Introduction

As prospective students navigate their higher education options, understanding the value and satisfaction a college provides becomes important. Using feature selection and machine learning methods on a given data set can determine university quality of life and predict how various institutions rank on the happiness scale. This model shows promising results and reflects the No.1 rule from the Google Developer Machine Learning Guide [1]: “Don’t be afraid to launch a product without machine learning.”

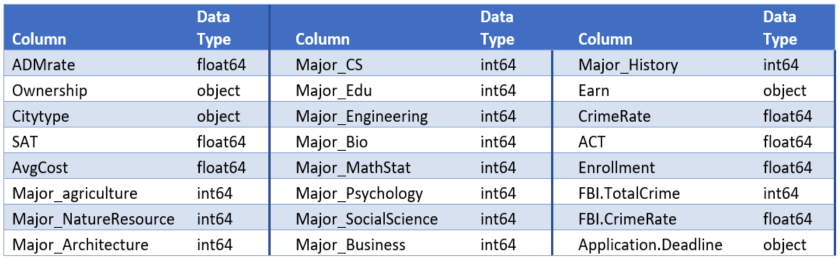
Previous studies [2][3] use data on criminal incidents, enrollment, and tuition to compare how well institutions are preparing students to be successful. Other studies [4] found that student happiness is associated with personal, familial, and social factors. The findings indicate that poor student physical fitness levels, reduced contact with family, and increased use of electronics may be the primary, underlying reasons for the decline in university student happiness in recent years. Statistical and machine learning scientists [5] have turned their attention to providing a self-assessment tool for schools to use in improving overall student satisfaction.

Happiness Model Using Machine Learning

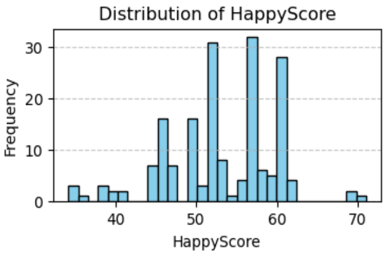
Different from social studies, the challenge of predicting college student happiness is formulated as a statistical and machine learning problem. By leveraging data-driven techniques, the various factors influencing student well-being can be analyzed.

Linear regression is used to predict a variable based on the value of other variables. In this case, the variables provided in the data set, including college enrollments, city type, major, SAT score, and crime rate, are used to predict student happiness. While one of the advantages of linear regression is that it is easy to understand and interpret, it is also sensitive to outliers and noises, and more importantly, it is not suitable to identify non-linear relationships among variables. However, linear regression can be used to find the importance of the features: the magnitude of the coefficients represents the relative importance of the features. Tree based supervised machine learning methods, like XGBoost (Extreme Gradient Boosting) and Random Forest, are also used for their ability to handle complex relationships in data, provide robust predictions, and effectively deal with both classification and regression tasks.

The training data set provided [6] contains 183 records and 25 data attributes with student happiness scores for each corresponding record. Pandas is used to read and load the csv file to a dataframe. The data types for each column are shown below; the y value and HappyScore are also analyzed as below. Even with limited data points, a normal distribution can be observed.



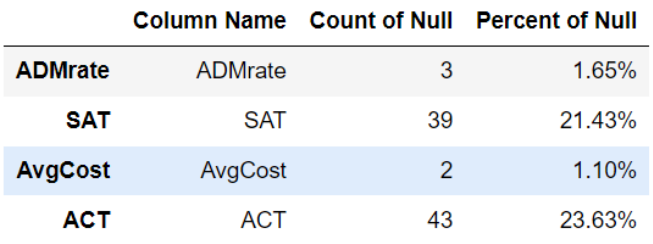
Column	Data Type	Column	Data Type	Column	Data Type
ADMrate	float64	Major_CS	int64	Major_History	int64
Ownership	object	Major_Edu	int64	Earn	object
Citytype	object	Major_Engineering	int64	CrimeRate	float64
SAT	float64	Major_Bio	int64	ACT	float64
AvgCost	float64	Major_MathStat	int64	Enrollment	float64
Major_agriculture	int64	Major_Psychology	int64	FBI.TotalCrime	int64
Major_NatureResource	int64	Major_SocialScience	int64	FBI.CrimeRate	float64
Major_Architecture	int64	Major_Business	int64	Application.Deadline	object



Various data explorations are performed to understand the data.

Step 1: Identifying empty/null value for each column

Several columns have null values, and in some cases, the number of rows with null values is close to ¼ of the total records. The below table shows those columns:



	Column Name	Count of Null	Percent of Null
ADMrate	ADMrate	3	1.65%
SAT	SAT	39	21.43%
AvgCost	AvgCost	2	1.10%
ACT	ACT	43	23.63%

The null values for ADMrate and AvgCost are set to the mean value (after major error is removed from step 2). The values for SAT and ACT are set to 0 for tree based machine learning methods, and set to the mean value for linear methods.

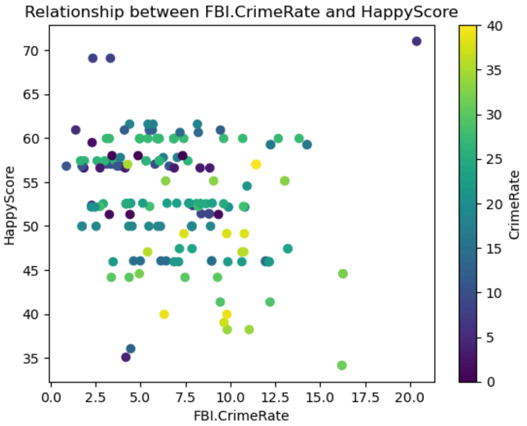
Step 2: Identifying major errors, and outliers; Analyzing categorical data

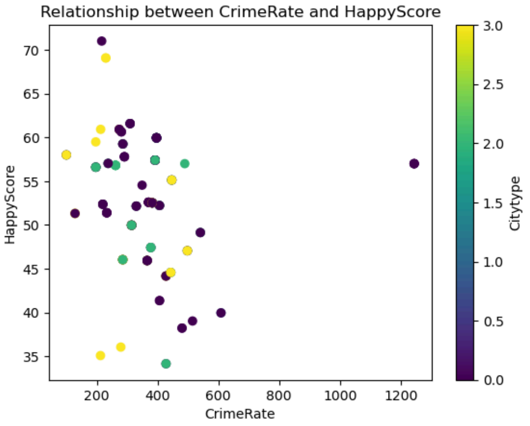
ADMrate has obvious errors in the provided data (greater than 8000). The outlier is set to the mean value of ADMrate. The label “Private for profit” does not exist in the testing data set, therefore all private-related labels are set to “Private”.

Step 3: Analyzing correlation

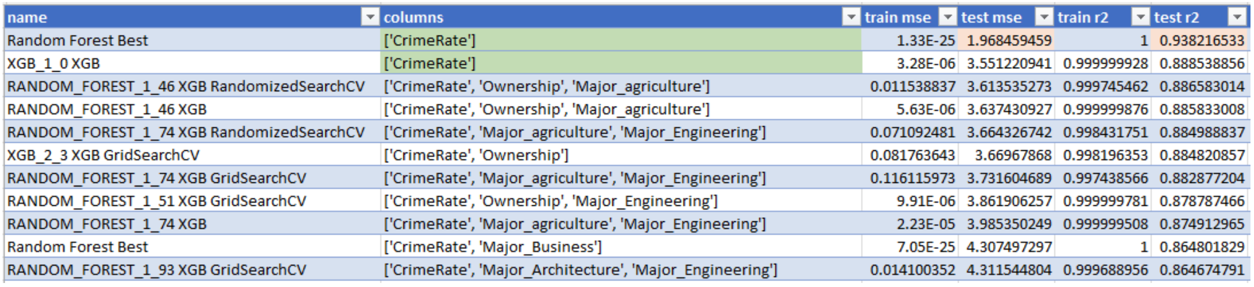
Correlation analysis can reveal meaningful relationships between different metrics or groups of metrics. Removing highly correlated features improves model performance, interpretability, and computational efficiency. SAT and ACT have 98% correlation: students that earn good scores on one test always achieve high scores on the other. Not surprisingly, standard tests are negatively correlated to ADMrate: the college with more advanced students displays a lower admission rate. Additionally, there are three columns related to crime: CrimeRate, FBI.CrimeRate and FBI.TotalCrime. Interestingly, they are not strongly correlated.

The left scatter plot below shows the relationship between FBI.CrimeRate and HappyScores as differentiated by the color of CrimeRate. Although a strong correlation cannot be found in the plot, the same color is always located in the same x-axis line. The right plot shows the relationship between CrimeRate and HappyScore. A strong linear relationship is observed: when the crime rate increases, HappyScore decreases with one outlier. From these observations, it can be determined that CrimeRate heavily influences HappyScore.



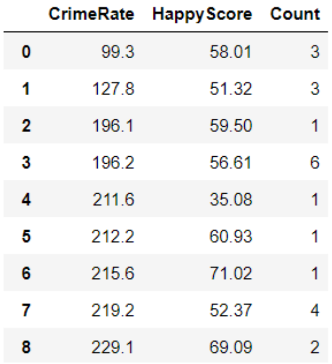


Feature selection focuses on removing non-informative or redundant predictors from the model [7]. As observed, Crime Rate is included in the feature list; the other two crime related features are excluded. Numerous combinations (more than 1000) are tested on three different methods (linear regression, XGBoost, and random forest) with parameter finetunings to include enrollment rate, citytype, SAT, ADMrate, AvgCost, Earn, and all the majors. Different max depths, number of trees and other parameters in XGBoost and random forest are tested. Parameter tuning using cross-validation tends to choose the smaller end of the parameters. GridSearchCV and RandomSearch are used to finetune the parameters. The MSE and R2 for each run are recorded. The MSE or R2 value decreases when more features are introduced in the training data set. The table below shows some of the features selected with parameter turing.

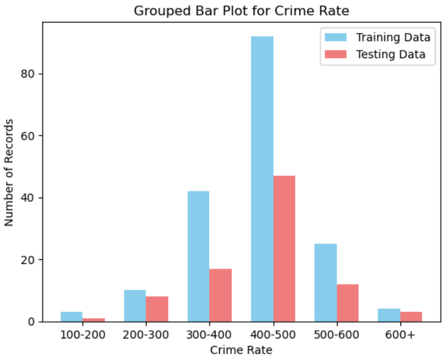


name	columns	train mse	test mse	train r2	test r2
Random Forest Best	['CrimeRate']	1.33E-25	1.968459459	1	0.938216533
XGB_1_0 XGB	['CrimeRate']	3.28E-06	3.551220941	0.999999928	0.888538856
RANDOM_FOREST_1_46 XGB RandomizedSearchCV	['CrimeRate', 'Ownership', 'Major_agriculture']	0.011538837	3.613535273	0.999745462	0.886583014
RANDOM_FOREST_1_46 XGB	['CrimeRate', 'Ownership', 'Major_agriculture']	5.63E-06	3.637430927	0.999999876	0.885833008
RANDOM_FOREST_1_74 XGB RandomizedSearchCV	['CrimeRate', 'Major_agriculture', 'Major_Engineering']	0.071092481	3.664326742	0.998431751	0.884988837
XGB_2_3 XGB GridSearchCV	['CrimeRate', 'Ownership']	0.081763643	3.66967868	0.998196353	0.884820857
RANDOM_FOREST_1_74 XGB GridSearchCV	['CrimeRate', 'Major_agriculture', 'Major_Engineering']	0.116115973	3.731604689	0.997438566	0.882877204
RANDOM_FOREST_1_51 XGB GridSearchCV	['CrimeRate', 'Ownership', 'Major_Engineering']	9.91E-06	3.861906257	0.999999781	0.878787466
RANDOM_FOREST_1_74 XGB	['CrimeRate', 'Major_agriculture', 'Major_Engineering']	2.23E-05	3.985350249	0.999999508	0.874912965
Random Forest Best	['CrimeRate', 'Major_Business']	7.05E-25	4.307497297	1	0.864801829
RANDOM_FOREST_1_93 XGB GridSearchCV	['CrimeRate', 'Major_Architecture', 'Major_Engineering']	0.014100352	4.311544804	0.999688956	0.864674791

CrimeRate is the main contributor to happiness score. Grouping is performed by crime rate and HappyScore in the training set, and the counts are aggregated. Surprisingly, the two attributes follow n-to-1 mapping perfectly (see left table below). This relationship is true for every crime rate. The distinct crime rates from the testing data is collected, and only two crime rates (total of 4 records) do not exist in the training set (195.5 and 399.9). The plot on the right shows that crime rates in training data and testing data follow a similar trend.



	CrimeRate	HappyScore	Count
0	99.3	58.01	3
1	127.8	51.32	3
2	196.1	59.50	1
3	196.2	56.61	6
4	211.6	35.08	1
5	212.2	60.93	1
6	215.6	71.02	1
7	219.2	52.37	4
8	229.1	69.09	2



Based on this observation, mapping is used for these 86 records. For the other 4 records (which are not found in the training set), XGBoost (trained from the training data set with the lowest MSE and highest R2 on the testing data set) is applied.

Conclusion

Statistical and machine learning methods are used to analyze college happiness score. A single contributor to HappyScore, CrimeRate, is identified. A mapping between the two attributes is created, and XGBoost regression is used to predict HappyScore for the values exclusively in the testing data. This information can be valuable for educational institutions aiming to enhance student satisfaction and well-being. The model developed in this study offers a practical tool for predicting happiness scores, providing a data-driven approach to assess and improve the overall quality of life for students.

References

1. Google Developers. (n.d.). Rules of Machine Learning. Retrieved from https://developers.google.com/machine-learning/guides/rules-of-ml (Visited on 1/12/2024).

2. U.S. Department of Education. (n.d.). College Scorecard Data Documentation. Retrieved from https://collegescorecard.ed.gov/data/documentation/ (Visited on 1/12/2024).

3. In Criminal Incidents at Postsecondary Institutions (Ch. 3, Section Campus Crime and Safety).

4. Princeton Review. (n.d.). College Rankings. Retrieved from https://www.princetonreview.com/college-rankings (Visited on 1/12/2024).

5. Sailaja, N. (2023). Happiness Index Prediction of Students Using Machine Learning. In Proceedings of the International Conference on Advances in Computer, Electrical, and Communication Systems (ICACECS 2023).

6. Veritas AI. (n.d.). National High School Data Science Competition. Retrieved from https://www.veritasai.com/national-high-school-data-science-competition (Visited on 1/5/2024).

7. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (p. 488).