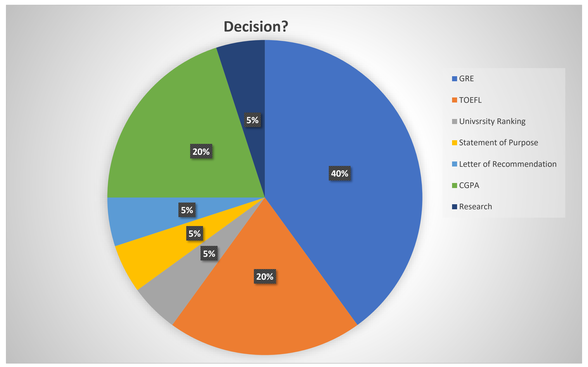
FORECASTING GRADUATE SCHOOL ADMISSION

GUIMIAO ZHANG
Department of Statistics

WASHINGTON STATE UNIVERSITY

BACKGROUND

- Have you ever considered graduate school?
- Have you ever curious about how does the admission committee decide on the applications?
- Which one or ones of the factors play the critical role on the decision?
- With the given information, can you predict the chance of admit?
- Do you think the chart below make sense?



KEY STEPS

- Graduate admission data from Kaggle
  - 7 parameters, considered important during the application for Masters Programs at University of California Los Angeles
    - GRE
    - TOEFL
    - University Rating
    - Statement of Purpose
    - Letter of Recommendation
    - College GPA
    - Research Experience
  - Chance of Admit – prediction by the applicant
- Explanatory search on the chance of admit and other predictors, normalized the data due to the high diversity of magnitudes, 1/3 of data will serve as test
- Linear regression
  - Regress the chance of admit by other predictors
  - Evaluate the model
- Logistic regression
  - Determine the threshold to set up admission status using the explanatory result
  - Classify the admission status by other predictors
  - Evaluate the model
- K-nearest neighbors
  - K-nearest neighbors regression
    - Evaluate the model
    - Compare the performance with linear regression
  - K-nearest neighbors classification
    - Evaluate the model
    - Compare the performance with logistic regression

EXPLANATORY RESULT

- 500 international applications were collected with no missing values.
- Self predicted chances of admit differ from 0.34 to 0.97 with median of 0.72. More than 75% self predictions are over 0.50 chance of admit and thus seems too optimistic.
- As only top students will be admitted to Grad School at UCLA, a different threshold was adopted other than self predictions.
- From the correlation heatmap, 3 factors are strongly correlated with admission chance, GRE, TOEFL & College GPA. Thus, only applicants with all three criteria at least 75% top will be considered as admitted students and yielding 87 out of 500 admissions.
- 17.4% of admission chance seems normal according to the recent statistics from top US universities.

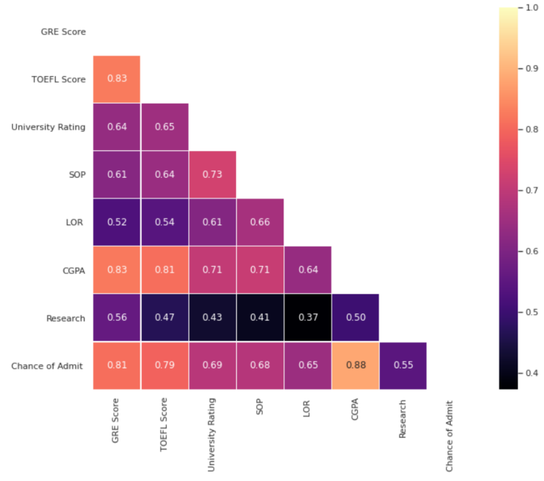


Figure 1 Correlation heat map

.

Table 1 General statistics summary of data

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

LINEAR MODEL / PREDICTION

- A linear model was established using all features to predict the chance of admit.
- The assumptions of linear model were roughly satisfied. However, only ~70% ($R^2 = 0.712, R^2_{adj} = 0.700$) variation of the data can be explained by the model.
- The performance is OK but not ideal. A perfect fitting model, all points in the last plot of figure 2 should line up on 45-degree line.

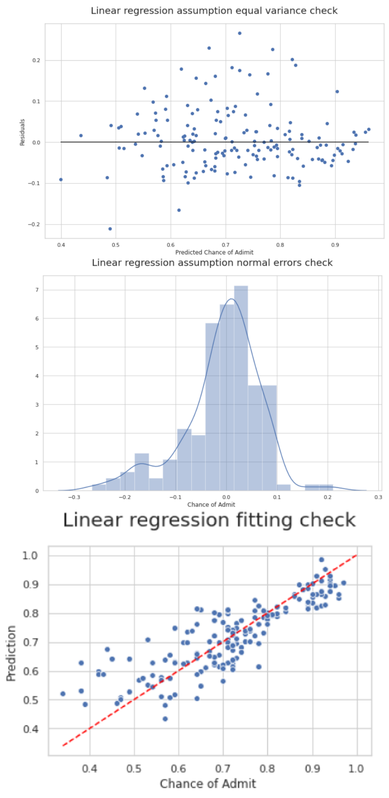


Figure 2 Primary result of Linear Regression Model

LOGISTIC MODEL / CLASSIFICATION

- 17.4% of admission chance seems normal according to the recent statistics from top US universities.
- Best parameters by grid search: C = 0.1, penalty = l2, no weight
- about 81.8% accuracy on test data

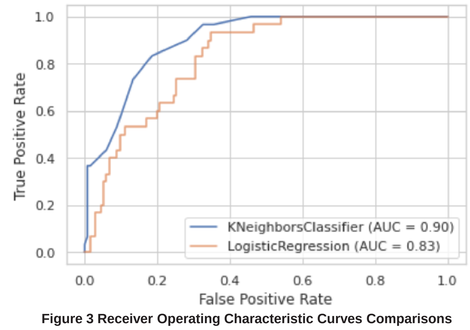


Figure 3 Receiver Operating Characteristic Curves Comparisons

KNN MODELS

- As linear regression provides numerical prediction on chance of admit and logistic regression provides classification of admission status, it is difficult to directly compare the performance of these two algorithms.
- We may involve other algorithms to make a comparison, KNN seems a good candidate as we can employ both KNN regression and KNN classification here.
- KNN regressor
  - Best parameters by grid search: n_neighbors = 4, weights = distance
  - ~ 61.7% of the data variation can be explained (R^2_{adj})
  - Mean squared error (MSE) is ~0.0076
- KNN classifier
  - Best parameters by grid search: n_neighbors = 19, weights = uniform
  - ~ about 86.1% accuracy on test data

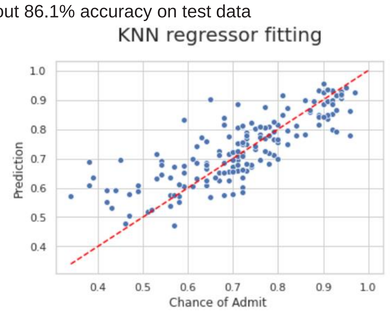


Figure 4 Primary result of KNN regressor

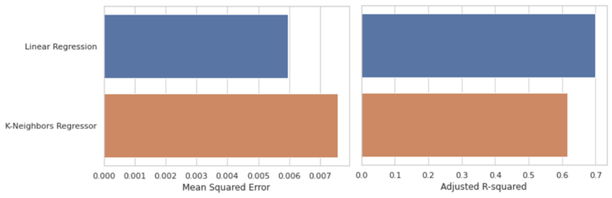


Figure 5 Linear regression and KNN regressor performance comparisons (up right)

SUMMARY

- Linear regression predicts the chance of admit with a smaller MSE and a larger adjusted R-squared than KNN regressor.
- Logistic regression performs worse than KNN classifier, if the accuracy and ROC curve are chosen as the evaluation criteria.
- Only 4 algorithms were applied here to forecasting the graduate school admission, if possible, more should be implemented to choose the “best” one.
- The threshold chosen to determine the admission status is super critical. If a different threshold is chosen, the classifiers performance may differ significantly.
- This prediction can give us only a general idea on the admission process. It can not be generalized too much as different graduate schools have different criteria to choose candidates.
- Now the question is: Do you consider graduate school ?

Graduate School
WASHINGTON STATE UNIVERSITY