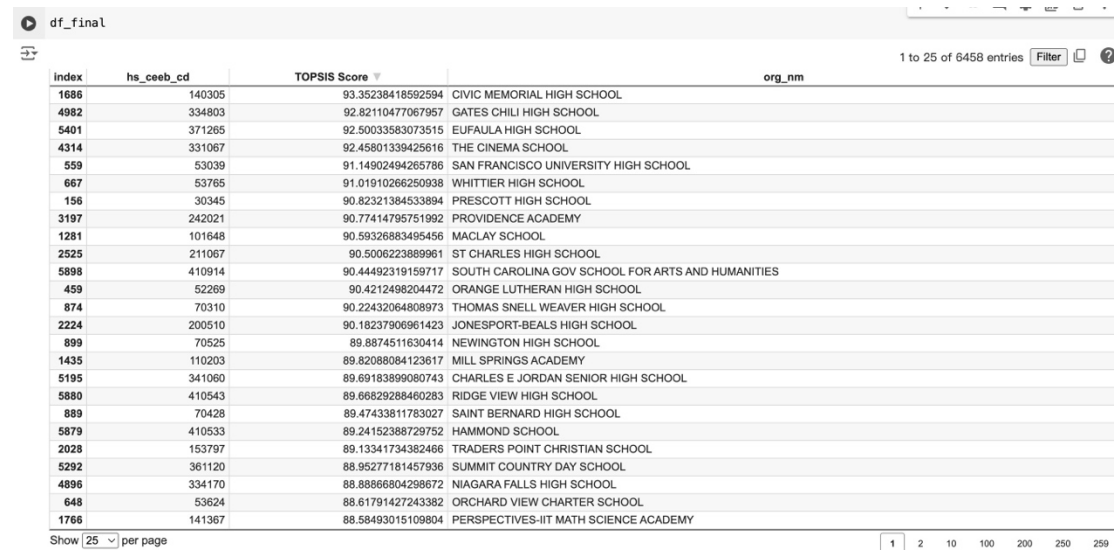
To make our label non-missing and to avoid missing data problems due to the dataset itself, we also combine the column [UC\_APPLS\_ENROLLED\_10TH\_DAY] in Enrollment Status dataset to check enrolled. If the value of [UC\_APPLS\_ENROLLED\_10TH\_DAY] is 1, we label the applicants as enrolled; if it is 0, we label the applicants as not enrolled.

And then we calculate the sum of the admitted applicants and enrolled

applicants for each high school, and the yield rate is  $\frac{\text{admitted count}}{\text{enrolled count}}$ .

## Result Analysis

We calculate the score of each high school, the high schools here are not all the high schools in the given data. The high schools that do not present here due to the missing value of some features or having no applicants of the terms we focused on. And the results of the score are shown partially in the image below. We can see that CIVIC MEMORIAL HIGH SCHOOL (140305), GATES CHILI HIGH SCHOOL (334803), EUFAULA HIGH SCHOOL (371265) are the top 3 high schools in our score mechanism.



index	hs_eeeb_cd	TOPSIS Score	org_nm
1686	140305	93.35238418592594	CIVIC MEMORIAL HIGH SCHOOL
4982	334803	92.82110477067957	GATES CHILI HIGH SCHOOL
5401	371265	92.50033583073515	EUFAULA HIGH SCHOOL
4314	331067	92.45801339425616	THE CINEMA SCHOOL
559	53039	91.14902494265786	SAN FRANCISCO UNIVERSITY HIGH SCHOOL
667	53785	91.01910266250938	WHITTIER HIGH SCHOOL
156	30345	90.82321384533894	PRESCOTT HIGH SCHOOL
3197	242021	90.77414795751992	PROVIDENCE ACADEMY
1281	101648	90.59326883495456	MACLAY SCHOOL
2525	211067	90.5006223889961	ST CHARLES HIGH SCHOOL
5898	410914	90.44492319159717	SOUTH CAROLINA GOV SCHOOL FOR ARTS AND HUMANITIES
459	52289	90.4212498204472	ORANGE LUTHERAN HIGH SCHOOL
874	70310	90.22432064808973	THOMAS SNELL WEAVER HIGH SCHOOL
2224	200510	90.18237906961423	JONESPORT-BEALS HIGH SCHOOL
899	70525	89.8874511630414	NEWINGTON HIGH SCHOOL
1435	110203	89.82088084123617	MILL SPRINGS ACADEMY
5195	341060	89.69183899080743	CHARLES E JORDAN SENIOR HIGH SCHOOL
5880	410543	89.66829288460283	RIDGE VIEW HIGH SCHOOL
889	70428	89.47433811783027	SAINT BERNARD HIGH SCHOOL
5879	410533	89.24152388729752	HAMMOND SCHOOL
2028	153797	89.13341734382466	TRADERS POINT CHRISTIAN SCHOOL
5292	361120	88.95277181457936	SUMMIT COUNTRY DAY SCHOOL
4896	334170	88.88666804298672	NIAGARA FALLS HIGH SCHOOL
648	53624	88.61791427243382	ORCHARD VIEW CHARTER SCHOOL
1786	141367	88.58493015109804	PERSPECTIVES-IIT MATH SCIENCE ACADEMY

## Summary

We first come out with the idea of clustering, but we think that clustering has some drawbacks. For example, there are 2 high schools, hs\_1 and hs\_2. hs\_1 has a little bit lower yield rate than hs\_2, and hs\_2 has a little bit shorter distance than hs\_1, and we cannot compare which of them has the higher score. So we think we need the method of calculating weights to score the high schools.

Our method also has drawback: the pairwise comparison matrix in AHP is subjectively constructed by us, so the weights we got are also subjective. One solution to this drawback is that we can let the experts to construct the matrix or to score the importance of each factor.

Apart from the modelling methods, we think that we can consider more factors that can decide whether to deploy marketing for a high school.

To conclude, we have developed a well-founded and transparent scoring system that admissions teams can refer to.