Project Overview: Professional Golf Forecasting

Brett Mele

MSDS498 Northwestern University

July 9, 2023

Overview

Analytics has changed the landscape and increased business opportunities across most major sports. Not only have *Moneyball*-like revolutions taken place within every major American team sport, but private companies have begun to reap the benefits of increasing the quality and quantity of data available. Driveline baseball has revolutionized private player development in baseball, offering a suite of data-driven products and services to amateur and professional athletes. TruMedia helps enterprise clients across five professional sports conduct in-depth analysis with their offerings of proprietary metrics and visualization tools. Pro Football Focus provides its clients with both charted in-game data and advanced metrics to better analyze American football.

Relative to other sports, professional golf has lagged behind the rest of the industry. Datagolf.com sells products primarily geared towards gaming, but otherwise enterprise analytical tools are hard to come by given the lack of available data and market opportunities for similar products. While the former may still be an obstacle, the latter should no longer be of concern. The PGA tour generated $1.59 billion in revenue in 2021, a 31% year-to-year increase (Byers 2022), and has increased its prize money pool to a record $415 million in the 2022-2023 season (Goldberg 2022). These increases are driven by growing interest in the sport, with viewership up 3-4% on average as of April 2022, including a 19% increase in viewers for the 2022 Masters (Carpenter 2023).

The growing revenue and interest in the game of golf presents a market opportunity for an enterprise golf analytics product. Our primary focus for this product will be to develop a suite of models and tools geared towards forecasting skill, tournament wins and earnings of golfers on the professional (PGA) and amateur (Korn Ferry) tours. While existing products exist for analysis and monitoring of player performance, forecasting and inference surrounding those forecasts are largely untapped areas of opportunity. Our aim is to develop an end-to-end architecture and working prototype that could eventually be offered to prospective clients, including individual golfers looking for an objective self-evaluation or enterprise clients looking to maximize the value of their sponsorships.

Project Objectives

The goal of this project is to develop an innovative decision making tool for professional golf, allowing us to be first to a growing market and create a valuable revenue stream. Model forecasts will help individual golfers set realistic goals, optimize training and make informed choices regarding their careers. Sponsors will be able to better understand which golfers are likely to generate a high return on investment relative to their sponsorship cost.

To achieve this goal we define our objectives as follows:

1. **Develop an enterprise architecture to collect and store golf data**
    - **a.** Extract, load and transform data from API endpoints into cloud data warehouse
    - **b.** Automate the system to collect new data and enable up-to-date model predictions

2. **Project golfer skill in terms of strokes gained over a four year time horizon for all active golfers on the Korn Ferry and PGA tours**
    a. Includes projections for the major components of golf skill: off-the-tee, approach, around-the-green and putting
    b. Provide objective comparisons to historical golfers as a tool for golfers to study and identify areas of improvement
    c. Create objective profiles of golfers relative to the PGA tour for each year within the time horizon

3. **Project golfer performance on the PGA tour and in major championships over a four year time horizon**
    a. Includes total number of wins each year on the PGA tour and in major championships

4. **Summarize current and project future earnings on the PGA tour and in major championships**
    a. Combine tournament winnings with projected tournament finishes
    b. Estimate sponsorship return on investment

5. **Surface model outputs and model derivatives in prototype dashboard**

Modeling Approaches & Technical Requirements

Modeling Approaches

We anticipate testing a number of approaches to building the aforementioned models. While our focus will be on maximizing predictive accuracy, we may favor model types that sacrifice accuracy for increased explainability given the models will be used for inference by clients.

We will experiment with time series (ARIMA) and regression (linear regression, GAM, XGBoost) models for forecasts of golfer performance. We may favor traditional regression approaches given our data does not lend itself to true time-series modeling, as there is no seasonality, there are few timesteps and we have grouped observations. Time permitting we may also explore hierarchical models given the grouped nature of the data, and bayesian models to allow for estimation of prediction uncertainty.

To project tournament and championship wins, we plan to test both parametric (logistic regression) and non-parametric (Random Forest, XGBoost) classification models, as well as simulation based approaches (Monte Carlo simulation). The problem likely lends itself to simulation given tournament finish is conditional on the performance of other players, but the approach may be difficult to implement within the timeframe.

For golfer comparisons we will employ unsupervised learning techniques. Specifically we will experiment with dimensionality reduction (principal components analysis) and clustering (K-Means, dbscan).

Technical Requirements

1. **Data Engineering**

Our ELT pipelines will be housed in a Postgres database hosted on Google Cloud Platform (GCP). Services will be built using Go, Python and Docker.

2. **Modeling**

Supervised and unsupervised learning models will be built in both R and Python.

3. **Prototype Dashboard**

The prototype application will be developed using R Shiny and deployed via Docker container on GCP.

Success Metrics and KPIs

Before going to market, it is important to ensure model quality. We will evaluate regression models using root-mean-squared error and classification models using log loss. We will use historical world golf and datagolf.com rankings in baseline models for comparison.

Once released, we will measure the success of our project via the following metrics:

1. **Revenue**

The primary goal of the product is to drive revenue for the company. Conservatively we hope to deliver at least $100,000 in revenue in our first year via a subscription based service.

2. **Customer Acquisition Cost**

As a new product we will incur marketing costs to reach prospective customers, which we will need to monitor relative to revenue increases.

3. **User Growth & Churn Rate**

We will monitor our growth in users and the rate at which users discontinue their subscription as an indication of the efficacy of our product.

4. **Daily Active Users**

Tracking active users will give an indication of user engagement and adoption of the product.

Project Plan

Timeline

Task	Week 1 (6/25)	Week 2 (7/2)	Week 3 (7/9)	Week 4 (7/16)	Week 5 (7/23)	Week 6 (7/30)	Week 7 (8/6)	Week 8 (8/13)	Week 9 (8/20)	Week 10 (8/26)
Research and project scope										
Project overview, plan, goals & KPIs										
Data collection										
Exploratory data analysis										
Data preparation										
Model experimentation										
Finalize models, aggregate outputs										
Prototype app/dashboard										
Deployment										
Final Paper & Presentation										

Description

Task	Description
Research and project scope	Literature review, explore potential data sources, project ideation
Project overview, plan, goals & KPIs	Define deliverables, project objectives and KPIs, and modeling approaches; create project plan and timeline
Data collection	Create initial pipelines to retrieve data from APIs and store in cloud database
Exploratory data analysis	Understand data, verify data quality, determine key relationships for modeling, evaluate feasibility of modeling approaches and adjust if necessary
Data preparation	Clean data, combine data sources, engineer features etc. for use in modeling
Model experimentation	Experiment with model types and targets, evaluate models with common error metrics
Finalize models, aggregate outputs	Determine final models to use in production, aggregate model outputs and derivatives
Prototype app/dashboard	Build prototype application to serve as proof-of-concept
Deployment	Deploy data pipelines, models and services
Presentation	Present results and final deliverables

REFERENCES

Byers, Justin, and Justin Byers. “PGA Tour Reports 37% Increase in Revenue During 2021.” Front Office Sports, December 20, 2022. https://frontofficesports.com/pga-tour-reports-37-increase-in-revenue-during-2021/#:~:text=The%20North%20American%20pro%20golf,the%20controversial%20LIV%20Golf%20League.

Carpenter, Josh. “PGA Tour Seeing Viewership Increases Midway through 2023 Season.” Sports Business Journal, April 12, 2023. https://www.sportsbusinessjournal.com/Daily/Issues/2023/04/12/Media/pga-tour-viewership-cbs-nbc-golf-channel-espn.aspx.

“Data Golf,” n.d. https://datagolf.com/.

Driveline Baseball. “About Us | Driveline Baseball,” August 3, 2022. https://www.drivelinebaseball.com/about/.

Goldberg, Rob. “PGA Tour Increases Prize Money to Record $415m for 44-Tournament 2022-23 Schedule.” Bleacher Report, August 1, 2022. https://bleacherreport.com/articles/10044029-pga-tour-increases-prize-money-to-record-415m-for-44-tournament-2022-23-schedule.

TruMedia. “TruMedia,” n.d. https://www.trumedianetworks.com/.

PFF. “NFL, Fantasy Football, and NFL Draft | PFF,” n.d. https://www.pff.com/.