Zach Printz

Golf Analytics

Professor Jesse Spencer-Smith Independent Study

# Predicting Golf Outcomes with Machine Learning

## Introduction

Predictive modeling in sports analytics has gained traction over the past decade with advancements in machine learning and data collection technologies. This study focuses on developing and comparing two types of models: a Random Forest model and an in-context learning model. The objective is to predict the outcomes of golf tournaments, leveraging player performance and tournament data from the PGA Tour website for the years 2023 and 2024. This paper details the methods employed, results obtained, and discussions on the effectiveness of the approaches.

## Methods

### Data Collection and Preparation

The foundation of predictive modeling in sports relies heavily on the availability and quality of data. For this study, comprehensive datasets for the years 2023 and 2024 were collected from the PGA Tour website, which provides detailed statistics on player performances and tournament outcomes. The data collection process involved the development of custom web scraping scripts, initiated by thoroughly inspecting the website's structure using browser developer tools. Two

specific scripts were crafted: one to extract detailed Strokes Gained statistics and another to gather leaderboard information including finishing positions and tournament scores. As the PGA Tour website underwent slight formatting changes in 2024, minor adjustments were made to the scripts to ensure accurate data extraction. Following collection, the data was wrangled and verified to address any inconsistencies or anomalies, ensuring the datasets were clean, structured, and ready for analysis.

**Selecting the Optimal Traditional Machine Learning Model**

The choice of the Random Forest model was grounded in a comparative analysis of various traditional machine learning models including LSTM, XGBoost, SVR, and neural networks. This decision was driven by Random Forest's superior performance in preliminary tests. The original models were curated with five independent variables: Strokes Gained: Total, Strokes Gained: Off the Tee, Strokes Gained: Approach, Strokes Gained: Around the Green, and Strokes Gained: Putting. These five features attempted to predict the dependent variable, finishing position, in each tournament.

In the initial comparisons, the Random Forest model demonstrated robust predictive accuracy with a Mean Squared Error (MSE) of 94.24 and an R-squared value of 0.78, outperforming all other models in terms of both error minimization and the proportion of variance explained in the dependent variable. These metrics indicated not only the model's effectiveness in fitting the training data but also its potential to generalize well to unseen data.

**Quantitative Random Forest Model Development**

After settling on a Random Forest model to carry out the predictions, the next step was the fine tuning of the model. This was a multi-step process, involving feature engineering, model training, and validation.

The effectiveness of the Random Forest model largely depended on the quality and relevance of the features used. Two novel features were engineered on top of the original five Strokes Gained categories to enhance the model's predictive capability:

- **Field Strength:** This feature quantified the overall competitiveness of each tournament by calculating the average 2023 finishing positions of all players in the field. This metric provided a comparative measure of difficulty for each tournament, reflecting the caliber of participants.
- **Recent Performance:** To capture the current form of each player, this feature averaged the finishing positions of a player's last eight tournaments. This moving average helped to incorporate the dynamic aspect of a player's current form into the model.

The model's performance improved with the addition of the above features, with the MSE dropping to 81.16 and the R-Squared rising to 0.81. The Random Forest model was trained using the 2023 data, which was segmented into training and testing subsets based on a 50/50 time-based split. This method ensured that the training data did not include any information from the future that could lead to data leakage and overfitting. The model utilized several hyperparameters, such as the number of trees and the depth of each tree, which were optimized through cross-validation to improve the model's accuracy and generalizability.

**Qualitative Random Forest Model Development**

Following the quantitative analysis, the study progressed to a qualitative Random Forest model to explore if data simplification through categorization could enhance predictive performance. To do so, the independent variables were converted from numeric values to categorical ones. The conversion of numeric data into categorical data involved segmenting each of the original quantitative features into the following quintiles.

- **Worst:** Bottom 20% of performances.
- **Bad:** 20-40% range.
- **Medium:** 40-60% range.
- **Good:** 60-80% range.
- **Best:** Top 20% of performances.

This categorization was applied uniformly across all Strokes Gained categories and additional features such as Field Strength and Recent Performance. It transformed the continuous variables into a set of discrete, ordinal categories. This method was anticipated to reduce the noise and variance within the data, focusing the model's learning on more distinct and arguably more interpretable patterns.

The qualitative Random Forest model used the same 2023 dataset, now transformed into categorical bands. The training and validation split remained identical to that of the quantitative model, as did the hyperparameters.

**Quantitative In-Context Learning Model Development**

After developing the Random Forest models, the next stage of the project was to explore the potential of in-context learning. This was first done using quantitative data. This approach aimed to leverage the power of a large language model (LLM), hoping to draw on its ability to utilize

more complex modeling techniques. ChatGPT 4 was the primary LLM used for the development of the quantitative in-context learning model. The initial plan was to use the full dataset for training, but due to capacity limitations, a reduction in data usage was necessary. Consequently, only 50% of the available data from the 2023 dataset was used in the training of the in-context learning model. Like the Random Forest models, the data for the in-context learning approach included the five Strokes Gained categories and the two engineered features: Field Strength and Recent Performance.

### Qualitative In Context Learning Model Development

Building on the experiences gained from the quantitative in-context learning and Random Forest models, the project then transitioned to developing an in-context learning model using qualitative data. The same dataset used in the qualitative Random Forest model was applied to the qualitative in-context learning attempt.

To maximize the chances of seeing strong performance with the model, multiple LLMs were used. First, like the quantitative in-context learning model, ChatGPT 4 was employed; however, as seen earlier, context window restrictions forced a shortening of the data to only 50% of the 2023 numbers. As a result, Google's Gemini 1.5 was also used. This LLM took in all the 2023 data, however, it came with its own limitations. Because Gemini 1.5 is still in testing, only one prompt is allowed, so there was no ability for the LLM to learn from its mistakes, as it could only predict the first tournament of the 2024 season.

## Results

### Overview of Model Testing

The performance of all developed models was evaluated based on their ability to predict outcomes for future tournaments, with distinct testing periods aligned to the specific data used during their training phases. The Random Forest models, both quantitative and qualitative, were tested using the 2024 data to assess their predictive power on future events after being trained with the comprehensive 2023 dataset. In contrast, the in-context learning models were primarily evaluated using 2023 tournament data, 50% of which was used for training, with the remaining 50% used for predictions. As mentioned earlier, the qualitative in-context learning model was also tested with Gemini 1.5, which allowed for all 2023 data to be used. As a result, it was tested on the first tournament of 2024, but no additional tournaments were tested due to model constraints.

**Evaluation Metric: Normalized Discounted Cumulative Gain (NDCG)**

The primary metric used to evaluate the model's performance was the Normalized Discounted Cumulative Gain (NDCG), a metric particularly suited for this type of predictive modeling where the order of predictions—i.e., player rankings—is crucial. NDCG provided a more relevant measure of success for the model by emphasizing the accuracy of the rank order of the predicted outcomes against the actual outcomes.

**Estimating NDCG of Pure Chance**

To establish a baseline for evaluating the effectiveness of the predictive models, the NDCG of pure chance was calculated. This process involved randomly permuting the player rankings for each tournament and then measuring the NDCG for these random rankings against the actual outcomes. This method simulates the expected performance of a purely random model, providing a critical reference point for assessing the true predictive power of the developed models. The

calculated NDCG for a model based purely on chance was consistently around 0.61, setting a threshold for acceptable model performance. Models scoring above this baseline demonstrate a predictive capability beyond random guessing, thereby validating their utility in practical applications.

**Quantitative Random Forest Model Performance**

Demonstrating solid predictive capabilities, the quantitative Random Forest model achieved an average NDCG of 0.84 across all 2024 tournaments. It performed best at the Arnold Palmer Invitational (NDCG of 0.99) and struggled at the Farmers Insurance Open (NDCG of 0.65), reflecting its variable performance across different events.

**Qualitative Random Forest Model Performance**

The qualitative Random Forest model outperformed its quantitative counterpart with an average NDCG of 0.88. The highest performance was noted at the Valspar Championship (NDCG of 0.99), with the lowest at The Sentry (NDCG of 0.73), showcasing consistent predictive accuracy.

**Quantitative In-Context Learning Model Performance**

This model faced significant challenges, culminating in a cumulative NDCG of 0.60. Its performance was particularly weak at the Charles Schwab Challenge (NDCG of 0.54) and slightly better at the Memorial Tournament (NDCG of 0.65), indicating difficulties in effectively utilizing quantitative data.

**Qualitative In-Context Learning Model Performance**

Like the quantitative version of this model, the qualitative version also struggled, achieving a cumulative NDCG score of 0.60. It saw its best result at the Mexico Open at Vidanta (NDCG of 0.77) and its worst at the Wells Fargo Championship (NDCG of 0.49).

Even when using the full 2023 dataset through Google's Gemini 1.5, this model performed poorly, achieving an NDCG of just 0.56 at The Sentry. This outcome, along with the quantitative in-context learning results, highlight the limitations of in-context learning under the current technological environment.

### Comparative Summary of Model Performance

The comparative analysis of the models against the benchmark of pure chance (NDCG of 0.61) as well as each other provides insightful conclusions. Both Random Forest models significantly outperformed the baseline, with the qualitative Random Forest model achieving the highest average NDCG of 0.88, indicating superior predictive accuracy and consistency across tournaments. In contrast, the in-context learning models, both quantitative and qualitative, struggled to exceed the performance of pure chance, with their best scores barely surpassing the baseline. This stark contrast highlights the effectiveness of the Random Forest approach, particularly when using qualitatively transformed data, which not only enhanced model performance but also improved the interpretability and robustness of the predictions. The qualitative Random Forest's success against the in-context learning models underlines the potential limitations of the latter in handling complex, multi-faceted sports data within the current constraints of LLMs.

## Discussion

### Implications of Qualitative Over Quantitative Random Forest Performance

The superior performance of the qualitative Random Forest model over its quantitative counterpart has important implications. It suggests that simplifying complex numerical data into categorical groupings can enhance predictive accuracy in scenarios where the precise magnitude of a variable is less critical than its relative ranking or classification. This approach reduces the noise and potential overfitting associated with highly granular data while retaining essential information for making effective predictions. For practitioners in sports analytics, this finding emphasizes the utility of qualitative assessments in predictive models, particularly in sports where the context and comparative performance metrics provide significant predictive value.

**Necessity of Technological Advancements for In-Context Learning Models**

The challenges faced by the in-context learning models highlight a clear need for technological advancements in this area. Current limitations in data capacity and context integration capabilities of available LLMs have restricted their effectiveness in sports analytics. To improve, there needs to be a development in models that can handle larger datasets with more complex, multi-dimensional contextual information. This advancement could potentially unlock new levels of predictive performance and offer more nuanced insights into player and game dynamics.

**Potential Limitations of the Study**

The study, while insightful, has several limitations that should be considered when interpreting the results:

1. **Lack of Generalization:** The models' ability to generalize to other golf seasons or tours may be compromised due to variations in player forms, tournament setups, and environmental conditions not captured fully in the training dataset. Expanding the

training data to include multiple seasons and varying professional golf tours can help improve the model's generalization capabilities.

2. **Overfitting to 2023 Data:** There is a potential risk that the models, especially the Random Forest ones, might be overfitted to the specific patterns and anomalies of the 2023 data. To mitigate this, future models could incorporate cross-validation across different golf seasons or employ regularization techniques to enhance their robustness and applicability to other datasets.
3. **Limited Contextual Understanding in In-Context Learning:** The in-context learning models demonstrated limited capacity to effectively integrate and utilize broader contextual information. Further technological advancements of these models could curate deeper contextual analysis.

**Areas for Future Exploration**

Given the findings and limitations identified, the future of predictive modeling in sports analytics, particularly for golf, appears ripe with opportunities for technological and methodological advancements. Over the next few years, significant technological advancements in machine learning and data processing are anticipated. These advancements could greatly enhance the capabilities of in-context learning models and other predictive methodologies that utilize LLMs, making them more effective and versatile.

The integration of a broader array of data sources, such as player interviews, social media sentiment, and biometric data, could also substantially enrich the datasets used for training predictive models. This expansion would likely lead to more nuanced and accurate predictions by capturing a wider range of factors that influence player performance and tournament outcomes.

Furthermore, investigating models that combine qualitative and quantitative data represents another promising area of future research. Such hybrid models could leverage the strengths of both data types, providing a more robust framework for making predictions. This approach might offer a more comprehensive tool for analysis, blending the depth of quantitative data with the broader insights provided by qualitative categorizations.

Overall, these future directions not only suggest pathways for enhancing the predictive accuracy of current models but also open possibilities for broader applications of machine learning in sports analytics and beyond. This could potentially transform how data-driven decisions are made in sports, improving strategies for player development, game planning, and talent scouting.

## Conclusion

This study has demonstrated the potential of machine learning models, particularly Random Forest and in-context learning, to predict outcomes in golf tournaments using data from the PGA Tour. The superior performance of the qualitative Random Forest model over its quantitative counterpart underscores the benefits of simplifying complex numerical data into categorical groupings, enhancing predictive accuracy and interpretability. While the in-context learning models faced challenges, particularly in integrating and utilizing complex contextual data, these setbacks highlight the need for further technological advancements in machine learning capabilities. Overall, the insights gained from this study not only contribute to the field of sports analytics but also pave the way for broader applications of advanced machine learning techniques in predictive modeling.