Predict Collegiate Basketball Profit

Standard Deviants - Ava Exelbirt, Yura Heo, Claire Li

2023-11-07

Introduction and data

In our research project, we aim to investigate and understand the factors influencing the profit generated by collegiate sport, specifically basketball, in the 2019 school year. We will be utilizing a dataset sourced from Tidy Tuesday, which provides a comprehensive collection of observations related to collegiate sports describing the sport players and the schools' financial investment in sports programs in the dataset called "sports.csv". Our primary motivation for this research is to gain insights into the determinants of collegiate basketball profit, which can be of significant interest to educational institutions and further local policymakers and sports enthusiasts when generating profit, making decisions, and organizing investment strategies in sports.

Our primary research question is as follows: **In the 2019 school year, how can we predict the profit (revenue - expenditure) in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, total count of students, percent of expenditures towards women's sports, and school classification name**.

Our hypothesis is as follows: **Participation rate, sector name, and gender ratio will be the most influential predictors for the total profit generated by the collegiate sport basketball in USD.**

We anticipated that participation rate would be a key predictor of total profit of the sport basketball because schools may allocate more money to this sport if there is more participation from the students. We also expected that the type of school (sector name) as well as school classification name would be a strong predictor of profit because different types of schools have varying levels of resources, alumni support, and participation rates in sports depending on school size and program. Next, we thought the total count of students would be relevant to predicting the profit because the size of the student body could affect the potential fan base, impacting the overall interest in collegiate basketball. Lastly, we included predictor variables that relate to gender such as gender ratio and percent expenditures towards women's sports because we assumed that sports could be hugely dependent on the demographics of players.

The data set was taken from TidyTuesday and was originally scraped from Equity in Athletics Data Analysis (EADA), a sector of the US Department of Education. The data is available on an online database found on the (EADA) website[1].

This data is submitted annually from colleges to the EADA. All co-educational postsecondary institutions that receive Title IV funding that have intercollegiate athletics programs are required by the Equity in Athletics Disclosure Act to submit this data[2]. These data are collected annually starting from 2003 to 2022, but for our specific data analysis we will only look at data taken from the school year of 2019 by cleaning the data to a new csv file which we will use for the rest of the project. The csv file from Tidy Tuesday contains thousands of observations from years 2015-2019, but we adjusted the data file to only include the year 2019 and will filter the data to only include basketball as a sport, as this is our population we are analyzing.

The dataset we will use from the data scraped on Tidy Tuesday include many observations regarding collegiate sports. The observations are both quantitative and categorical and measure characteristics of different school's spending, revenue, populations, locations, and sports. These observations include variables such as the name of the city which a school is in, the state, the school name, school classification (e.g. which NCAA Division), school sector name (e.g. 4 year accredited university), sport. Most of the quantitative variables are split between men and women. For example, there is total male population and total female population. There are also observations for the total amount for each of these above variables which includes both men and women; for example, total expenditures for both men and women.

The key variables we will use are **participation rate** which is the total percentage of men and women students who participate in sports, **sector school name** which is the type of school for example, public, 4-year or above, **gender ratio** which is the male population count divided by female population count, **total count of students** which is the total amount of students enrolled in the college, **women expenditure percent** which is the percentage of expenditures that goes towards woman's sports (which we calculated by taking a school's expenditures of women and dividing it by the school's total expenditures), **school classification name** which is a school's sports classification for example NCAA Division I-FCS. Then, our response variable is **profit** which is calculated by the total revenue of the school for basketball minus the expenditures of the school for basketball.

For data cleaning, we filtered for observations that have `Basketball` as the sports. A number of schools do not have a basketball team and show missing values in predictor variables we need in the model like participation rate, revenue, and expenditure. Therefore, we filtered out any school that does not have a basketball team, mostly public, 2-year schools. Additionally, We created new variables by mutation to turn variables involving revenue and expenditure into the unit of millions and total student count into the unit of hundreds, so that we can get

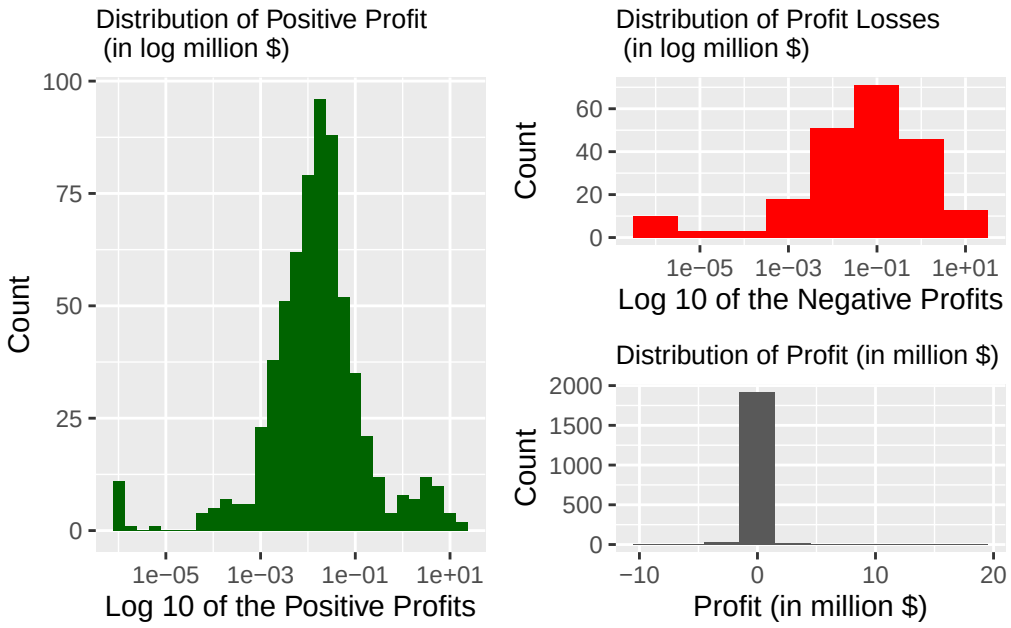
[1] https://ope.ed.gov/athletics/#/datafile/list

[2] https://knightnewhousedata.org/about-the-data#:~:text=The%20data%20is%20available%20via,student%20aid%20programs)

n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
57	0.0273844	0.91862	-9.67379	0	0	0.004166	16.89384

larger coefficients for better modeling. Lastly, we created the new response variable of profit and new predictor variable of percentage women expenditure.

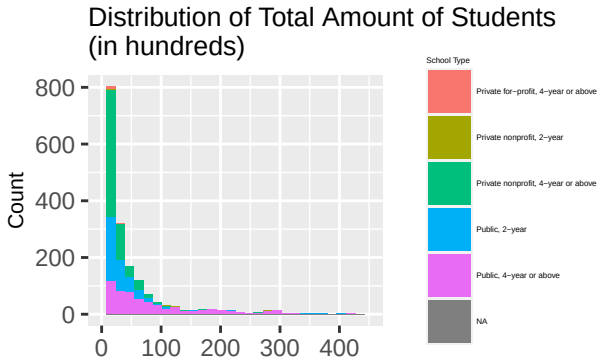
Distribution of the response variable: profit generated by college basketball

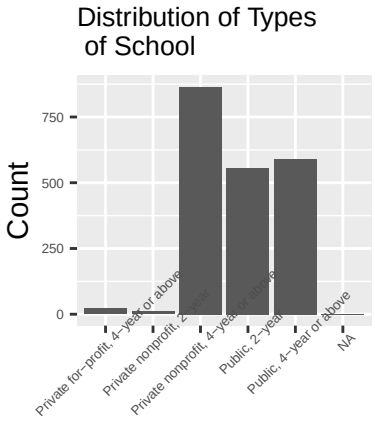


The distribution of profit is approximately normally distributed with a slight right skew. However, many observations centered around 0$ in profit. The median of the data is at 0$, and the mean is about 0.027 million USD. Since the response is approximately normal, we can use the mean as the center of the data. The range is from -9.6738 million USD to 16.8938 million USD. We also split the distribution into positive and negative profits and took the log of the x axis to see a more detailed distribution. It seems that positive profits are normally distributed while negative profits are left skewed. There are 57 missing values of profit.

Distributions of total amount of students (potential quantitative predictor variable) and type of school (potential categorical predictor variable)

n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
0	41.73254	59.4928	0.5	10.6025	20.055	46.33	662.79



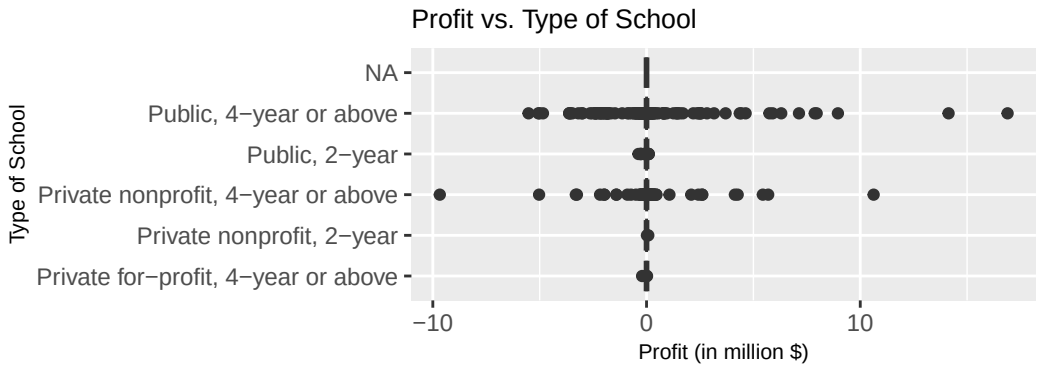


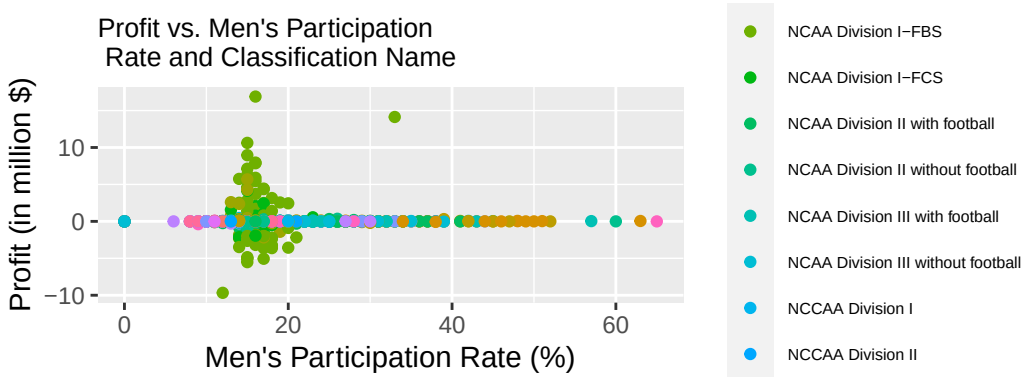
	n	p	nmiss
Private for-profit, 4-year or above	22	0.011	1
Private nonprofit, 2-year	11	0.005	NA
Private nonprofit, 4-year or above	865	0.423	NA
Public, 2-year	556	0.272	NA
Public, 4-year or above	591	0.289	NA

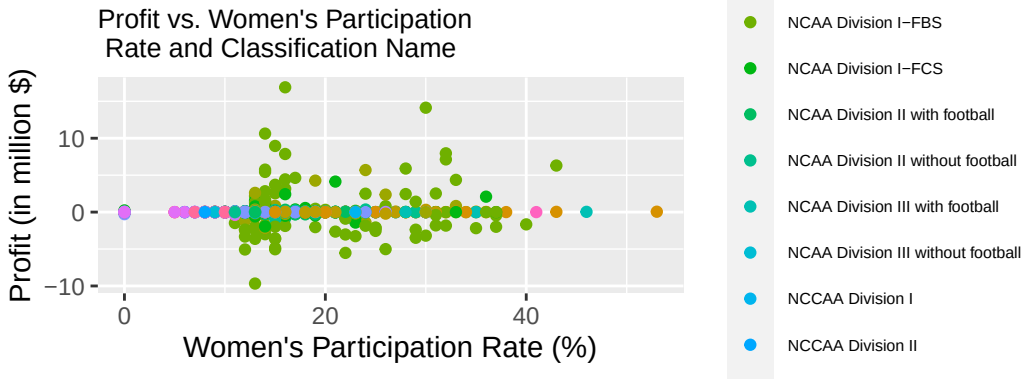
The distribution of total amount of students on college sports is right-skewed with most of the amount of student values in the lower range, while a number of observations have very high values that make them outliers. Given the apparent skewness, the center is the median of 2,005.5 students. Since the distribution is skewed, the IQR is used as a more reliable measure of spread which is Q3 - Q1 = 4,633 - 1,060.25 = 3,572.75 students. You can also see that most of the schools with very high number of students are public, 4 year or above colleges.

There are 5 types of schools. Private for-profit, 2-year and private nonprofit, 2-year have low numbers of observations. Private nonprofit, 4-year or above, public, 2 year, and public, 4-year or above have comparably higher number of observations than the other two, with private nonprofit, 2-year having the highest number of observations. There is one missing value in types of school.

Relationship between Profit and a Categorical and Quantitative Predictor





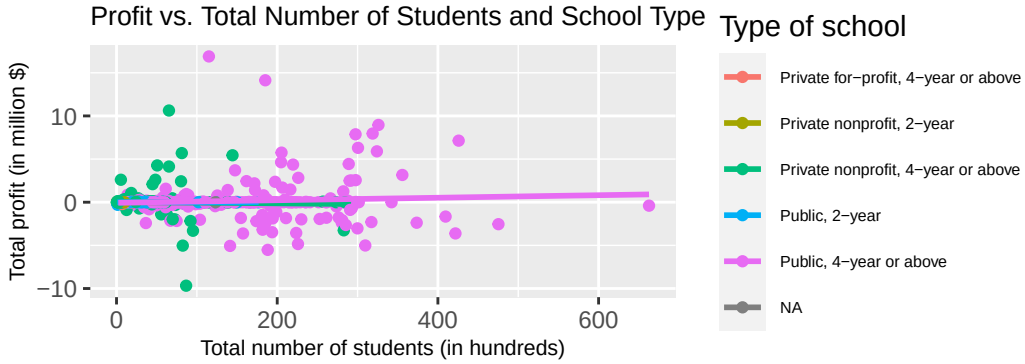


Looking at the relationships above, it seems that the medians of each type of school's profit is around $0. We can also see that Public, 4 year or above colleges have the highest recorded profit in USD, and that Private nonprofit, 4 year or above colleges have the lowest profit in USD.

Looking at the participation rate plots, there seems to be slightly more participation of women than of men, as many of the data points are more spread out to the right of the graph for women while there is a conglomerate of data points centered around 20% for men's participation rate. There seems to be a very weak, almost negative, linear relationship between participation rate

versus profit, as most of the data points are centered around $0 profit regardless of participation rate. However, looking at the men's participation rate specifically, it seems schools classified with sports as NCAA Division I-FBS have the highest profits and highest range of profits. Schools classified as NAIA Division II and NCAA Division II without football seem to have high men's participation rates. For women's participation rate, it shows the same pattern as men's, However, there tends to be greater participation rate of women than of men in NCAA Division I-FBS schools.

A potential interaction effect between total number of student and type of school



The lines are not parallel indicating there is an interaction effect. The slope of total number of student differs based on the type of school.

Methodology

We planned to use a multiple linear regression to predict the profit in USD of the collegiate sport basketball using participation rate, school sector name, gender ratio, percent women expenditure, total count of students, and school classification name. For the predictor gender ratio, we mutated the data and divide the total male student count by total female student count. For the predictor percentage of women expenditure, we mutated a variable that divides a school's woman's expenditure by total expenditure.

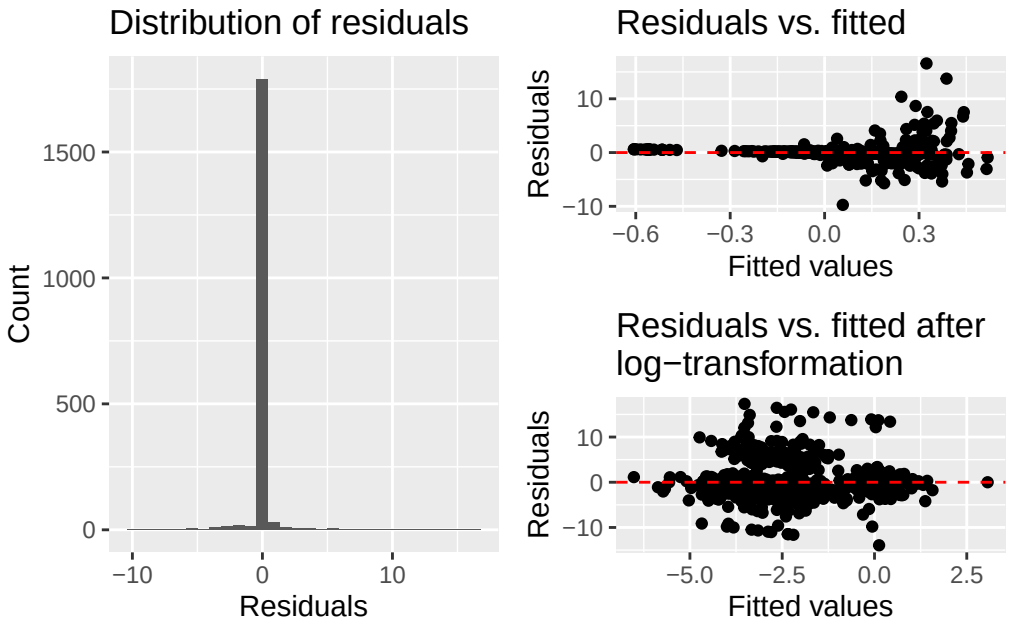
For the response variable, we predicted the profit instead of the revenue variable or the expenditure variable because we anticipated that they will have strong correlations. As schools allocate significant financial resources to their sports programs, the expectation is that these expenditures will have a direct impact on the overall revenue generated through different resources such as ticket sales.

We anticipated interaction effects. Specifically, we believe the participation rate and the school sector type may have a correlation, since the participation rate to be very dependent on the levels of investment, competitive levels, institutional culture, and student demographics which vary based on the type of institution(sectorname).

To avoid overfitting the data, we then split the dataset into training and testing data, and we used cross validation of 20 folds on the traning data to fit models using our recipe. We calculated AIC and BIC to evaluate each model fit and selected our final model.

Checking Model Conditions:

We first checked the conditions for inference with the full model including all available predictors on the full data.



Linearity is not quiet met as there seems to be a negative linear relationship between the residuals and fitted values as there is a conglomerate around the y intercept. There does not seem to be such a random scatter of points. Constant variance is met because there is a relatively even amount of residuals above and below the horizontal line. The normality condition is satisfied as the residuals are normally distributed and the sample has more than 30 observations. Independence is also satisfied because based on how the data was taken, collegiate information about one school is independent of another school.

Since linearity is not satisfied, we log transformed the response variable `profit` where we can see the residuals vs. fitted values show a random scatter and linearity is satisfied. Hence, we use `log_profit` as the response variable for fitting models.

To prepare the variables for our analysis, we used a recipe. First, we dropped all the NA values using `step_naomit()` to ensure that the values we do use are accurate. For example, the data set has observations with NA values in women's participation rate. However, when the sum participation rate for women is calculated, the data set represents this sum with 0. This 0 can skew the regression model when the true participation rate of women may not

be 0. Since we are using total participation rate in our model, it is best to drop NA values. Also, considering that the data set has 775 observations after all NA's are dropped, we still have a lot of reliable data to use for modeling. Most of the school's with missing data were school's without a basketball team. However, there are potentially some school's that did have a basketball team and had other NA values that got dropped that we will analyze more in discussion section. Then, we used `step_zv()` to remove all predictors that contain only a single value. We created dummy variables using `step_dummy()` for all nominal predictors which are Classification name and Sector name. Lastly, we used `step_center()` to mean center our quantitative predictors.

.metric	.estimator	mean	n	std_err	.config
rmse	standard	4.14	20	0.180	Preprocessor1_Model1
rsq	standard	0.09	20	0.019	Preprocessor1_Model1

mean_adj_rsq	mean_aic	mean_bic
0.077	3443.95	3545.167

.metric	.estimator	mean	n	std_err	.config
rmse	standard	4.139	20	0.173	Preprocessor1_Model1
rsq	standard	0.089	20	0.020	Preprocessor1_Model1

mean_adj_rsq	mean_aic	mean_bic
0.072	3443.531	3527.144

.metric	.estimator	mean	n	std_err	.config
rmse	standard	4.16	20	0.155	Preprocessor1_Model1
rsq	standard	0.06	20	0.009	Preprocessor1_Model1

mean_adj_rsq	mean_aic	mean_bic
0.047	3446.948	3473.352

We compute the AIC, BIC, and adjusted R^2 values of all 3 models. Aiming for a parsimonious model, we chose the third model with the least number of terms and lowest BIC.

Our final model uses the predictor of **total participation rate, total count of students, gender ratio, and percentage of women expenditure**.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.619	0.181	-14.498	0.000
total_partc	0.065	0.032	2.071	0.039
ef_total_count	0.006	0.003	2.318	0.021
gender_r	-0.363	0.394	-0.921	0.358
perc_wom_exp	-7.178	2.013	-3.566	0.000

- Model fit statistics

.metric	.estimator	.estimate
rmse	standard	5.111739

.metric	.estimator	.estimate
rmse	standard	5.056632

AIC	BIC	adj.r.squared
3627.979	3654.691	0.0468664

The final model has a AIC of 3627.979, a BIC of 3654.691, and an adjusted R^2 of about 0.047.

To see if there are interaction terms, we used `step_interact()`. Because we took out the sector name variable, we decided to check if there were any interaction effects between gender ratio and percentage of women expenditure with these variables with both variables having a relationship to gender.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.623	0.181	-14.487	0.000
total_partc	0.064	0.032	2.040	0.042
ef_total_count	0.006	0.003	2.233	0.026
gender_r	-0.350	0.396	-0.883	0.378
perc_wom_exp	-7.676	2.399	-3.199	0.001
gender_r_x_perc_wom_exp	-1.493	3.908	-0.382	0.703

Because the p value for `gender_r:perc_wom_exp` is about 0.703, we fail to reject the null hypothesis and can conclude that there are no significant interactions between the two variables.

Results

Our initial model including total participation rate, school sector name, gender ratio, total student count, classification name, and percentage of women expenditure has a AIC of 3443.95 and a BIC of 3545.167. When we reduce the sector name variable, the AIC decreases to 3443.531 and the BIC decreases to 3527.144. In accordance to parsimony, we choose our final model that further reduce the model by taking out the classification name variable since it has the lowest BIC. We decide not to further reduce the terms in the model to have sufficient predictor variables for the response variable.

	x
total_partc	1.028
ef_total_count	1.148
gender_r	1.019
perc_wom_exp	1.174

There is no variable with $VIF > 10$ that indicates concerning multicollinearity, so no apparent issue with multicollinearity is found.

.metric	.estimator	.estimate
rsq	standard	0.018468

.metric	.estimator	.estimate
rsq	standard	0.002249

The RMSE for the training data predicting log-profit is about 5.112, which means that the RMSE for profit would be about $166 million. The RMSE for testing data is about 5.057, meaning that the RMSE for profit on the testing data is about $157million. While there is a difference of about $9 million, considering that the range of the data is in the hundreds of millions of USD, this is not a jarring sign that the model overfits the data. Therefore, we can reasonably conclude that the model can be generalized to new data. However, looking at the RMSE and R^2 of the training data, we may not conclude that this model performs well. We can see that there is a very small R^2 for the model showing that only about 1.85% of the variability in the profit can be explained by variables in the model.

Our final model includes total participation rate, total count of students, gender ratio, and percentage of women expenditure to predict the log of profit.

$$\widehat{log\_profit} = -2.619 + 0.065 * total\_participation\_rate + 0.006 * total\_student\_count - 0.363 * gender\_ratio - 7.178 * percent\_women\_expenditure$$

To answer our research question, **we can predict profit in USD of the collegiate sport basketball for schools in the year 2019 with predictors of student participation rate, total student count, gender ratio, and the percent of a school's expenditures that are spent towards women's sports**.

Regarding our hypothesis, our final model shows that only the effect of participation rate is significant in predicting a school's collegiate sport basketball profit. For every one percent increase in a school's sport's total participation rate, we expect on average that the school's profit from basketball in the year 2019 will multiply by a factor of 1.067, holding all else constant. We reduce the hypothesized variable sector name early using cross validation.

Interestingly, while the effect of gender ratio is not significant with a p-value larger than 0.05, the percentage of women expenditure is a significant predictor that we did not hypothesize would be influential. For every one percent increase in a school's expenditures that goes towards women's sports, we expect on average that the school's profit from basketball in the year 2019 will multiply by a factor of 0.001, holding all else constant. Total student count is another significant predictor. For every 100 student increase in a school's population count, we expect on average that the school's profit from basketball in the year 2019 will multiply by a factor of 1.006, holding all else constant.

Discussion + Conclusion

Our research concludes that the major three significant predictors of collegiate basketball profits are total participation rate, total count of students, and percentage of women expenditure.

There were some limitations. Regarding the model performance, since it has a fairly low R^2 of 0.0185 and high rmse of $166 million, our model does not generate very accurate predictions of collegiate basketball profits. As of the complexities of sports profits and the amount of variability such as school reputation and celebrity effect not simply accountable by statistics, it is fair that our model only explains a small amount of variability in the data. Also, there may be more predictors outside the scope of our dataset that influence the profit of collegiate basketball. We chose the school year 2019 for our analysis because it was the most recent available data in the database, which we believed reflects the latest conditions and trends in collegiate basketball. However, towards the end of the 2019 school year was when COVID 19 may have had an effect towards the end of the basketball season generally from November-March. We acknowledge that the data collected during this timeline may not be applicable for other years without COVID, influencing factors such as participation rates, revenue, and expenditure in collegiate basketball, and further anticipate that these factors may decrease during COVID because the pandemic likely introduced restrictions on gatherings and financial constraints on institutions. Lastly, the schools with missing data were those that did not have basketball teams. Since we are predicting profit from collegiate basketball, these schools are not in our population of interest and therefore should be removed. Potential limitations would be that we lose the data about which schools do not have a basketball team. However, looking at our data set the school's without a basketball team are mostly public 2-year institutions. There are less observations about public 2-year institutions to generalize the data on these types of schools, but our filtered data set still does include a large amount of observations from these schools.

In conclusion, we anticipate that the implications of our research predicting profit generation in collegiate basketball holds profound meaning in domains such as academia, sports management, and educational governance to foster an environment conducive to the holistic development of student-athletes. As a future direction, we would like to not just look at trends from the 2019 school year, but delve into longitudinal trends within collegiate basketball profitability by analyzing data spanning multiple seasons. Furthermore, we could extend the scope of our model to also predict profit from different collegiate sports like baseball or football.