Fall 2022 – MSBX 5415

Term Project

Law School Admissions for Bar Passage

Team Members

Danielle Allen, Morgan Likens, Micaiah Lowe, & Samantha Fildish

University of Colorado Boulder

1. Introduction

1.1 Business Understanding

Each law school student who is accepted into a program must first take the LSAT and submit their LSAT score and Undergraduate GPA (UGPA) as part of their application package to law school. By using the LSAT score and UGPA, we attempted to predict whether a student would or would not pass the bar exam. If successful, this could assist law school admissions offices when selecting whether to admit a student into the program or not. Additionally, we explored the dataset provided by LSAC for other important factors contributing to or hindering passing the bar exam, e.g. does race, gender, or background have a correlation with a range of LSAT scores and how high is their bar passing rate.When reviewing admissions files, a model that could predict the probability of success of a potential law school student could be useful in situations where admissions are limited and the admissions office is looking for candidates that have the highest chance of succeeding and passing the bar exam.

1.2 Data Understanding

Our dataset was downloaded from Kaggle. However, the dataset was originally collected by the Law School Admissions Council(LSAC) for a study called 'LSAC National Longitudinal Bar Passage Study' in 1998. According to the study website “From 1991 through 1997, LSAC tracked some twenty-seven thousand law students through law school, graduation, and sittings for bar exams. The result was the most comprehensive database that exists on the demography, experiences, and outcomes of a large cohort of aspiring lawyers” [1]. The purpose of the original study completed by LSAC began “primarily in response to rumors and anecdotal reports suggesting that bar passage rates were so low among examinees of color that potential applicants were questioning the wisdom of investing the time and resources necessary to obtain a legal education."[2] Ultimately, the data we analyzed was provided by students, their law schools, and the state boards of bar examiners during the time frame indicated by the study. We have detailed demographic information included in this dataset because the goal of the LSAC was to both report the general bar examination outcome data, and to analyze factors including ethnicity and gender that might explain differences in outcomes of taking the bar exam. However, it does seem that we had a pared down dataset compared to the full study report. Information including Jurisdiction, Region, and even more detailed Ethnic Groupings were part of the original sample data. According to the study, the data available represents more than 93% of the students who entered law school in 1991 and are known to have graduated.

1.3 Data Preparation

The original dataset we selected had 39 variables to work with, the focus being on "pass_bar" (1 being a pass and 0 being a fail). The rest of the dataset contained the following student information; sex, race, undergraduate GPA (UGPA), DOB year, LSAT scores, if they had to take the bar multiple times to pass, whether or not they were part time students, their family income as a quantile, as well as the law school rankings by decile of each candidate. One flaw in using this dataset to predict our model for current use, is that the 1991 LSAT score was on a different scale than the modern LSAT, it switched from a 48 point system after 1991 to a 120-180 point system. In order to utilize this model in practice, we would need to convert the old scores to the new measuring system.

To begin our descriptive analytics, we first examined the breakdown of the data by gender to understand how many responses we had for male (11,501) and female (8,926) and how many of each passed or failed the bar exam. Overall, the pass rate for female students was 94% while the pass rate for male students was 95%. Additionally, we looked at the average UGPA and

LSAT scores for females and males and found that males had a slightly higher average LSAT score while females had a slightly higher UGPA [Fig. 1]. Second,we examined the data by each race represented in the data. We looked at the total number of students who passed and failed, their average UGPA and average LSAT score grouped by race. Finding that the group of white students had the highest LSAT scores as well as UGPA (closely followed by the group of Asian students) [Fig. 2]. These summary statistics identified that our data set is skewed toward data gathered for white students (representing 84% of the sample) which may have an impact on the results identified above.

When we first began the exploratory analysis of our dataset we decided to omit rows that had values missing. In doing so, only 1,980 rows of the dataset were lost, most of which belonged to those who passed the bar. This left us with an estimated 20,000 observations remaining, which is a meaningful amount of data to work with. Many of the original variables, race, gender, UGPA, and full or part time students, were redundant, to avoid this we either removed or transformed the variables. Initially, we had a total of 7 variables that all referred to race, 4 of which were dummy variables; we omitted all but a single numerical column representing race, then created dummy variables with binary values for each race [Fig. 3].

There were also quite a few variables with unclear meanings, our original data set only contained descriptions for 4 of the 39 variables. Our first analysis looked at the columns decile1, decile1b, decile3, zfygpa, and zgpa. Looking at the range and values of zfygpa and zgpa, we surmised the values were a z-score of some data; however, neither zfygpa nor zgpa had a high correlation with UPGA. Upon a larger correlation table, we noticed zfygpa correlated above 0.99 with decile1 and zgpa correlated above 0.99 with decile3 [Fig. 4]. We made an educated assumption that zfygpa represents the z-score of the first year GPA (L1 students), and that zgpa is the z-score for the final year GPA of each student (L3 students) [Fig. 5].

Next, neither the attributes fam_inc (family income) nor tier were given descriptions, considering their names we decided to explore what they could be as they could prove statistically significant in our model. Knowing that economists create quantiles for income brackets, we assumed that fam_inc is the quantile ranking of family income for each student. For fam_inc the proportion of students who passed the bar steadily increased from the lowest value 1, where 86% of the sample passed, to the highest value 5, where 96% had passed. Inversely, there was a steady decrease for not passing, from 14% to 3%.

We then focused our efforts on the variable tier. Determining the intended meaning of tier proved more difficult than family income. After an in-depth analysis, the proportions of tier closely resembled those who passed or did not pass the bar based on family income, with the only exception being the lowest tier, 23% of those in tier 1 did not pass the bar. Apart from age, which decreased, the mean UGPA, LSAT, and family income increased as the tier value increased. We hypothesize that tier stands for the law school ranking of each student, where 1 is the lowest and 6 is the highest.

Our first analysis of the variable cluster showed no clear distribution, trends with means, correlations with other variables, or proportions of those who passed or did not pass the bar. After spending more time than necessary examining this variable, we eventually discovered that cluster was in fact the tier of each student just randomly rearranged, where cluster 1 represented tier 4, cluster 6 tier 1, and so on. This further supported our assumption of what tier represented. This displayed how a higher tier represents a higher caliber of student, which we assume is taken from their law school ranking.

To prepare for deeper analysis we created two "cleaned" datasets. One with just basic adjustments to column names for understandability, standardizing dummy variables, as well as reducing variables that represented each race from 8 variables to 5. We used this first dataset to complete exploratory data analysis. The second cleaned dataset was geared toward modeling and included the adjustments made to the first cleaned dataset. Additionally, there was also the removal of the unnecessary variables from the dataset, inclusion of a calculated column to represent age (based on the DOB year provided), and a creation of two new columns utilizing the z-score to standardize the UGPA and LSAT variables for our analysis. Our final cleaned dataset, included the following information: the law school rankings by decile of each candidate, LSAT scores,undergraduate GPA (UGPA), family income quantile, tier, sex, race, age, and whether or not they were part-time or full-time students. With our prepared data, we then split the data into training and validation sets using an 80:20 split and began modeling.

2. Modeling

In an article by Baptiste Rocca, a data scientist at ManoMano, he writes: "When using a resampling method [for an imbalanced class distribution] . . . we show the wrong proportions of the two classes to the classifier during the training. The classifier learned this way will then have a lower accuracy on the future real test data than the classifier trained on the unchanged dataset" [3]. He concludes that one should use caution when employing undersampling or oversampling, and, furthermore, they should understand the impacts that resampling methods have on the distribution of data itself.

Due to a moderate imbalance on our dataset, only 5% of the data are classified as not passing the bar, we evaluated multiple types of models, with some employing oversampling and others trained on the imbalanced dataset. The base metric we used to compare all models with each other was an F1 score of the negative class (F1_0). We chose this metric since the negative class, pass_bar with values 0, is the minority class; we wanted our results to have the best performance on the minority class as opposed to the majority class. Additionally, we desired to learn whether oversampling could improve model performance on unaltered testing data, and, if not, what other models, or cost / loss functions had the highest predictive success for both classes, specifically the minority class.

2.1 Modeling: Logistic Regression and Backward Selection

To start our modeling efforts we built a logistic regression model, using only LSAT score and UGPA as independent variables, pass_bar as the dependent variable, and a threshold of 0.5. Both features were listed as significant, and the overall model accuracy was 95%, however the overall specificity was very low at 2.9%. The F1_0 for this model was 6%. With additional inspection of the results using a confusion matrix, we discovered that the model predicted that nearly every student would pass the bar exam. We then created a second logistic regression model using all features to make our predictions and noticed the same issue, nearly all students were predicted to pass. While this also gave us an accuracy of 95%, we could again determine that the model is biased towards predicting a student will pass the bar.

Due to our model predicting mostly passing results, and the knowledge that our data is skewed towards passing, we decided to use the upsampling technique on the training data to even out our proportions and balance the bias for better results. Once we upsampled our training data, we trained a logistic regression model using all variables on the upsampled training dataset. This time, we saw a drop in overall accuracy to 78%; however, our model was more accurately predicting both passes and failures. The F1_0 of the upsampled logistic model increased to 27% with overall specificity increasing to 81%. The improvement in specificity showed that our model was learning how to better predict when a student would fail the exam, and was accurately predicting this class 81% of the time, compared to the earlier 2%.

Our next task was to reduce the number of variables being used in the model to find the best set of variables. We implemented backward stepwise logistic regression, which consisted of starting with all available variables and removing variables to reduce the number of predictor variables. Upon completion of the backward selection, we found the best performing model to include the features; decile3, decile1, lsat, ugpa, tier, zdiffgpa, other, asian, black, hispanic, female, y.fulltime, and age. Taking these features, we built the model we would use for final predictions. We decided to use the upsampled training dataset to train the top performing model. Prior to using our model for predictions, we examined the model summary to analyze the variable coefficients. Of the variables selected those that had a positive impact on probability of passing the bar included; decile3, decile1, lsat, ugpa, tier, zdiffgpa, fulltime. This suggests that if you have a higher lsat score and gpa, are in a higher ranked tier or decile, and are a fulltime student, you have a better chance of passing the bar exam. Logically, this made sense. On the other hand, the features that had a negative impact on the probability of passing the bar exam included; the race other, race asian, race black, race hispanic, sex female, and age. Those students included in these race and sex categories would have a lower probability of passing the bar exam. This could allude to the feasibility of the initial study by the LSAC investigating the impacts of race on the bar exam, but could also be affected by the fact that within our dataset a majority of the students fall into the category of white males. Based on these findings, inquiries into the accuracy of the LSAT and / or the bar exam, as they stand, may be warranted if certain groups of students are left at a disadvantage because of factors such as race, gender, or background, which we explore further in cluster analysis (section 2.8).

After fitting our model and exploring the model summary, we made predictions with our unaltered test dataset, using a threshold of 0.5. Our prediction accuracy using this model was 78%. The F1_0 and specificity stayed about the same as the upsampled model using all features. Since we were able to reduce the number of features without degrading performance, we decided to state the model identified using backward selection trained on the upsampled dataset was the best performing basic logistic regression model. We also generated an ROC Curve and it shows the AUC of 0.791 [Fig. 6].

In conclusion, the prediction accuracy of our final model decreased compared to the simple model where nearly every student was predicted to pass the exam.Although other important metrics, such as specificity, did improve [Fig. 7]. Our biggest predictive error falls into the category of False Negatives, or type II errors. Out of the 3884 instances that did pass the bar exam, we predicted that 859 of those instances would not pass. Out of the 208 that did not pass the bar exam we only predicted that 41 would pass.

2.2 Modeling: Threshold Optimization with Logistic Regression

After oversampling the data with logistic regression, we decided to keep the original distributions of the two classes intact and run another logistic model, this time with a focus on increasing the threshold for predictions. Using the pROC package, we plotted the ROC for our model and found the optimal threshold to be 0.9564165. The threshold increased logistic model scored higher for recall on class 1 (2% increase) and precision for class 0, the latter increasing from 7% to 72%; however, it lowered, roughly, recall for class 0 by 30% and precision for class 1 by 20% with a F1_0 of 21.6%. Although the F1_0 of the threshold logistic model dropped, the model predicts positive cases with a 98% accuracy (recall), in other words, when it does predict a positive we can be 98% sure it is a positive. Additionally, we can trust this model to continue to perform well for the positive class on unseen data as no resampling techniques were employed on the training data.

2.3 Modeling: Focal Loss

After optimizing our logistic regression model, we decided to go further by creating a single layer neural network with sigmoid activation (a basic logistic regression model), which would utilize focal loss in training the weights. Focal loss was first introduced by Lin et. al, from Facebook, one of its many benefits is that it increases the weight of a majority class [4]. First we built a logistic regression model with Pytorch, then we expanded our single-layer model into a three-layer model, turning our logistic regression model into a neural network. In lieu of cross-entropy, we utilized focal loss in training both the logistic model and neural net.

Our best results in applying focal loss came with simple logistic regression and not a neural network. The logistic model outperformed the neural net with an overall accuracy of 94% and F1_0 of 36.9%, while the neural net performed with a 94% and 27%, respectively.

2.4 Modeling: Fine-Tuning XGBoost with and without Resampling

Next, we examined the performance of a base XGBoost model, 4 trees with a learning task of binary logistic regression. The base model performed roughly the same as our base logistic regression model with a F1_0 of 13%. Using the original data (unaltered with resampling techniques), we created a 10,000-row data frame with randomly generated parameters as columns for max_depth, eta, subsample, colsample_bytree, and min_child_weight and saved the results of the highest accuracy and AUC for the parameters of each row. Filtering the data frame based on highest AUC, we reran the XGBoost with the parameters for the highest AUC. The results showed that, except for two observations, the model predicted every observation as having passed the bar. For the tuned model, specificity dropped to 0.4%, which was a significant decrease from the 7% of the base model.

The best performing XGBoost we had was when we used the parameters from our cross-validation search and scaled the weights of the positive class. XGBoost documentation recommends using the proportion of the sum of negative instances over the sum of positive instances for this number (XGBoost, “XGBoost Parameters”, 2022). In this scaled model, the specificity increased to 68% with a negative predictive value of 14% (F1_0 increased to 24%), and an overall accuracy of 77%.

Due to many articles discussing the success of SMOTE for generating synthetic data on the minority class – in our case class 0 – we decided to train a final XGBoost model on oversampled training data via SMOTE, and see how the trained model would perform on an unsampled, i.e. SMOTE was not used, testing set. Keeping a max depth of 4, we also scaled the weights of the majority class, the idea was that a higher weight on the majority class coupled with upsample, implementing SMOTE, would create a model with better specificity and a negative predictive value. The SMOTE XGBoost model was tested on a testing set that SMOTE was not applied to, so the class distribution still existed. Compared to previous XGBoost models, the performance of the SMOTE XGBoost did not produce a significant increase in performance metrics, accuracy dropped to 69%, specificity increased to 70%, while the negative predictive value decreased to 10%, and our assessing metric of F1_0 decreased to 19%. Interestingly enough, the most important feature SMOTE XGBoost identified, by a significant gap, was ugpa, which was ranked third, fourth, or fifth in the unsampled XGBoost models without SMOTE. Since we know that SMOTE distorts the real class distribution and model performance did not increase, we decided to neither use these results of important features nor the model for SMOTE XGBoost.

Despite the lack of performance with XGBoost, our best insight from XGBoost came with feature importance. All three upsampled models showed the attributes lsat, zdiffgpa, ugpa, white, and age as import features, while the latter two models showed black, tier, fam_inc, y.fulltime, and female as additional important features.

2.5 Modeling: K-Nearest Neighbors with Feature Importance

To work with K-Nearest Neighbors (KNN), we first scaled the training and testing data, then partitioned data into two new training and testing sets, one with all important features specified in each XGBoost model (including additional important features found in only 2 models), and one with only the important features consistent across all three XGBoost models. Using a for loop, we iteratively ran a KNN model with k values of 1 to 51 and assessed the test accuracy of each, we received the highest results with all important features (lsat, zdiffgpa, ugpa, black, white, age, tier, fam_inc, y.full_time, and female) with a k value of 5. This final KNN model had a total accuracy and sensitivity approximately close to the threshold-increased logistic model; however, its negative predictive value was higher, at 50%, while its specificity was lower, at 0.7%, the F1_0 of the KNN model was 1.4%.

2.6 Modeling: Decision Trees with Boosting

To begin our analysis with decision trees we started with a basic classification tree model as the baseline. We utilized all features except for p.bar 1, p.bar 2, and ID. In order to build a larger tree we used the control argument in the model and set the number of rows in the dataset as the nob argument and a mindev value of 0.001. When tested on our test dataset the baseline model reports an accuracy of 94% and an F1_0 value of 29%. Once we had our baseline model built and tested we used cross validation and pruning to improve the model. Cross validation identified that 6 was the optimal number of nodes so we pruned the tree with the best argument set as 6. Evaluation of the pruned model on our test data reports an accuracy of 94% and F1_0 score of 23%, which is slightly better than our unpruned model.

After building and pruning our base model we moved on to look at using a Gradient Boosting Machine (GBM) model instead of just a basic tree. We tried multiple different values for the parameters n.tree and determined that 2000 trees with a threshold of 0.2 produced the best results with an accuracy of 95% and an F1_0 score of 4.1% when tested. Our GBM model also identifies the top four important features as being lsat, decile3, zdiffgpa, and age with lsat having a much higher relative influence than the next three features.

Alongside our GBM model we also built an Adaboost model. We set the boost argument to true so a bootstrap sample of the training set would be used and set the mfinal argument to 100. This model results in an accuracy of 95% and F1_0 score of 27%. The Adaboost model identified the top four most important features as decile3, lsat, ugpa, and zdiffgpa, from greatest to least, with decile3 having a significantly higher relative importance in comparison with the next three features. Our Adaboost model has the highest accuracy of the tree models but does have a higher F1_0 score than the pruned model and GBM model.

2.7 Modeling: Nearest Centroid

Based on the important features ascertained in sections 2.4 and 2.6, as well as our models that utilized trees, we used this information create a nearest centroid(NC) model , also called the Rocchio classifier, from the lolR package in R, which is somewhat similar to a KNN classifier [5]. Considering that KNN performed best with all the important features we found with XGBoost, and NC's closeness with KNN, we ran a base NC model with all the important features found from XGBoost (lsat, zdiffgpa, ugpa, black, white, age, tier, fam_inc, y.full_time, and female).

The initial results from our base NC model were higher than most models thus far with an overall accuracy at 82%, specificity at 61%, negative predictive value at 15%, and a balanced accuracy of 72%, the latter being computed from the caret's confusionMatrix. After creating a base model, we ran a second model based on the top important features found across all XGBoost models (lsat, zdiffgpa, ugpa, white, age, black), for this second NC model performance results decreased, overall accuracy increased to 83%; however, balanced accuracy decreased to 69%. At this point, we began to fine-tune our NC model by deciding which features to use based on the models in sections 2.4 and 2.6, the aim being the highest balanced accuracy we could achieve.

To continue fine-tuning our NC model, we looked at the important features found in trees (decile3, lsat, ugpa, and zdiffgpa), which increased performance with an overall accuracy at 75%, specificity at 80%, negative predictive value at 13%, and a balanced accuracy of 77%. Considering there were less features in the important features found in section 2.6 than 2.4, we iteratively removed features from those found with XGBoost one-by-one, we observed an increase in balanced accuracy when features "white" and "ugpa" were removed. After iteratively removing each feature from the XGBoost models, we added decile3 as an additional feature to the XGBoost important feature list, since that was the only important feature generated from section 2.6 that 2.4 did not have. Adding decile3 to the XGBoost features increased balanced accuracy to 77%, the same as the NC model based solely on the decision tree features.

We concluded with a final round of fine-tuning. We iteratively added each race, removed zdiffgpa – which was originally a calculation of the difference between decile3 and decile1 – and added decile1. The final NC model had a feature list of lsat, decile1, decile3, black, other, asian, hispanic, age, tier, fam_inc, y.fulltime, and female, and had the highest performance of all of our

NC models based on balanced accuracy: overall accuracy at 79%, specificity at 77%, negative predictive value at 15%, a balanced accuracy of 79%, and F1_0 of 26%.

2.8 Modeling: Unsupervised Learning with Clustering

In an effort to understand our data in more depth, we explored what trends might appear while using unsupervised clustering methods. The goal of seeing our data through the lens of K-means clustering was to identify any more meaningful information that might influence our predicting models. After the exploratory data analysis, we selected LSAT, UGPA, family income, and tier as the first subset to review. We started by taking a random sample of scaled data that was about a quarter of our dataset (5,000 data points) as our clustering set. While this modeling method can handle large amounts of data, it was computationally expensive for a laptop to use the full dataset.

Because we were unsure of how this data would group, we ran through the model and plots for centers ranging from 2 to 7 as a preliminary exploration. Each model began with 25 random sets, and the maximum number of iterations was 1,000 (applied to all K-means models mentioned). We kept the default algorithm after reviewing that Hartigan-Wong generally generates the best results if you have multiple random sets represented in the model. After reviewing the preliminary plots, we used a couple of different methods to choose our final number of clusters. Running through different NbClust options, we reviewed the suggested number of clusters from the Elbow method, Silhouette method, as well as the dendrogram plot height differences. We ultimately decided on 4 clusters and went on to review what their means meant for our dataset [Fig. 8]. While not going into detail for each cluster, overall cluster 1 had the best performance out of all our input variables and applied to our random dataset showed the highest pass_bar rate as well. Cluster 2 was a grouping of poor performance on LSAT, UGPA, tier, and not surprisingly pass_bar. While cluster 3 had the highest family income, it did not seem to be an indicator of performance, with all other indicators falling in line with the dataset means. Cluster 4, on the other hand, had the lowest family income represented, the students were also slightly below average on the LSAT and tier indicators, but consistent with the overall dataset pass_bar rate of 95%. From the clusters above, we did not gain too much new insight into the data, they seemed to fall in line with the distribution of the dataset.

After looking at the above clusters, we wanted to dig into the questions that prompted the LSAC study to begin with and removed all white students to create a new dataset. The dataset of non-white students contained 3,294 observations and showed a pass_bar rate of 86% overall. Looking at the means of this dataset, LSAT, UGPA, tier, and family income were all below the averages of the dataset as a whole. This time, we included race as a part of the data we wanted clustered as well. With the settings mentioned above and after running an Elbow and Gap-Statistics method on the new scaled dataset we settled on 3 clusters [Fig 9]. Cluster 1 was inclusive of all races (other, Asian, black, Hispanic), and was the poor performance grouping with all means below the averages for this smaller dataset. Cluster 2 consisted of other and Asian students, the means were higher for LSAT, UGPA, tier, and family income than the whole dataset averages, however the pass rate was slightly lower at 94%. Cluster 3 represented black and Hispanic students, and while they had the highest average tier rankings of these clusters, they were still below the averages of the whole dataset on the other factors. Because the distribution of data shifts in removing white students, we spent time digging into the different race categories and trying to understand their different outcomes.

While we cannot conclude a causation between family income and passing the bar for non-white students, there is a correlation. Asian students had the most similar averages for LSAT, passing rate, and family income to white students, showing they perform relatively the same in this dataset, other groups were less so. Black and Hispanic students had noticeably lower average LSAT, UGPA, and passing rates; however, they also had lower average family incomes. The data displays a strong correlation between low family income and not passing the bar, which is not surprising since the lack of resources for lower income students has been shown to negatively impact UGPA, availability to study for the LSAT, and the institutions they are admitted to for law school [6]. We would say that while family income might be less influential on the dataset as a whole, it does seem to have a greater impact for those of black and Hispanic background. While it does not appear that the bar itself is inherently skewed toward passing white students, studies show that background factors such as accessibility to resources are a part of the equation and should be considered, especially if makers of the bar exam wish to see an increase in black and Hispanic people passing the bar.

3. Evaluation

We evaluated our models on a number of different metrics, such as overall accuracy, misclassification rate, recall, precision, specificity, ROC Curves, and the F1_0, the latter being our main evaluator. At the start of the project, our hope was to correctly predict if a student will eventually pass the bar using the ground truth, pass_bar. In other words, certain disadvantages may prevent a student from passing the bar the first time but over time they might eventually pass. If successful, our best performing model could help law schools identify a portion of students that they normally would reject; not that our model should be used as the deciding factor in a student’s application, but it can help admission officers make informed decisions.

The return on investment (ROI) of implementing our model can be measured on the number of students who are predicted to eventually pass the bar but would normally be denied entry into the law schools based on current admission's standards. The schools who use this model should also be collecting data on their students who do or do not pass the bar after graduation to monitor if there is an increase or decrease in bar passage after they introduce our model in the admissions process.

Despite practical applications, it is important to state: Although our best performing model we created – logistic regression with focal loss – had a recall of 97% and a F1_0 of 36.9%, which was the highest F1_0 we could achieve. The best model we produced was good at predicting if students would eventually pass the bar, no matter how many times they might take it, but it was worse than random guessing when it came to predicting if students would not pass the bar. That said, law schools could implement our model to gain information on if a student is predicted to pass, but should all together ignore predictions of not passing the bar.

After working with many different models in supervised learning and evening running cluster analysis in unsupervised learning, we believe that to better predict if students will both pass and not pass the bar, further research must be conducted, we need more data. Additionally because of our findings with unsupervised learning and descriptive analytics, we believe this research should focus on collecting more data from all races, not just the majority race, white.

4. Deployment

The business use case for implementing our models lies in potential improvements made to the admissions process, e.g. the recall of our best model was 97%, meaning that when that model did predict someone as passing the bar, it is highly likely they will pass. Admissions officers can use our model to make informed decisions when they are evaluating if an applicant will be successful at their school and eventually pass the bar. This could result in improvements to the time it takes for review as well as help identify students who may struggle to pass the bar the first time, e.g. due to family income or cultural background, but will eventually pass the bar.

In our analysis we found that, among other features, a student's LSAT score, undergraduate GPA (UGPA), and L1 and L3 school rankings are good predictors for their future ability to pass the bar exam. Both the LSAT score and UGPA of an applicant is information available prior to admitting a student. While admissions are not exclusively determined by LSAT scores and UGPA they currently play a primary role in the decision making process [7]. To train our best model to information available before the application process, we removed zdiffgpa, decile1, and decile3, accuracy remained the same and our F1_0 dropped to 31%. Once a prospective student applies, a school could then take their LSAT score and undergraduate GPA and apply our model to estimate if the student will ultimately be successful in passing the bar exam after completing their studies. This evaluation can then be taken into consideration when reviewing the entirety of the student's application for admissions.

School administrators should be aware that the outcomes from the model should not be the only thing taken into consideration for admitting applicants. The results should be used as a facet of a more comprehensive, holistic review of a student’s application to ensure that all factors, such as background, extracurricular activities, recommendations, etc., are fully considered when determining an admission decision. This is extremely important as the data used to build the model skewed heavily towards students who were white, male, and full-time. If law schools desire to use a model with high predictions for both classes and all applicants, we will need a more normally distributed dataset. While the current model can provide useful information to be included in the decision making process, it should not be given more weight than any other decision factor.

As mentioned briefly above, using our model to inform admission officers does come with risks. First, our best performing model did worse than random guess at predicting if a student will not pass the bar. And second, the school would be introducing a decision making factor that could be generalized and not be representative of the entire population that applies to their school – considering most of the data was collected in the 90s and dominated by male, white, full-time students. This is seen in the use of predictive models used by other higher education institutions to determine if a student will succeed as well as what majors would be ideal for them. While many of these predictive models are designed to help eliminate bias from administrators they often instead introduce patterns of bias. This bias can come from the data used to build the models to the interpretations of the results [8]. To help mitigate both risks discussed, we highly recommend using the results of our model as a single component part of a larger evaluation of a student's background and qualifications for an admission's decision.

Bibliography

[1]”Project SEAPHE: Databases”. Seaphe.org, 2022.
https://www.seaphe.org/databases.php

[2] Wightman, Linda, and Henry Ramsey. *“*LSAC National Longitudinal Bar Passage Study a Publication of the Law School Admission Council”. Law School Admission Council, 1988.
https://archive.lawschooltransparency.com/reform/projects/investigations/2015/documents/NLBPS.pdf

[3] Rocca, Baptiste. “Handling Imbalanced Datasets in Machine Learning”. Medium: Towards Data Science, 2019.
https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

[4] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. “Focal Loss for Dense Object Detection”. ArXiv:1708.02002, 2018.
https://arxiv.org/abs/1708.02002

[5] Balasubramanian, Vineeth, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. 115-130. Boston: Elsevier, 2014.

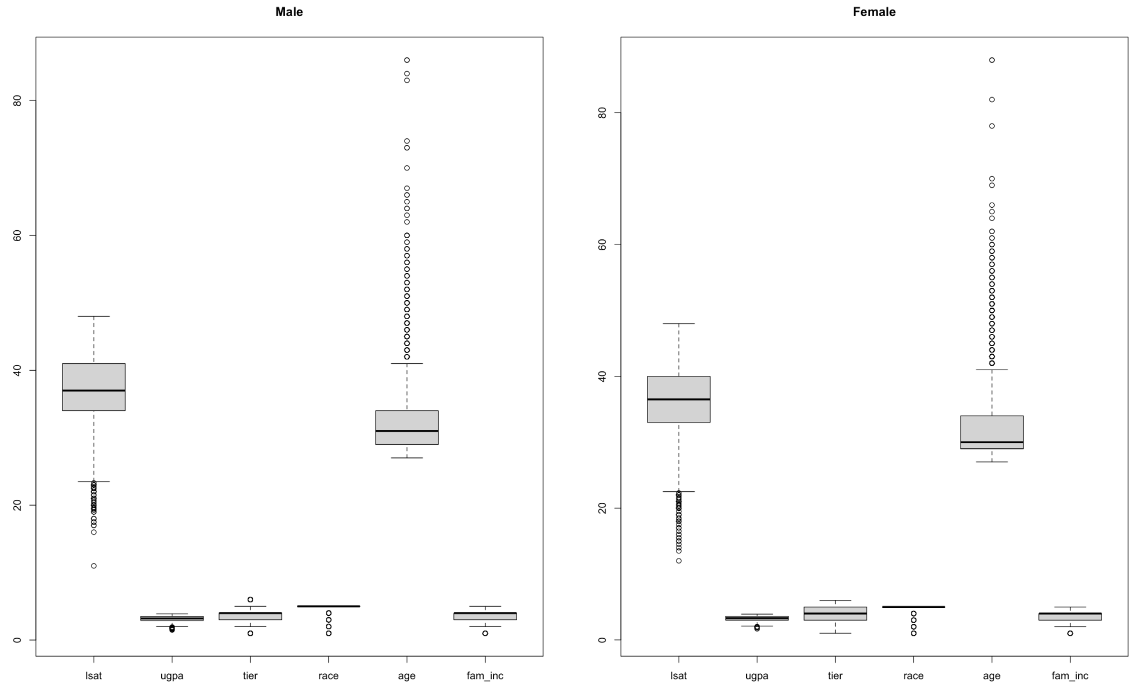
[6] United Negro College Fund. “K-12 Disparity Facts and Statistics”. UNCF, 2020.
https://uncf.org/pages/k-12-disparity-facts-and-stats

[7] Kowarski, Ilana. “5 Traits that Help People Get into Top Law Schools”. U.S News, 2018.
https://www.usnews.com/education/best-graduate-schools/top-law-schools/articles/2018-04-23/5-traits-that-help-people-get-into-top-law-schools

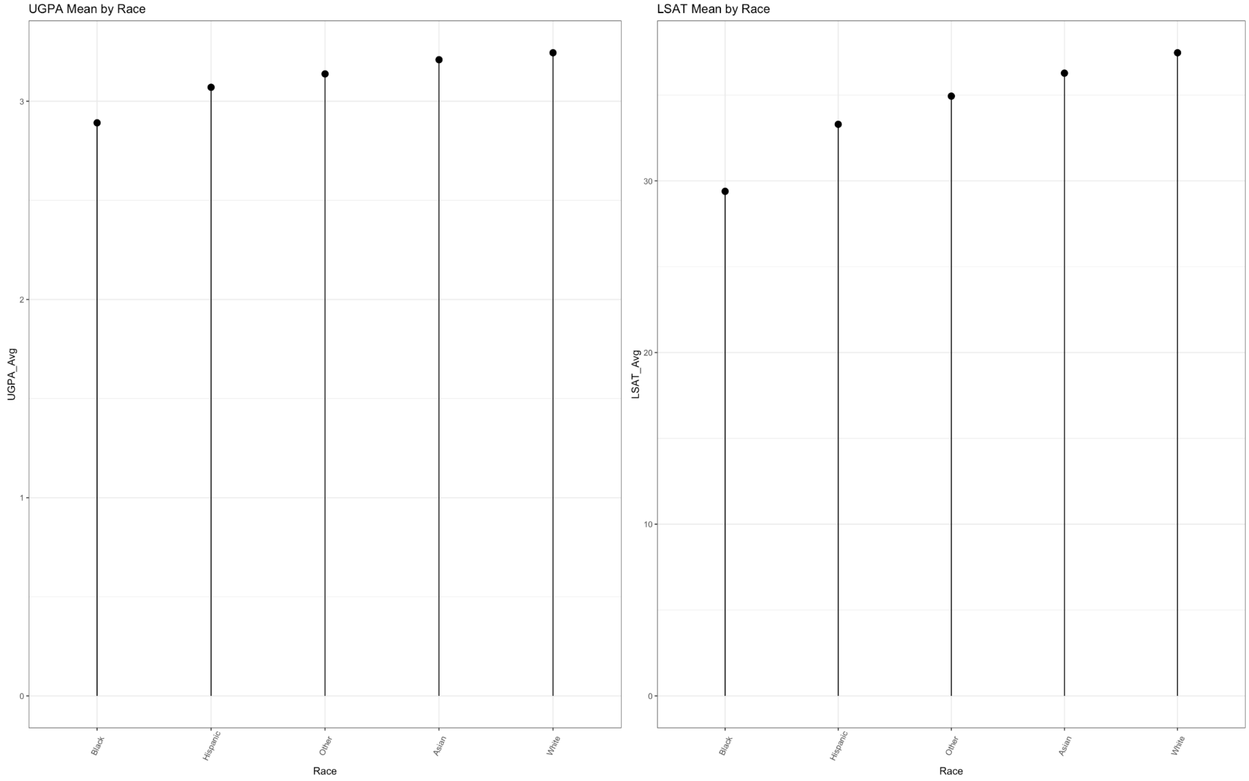
[8] Swauger, Shea. “The next normal: Algorithms will take over college, from admissions to advising”. Washington Post, 2021.
https://www.washingtonpost.com/outlook/next-normal-algorithms-college/2021/11/12/366fe8dc-4264-11ec-a3aa-0255edc02eb7_story.html

Appendix

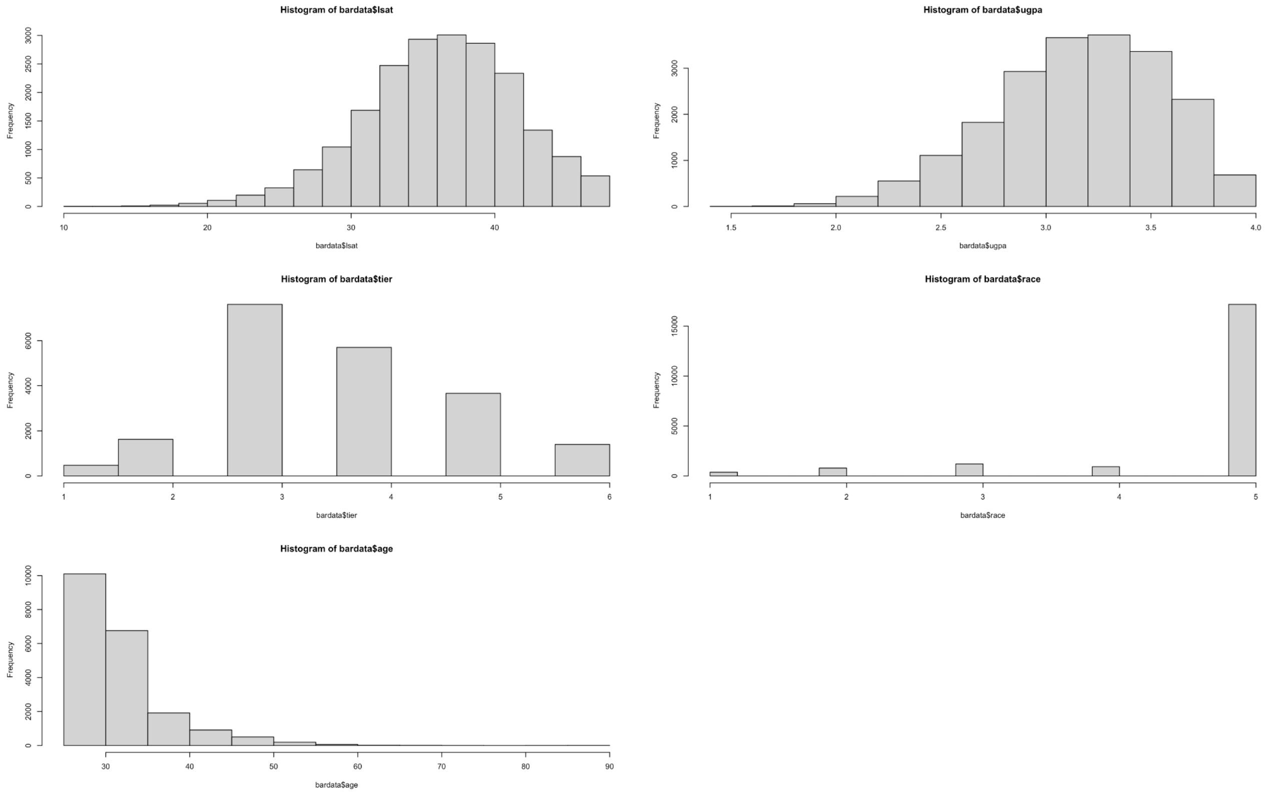
[**Figure 1**] Male versus Female review -
From left to right: LSAT, UGPA, Tier, Race, Age, Fam_Inc



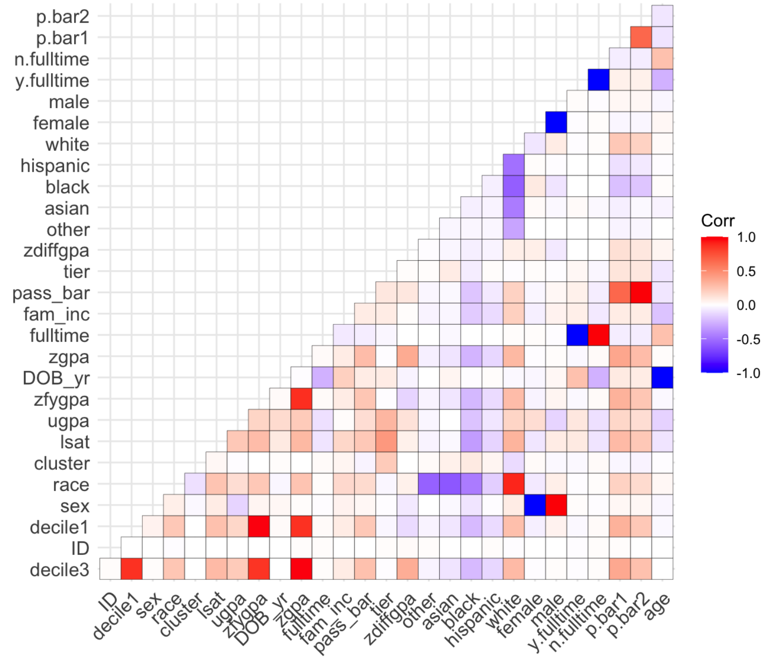
[**Figure 2**] Average UGPA and LSAT by Race



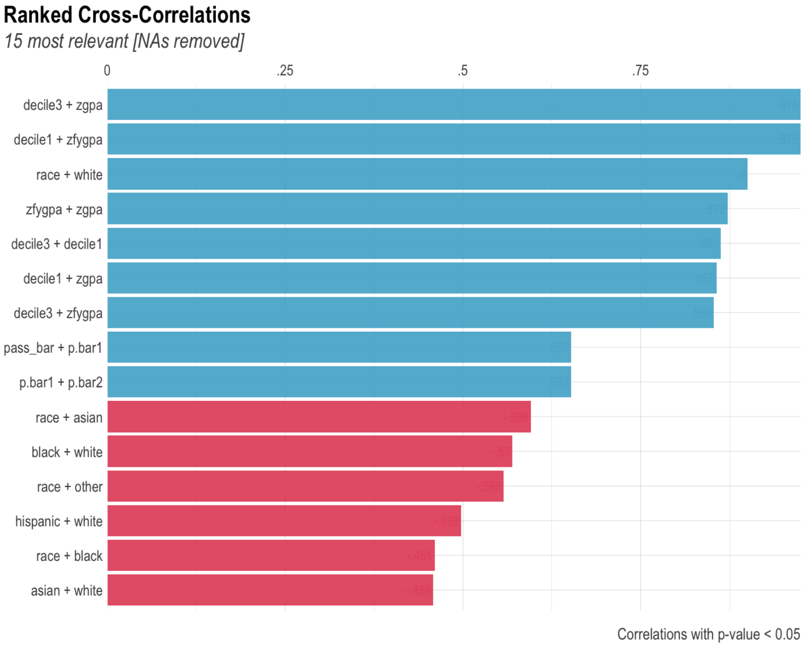
[**Figure 3**] Histograms - from left to right by row: LSAT, UGPA, Tier, Race, Age



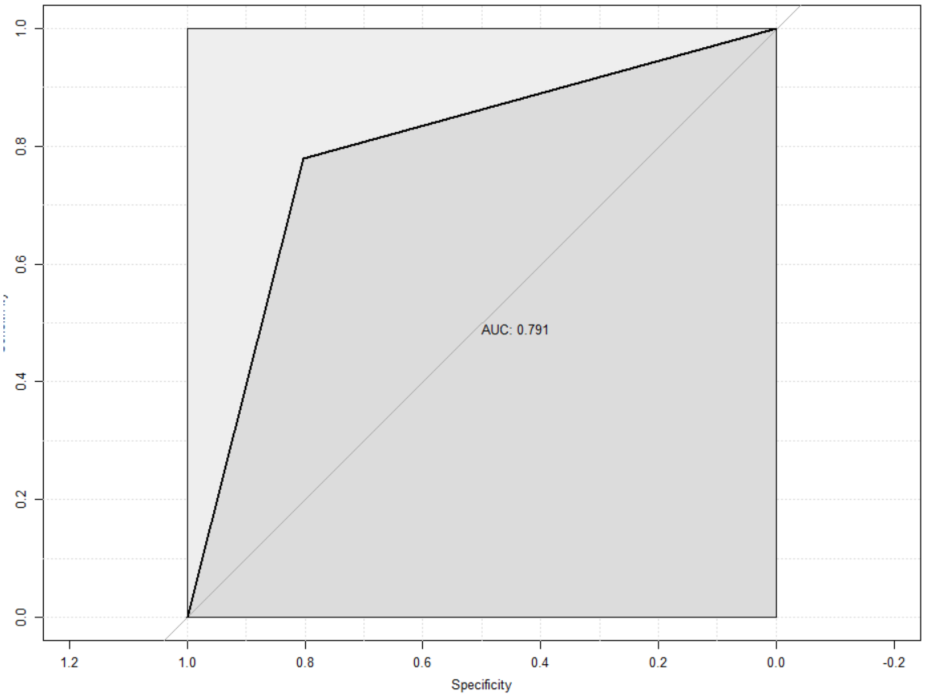
[**Figure 4**] Correlation Plot



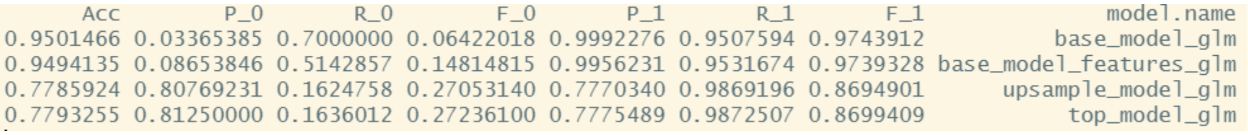
[**Figure 5**] Top correlations - Listed



[Figure 6] - ROC Curve for Final Logistic Regression with Backward Selection Model



[**Figure 7**] Model Results - Logistic Regression and Backward Selection

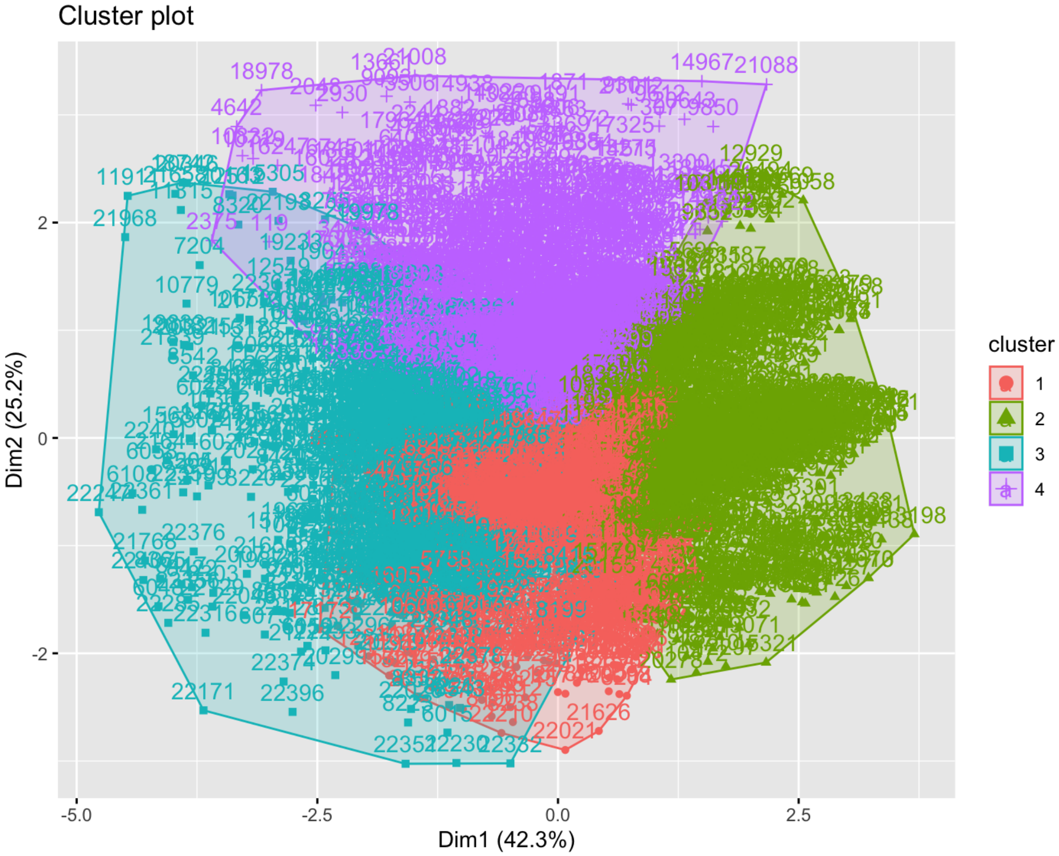


```
      Acc          P_0       R_0         F_0       P_1       R_1       F_1                model.name
0.9501466 0.03365385 0.7000000 0.06422018 0.9992276 0.9507594 0.9743912            base_model_glm
0.9494135 0.08653846 0.5142857 0.14814815 0.9956231 0.9531674 0.9739328 base_model_features_glm
0.7785924 0.80769231 0.1624758 0.27053140 0.7770340 0.9869196 0.8694901        upsample_model_glm
0.7793255 0.81250000 0.1636012 0.27236100 0.7775489 0.9872507 0.8699409             top_model_glm
```

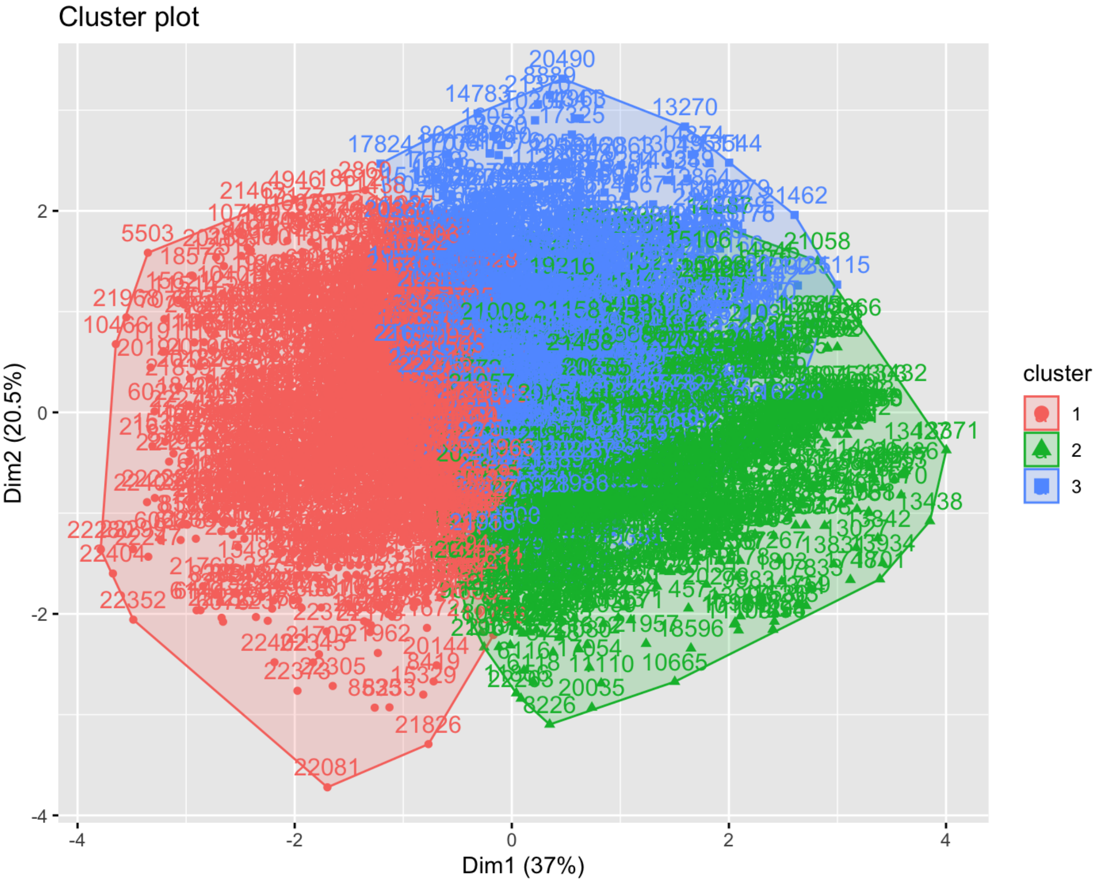
Description of Model Result Fields

Acc: Overall Accuracy of the Model
P_0: Precision of Class 0
R_0: Recall of Class 0
F_0: F Score of Class 0
P_1: Precision of Class 1
R_1: Recall of Class 1
F_1: F Score of Class 1

[**Figure 8**] 4 Clusters - 5,000 random observations from dataset



[**Figure 9**] 3 Clusters - white students removed



Team Contribution Statement

- **Morgan:** Data Understanding (writing), Data Preparation (writing), Modeling: Unsupervised Learning with Clustering (writing and code), Visualizations (writing and code)
- **Samantha:** Introduction (writing), Business Understanding (writing), Data Understanding (writing and editing), Modeling: Logistic Regression and Backward Selection (writing and code), editing of final document
- **Micaiah:** Modeling: Threshold Optimization with Logistic Regression (writing and code), Modeling: Focal Loss (writing and code), Modeling: Fine-Tuning XGBoost with and without Resampling (writing and code), Modeling: K-Nearest Neighbors with Feature Importance (writing and code), Modeling: Nearest Centroid (writing and code), editing of final document
- **Danielle:** Evaluation (writing), Deployment (writing), Descriptive Analytics (code), Modeling: Decision Trees with Boosting (writing and code)