Case Study – Predicting the NCAA 2019 Men's Basketball Championship Rubric

DS 4002 - Fall 2022 - Catherine Schuster
Submission format: PDF document, source code file, and completed bracket.

Group Assignment

General Description: This assignment will have you create an algorithm to model and predict the entire 2019 NCAA bracket.

Prerequisites: Intro machine learning, statistics, or data science course. Introduction coding course (R or Python). A basic understanding of statistical machine learning methods.

Why am I doing this? This assignment is an opportunity to learn first-hand how to decide on the appropriate machine learning model for your goals, and how to optimize this model with feature selection and fine-tuning techniques. You will then apply high-level data science thinking by communicating your results in context and reflecting on your work.

What am I going to do? Produce a methodology to predict the winner of each game in the NCAA 2019 tournament. You have been given three datasets to work with for this project that contain individual game statistics and average across-season statistics from 2003-2019. These can be found in the case repository. You may collect additional data if you feel that you need it. Your work will be assessed by the thoughtfulness of your methodology, your choice of model, and how the features used align with the goal of predicting the outcome of basketball games. You will NOT be graded by the accuracy of your predictions. Produce:

- Document explaining your methodology, analysis, limitations, and conclusions.
- Your completed bracket.
- Well-documented source code containing data cleaning, EDA, model building, and analysis.

Constraints:

- You are given a starting bracket with the initial 32 NCAA 2019 Tournament games.
- You are only required to determine the winner of each NCAA game. The predicted score of each game is not required or necessary.
- Your model should not use the performance result of a game after it has happened to predict the outcome of the same game.
- You should only use one model to predict the outcome of each game, i.e. do not use different metrics / techniques for each match-up.

Tips for success:

- Think about the context of the problem before diving into your methodology. Does the question you are trying to answer require a classification based approach (win/loss), or a

regression based approach (point differential)? What attributes are most important to answer your question?

- Consider the difference between *flexible* and *inflexible* statistical machine learning methods.
    - Inflexible methods make an assumption or use subject-based knowledge about the functional form of your response variable. If the assumed functional form is too far from the true response variable's distribution, then the estimate will be poor, leading to low accuracy. Inflexible methods typically have higher bias and lower variance on test data than flexible methods. They are easier to interpret, and do not need a large sample size.
        - Examples of inflexible methods:
            - Logistic regression (binary response variable)
            - Linear discriminant analysis (binary or multiclass response variable)
            - Ordinary least squares regression, lasso regression, or ridge regression (quantitative response variable)
    - Flexible methods do not make any assuption about the fuction of your response variable, and seek to get as close to the data points as possible. Flexible methods typically have lower bias and higher variance on test data than inflexible methods. Flexible methods need a lot more observations to accurately estimate your response variable. These methods are harder to interpret and more likely to overfit your data.
        - Examples of flexible methods:
            - K-nearest-neighbors
            - Classification or Regression tree based methods
            - Bagging, boosting, and random forest.
- Beyond using the optimal method for predicting each game outcome, choosing the appropriate attributes, statistics, and methods are just as, if not more important.
    - Examples of predictor variables:
        - Cumulative metrics (i.e. the average free throw percentage from the last N games preceding the current game)
        - Across-season aggregated metrics
        - Difference in seeding ranks
        - Consensus rankings
- You have been given an abundance of variables to use. Remember that a simpler model is more reliable than a complex model when prediciton accuracy is comparable. Use variable selection as an advantage for interpretation, correcting multicolinnearity, and to avoid overfitting.

How will I know I have Succeeded? You will meet expectations on Predicting the NCAA 2019 Men's Basketball Championship when you follow the criteria in the rubric below:

Spec Category	Spec Details
Methodology Document	Imagine that this requirement is your official report you will turn in to fulfil your manager's request. Submit a pdf document with a title, your name, the course name, and the date, that describes your approach, findings, limitations, and conclusions, with four sections. 1. Executive Summary: a high level explanation of the goal of this project and the work you completed. Use 2-4 short and precise sentences. 2. Methodology: a thorough description and rationale of how you arrived at an optimal model and which features you chose to use in the selection process. Include the choices you made to fine-tune and improve your model, if necessary. 3. Analysis: a thorough analysis of the game outcomes arrived by your model. Include any EDA you performed to understand your data and arrive at your final model. How do these outcomes compare to the true game outcomes? What were the strengths and weaknesses in your choice of model? Include model performance visualizations, if appropriate. 4. Conclusion: a high-level summary of what you learned from this project. Include a summary of your key-findings and its significance, in context. Describe how these conclusions compared with the true outcome of the tournament. If your findings were not what you expected, explain why. Lastly, list the challenges you faced, how you overcame them, and any limitations you discovered in your work.
Bracket	The completed bracket template using the outcomes that you predicted to fill in each game slot. Include your name at the top of this template.
Source Code	A well-commented, well-documented .r, .rmd, or .py file containing the code you used to

	execute your methodology. This includes data cleaning, exploratory data analysis, model building, model testing, and post-model analysis.

Acknowledgements: This rubric is built from the many rubrics provided by Professor Loreto Alonzi in DS 4002. This structure is pulled directly from Streifer & Palmer (2020).