Golf Performance Analysis and Prediction

TDDE64 Sports Analytics

Filip Malm-Bägén

May 28, 2024

Contents

1 Introduction

In the field of sports analytics, golf holds a unique position due to its complexity and the multitude of factors affecting performance. This report focuses on the complex world of golf performance analysis, aiming to predict outcomes and enhance player strategies using advanced data science techniques. The significance of understanding and predicting golf performance is underscored by the sport's competitive nature and the financial implications tied to player success and event outcomes.

Traditionally, golf analytics has focused on straightforward statistical data such as driving distance, putting accuracy, and greens in regulation. However, with the advent of machine learning and big data, the scope for analysis has expanded dramatically [1]. This project builds upon these foundations by employing sophisticated algorithms to not only understand past performance but also predict future results. Such predictive analytics serve as a crucial tool for players and coaches alike, offering insights that are not immediately apparent from basic statistics.

Previous solutions in golf performance analysis have primarily utilized regression models and time-series analysis to predict player performance based on historical data. An early example of this is the development of the ShotLink system, which evolved from a simple scoring system to a sophisticated data collection and analysis tool used widely across the PGA Tour [2]. More recent approaches have integrated machine learning techniques, including decision trees and neural networks, to provide more accurate predictions and uncover deeper patterns in player behavior and game dynamics. These advancements have allowed for a more granular analysis of performance metrics, offering deeper insights that go beyond traditional statistical methods [3].

This report is about synthesizing traditional statistical techniques with modern data science to offer a comprehensive analysis of golf performance. The aim is to to contribute to the field of sports analytics, providing stakeholders with tools that can refine training programs, improve player selection, and enhance game strategy.

2 Background

The domain of golf analytics transcends mere statistical record-keeping to embrace a multifaceted exploration of how data can enhance player performance and strategy formulation. In professional golf, particularly on the PGA Tour, players, coaches, and analysts seek to understand and optimize every aspect of play—from swing mechanics to course strategy—through the lens of data [4].

2.1 Importance of Performance Metrics in Golf

Golf is a sport where the margin between winning and the second-best can be extraordinarily slim. A single stroke can differentiate the winner of a major tournament from the rest of the field, making precision and strategy paramount. In this context, detailed performance metrics such as scoring average, driving accuracy, strokes gained, and other statistical measures become invaluable. They allow for the quantification of a player's performance relative to the field and can highlight strengths to be leveraged and weaknesses to be addressed.

1. **Scoring Average**: This is perhaps the most direct indicator of a golfer's performance, representing the average number of strokes per round. It is a critical metric because it aggregates a player's performance across different courses and conditions, providing a consistent basis for comparison.
2. **Driving Distance and Accuracy**: These metrics assess how far and accurately players can hit the ball off the tee. While longer drives can significantly advantage a player by reducing the length of subsequent shots, accuracy ensures that the ball lands in favorable positions, thus avoiding hazards and rough terrain.
3. **Strokes Gained**: Developed as a more sophisticated measure of a player's performance, strokes gained metrics compare a player's performance to a professional baseline across different aspects of the game: tee-to-green, putting, and overall. These metrics have revolutionized performance analysis by providing more granular insights into where a player is gaining or losing strokes against the field.
4. **Fairway and Green Regulation Percentages**: These are critical for strategy as they measure a player's ability to hit fairways and greens in regulation, which are key determinants of scoring opportunities.

2.2 Evolution of Data Analytics in Golf

The advent of machine learning and advanced statistical techniques has propelled golf analytics into a new era. Traditional methods often relied on basic descriptive statistics to understand past performances [3]. However, contemporary approaches, as explored in this report, employ predictive models that not only analyze past data but also predict future outcomes. These models can identify patterns and insights that are not immediately apparent from raw data alone.

3 Algorithms

In the pursuit of extracting meaningful insights from the PGA Tour dataset, the approach integrated several sophisticated algorithms, specifically chosen for their applicability to the complexity and characteristics of the data involved. This section provides an overview of the primary algorithms employed—Random Forest and Support Vector Machines (SVM)—and rationalizes their selection based on their suitability for dealing with the challenges inherent in golf performance data.

3.1 Random Forest

Random Forest is an ensemble learning technique that builds upon the simplicity of decision trees by creating multiple trees and merging their outputs to get a more accurate and stable prediction. It does not assume linear relationships between features, which is crucial given the non-linear correlations often found in golf data, such as between driving distance and scoring. Random Forest also offers excellent capabilities for handling large data sets with numerous variables, allowing us to gauge the most influential features affecting a player's performance. Moreover, it reduces the risk of overfitting, a common problem with single decision trees, by averaging multiple trees that individually might overfit the data [5].

3.2 Support Vector Machines

Support Vector Machines (SVM) offer a robust way to classify data by finding the best hyperplane that separates data into classes. For the data, the ability of SVMs to operate effectively in high-dimensional spaces is invaluable, especially after the application of polynomial feature engineering that expands the feature space significantly. The SVM's capacity to maximize the margin between class boundaries ensures that the model has a strong generalization, reducing the risk of overfitting. Furthermore, the flexibility offered by the kernel trick allows SVMs to adapt to the nonlinear aspects of the data, accommodating the complex relationships among features [6].

3.3 Selection Rationale

The selection of Random Forest and SVM was driven by the specific challenges posed by the golf performance data, characterized by its complexity and the intricate relationships within. The robust generalization capabilities of these algorithms make them particularly apt for this analysis, ensuring that the models are not only accurate but also reliable across various scenarios. The combination of Random Forest and SVM harnesses their complementary strengths—feature handling, robustness against overfitting, and flexibility in modeling nonlinear relationships—making them ideal for the objective of predicting golf performances with high precision.

4 Test Data and Test Setup

To evaluate the effectiveness of the predictive models—Random Forest and Support Vector Machines (SVM)—it was imperative to establish a robust testing framework. This involved careful selection and preparation of the test data, as well as designing a test setup that accurately measures the models' performance in realistic scenarios. This section describes the test data and the setup used to evaluate the algorithms.

4.1 Dataset Description

The dataset used for this analysis consists of PGA Tour player data from 2010 to 2017, containing 2044 rows and 15 features. These features capture various aspects of player performance and achievements throughout the season. Below is a brief overview of the key features:

- **Name**: Name of the golfer.
- **Country**: Home country of the player.

- **Year**: The year of the tournament.
- **Rounds**: Number of PGA tour rounds played in that year by that particular player.
- **Scoring**: Average score per round played for that year.
- **Driving Distance**: Average drive distance calculated from two holes per round, chosen to negate wind effects.
- **FWY_Percentage**: The percentage of time a tee shot comes to rest in the fairway.
- **GIR_Percentage**: Green in Regulation (GIR) percentage, indicating how often a player's ball touches the green after the GIR stroke.
- **SG_P (Strokes Gained Putting)**: Strokes gained putting, calculated by comparing a player's putts from a specific distance to a baseline.
- **SG_TTG (Strokes Gained Tee to Green)**: Average per round of how a player's strokes compare to the field average, excluding strokes gained putting.
- **SG_T (Strokes Gained Total)**: The per round average of the number of strokes the player was better or worse than the field average.
- **Points**: FedExCup points earned.
- **TOP_10**: Yearly count of a player's top 10 finishes.
- **Money**: The amount of money the player earned in that year.
- **1ST**: The number of wins the player had in that year (target feature).

4.2 Test Data Selection

The test data comprised a subset of the PGA Tour dataset, specifically isolated to ensure no overlap with the training dataset. The dataset contained 2044 rows of player data with 15 features each from PGA Tour players through the years 2010-2017. The data derived from PGAtour.com [7]. For a realistic assessment, data from the 2017 season was selected as the test set. This approach allows the models to be evaluated on the most recent data, reflecting their ability to generalize to new, unseen data. The 2017 season was chosen because it represents the culmination of the trends and patterns observed over the previous years, and testing on this data provides insights into the models' predictive power under the latest competitive conditions. The features "Country" and "Name" were dismissed from the training set due to the fixed property. 17% of players has won at least one competition, while the rest has not.

4.3 Test Setup

The testing setup was designed to rigorously evaluate the predictive accuracy of the models and their ability to generalize across different metrics of golf performance. The primary metrics for evaluation were accuracy, precision, recall, and the Receiver Operating Characteristic Area Under the Curve (ROC AUC). These metrics were chosen because they provide a comprehensive view of model performance, especially in the context of an imbalanced dataset like ours where the proportion of tournament winners to non-winners is skewed.

- **Accuracy** measures the overall correctness of the model in predicting winner and non-winner classes.
- **Precision** reflects the accuracy of positive predictions, which is crucial for minimizing false positives in winner predictions.
- **Recall** indicates the model's ability to detect all actual winners, an essential factor in sports analytics where missing a potential winner is more detrimental than falsely highlighting a non-winner.
- **ROC AUC** provides a single measure of performance at various threshold settings, which is particularly useful for imbalanced datasets. It helps in understanding how well the model can distinguish between the two classes under different discrimination thresholds.

For the test execution, data was first preprocessed in the same manner as the training set to maintain consistency. This included scaling of features and application of the same feature engineering steps used to prepare the training data. Specifically, the following steps were performed:

1. **Feature Scaling**: The test data was scaled using the StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1, matching the preprocessing applied to the training data.

2. **Feature Engineering**: The same domain features (such as *MONEY_PER_ROUND* and *SG_SUM*) and polynomial features were generated for the test data as were done for the training data. This ensured that the test data was in the same format and feature space as the training data.

3. **Data Splitting**: The test data was isolated by filtering out the 2017 season from the dataset, providing a robust evaluation set.

Models were then applied to the test data, and predictions were generated for each model. The specific models tested included the baseline model using the original features, the domain feature model with additional engineered features, and the polynomial feature model incorporating non-linear interactions and higher-order terms.

4.4 Model Evaluation

Each model's predictions were compared against actual outcomes to compute the evaluation metrics mentioned above. The results of these metrics provided the necessary insights into each model's performance, highlighting strengths and weaknesses in their ability to predict golf performance outcomes.

This methodical approach to testing ensures that the conclusions drawn from this study are well-supported by empirical evidence, thus providing reliable guidance for future analytics applications in golf performance prediction.

5 Test Results

The evaluation of the Random Forest and Support Vector Machines (SVM) models on the 2017 PGA Tour dataset yielded insightful results. This section presents the key performance metrics and feature importance for both models.

While the dataset was examined, several interesting key interests was found. Figure 1 shows how the scoring average through the years. The Scoring Average for players on the PGA Tour has not varied much over the 8 years in the dataset. The highest scoring averages came in 2010 and 2016 at about 71 strokes per round, and the lowest came in 2014 with roughly 70.84 strokes per round. Figure 2 shows that tournament winners tend to drive the ball further than non-tournament winners.

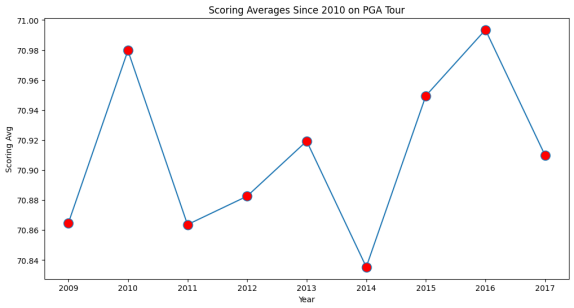

Figure 1: Scoring Average by Year

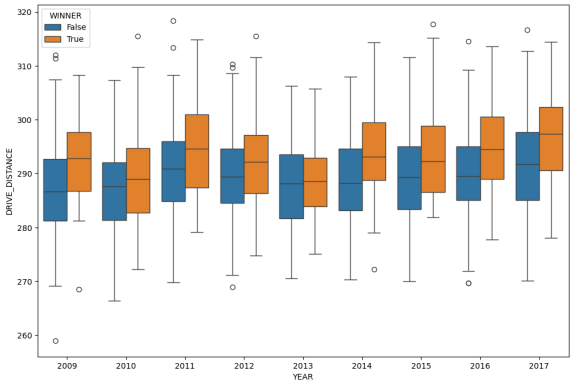

Figure 2: Boxplot of Drive Distance by Year and Winner Status

Figure 3 displays that the average distance is negatively correlated with the percentage of fairways hit. This is expected as players who hit the ball further are more likely to miss the fairway.

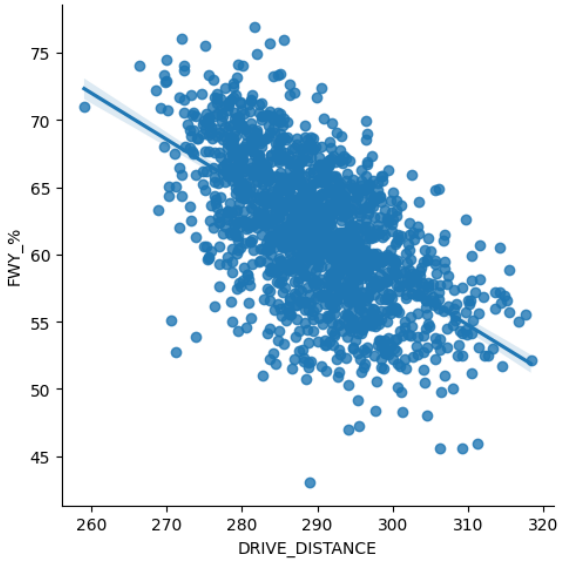

Figure 3: Correlation between Driving Distance and Percentage of Fairway Hits

The correlation of each feature in the feature set can be seen in the correlation matrix heatmat in Figure 4.

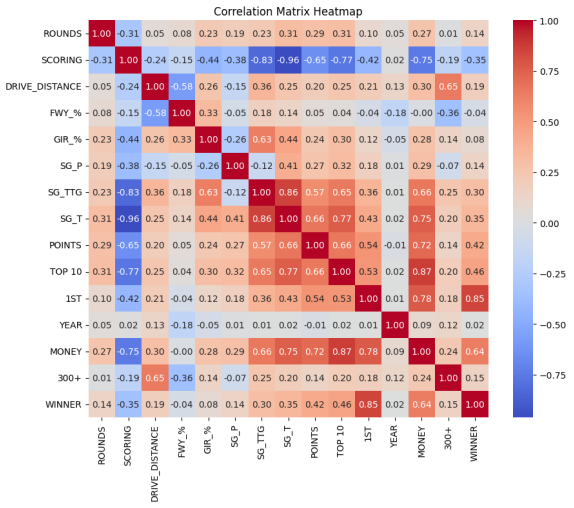

Figure 4: Correlation Matrix Heatmap

From this data, the most prominent features which yields to winning at least one tournament can be seen in table 1.

Feature	Imp.	Abs. Imp.
MONEY	3.312462	3.312462
TOP 10	-1.235560	1.235560
SG_T	-0.384802	0.384802
SG_TTG	-0.307805	0.307805
YEAR	-0.265861	0.265861
POINTS	0.249334	0.249334
SCORING	0.197283	0.197283
SG_P	-0.143385	0.143385
ROUNDS	0.109266	0.109266
300+	0.053877	0.053877
DRIVE_DIST.	-0.030821	0.030821
FWY_%	-0.019446	0.019446
GIR_%	-0.000605	0.000605

Table 1: Feature Importances with their Absolute Values

5.1 Random Forest Results

The Random Forest model demonstrated robust performance across all evaluated metrics. Three different feature sets were evaluated: Baseline Model, Domain Feature Model, and Polynomial Feature Model. The results are summarized in Table 2.

Model	ROC AUC Score
Baseline Model	0.82
Domain Feature Model	0.84
Pol. Feature Model	0.86

Table 2: ROC AUC Scores for Random Forest Models

The Polynomial Feature Model achieved the highest ROC AUC score of 0.8577, indicating strong discriminative ability between the winner and non-winner classes. The performance metrics for each model are detailed below:

Metric	Baseline Model
Accuracy	90%
Precision (True)	0.73
Recall (True)	0.70
F1-score (True)	0.72
ROC AUC Score	0.82

Table 3: Performance Metrics for Baseline Model

Metric	Domain Feature Model
Accuracy	91%
Precision (True)	0.75
Recall (True)	0.73
F1-score (True)	0.74
ROC AUC Score	0.84

Table 4: Performance Metrics for Domain Feature Model

Metric	Poly. Feature Model
Accuracy	91%
Precision (True)	0.74
Recall (True)	0.77
F1-score (True)	0.76
ROC AUC Score	0.86

Table 5: Performance Metrics for Polynomial Feature Model

5.2 Support Vector Machines (SVM) Results

The SVM model, equipped with a linear kernel, demonstrated strong performance across all evaluated metrics. The results are summarized in the tables below.

Metric	SVM Model
Accuracy	92%
Precision (True)	0.82
Recall (True)	0.73
F1-score (True)	0.78

Table 6: Performance Metrics for SVM Model on Test Data

Metric	SVM Model
Accuracy	91%
Precision (True)	0.76
Recall (True)	0.76
F1-score (True)	0.76

Table 7: Performance Metrics for SVM Model

Metric	CV SVM Model
Average Accuracy	0.91
SD of Accuracy	0.0098

Table 8: CV Results for SVM Model

The SVM model achieved an overall accuracy of 92% on the test data, indicating a high rate of correct classifications. The precision, recall, and F1-score for the winner class (True) were 0.82, 0.73, and 0.78, respectively. On the 2017 data, the SVM model also performed well, with an accuracy of 91% and consistent precision, recall, and F1-score values for the winner class. The cross-validation results further support the model's robustness, with an average accuracy of 0.91 and a low standard deviation of 0.0098.

6 Discussion

The purpose of this study was to identify the most influential features that lead to winning in PGA Tour golf tournaments and to evaluate the effectiveness of machine learning models in predicting winners based on these features. This discussion section analyzes the results of the Random Forest and Support Vector Machines (SVM) models, compares them with previous studies, and highlights key insights.

6.1 Analysis of Results

The results from both the Random Forest and SVM models indicate a strong ability to predict tournament winners, with accuracies exceeding 90%. The performance metrics, including precision, recall, and F1-score, further demonstrate the robustness of these models.

6.1.1 Random Forest Model

The Random Forest model showed an overall strong performance across all feature sets. The Polynomial Feature Model achieved the highest ROC AUC score of 0.86, indicating its superior discriminative ability. The key performance metrics for the Polynomial Feature Model were

- **Accuracy**: 91%
- **Precision (True)**: 0.74
- **Recall (True)**: 0.77
- **F1-score (True)**: 0.76

These results suggest that incorporating polynomial features improves the model's predictive power by capturing more complex relationships within the data.

6.1.2 Support Vector Machines Model

The SVM model also demonstrated robust performance, with an overall accuracy of 92% on the test data. The precision, recall, and F1-score for predicting winners were:

- **Precision (True)**: 0.82
- **Recall (True)**: 0.73
- **F1-score (True)**: 0.78

The SVM model's cross-validation results further support its robustness, with an average accuracy of 0.91 and a low standard deviation of 0.0098, indicating consistent performance across different subsets of the data.

6.2 Key Insights and Implications

The feature importance analysis revealed several key insights:

- **Money**: This feature had the highest positive importance, indicating that higher earnings are strongly associated with winning. However, one can argue that winning leads to a high amount of money and not the other way around.

- **Top 10 Finishes**: Despite having a negative coefficient, this feature was still among the most influential. This negative coefficient suggests that while players who frequently finish in the top 10 are often high performers, consistently finishing in the top 10 without winning may not directly predict winning. This could imply that occasional higher placements (i.e., winning) are more critical than consistent near-wins.

- **Strokes Gained: Tee-to-Green (SG_TTG)** and **Strokes Gained: Total (SG_T)**: These were also significant, underscoring the importance of overall skill and consistency in different aspects of the game.

- **Green in Regulation Percentage (GIR_%)** : This was the least influential feature with a very low absolute importance value. While hitting greens in regulation is fundamental in golf, it may not strongly predict winning compared to metrics like strokes gained or earnings. Since the dataset comprises tour-level players, most are proficient at hitting greens, reducing this feature's variability and predictive power.

These findings have practical implications for players, coaches, and analysts. Understanding the key performance metrics that lead to winning can help in developing targeted training programs and strategies to improve player performance. However, it turned out that the most influential features leading to a win were not directly related to the game itself, but rather to the outcomes of the tournament, such as "Money" and "Top 10 Finishes."

A more interesting approach would be to run the algorithms without these post-tournament features. This could better reflect the true importance of each feature in relation to one another and provide trainers and players with insights into the most critical aspects of performance that can be directly improved upon. Given that "Money" and "Top 10 Finishes" had such significant importance compared to other features, excluding them would likely reduce accuracy. To compensate, a larger dataset with more players and additional features would be necessary.

6.3 Comparison with Previous Studies

Previous studies in golf analytics have primarily utilized regression models and time-series analysis to predict player performance based on historical data. More recent approaches have integrated machine learning techniques, including decision trees and neural networks, to provide more accurate predictions and uncover deeper patterns in player behavior and game dynamics.

The findings are consistent with these trends, demonstrating that machine learning models, particularly ensemble methods like Random Forests and linear classifiers like SVMs, can effectively predict tournament outcomes. The integration of advanced metrics such as 'Strokes Gained' further enhances the predictive power of these models, aligning with the literature that highlights the importance of these metrics in modern golf performance analysis.

6.4 Conclusion

This study demonstrated the effectiveness of Random Forest and SVM models in predicting PGA Tour tournament winners based on historical performance metrics. The results highlight the importance of key features such as money earned, top 10 finishes, and strokes gained metrics. However, it is worth noting that the most influential features determining the outcome of a tournament were determined after the winning. Either way, the findings provide valuable insights for enhancing player performance and developing strategic training programs. Further research could build on these results by incorporating additional features, removing withdraw other features and exploring advanced machine learning techniques.

7 Future

The application of Random Forest and Support Vector Machines (SVM) in golf performance analysis enhances methodologies and paves the way for improvements. Future enhancements could involve fine-tuning model parameters, developing

complex features, and integrating ensemble methods to improve predictive accuracy.

Expanding analytics to include biometric sensors and weather conditions could offer a holistic view of performance influencers. Adapting models for real-time analytics during tournaments could provide immediate strategic insights, while tailoring models to individual players would enable personalized training and recommendations.

Applying these advanced analytics methods to youth and amateur sports could democratize access to professional-level insights, transforming training methodologies. Injury prediction and prevention is another area ripe for exploration, where analytics could impact player longevity and well-being.

Advancements in AI, deep learning, IoT, and wearable technology could unlock deeper insights from complex datasets. By building on current models and embracing new technologies, golf performance analytics is set to significantly broaden, enhancing strategic and training capabilities across all sport levels.

References

[1] Guillermo Martinez Arastey. *THE INCREASING PRESENCE OF DATA ANALYTICS IN GOLF*. URL: https://www.sportperformanceanalysis.com/article/increasing-presence-of-data-analytics-in-golf (visited on 05/22/2024).

[2] *TURN DATA INTO INFORMATION, INFORMATION INTO KNOWLEDGE, KNOWLEDGE INTO ENTERTAINMENT*. URL: https://shotlink.com/about/history (visited on 05/23/2024).

[3] Mark Lamport-Stokes. *THE INCREASING PRESENCE OF DATA ANALYTICS IN GOLF*. URL: https://www.firstcallgolf.com/features/feature/2024-05-02/how-ai-may-impact-golf-in-the-near-future (visited on 05/22/2024).

[4] Spencer Hong. *From the Fairway to the Spreadsheet: The Rise of Golf Analytics and Data-Driven Strategies*. URL: https://www.alteryx.com/blog/from-the-fairway-to-the-spreadsheet-the-rise-of-golf-analytics-and-data-driven-strategies#:~:text=The%20Importance%20of%20Data%20%26%20Analytics%20for%20Golfers&text=Data%20analytics%20can%20be%20used,are%20they%20from%20the%20green%3F (visited on 05/22/2024).

[5] Niklas Donges. *Random Forest: A Complete Guide for Machine Learning*. URL: https://builtin.com/data-science/random-forest-algorithm# (visited on 05/22/2024).

[6] Anshul Saini. *Guide on Support Vector Machine (SVM) Algorithm*. URL: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/ (visited on 05/22/2024).

[7] URL: PGAtour.com (visited on 05/22/2024).