Final Report: NCAA March Madness Predictor

Andrew Bryan

Problem Statement

Baylor University, aiming to enhance basketball strategy and recruitment, plans to leverage historical NCAA Basketball data from the 2013/2014 to 2017/2018 seasons, sourced from Google Cloud. The project's success hinges on identifying insights that aid in accurately predicting game winners, particularly in NCAA Tournament matches.

The scope encompasses analyzing all college basketball data with a focus on regular season games to inform tournament predictions. Despite constraints such as limited online data availability beyond the specified seasons, the project aims to develop a deployable machine learning model capable of providing actionable insights.

Stakeholders including Baylor's basketball head coach and Athletic Director anticipate utilizing the findings to inform strategic decisions and player recruitment efforts, with primary deliverables being a presentation deck summarizing key insights and a deployable predictive model supported by comprehensive documentation.

Data Wrangling and Cleaning

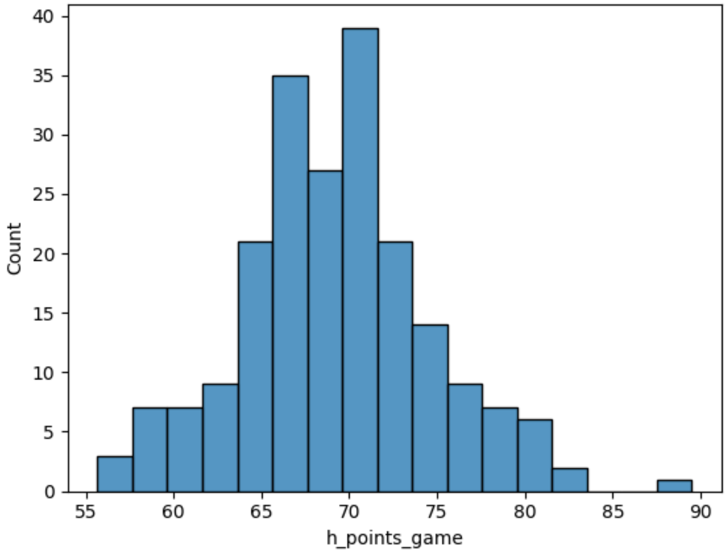
The dataset obtained from Google Cloud consists of 29,805 lines of individual game statistics spanning four years. A thorough investigation was conducted to assess the usability of the data and to understand its contents. The initial step involved identifying the number of null entries. These were addressed by either filling in the null values with the median, zero, or removing the rows entirely. It was found that several rows only contained box scores, which were insufficient for the machine learning model since comprehensive game statistics were required as features.

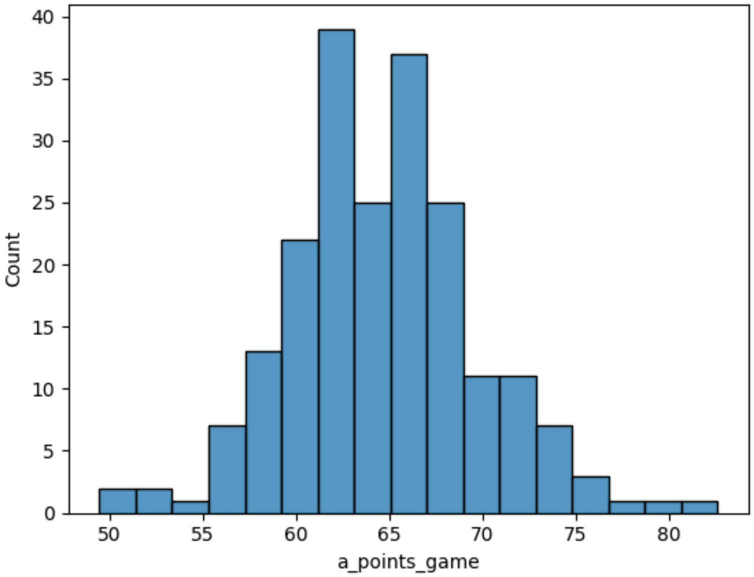
Moreover, the dataset included an extensive array of extraneous columns, such as team logos, possession arrows, and venue addresses, which were irrelevant for this project's goals. These columns were subsequently removed to streamline the dataset. Further scrutiny revealed that the dataset encompassed games and teams from D2, D3, and other divisions not pertinent to D1 men's basketball, the focus of this project. Rows containing data from these divisions were also eliminated, resulting in a more focused and relevant dataset for the analysis.

Exploratory Data Analysis

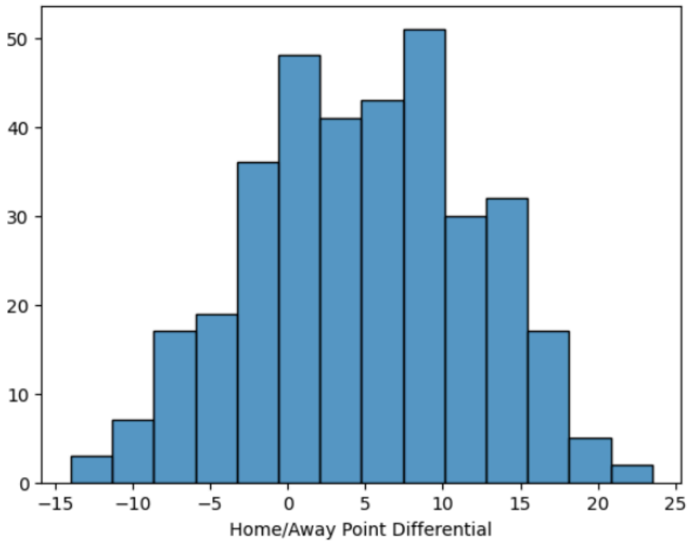
The exploratory data analysis (EDA) step in this project took a unique approach since most data relationships were already established. However, a critical aspect of the categorical models was examining whether a team's home performance statistically differed from its away performance. This analysis would guide how features were structured in the model, determining whether to use aggregated statistics across all games or to separate them into home and away categories.

The following histograms compare the average points scored by home and away teams across the league.





These histograms revealed that the normal distribution of points scored by away teams is typically about 5 points lower than that of home teams. This observation was supported by an analysis where the average home and away scores for each team were calculated, the differences were computed, and an average difference of approximately 4.99 points was observed. However, further detailed investigation showed significant variability among individual teams. Some teams scored up to 20 points more at home than away, while others scored up to 10 points more when playing away.



This variability suggests that home and away performances significantly affect teams differently and should be treated as separate features in the predictive model. The histogram clearly indicates that a blanket approach to team performance, without accounting for venue, would overlook crucial nuances in the data.

Model Preparation and Pre-Processing

This phase was the most complex and substantial part of the project, requiring extensive feature engineering to prepare the data for modeling effectively.

Initially, it was determined that the project's objectives could be best met through both a categorical model or a linear regression model. This decision was critical because each type of model necessitates a distinct set of features. Consequently, the approach to handling the dataset diverged significantly based on the chosen model type.

For the linear regression model, the preparation process was relatively more straightforward. Key features relevant to this model were carefully selected from the original dataset. Dummy variables were created for categorical features such as location and team name, which are essential for handling non-numeric data within a regression framework. The target variable for the linear regression was defined as the statistical fields for each game, which were then extracted from the dataset. These steps collectively formed the foundation for developing the linear regression model, focusing on predicting game outcomes based on the defined features and target variable.

For the categorical models in this project, substantial effort was invested in constructing an effective feature class. Initially, the raw game statistics were not suitable for direct use in modeling; however, they were identified as crucial features. To make these statistics usable, aggregated metrics such as the mean, median, maximum, and minimum were computed for each team's performance in both home and away games, covering key statistics like points, rebounds, assists, and three-pointers.

A new table was then assembled, tailoring the data to reflect whether teams were playing at home or away, using the appropriate aggregated statistics for each game instance. This approach ensured that all pertinent statistics were utilized, enhancing the model's predictive power by providing a comprehensive view of team performances.

Additionally, in a similar fashion to the preparation for the linear regression model, dummy variables were created for categorical features such as game location and the names of the home and away teams.

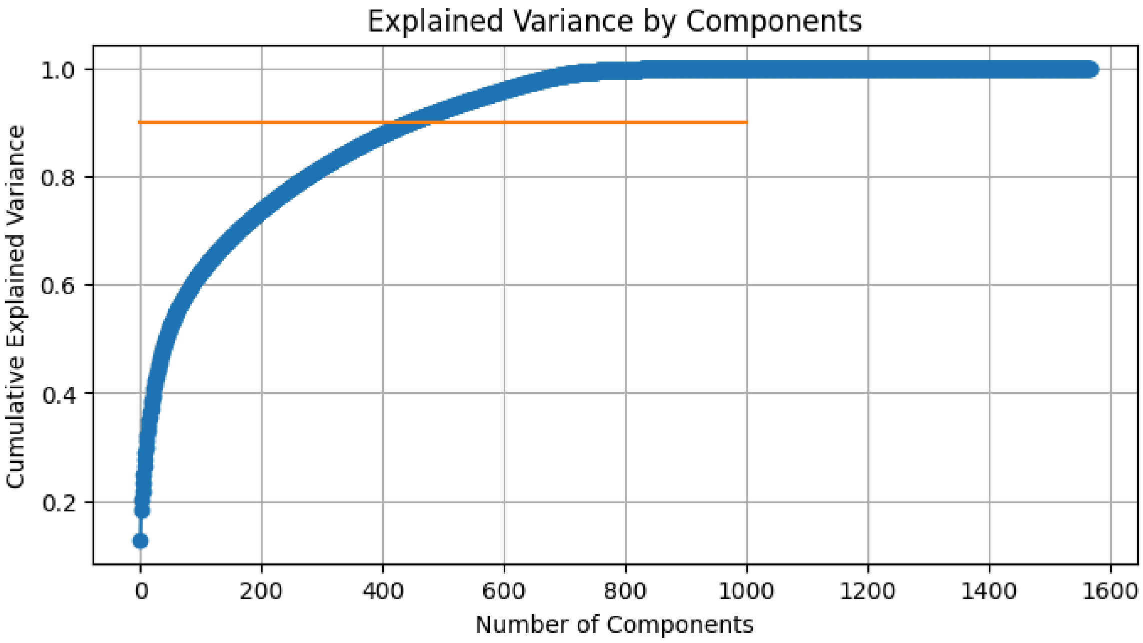
The target variable for the categorical models was constructed by calculating the point differential for each game and then assigning a binary outcome: 1 for a home team win and 0 for an away team win. This binary target variable serves as the basis for predicting game outcomes, using the structured and enriched data prepared through these steps.

Modeling

Starting with the categorical models, the feature class grew considerably due to the creation of dummy variables for team names and locations, along with the aggregation of extensive data. This resulted in a feature set of 1600 dimensions, posing challenges for model performance and interpretability. To mitigate this issue, Principal Component

Analysis (PCA) was identified as a suitable technique to reduce dimensionality while preserving the majority of the variance in the data.

To determine the optimal number of principal components to retain, the cumulative explained variance ratio was analyzed. The goal was to retain enough components to capture at least 90% of the variance in the original data. Through analysis, it was found that retaining 453 components would preserve approximately 90% of the variance, making it a suitable threshold for dimensionality reduction. This approach ensured that the data's essential characteristics were preserved while significantly reducing the number of features, thereby enhancing model performance and interpretability.



Both random forest and nearest neighbor models were applied to the data, with accuracy being the key scoring metric for evaluation. Given the nature of the data, neither false positives nor false negatives pose significant harm. However, it might be advantageous to prioritize recall, as over-predicting 0s could imply over-predicting upsets. Consequently, leaning towards over-predicting the expected winner to win may prove more beneficial. Nevertheless, overall accuracy remains the most important metric.

For the random forest model, grid search cross-validation was employed to test various values for the hyperparameters. While the model never quite surpassed 70% accuracy, several different hyperparameter combinations achieved a 69% accuracy. Examples of these combinations are provided below:

```
Estimator Number: 50
Max Depth: 20
Min Samples per Split: 5
Accuracy: 0.6940298507462687
Precision: 0.6953125
Recall: 0.978021978021978
F1 Score: 0.8127853881278538

Estimator Number: 50
Max Depth: 100
Min Samples per Split: 10
Accuracy: 0.6940298507462687
Precision: 0.6893939393939394
Recall: 1.0
F1 Score: 0.8161434977578476

Estimator Number: 200
Max Depth: 50
Min Samples per Split: 2
Accuracy: 0.6940298507462687
Precision: 0.6893939393939394
Recall: 1.0
F1 Score: 0.8161434977578476
```

```
Estimator Number: 50
Max Depth: 50
Min Samples per Split: 10
Accuracy: 0.6940298507462687
Precision: 0.6984126984126984
Recall: 0.967032967032967
F1 Score: 0.8110599078341014

Estimator Number: 100
Max Depth: 100
Min Samples per Split: 2
Accuracy: 0.6940298507462687
Precision: 0.6923076923076923
Recall: 0.989010989010989
F1 Score: 0.8144796380090498
```

For the nearest neighbor model, grid search cross-validation was also conducted, and the best result is outlined below. While it yielded a slightly less optimal model with a more balanced trade-off between precision and recall, it ultimately appeared to be a less effective model overall.

```
Number of Neighbors: 13
Accuracy: 0.6865671641791045
Precision: 0.6991869918699187
Recall: 0.945054945054945
F1 Score: 0.8037383177570093
```

Finally, the linear model was attempted using the feature set and target variables defined in the preprocessing step. Unfortunately, the linear regression model did not perform as well as the others. The resulting R-squared value was negative, indicating that this is actually a very poor model.

Consequently, the random forest model produced the best results. The table below showcases the outcomes from one of the top-performing random forest models.

Home Team	Away Team	Actual Result	Predicted result
USC	UNCA	1	1

BSU	WASH	0	1
OKST	FGCU	1	1
MARQ	HARV	1	1
UTAH	UCD	1	1
ND	HAMP	1	1
STAN	BYU	1	1
PSU	TEM	1	1
ORE	RID	1	1
WKU	BC	1	1
MTU	UVM	1	1
LOU	NKU	1	1
BAY	WAG	1	1
MSST	NEB	1	1
LSU	ULL	1	1
SMC	SELA	1	1
MARQ	ORE	1	1
BAY	MSST	0	1
LOU	MTU	1	1
ND	PSU	0	1
OKST	STAN	1	1
SMC	WASH	1	1
UTAH	LSU	1	1
USC	WKU	0	1
LOU	MSST	0	1
MARQ	PSU	0	1
SMC	UTAH	0	1
OKST	WKU	0	1

SF	UNT	1	1
LIB	UIC	0	0
UNCO	SHSU	1	1
UNT	SF	1	1
UNCO	UIC	1	1
UNT	SF	1	1
UCLA	SBON	0	1
NCCU	TXSO	0	1
ASU	SYR	0	1
RAD	LIU	1	1
GONZ	UNCG	1	1
MIA	L-IL	0	1
TENN	WRST	1	1
KU	PENN	1	1
DUKE	IONA	1	1
URI	OKLA	1	1
VILL	RAD	1	1
VT	ALA	0	1
TTU	SFA	1	1
FLA	SBON	1	1
ARIZ	BUFF	0	1
UK	DAV	1	1
OSU	SDST	1	1
HOU	SDSU	1	1
MICH	MONT	1	1
HALL	NCST	1	1
MIZZ	FSU	0	1

XAV	TXSO	1	1
CIN	GAST	1	1
UVA	UMBC	0	1
CREI	KSU	0	1
TXAM	PROV	1	1
UNC	LIP	1	1
MSU	BUCK	1	1
TCU	SYR	0	1
ARK	BUT	0	1
PUR	CSF	1	1
CLEM	NMSU	1	1
AUB	COFC	1	1
WICH	MRSH	0	1
WVU	MURR	1	1
NEV	TEX	1	1
UNC	TXAM	0	1
CIN	NEV	0	1
KU	CLEM	1	1
DUKE	SYR	1	1
KENT	FSU	1	1
PUR	TTU	0	1
MICH	FLA	1	1
VILL	TXTECH	1	1
KU	DUKE	1	1
MICH	LOY	1	1
VILL	KU	1	1
MICH	VILL	0	1

Future Research

This project initiates a preliminary exploration into predicting college basketball game winners, utilizing basic modeling techniques. To enhance and refine these models in the future, several strategies could be considered:

1. Refinement of PCA Strategy:

- Varying Explained Variance: Instead of adhering to a static explained variance threshold (e.g., 90%), experimenting with different levels might yield better performance. Lowering the variance threshold could simplify the model without significant loss of information.
- PCA with Cross Validation: Implementing a method akin to cross-validation to test various numbers of principal components in conjunction with random forest modeling could identify a more optimal configuration that balances complexity and performance.

2. Feature Selection:

- Selective Feature Inclusion: The initial model did not eliminate any game statistics when forming the feature set. Employing domain knowledge to discard less impactful features could focus the model on more predictive elements, potentially improving accuracy.
- Correlation-Based Feature Selection: Choosing features based on their correlation to game outcomes can streamline the model and emphasize variables that directly influence wins.

3. Incorporation of Recent Performance:

- Categorical Modeling with Recent Trends: Developing models that specifically incorporate recent performance metrics, such as wins in the last 5 or 10 games and scoring trends, can add a dimension of "recency effect." This approach might capture the dynamics of streaks or impacts of recent team changes such as injuries.

These proposed enhancements are just a starting point. The field is rich with potential for integrating more sophisticated statistical models and machine learning techniques to continually refine predictive accuracy.