

Predicting High School Alumni Success

ISYE/MGT 6748 OAN Fall 2020

Presentation by KIPP - Team A
Rajeev Ranjan Singh and Pragyan Shrivastava

Table of Contents

PREDICTING HIGH SCHOOL ALUMNI SUCCESS	1
ISYE/MGT 6748 OAN FALL 2020.....	1
1 EXECUTIVE SUMMARY.....	3
2 PROBLEM STATEMENT / OBJECTIVE.....	3
3 SURVEY	3
4 PROPOSED METHOD	3
4.1 SOLUTION APPROACH	3
5 EXPERIMENTS / EVALUATION	4
5.1 DATA SOURCES AND FEATURES.....	4
5.2 DATA TRANSFORMATION AND CORRELATION	4
5.3 EXPLORATORY DATA ANALYSIS.....	5
5.4 ACT DATA ANALYSIS	6
5.5 CORRELATION MATRIX	10
5.6 DATA MODELLING AND EVALUATION	11
5.7 FEATURE ANALYSIS.....	14
6 CONCLUSION AND DISCUSSION	15
7 WORKS CITED	16

1 Executive Summary

KIPP Metro Atlanta Schools (MAS) is a network of 11 schools located in 7 campuses in the city of Atlanta, Georgia. KIPP schools are open-enrollment public charter schools, serving over 5000 students from educationally underserved communities. KIPP's vision is "Every child grows up free to create the future they want for themselves and their communities." [1]

Over the years, KIPP schools have performed better than the Atlanta Public Schools in terms of Georgia Milestones [2]. A success in high school may be defined as how well the students are prepared for college. As part of this project, the team wanted to understand which students are succeeding at KIPP and how should we define success for a group of students. The team wanted to focus on various demographic, school performance indicators and attendance information impact success. This way we could identify the driving factors so that school can focus its resources optimally on areas that make most positive impact.

The team found the data quantity was insufficient and somewhat unbalanced to drive towards reliable conclusions in establishing trends and identifying clear indicators.

2 Problem Statement / Objective

We would explore high school data, demographic, enrollment, ACT score, absence data and alumni success data. The team defined a successful high school experience as one where a student graduates to attend a four-year university. We will explore how well prepared KIPP students are prepared for life after high school and will predict success and failure for KIPP students.

3 Survey

While there are many data-centric analytics and research papers available that measure student success in colleges, there is not enough material available that can objectively identify criteria that drive a success of student after high schools. This project aims to combine robust machine learning algorithms with the unique feature set provided by KIPP to draw insights on student performance through high school and predict success after high school.

4 Proposed Method

4.1 Solution Approach

Looking at the problem statement and after understanding the current business processes, the following stepped approach towards a smarter site selection is proposed:

1. We began by using the demographic data for high school students and joining it with the alumni student success data.
2. Used enrollment dataset to find how many years a student has been enrolled at KIPP and created a new field for year enrolled.
3. We utilized the ACT dataset and created a new categorical variable based on composite score to create 5 class: failure, poor, satisfactory, good and excellent. We used the mean and 1 standard deviation range for 'satisfactory'.

Excellent	<i>greater than mean + 2 s.d.</i>
Good	<i>between mean + 1 s.d. and mean + 2 s.d.</i>
Satisfactory	<i>mean \pm 1 standard deviation</i>
Poor	<i>between mean - 1 s.d. and mean - 2 s.d.</i>
Failure	<i>less than mean - 2 s.d.</i>

4. We used absence dataset by calculating percentage for attendance and used this as a feature.
5. At last, we have used enrollment status from alumni status dataset and created two classes by mapping Graduating and Graduated to success and everything else as failure.
6. We have run correlation for these features and observe the correlation. After reviewing it, we dropped the newly created categorical variable in step3 from models.
7. We modeled the data with linear, logistic, and svc ,adaboost, xgboost classifiers to predict the chance of success of a student at KIPP.

5 Experiments / Evaluation

5.1 Data Sources and Features

1. ACT Scores: Composite and subject ACT scores and practice tests
2. Alumni Success: College matriculation, military, etc.
3. AP Scores: Scores are 1-5, exams are given in a wide variety of subject areas
4. Attendance: Enrollment and attendance data
5. MAP Scores: Measures and projects growth across the year
 - a. KIPP-specific: Typical growth vs Tiered growth

5.2 Data Transformation and Correlation

We completed the exploratory data analysis in the following steps:

1. Started with High School Demographic information
2. Then, combined the Attendance Data
3. Picked up Student Success Data
4. Combined Demographic and Success data

5. Calculated the number of years at KIPP
6. ACT Score Data – transposed the columns and dropped columns with null values
7. Eventually, combined all AP Exam Data

As we made progress on this path, we drew charts and visualizations to ensure we are not missing any data anomaly or finding outliers. One of the biggest challenges we faced in training our models was that the “Alumni Success” data provided by KIPP only tracked 855 student’s performance after high school. While we had features for almost 2100 high school students, limitations in the result of their high school experience restricted our ability to draw conclusions.

The years enrolled at KIPP data, calculated by tracking a student through all the KIPP schools he or she attended, was especially useful in our analysis. Research published at the University of North Carolina has found that “Students who repeat a grade prior to high school have a higher risk of dropping out of high school than do students who are continuously promoted.” (Stearns, 2007)

5.3 Exploratory Data Analysis

We began by plotting the occurrence various independent features to determine if there could be any factors not provided in the dataset that could play a significant role on student success at KIPP.

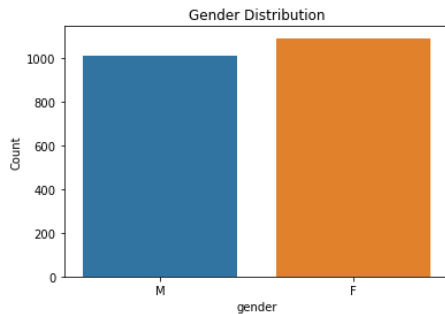


Figure I

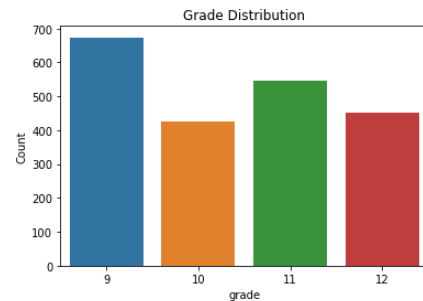


Figure II

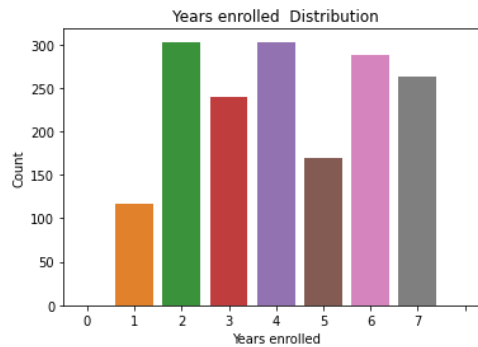


Figure III

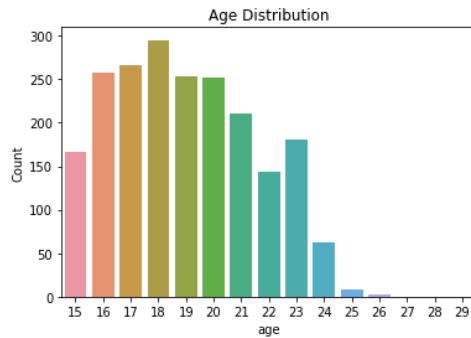


Figure IV

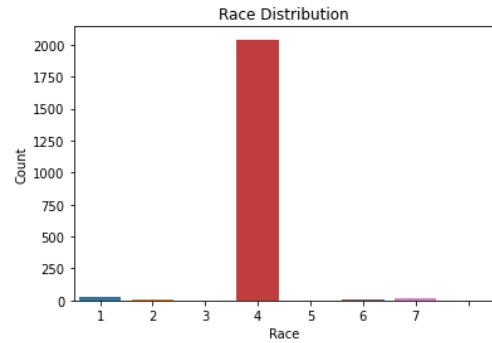
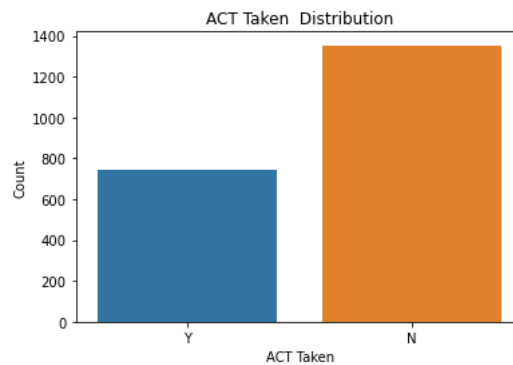


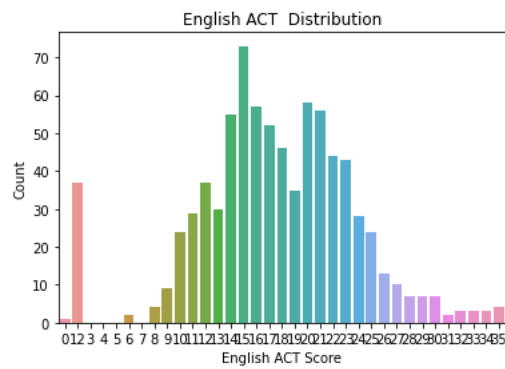
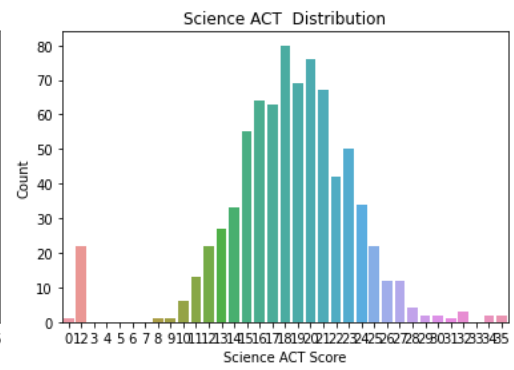
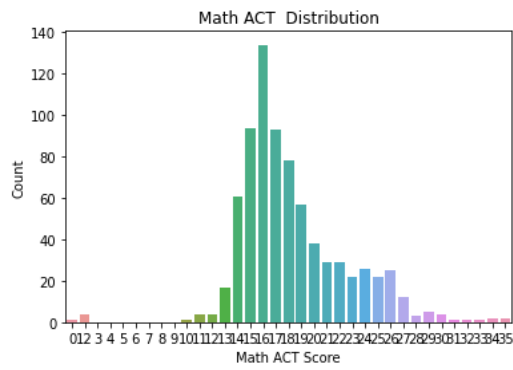
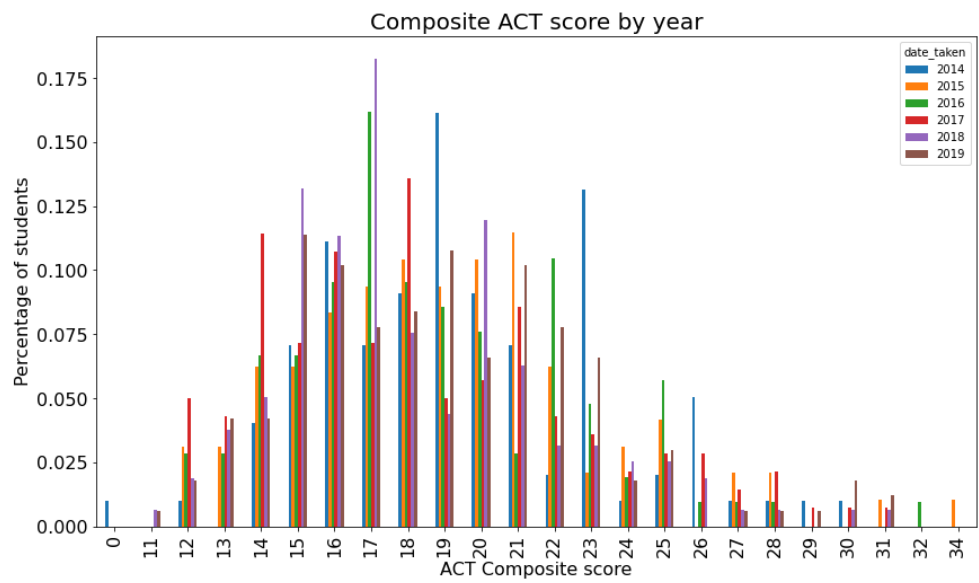
Figure V

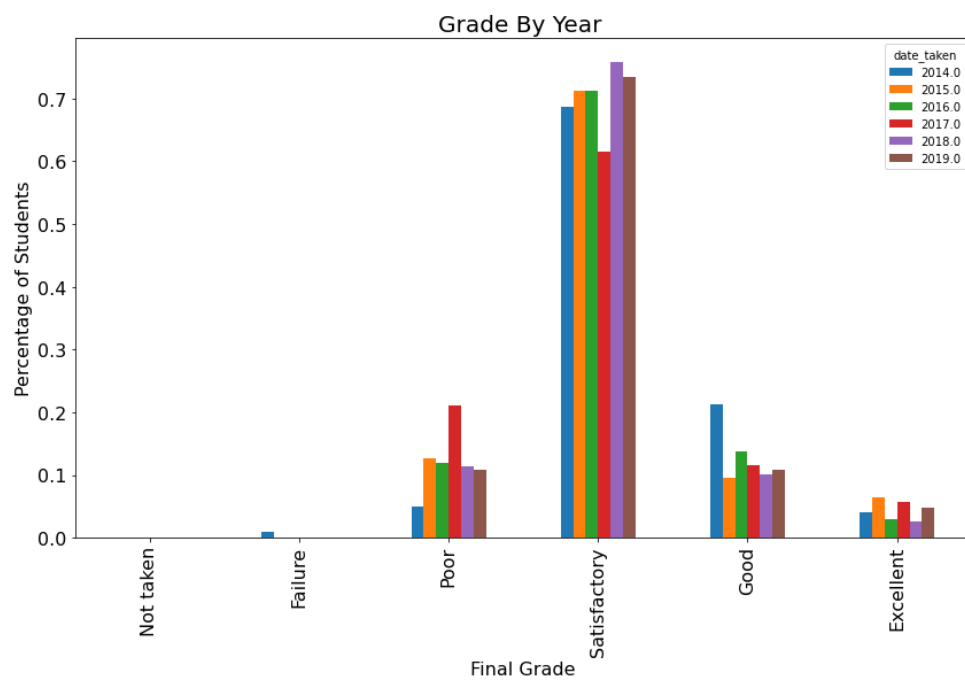
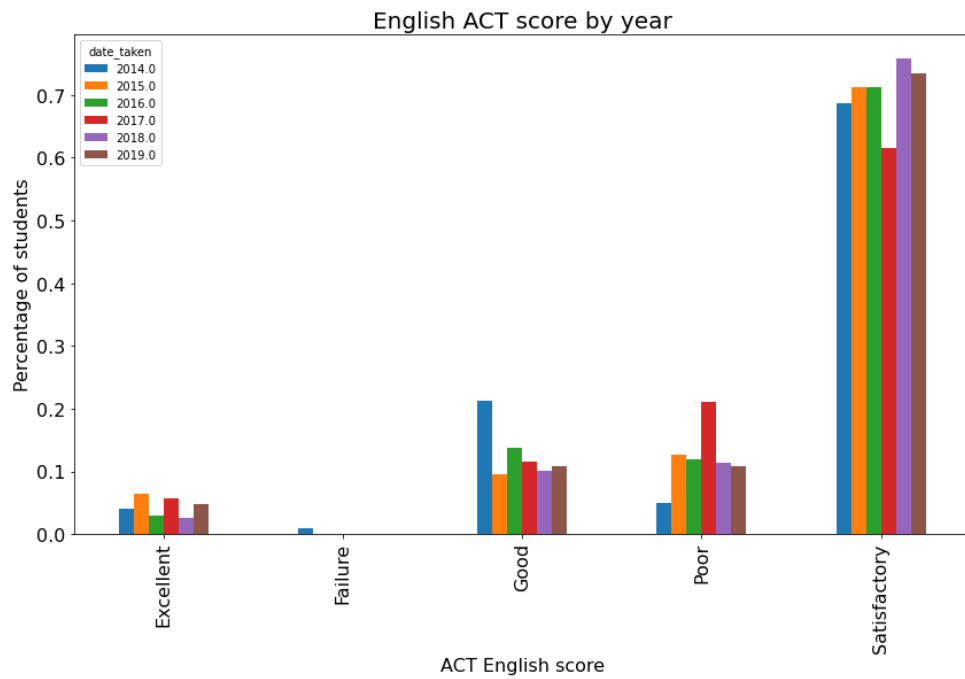
As seen in figures I and II, we have a fairly even distribution in gender and grade level for all high school students provided. Figure V tells us that the majority of students enrolled at KIPP schools are Federal Ethnic Codes 4 and 1, corresponding to African American and Latino students.

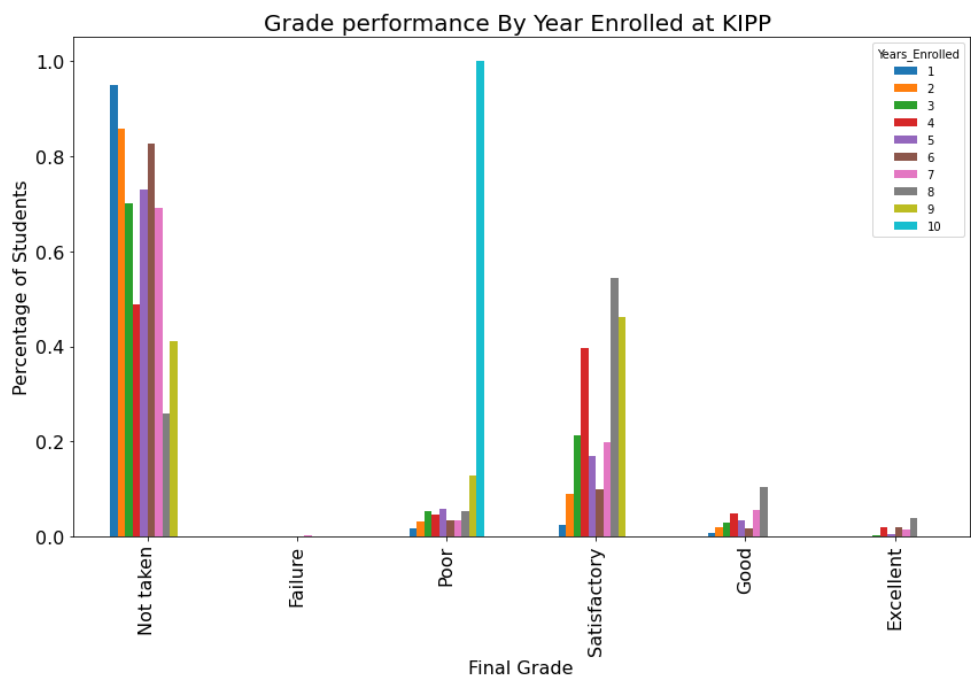
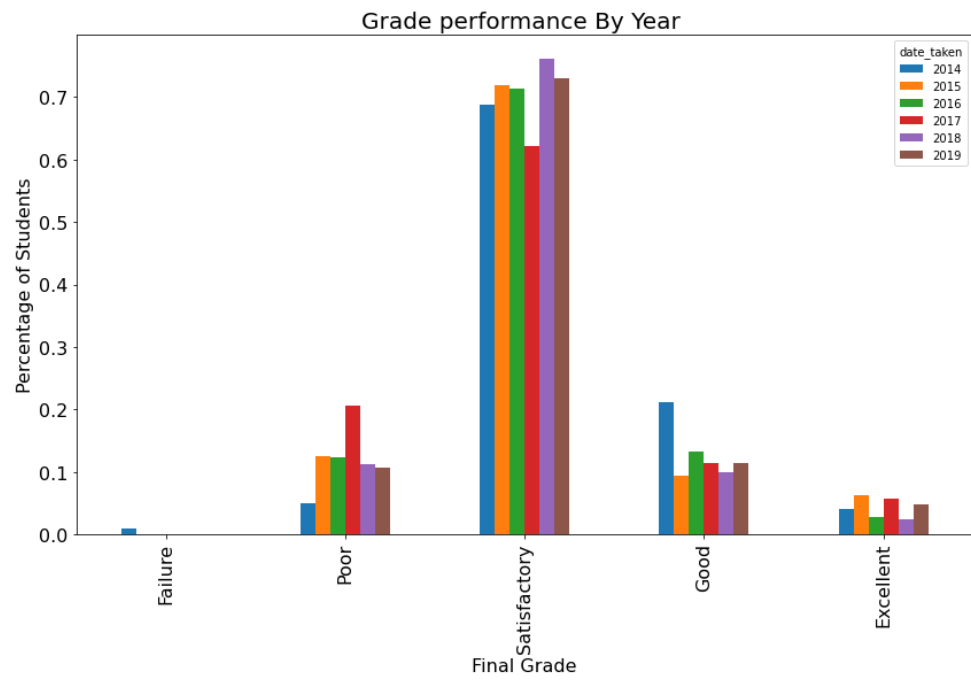
5.4 ACT Data Analysis

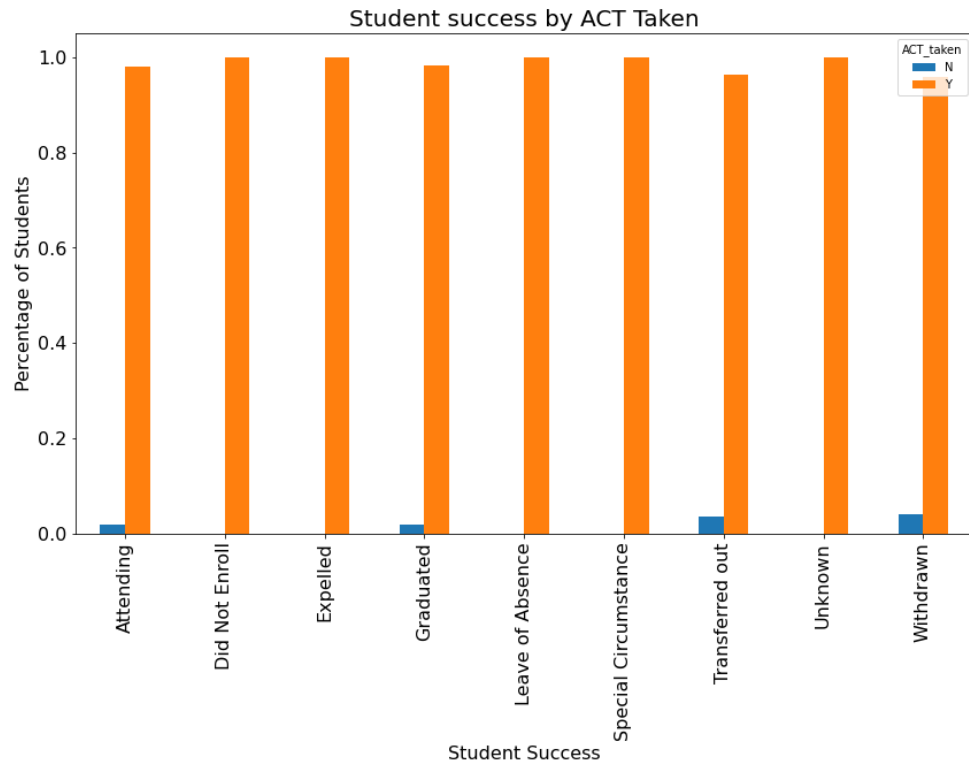
The American College Test (ACT) is a standardized examination taken by many students at KIPP Schools. Below are some of our graphs analyzing the distribution of ACT scores, along with the impacts of certain factors on ACT performance.











5.5 Correlation Matrix

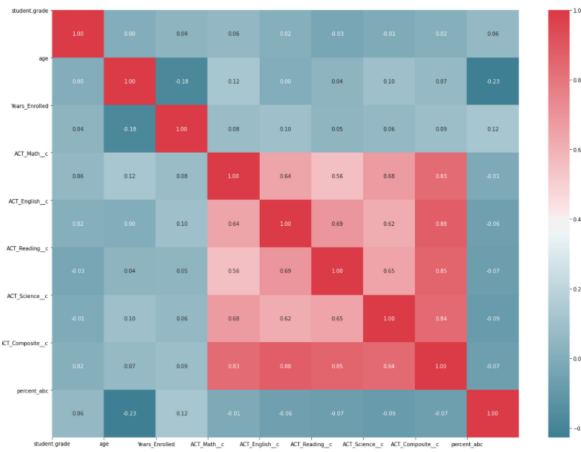


Figure VI

Above correlation plot shows correlation between different type of ACT score but stronger co-relation between composite score and all other scores. Based on this observation we decided to drop the newly added categorical variable for performance grade. However, we kept the composite score in final model.

5.6 Data Modelling and Evaluation

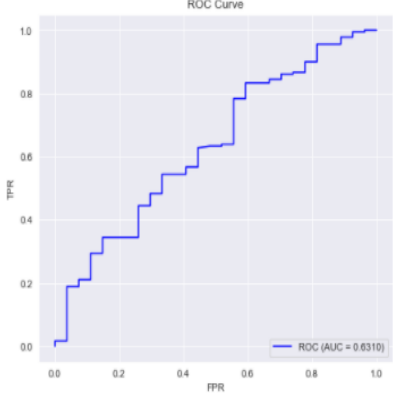
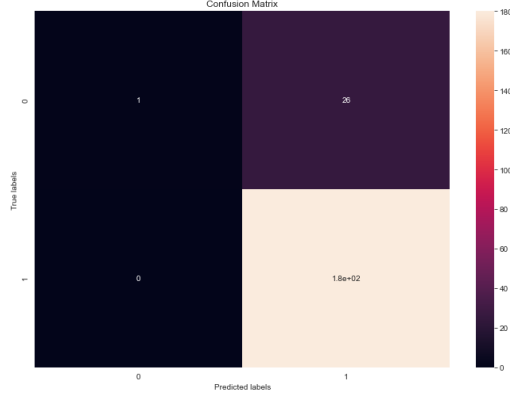
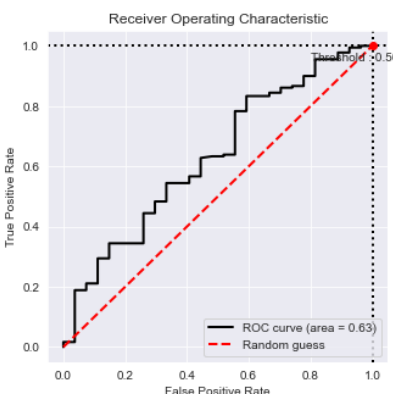
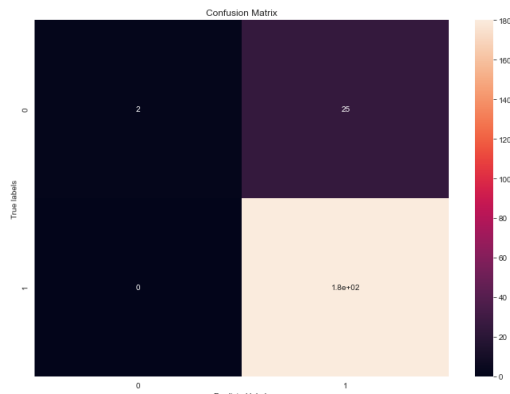
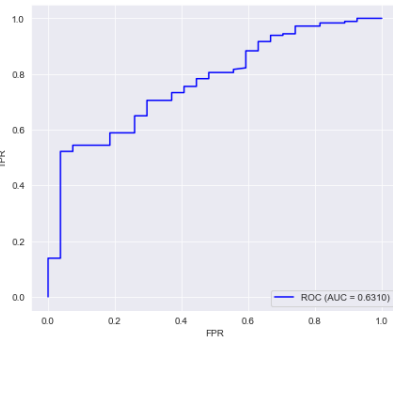
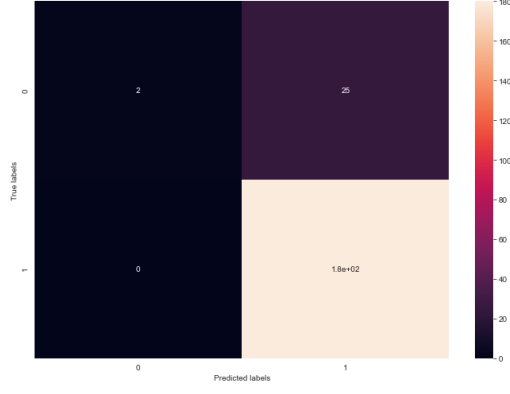
This table represents accuracies of all models tested

Models	Logistic Regression Model	Binary Classification	Random Forest	SVM Model	Decision Tree Model	ADABOOST Model	XGBoost Model
Accuracy Score	.87	.87	.88	.87	.77	.87	.89

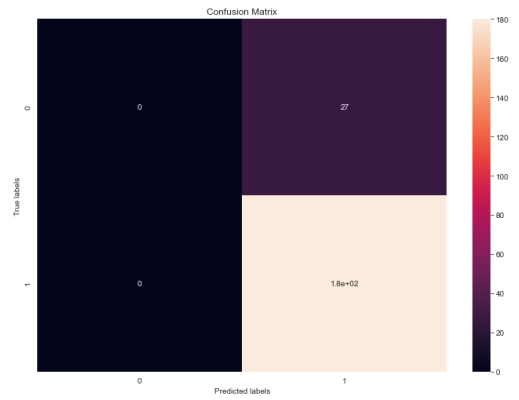
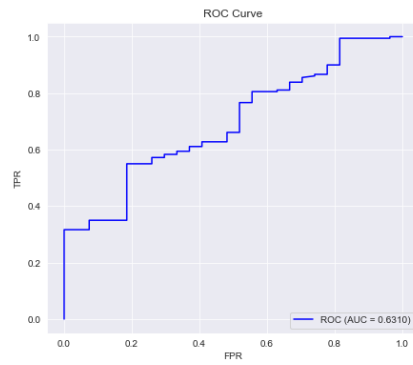
We experimented with all the feature extracted in data analysis section, such as age, gender, grade, all ACT scores, newly created categorical variable from ACT composite score, Years enrolled at KIPP . This resulted in accuracy ~87% for most models. Next we used student absence data and calculated the average percentages of student's attendance and used it as feature and removed the categorical variable that we created from ACT composite score. Now this resulted in slightly better performance by xgboost model. Also feature engineering plot did indicate that gender and grade contribute less than other variables toward explaining the variability in the model.

So as per observations over the above models , xg-boost outperformed all other models. As well as this, the model is more generalized based on our observation training test score which has small difference in their prediction.

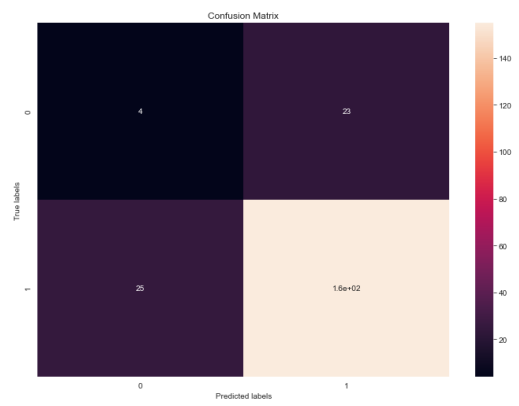
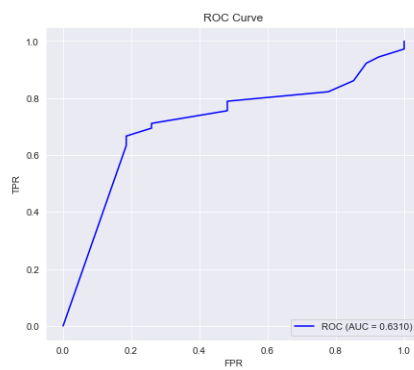
Attached below is a table containing the models we trained upon the KIPP dataset, along with their Receiver Operating Characteristic (ROC) curve. We used these charts to determine the highest performing model.

Model Type	Receiver Operating Characteristic Graph	Confusion Matrix
Logistic Regression Model	 <p>ROC Curve</p> <p>ROC (AUC = 0.6310)</p>	 <p>Confusion Matrix</p> <p>True labels: 0, 1</p> <p>Predicted labels: 0, 1</p> <p>Counts: (0,0)=1, (0,1)=20, (1,0)=0, (1,1)=1.9e+02</p>
Binary Classification	 <p>Receiver Operating Characteristic</p> <p>ROC curve (area = 0.63)</p> <p>Random guess</p>	 <p>Confusion Matrix</p> <p>True labels: 0, 1</p> <p>Predicted labels: 0, 1</p> <p>Counts: (0,0)=2, (0,1)=25, (1,0)=0, (1,1)=1.9e+02</p>
Random Forest	 <p>ROC Curve</p> <p>ROC (AUC = 0.6310)</p>	 <p>Confusion Matrix</p> <p>True labels: 0, 1</p> <p>Predicted labels: 0, 1</p> <p>Counts: (0,0)=2, (0,1)=25, (1,0)=0, (1,1)=1.9e+02</p>

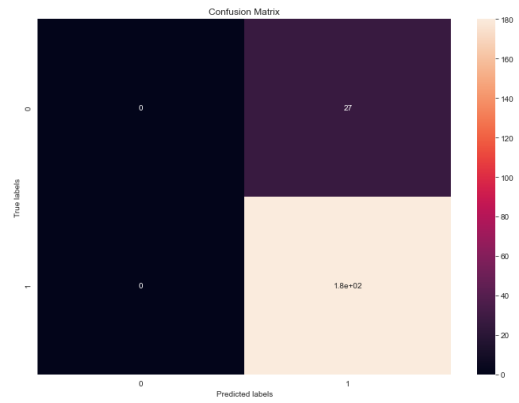
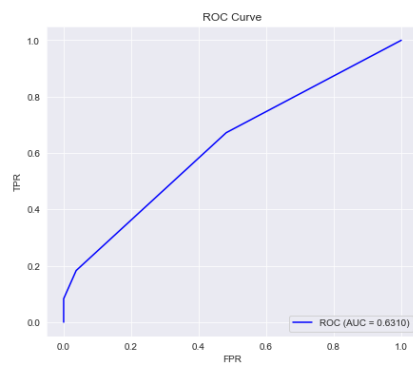
SVM Model



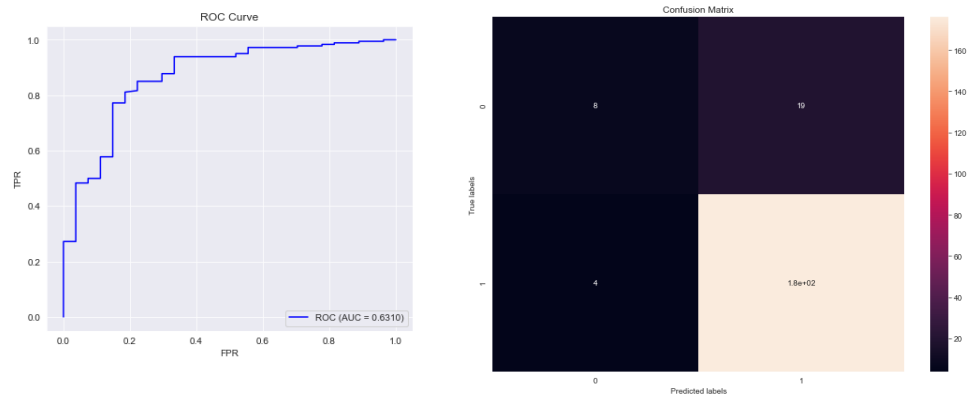
Decision Tree Model



ADABOOST Model



XGBoost Model



Classification Report for the XGBoost Model

	precision	recall	f1-score	support
Success	0.67	0.30	0.41	27
Failure	0.90	0.98	0.94	180
accuracy			0.89	207
macro avg	0.78	0.64	0.67	207
weighted avg	0.87	0.89	0.87	207

Fbeta score = 0.7250000000000001

5.7 Feature Analysis

The “plot_importance” method provided by the XGBoost library within SK-Learn was used to derive any individual features importance on the overall outcome of the model.

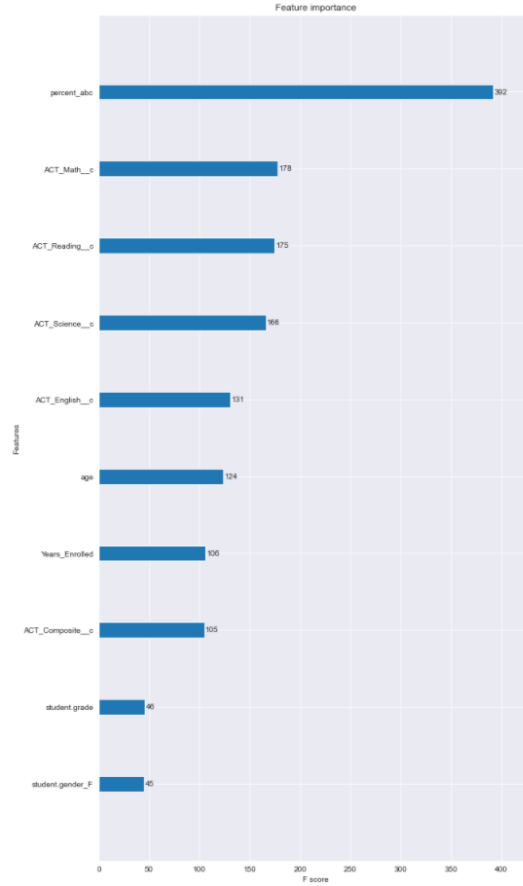


Figure VII

As seen in figure VII, we determined the most important factor in predicting if a student will graduate from a KIPP high school to be the total percentage of time they are absent from school.

6 Conclusion and Discussion

In our first iteration of model evaluation, we did not include attendance data and found around 87% accuracy. As a second step, in our second iteration we included attendance data and our model detected it as the most important feature, but that did not change the model performance. Most of the models have similar accuracy of about 87%, but results look like biased toward success. This may be because we have very limited data for student's success. We were limited to only 828 student records which is not many outcomes upon which these models can train. Two classes are also very unbalanced in the dataset. Also, there are other important factors that can play important role in predicting student success such as economic status, family background, residential location, family life behavior etc. One thing that we can improve is by utilizing AP scores and understand the pattern for students who have taken AP courses. It is also to keep in mind important aspects about the facilities, extracurriculars, and teaching staff available at the various KIPP institutions we studied. Papers

published in the journal *Sociology of Education* cite major differences in the learning environments and educational opportunities provided at suburban and urban high schools (Rumberger, 2000).

7 Works Cited

- Cooney, P. (2017, September 06). *What actually predicts college success?* Retrieved from <https://michiganfuture.org/2017/09/actually-predicts-college-success/>
- Jaschik, S. (2017, September 25). *Ninth-Grade Marks as Predictor of College Success.* Retrieved from <https://www.insidehighered.com/admissions/article/2017/09/25/study-finds-ninth-grade-marks-predict-college-enrollment-and-success>
- KIPP Metro Atlanta Schools.* (2020, 11 28). (KIPP MAS) Retrieved from <https://kippmetroatlanta.org/about/>
- KIPP Results.* (2020, 11 28). (KIPP MAS) Retrieved from <https://kippmetroatlanta.org/about/results/>
- Rosa, S. D. (2020, January 29). *www.educationdrive.com.* Retrieved from <https://www.educationdrive.com/news/high-school-gpa-5-times-more-likely-to-predict-college-success-than-act-sco/571287/>
- Rumberger, R. W. (2000). The Distribution of Dropout and Turnover Rates among Urban and Suburban High Schools. *The Sociology of Education*, 39-67.
- Schools, A. P. (2020, 11 28). *Georgia Milestones.* (Atlanta Public Schools) Retrieved from <https://www.atlantapublicschools.us/Page/48440#:~:text=The%20Georgia%20Milestones%20Assessment%20System,grades%203%20through%20high%20school.&text=Importantly%2C%20Georgia%20Milestones%20is%20designed,their%20next%20level%20of%20learning.>
- Stearns, E. (2007). Staying Back and Dropping Out: The Relationship Between Grade Retention and School Dropout. *Sociology of Education*, 210-240.