Optimizing Spending to Improve Athletic Department Performance

Team Members: Hari Ilangovan, Morissa Chen, Saharsh Chordia, Zisis Daffas

Abstract

Collegiate athletics has a considerable impact both on institutions and student life due to large revenue and expenditure figures. In the present study, an investigation into the relationship between athletic budget management and institutional athletic performance is conducted. Using Capital One Cup ranking score as a metric of performance, several data mining techniques are applied in order to identify significant financial indicators of collegiate athletic success. The results demonstrate the importance of investing in medical expenses and increasing revenues in order to improve overall athletic department performance. Moreover, the analysis highlights the necessity of data collection beyond expenses and revenues, which will enable future researchers to enhance the predictive power of the formulated models.

Introduction

The purpose of our project is to understand how athletic budget allocation affects institutional athletic performance. Previous studies focused on the budget implications on collegiate football success, but the objective of this study is to take a holistic approach that looks at success across all sports[1]. College athletic programs generate billions in revenue nationwide and require significant funding to operate effectively. Each institution has many factors to consider when managing their funds, which include the sports to invest in, how much to pay operational staff, and how many scholarships to give out[2]. This study required the collection and aggregation of revenue and expense data from large collegiate programs in order to understand their budget allocation. Subsequently, the financial information for each institution was linked to a quantifiable metric that describes the program's athletic performance relative to its competition. Ultimately, this study focused on predicting how an institution will perform in all sports based upon the management of its athletic funds, which provides a method for improving their athletic department's performance.

Final Report

Data Cleaning and Preparation

Data Sources

This project utilizes data from three different sources. The first data set contains information about college athletic department revenue and expenses obtained from the publicly available College Athletics Financial Information (CAFI) Database. This provided financial information on public schools for six years ranging from 2010 to 2015. In order to include non-financial data regarding these colleges we found demographic information from the College Scorecard. This data set was used to obtain the tuition of schools and the undergraduate student population. Lastly, the metric used to quantify collegiate athletic performance is the Capital One Cup ranking score. The ranking score for each institution is publicly available for the same range of years that financial data are available. These rankings assess athletic success across all sports based on the final standings of NCAA championships. Using a two-tiered scoring system both men's and women's sports are divided into two groups. This grouping is performed based on popularity and pool of competition and sports that belong to different groups are weighted differently. By joining these three data sets, the data include 15 variables related to each college and its spending and 1 response related to athletic performance.

Data Cleaning and Integration

To integrate these three data sets, a full inner join was first used to combine the CAFI data and Scorecard data on University ID and year. Since the Capital One cup ratings were scrapped from PDFs, this dataset required us to create a mapping of colloquial university names to official institute names. Then, a full inner join on official institute names and year was conducted between the joined CAFI and Scorecard data to produce our final dataset. Due to the limitations of privacy regulations for private schools, our final dataset was only able to include public universities. This resulted in a significantly smaller dataset than initially anticipated with 472 unique observations and 18 different variables; the response variable is total Capital One Points between men's and women's sports.

Exploratory Data Analysis

The exploratory data analysis (EDA) performed on the cleaned data set examined both a continuous and discrete (categorical) response based on the total Capital One *Total Points* value.

As depicted in Fig 1, the *Tuition*, *Undergraduate Size*, and *Total Points* variables are not significantly correlated with each other or any other variables. The remaining variables, how-

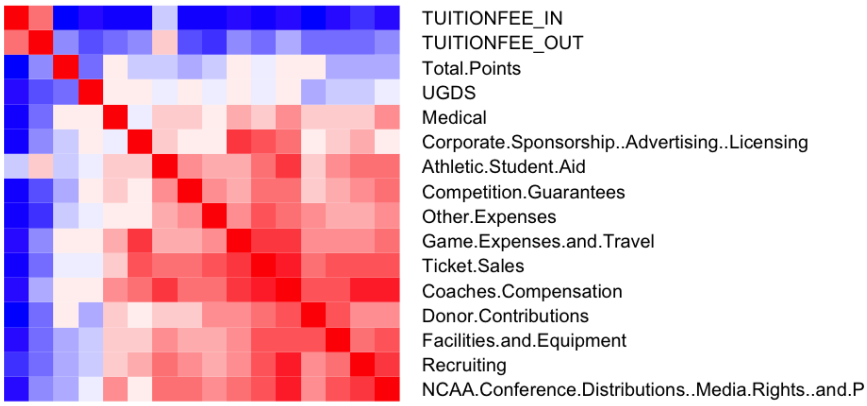

Figure 1: Correlation Heatmap with blue for negative correlation and red for positive correlation

ever, show significant positive correlation with each other. These variables are all related to revenue and expense behavior of the universities measured in USD. These positive correlations may result in multicollinearity when fitting the model, so a Variance Inflation Factor (VIF) analysis was performed on these covariates.

Variable	*VIF*
Coaches Compensation	**12.25**
Ticket Sales	**7.10**
NCAA Conference Distributions	**5.90**
Recruiting	4.87
Game Expenses and Travel	4.85

Table 1: Top five independent variables based on VIF

As shown in the table 1 above with the results of VIF analysis, there are 3 variables with VIF greater than 5. The *Coaches Compensation* variable has a VIF exceeding 10, so the remaining independent variables will explain the majority of the same variability in the response. Therefore, the *Coaches Compensation* variable is removed from the data set in the subsequent analysis for the continuous response modeling.
Due to the differing scales for each of the monetary variables, multiple normalization techniques were applied and analyzed according to their distributions. Using the *Coaches Compensation* variable against the response as an example to show the general behavior of all financial predictors, in Fig 2 panel 1 the simple linear regression and distribution show a megaphone effect. This effect is corrected for using a log transform on the response and predictor. The skew of histogram is decreased in the log normalized results in panel 2 of Fig 2.

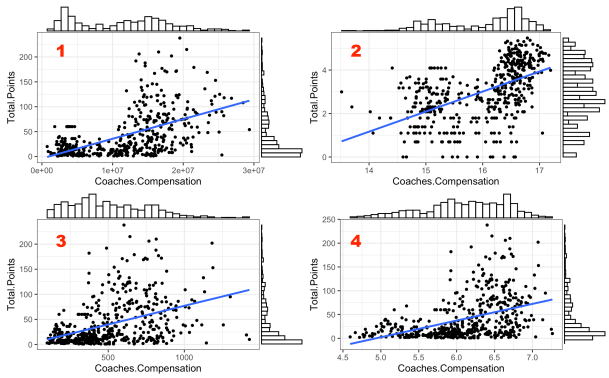

Figure 2: Plot of *Coaches Compensation* against response - *Coaches Compensation* used for reference because of multicollinearity with other financial predicting variables

In panels 3 & 4 of Fig 2 a new feature from *Coaches Compensation* is created by using *Undergraduate Student Population Size*. This is achieved by dividing the compensation of each school by its respective student size. This approach shows a long right tail in panel 3. In panel 4, a log transform is applied to the new feature and response, which shows a much more normal distribution of the predictor and a short tail.

These transformations were all applied in the subsequent analyses, but the approach in panel 3 that leveraged normalization by the *Undergraduate Student Population Size* is the only transform that proved useful. The log transform is explored but did not produce additional prediction power or interpret-ability.

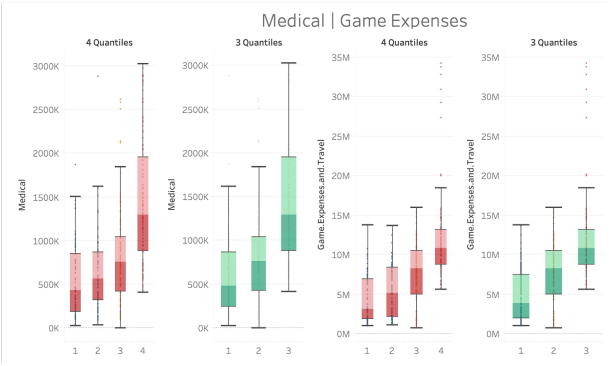

Figure 3: Tertile versus Quantile comparison for the *Medical Expenses* and *Game Expenses* variables

As shown in Fig 4, the quartile split of the response resulted in similar 1st & 2nd quartiles based upon the means, and quantile distribution between each of these groups. Using tertiles as a result of combining groups 1 and 2 from the quartiles, the box-plot analysis identified 3 disparate groups split at the 50th and 75th percentiles of response. Using a visual analytic approach, the box-plots are shown to have unequal means between groups. As a result of

this split, the total size of group 1 is larger as a result of the large portion of low scoring schools in capital one standings.

Methods

The explanatory data analysis described in the above section revealed the existence of high non-linearity in the relationship between the response and predicting variables. As a result, a basic linear regression model performed poorly. In order to address this non-linearity and capture more variability several transformations of the response and predictors were applied, along with the incorporation of some interaction terms. However, none of these techniques led to noteworthy improvements. This was mainly due to the fact that all quantitative predictors of the dataset are measured in the same unit (i.e. USD) and are highly correlated to each other. Tree-based and non-parametric models provided slightly better results than other non-linear methods, such as piecewise polynomial regressions (splines), but none of these results were satisfactory. The three best-performing models in terms of test prediction error are summarized below:

- Random Forest: An ensemble method known to have strong predictive power, Random Forests are built by creating many decisions trees using bootstrapped training data. Their strong predictive power is due to the many trees being uncorrelated. This happens because at each split in each tree, only a small subset of the predictors are allowed to be split on. In order to determine the number of variables to split on in each tree and the number of trees to create, as well as the minimum node size (i.e. the minimum number of samples in each node), the testing error was used as a metric of comparison. To be more specific, multiple random forest models were trained using different combinations of the above parameters and the model with the smallest testing error was chosen as the final one. The best Random Forest model built 1000 trees using 5 variables in each tree, and a minimum node size equal to 3.

- Boosting: Similarly to Bagging, Boosting is an ensemble technique that uses trees as building blocks to construct more powerful prediction models. However, in contrast to bagging, Boosting does not involve bootstrap sampling and trees are grown sequentially using information from previously grown trees. Following the same process which was described above for the Random Forest model, the parameters of the Boosting model were tuned based on the testing error. In particular, the Boosting model with the smallest testing error grew 1000 trees with a learning rate equal to 0.12 and maximum depth of each tree equal to 6.

- KNN Regression: One of the simplest non-parametric methods, KNN regression generally outperforms linear models under high levels of non-linearity. This machine learning algorithm estimates the response corresponding to each data point by using the average of the responses of its k-nearest points. To find the best value of k, 15 different values

of this parameter were tried and the corresponding models were compared based on testing error. From this analysis, the KNN regression model with the lowest testing error used k=3 neighbors.

Given the small size of the examined dataset, a Monte Carlo Cross-Validation algorithm was implemented in order to evaluate these 3 models. Each model was trained 100 times and the testing errors for each model were calculated. Each train / test iteration used a different train and test dataset. This was accomplished by changing the seed in R for each iteration, thus having different observations randomly chosen for each dataset. Then, the mean and variance of those 100 testing errors was calculated to evaluate the models. This method provides a better evaluation of model performance, because it enables a random sub-sampling of the data into train and test sets multiple times, thus allowing the models to fit and test against different underlying patterns.

The results obtained from this analysis were extremely poor and indicated that all models were governed by high uncertainty in terms of outcomes (see Fig. 5 in Appendix). Hence, it was concluded that with the current nature of the dataset it was not possible to get adequate results by predicting Capital One Points as a continuous variable. Consequently, we decided to group the teams into buckets (or tiers). We initially explored splitting the schools into four groups based on their quartiles, but after analyzing different histograms and boxplots of the predictors (see EDA section), there didn't seem to be a meaningful difference between the first and second quartile team's points. Therefore, we grouped schools below the 50th percentile together as below average schools, schools between the 50th and 75th percentile as average schools, and schools above the 75th percentile as above average. After splitting the data into these groups, several classification models were trained and tuned to select the one with greatest predictive power. Then, we ran again a Monte Carlo Cross-Validation Algorithm to calculate the testing error for the top 3 models, which were the following:

- K Nearest Neighbor: The K-Nearest Neighbor's model takes the k nearest point's classification and assigns the new point to the class where the majority of its neighbors lie. In the KNN function used in this analysis, the distance calculation used was Euclidean distance (L2-norm). This method is effective for smaller datasets but can be sensitive to outliers. When training this model, we iterated through 15 values of k and calculated the testing error for each of the models. From that analysis, the KNN model had the lowest testing error at k=5 neighbors.

- Random Forest: Following the same methodology with that described in the Random Forest section for modeling with continuous response, the best Random Forest model for classifying which group the school falls into, built 500 trees using 4 variables in each tree and a minimum node size equal to 3.

- Support Vector Machine: A Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane, which categorizes observations based on a

certain number of attributes. Although this algorithm is mostly used for binary classification, it can be extended to the more general case in which there are more than two classes. In the present work, this was achieved by using the "one-versus-one" classification approach of the ksvm package of R statistical software. The two most important tuning parameters of support vector machine models are the cost of constraints (C) and the kernel function. More specifically, the parameter C determines the number of the violations to the margin (and to the hyperplane) that are tolerated, controlling the bias-variance trade-off of the statistical learning technique, while kernel functions enable the enlargement of feature space in order to address the problem of non-linear boundary between the classes. After iterating through different values of C (from 0.001 to 1000) for several kernel functions, the suitable order of magnitude of this parameter was found based on testing error. Then, using again the testing error as a performance metric, the best value of C and the corresponding kernel function were determined. Fig. 4 illustrates the variation of prediction accuracy for different kernels. It is shown that the highest accuracy is achieved by using a radial kernel with C equal to 8. Moreover, it shall be noted that although the radial kernel is a substantial improvement over the linear kernel "vanilladot", all other kernels provided same (polynomial kernel "polydot") or worse results than the linear classifier.

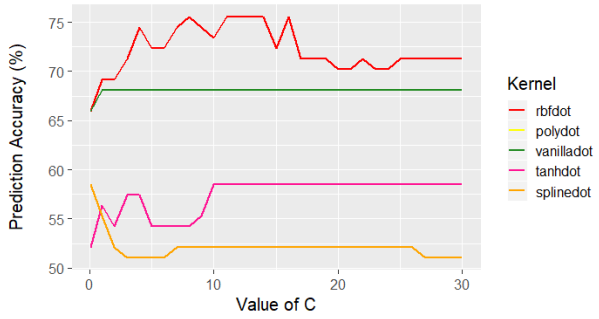

Figure 4: Accuracy of kernel functions for different values of C

Results

The mean squared error of the models used to predict Capital One Cup points as a continuous variable are summarized below:

Model	RMSE - Mean	RMSE - Variance
Gradient Boosting	30.16	8.26
KNN Regression	37.78	14.09
Random Forest	29.06	9.87

The following table displays the testing error of the classification models used to predict a school's performance in the Capital One Cup rankings:

Model	Testing Error - Mean	Testing Error - Variance
Random Forest	28.0%	0.10%
KNN	32.4%	0.20%
SVM	28.4%	0.14%

Based on the above results, we determined that the Random Forest model was our best performing model. By looking at a variable importance metric (Fig 7 in Appendix), which is determined by measuring the average amount of decrease in the Gini index by splits over a given predictor, our top predictive variables were:

- Corporate Sponsorship, Advertising & Licensing
- Game Expenses and Travel
- Other Expenses
- Ticket Sales
- Medical Expenses

Recommendations

From the important variables identified above, we believe that there are three key predictors that schools can really focus on improving to increase their athletic success.

- Corporate Sponsorship, Advertising & Licensing is the lowest proportion of revenue (less than 11% for all teams see Fig 6 in Appendix) and the most important predictor

for classification. We recommend teams spend more resources on acquiring corporate sponsorships, advertising and licensing. This influx in revenue can lead to increased spending attract better athletes.

- Ticket Sales: every school should have the goal to increase this revenue to improve athletic department performance as the better performing schools have a higher percentage of revenue coming from Ticket Sales than worse performing schools. This could involve running promotions such as meet and greets with players and coaches or field access when purchasing tickets

- Medical Expenses is a strong indicator of a successful athletic school. By keeping athletes healthier, teams can expect more success on the field and therefore should try to increase Medical Expenses by investing in top trainers and recovery equipment.

Though Game Expenses & Travel is an important contributor to Capital One success, it is difficult to control as Game Expenses & Travel are a result of a team's schedule which is made by conference administrators. Therefore, we do not recommend schools allocating any resources to optimize this spending.

Lessons Learned - Conclusions

After developing our models and getting poor predictive results, we identified a few key problem areas to further improve our analysis in future iterations. First, we determined that high-level budgetary data alone is not enough to predict success and more granular budget data would be needed to understand a school's athletic prowess. For example, revenue and spending at the sport level (i.e. Football, Tennis, Baseball, etc.) would be helpful to understand where the top teams are investing the most money. Moreover, since normalization improved model predictions, we suggest applying similar normalization techniques with respect to other features related to student, athlete and faculty size. Next, we believe there are many other categorical variables needed to better predict success. Variables such as coaching experience, recruiting strength and number of upperclassmen are extremely crucial to athletic success, but are difficult to collect. Finally, due to vast differences amongst schools athletic departments, we believe schools will be more successful analyzing their own historical budgets to optimize future budgets. For instance, schools in different conferences have extremely different revenue sources through TV deals and endorsements at the conference level. This allows schools to spend more on athletics and have more sports teams overall than schools that do not have as much revenue from their conference. This inherent bias makes it difficult to identify trends across schools nationally. Therefore, in future analyses, we recommend that schools analyze their own historical budgets. By looking back 10-20 years, schools can identify what variables led to successful years and can work to optimize those factors going forward.

References

[1] Mirabile, McDonald, and Mark Witte. "Can Schools Buy Success in College Football? Coach Compensation, Expenditures and Performance." IDEAS Working Paper Series from RePEc (2012): IDEAS Working Paper Series from RePEc, 2012. Web.

[2] Bradley, Robert. "A Comparison of Athletic Training Program Financial Resources." The Sport Journal 13.1 (2010): The Sport Journal, Wntr, 2010, Vol.13(1). Web.

Appendix

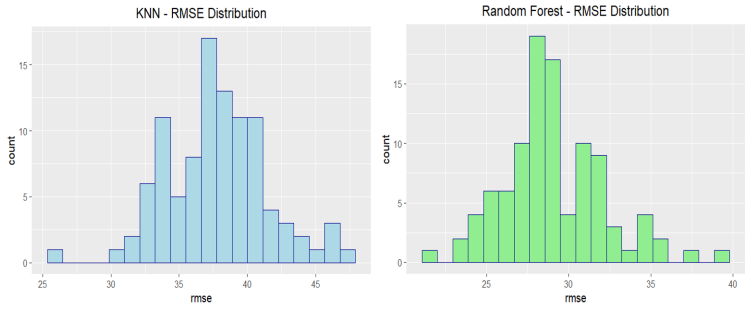

Figure 5: RMSE Histograms - Monte Carlo Cross Validation

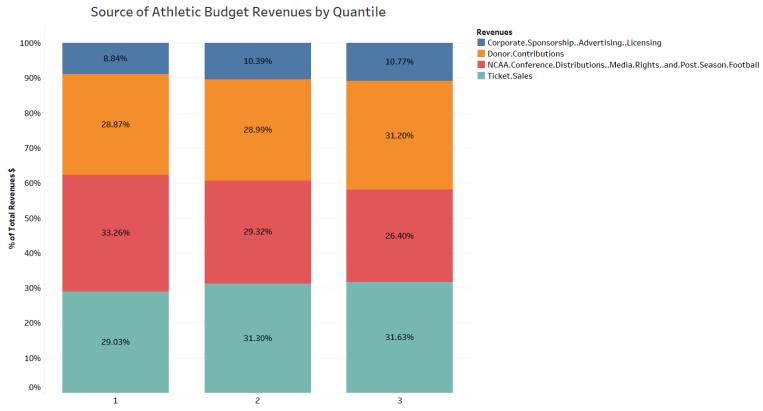

Figure 6: Distribution of revenue over tertiles

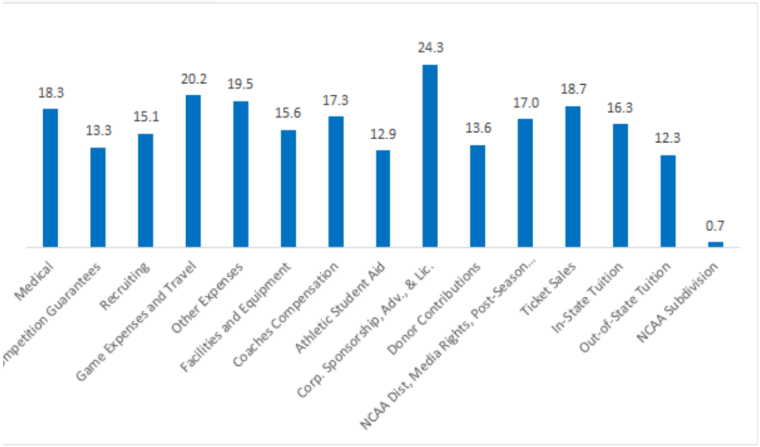

Figure 7: Variable Importance based on Gini Index from Random Forest