# *Application Data's Relation to Student Success*

Abdullah Nahiyan, Daíre O'Gorman, Stephanie Stewart

## Abstract

We analyzed and visualized the profile of admitted students at NC State University (NCSU) over the time period of Fall-2009 to Fall-2012. In this time period, we summarized and visualized the GPA, SAT and ACT scores of admitted students in several groups. We found statistically significant differences in scores between groups. By using exploratory factor analysis (EFA), we also discovered that two underlying factors- high school factor and test scores- play a major role in admission decision at NC State University (NCSU). We also simplified and visualized this model through confirmatory factor analysis (CFA). We then attempt to use classification in order to predict individual student success and overall graduation rates however are hampered primarily by technical limitations at hand. Some models do provide reasonable levels of accuracy but ultimately none reach a satisfactory level. To say that, we found that each of the model fit - Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbor (KNN), and Regression Tree (RPART)- has certain advantages and disadvantages to predict a student's success in graduation based on our large dataset.

Key words: Education, Admissions, SAT, ACT, PCA, EFA, CFA, Classification.

# 1. Introduction

College undergraduate admission is the gateway of opportunity into institutional higher education. The benefit of higher education makes admission very competitive and is dependent on various qualitative and quantitative criteria of an applicant. From the limited snapshot provided in an application one must decide which students to admit; therefore, it is important to know which components of an application are most relevant to success. We aim to shed light on shifting trends in this area in order to maximize success for the students and the college itself. Our data consists of admission and graduation information for all newly admitted freshmen enrolled at NC State University (NCSU) from Fall-2009 to Fall-2012, excluding athletes. Athletes were removed due to a difference in admissions process, and to resolve privacy issues. Variables include minority status, gender, High School GPA (weighted and unweighted), High School percentile, SAT scores, ACT scores, transfer/AP credit, applied college, residency, county distress tiers[1], and years to graduate. As the data consists of many numeric variables, one of our goals is to reduce the number of numeric variables of this data structure. We also visualized all the possible scores of admitted students among groups and tried to hypothesize underlying major factors behind theses scores. Additionally, we examined the profile of admitted students, and the relationship between students' applications and academic success. In this instance, student success is defined by the time it took for a student to graduate as this is the measure subject to the least subjectivity available to us. We classified the students into groups based on the students' graduation duration. While this is not an all-encompassing factor of student success (as it is often impacted by changing major, studying abroad, etc), graduation

[1] https://www.nccommerce.com/grants-incentives/county-distress-rankings-tiers

rates are often used as a benchmark of institutional quality, and is a target for improvement within the university.[2] To compute all the statistics, we will use publicly available statistical software R (https://www.r-project.org/) and documented packages.

# 2. Methodology

## 2.1 Basic Summary Statistics

Since our data consists of many valuable indicators of admitted students at NC State University, we visualize those indicators e.g., High School GPA, SAT scores and ACT scores among groups through several boxplots. To validate the visualization, we applied some test statics, mainly t-test, to validate the data visualization. We also looked at correlations between these indicators via correlation plot.

## 2.2 Principal Component Analysis

To reduce the dimension of the data structure, we performed a Principal Component Analysis (PCA) on our observational admission data set. Prior to that, we scaled the variables since they had a wide range of standard deviations. Additionally, we ran into issues with missing values; many students take either the SAT or ACT, and, had one or the other missing. For these students we calculated the missing test score based on the SAT/ACT concordance tables.[3]

## 2.3 Factor Analysis

Factor Analysis reveals the underlying few significant factors among many manifest variables. We only took numeric variables- high school GPA in scale of 4.00, total SAT score, ACT score with concordance values in missing observations, high school position percentile and transfer credit amount- as a dataset for factor analysis. We discarded individual verbal and math SAT

[2] https://www.northcarolina.edu/sites/default/files/unc_retention_and_graduation_report_2014rev.pdf
[3] https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf

scores, as they are redundant here. Then we use 'factanal' function in statistical software R to run an exploratory factor analysis (EFA) to find out underlying factors. To say that, ultimately by default we use default maximum likelihood estimation based approach for this EFA. After finding out two factors, which is statistically significant, we fit two hypothesized models into a confirmatory factor analysis (CFA) using package 'sem', and visualize the models by using package 'DiagrammeR'.

### 2.4 Classification

Our goal with classification is to develop a model that will predict individual student's success and overall graduation rates based on applicant data. We will first create training and testing datasets and examine data structure to determine possible classification methods, then classify the students into groups based on years to graduate using an appropriate classification method after assessing various appropriate methods.

## 3. Results

### 3.1 Basic Summary Statistics

The dataset shows us that the average GPA, SAT scores and ACT scores are 3.605, 1198 and 25.97, respectively. The median GPA, SAT and ACT are 3.63, 1190 and 26, respectively. The group wise comparison reveals that the average GPA of male students (3.59) is lower than that of female students (3.62). On the other hand, admitted male students have an average of 1221 in SAT and 26.43 in ACT, where admitted female students have an average of 1170 in SAT and 25.5 in ACT. First generation students have an average GPA of 3.63, and non-first generation students have an average GPA of 3.60. All these comparison between two groups are validated by significance in t- test. We found by visualizing the data that average GPA, SAT and ACT are in increasing trend over the time period of fall'09 to fall'12. We also found that the average

GPA, SAT and ACT scores are higher for students who complete their study in four years than students who drop out and complete study in 5/6 years. (Figure 01) The non-overlapping notches in boxplot indicates that there is strong evidence of different median among groups.

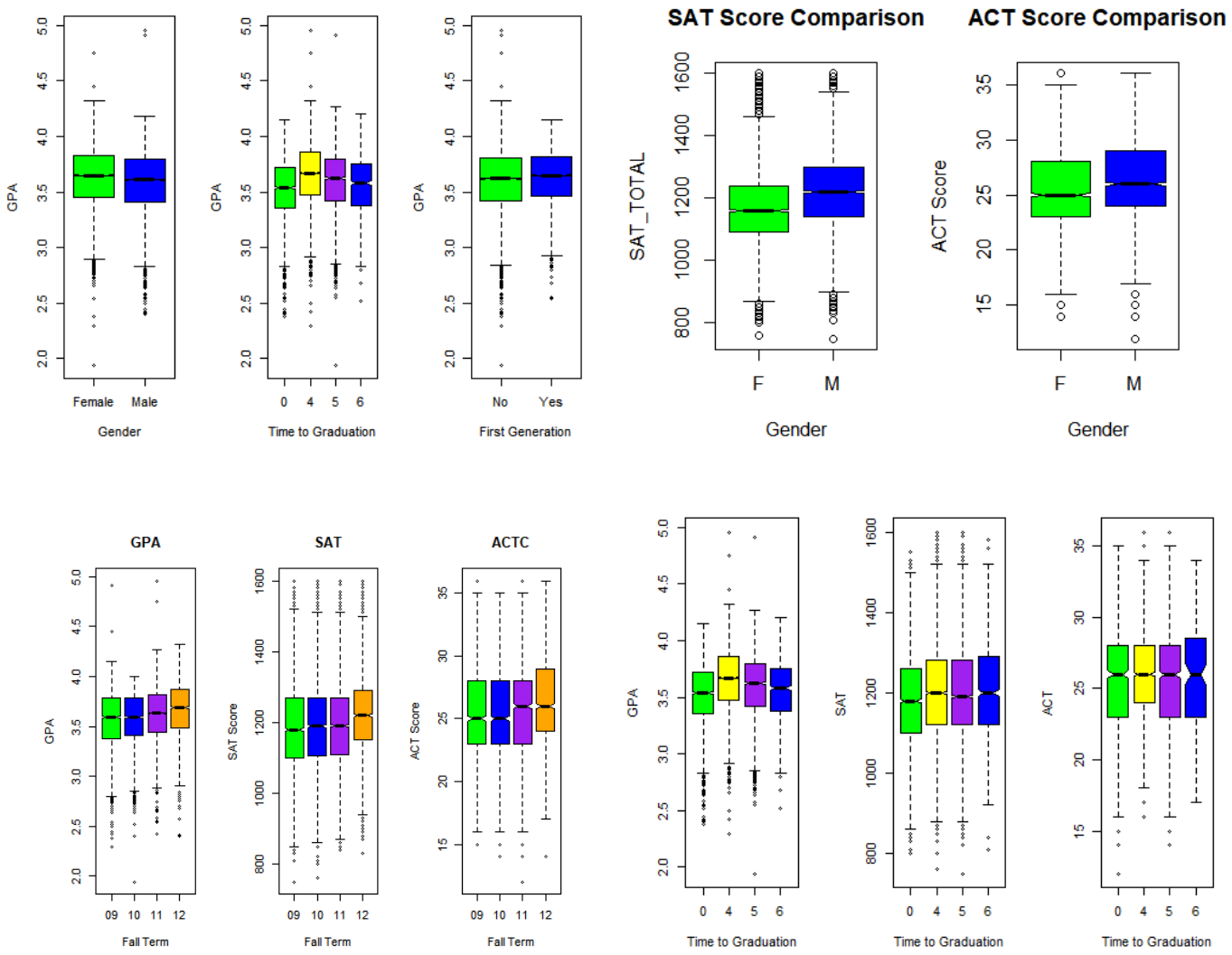


*Figure 01: High School GPA Comparison Among Gender, Graduation Time and Generation (Upper Left), SAT and ACT Score Comparison Between Male and Female (Upper Right), GPA, SAT and ACT Comparison over 4 years (Lower Left) and GPA, SAT and ACT Comparison among time take to graduate*

## 3.2 Principal Component Analysis

To reduce the dimension of the data structure, we performed a Principal Component Analysis (PCA) on our observational admission data set. After examining the proportion of variance and scree plot below, we decided to retain the first three principal components, explaining 82.7% of the variability.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8
Standard deviation     1.9832 1.3814 0.87978 0.74654 0.57507 0.50460 0.47542 0.1266
Proportion of Variance 0.4916 0.2385 0.09675 0.06967 0.04134 0.03183 0.02825 0.0020
Cumulative Proportion  0.4916 0.7302 0.82691 0.89658 0.93792 0.96974 0.99800 1.0000
```

*Table 01: PCA Summary*

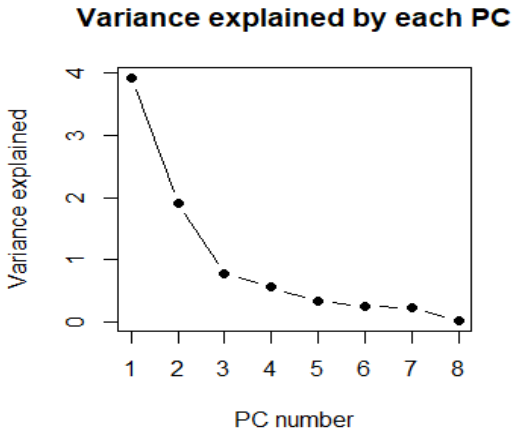


*Figure02: PC's to Retain*

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| HS_GPA | 0.3517489 | 0.42568422 | -0.003867612 |
| HS_GPA_UWH | 0.2782877 | 0.51246924 | 0.140488756 |
| satm | 0.3917442 | -0.20060840 | 0.127840015 |
| satv | 0.3783839 | -0.26407945 | 0.105961239 |
| sat_total | 0.4529768 | -0.27351379 | 0.137587835 |
| actc | 0.4102285 | -0.26801445 | 0.137792245 |
| HSPCT | 0.2350226 | 0.54642551 | 0.051582345 |
| trf_credits | 0.2710716 | -0.03105084 | -0.955038870 |

*Table 02: Retained PC1, PC2 and PC3*

Examining these principal components, PC1 can be interpreted as a general weighted sum, PC2 as a High School grade component, and PC3 as a contrast between transfer credit and high school performance.

## 3.3 Factor Analysis

The correlation plot using the package 'corrplot' groups the manifest variables clearly into two groups. Thus, we initially assumed that there are in fact two factors controlling all the manifest variables. However, prior to running a two factor FA, we run a three factor FA. With no surprise, the three factors model was not executable in R, since three factors are too many for the five

manifest variables. And, then the two factor FA confirms there are two underlying factors indeed. We accept the two factors model because the p-value of the model is 0.073, which indicates that the two factors are sufficient. By observing the factor loadings, we find out that first factor controls SAT, ACT and transfer credit, whereas the second factor controls high school GPA and position percentile. We also looked at the uniqueness of each manifest variable. The uniqueness of each variable is small enough to explain, whereas the uniqueness of transfer credit is higher, 0.814.

```
Call:
factanal(x = std.dat1, factors = 2)

Uniquenesses:
 HS_GPA_UWH    SAT_TOTAL         ACTC        HSPCT trf_credits
      0.506        0.081        0.272        0.043        0.814

Loadings:
             Factor1 Factor2
HS_GPA_UWH   0.187   0.677
SAT_TOTAL    0.955
ACTC         0.850
HSPCT                0.976
trf_credits  0.402   0.156

                Factor1 Factor2
SS loadings       1.836   1.448
Proportion Var    0.367   0.290
Cumulative Var    0.367   0.657

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 3.21 on 1 degree of freedom.
The p-value is 0.073
```

*Table 03: Factor Analysis*

```
> summary(pca.dat1)
Importance of components:
                          PC1    PC2    PC3     PC4    PC5
Standard deviation     1.5356 1.1857 0.8577 0.56604 0.4243
Proportion of Variance 0.4716 0.2812 0.1471 0.06408 0.0360
Cumulative Proportion  0.4716 0.7528 0.8999 0.96400 1.0000
> pca.dat1
Standard deviations (1, .., p=5):
[1] 1.5356448 1.1856528 0.8576931 0.5660354 0.4242510

Rotation (n x k) = (5 x 5):
                PC1        PC2          PC3         PC4         PC5
HS_GPA_UWH  0.4011057 -0.5645728 -0.121908722 -0.70909957 -0.05184421
SAT_TOTAL   0.5366627  0.3561984 -0.258488335  0.01177735  0.71983395
ACTC        0.5187340  0.3792364 -0.318260412  0.09666538 -0.69026149
HSPCT       0.3571609 -0.6218695 -0.007056683  0.69635285  0.02751851
trf_credits 0.3930121  0.1543964  0.903870875 -0.05279782 -0.04396679
```

*Table 04: PCA of 5 variables*

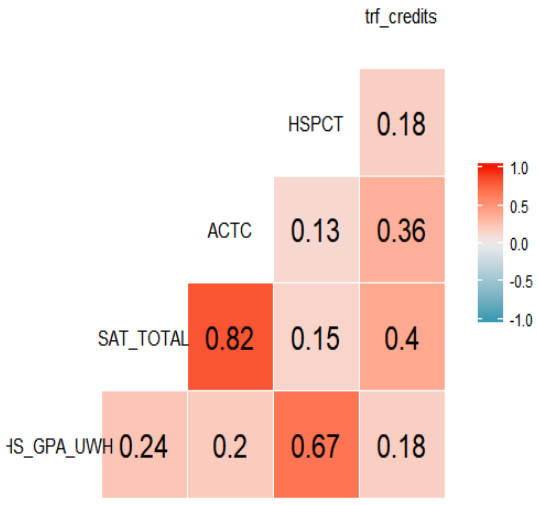


*Figure 03: Correlation Plot of 5 Variables*

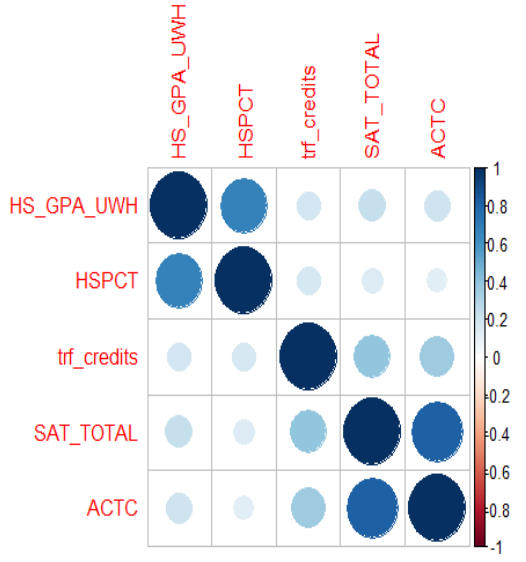


*Figure 04: Grouped Correlation Plot*

Following the revelation of two factors, we hypothesized that high school GPA and percentile are controlled by high school performance, and SAT, ACT and transfer credits are controlled by test score performance (Model 1). We also fit another model (Model 2) where we included transfer credits into high school performance factor. We feed the models into a CFA and visualize through the graphical diagram (Figure 05 -06). Even though our model fits are not statistically significant by CFA due to a p value of 0. But the model 2 is better fit than model 1 according to the AIC and BIC.

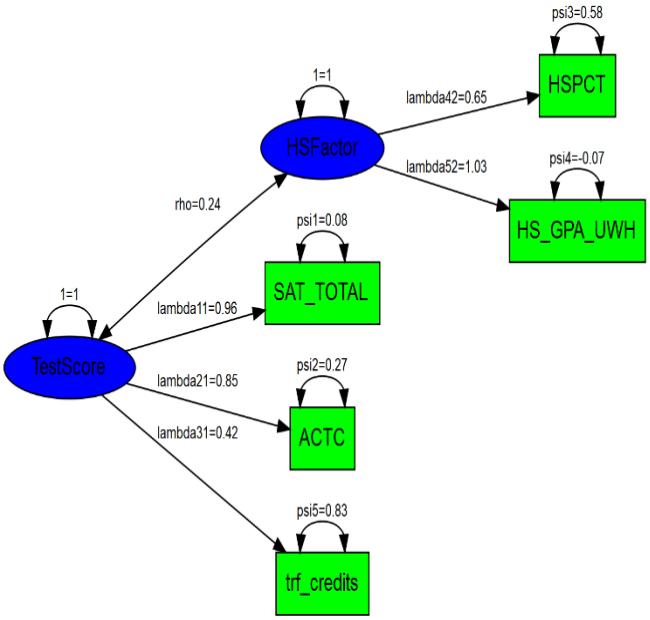


*Figure 05: CFA Model 01*

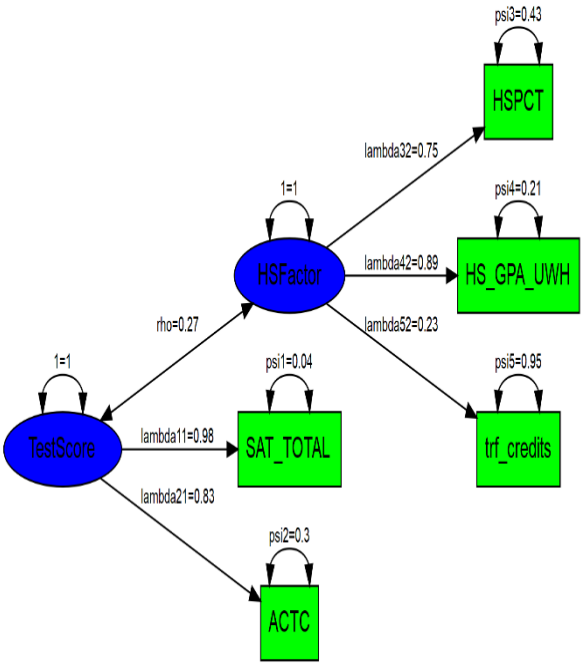


*Figure 06: CFA Model 02*

## 3.4 Classification

We first examined the relationship between our variables and groups, and then fit the following models using our training dataset: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbor (KNN), and Regression Tree (RPART).

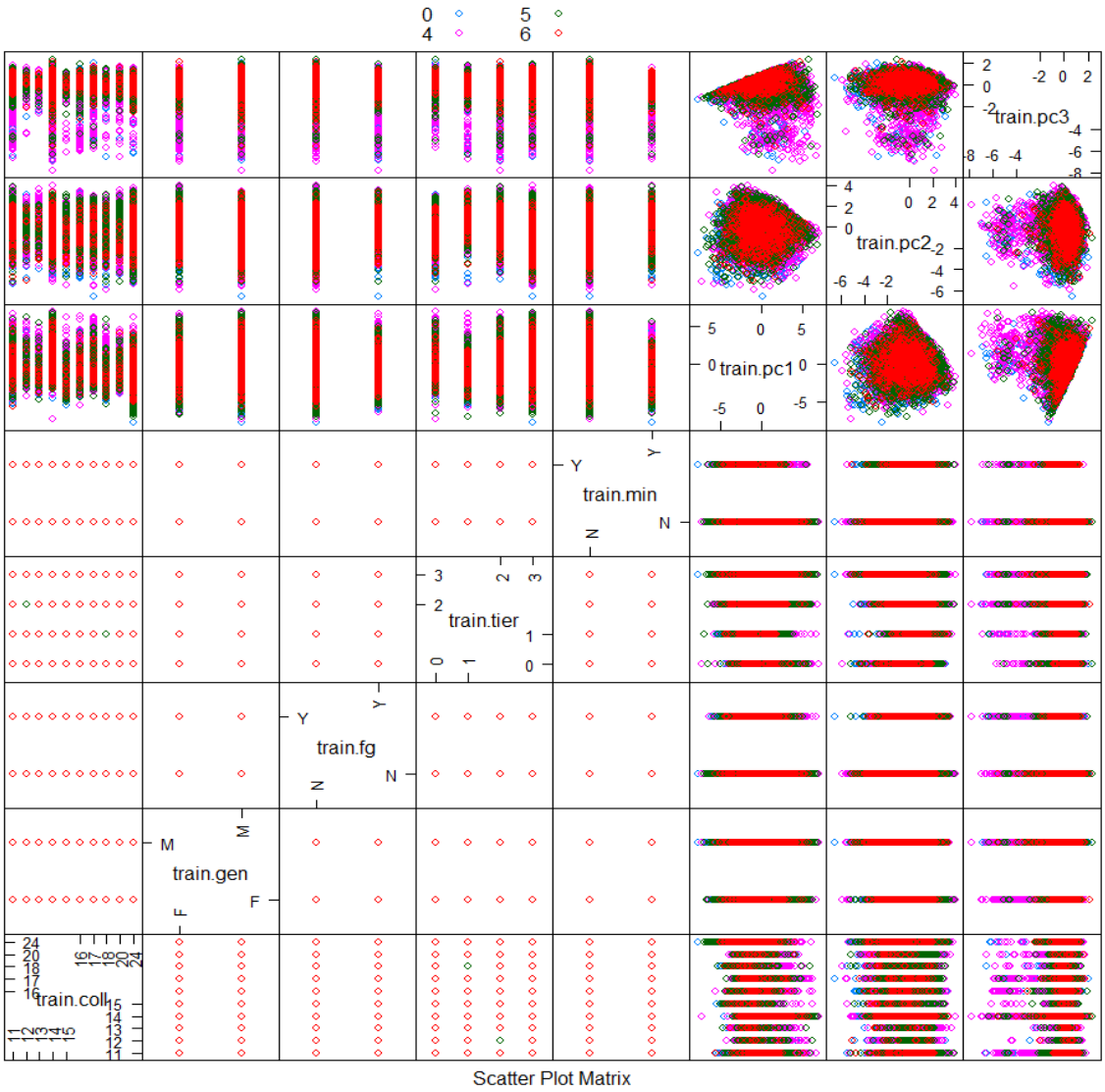


*Figure 07: Scatter Plot Matrix*

We also considered SVM, which seemed like a potentially good fit due to the non-linear appearance of our data, seen above. SVM however is not well optimized for large datasets[4], and our computers did not have the necessary resources for processing our data.

The resulting accuracy tables for our trained models are below.

[4] https://pdfs.semanticscholar.org/975e/2e0204cb7a37f6b873795c425616a8678178.pdf

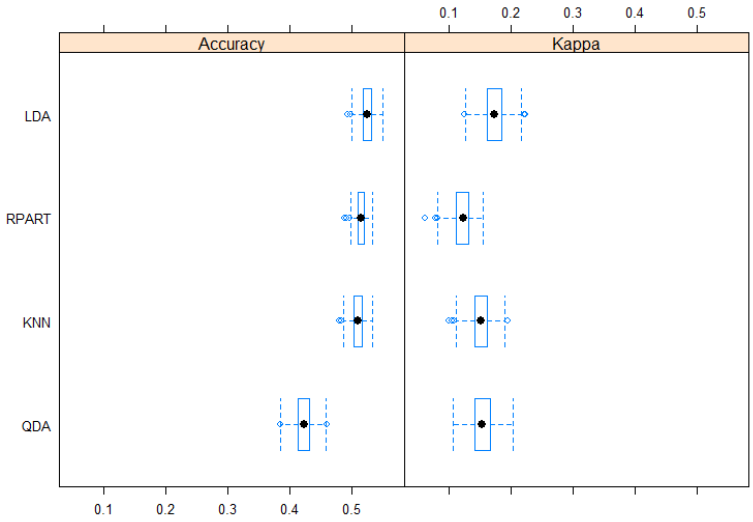


*Figure 08: Model Accuracy For Graduation Years*

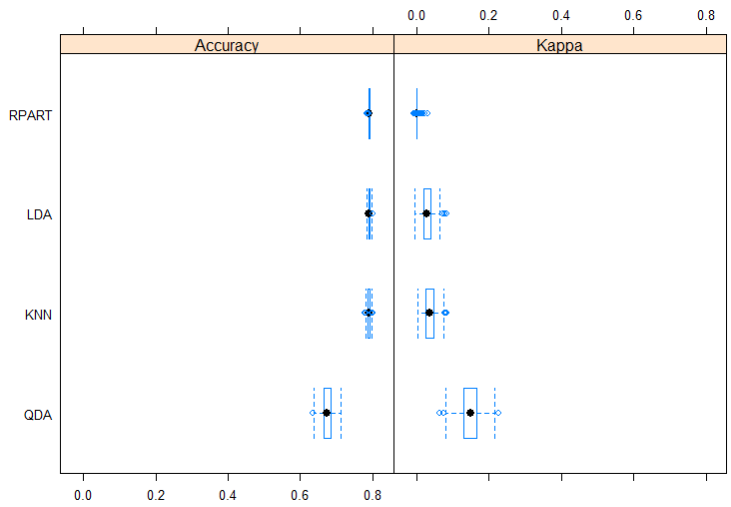


*Figure 09: Model Accuracy For Graduation Status*

In our initial model LDA and RPART had the highest accuracy, however this accuracy was only slightly higher than 50%. Examining the the confusion matrices created from our test dataset, we see similar accuracy, but LDA is better at predicting if a student will not graduate.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    4    5    6
         0  107   83   81   18
         4  459 1471  575   75
         5  163  169  228   53
         6    0    0    0    0

Overall Statistics

               Accuracy : 0.5187
                 95% CI : (0.5019, 0.5354)
    No Information Rate : 0.4948
    P-Value [Acc > NIR] : 0.002584

                  Kappa : 0.1575

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 4 Class: 5 Class: 6
Sensitivity           0.14678   0.8537  0.25792  0.00000
Specificity           0.93389   0.3695  0.85181  1.00000
Pos Pred Value        0.37024   0.5702  0.37194      NaN
Neg Pred Value        0.80520   0.7206  0.77135  0.95807
Prevalence            0.20936   0.4948  0.25388  0.04193
Detection Rate        0.03073   0.4225  0.06548  0.00000
Detection Prevalence  0.08300   0.7410  0.17605  0.00000
Balanced Accuracy     0.54033   0.6116  0.55486  0.50000
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    4    5    6
         0   11    8    8    0
         4  542 1536  629   84
         5  176  179  247   62
         6    0    0    0    0

Overall Statistics

               Accuracy : 0.5152
                 95% CI : (0.4985, 0.5319)
    No Information Rate : 0.4948
    P-Value [Acc > NIR] : 0.008433

                  Kappa : 0.1239

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 4 Class: 5 Class: 6
Sensitivity          0.015089   0.8915  0.27941  0.00000
Specificity          0.994188   0.2865  0.83949  1.00000
Pos Pred Value       0.407407   0.5503  0.37199      NaN
Neg Pred Value       0.792185   0.7294  0.77395  0.95807
Prevalence           0.209362   0.4948  0.25388  0.04193
Detection Rate       0.003159   0.4411  0.07094  0.00000
Detection Prevalence 0.007754   0.8016  0.19070  0.00000
Balanced Accuracy    0.504639   0.5890  0.55945  0.50000
```

*Table 05: Confusion Matrix*

If we mapped the confusion matrix for overall graduation status instead of year, as seen below using the LDA model, the accuracy is much higher.

| | Not Graduating | Graduating |
|---|---|---|
| Not Graduating | 107 | 182 |
| Graduating | 622 | 2571 |

Given this, in an attempt to create a model with increased accuracy we also examined overall graduation status, regardless of years to completion. In these models RPART, LDA, and KNN all had an acceptable accuracy of around 80%. Examining the the confusion matrices created from our test dataset we see KNN has marginally higher accuracy, however none are very good at predicting when a student will not graduate. For example, RPART predicts everyone will graduate, and since 79% of the population graduates, it still has relatively high accuracy.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0    0    0
         1  729 2754

               Accuracy : 0.7907
                 95% CI : (0.7768, 0.8041)
    No Information Rate : 0.7907
    P-Value [Acc > NIR] : 0.5099

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7907
             Prevalence : 0.2093
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

RPART

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0   32   29
         1  697 2725

               Accuracy : 0.7916
                 95% CI : (0.7777, 0.8049)
    No Information Rate : 0.7907
    P-Value [Acc > NIR] : 0.4601

                  Kappa : 0.0503

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.043896
            Specificity : 0.989470
         Pos Pred Value : 0.524590
         Neg Pred Value : 0.796318
             Prevalence : 0.209302
         Detection Rate : 0.009187
   Detection Prevalence : 0.017514
      Balanced Accuracy : 0.516683

       'Positive' Class : 0
```

LDA

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0   21   30
         1  708 2724

               Accuracy : 0.7881
                 95% CI : (0.7742, 0.8016)
    No Information Rate : 0.7907
    P-Value [Acc > NIR] : 0.6551

                  Kappa : 0.0272

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.028807
            Specificity : 0.989107
         Pos Pred Value : 0.411765
         Neg Pred Value : 0.793706
             Prevalence : 0.209302
         Detection Rate : 0.006029
   Detection Prevalence : 0.014643
      Balanced Accuracy : 0.508957

       'Positive' Class : 0
```

KNN

*Table 06: Confusion Matrix*

As such, we were unable to create a satisfactorily accurate model for classifying students based on anticipated graduation status.

# 4. Discussion

Initial inquiry of data suggests that the average GPA, SAT and ACT scores are very good criteria to assess the admission probability of a student at NC State University. However, the average GPA, SAT and ACT scores are in increasing trend from fall-2009 to fall-2012 among the admitted students. Admitted male students during this time period has lower high school GPA than female students, which was compensated by higher average SAT and ACT scores of male students than female students. Breaking the SAT scores down to each component of the exam may also have been beneficial in measuring success. This immediately shows up the differences and subjectivity in measurements of success and contributed to our selection of graduation as a measure of success. Also, first generation students have higher average GPA than non first generation students. Other factors relating to this information were unfortunately unavailable to us and would have added another facet to our analysis but data restrictions prohibited that. We attempted to run Hotelling's T2, ANOVA and MANOVA tests on this multivariate data, but the missing values did not permit us to do so. Thus, missing values in the dataset was an impediment against some analysis and could perhaps indicate a need to revise data collection methods. Factor analysis reveals that two factors- high school performance and test scores- controls all the manifest variables, which is confirmed by EFA statistically and significantly. However, following EFA, our hypothesized model did not fit well significantly through CFA. Even though we think the model fit is explainable and easily interpretable through visualization. This may be due to cross loading or correlation between items in the factors or variables which may require further analysis to truly understand. We fitted transfer credits into two different model differently, where Model-02 has lower AIC/BIC value than Model-01. But we prefer Model-01, since it explain the model well taking account of the lower communalities of each variables, and

also, transfer credits grouped well with test score factor well according to EFA and correlation plot.

The aim of this project was to identify the key factors which indicate that a prospective applicant will indeed be successful in a university environment. Our goals were to not only identify these but to also assess the importance of each factor relative to the rest in a manner so that they can be used beyond an academic setting.

We were unsuccessful in finding a model to predict graduation status based on academic careers. As mentioned in 3.4 we were unable to create a model using SVM; further testing could be done to examine if a better model could be created using this algorithm in particular if improved technology were at hand. It could also be that no model using application data would be accurate, as there are too many external factors impacting graduation (eg. work, family life, financial burdens, etc.) Any similar studies found during research, particularly those aimed at predicting college GPA, had these mentioned factors and others, including race and ethnicity, for example, which were not available to us and proved to be pivotal in their findings. It would be useful to investigate this further, because if available application data is not a satisfactory indicator of student success, then should it be used to determine admission and what refinements can be made if they are even possible?

# 5. References


1. https://www.nccommerce.com/grants-incentives/county-distress-rankings-tiers
2. https://www.northcarolina.edu/sites/default/files/unc_retention_and_graduation_report_2014rev.pdf
3. https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf
4. https://pdfs.semanticscholar.org/975e/2e0204cb7a37f6b873795c425616a8678178.pdf
5. https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=4546