2015 International Conference on Computational Science and Computational Intelligence

Mining Trackman Golf Data

Ulf Johansson*, Rikard König*, Peter Brattberg*, Anders Dahlbom†, Maria Riveiro†
*Department of Information Technology
University of Borås, Sweden
Email: {ulf.johansson, rikard.konig, peter.brattberg}@hb.se
†School of Informatics
University of Skövde, Sweden
Email:{anders.dahlbom, maria.riveiro}@his.se

***Abstract*—Recently, innovative technology like Trackman has made it possible to generate data describing golf swings. In this application paper, we analyze Trackman data from 275 golfers using descriptive statistics and machine learning techniques. The overall goal is to find non-trivial and general patterns in the data that can be used to identify and explain what separates skilled golfers from poor. Experimental results show that random forest models, generated from Trackman data, were able to predict the handicap of a golfer, with a performance comparable to human experts. Based on interpretable predictive models, descriptive statistics and correlation analysis, the most distinguishing property of better golfers is their consistency. In addition, the analysis shows that better players have superior control of the club head at impact and generally hit the ball straighter. A very interesting finding is that better players also tend to swing flatter. Finally, an outright comparison between data describing the club head movement and ball flight data, indicates that a majority of golfers do not hit the ball solid enough for the basic golf theory to apply.**

***Keywords*—Data mining, Golf, Trackman**

I. Introduction.

The purpose of data analysis ("data mining") is to find potentially valuable information hidden in the data. When analyzing data, patterns will always be found, but a majority of these relationships are often trivial. Such patterns are not interesting in itself, but can be used to validate the technique used, or the model produced. In fact, we would be very surprised if the analysis failed to find the most obvious patterns. Looking at identified unexpected patterns, it must be remembered that most techniques are prone to overfitting the data, resulting in spurious relationships. With this in mind, it is often necessary to utilize domain knowledge for determining whether discovered patterns are in fact novel and general. Finally, it must be noted that findings often have to be actionable in order to be valuable.

In this application paper, we combine descriptive statistics and machine learning for analyzing data from golf swings. Findings are related and contrasted to golf theory. The overall purpose is to discover non-trivial, general and, where possible, actionable findings about what separates skilled golfers from poor.

II. Background.

The golf swing is a very complex motion, and the club head moves at high speed, making it extremely hard to analyze swings visually. Nevertheless, this is the traditional method, still used by most golf instructors. During the last decade, affordable high-speed cameras have become available to the public, making it slightly easier for the instructors, but it is is only recently that new and innovative technology like the Trackman launch monitor radar has made it possible to evaluate and analyze golf swings quantitatively.

Golf has a unique handicap system, designed to let golfers of different skill compete against each other on equal terms. There are a couple different handicap systems, but the EGA (Europe) and the USGA (USA) are the two dominant. A golfer's handicap, which is a measure of his/her skill, is a numeric value, typically between 0 and 36[1]. In handicap competition golfers deduct their handicap from the gross score producing a net score, which is the final result.

Broadie [1], divides golf shots on a course into four different categories:

- *Long game:* shots longer than 100 yards
- *Short game:* shots shorter than 100 yards, not including sand shots
- *Sand game:* shots from bunker no longer than 50 yards
- *Putting:* shots on the green.

Naturally, mastering all parts of the game are vital to become a skilled golfer; as an example, almost half of all shots, for most golfers, are actually putts. Nevertheless, Broadie's argument that the long game is by far the most important, is widely accepted.

The golf swing can be analyzed and measured using different metrics and terminology, but in this study we use Trackman data and Trackman terminology. The current version of Trackman produces, for each shot, nine values related to the club head movement around impact, and 14 values related to the ball flight:

- *ClubSpeed* - Speed of the club head at impact.
- *AttackAngle* - Vertical movement of the club through impact.
- *ClubPath* - Horizontal movement of the club through impact.
- *SwingPlane* - Bottom half of the swing plane relative to ground.

[1]It is possible to have a handicap lower than 0 – which is called a plus handicap

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.77

- *SwingDirection* - Bottom half of the swing plane relative to target line.
- *DynLoft* - Orientation of the club face, relative to the plumb line, at impact.
- *FaceAngle* - Orientation of the club face, relative to target line, at impact.
- *FaceToPath* - Orientation of the club face, relative to club path, at impact.
- *BallSpeed, BallSpeedC* - Ball speed instant after impact, speed at landing.
- *SmashFactor* - Ball speed / club head speed at instant after impact.
- *LaunchAngle* - Launch angle, relative horizon, immediately after impact.
- *LaunchDirection* - Starting direction, relative to target line, of ball immediately after impact.
- *SpinRate* - Ball rotation per minute instant after impact.
- *SpinAxis* - Tilt of spin axis. (+) = fade / slice, (-) = draw / hook.
- *VertAngleC* - Ball landing angle, relative to ground at zero elevation.
- *Height, DistHeight, SideHeight* - Maximum height of shot at apex, distance to apex, apex distance from target line.
- *LengthC, LengthT* - Length of shot, C = calculated carry at zero elevation, T = calculated total including bounce and roll at zero elevation.
- *SideC, SideT* - Distance from target line, C = at landing, T = calculated total including bounce and roll. (+) = right, (-) = left.

When looking at Trackman values, it is fair to say that a few attributes, i.e., face angle, club path, face-to-path and the angle of attack, are considered to be the most important. In fact, they are often referred to as the fundamental attributes. More specifically, using fundamental golf theory, the starting direction of the shot is, to a large degree, based on the face angle [2], while the face-to-path will determine the curvature of the shot. In practice, only these few numbers are frequently used when analyzing golf swings in order to give instructions. It must be noted, however, that this analysis is rather crude, especially if the angle of attack is left out. In addition, these attributes only explain the start of the ball, and the curvature of the shot, if the ball is hit in the sweet-spot of the club. If the ball is, in fact, hit towards the heel or the toe of the club, other forces apply, e.g., the *gear effect*.

In this study, we will combine descriptive and modeling techniques to predict and explain a player's skill (handicap), based on Trackman data. This is clearly a challenge since while the handicap summaries the entire game, i.e., long game, short game, putting etc., Trackman data cover only the long game. In addition, a handicap is not necessarily (for a number of reasons) accurate. Nevertheless, it is fair to say that for amateur players, the handicap is the most obvious proxy for skill.

A. Related work.

There are a few scientific studies analysing golf swings quantitatively, typically using high speed video, see e.g., [3][4]. However, due to the tedious manual labor required for video analysis, these and similar studies have been restricted to between 20 and 45 players. Using a small number of players, and often also just a few swings for each player, makes it very hard to perform a rigorous analysis of the interactions between different swing-related variables. Because of this, previous studies have focused on analysing single variables using statistical techniques. In [5], which is the only identified study using a larger group of players, 285 players each hit 10 shots with the driver. All shots where recorded using high speed cameras, but also with Trackman. Statistical tests were used to show that players with lower handicaps (in that study under 11.4) were significantly more consistent regarding a large number of variables related to the movement of the club near impact.

III. Method.

The overall purpose of this study is to utilize data analysis to identify important aspects separating skilled golfers from poor. This study contains, all-in-all 275 male golfers, each hitting five shots with a 7-iron and five shots with a driver. A large number of Trackman attributes were recorded for each shot, and the last shot with each club was also recorded using high-speed cameras for future analysis

In order to find and identify (preferably non-trivial) aspects separating skilled and poor golfers, three different analyses were performed. First we group the players based on handicap and identify, using descriptive statistics, the key differences between the groups. For this analysis we used four different groups:

- *Low* - handicap better than 4.5. In Sweden, handicaps below 4.5 are only modified based on scores obtained in competition.
- *Single* - handicap between 4.6 and 9.9. Sporting a "single digit handicap" is the goal for many recreational golfers.
- *Average* - handicap between 10.0 and 18.0.
- *High* - handicap over 18.0

Second, we use correlation analysis, between attributes and handicap, showing which attributes that are the most important. Finally, we apply predictive regression – generating and interpreting models directly relating the handicap to Trackman attributes.

In the analyses we elected to look at data for the 7-iron and for the driver separately. When producing the profile for each player, the actual values for the "median shot" (based on carry length) was used, instead of averaging over different shots, which could be a dubious procedure. In total, 55 independent variables were used for each player and club. For completeness, the attributes used are the 23 described above, but for nine specific attributes, that produce both positive and negative values (ClubPath, SwingDirection, FaceAngle, FaceToPath, LaunchDirection, SpinAxis, SideHeight, SideC, SideT) the original attribute was replaced with two new; one holding the absolute value of the original attribute, and one (Boolean) representing whether the attribute value is positive or negative. With this pre-processing, which is a standard technique, it becomes possible to analyze absolute values, disregarding the signs. As an example, for face angle, the

absolute deviation from the target line (a square blade at impact) may very well be more important than whether the club face is pointing to the left or to the right. In addition, the standard deviation over all five shots was also used for each of the original attributes (to measure the consistency of the player) resulting in another 23 attributes.

For the predictive modeling, two different machine learning techniques were used; 100 tree random forests [6] and single regression trees. Random forest is a robust state-of-the-art predictive modeling technique, normally producing very accurate models. Unfortunately, these models are, due to their complexity, opaque. Single regression trees are, on the other hand, interpretable, making it possible to analyse the relationships in the predictive model. All experiments were performed in MatLab, so the regression trees used were built using the MatLab version of CART [7], called *rtree*. In order to produce more comprehensible trees, the built-in pruning procedure was applied, using an internal 5-fold cross-validation. All other parameter values were left at their default vales. For the modeling, three different data sets were used:

- *All* - The 55 input attributes, as described above.
- *Club data* - Only attributes related to the movement of the club, but including their standard deviations.
- *Single shot* - Club and flight data, but no standard deviations.

For the evaluation of the predictive performance, standard leave-one-out cross-validation was used.

IV. Results.

To identify key differences between the four groups described above, and to understand the data better, we start by showing some descriptive statistics. More specifically, a number of Trackman attributes are presented in bar charts below. Each chart shows mean values and (mean) standard deviations for each of the four handicap groups. Naturally, as in most data analysis, a majority of the findings are as expected, i.e., trivial. As an example, Figure 1 below shows that better players have a higher club speed, and that their club speed is more consistent.

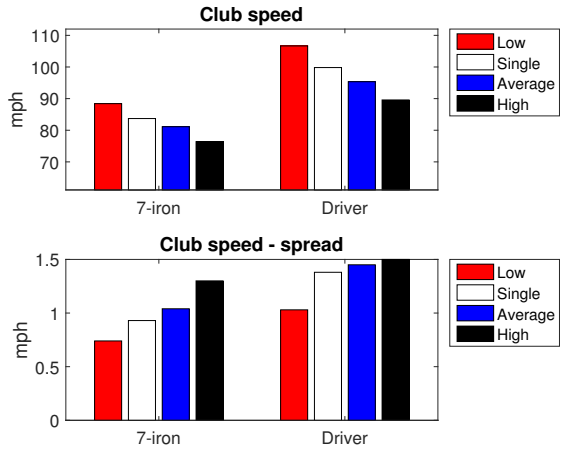

Fig. 1. Club speed

Generally, higher club speeds translate into longer shots, as seen in Figure 2 below. From this bar chart it can be seen that the average carry length for golfers with low handicaps is almost 250 meters with the driver, and close to 165 meters with the 7-iron. For high handicappers, on the other hand, the average driver shot is just over 160 meters, and they hit the 7-iron approximately 125 meters. Interestingly enough, when looking at the standard deviations, the picture is very clear; the lower the handicap, the more consistent is the carry length. A player with a low handicap has a standard deviation of approximately five meters (5.4) with the 7-iron, and less than 10 meters (9.3) with the driver. The corresponding numbers for a high handicapper are 24.1 m for the driver and 11.0 m for the 7-iron.

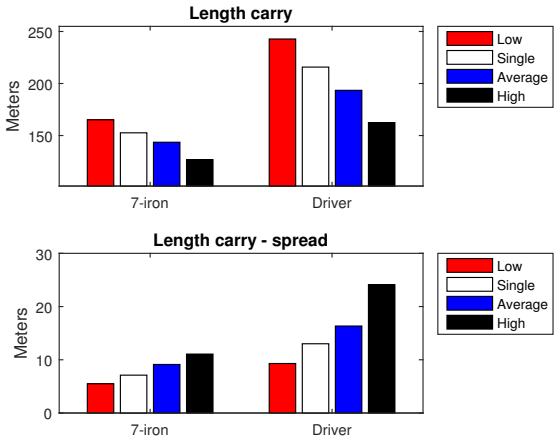

Fig. 2. Carry distance

Looking at a couple of the fundamental properties, Figure 3 below shows that better players tend to have smaller and more consistent face angles. i.e., they are able to regularly start the ball very close to the target line.

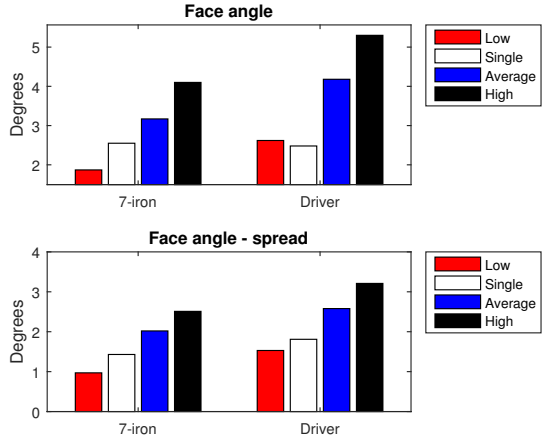

Fig. 3. Face angle

For the other three groups, however, the control of this

vital aspect of a golf shot rapidly decreases. High handicappers clearly struggle with delivering the club face square to the target line, and they are also extremely inconsistent between shots.

Figure 4 below shows the face-to-path attribute. As described above, face-to-path determines the curvature of the shot; the higher the value the more pronounced left-to-right or right-to-left movement. Again, it is obvious that better players are more consistent, but as seen in the chart, they also hit the ball straighter.

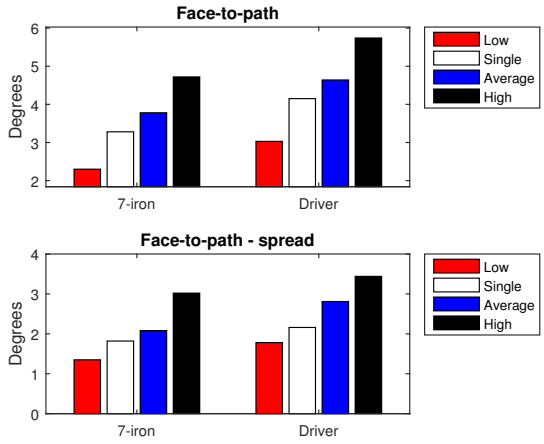

Fig. 4. Face-to-Path

Obviously, there are few surprises so far. Looking at angle-of-attack in Figure 5 below, however, we see that while the results for the 7-iron are as expected, i.e., better players hit more down on the ball, the results for the driver are a bit mixed. Since it is well-known that golfers who want to maximize the carry length should hit up on the ball when using a driver, with recommended values between +2 to +5 degrees, we might have expected the best players to, at the very least, have positive numbers.

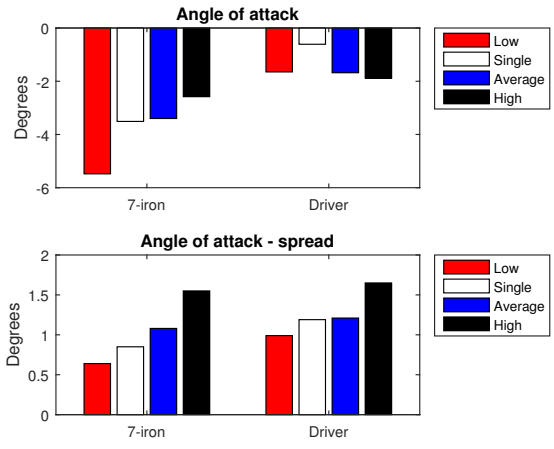

Fig. 5. Angle-of-Attack

The very straightforward explanation is, though, that better players sacrifice some length for increased control, since hitting more up on the ball would also (significantly) increase the side spin. It may be noted that the average angle-of-attack with drivers for players on the PGA tour is -1.3 degrees, while the corresponding number for the women golfers at the LPGA tour is +3.0 degrees, so it appears that the female professionals generally look to increase the shot distance, even at the risk of introducing more side spin.

Figure 6 below shows something very interesting, i.e., that better players tend to have a flatter swing plane. It may be argued that this is another trivial finding - after all beginners are know to have much too upright swings, typically resulting in the dreaded "over-the-top movement" producing weak and sliced shots. Still, what is seen here is that a flatter swing also discriminates the really good golfers (low and single handicappers) from average golfers. Actually, especially when looking at the driver, there is even a difference between the low handicappers and the single handicappers.

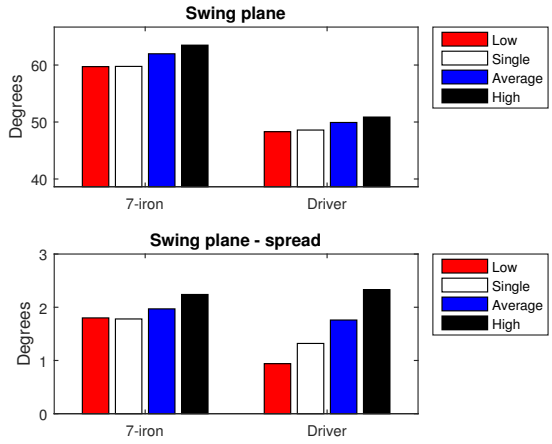

Fig. 6. Swing plane

In the next step, we investigate the basis for the fundamental theory described above. More specifically, we will compare face-to-path values with spin axis values. As described above, a negative fact-to-path should produce a draw, while positive values indicate a fade. The actual ball flight, however, is determined from the spin axis, where, again, a negative value indicates a draw and a positive value a fade. For this analysis, it must be noted that we look at individual shots, i.e., we use five shots per player and club. In addition, we disregard all magnitudes, so we just look at the proportion of all shots that are draws, i.e., have negative values for face-to-path and spin axis, respectively. Figure 7 below shows both the club data, as measured using face-to-path and the actual ball flight measured using spin axis. If the fundamental theory applied, we would expect very similar results. As seen in the chart, however, the differences are huge. According to the fundamental theory, a large majority of all shots should be draws, i.e., the face-to-path is negative, while in fact a much larger proportion of actual ball flights are left-to-right. This should be analyzed further, but there are a couple of possible explanations; first of all, the fundamental theory is somewhat simplified, it does

not, for instance, consider the downward motion of the club, which, however, as long as it is negative, would add to the club path making the face-to-path even more negative. With this in mind, we believe that the major reason for this finding is that a very large proportion of shots are not hit in the center of the club, i.e., a large majority of golfers do not hit the ball solid enough for the basic golf theory to apply.

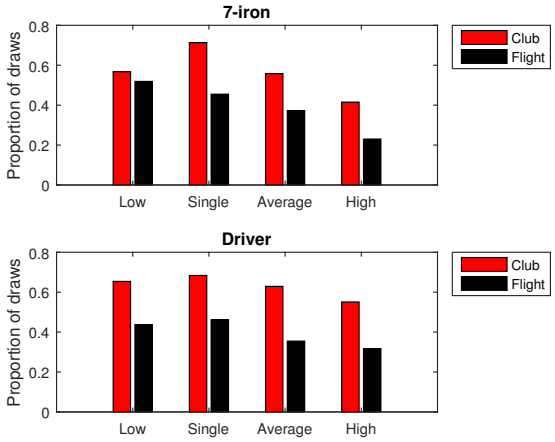

Fig. 7. Club vs. Flight data

Summarizing the descriptive statistics, most findings are as expected, thus validating the approach. Some interesting observations are that better golfers deliver the club head more square to the target (face angle) and have a straighter ball flight (face-to-path). It is also obvious that players with lower handicaps are more consistent, in every aspect. Hitting down with the irons is an indicator of a better player – but the results for the driver are inconclusive. The two most interesting observations are that better players tend to have a flatter swing and that the results indicate that many players don't hit the ball well enough for basic golf theory to apply.

We now analyze which attributes that have the most influence on the handicap. Before presenting correlations between the different Trackman attributes and the handicap we look at a couple of scatter plots. Figure 8 below shows club speed (on the x-axis) and smash factor (on the y-axis).

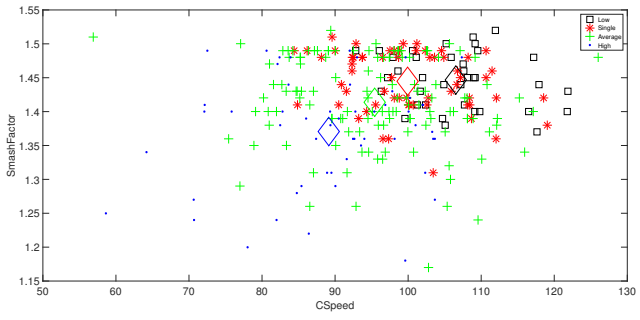

Fig. 8. Club speed and Smash factor using driver

Naturally, each point represents one golfer. In addition, the diamonds show the average value for each group. As expected, low and single handicappers tend to be in the top right corner, i.e., they have a higher club speed but they also hit the ball better. As a side note, the golfer with the highest club speed (and a fairly high smash factor) who actually has a handicap over 9.9, turned out to be a player competing in long-driving.

Figure 9 below shows face-to-path values, both the absolute numbers and the standard deviations. Here, the better players are mostly in the lower left corner, i.e., they have small face-to-path values, and they are very consistent. An interesting observation is that a large majority of all the golfers in the study actually have fairly small values, i.e., clearly under 5 degrees, meaning that their shots are rather straight. High handicappers, however, both have high absolute numbers (they will hit the ball with a lot of side spin) and are very inconsistent, i.e., the ball flight will differ a lot between different shots.

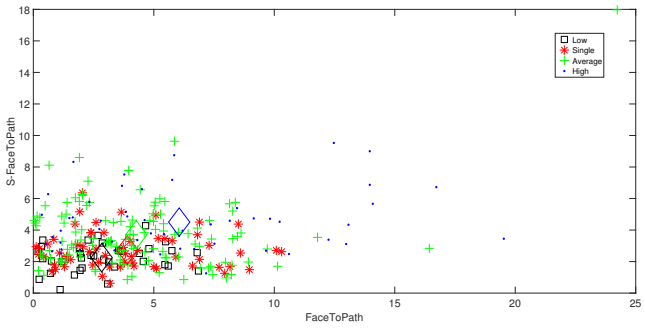

Fig. 9. Face to path - absolute values and standard deviation

Table I below shows Pearson correlation coefficients between a number of Trackman attributes and the handicap. It must be remembered that a low handicap indicates a better player. We immediately see that the single attribute most strongly correlated with the handicap is the carry length with the driver. Club speed and smash factor are also very important. Looking at the more technical attributes, it is very interesting to see that face angle and spin axis are quite important; obtaining similar correlations as side deviation. Again it is interesting to see that a flatter swing plane is a fairly strong indicator of a better golfer. The most important observation is, however, the fact that consistency, especially with regard to key attributes like face angle, face-to-path and smash factor, is so important.

TABLE I. CORRELATIONS BETWEEN ATTRIBUTES AND HANDICAP

	Mean values		Standard deviations	
	Driver	7-iron	Driver	7-iron
Club speed	-0.44	-0.44	0.10	0.28
Length carry	-0.65	-0.55	0.39	0.27
Side deviation	0.22	0.28	0.08	0.24
Height	-0.26	-0.44	0.18	0.23
Smash factor	-0.50	-0.39	0.43	0.41
Angle of attack	-0.07	0.34	0.30	0.35
Swing plane	0.25	0.36	0.41	0.18
Club path	0.10	0.10	0.24	0.22
Face angle	0.37	0.21	0.41	0.52
Face-to-path	0.20	0.23	0.39	0.44
Spin axis	0.29	0.35	0.10	0.38

So, the observations from the initial descriptive analysis are

confirmed; consistency is a key indicator of a better golfer. In addition, more skilled golfers have much better control of the club head (face angle) and the curvature of the shot (face-to-path), and also hit the ball more solid (smash factor). Finally, poor golfers tend to have steeper swing planes.

Turning to the results for the predictive modeling, Table II below shows mean absolute errors and determination coefficients. Starting with the random forest, the model it actually able to explain 50% of the handicap, when allowed to use all data. Naturally, a MAE of 4.67 means that if we take a random golfer, not part of the training data, and let him hit five shots, the model will predict his handicap based on the resulting Trackman attributes, and this prediction will on average differ from his actual handicap with 4.67 shots. This may appear to be a rather inaccurate model, but it must again be remembered that the model only has access to data from the long game, and there are several issues related to using the handicap as a measure of skill. Nevertheless, we also performed a very limited follow-up, where a human expert (a PGA teaching pro) took the role of the model, and tried to guess the handicap for a number of players, based either on just looking at the swings (in slow-motion videos) or by manually analyzing the Trackman data. The result was that the human expert was less accurate than the random forest on average, but much better at estimating the handicaps for really good or really poor players. An interesting observation is that even when restricted to using only club data or when making the prediction based on only one shot, the random forest is still able to obtain very similar performance.

TABLE II. RESULTS PREDICTIVE MODELING

	Random forest				Regression tree			
	Driver		7-iron		Driver		7-iron	
	MAE	R2	MAE	R2	MAE	R2	MAE	R2
All	4.67	0.50	4.76	0.45	5.73	0.30	5.09	0.38
Club data	4.96	0.44	4.79	0.46	5.70	0.27	5.97	0.17
Single shot	4.88	0.43	5.19	0.37	5.79	0.27	6.12	0.18

Figure 10 below shows the eight most important features for 7-iron, as identified by the standard feature importance ranking procedure inherent in random forest.

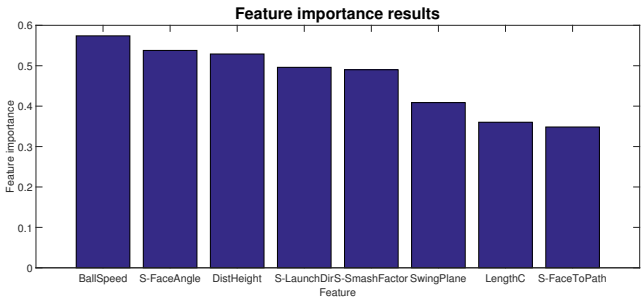

Fig. 10. Feature importance

All features starting with s- are standard deviations, so for the random forest, the most important feature, after ball speed, was face angle consistency. In fact, several of the other most important features also relate to consistency, with regard to face-to-path, smash factor and launch direction. Interestingly enough, this analysis too identifies the swing plane as a very important attribute. The last two attributes found are directly related to shot distance, i.e., LengthC and DistHeight.

Figure 11 below shows an induced regression tree using all data. As indicated by the different colors, four different parts of the tree can be identified. First of all, players with a low ball speed are all predicted to have a fairly high handicap. Looking at the right part, we see that the two key splits both relate to consistency attributes. The best group of players in this tree, consequently, has a high ball speed and are consistent with regard to launch direction and vertical angle, i.e., they are able to repeat the the start of the ball flight between shots.

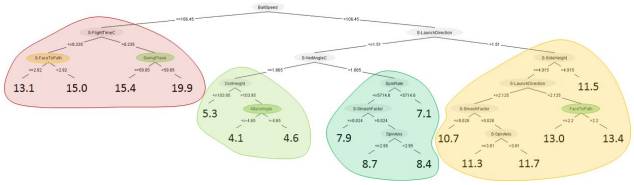

Fig. 11. Regression tree using all available data for 7-iron. MAE 5.1

V. CONCLUDING REMARKS.

The results presented in this paper show that data analysis is a tool that can be used, in different ways, to increase the understanding of what separates skilled golfers from poor players. Specifically, random forest models were able to predict the handicap – based on Trackman values – with performance at the very least comparable to human experts. The interpretable models were slightly less accurate than the opaque, but on the other hand small enough to be comprehensible. Using these predictive models, descriptive statistics and correlation analysis, one key finding was that the most distinguishing property of better golfers is that they are more consistent, in all aspects. Looking at specific attributes, better players demonstrate superior control of the club head at impact and produce straighter shots, as indicated by small face angles and face-to-path values. An interesting and novel observation was that better players also tend to have flatter swing planes. Finally, it was showed that a majority of golfers do not hit the ball solid enough for the basic golf theory to apply.

REFERENCES

[1] M. Broadie, "Assessing golfer performance on the pga tour," *Interfaces*, vol. 42, no. 2, pp. 146–165, 2012.

[2] F. Tuxen, "The secret of the straight shot ii," Trackman News, Tech. Rep., 2009.

[3] A. Fradkin, C. Sherman, and C. Finch, "How well does club head speed correlate with golf handicaps?" *Journal of Science and Medicine in Sport*, vol. 7, no. 4, pp. 465–472, 2004.

[4] M. Sweeney, P. Mills, J. Alderson, and B. Elliott, "The influence of club-head kinematics on early ball flight characteristics in the golf drive," *Sports Biomechanics*, vol. 12, no. 3, pp. 247–258, 2013.

[5] N. F. Betzler, S. A. Monk, E. S. Wallace, and S. R. Otto, "Variability in clubhead presentation characteristics and ball impact location for golfers' drives," *Journal of sports sciences*, vol. 30, no. 5, pp. 439–448, 2012.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.