ORF 350 Homework 7 - Adam Chang and Chris Evanko

1. Introduction and Problem Statement

We are trying to understand admissions into graduate programs. To address this we seek to answer a few sub-questions - notably, what factors/parameters are most important in determining an applicants chance of admission? And how does the decision making process work from the school's perspective? To answer this problem, we will use a dataset about Indian graduate school applicants, which includes information about test scores, writing strength, undergraduate experience, etc, and chance of admission (more detailed description below). This data will help us to explore our questions about higher education admissions in a very specific context (Indian students applying to graduate school). Because the data is specific in scope and limited in size, we expect our results to be non-generalizable to the broader admissions questions, though we do expect to gain insights into how to better understand the admissions process by employing and comparing a variety of analysis methods, including linear regression, clustering, and decision trees.

2. Dataset and Methods

The dataset we are dealing with is called Graduate Admissions, and was uploaded to Kaggle.com by Mohan S. Acharya. The dataset contains 500 data entries, each of which describe information about a particular undergraduate university student from India applying to graduate school. There are 7 parameters for each entry in this dataset which are GRE Scores (out of 340), TOEFL Scores (out of 120), Undergraduate University Rating (out of 5, 5 being the highest/best), Statement of Purpose Strength (SOP) and Letter of Recommendation Strength (LOR) (both out of 5), Undergraduate GPA (CGPA) (out of 10), and Research Experience (either 0 or 1). The dependent/response variable is chance of admission, given as a probability.

In order to address the first question (factors determining admission), we first give scatter plots of admit probability over each feature, giving us a general impression of the data and helping us form initial hypotheses. We employ several linear regression models to address this question, including OLS, Ridge Regression with 10 fold CV, and Lasso Regression with 10 fold CV. We also use PCA to examine variation in the students over admissions factors, and attempt to combine regression models with lower-dimensional data (projected using PCA).

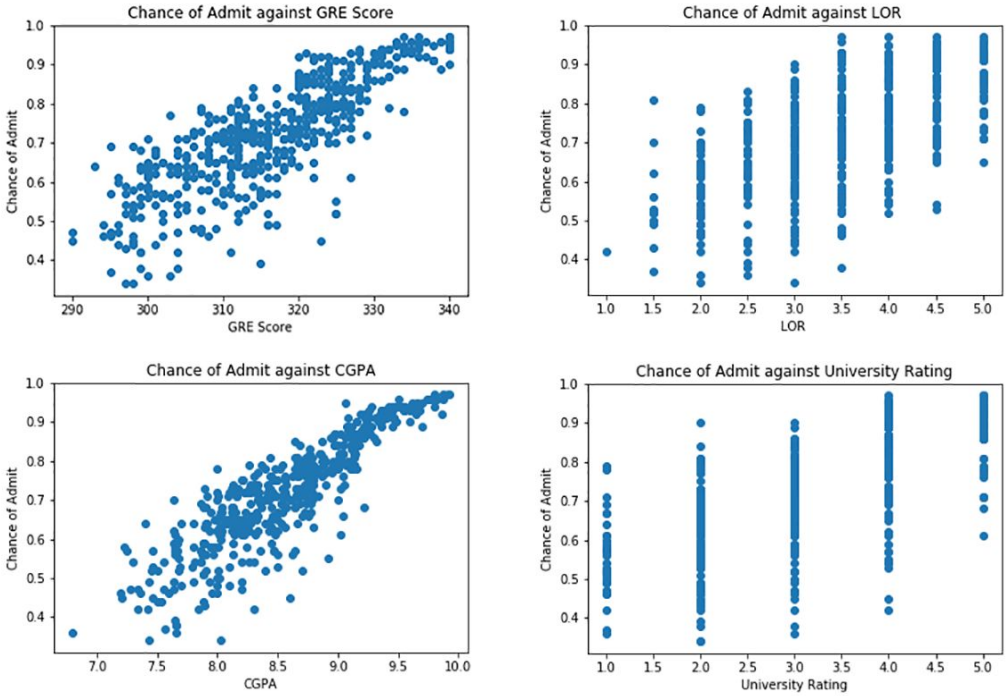
Next, we use a series of regression decision tree methods a la HW 4 to address the second question (how decision making may work). We construct these decision trees restricting on maximum depth, maximum number of leaf nodes, and minimum samples per leaf. We also prune our minimum sample per leaf tree using 5 fold cross-validation using the same cost complexity pruning routine from HW 4.

To better elucidate both of these questions, we then explore the dataset using K-Means clustering. We compare intra-cluster variance in admissions probability, and move on to predicting admissions outcomes using K-Means clusters.

In the final part of our analysis, we turn to classification over our dataset, splitting students by if their admissions probability was over the median. We used linear classification methods such as linear SVM with hinge loss and regular linear logistic regression (including over PCA), as well as K-Means clusters.

3. Analysis and Results

With Occam's razor in mind, we began fitting various regression models to the data, since in the context of the features, we expect some kind of linear dependence (i.e., higher exam scores/higher GPAs should strictly lead to higher acceptance probability). In order to get a rough idea of the relationship between each parameter and the chance of admission, we will give scatterplots of each variable against chance of admission, a la the procedure in HW 2.



Our regression coefficients for a basic **linear regression** were as follows, with the regression over scaled features (std. Normal scaling) in yellow:

GRE	TOEFL	Univ. Rat.	SOP	LOR	CGPA	Research	Intercept
0.0018	0.0025	0.0088	0.0008	0.0180	0.1167	0.0219	-1.219
0.0197	0.0145	0.0100	0.0008	0.0164	0.0690	0.0108	.7271

In this most basic regression, we obtain a test MSE of 0.0044 - not bad! We then perform **ridge** regression and **Lasso**, using a 10-fold CV grid search for the regularization parameter. We obtain an alpha of 6.46 for ridge and ~0 for Lasso, with both yielding MSE of around 0.0052. The coefficients below are over std. Normal scaled features.

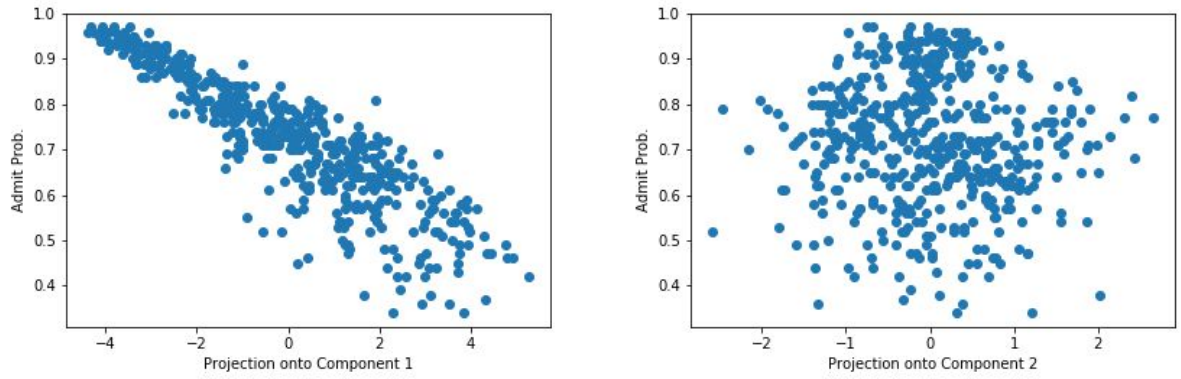
	GRE	TOEFL	Univ. Rat.	SOP	LOR	CGPA	Research	Intercept
Ridge	0.021	0.015	0.010	0.002	0.017	0.065	0.011	0.727
Lasso	0.020	0.015	0.010	0.001	0.016	0.069	0.011	0.727

ORF 350 Homework 7 - Adam Chang and Chris Evanko

Across these various regression methods, our coefficients end up being quite similar.

We next use **PCA** to examine variance within the sample data. We found that the first component explained most of the variance, while the remaining six explained little of it (0.15-0.56 explained variance). Coefficients for the first two components are shown below, all features being standard normal scaled.

Comp.	GRE	TOEFL	Univ. Rat.	SOP	LOR	CGPA	Res.	Exp. Variance
1	-0.404	-0.401	-0.383	-0.385	-0.347	-0.421	-0.289	4.74
2	-0.275	-0.111	0.250	0.343	0.426	-0.015	-0.742	0.74



Graphing admit probability over projections onto these components (one at a time), only the projection onto the first component suggested predictive ability (a modest linear relationship). We tested linear regression over these components (using 1-7 components), and found that only the first component had significant predictive power. Specifically, R^2 = 0.78 (and only increased slightly to 0.80 with seven components), and test MSE = 0.0062 (decreasing to 0.0052 with seven components). Below we give the combined PCA/LR result (regressing over projection onto first component) , and compare it to linear regression.

	GRE	TOEFL	Univ. Rat.	SOP	LOR	CGPA	Research (0/1)	Intercept
Lin Reg.	0.0197	0.0145	0.0100	0.0008	0.0164	0.0690	0.0108	.7271
PCA/LR	0.022	0.022	0.022	0.022	0.019	0.024	0.015	.727

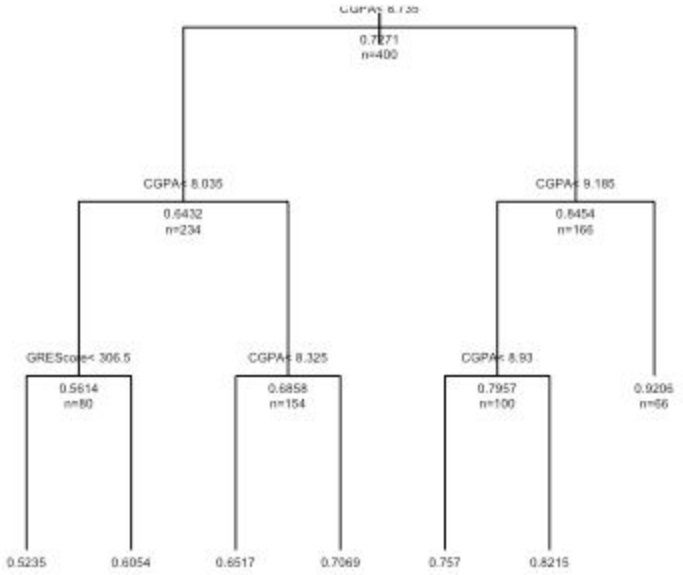
We move on to regression **decision trees**. Fitting trees with maximum depth of 2 and 3, we found that trees were almost entirely based off of CGPA, setting an initial threshold and then various sub-thresholds. At depth 3, the tree begins to depend on GRE, but only at a sub-level (low CGPA). Next, regularizing by number of leaf nodes (= 6), we find that CGPA (and partially

GRE) continues to dominate the decision tree. All trees achieve test MSE of around 0.006, higher than in regression, but there are clear benefits in offering an alternative interpretability.

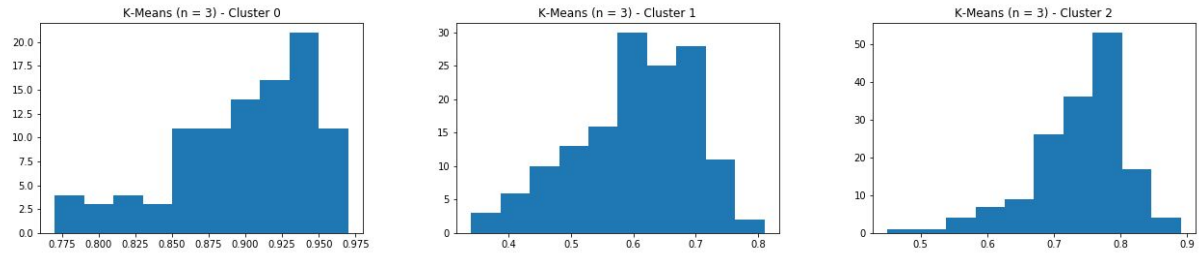
A sample decision tree (for 6 leaf nodes max.) is given below:
Root: If CGPA <= 8.73499965668 go to node 1 , else go to node 2
Node 1 : If CGPA <= 8.03499984741 go to node 3 , else go to node 4
Node 2 : If CGPA <= 9.18499946594 go to node 5 , else go to node 6
Node 3 : If GRE Score <= 306.5 go to node 7 , else go to node 8
Node 4 : If CGPA <= 8.32499980927 go to node 9 , else go to node 10
Node 5 : Predict Chance of Admit = 0.7957
Node 6 : Predict Chance of Admit = 0.920606060606
Node 7 : Predict Chance of Admit = 0.523488372093
Node 8 : Predict Chance of Admit = 0.605405405405
Node 9 : Predict Chance of Admit = 0.651694915254
Node 10 : Predict Chance of Admit = 0.706947368421

Inspired by HW 4 (aren't we all), we tried pruning a tree, first employing 5-fold CV to determine the proper tuning penalty. We obtain the tree shown below, with a test MSE of still around 0.006, and a very similar structure to previous trees, with CGPA still dominating.

Pruned Regression Tree for Chance of Admit



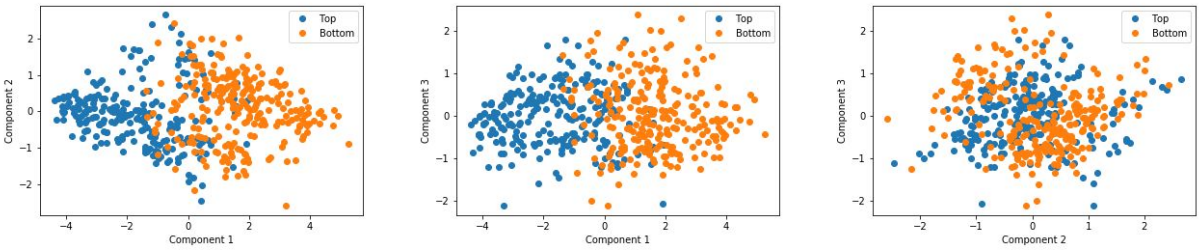
Next, we use **K-Means clustering** to predict admissions probabilities (using the average of in-cluster samples). Before we begin, we first check what clusters look like, by plotting histograms of admissions probabilities in each cluster, with clustering only based on the seven factors above (and not admit probability). We show the result for three clusters on the following page; we note that there is **huge** variance within clusters on admissions probability, suggesting that clustering methods may not be of much use.



(The x-scale goes from 0.775-0.975 in the first cluster, but 0.3-0.8 in the other two clusters.)

Still, as ever diligent students of big data, we try our best. We determine an optimal number of clusters by using 5-fold CV on the training set, and settle on using 4-6 clusters. Using 6 clusters, we achieve the worst MSE so far, of around 0.0076.

We move on to various classification methods. First, we try to classify students as either “top” or “bottom” (vs. median) using logistic regression on PCA-projected lower dimension data. Below, we present these two groups in various component spaces 1 and 2, 1 and 3, and 2 and 3).



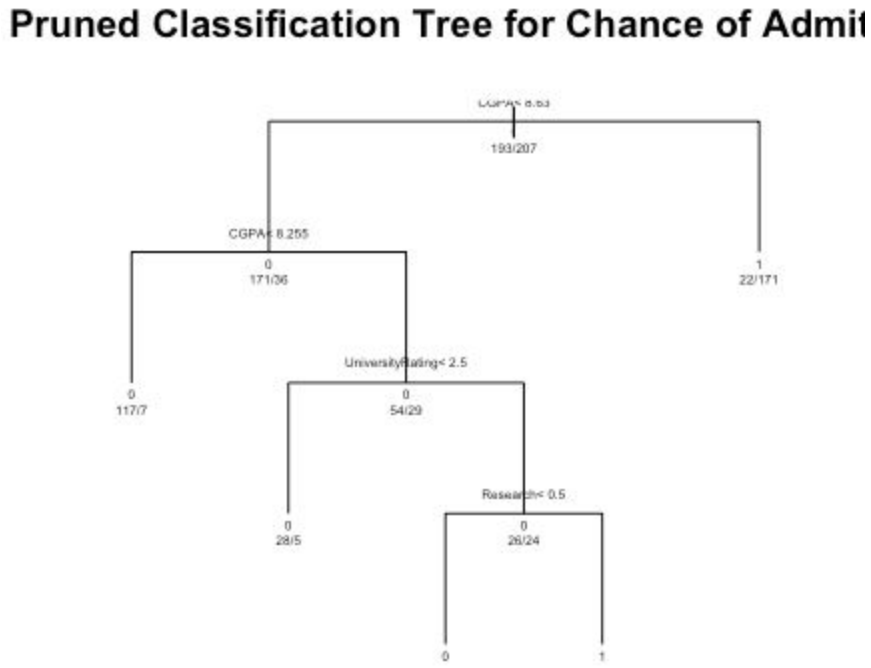
From these graphs, it’s clear that component 1, again, seems to have major predictive power in classifying students’ admit probabilities. Performing a **logistic regression**, we obtain that using only component 1 results in test accuracy (0 vs. 1 of being in the top half) of 0.89, around the same test accuracy as using just a logistic regression. Adding more components barely increased test accuracy.

We next try K-Means clustering with classification, simply classifying students based on the mode of other students in the same cluster. We obtain a test accuracy of 0.9, similar to logistic regression and PCA + logistic regression, suggesting that K-Means is much better at this basic classification than at a complicated regression problem. Using a simple hinge-loss SVM classifier, we also obtain similar test accuracy of 0.9.

Moving on to a decision tree, we obtain some novel results compared to the regression decision tree. Treated as a classification problem, the decision tree yields a much more encompassing set of features that it considers.

Our classification decision tree (pruned by the same method described earlier) is shown on the next page, but while CGPA does dominate initially, other factors do come into play at deeper levels. Specifically, for students with a low, but not extremely low GPA, having a high university

rating (> 2.5) and research experience (> 0.5, or basically “yes”) results in a positive admissions decision, or at least probability above the median. We obtain a similar test MSE as before.



4. Conclusions

Our conclusions to our analysis was fairly straightforward. As we found in our initial regression methods and then confirmed with our decision tree analysis, we found CGPA to be by far the most important factor in determining probability of admission. This is unsurprising as we would expect GPA to be extremely important, especially in comparison to the other features we were considering. Additionally, we found that the decision process was similarly dominated by GPA. We were able to see that other features made an impact in our regression trees when complexity increased (i.e. GRE score and TOEFL score). Using PCA, we were able to see that students truly varied among one axis, that was a linear combination (in rather equal parts, after features were scaled to standard normal) of the different admissions factors.

When we treated the problem as a classification problem, we found that our decision trees included several other factors such as LOR, University Rating, and Research). That university rating and research experience could alter the classification of a certain group of students (those with low, but not extremely low, GPA) informs us of how different factors may play a role in admissions; in other words, attending a good school and having research experience can compensate for having a low GPA, which makes sense to us.

Overall, we were able to get fairly satisfactory answers to our initial questions. That being said, we recognize that this analysis was a very exploratory investigation into a much broader problem of understanding higher education admissions processes. If we were going to continue research into this problem, we would focus on expanding our data in two ways. First, we would increase the feature space significantly, for example trying to acquire data on things such as extracurricular activities (perhaps by number or kind), work experience, application essays, etc etc. Ultimately, we could try to look at something like the Common App (used for undergraduate admissions) and attempt to compile data for all the features included. This would allow for a more complex and interesting analysis, testing validity of admissions process claims such as “holistic admissions processes.” The other main way to expand our research is by expanding the scope of the data itself. With the dataset used here, we were limited to Indian students applying to graduate school. If we had data on students applying to undergraduate universities across the US or world, we could attempt to make broader conclusions about the admissions process on the whole, which we simply could not do with the data we used here.

5. Appendix

Dataset: https://www.kaggle.com/mohansacharya/graduate-admissions