IT5006 Project Report

Building a University Recommendation Engine

Chong Si Qing (A0158925R)
Lee Ming Xuan (A0002533Y)
Premi Jeevarathinam (A0268262Y)
Venessa Tan (A0268463U)

1. Kaggle Submission

1.1 Initial Approach

In the initial submission, a basic model was employed using Ordinary Least Squares (OLS) regression. The correlation between features was analysed to identify highly correlated ones, and LASSO regression was used to determine the significance of each feature's impact on completion rate. The selected features were then fitted as linear features to OLS regression and checked for linearity and heteroscedasticity. The residual plot revealed some form of non-linearity, which prompted the model to be fine-tuned. To address this issue, polynomial features (i.e. squares of certain features) were added to capture the non-linear relationship between the predictors and the response variable.

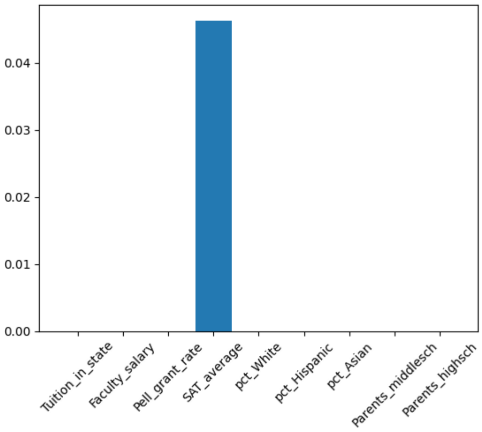


Fig. 1: Lasso Regression Results

1.2 Evaluation of Initial Approach

One point of interest in our initial data exploration was the discovery that many of the features had weak significance to completion rate (Fig. 1). Hence tuning the OLS model with polynomial features did not help the weak features. While there was improvement in the R-score, it was not possible to push it much higher (from ~0.65 - 0.7 to 0.7 - 0.75). There was also risk of overfitting if we were to keep tuning the OLS model.

1.3 Improvements to Model

To address weak features, some heteroscedasticity and variance in our dataset, the team explored other models and opted for XGBoost. Boosting, regularisation, and bagging were techniques deployed in XGBoost which helped to address these issues. We also tuned the hyperparameters to prevent overfitting. The XGBoost model improved prediction accuracy, through the increase in mean R-score from K-fold cross validation while ensuring that there was no overfitting as there was no increase in the standard deviation of the R-score.

Model	Mean R-Score	Std R-Score
1st submission (OLS with poly features)	0.768	0.021
XGBoost	0.828	0.023
XGBoost after tuning	0.849	0.020

Table 1: OLS Regression Results

1.4 Conclusion

A significant improvement in the R-score was observed when moving from the polynomial model to XGBoost. However, some heteroscedasticity was still evident in the results at the highest and lowest ends of the completion rate. This observation is likely due to the fact that the top and bottom schools have more distinct profiles, making it generally easier for models to predict them. It was noted that this heteroscedasticity is a feature of the data and further tuning should be avoided to prevent overfitting.

2. College Scorecard Data Analysis

2.1 Motivation

Choosing a university is a crucial decision for those considering higher education. With student debt being at an all-time high of $1.6 trillion in the US, students have the difficult task of weighing their potential earning power against the present cost of studying. The decision on what university to choose is further complicated by other factors, such as one's likelihood of completing their studies and repaying their loans. As such, we wanted to develop a recommendation engine to help students make better-informed decisions on their higher education options.

2.2 Objectives and Outcomes

Our recommendation engine will take in a student's profile, which includes their SAT scores, race, family income, desired field of study, desired region and desired locale, and provide a list of recommended universities that best fit their profiles, as well as their projected earnings from this selection of schools.

2.3 Dataset

We used data from the College Scorecard dataset, which contains a wide range of information on colleges and universities in the United States. We have chosen to use institution-level data due to its richness in features including institutional data, students' profiles, and domains of study, all of which are important variables in predicting earnings. We extracted the following features for analysis:

- **Location of college**: State, Region, Latitude/Longitude, Locale
- **Attributes of College**: Full-time Faculty Rate, Student Expenditures, Faculty Salary, Admission rate, Completion rate
- **Background of students**: Percentage of each race in institution, Parent's education, Family Income
- **Field of study**: Course offered by university

To ensure that our analysis is targeted, we limited our scope of data to universities that predominantly offer undergraduate programs. Finally, to accurately predict earnings based on recent trends, we extracted the most recent data. We decided to not use time series data as we would need other features to help predict the growth of earnings from year to year (e.g. economic situation) which is outside the scope of this model.

2.4 Our Approach

Our approach can be summarised as follows:

1. **Train a regression model using university-specific data to predict earnings:** We first train a regression model using university-specific data i.e. the 'School' dataset, with earnings as the target. Features include Admission Rate, Average SAT score, Completion rate, Demographics (Percentage of each race in institution), Parents' education, Full-time Faculty Rate, Expenditures per student, Faculty salary, Locale and Region. The regression model will thus be built based on all the colleges within the dataset.
2. **Train a second model using student-specific data to find the most similar colleges based on a student's profile:** We train an unsupervised learning model using student-specific data such as SAT score, race, family income, parents' education levels, desired field of study, desired region of study (e.g. Southwest, Northeast) and desired locale of study (e.g. City, Suburb). After filtering out schools that fit the student's preferences (i.e. field, region and locale of study), the model will be trained using the remaining features i.e. the 'Student' dataset, and will provide a list of most similar schools based on the student's profile.
3. **Combine the 2 models to predict earnings based on a student's profile and preferences:** When a student provides their inputs, we first use the clustering model to provide a list of similar schools. From the list of schools, we then generate an aggregated feature vector that can be fitted into the regression model to predict the student's projected earnings.

2.5 Data Preprocessing

To ensure the quality and usability of the data for our analysis, we went through through a series of preprocessing steps:

1. **Handling Missing Values:** To address the issue of missing values in the dataset, we dropped all rows with NA values, and as a result reduced the number of colleges from around 2,000 to between 800 and 1,000. While this represented considerable data loss, this step was necessary to prevent any biases or inaccuracies that may arise from using incomplete data.

2. **Aggregating some of the features**: The original dataset contained 12 different categories for the Locale variable. To streamline our analysis and filtering, we aggregated these categories into four broader groups (City, Town, Suburb, and Rural). Similarly, we condensed the 50 US states into five regions (Northeast, Southeast, West, Southwest and Midwest) to facilitate the filtering process. This step ensures that a student will have a wider range of colleges to choose from based on a broader region rather than a single state.
3. **Feature Scaling**: Finally, we performed feature scaling on both School and Student datasets using a Standard Scaler, which ensured that no particular feature would have a stronger influence on our models due to differences in the scales of the input variables, thus mitigating potential biases in the model's predictions.

2.6 Feature Selection

To select the features for our regression model, we first visualised each feature’s relationship to earnings, and identified that the variables most correlated with earnings were Faculty Salary and average SAT scores. We then performed feature correlation analysis and dropped features that had a high correlation with each other (>0.7) to reduce collinearity in our model. However, we decided to retain both Faculty Salary and SAT scores for testing with the regression model, as both features showed a strong correlation to earnings. Finally, we performed LASSO regression on two versions of the dataset, one with Faculty Salary and the other with SAT score, to determine the significance of each feature whether we could further drop some of them.

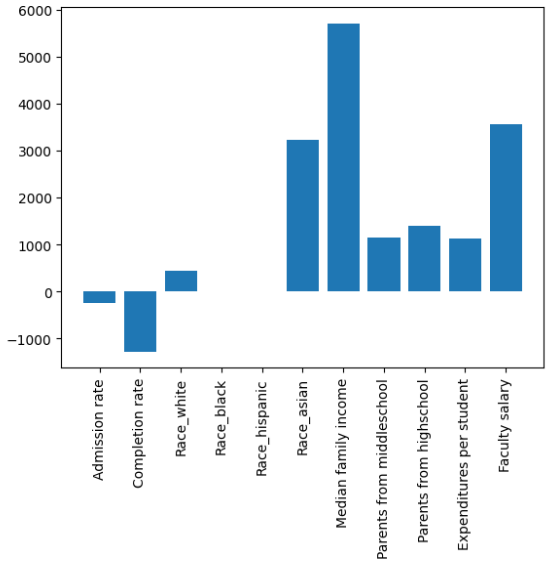


Fig. 2: LASSO result for dataset with Faculty Salary

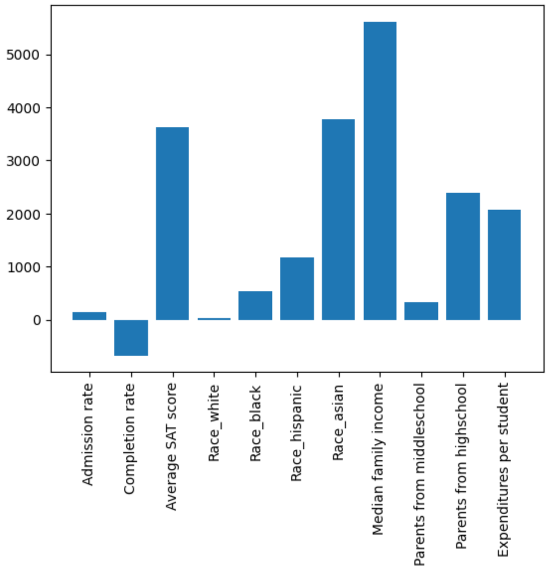


Fig. 3: LASSO result for dataset with SAT score

2.6 Model Selection and Tuning

2.6.1 Clustering Model Selection

For the unsupervised learning model, we chose to go with K-Nearest Neighbours. While there were many other recommended approaches in literature, we were unable to use them as we did not have granular data about the student level of the colleges. To tap on other recommender models (e.g. content-based filtering), it would be necessary to compare individual student’s profiles.

As we did not have a benchmark to compare with, the distance score from K-Nearest Neighbours wasn’t helpful in quantifying how accurate or good the model is. However, the recommended colleges did have similar SAT scores, race and family background to a student’s profile, which we felt was sufficient.

We choose to select the top 15 schools from our K-Nearest Neighbour model to form a student’s profile that would fit into the regression model. This ensured that the predicted earnings had about an average of 5% difference (capped at 10% difference) from the average median earnings of the recommended colleges from the K-Nearest neighbours model which we felt was sufficiently accurate.

Due to the limited dataset of colleges, we can sometimes be left with less than 15 colleges after filtering the full list by a student’s subject and location of interest. Even having only 50 colleges for the model can sometimes result in huge variance in the profiles of schools recommended due to the nature of the dataset. This can be observed when measuring standard deviation of the median family income of the recommended schools.

2.6.2 Regression Model Selection and Hyperparameters Tuning

For the regression model, we had picked 4 models:

1) **ElasticNet** - this is a linear regression model with both L1 and L2 regularisation technique. We felt regularisation would be important as there were multiple weak features that we would like to include in the model, and this could help reduce overfitting
2) **AdaBoost** - this model uses gradient boosting with the decision tree as weak learners. We hope boosting could help to improve parts of the data that might be skewed (e.g. top and bottom colleges)
3) **RandomForest** - this model uses bagging with the decision tree as learners. We felt that bagging will be useful given the variance of median earnings in similar profile of colleges.
4) **XGBoost** - an ensemble technique using all 3 techniques (i.e. regularisation, boosting and bagging) from the previous 3 models with the base decision tree model. We hoped to combine all 3 techniques to further improve our model.

From using the above models with default hyperparameter, it was clear that the feature set with SAT performed better (R-score was generally higher and there was less overfitting, see figure below) and the Random Forest and XGBoost models gave better results. This would suggest that our hypothesis that bagging will be important due to the large variance of median earnings of schools with similar profiles is accurate.

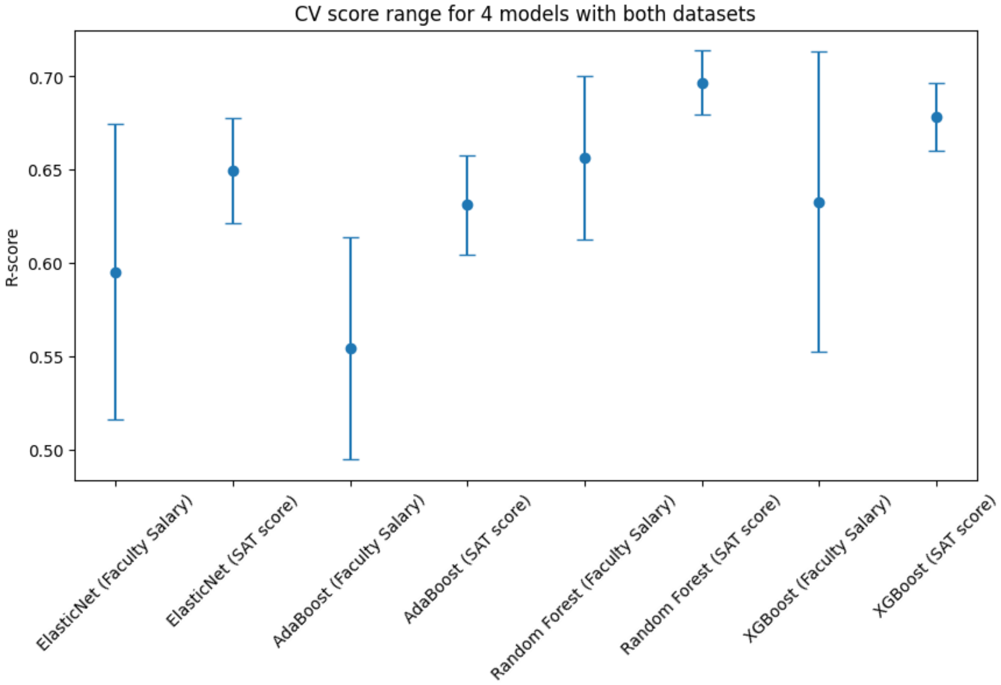
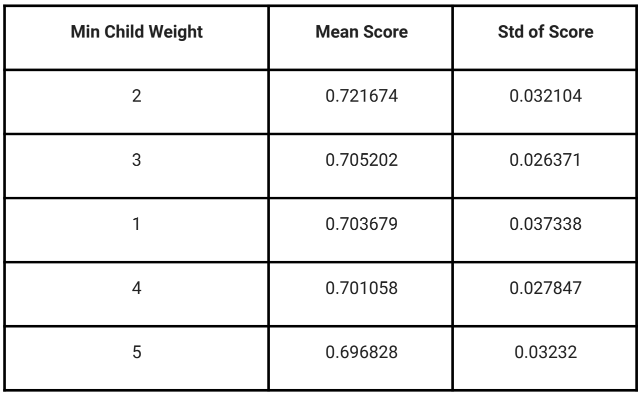


Fig. 4: Result from basic model testing

While we had tried to drop different features in the dataset to improve the R-score before hyperparameters tuning, it did not help much and so we kept all the features.



Min Child Weight	Mean Score	Std of Score
2	0.721674	0.032104
3	0.705202	0.026371
1	0.703679	0.037338
4	0.701058	0.027847
5	0.696828	0.03232

Table 2: Sample of Hyperparameter Tuning results

When tuning hyperparameters, there is usually a tradeoff between accuracy and overfitting. As seen from the scores in table 2, while mean score improved when changing the child weight from 3 to 2, standard deviation increased (i.e. more overfitting).

We used the grid search technique for hyperparameter tuning, and landed with XGBoost as the best model with a decent mean R-score of ~0.72 and standard deviation of ~0.03 from K-fold cross validation. Random Forest also produced decent results of ~0.71 mean R-score with ~0.03 as standard deviation. Samples of results from our model can be found in Annex A and B.

2.7 Qualitative Analysis

2.7.1 Cost and Benefit

In this paper, we have adopted a two-model approach, utilising K-Nearest Neighbours clustering to identify closest schools, and XGBoost to predict earnings, returning a list of top recommended colleges that represent the best potential return on education investment.

We are aware that there are alternative approaches to developing a recommender system, including Content-Based Filtering and Matrix Factorisation. Such processes can be more computationally efficient than our two-step approach and if conducted on a larger scale/in the long-term, could be cheaper to run. However, in the context of this project - helping prospective students make a well-informed decision on a small-scale, we have prioritised providing a highly-personalised and explainable model. Further, we recognise that there are source dataset limitations: in particular, the lack of granular data on a user-college interaction level, which impacts the effectiveness of a Matrix Factorisation approach. In addition, a Content-Based Filtering approach may also result in over-specialisation, and limit the appeal of the model as a discovery tool.

2.7.2 User Feedback

We interviewed three individuals who have already been through the US college research process, to obtain their feedback on the relevance and usefulness of our recommendation engine, and how they compared to their college search experience. The individuals found the recommender useful as a discovery tool, and would be interested to research more about the top recommended schools based on earnings. However, they also did point out the complex nature of college school selection - beyond predicted earnings, choice of college is also affected by factors such as cost, programme, and school brand. Further details can be found in Annex C.

2.7.3 Evaluation and Potential Applications Going Forward

Thus far, we have focused on potential earnings as a key determinant of college choice (since it ties closely with the cost of education). Referencing user feedback, we note that choice of college can be driven by other motivating factors. Going forward, the recommender could be tweaked to determine other predictive outcomes such as likelihood of college completion and debt repayment.

In addition, we recognise that we have featured a limited number of case studies. If a longer-term study were possible, deeper analysis can be done via A/B testing to compare the relevance and usefulness of the model against traditional methods of college research and selection. A/B testing requires time as it involves following subjects through: i) the college research and selection process and ii) the completion of their college education and post-employment earnings. This could be a more robust way to better measure the performance of the model.

2.8 Conclusion

In summary, we have developed a recommendation engine that takes in a student's unique characteristics, and provides them with a personalised list of top universities that are most likely to provide the best return on their education investment.

As there is no one-size-fits-all college, our recommender serves to provide prospective students an additional discovery tool that is data-based and customised to each users' unique background, to help them make an informed decision.

With this use case in mind, we have prioritised personalisation and explainability. From user feedback, results are promising in terms of exposing users to new colleges that they have not previously considered. Enhancements (weighting, multiple option selection) can be made to allow for more flexibility in selection and more fine-tuned results. As complex motivations are involved in college selections, there could be other applications for the recommendation engine going forward, such as tweaking to predict college completion and debt repayment instead.

3. Annex

Annex A: Random Student Profiles fitted into K-Nearest Neighbour and Regression Model

No. of Schools left after filter	Avg difference of SAT Score between student and schools recommended	Standard deviation of SAT Score for schools recommended	Avg difference of family income between student and schools recommended	Standard deviation of family income for schools recommended	Mean earnings of recommended schools	Predicted Earnings	% diff between predicted earnings and mean earnings
25	15.14	154.95	10.96	8587.3	48098.13	46997.06	2.29
43	5.89	84.15	23.55	12693.5	45688.53	48178.42	5.45
43	7.64	128.6	15.49	13886.8	48231.53	46784.22	3.00
128	8.44	110.56	9.74	8691.49	51595.07	51907.80	0.61
58	11.54	63.94	14.81	10414.1	37909.47	39409.39	3.96
43	10.50	128.6	14.83	13886.8	48231.53	46784.22	3.00
81	10.68	121	76.76	15098	56974.27	59659.59	4.71
27	10.13	111.18	37.46	16812.6	43286.73	45760.93	5.72
43	10.23	128.6	15.87	13886.8	48231.53	46784.22	3.00
43	9.01	128.6	45.43	13886.8	48231.53	46784.22	3.00

Annex B: Example list of recommended schools for 1 student

School Name	Average SAT score	%White	% black	% hispanic	% asian	Median family income	Parents from middle school	Parents from highschool	Parents from college	Median earnings
Stevens Institute of Technology	1415	0.5854	0.023	0.1336	0.1846	117032	0.01019	0.14475	0.84506	82237
Fairfield University	1281	0.7766	0.0164	0.0713	0.0266	117152	0.01141	0.15403	0.83456	66244
Lafayette College	1360	0.6669	0.0498	0.0751	0.0434	111341	0.02511	0.12329	0.8516	66921
Villanova University	1406	0.7245	0.0499	0.0913	0.064	110409	0.01187	0.11387	0.87426	77358
University of Delaware	1254	0.6771	0.0626	0.0915	0.0529	107127	0.0086	0.16234	0.82906	57091
The College of New Jersey	1240	0.6207	0.0554	0.1512	0.1164	106390	0.0114	0.19382	0.79477	57264
University of New Hampshire-Main Campus	1197	0.8358	0.0105	0.0357	0.0296	89334	0.00717	0.16731	0.82551	50047
Tufts University	1474	0.5077	0.0454	0.0843	0.1513	85929	0.0197	0.13326	0.84705	54590
York College of Pennsylvania	1096	0.777	0.057	0.0765	0.0206	81845	0.00821	0.29502	0.69677	45864
University of Pittsburgh-Johnstown	1108	0.8058	0.0408	0.054	0.0212	81308	0.00649	0.20318	0.79033	48483
University of Pittsburgh-Greensburg	1085	0.779	0.0651	0.0579	0.0322	81308	0.00649	0.20318	0.79033	48483
University of Rhode Island	1183	0.7351	0.0525	0.1072	0.0338	80297	0.02163	0.24166	0.73672	51827
Elizabethtown College	1176	0.8551	0.0273	0.0522	0.0255	79305	0.01616	0.28341	0.70043	48470
Skidmore College	1324	0.6146	0.0556	0.1014	0.0544	78843	0.02247	0.12584	0.85169	46423
Widener University	1126	0.7135	0.1263	0.0593	0.0338	77624	0.01071	0.28967	0.69962	53312

Annex C: Feedback from User Interviews (Anonymised)

	Wendy	**Stella**	**Bill**
Inputs 1. SAT score (1600-point scale) 2. Race 3. Family income (USD) 4. Parents' education level 5. Desired region of study 6. Desired locale of study 7. Desired degree	1. 1470 2. Asian 3. 105000 4. College 5. Northeast 6. City 7. Bachelor in social sciences	1. 1580 2. Asian 3. 180000 4. College 5. Northeast 6. City 7. Bachelor in psychology	1. 1550 2. Asian 3. 48000 4. High school 5. Southeast 6. City 7. Bachelor in engineering
Output (top 15 colleges, ranked by predicted median earnings)	1. Carnegie Mellon University 2. Lehigh University 3. Rensselaer Polytechnic Institute 4. Boston College 5. College of the Holy Cross 6. Union College 7. Providence College 8. Saint Joseph's University 9. Brown University 10. Fordham University 11. University of Scranton 12. Duquesne University 13. Muhlenberg College 14. University of New Hampshire at Manchester 15. Emmanuel College	16. Carnegie Mellon University 17. Lehigh University 18. Rensselaer Polytechnic Institute 19. Boston College 20. College of the Holy Cross 21. Union College 22. Providence College 23. Saint Joseph's University 24. Fordham University 25. University of Scranton 26. Duquesne University 27. Muhlenberg College 28. University of New Hampshire at Manchester 29. Emmanuel College 30. University of Vermont	1. Auburn University 2. North Carolina State University at Raleigh 3. Eastern Mennonite University 4. University of Kentucky 5. University of Georgia 6. Virginia Polytechnic Institute and State University 7. Georgia Institute of Technology-Main Campus 8. Tulane University of Louisiana 9. University of Arkansas 10. James Madison University 11. Bellarmine University 12. University of South Carolina-Columbia 13. The University of Alabama 14. West Virginia University 15. Duke University
Feedback			
Which college did you actually attend?	National University of Singapore	Yale University	National University of Singapore
What influenced your choice of college?	Location - based in my home country. University rankings - highly ranked university & political science programme on global indices	For me the top consideration was programme, because I already know my post grad plans (scholarship bond). Things like professors, research programmes, breadth of courses, proximity to city (NYC), and brand were important to me.	Mainly finances, if I go over [to the US] I need to pay ~250k on my own.
How did the recommended colleges compare to your choice of college? Did you come across them during your college	Carnegie Mellon, Brown & Boston College (i.e. more internationally renowned colleges) came up in my college search	Did not apply to any on the list, but considered CMU and toured Boston College.	It might be because of the selected region, but of all of the schools provided, only North Carolina, Georgia Tech, and Duke are aligned to my SAT score and in terms of standard of engineering education.

search process?	process --Not as familiar with the other recommendations they don’t typically come up for international students. In comparison to my choice of college -- some had a less global programme.		
Were the recommended colleges relevant?	Somewhat -- perhaps more relevant for candidates who prioritize/ search for universities based on location instead of other factors (e.g. global recognition, university rankings).	Not really, because I did not focus on earnings as a selection criteria (special case due to scholarship).	A few are relevant.
Would having such recommendations have been useful in your college search?	Yes -- for searching for US colleges.	Not really, because I was applying to a few countries [beyond the US].	I will at the very least google and check out the top 3 based on the predicted earnings, then I may consider. So I think it helps as a discovery tool.
Any suggestions for improvement?	I would appreciate a recommendation that is not limited by location. (i.e. recommendations that prioritise the SAT score > global recognition/ranking > earnings > location)	Family income may matter more or less depending on whether the school is need-blind, need-aware, or other financial aid arrangements.	In terms of UI/UX, the flexibility to choose multiple regions of study or weight factors as I’m open to multiple. The list provided is limited because if we go by SAT score to Uni tiering, I should qualify for some tier 1 or tier 1.5 schools, but a lot of the recommended schools are tier 2 or 3.