Branden Williams

Machine Learning

Dr. Jorge Colazo

May 8, 2020

Admissions Project Report

**Background**

The admissions process at most Universities occurs in stages. Colleges will seek out prospective students for their institution and begin to collect data on these students. I personally work in the records and data department at Trinity as a student worker so I am responsible for entering data we collect about prospective students such as their academic interests, address, email, phone numbers, and academic factors such as their GPA. When students are seniors in high school they begin applying to colleges and typically institutions will judge their applications on their GPA, SAT/ACT scores, extracurriculars, sports, and fine arts. Students in Texas typically apply to schools using either the common app or Apply Texas.

There are three types of applications students can file under, early action, early decision, and regular decision. Early action simply means that the student is applying earlier than the regular decision deadline and typically indicates more interest in the school than students who submit regular decision. The most binding of these three is early decision which requires that if the student is accepted, they must enroll in that school. In the admissions process early decision is very useful because those students are almost guaranteed to accept their offer of admission. Schools often admit more students than they expect to accept the offer to ensure that they reach class size targets, as many students who are admitted end up not accepting their admission. Lastly, regular decision is the baseline choice that has no special meaning, it is just a normal application with no strings attached.

Many schools use an academic index to screen out applicants that they do not believe will be academically able to thrive at their institution. I did some research into this and came across what is known as the Ivy Academic Index that many Ivy League Schools use to determine which applicants’ applications should be looked further into and which ones they are not interested in. I decided to add this to my dataset but changed the formula around a little as the admissions data did not have SAT subject tests scores for the admitted students. I had the theory that many of the students who did not accept their admission would be those who scored higher on this index as they probably chose to go to an even better academic institution.

**Data Cleaning**

Using what I learned from the Kaggle assignment the cleaning process this time around was far easier to navigate, particularly using stringsAsFactors = T saved me a lot of trouble. I deleted NA columns and redundant variables to narrow down the data I was working with, for example I deleted the ACT sub scores columns as they averaged up to the total ACT composite column. After this I went ahead and changed the data types of a few of the columns to match up with what I thought they should be and filled in all the NA values in the dataset.

I decided that I wanted to be able to use the data from the columns containing dates and times such as InquiryDate so I set out to use regex expressions to clean the data into a usable form. After messing around with it for a while I was able to change the first two date columns to years and the Submitted column to month as that would give me information I could use in regards to how early the student submitted their application (the theory here was students who were more interested in coming to trinity most likely would have applied earlier than those who were not). Probably the biggest headache I encountered when cleaning the data was the mess that was the Legacy and Athlete columns. Admissions decided to throw unrelated information into the same column and then sometimes duplicate the exact same info in the next column. Thankfully I was able to make the data usable by combining the Legacy and Athlete columns using the unite function from tidyr and then used regex to clean up the data and separate the new data into 5 usable columns. I also realized at this point I should remove the sport.1.sport column as my new athlete column effectively gives me the same information as I did not think the exact sport a person played would have a large impact on their acceptance of admission.

The last bit of cleaning I did involved implementing the Ivy Academic Index and making it work with the data I had. Since I did not have any info on SAT subject tests in the data, I used only SAT/ACT scores and GPA and weighted them as being worth 120 each. In cases where a student took both the SAT and the ACT, I took the higher of the two academic indexes for that student. For the clustering methods and neural network I also used one hot encoding on the data.

**Clustering**

After the data was clean I one hot encoded it and ran it through a k means model. I determined using the elbow method that the optimal number of clusters was 3 or 4 and decided to use 4 for my model. The following groups were observed:

* **Group 1**: The students in this group came from mid-sized schools and were average students in the GPA spectrum. 18% of these students were athletes which was the highest percentage of athletes in the 4 groups. The majority of these students were White or Asian and also tended to be slightly more female. 84% of these students applied early action 1 or 2 and only 39% of them visited campus. These students were also overwhelmingly from Texas.
* **Group 2**: The students in this group came from large schools and were generally the least academically inclined of the 4 groups. This group contained the highest percentage of legacy students and this shows legacy students tend to be less qualified for Trinity than their non-legacy peers. This group was overwhelming white and female with 71% of the students being white and 62% of the students being female. The students in this group also seem to have been wanting to pursue careers in STEM fields with their top 3 academic interests being Pre-Med, Engineering Science, and Biology.
* **Group 3:** The students in this group came from small schools with the average class size being 162. They had the highest average GPA and Ivy Academic Index of the 4 groups with 3.81 and 216.95 respectively. The students in this group were 77% white with the 26% of these students identifying as Hispanic/Latino, this group also had the least amount of Asian representation with only 10%. On trinity’s academic index scale 67% of these students were either classified as a 1 or a 2. A staggering 84% of these students declined admission, this leads me to believe that this is the group of students who went to higher institutions.
* **Group 4:** Group 4 can best be characterized by saying it is the average group. The students in this group came from mid-sized schools and they had average GPAs. They tended not to be athletes with only 13% of the students playing a sport. Furthermore, this is the most even gender split of the groups with 52% being female and 48% being male. The most common academic index in this group was the middle of the road 3.

After running my k means clustering algorithm, I moved on to doing hierarchical clustering using an average linking method. I settled on doing two groups here because 3 and 4 were having some errors. The following groups were observed:

* **Group 1:** The students in this group came from mid-sized schools. They consisted of 53% female and 47% male. This group had more non-Texas students than group 2 as over 10% more of these students used the common app rather than Apply Texas. They consisted of mostly white students with 1/5 being Hispanic/Latino. The most common academic index was 3.
* **Group 2:** The students in this group came from large schools. 64% of these students were female and 80% of these students first source summary was from CBINQ. 15% of these students were athletes.

**Admission Acceptance Models**

The first model I applied to predict which students would and would not accept their offer of admission was Logistic Regression. Unsurprisingly, keeping my model simple with logistic regression provided the best kappa value I attained which was **0.5485215.** Effectively this model was a decent predictor but nothing too special, like all of the following models it is much better at predicting who won’t accept admission rather than who will.

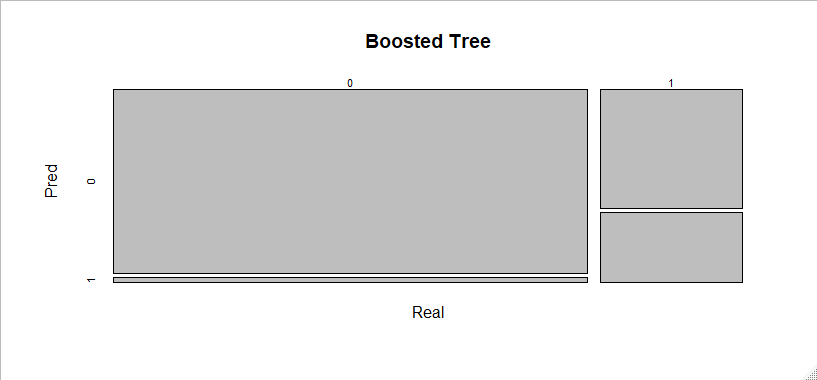


For my random forests I used 4 different models. All four of the models performed worse as far as kappa is concerned in comparison to the logistic regression model.

**Outcomes**

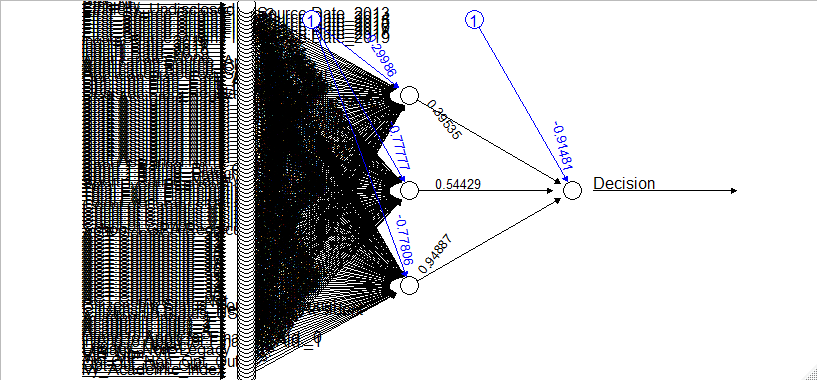
* Bagging w/ All variables - 0.4859315
* Bagging w/ Random variables - 0.5214963
* Bagging w/ Importance - 0.4936309
* Boosting - 0.4936309

While none of these models outperformed the logistic regression model in relation to kappa. I did find something interesting with my boosted random forest model. The model correctly predicted 97% of true negative, meaning of course that the model can say with high certainty who WON’T accept the offer of admission. Knowing this, someone using this model could say with decent accuracy who would accept the offer by looking at the students the model did not predict as declining admission.



Lastly, I used a neural network with 3 neurons. I attempted to use larger numbers of neurons, but the result was actually worse than just keeping it simple. That being said, with a kappa of **.4927227** the model still did not outperform basic logistic regression.

Neural Network



**Conclusion/Personal Reflection**

In conclusion these models served as mediocre predictors of students who would **Accept** the offer of admission. However, the boosted random forest model was extremely accurate at predicting who would **Decline** which is incredibly useful as it would allow admissions to have a pretty good idea of the class size they can expect. Personally, I am fairly satisfied with my work on this project. I spent way more time cleaning the data and doing interesting things with it like using external data such as the Ivy Academic Index and using regex expressions to help clean. I found the neural network to be the most challenging model to get to work and that caused a big headache. Hopefully now that I have done it once I can perform better with it the next time I employ it. Lastly, thank you for all that you have taught me this year! I thoroughly enjoyed this class and I am super happy that I know so much about predictive analytics as its always been something I really loved.