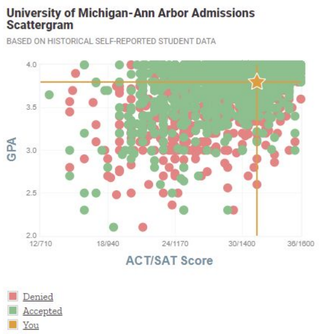
Assess the probability of admission based on various characteristics of the student

Group 26: Zenan Wang
Poster presentation of final project for DSAN 5300

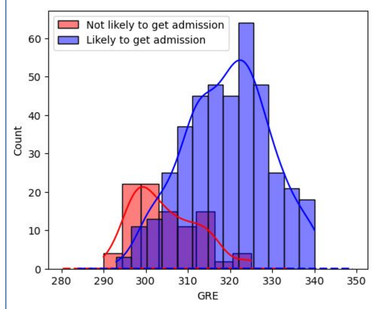
Abstract

In today's competitive academic landscape, gaining admission to esteemed educational institutions is a significant milestone for students worldwide. With the increasing number of applicants and limited available slots, the admissions process has become more nuanced, often relying on various factors to assess a candidate's potential.

This project delves into the realm of predictive modeling to gauge students' chances of admission to schools based on a range of essential features, such as CGPA, TOEFL scores, and GRE scores. By leveraging advanced statistical techniques and machine learning algorithms, this study aims to develop a reliable predictive model that can assist both students and educational institutions in making informed decisions during the admissions process.

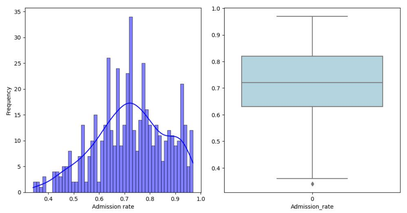


Exploratory data analysis



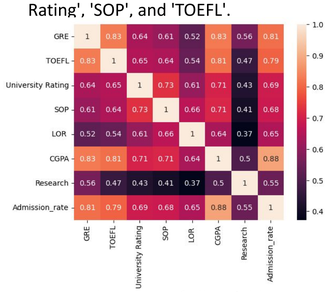
Initially, I create visual representations for the distribution of individual student features and then partition the data into two groups: admitted and denied students. On the left side, you can see the distribution plot for GRE scores.

Next, I visualize the distribution of the target variable, 'admission chance', find no outliers, and see the data to be normally distributed.



Distribution of target variable

I also fit a heatmap to find variables that are strongly correlated with the target variable, which are 'CGPA','GRE', 'University Rating', 'SOP', and 'TOEFL'.



heatmap of all variables

Methods and Materials

Variable Name	Type	Description
GRE	Numerical	GRE Scores (out of 340)
TOEFL	Numerical	TOEFL Scores (out of 120)
University ranking	Numerical	University Rating (out of 5)
SOP	Numerical	Statement of Purpose Strength (out of 5)
LOR	Numerical	Letter of Recommendation Strength (out of 5)
CGPA	Numerical	Undergraduate GPA (out of 10)
Research	Binary	Research Experience (either 0 or 1)
Chance of admit	Numerical	Chance of Admit (ranging from 0 to 1)
Chance of admit (binary)	Binary	Set chance of admit to be 1 if the value >= 0.7, 0 otherwise

Data description

Models and criterion I have used

I began by fitting a linear model using several parameters that I believe are strongly correlated with the target variable. Next, I employed random forest and decision tree models, creating feature importance graphs and plotting predicted values against actual values. Moving on, I combined both classifier and regressor models. Initially, I trained various classifiers—logistic regression, LDA, QDA, and SVM—on a modified binary target variable, using ROC curves and confusion matrices to visualize their performance. Then, I utilized regressors—ridge regression, gradient boosting, and SVR—on the original continuous target variable, generating scatterplots of predictor values against actual values and residual plots for each model. At last, I put into action an ANN model, trained it on the testing set, and recorded the final testing loss. Across all models, I utilized MSE as the loss criterion, and for certain ones, I employed the R-squared value to evaluate their predictive performance.

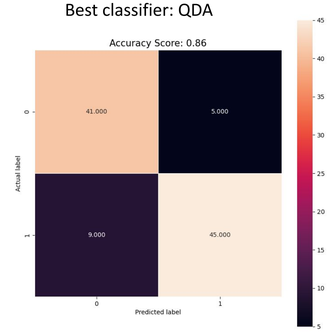
Results

Model name	Model type: regressor or classifier	R-square value (If N/A, AUC value will be used)	Mean squared error
Linear regression	regressor	0.772	0.0043
Decision tree	regressor	0.679	0.0066
Random forest	regressor	0.792	0.0042
ANN	regressor	N/A	0.0037
Logistic	classifier	N/A, AUC = 0.92	0.15
LDA	classifier	N/A, AUC = 0.95	0.15
QDA	classifier	N/A, AUC = 0.96	0.14
SVM	classifier	N/A, AUC = 0.95	0.15
Ridge	regressor	0.818	0.0037
Gradient boosting	regressor	0.782	0.0045
SVR	regressor	0.649	0.0072

Best classifier:

I'll extract the top-performing model for both classification and regression, and present the results.

Best classifier: QDA



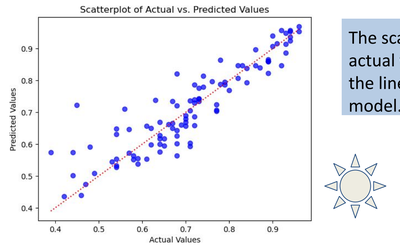
Confusion matrix

	precision	recall	f1-score	support
0	0.82	0.89	0.85	46
1	0.90	0.83	0.87	54
accuracy			0.86	100
macro avg	0.86	0.86	0.86	100
weighted avg	0.86	0.86	0.86	100

Classification report

Results continued:

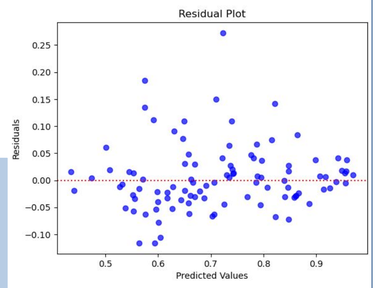
Best regressor: Ridge
(Although Ridge and ANN share the same MSE value, ridge has a smaller computation burden)



The scatterplot depicting predicted values against actual values shows points that closely align with the line of slope 1, suggesting a strong fit of the model.

Scatter Plot of predicted vs. actual values

A residual plot can indicate overfitting when it shows a pattern of non-random residuals. In an overfit model, the residuals may display a systematic pattern, such as a curve or a trend, instead of being randomly scattered around the horizontal axis. In contrast, this model exhibits random and evenly distributed residuals in the plot.



Residual plot

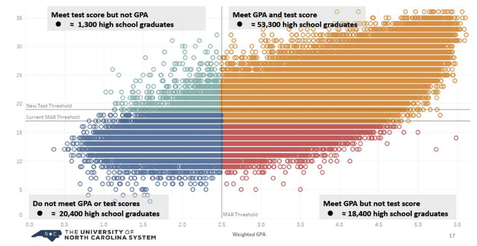
Conclusions and future improvements

Conclusions:

The optimal model for classification is QDA, while for regression, it's ridge regression. However, I lean towards using a regression model over a classifier because the 0.7 threshold used for classification could be arbitrary. A regressor offers higher accuracy and can more effectively portray a student's probability of admission compared to a classifier.

Future improvements:

Since the data's scale is relatively small, it might not accurately represent the nationwide trend. Additionally, the university rating scale of 1 to 5 may not sufficiently capture the school's overall reputation. In my view, the model would benefit from the inclusion of more predictors, such as the difficulty level of electives and the alignment between the student's high school education and their intended major.



Contact

Author name: Zenan Wang
Email address: zw417@georgetown.edu
phone: 984-261-5597

References

Data platform:
1. https://www.kaggle.com/datasets/mohansacharya/graduate-admissions/data

Figure sources:
1. https://www.thecollegefundingcoach.org/how-to-check-your-admissions-probability-for-college/
2. https://www.theglobeandmail.com/featured-reports/article-admissions-advice-tips-to-get-your-application-noticed/
3. https://journalnow.com/townnews/school/the-syllabus-the-unc-system-will-consider-a-new-minimum-freshman-admissions-requirement/article_52de7d69-8076-57e0-bf9d-b2e0817a9c50.html