# Predictive Analysis of Player Success Across Golf Courses

by Cole Whitelaw

## Problem Definition & Background:

Golf performance on the PGA Tour is influenced by a wide range of factors. Each shot demands varying degrees of power, precision, and technique, making every round unique. While golf statistics have been extensively studied, many unknowns remain when it comes to identifying the most critical variables that drive a player's success [1-3]. One of the biggest challenges in predicting performance is the variability of the courses themselves. Each tournament is held on a different course, each with distinct attributes and levels of difficulty [4]. This variability can significantly impact a golfer's preparation, as a player who excels on one course may struggle on another.

Our objective is to develop a model that accounts for course variability, allowing us to determine which player statistics, or combinations thereof, have the most significant impact on success. Additionally, we aim to explore whether this model can accurately predict player success (defined as frequency of top 10 finishes) in a given season by using a player's statistics for that given year. If our model proves to be accurate, we can determine which features or statistics have the largest impact by extracting feature importance from our model. By factoring in both course and player variables, we hope to provide deeper insights into the relationship between a golfer's skill set and their success on different types of courses, ultimately leading to more accurate predictions for future models.

If successful, this research could offer invaluable insights for players, coaches, and caddies, enabling them to better tailor their training and strategies to the specific challenges of each course [5]. However, due to the inherently unpredictable nature of golf, it's possible that no strong correlation exists between player statistics and performance across varying courses. Our goal is to conduct a thorough, in-depth analysis to test this hypothesis and uncover whether meaningful relationships can be identified between key player metrics and tournament outcomes.

## Description of Dataset:

Our dataset comprises player statistics from 2013 to 2024, collected via the SportsRadar API. It includes over 20 golfer-specific statistical metrics and 3,013 rows of historical player data. Course-specific statistics were obtained from DataGolf.com, covering 98 golf courses with over 30 variables per course. Courses with missing data were excluded. To ensure quality, we conducted background research on each data source.

Player leaderboard data, including the top 10 finishers from each tournament between 2013 and 2024, was scraped from pgatour.com. This dataset contains 6,139 rows, detailing player names, finishes, tournament names, and years. Additionally, tournament-course mapping data for events held between 2013 and 2024 was also extracted from pgatour.com to link tournaments to their respective courses.

## Methods & Experiment Setup:

To maximize the utility of our data, we employed a combination of supervised and unsupervised machine learning techniques. First, we grouped golf courses into clusters using K-means clustering. Given the high dimensionality of the course data, we applied a Principal Component Analysis (PCA) to reduce the number of variables, extracting two principal components for each course. These components served as inputs for the K-means algorithm, which classified the courses into two clusters.

After clustering, we created a dataset linking course names to their respective clusters and merged it with leaderboard and tournament-course data. This resulted in a comprehensive dataset containing player names, finishes, course names, course clusters, tournament names, and years. From this dataset, we calculated how many times each player finished in the top 10 in a given year for each course cluster, creating an outcome variable for our player statistics dataset. To generalize the data across seasons, players statistics were unique to each year.

We then used an XGBoostClassifier, a gradient-boosting algorithm, to predict player performance. XGBoostClassifier was selected for its robustness with tabular data, ability to handle missing values, feature importance insights, and high accuracy. Two models were trained, one for each course cluster. Players were classified based on their number of top-10 finishes in a given year, with ≥ 3 finishes forming class 0 and < 3 finishes forming class 1. The threshold of three finishes was chosen to balance class distribution while maintaining meaningful outcomes. Input variables for the models were player statistics for a given year, and the outcome variable was the player's class. Each section of cluster data was also split into training and testing sets using an 80-20 ratio with a random state of 42 for cross-validation.

To evaluate our entire model, we conducted backtesting using data from 2022 and 2023, training each model on prior years' data and predicting outcomes for the subsequent year. Backtesting for 2024 was not performed, as the season had not yet concluded at the time of this analysis. Predictions were made by calculating the average probability of a player belonging to the ≥ 3-cluster-wins class across the two models. Players with the highest average probabilities were identified as the most likely to achieve the most top-10 finishes for the given season. Finally, we assessed feature importance for each model to determine the player statistics most influential for success on specific course clusters, providing insights into the key skills required for different types of courses.

## Results:

When conducting our PCA we identified two primary components that characterize the attributes within the 98 golf courses. Principal Component 1 (PC1) represents courses with increased overall scoring difficulty, particularly on par-3 and par-4 holes, where players may struggle to score low due to factors such as complex green placements, protected greens, and potentially longer yardages. In contrast, Principal Component 2 (PC2) reflects courses that demand high shot precision, making it more challenging for players to stay within fairway boundaries and avoid penalty zones, thus requiring strategic play to avoid costly mistakes. Together, these components highlight key difficulty dimensions—scoring challenges from hole design and green protection (PC1) and the need for precise shot-making to navigate fairways and hazards (PC2).

Using the two principal components, we created an elbow plot to visualize the within-cluster sum of squares (inertia) for different values of k = n (not shown). After reviewing the elbow plot, we concluded that 2 clusters would be the optimal choice, as it balances the compactness of the clusters with simplicity, based on the clear elbow point at k = 2. After the number of clusters were chosen, we performed K-means clustering using the 2 principal components calculated with the course data. Results are displayed in Figure 1.

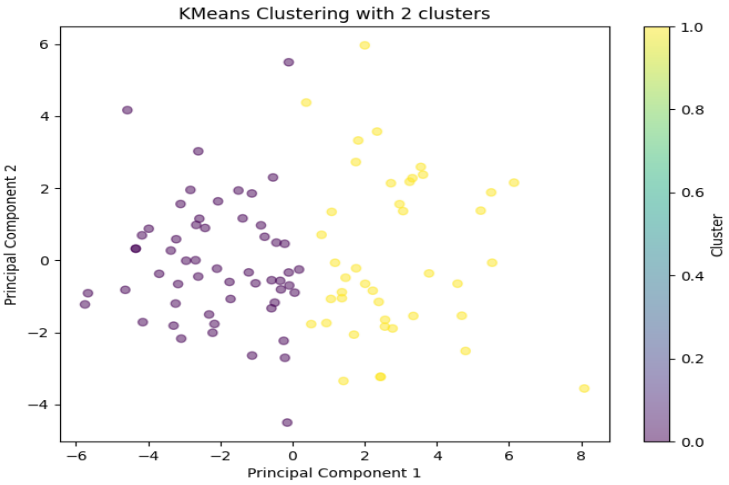

**Figure 1: Clusters graphed by Principal Component 1 and 2**

Cluster 0 (Yellow) represents the more difficult courses with higher PC1 values, indicating increased scoring difficulty, and variable PC2 values, which suggest a range of shot precision challenges. These courses are characterized by a combination of tough par-3 and par-4 designs, protected greens, and potentially narrow fairways. Cluster 1 (Purple) represents less difficult courses with lower PC1 values, meaning reduced scoring challenges, and variable PC2 values, reflecting a diverse range of precision requirements.

After assigning classes within our aggregated data set, we applied a XGBoostClassifier algorithm for each cluster. Confusion matrices and Heat Maps for the classification report were created for both clusters and represented in Figure 2 and Figure 3.

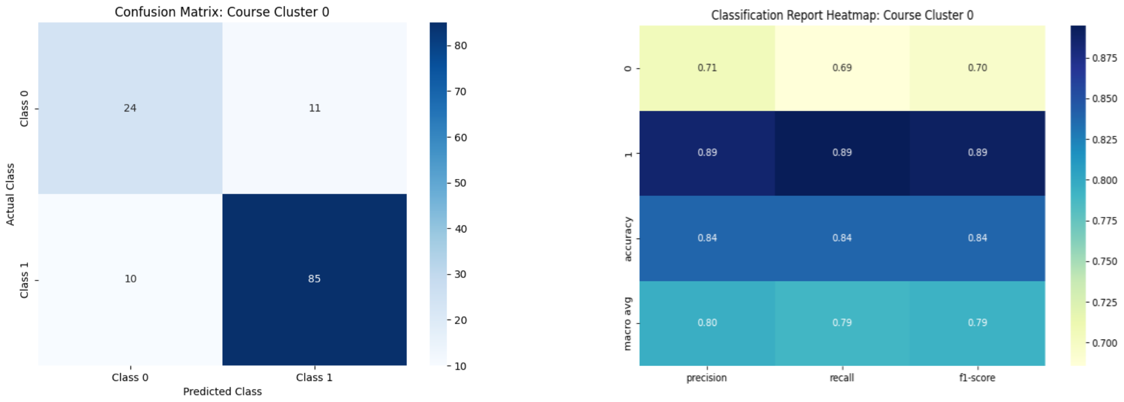

**Figure 2: Confusion Matrix and Classification Report Heatmap for XGBoostClassifier: Course Cluster 0**

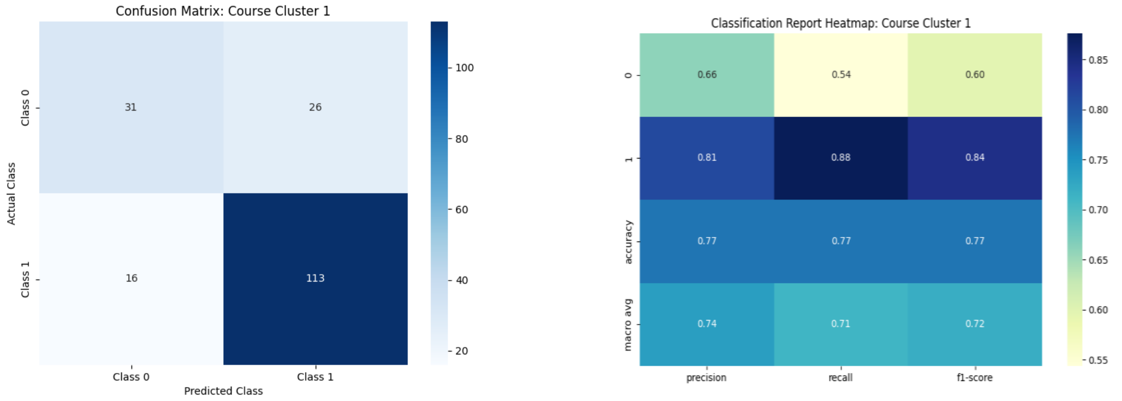

**Figure 3: Confusion Matrix and Classification Report Heatmap for XGBoostClassifier: Course Cluster 1**

Model reports were promising, especially considering the unpredictable nature of golf, with accuracies of 0.84 for course cluster 0 and 0.77 for course cluster 1. Although these reports are based on binary classification, we used the probability of being classified as class 0 to estimate the likelihood that a player will achieve more than 3 top 10 finishes on a specific cluster, in a given season.

Our top 10 predictions were determined by averaging the probability of being in class 0 from both models. Players with the highest average probability were identified as the predicted top performers with the most top 10 finishes. Prediction results for 2022 and 2023 are shown in Figure 4, with the model achieving strong accuracy: 70% for 2022 and 80% for 2023. Notably, the model tends to generally assign higher probabilities to players with more wins and lower probabilities to those with fewer wins. Feature importance for both course clusters, trained through 2023, was subsequently extracted and is presented in Figure 5.

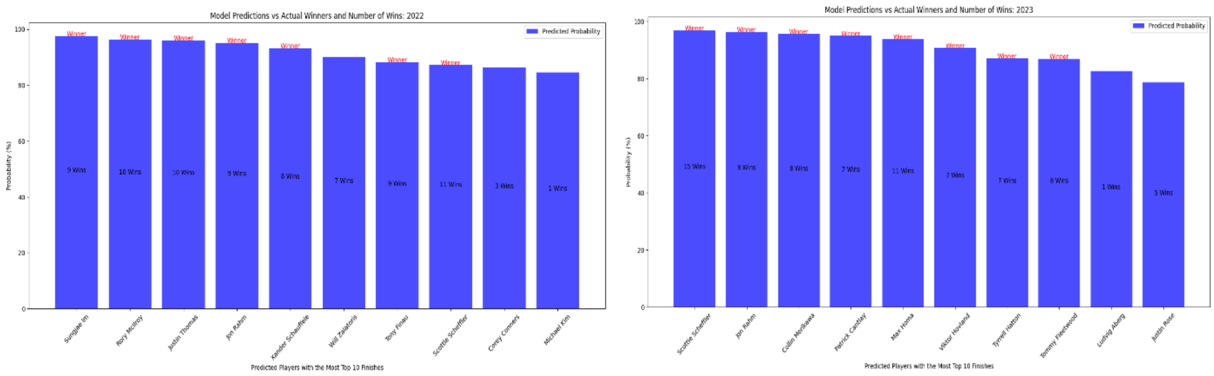

**Figure 4: Model prediction results for both years**

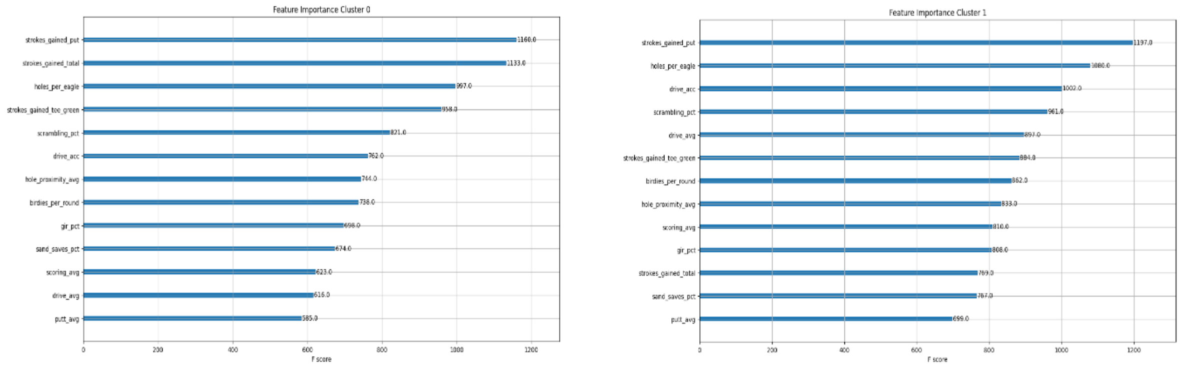

**Figure 5: Feature Importance for 2023**

## Observations & Conclusion:

This analysis highlights the key insights gained from using our model to identify predictors of top 10 finishes in a given year. The features presented in Figure 5 reveal the most influential factors in determining success, offering valuable guidance for players, coaches, and caddies. Notably, the key predictors differed between the two course clusters, reflecting varying requirements for success.

On the more difficult courses (Cluster 0), strokes gained variables were among the most heavily weighted, with strokes gained putting, strokes gained total, and strokes gained tee-to-green ranking highly, alongside holes per eagle. These results suggest that a combination of scoring efficiency, shot precision, and the ability to capitalize on eagle opportunities is critical for performing well on these challenging courses. In contrast, on the more variable courses (Cluster 1), strokes gained putting emerged as the most significant factor, followed by holes per eagle, driving accuracy, and scrambling percentage. This indicates that success on these courses may rely more heavily on short-game precision, consistency, and recovery skills.

Despite the promising results, several limitations must be noted. One key issue was the course groupings; while clusters showed moderate differences, some courses in separate clusters displayed similarities, potentially reducing precision. Additionally, the analysis only included tournaments on courses with available data, excluding others that contributed to players' overall season statistics. This may have limited the scope of the performance picture.

Another limitation was the reliance on season-long averages for predictions. Player statistics naturally fluctuate after each tournament and vary before specific events, making real-time or tournament-specific data potentially more accurate for identifying variable importance. Additionally, training our model using end-of-season statistics to predict player success for a year that has already concluded reduces the model's predictive power and limits its applicability to real-time or future forecasting scenarios. This model may be more applicable for mid-year or three-quarter-year data, where players have had sufficient time to develop a clear statistical profile. However, since the primary objective of this analysis was feature extraction rather than prediction, this limitation is less critical in this context. Nonetheless, incorporating real-time or tournament-specific data could be a valuable future direction for enhancing the analysis. Unfortunately, we did not have access to this data at the time of the study.

Prediction accuracy was used to identify patterns in the statistics of successful players, ultimately highlighting the most significant features that differentiate top performers. This analysis demonstrates that it is possible to predict top 10 finishes using players' year-average statistics while pinpointing key predictors that set successful players apart. While further investigation is needed to understand why certain metrics, such as holes per eagle, were weighted so heavily, these findings provide valuable insights for players, coaches, and caddies. They may also serve as valuable guidance for developing preparation strategies tailored to specific course characteristics.

Additionally, these features may contribute to future models for predicting player success on specific courses, emphasizing the varying skill requirements across different course types. Domain expertise is necessary to better understand why certain features were more heavily weighted, but this analysis offers an unbiased, data-driven approach to identifying critical factors for success. Overall, this approach presents a novel perspective for guiding player development and preparation strategies for different types of golf courses on the PGA tour.

**References:**

[1] D. S. Belkin, B. Gansneder, M. Pickens, R. J. Rotella, and D. Striegel, “Predictability and Stability of Professional Golf Association Tour Statistics,” Perceptual and Motor Skills, vol. 78, no. 3_suppl, pp. 1275–1280, Jun. 1994, doi: https://doi.org/10.2466/pms.1994.78.3c.1275.

[2] T. N. Dorsel and R. J. Rotunda, “Low Scores, Top 10 Finishes, and Big Money: An Analysis of Professional Golf Association Tour Statistics and How These Relate to Overall Performance,” Perceptual and Motor Skills, vol. 92, no. 2, pp. 575–585, Apr. 2001, doi: https://doi.org/10.2466/pms.2001.92.2.575.

[3] R. J. Quinn, “Exploring Correlation Coefficients with Golf Statistics,” Teaching Statistics, vol. 28, no. 1, pp. 10–13, Feb. 2006, doi: https://doi.org/10.1111/j.1467-9639.2006.00229.x. y6JIwWS65ispZ6 (accessed Oct. 08, 2024).

[4] “Golf Stat and Records | PGA TOUR,” Pgatour.com, 2024. https://www.pgatour.com/stats/course/toughest-course (accessed Oct. 08, 2024).

[5] S. Deakins, “Why PGA Tour Golfers Turn to Statisticians for a Lower Score,” Under Par Performance Golf, Dec. 20, 2022. https://underpargolfapp.com/blogs/playing-strategies-for-golfers/why-pga-tour-golfers-turn-to-st atisticians-for-a-lower-score?srsltid=AfmBOopCzljRYZqRCwyWXcMFwtyVOygYf3ZuicCb