College Basketball Predictions

By: Ben Oberg & Michael De Simone

Introduction

- Describing the Project:
  - Analyzing the impact of key statistics on college basketball team performance.
  - Seeking to contribute to the evolving realm of sports analytics.
  - Building a model to predict winning while focusing on the regular season.
- Justification:
  - With the rise of analytics in sports, identifying crucial stats is imperative.
  - Aim to determine statistics crucial for success and their optimal thresholds.
  - Ultimately, creating predictive models for team performance.

Related Works

- We found a few research papers that also aimed to predict college basketball games:
  - [1] Created multiple models with logistic regression being the most successful for predicting games. Their model finished with a .6 RMSE. They built their model in multiple steps, first with one team, then one conference, then all of college basketball.
  - [2] looked at head to head matchups from 2008 to 2013 and tried to determine who would win in these matchups. They took away two key points, with the first being that the variables selected are far more important than the models used. They also found there to be a certain point where the models can not get better due to the randomness of the sport.
  - [3] Utilized 5 different models to predict winning: Neural Networks(67%), SVM(65%), KNN(63%), Logistic Regression(63%), and RFs(61%).

Dataset

- Link: https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset
- Our Dataset contains statistics from the 2013-2023 seasons of division 1 college basketball, with only 2020 being removed from the dataset.
- Each year is contained in its own dataset, so we have 10 datasets with around 350 teams in each dataset, so we have over 3500 instances to work with.
- The dataset contains 23 variables ranging from the teams names, amount of wins, to specific stats such as adjusted offensive efficiency.

The Variables

- Our target variable in the dataset was obviously wins as we set out to determine what affects winning.
- To be able to determine the variables that affected winning we needed to remove a few variables:
  - Within the 23 variables there were 2 that needed to be removed right away, these were the team's name and the conference which the team played in.
  - Next there were 2 more variables, games played and wins above bubble, two variables that directly correlated to winning, thus they needed to be removed.
  - Lastly there was their tournament seeding and how they performed in the tournament, two other variables that also directly correlated to winning.
  - So we were left with 1 target variable(wins) and 16 predictor variables which were: ADJOE, ADJDE, Barthag, EFG_O, EFG_D, TOR, TORD, ORB, DRB, FTR, FTRD, 2P_0, 2P_D, 3P_O, 3P_D, ADJ_T

Exploratory Data Analysis

Correlation to V4(Wins):

Four most correlated variables:

ADJOE: Adjusted Offensive Efficiency

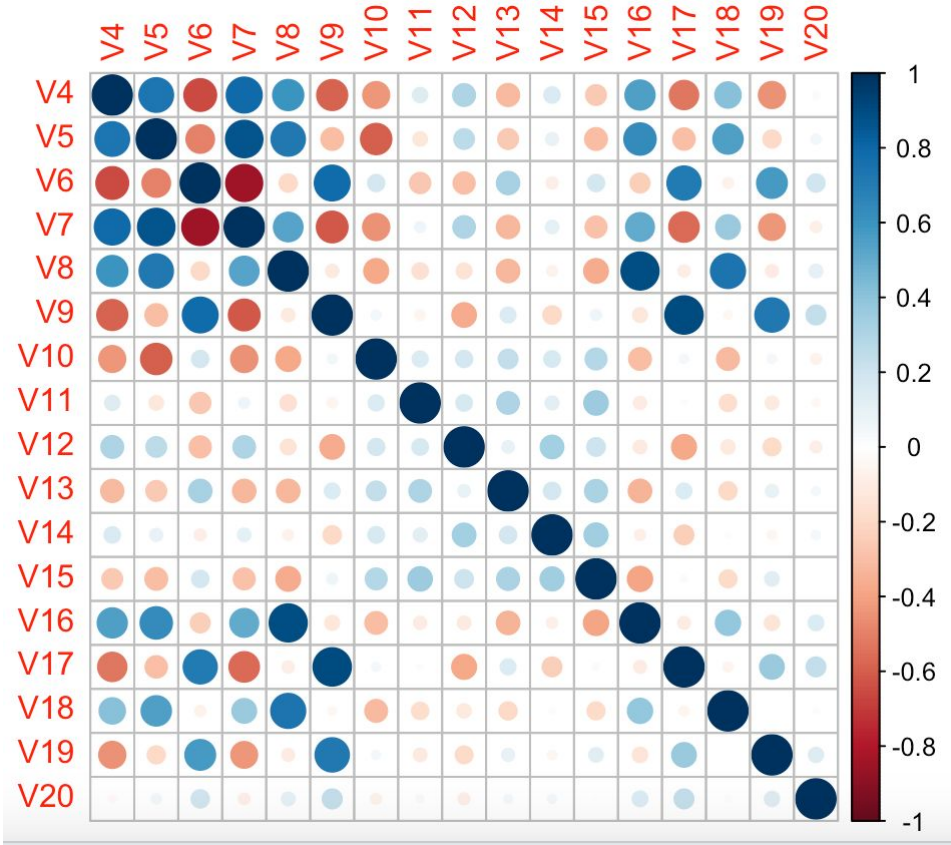
BARTHAG: Chance of beating an

Average D1 team

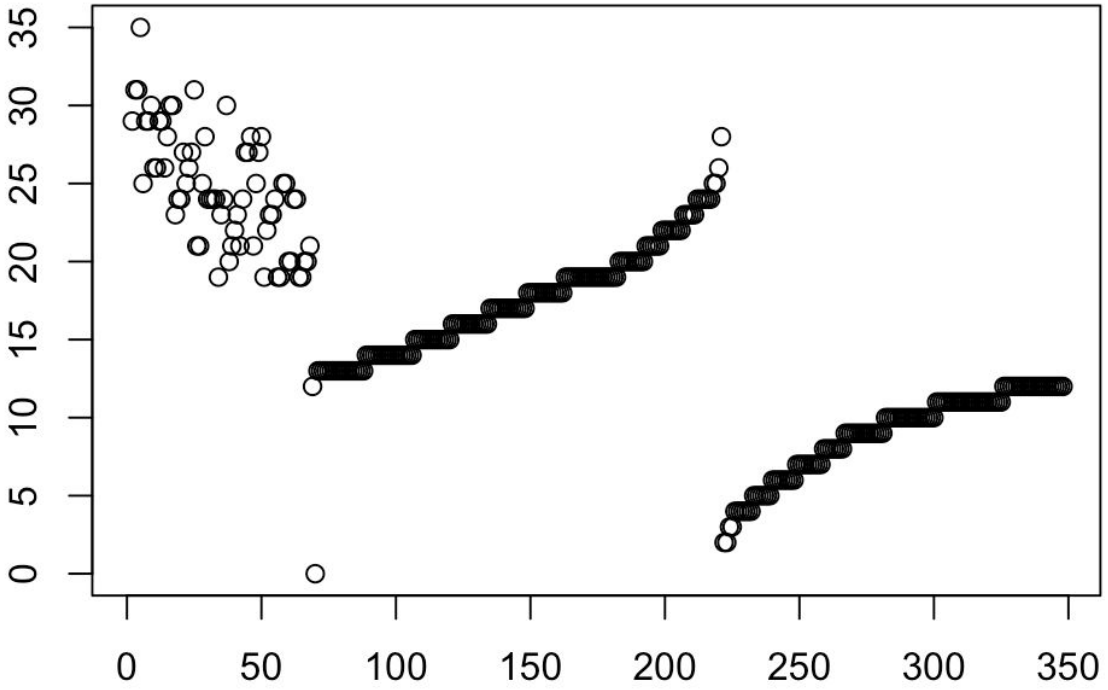
EFG_O: Effective Field Goal Percentage

Shot

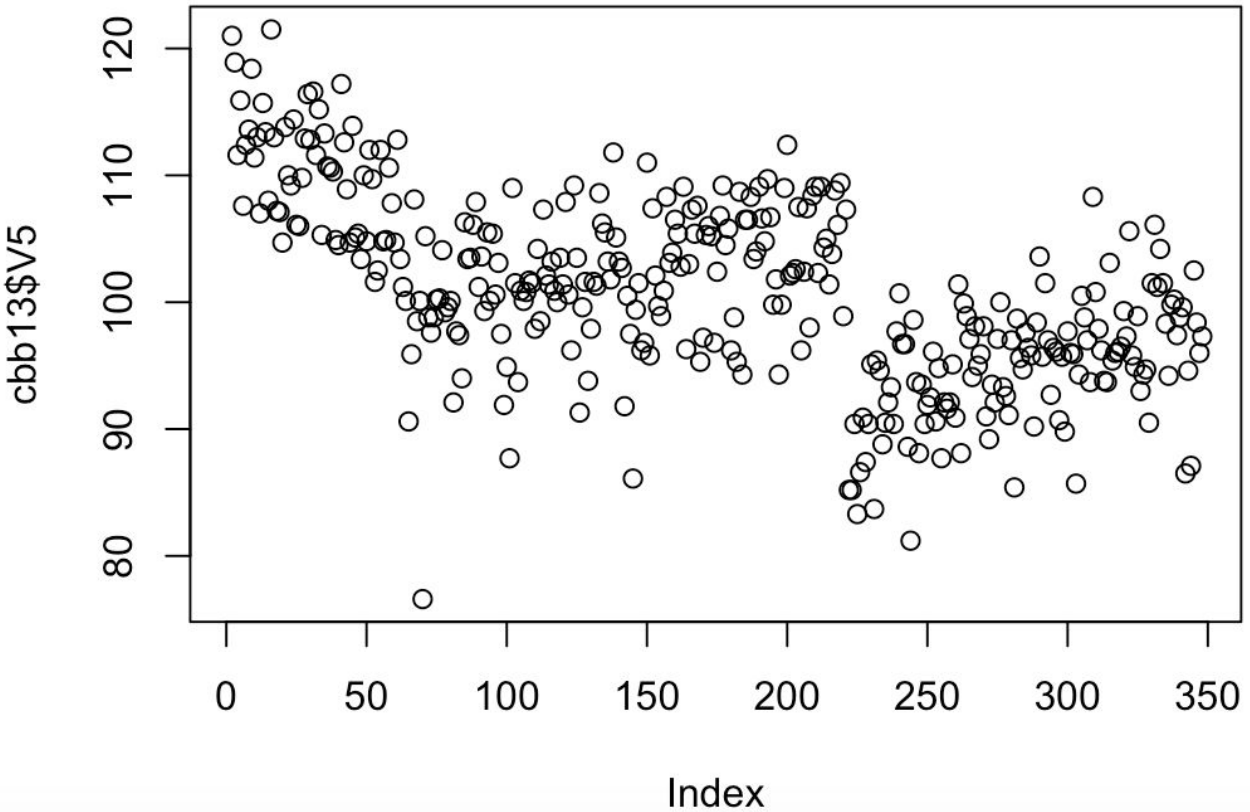
2P_O: Two-Point Shooting Percentage



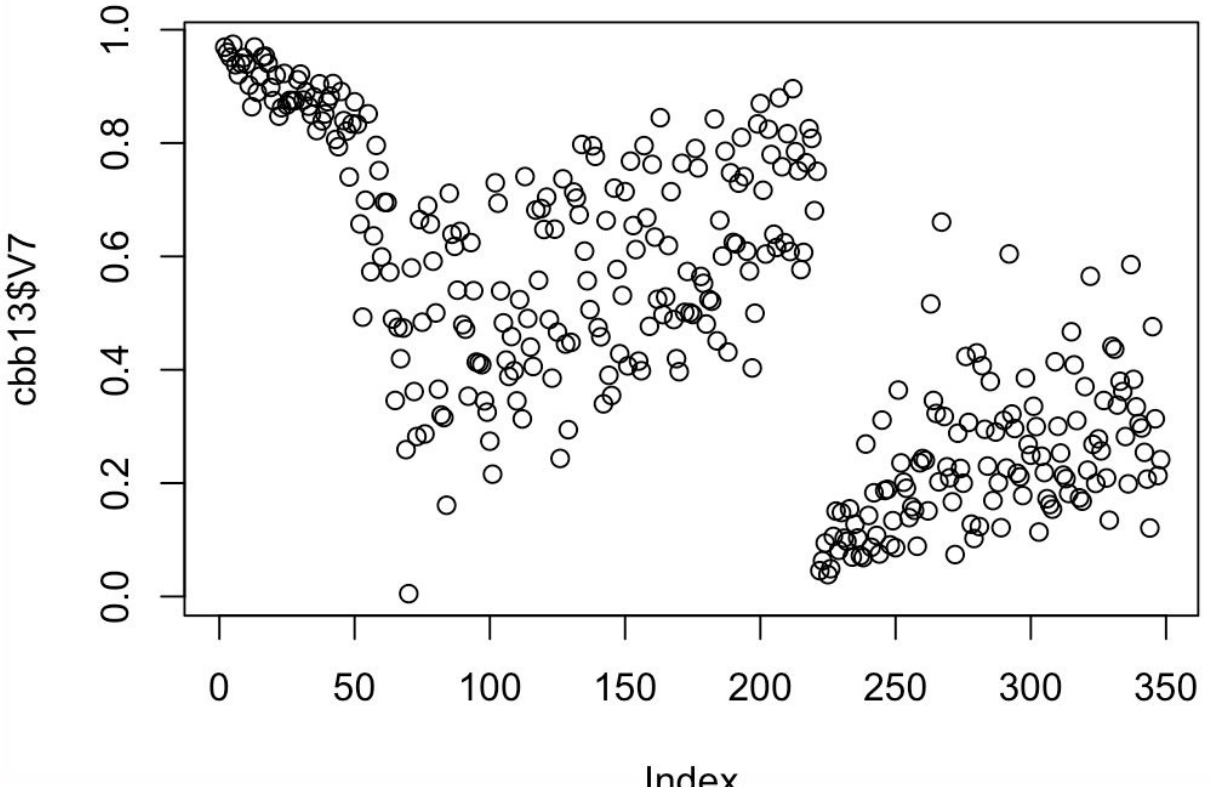
Wins Graph



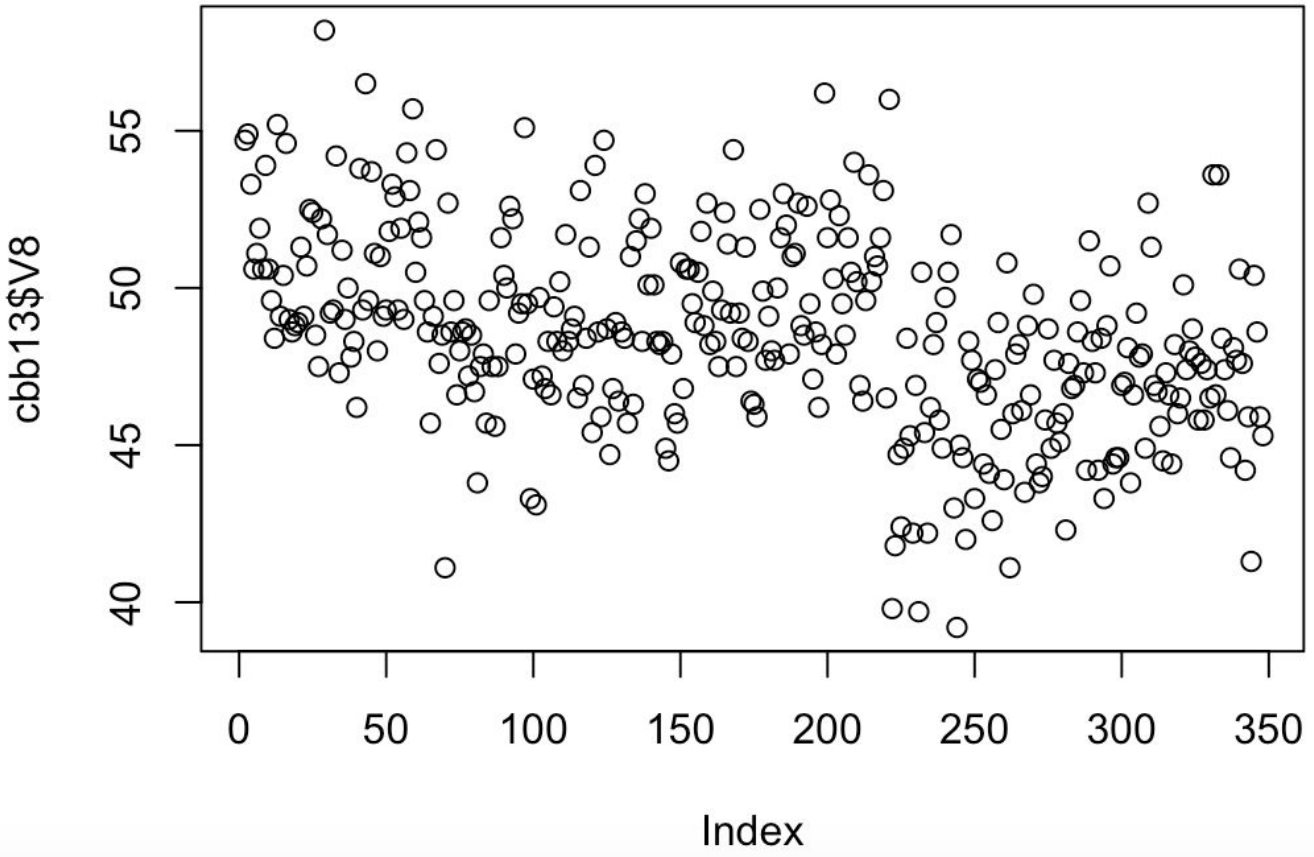
ADJOE



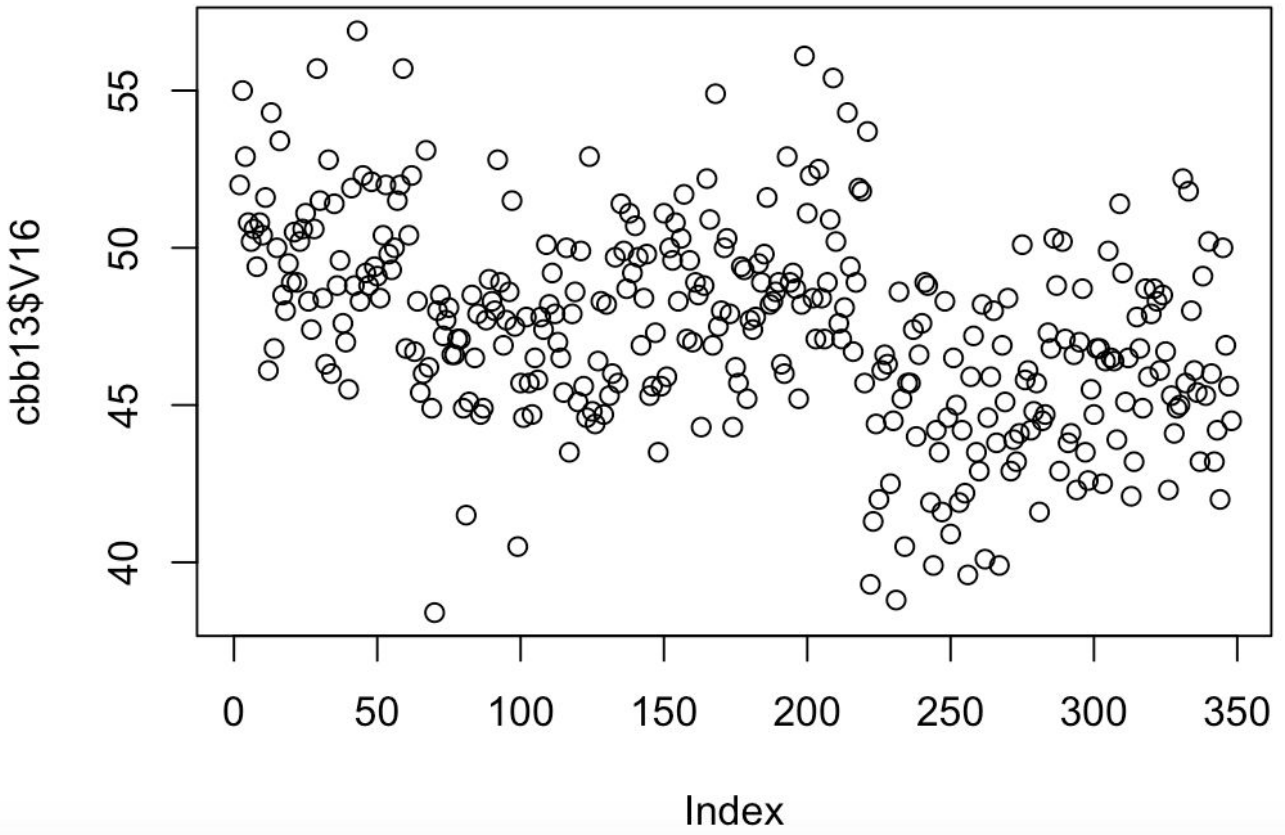
BARTHAG



EFG_O



2P_O



Methods

- Feature Selection
  - Random Forest Selection
  - Correlation Based
  - Variance Based
- Cross Validation
  - 10 Fold Cross Validation
- Models Used
  - Linear Regression
  - Random Forests
  - XGBoost

Feature Selection

- Random Forest selected
  - Adjusted Offensive & Defensive Efficiency, BARTHAG, Effective Offensive & Defensive FG %
- Correlation with a cutoff of 70%
  - Adjusted Offensive and Defensive efficiency, BARTHAG, and Effective FG %
- Variance
  - Using low variance no variables were selected so we proceeded with correlation for our linear regression models

Cross Validation

- For both models we used a simple 10 fold cross validation in order to get a better understanding of the model's potential performance
- We used cross validation in order to test our models further and prevent overfitting

Random Forest

- With Feature Selection & Cross Validation we had an RMSE of: 3.42 and the most important variables were as expected: Adjusted Offensive & Defensive Efficiency, BARTHAG, Effective Offensive & Defensive FG %
- Without Feature Selection & Cross Validation we had an RMSE of: 3.06 and the most important variables were: BARTHAG, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, 2P %, Effective FG %, and Effective FG % against.

Logistic Regression

- We utilized the correlation feature selection for one of our logistic regression models. The variables we used for this were Adjusted Offensive & Defensive Efficiency, BARTHAG, and Effective 2P %.
- With this set up we had an RMSE of 3.575 and the most important variables in order were Offensive Efficiency, Effective 2P %, BARTHAG, Defensive Efficiency. The last 2 had negative coefficients.
- Our next model we utilized 10 cross validation but this worsened our RMSE to 3.7
- Lastly we used a model with no cross validation or feature selection. This had an RMSE of 2.83, The top 4 features were BARTHAG, Steal rate, Offensive rebound rate, and Effective FG % in that order.

XGBoost

- We also utilized XGBoost to create a model but found little success when utilizing feature selection and cross validation and when not using them.
- Due to these reasons we moved forward with Logistic Regression and Random Forest.

Conclusion

- We came into this project looking to build models that can predict how much a team wins in college basketball based off their statistics, while also learning what statistics affect winning the most.
- I believe we succeeded in building multiple models that can predict winning
- As far as the most important variables a few consistently showed up: BARTHAG, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective FG %, and Effective FG % against. 2P % also showed up sometimes too.
- This is not too surprising as I expected these stats to be the most important in the dataset.

Future Works

- We plan on testing our models in multiple ways
  - The statistics are being released for the 2024 season which just concluded, We plan on seeing how accurate are models are with this new dataset.
  - We also plan on looking at the data halfway through next season and attempting to predict how many more wins each team will have.
- We also intend to predict winners in head to head matchups next year based off the most important statistics to see if they truly determine winning.
- We also plan on attempting to create models with more success by using various methods and new models entirely.

References

- [1] Mente, S. N. (2022). Accuracy of machine learning algorithms in predicting college basketball games. Bachelor's thesis, Department of Computer Science, University of Virginia. https://libraetd.lib.virginia.edu/downloads/sn009z84g?filename=Mente_Sindhura_Accuracy_of_Machine_Learning_Algorithms_in_Predicting_College_Basketball_Games.pdf
- [2]Zimmermann, A., Shi, Z., & Moorthy, S. (2013). Predicting college basketball match outcomes using machine learning techniques: Some results and lessons learned. Research Gate https://arxiv.org/abs/[arXiv_ID]
- [3]Kim, J. W., Magnusen, M., & Jeong, S. (2023). March Madness prediction: Different machine learning approaches with non-box score statistics. *Journal of Multivariate Analysis*, https://doi.org/10.1002/mde.3814