DEVELOPING A RECOMMENDATION SYSTEM FOR COLLEGIATE GOLF RECRUITING

A Research Paper submitted to the Department of Engineering Systems and Environment
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

Joshua Barnard
Vienna Donnelly
Ava Jundanian
Rachel Kreitzer

By

Michael Bassilios

April 12, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

Developing a Recommendation System for Collegiate Golf Recruiting

***Abstract*—In the world of college sports, the process of recruiting players is one of the most important tasks a coach must tackle. With only 6% of the 8 million high school athletes earning spots on NCAA teams, finding and selecting the right players can be incredibly challenging even with the availability of widespread data. Some sports, like football and basketball, have found great success using predictive analytics to estimate success in college. These efforts, however, have not yet been extended to other sports, such as golf. Given the vast amount of data available to the public on junior golfers, there is clear potential to bring analytics to college golf recruiting. We partnered with GameForge, a leading golf analytics company, to create a recommendation tool for college coaches, one that leverages the already existing data on high school and collegiate golfers and a variety of predictive models to display athletes we believe would best fit in a certain college program. A systems analysis approach was taken to find the factors that most accurately predict a high school player's success in college golf. This was done with a variety of models including the forecasting of probability of a high school athlete being a top ranked college golfer, the finding of players with a similar performance to another desired player, and the predicting of a junior golfer's scoring performance and development during the remainder of their high school career and during college. Using these models, we identified several factors that are predictive of player similarity and performance. The research team iteratively developed these models to be used in conjunction with each other in order to provide meaningful, and understandable recommendations to a college coach on which players they should recruit to maximize success.**

Keywords—golf, k-means, linear regression, sports analytics, sports recruiting

I. Introduction

In the last few decades the field of sports analytics has grown from an antiquated system, using personal experience and popular opinion to determine desirable players, into a cutting-edge field that is using big data to drive decisions [1]. It is no longer enough to just be a winning team. Coaches want to know why they are winning, what makes their players successful and how they can continue to be successful as their players come and go. The implementation of data analytics is steadily permeating the training and recruitment of players at both the collegiate and professional level for a variety of sports, primarily baseball, basketball and football [2][3]. College golf is no exception to this paradigm shift to using analytics to characterize and improve player performance. Golf recruiting practices, however, continue to rely heavily on public rankings and coaches collecting player information themselves by researching players, attending tournaments and making personal connections. Coaches must prudently choose players who they believe will contribute to the future achievement of their team, sometimes years in advance of when the players will actually be eligible for collegiate play. At present, the only available resources for coaches to identify junior talent and judge players' skill level are junior golf ratings created by organizations like the American Junior Golf Association (AJGA). The system of college golf recruiting is currently fragmented and tedious; great difficulty exists in determining which players will best fit a particular college team. Coaches would benefit greatly from a recommender tool that predicts a player's future performance and identifies desirable players. However, no such system currently exists for golf recruiting. The creation of an analytics driven recommender system will simplify the recruiting process, not only making the process more convenient, but also granting coaches more confidence in their recruits.

The purpose of a recommender system is, "to generate meaningful recommendations to a collection of users…. which can be used to identify well-matched pairs" [4]. The goal of the research is to assist GameForge, a golf analytics company that seeks to help golfers improve their game, in creating a recommender system to streamline the college golf recruiting process. The uses of this recommender system are twofold: (1) to provide college coaches with information on junior golfers that may fit their team; and (2) to match junior golfers with college teams that best fit their preferences and skill level. Because no recommender system exists that streamlines the college golf recruiting process for players and coaches, the recommender system outlined in this paper will be the first to simplify the college golf recruiting process, saving money and time for all stakeholders involved.

Six models, outlined in Table I, were created to provide insight on the performance of junior and collegiate golfers. The first two models, the team ranking model and college performance predictor, were used to discern important predictors to be used in the four subsequent models. The use of each of these models in conjunction with one another will provide coaches a comprehensive description of the prospective players they are interested in as well as alert them to players they might not have discovered on their own.

This paper will first detail related application of data analytics in golf and career recruiting. Following the explanation

of related work, the general approach to our analysis will be explained, including the variables used and results of each model created.

TABLE I. DEVELOPED MODEL DESCRIPTIONS AND OUTPUT

Model	***Description***	***Output***
Team Ranking Model	Determines which college golf metrics are indicative of team success	List of variables that are significant in predicting team rank
College Performance Predictor	Attempts to predict a junior golfer's performance in the significant college metrics found in the Team Ranking Model	35 linear regression models with variable significance
Division I Predictor	Determines whether a junior player will play on a Division I golf team	Probability that a junior will play on a Division I team
Top X Classifier	Predicts generally where a junior player will rank once she is playing golf in college	Confidence a player will be within a given range of ranks
K-Nearest Neighbor Lookalike	Determines the K players who are most similar to an inputted player of interest	List of K lookalike players
Win Shares Simulation	Predicts team success when adding current junior players, team performance over time	Wins with player Team metrics as players are added

II. RELATED WORK

Recommender systems are heavily relied on in employee recruiting "in order to generate personalized recommendations of candidates and jobs" [5]. In 2019, 99% of Fortune 500 companies used applicant tracking systems, a type of recommendation software, to manage their talent acquisition [6]. The use of recommender systems gives way to a new form of recruiting whereby companies can identify individuals who possess the highest likelihood for success in their organization while also considering far more applicants than ever before. This paradigm shift in the practice of professional recruiting can also be adopted by sports teams looking to recruit new players. The harnessing of athletic data to inform and characterize team decisions has become incredibly prevalent in professional sports, to the point where each major professional sports team now possesses an analytics expert on staff [1]. For instance, a study using predictive modeling and data analytics was conducted to help the University of Virginia football team in their recruiting process, giving their coaches a "competitive advantage" in the recruitment of new players [7]. Professional golfers are also beginning to leverage the power of analytics to improve their game. Shotlink, a ball-tracking database created by the Professional Golfers Association (PGA), has afforded professional players the opportunity to improve their game by better understanding their performance in PGA tournaments [8]. The Golfmetrics program was also developed to record golfer's shot data in order to discern performance patterns prevalent in golfers at both professional and amateur levels [9]. GameForge, too, offers professional and collegiate golfers a way to leverage their personal performance data into targeted training regimes designed to enhance player performance [10][11]. The power of data analytics has been applied to golf, but only in the context of improving performance, and has not yet been leveraged in the process of collegiate level golf recruiting. The successful implementation of predictive modelling in recruitment by the University of Virginia football team underscores the power a comprehensive recruitment recommender system can have for a collegiate program and suggests the need to expand the practice of predictive modelling to other collegiate athletic programs.

III. APPROACH AND RESULTS

A. General Methodology

The analysis was focused entirely on female golfers due to a need for external model validation from our client, who is more specialized in women's golf. The analysis was divided into two main tasks. The first task was to determine the college golf metrics that are significant in predicting college success both for individual players and the team as a whole. To perform this task, success for a collegiate golfer and a collegiate golf team was defined. For players, college rating and ranking is used as a measure of success. Both of these metrics are similar because the player rank is determined by simply sorting the player ratings from least to greatest. Note that for both of these metrics, a smaller numerical value is preferred (e.g. rank 1 is better than rank 2 and a rating of 70 is better than a rating of 71). For college teams, only team ranking was used as a measure for success because teams are not given ratings. The reason these two metrics were chosen to measure success is because of their widespread acceptance. Every golfer and coach will have certain metrics that they want to focus on or improve, such as tournament wins, mean score, or consistency and these will vary for each player and coach. Ultimately, a high ranking or rating generally indicates that a player or team is successful as a whole. Additionally, these two metrics encompass other widely used metrics, because if a player or team plays well and wins tournaments, they will have good ratings and rank.

The second task was to determine the junior golfers that a college team should recruit. Many predictive models were developed using the junior players' AJGA data to predict their success and performance in college golf. The use of many models gives users of the recommender system a more nuanced, wide view on a player's predicted performance. The system produces a variety of clear, understandable metrics on each junior player—likelihood of being Division I caliber, predicted college rank, lists of look-alike golfers, and simulated wins.

B. Variable Selection

There were ten main predictor variables used for college players and twenty-one predictor variables used for junior players. Shown below in Table II, variables were created to break down performance first by par value, then again on relative length of the hole for par four and five. This allows the examination of the effects of hole yardage on a player's score relative to par while holding constant the par value. This provides the insight that having a good score relative to par on the longer holes has a greater impact on a player's rating.

TABLE II. DATA SOURCES AND VARIABLES CREATED

Data Source	***Variables Present***	***Variables Created***
AJGA (junior data)	Graduation Year	Yearly Improvement in Below Variables Tournament Holes Played
AJGA (junior data)	Year Player Name Tournament ID Round Number Hole Number Hole Yardage Hole Par Hole Score Score Relative to Par	Par 3 Mean (P3M) Par 3 Variance (P3V) Par 4 Short Mean (P3SM) Par 4 Short Variance (P4SV) Par 4 Long Mean (P4LM) Par 4 Long Variance (P4LV) Par 5 Short Mean (P5SM) Par 5 Short Variance (P5SV) Par 5 Long Mean (P5LM) Par 5 Long Variance (P5LV)
Golfstat (college data)	College Attending	

Cutoff values for long yardage are set at 350 yards and 480 yards, for par four and par five respectively, based on the median of the Gaussian histograms shown in Fig. 1 and validated with client input.

For each of these groups, every player's mean score relative to par as well as the variance of their score relative to par was calculated. Mean score relative to par is the main performance metric used for the players. Variance was included as a metric to evaluate a player's consistency. This resulted in ten total predictors for every college player as shown in Table II.

Using the same process, the same ten predictors are found for the junior players. Junior metrics are further augmented with additional eleven additional potential predictors. Three factors directly related to golf metrics determine a junior player's success in the college game. The first factor is a junior player's performance on the course, given by the mean and variance of her score relative to par, as described earlier. The second factor is a player's improvement over time. If a player tends to show consistent improvement as a junior golfer, this could improve her chance of having success in college compared to a player that does not show improvement or even regresses. To calculate this metric, the junior player data was divided into four groups, one for each year of high school, freshman through senior. Doing so resulted in a calculation of the change in each of the players' ten metrics from one year to another. The third factor came from the idea that experience as a junior golfer could influence a player's success in college golf. This experience factor was represented by including the total number of holes played by the junior golfer in each year of school.

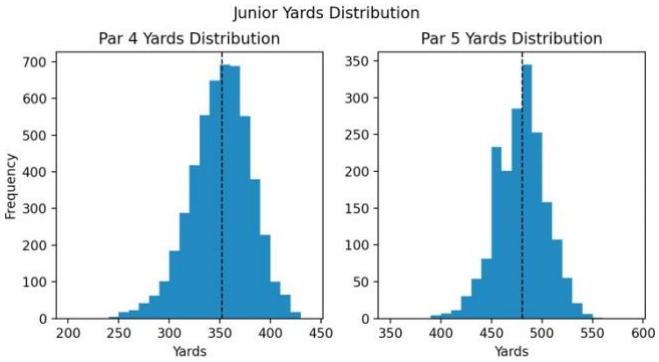

Fig. 1. Junior Yards Distribution for Par 4 and Par 5

Thus, for each player's school year, the dataset includes twenty-one metrics: ten metrics defining mean and variance of score relative to par for each hole type, ten metrics that are the deltas of each of the first ten metrics, and finally a metric totaling the number of holes played that year. These variables and datasets were used throughout the remainder of the analysis.

C. College Player Rating and Team Ranking Models

Using the success objectives of player rating and team ranking, we determined the significant golf metrics for predicting those objectives. All ten collegiate player predictors were used in several models to determine which metrics were most indicative of college success. The first model was a linear regression model that used player rating as the response variable and the ten listed predictors as the independent variables, as shown in Figure 2. The second model was an ordinal logistic regression model that used team ranking as the response variable and each of the ten predictors as the independent variables. However, each predictor was averaged by team rather than by player.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.325e+01  2.212e-01 331.113  < 2e-16 ***
P3M          2.712e+00  3.234e-01   8.386 2.53e-16 ***
P3V          4.846e-01  2.660e-01   1.822  0.06891 .
P4SM         3.128e+00  3.060e-01  10.222  < 2e-16 ***
P4SV        -4.557e-05  2.308e-01   0.000  0.99984
P4LM         4.927e+00  3.616e-01  13.625  < 2e-16 ***
P4LV        -6.475e-01  2.432e-01  -2.662  0.00793 **
P5SM         2.088e+00  2.055e-01  10.162  < 2e-16 ***
P5SV        -1.526e-01  1.256e-01  -1.215  0.22460
P5LM         2.698e+00  2.293e-01  11.765  < 2e-16 ***
P5LV        -3.944e-01  1.301e-01  -3.030  0.00253 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.844 on 740 degrees of freedom
  (150 observations deleted due to missingness)
Multiple R-squared:  0.7835,    Adjusted R-squared:  0.7806
F-statistic: 267.9 on 10 and 740 DF,  p-value: < 2.2e-16
```

Fig. 2. Results of Linear Regression on Player Rating

The results of both models were fairly similar, with each of them having nearly identical significant predictors. The significant predictors at the 0.05 significance level were: P3M, P4SM, P4LM, P5SM, P5LM, and P5LV, noted by asterisks in Figure 2. Note that P4LV was not significant in the team ranking model and therefore not included in subsequent models. The similarity between the models indicated that the metrics that make a college golfer successful are the same ones that make a college team successful. These models helped determine which college metrics were most important in allowing a player and team to be successful in collegiate golf. Using this information, it was possible to determine which junior golfers to recruit.

D. College Performance Predictor Model

Using the six significant metrics found in the prior model, we created 35 linear regression models to predict a junior player's performance in those metrics using her AJGA data. Players used in this model were ones that had both AJGA data and college data. The AJGA data was divided into the four years of high school and a fifth data set that had each player's four-year average for each metric. Using each of those five datasets, our goal was to directly predict each of the six significant college metrics that were found earlier and the player's college rating. In other words, there were seven models for each school year and seven models for the players' average junior data. Of those 35 models, most of them had R^2 values of around 0.25 to 0.30. The toughest college variable to predict was P5LV, which had R^2 values under 0.10 for each of its five models. The best model was using the players' average junior data to predict their college rating directly, which had an R^2 of about 0.42. Overall, the results of these models show that it is very difficult to accurately predict players' performance in college in specific metrics just from their AJGA data. However, the models did still result in some significant AJGA metrics that may be more indicative of collegiate golf success. For instance, the number of holes played was a significant predictor in many of the models, showing that experience as a junior is important for success in collegiate golf.

E. Division I Model

Because player caliber varies by division, a model was created to differentiate junior players who have the potential to join a Division I team versus junior players who are more likely to join Division II or III teams. This model uses players' junior AJGA data to predict the probability of a golfer playing on an NCAA Division I team. This model was created using CART classification with an entropy node splitting method and using players' P3M, P4SM, P4LM, P5SM, P5LM, and P5LV as the predictor variables. These variables were used because they were found to be significant in predicting college player rating, and, therefore, significant in predicting collegiate success. Players with data outside of 5 standard deviations of the mean for the predictor variables were not included in the model because such extreme values decreased the accuracy of the model. The CART classification assigned each player in the dataset a probability of playing on a Division I team based off of her junior performance. The model was cross-validated with junior players' actual division assignment and 30% of the data was used for testing while the remaining 70% was used for training. As shown in Table III, the model predicted with 88.6% accuracy that a player would play Division I and 86 % accuracy that a player would not play Division I.

TABLE III. CROSS VALIDATED RESULTS OF DIVISION I CLASSIFIER MODEL

Statistics	***Training (%)***	***Test (%)***
True positive rate	91.2%	88.6%
False positive rate	3.2%	14%
False negative rate	8.8%	11.4%
True negative rate	96.8%	86%

F. Top X Classifier Model

One of the goals of this recommender system is to predict if a current junior player will succeed as a player in college. To satisfy this goal, a model was created to predict the college rank of a current junior woman golfer. The model included ranks of women who are currently playing or previously played in college and their respective junior data, which included all of the predictors described above. Initially, the model was designed to predict, using a player's junior data, if she would be a Top 25 ranked college player or not. Additional models were created that classified players as Top 50 or not, Top 75 or not, Top 100 or not, and Top 150 or not.

For these models, good performance was defined as high recall and high precision. Due to the large imbalance of the classes (there were many more players not in the Top X than in the Top X), accuracy was not considered. Recall provided the percentage of the Top X players that were correctly classified and precision gave the percentage of actual Top X players that were classified as Top X. Given the goals of the system, recall was determined to be the most important metric.

Ten-fold cross-validated ensemble models using clustering methods, random forests and neural nets were run and evaluated based on their performance. For some models, a more complicated ensemble model performed better and for others, simply using a clustering method provided the most optimal results. Despite better performance on the Top 75 and Top 100 models using an ensemble method, a clustering model performed nearly as well and was chosen in order to increase interpretability. The Top 50 model confusion matrix is shown in Table IV.

TABLE IV. CROSS VALIDATED CONFUSION MATRIX: TOP 50

	True Top 50	True Not Top 50	*Precision*
Pred. Top 50	122	8	93.85%
Pred. Not Top 50	15	1214	98.78%
Recall	89.05%	99.35%	

The classifier models discussed have all been binary classifiers with only two states, Top X or not. However, in order to better understand how the players were being classified and where errors were occurring, a multi-state classifier was built. The states in this model were bins of college ranks listed here: 0-25, 26-50, 51-75, 76-100, 101-150, >150. A clustering method

was used for this model as well due to the quality of performance and the explanatory capability. The resulting class precision ranged from 60% to 98% (Table V).

TABLE V. MULTISTATE CLASSIFIER CROSS VALIDATED CONFUSION MATRIX

	True 25	True 50	True 75	True 100	True 150	True >150	Class precision
Pred 25	61	8	0	4	0	0	84%
Pred 50	10	43	4	0	0	2	73%
Pred 75	0	15	52	4	4	6	64%
Pred 100	0	0	16	26	1	2	58%
Pred 150	0	0	2	13	51	7	70%
Pred >150	0	0	2	4	12	1010	98%
Class recall	86%	65%	68%	51%	75%	98%	

G. K-Nearest Neighbor Lookalike Model

A goal of the GameForge research effort was to accurately identify a group of players with similar attributes to a desirable player, or player of interest, indicated by coaches using GameForge. With this tool, coaches would be able to identify players who possess similar attributes as other top players in golf. The tool could also be used by coaches to find new players to replace the graduating members of their teams.

The Lookalike model was built using a K-nearest neighbor algorithm which calculates the Euclidean distance between the player of interest and every other player within the given dataset. The algorithm then returns the K players with the smallest distance to the player of interest. Coaches using this tool will be able to specify the player of interest and the number of lookalikes to return, K.

In order to validate the results provided by this model several samples of results were generated and informally checked by the GameForge team and several of their clients.

H. Win Shares Simulation Model

Another goal of this system is to investigate how a recruit will improve or change the current team, so a model that predicts the team's performance over time was constructed. This model simulates a given team's current performance hundreds of times based on the type of tournaments the team participates in throughout the season. This is then compared to the relative performance as new members are added to and removed from the team. This allows the client to not only look at the team when adding one or two players in a single year, but to also see how the team evolves over time.

Using the K-Nearest Neighbor Lookalike Model we are able to create a prediction of future college performance of current junior players. We can restrict the lookalikes of junior players to only current college player's data points from when they were a junior. This allows us to approximate the performance of junior players by the time they reach college by using a weighted average of the K-nearest neighbors. These players' performance is then simulated and their wins over the initial wins, improvement from the previous year, and trends in average team statistics is predicted. Because the tool is based on a simulation model, it will perform exactly as predicted, given the assumption that the underlying distribution of scores is Gaussian (Fig. 3).

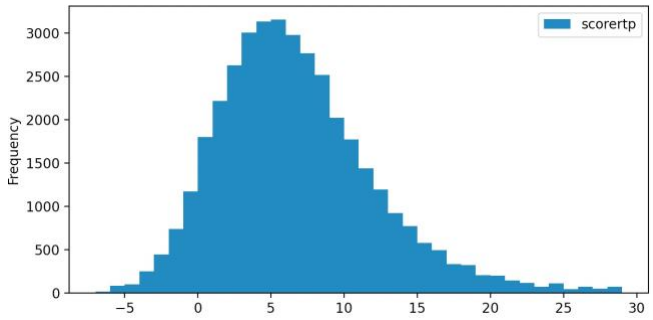

Fig. 3. Distribution of Tournament Scores

Due to the limited college data, the predictive element of the model is difficult to validate; however, because it is based upon real world validation in the K-Nearest Neighbor model, it is expected to predict correctly and can be validated as more data is collected in the GameForge system.

I. Dashboard

The college player rating and team ranking regressions provided insight for the other four models developed. These remaining models are organized into a dashboard to allow users of the recommender system a full view of potential recruits. By incorporating these four models, the user receives multiple viewpoints that provide clear, understandable success metrics on potential recruits.

IV. CONCLUSION

From the models, six metrics were found to be significant in predicting success in college golf: P3M, P4SM, P4LM, P5SM, P5LM, and P5LV. However, while it is difficult to predict a junior player's performance in any of those six metrics in college, the models developed gave insight into the important features. Using a player's AJGA data, we were able to predict that she will be a Division I player with 88.6% accuracy. Additionally, we were able to claim with precision of 84% that she would be ranked in the Top 25. The lookalike model returns a list of players that are most similar to a given player of interest by measuring the Euclidean distance between all player's significant metrics and returning the K closest players. The Win Shares Simulation model allows coaches to view the impact of adding certain recruits to their team's performance over time. These models were designed with the purpose of building on each other to provide as much information as possible to the coaches about which players would be best to recruit. Most of the time, the top ranked junior golfers are well-known and highly sought after. However, by using these models, we believe that some of the lesser-known junior players that have the potential to be great college players can be found. Furthermore,

players can have a better idea of which golf metrics tend to lead to a higher chance of success.

These models still have room for improvement, and it is necessary to continue to validate them using future data to ensure their proper functionality. One crucial step is to validate the model behavior when expanding application to men's golf. Given the relative lack of college data, models that use college player data should be closely monitored as more data is collected. Additionally, as the GameForge network grows and more players are added into the system, more detailed analysis of both junior and collegiate players can be conducted to better understand the links between junior performance and collegiate success, bringing in raw physical metrics such as swing speed and length of shot. Also, these models were built using women's data exclusively. In the future, these models should be expanded to provide predictive metrics for male golfers as well. Lastly, these models do not address the qualitative components of recruiting like player and coach preferences. Player and coach preferences could be built into the GameForge recruiting tool by getting coaches and players to fill out surveys and having prospective players enter their desired characteristics of a collegiate program when joining the GameForge recruiting tool. We could then ask players and coaches the importance of each characteristic, giving us weights to attach and then finding the optimal qualitative match to use as an additional category when evaluating the quality fit between a player and program.

ACKNOWLEDGMENT

We would like to thank GameForge for sponsoring and supporting this project. In particular, we would like to thank Mark Sweeney and Brian Bailie for their continuous support, advice, and engagement throughout the project.

REFERENCES

[1] L, Steinberg, "Changing the game: The rise of sports analytics" Forbes. Retrieved from https://www.forbes.com

[2] J. Hoege et al., "An Interdisciplinary Approach to Sports Analytics in a University Setting," 2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2020, pp. 1-6, doi: 10.1109/SIEDS49339.2020.9106647.

[3] J. E. Blanchfield et al., "Developing Predictive Athletic Performance Models for Informative Training Regimens," 2019 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2019, pp. 1-6, doi: 10.1109/SIEDS.2019.8735633.

[4] P. Melville, V. Sindhwani, "Recommender Systems," 2011 Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_705

[5] Y. Lu, S. El Helou, & D. Gillet, "A recommender system for job seeking and recruiting website," 2013 Proceedings of the 22nd International Conference on World Wide Web, pp. 963-966.

[6] R. Ryan, "Want to be noticed by recruiters? Try this resume strategy to get through the applicant tracking system." Forbes. Retrieved from https://www.forbes.com

[7] K. Peng et al., "Predictive analytics for University of Virginia football recruiting," 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2018, pp. 243-248, doi: 10.1109/SIEDS.2018.8374745.

[8] G. M. Arastey, "The increasing presence of data analytics in golf." 2020 Sport Performance Analysis. Retrieved from https://www.sportperformanceanalysis.com

[9] M. Broadie, "Assessing golfer performance using golfmetrics." *Science and golf V: Proceedings of the 2008 world scientific congress of golf*. St. Andrews: World Scientific Congress of Golf Trust, 2008.

[10] K. Rohrer et al., "Developing State-Based Recommendation Systems for Golf Training," 2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2020, pp. 1-7, doi: 10.1109/SIEDS49339.2020.9106646.

[11] https://mygameforge.com/landing