Final Project Report

Predicting the Golf Player’s Performance on a PGA Tour Tournament

BIA 652 Multivariate Data Analytics

May 20, 2021

Shashank Chaudhary, Nishitha Reddy Dodda, Hara Naga Teja Batchu

Table of Contents

Introduction and Problem statement………………………………………………………………... 3

Model Approach …………….......……….……………………………….………………………… 5

Exploratory data analysis.......……….……………..…………………………………………… 9

Data cleaning and Variable description……..…….……………………………………………… 12

Pre-processing and Feature selection ....…...……..………………………………………….... 15

Model Analysis ……………………………………………………………….……………18

Recommendation…………………………………………………………………………...21

References……………………………………………………………………………….22

INTRODUCTION

The PGA-Tour is the most established and prestigious golf tournament circle in world. Its tournaments are mostly hosted in the United States, but also in other countries, e.g. Australia, China and England. The total purse for the 2021-2022 season of the PGA-Tour is over 400 million USD. (pgatour.com, 2021) Similar to other sports, betting on the outcomes of golf tournaments is huge business. However, due to the nature of golf tournaments, it is historically very difficult to predict the outcomes of tournaments. For example, in any 2- player tennis match you can only have two outcomes: Player One or Player Two. When betting on the winner of this tennis match you only need to look at the odds for two players. On the other hand, in PGA-Tour tournaments you will have usually between 132 and 156 participants competing for the title over the four rounds played during the tournament. This makes it very difficult to predict the winners of each PGA-Tour tournament.

Golf is played outdoors on a golf course. A golf course typically consists of 18 holes, though 9 hole courses exist where each hole is played twice. A player has completed a round of golf when they have played all 18 holes. Each hole is given a notional score called “Par” which can be considered to be the typical number of strokes (the number of times a golfer hits the ball with a club) that a good player should take in order to complete the hole. If a player takes more strokes than par for the hole, they have completed the hole “over par”. Similarly, if they complete the hole in less stokes than par they have completed the hole “under par”. All holes on a golf course will have par scores of 3, 4 or 5 strokes, which typically indicate the length of the hole - the lower the par score, the shorter the distance between the tee box and the front of the green.

PROBLEM STATEMENT

Golf has been a thoroughly studied game since its inception over 300 years ago. Every amateur and even every professional ask how they can improve their game, lower their scores, and in doing so, make more money or have more fun playing the game.

Surprisingly enough, a select handful of researchers have tried to answer this very innocuous question in an extremely scientific manner. Using multiple regression analysis independent researchers have found similar results through a multitude of different variables including driving distance, driving accuracy, putts per round, sand saves, eagles, birdies, top ten finishes and of course tour winnings.

In the years since the different dataset has been made available to the academic community, it appears that the scope of research carried out using the data has been of a narrow focus. Examples include the creation of new statistics, or creating new ranking lists for events that have already happened in previous seasons. The key research problem of this dissertation is to assess if the application of Data Analytical techniques to the collected data can predict the performance of professional golfers in a tournament.

Through this project, we are aiming to explore more on attribute selection to have perfect fit model and compare the results. We want to run through different linear regression models and compare the accuracy of each. We also intend to use logistic regression to predict if the player is a potential winner or not.

Performance Prediction Part of the challenge of this project will be to create suitable metrics that accurately describe an association which is pertinent to the problem to be solved, for example predicting if a golfer will miss the cut or not. This will be a significant part of the experiment and the identification of key metrics, either derived or transformed from the source data, or the original data items, will be the key to success of the experiments.

Psychology Another of the main challenges will be the creation of successful models that may not be able to take into account any psychological aspects of performance. It is theorized that by measuring average past performance, it will smooth out any "bad days" (outliers) in a golfer's performance. However, it will not be possible to incorporate the likelihood of a golfer having once-off issues, or possibly, the beginning of ongoing psychological / off-the-course pressures such as those experienced by Rory McIlroy in the 2013 season.

MODEL APPROACH

Regression Analysis

Regression is a way of predictive modeling using statistical methods that help us to analyze and understand the relationship between two or more variables of interest. It is a process of getting the response variable as a function of different attributes that matter in its prediction. It is a conglomerative analysis which includes the process right from feature selection, model fitting, Prediction, finding the accuracy of the model which makes it long lasting.

There are different types of regression analysis like linear regression, polynomial regression, logistic regression, discriminant analysis. For this dataset from PGA tour, we have tried different regression approaches like Lasso, Ridge and Logistic Regression, and ultimately found the logistic regression best fits the dataset.

Logistic Regression

A statistical model typically used to model a binary dependent variable with the help of logistic function. It establishes a relationship between dependent and independent variables. Another name for the logistic function is a sigmoid function and is given by:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1}$$

This function assists the logistic regression model to squeeze the values from (-k, k) to (0, 1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification. In our analysis, we majorly focused on the binary classification of being a winner in the tournament or not.

The reason behind this model is that just like Linear Regression, logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task.

So it all start with a linear function p(x) and then using log function with p(x) we are able to bound this function to 0 to 1. So the function will be like

$$\log\frac{p(x)}{1-p(x)} = \alpha_0 + \alpha . x$$

Since Logistic regression predicts probabilities, we can fit it using likelihood. Now the likelihood can be written as:

$$L(\alpha_0, \alpha) = \prod_{I=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i}$$

Further, after putting the value of p(x):

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} -\log 1 + e^{\alpha_0+\alpha} + \sum_{i=0}^{n} y_i(\alpha_0 + \alpha . x_i)$$

In order to increase the probability of occurring we can use Maximum Likelihood function and differentiating the equation with respect to different parameter and setting it to zero.

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=0}^{n} \big(y_i - p(x_i; \alpha_0, \alpha)\big) x_{ij} = 0$$

Assumptions:

- The dependent variable is categorical. Dichotomous for binary logistic regression and multi-label for multi-class classification
- Attributes and log odds i.e. log (p / 1-p) should be linearly related to the independent variables
- VIF eliminates multicollinearity
- In binary logistic regression class of interest is coded with 1 and other class 0
- The players with >6 “top 10” score would be the probable winner and hence been indexed the binary value of 1

The objectives of this project is to derive value and insight from the use of Machine Learning techniques in order to predict professional golfer performance - specifically whether or not a golfer will have top 10 finish or win.

- These objectives will be achieved via completion of the following milestones:
- Prepare and transform the collected data for use.

- Generate “Golfer Analytical Records” from which models will be trained
- Design and build predictive models.
- Design and implement statistical experiments.
- Evaluate the success or otherwise of the models and experiments deriving insight from the results.

MODEL ACCURACY

As an important step of any predictive modelling, we have verified the accuracy of our model using the following techniques. Here is our understanding on the same.

Mean Squared Error:

The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs.

Confusion Matrix:

It is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. This consist of four different parts:

- True Positive (TP) - These are cases in which we predicted yes, and that’s actually yes.
- True Negative (TN) - We predicted no, and it’s actually no.
- False Positive (FP) - We predicted yes, but they actually was no.(Type 1 Error)
- False Negative (FN) - We predicted no, but actually it is yes.(Type II Error)

ROC Curve and AUC:

An ROC Curve (receiver operating characteristic curve) is a graph showing the performance of a classification model with the help of True Positive Rate and False Positive Rate at different classification thresholds. AUC stands for Area under the ROC curve measures the entire two-dimensional area underneath the entire ROC curve.

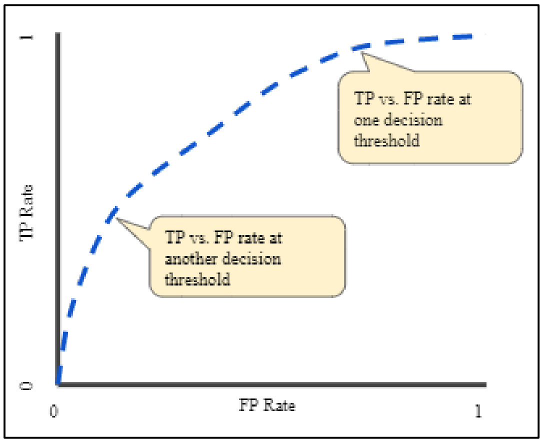


Figure 1: ROC Curve

EXPLORATORY DATA ANALYSIS

Data Set Description

The analyzed dataset is collected using scrape dataset from the kaggle website , (https://www.espn.com/golf/statistics/_/year/2020 and PGA Tour official website (https://www.pgatour.com/content/pgatour/stats/stat.120.y2020.eoff.t033.html).

Attribute	Data Type	Explanation
Player Name	STRING	Name of Player
Rounds	INT	Total Number of Rounds played
Fairway Percentage	FLOAT	Percentage of time ball lands on fairway
Year	INT	Year of Rounds played
Avg Distance	FLOAT	Average Driving Distance
Gir	FLOAT	Green is Regulation
Average Putts	FLOAT	Average number of putts taken
Average Scrambling	FLOAT	Average number of time par made when player miss the Green
Average Score	FLOAT	Average Score of all the rounds
Points	FLOAT	Total number of points assigned for all tournaments
Wins	INT	Total number of wins
Top 10	FLOAT	Total number of Top 10 finish in the tournament
Average SG Putts	FLOAT	Strokes gained: putting measures how many strokes a player gains (or loses) on the greens.
Average SG Total	FLOAT	Strokes gained: putting measures how many strokes a player gains (or loses) on the greens.
SG:OTT	FLOAT	Strokes gained: off-the-tee measures player performance off the tee on all par-4s and par-5s.
SG:APR	FLOAT	Strokes gained: approach-the-green measures player performance on approach shots. Approach shots include all shots that are not from the tee on par-4 and par-5 holes and are not included in strokes gained: around-the-green and strokes gained: putting. Approach shots include tee shots on par-3s.
SG:ARG	FLOAT	Strokes gained: around-the-green measures player performance on any shot within 30 yards of the edge of the green. This statistic does not include any shots taken on the putting green.
Money		The amount of prize money a player has earned from tournaments

Table 1: Attributes in the data set

We have collected data for PGA tour for 2010-19 season. Our datasets have top performing golf players with 18 attributes, which are explained below. We have changed Top 10, wins and year to int datatype.

The dataset consists 2312 rows each representing the various attributes as described in the previous section and their corresponding scores. The description statistics of all the attributes are as follows:

Feature	count	mean	std	min	25%	50%	75%	max
Rounds	1678	78.71	14.27	45	69	79.5	89	120
Fairway Percentage	1678	61.44	5.06	43.02	57.94	61.43	64.91	76.88
Year	2312	2014	2.58	2010	2012	2014	2016	2018
Avg Distance	1678	290.81	8.92	266.4	284.9	290.55	296.4	319.7
gir	1678	65.66	2.75	53.54	63.83	65.79	67.58	73.52
Average Putts	1678	29.16	0.52	27.51	28.81	29.14	29.52	31
Average Scrambling	1678	58.12	3.38	44.01	55.9	58.28	60.42	69.33
Average Score	1678	70.92	0.7	68.7	70.49	70.9	71.34	74.4
Points	2296	481.66	463.07	1	113	381.5	676	4169
Wins	293	1.22	0.57	1	1	1	1	5
Top 10	1458	2.78	1.9	1	1	2	4	14
Average SG Putts	1678	0.03	0.34	-1.48	-0.19	0.04	0.26	1.13
Average SG Total	1678	0.15	0.7	-3.21	-0.26	0.15	0.57	2.41
SG:OTT	1678	0.04	0.38	-1.72	-0.19	0.06	0.29	1.49
SG:APR	1678	0.07	0.38	-1.68	-0.18	0.08	0.32	1.53
SG:ARG	1678	0.02	0.22	-0.93	-0.12	0.02	0.18	0.66
Money	2300	1124903	1354085	5520	185325.8	699442.5	1526660	12030465

Table 2: Description statistics of attributes in data set

We have arrived at the distribution of all the attributes. We observed that Rounds, fairway percentage, average distance, girl, average putts, average scrambling, average score are all normally distributed.

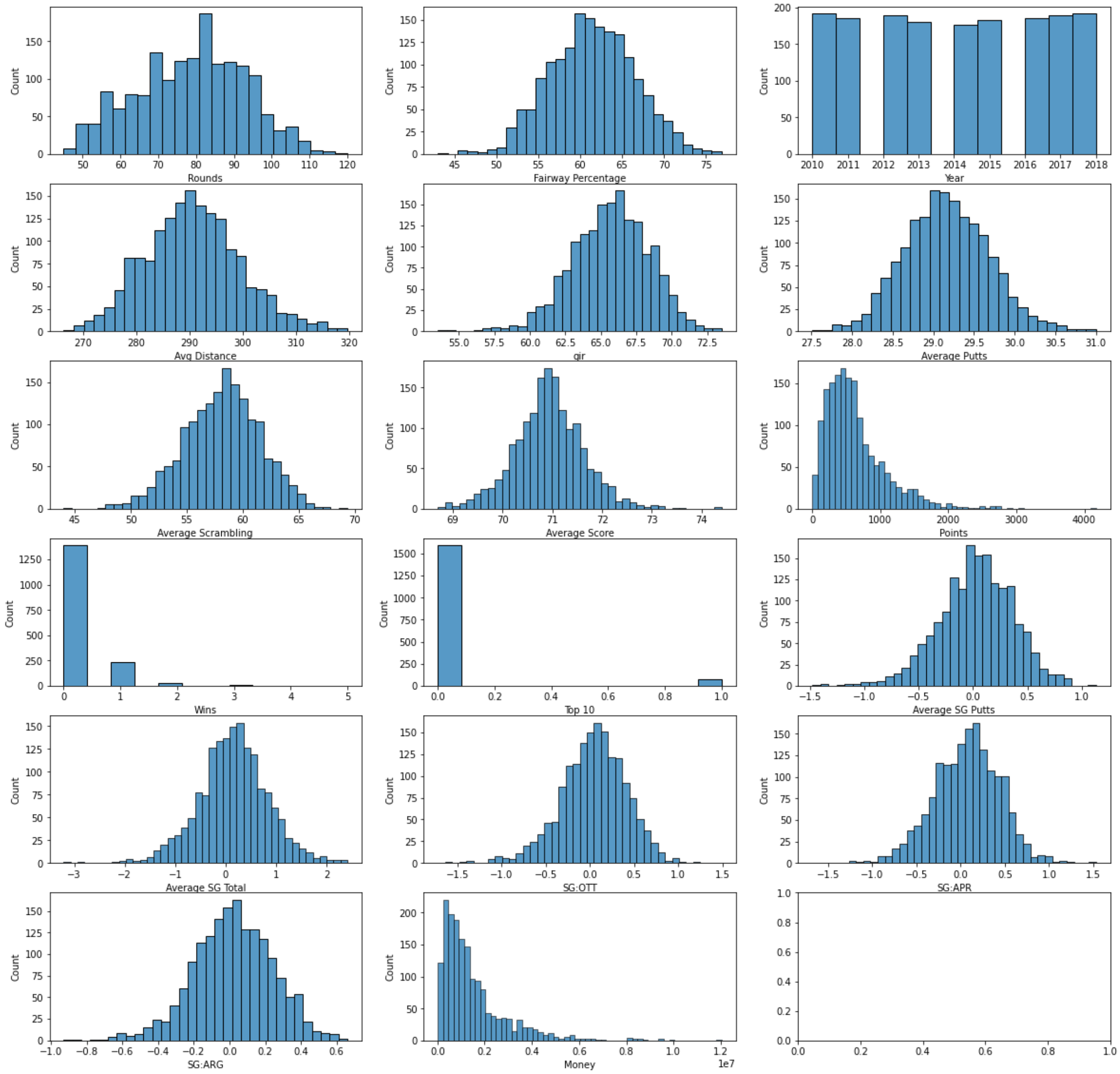


Figure 2: Distribution of each attribute

DATA CLEANING

As a part of data cleaning process, we have handled all the Nan values by replacing them by 0 and changing the data type of Top 10, Wins and Rounds to int. To confirm the cleaned data, we checked the info of the data, which displays the data type of each attribute. For Nan, we calculated the sum of Nan in the complete data set after cleaning, which came down to 0.

Feature	Sum(Nan)
Player Name	0
Rounds	0
Fairway Percentage	0
Year	0
Avg Distance	0
gir	0
Average Putts	0
Average Scrambling	0
Average Score	0
Points	0
Wins	0
Top 10	0
Average SG Putts	0
Average SG Total	0
SG:OTT	0
SG:APR	0
SG:ARG	0
Money	0

Table 3: Sum of Nan after data cleaning

INDEPENDENT AND DEPENDENT VARIABLES

Dependent Variable:

This is a variable which is used as a resultant variable in the model. This is predicted using all the independent variables. In our case dependent variable is "**Top 10**"

Finishing top 10 in an event represents a very positive performance for a Tour player.

Independent Variable:

1. Average Distance

- The average driving distance is key parameter for deciding the performance of golfer. Accurate and long hit can save shots for players.

2. GIR (Greens in Regulation)
   - The reason for this is because your chances of making a par (or better) dramatically increase when your ball is on the putting surface versus being in the rough or a sand trap
3. Average Putts
   - Putting is important. Regardless of skill level, putting accounts for approximately 43 percent of your total strokes, taking into account your good putting days and the ones where you're ready to snap your flagstick over your knee. Lower this percentage and your scores will go down
4. Average Scrambling
   - The Scrambling stat was developed to measure how often a golfer avoids bogey after missing the green with their approach shot. That requires a golfer to chip or pitch on to green and then hole a par putt. Scrambling is simply calculated by dividing successful scrambles (par or less) by total GIR missed.
5. Average Score
   - All the score of the rounds played by a golfer divided by number of rounds. Lower average score indicates that golfer is good golfer and vice versa.
6. Average SG putts
   - Strokes gained: putting measures how many strokes a player gains (or loses) on the greens.
7. Average SG Total
   - Strokes gained: putting measures how many strokes a player gains (or loses) on the greens.
8. SG OTT
   - Strokes gained: off-the-tee measures player performance off the tee on all par-4s and par-5s.
9. SG APR
   - Strokes gained: approach-the-green measures player performance on approach shots. Approach shots include all shots that are not from the tee on par-4 and par-5 holes and are not included in strokes gained: around-the-green and strokes gained: putting. Approach shots include tee shots on par-3s.

10. SG ARG

- Strokes gained: around-the-green measures player performance on any shot within 30 yards of the edge of the green. This statistic does not include any shots taken on the putting green.

Visualization:

Below we have independent variable named **Avg Distance**, **Average Scrambling**, **Average Putts**, **Average Score** and **gir,** that has a relation with the dependent variable **Top 10 finishes.**

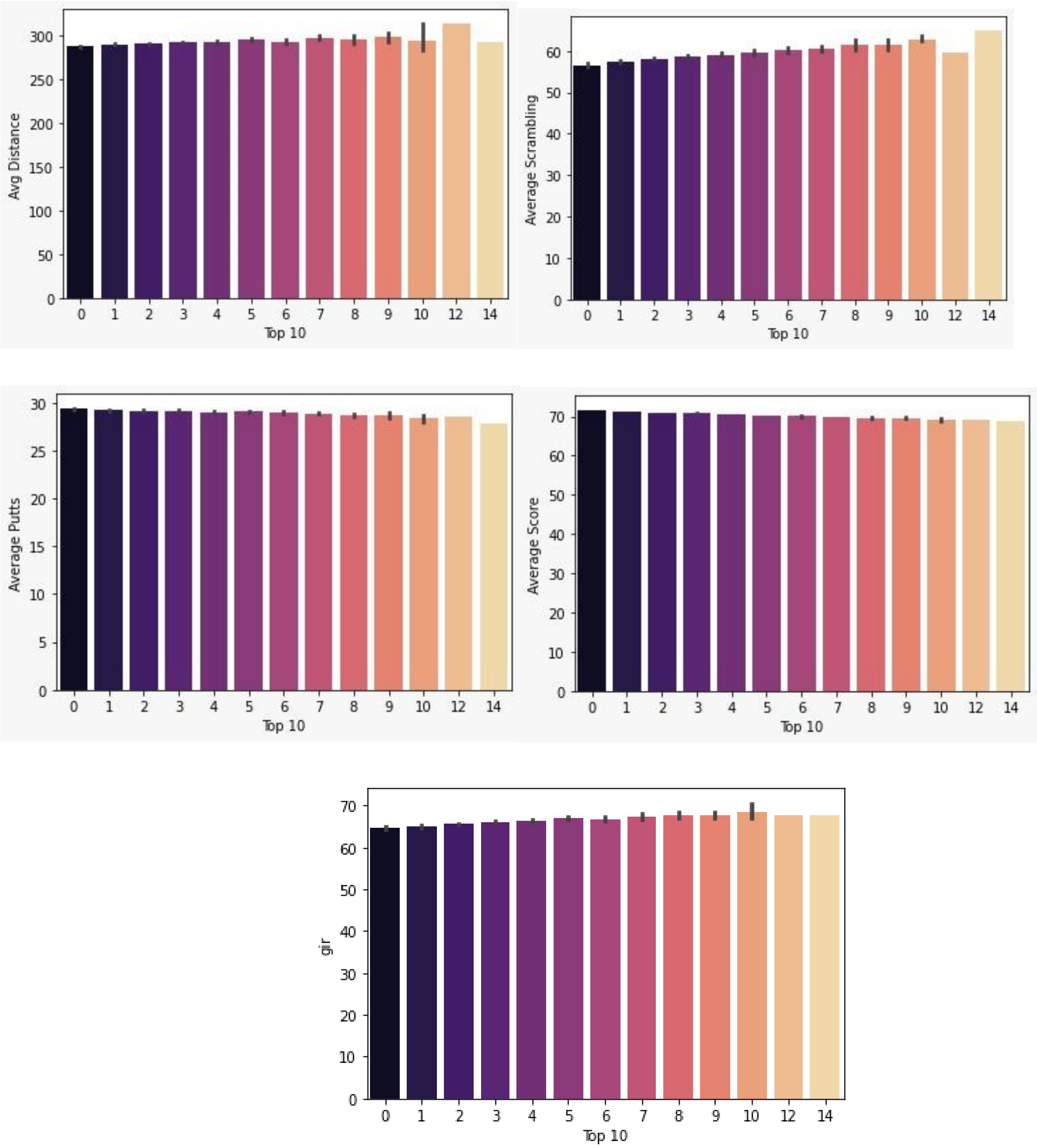


Figure 3: Attribute variation with response variable

PRE-PROCESSING

Transforming Variable:

- According to exploratory data analysis we can observe that not all the categories of dependent variable ('quality') has sufficient amount of data. Thus it would lead to unnecessarily complex analysis.
- As we analyzed the data further, we could see that there are very few numbers of players who have the highest number of Top 10, to eliminate this discrepancy or any kind of biasness, we have categorized the Top 10 score of 0 to 6 into one category and the rest to other and found the total number in each of these categories.
- Thus it is suitable to change the variable to binary which is if Top10 finish more than 6 it will be considered Top players otherwise Average players (based on 8 years of data). This can be changed based on number of years' data considered.
- Hence the final distribution of the dependent variable ('Top 10') is

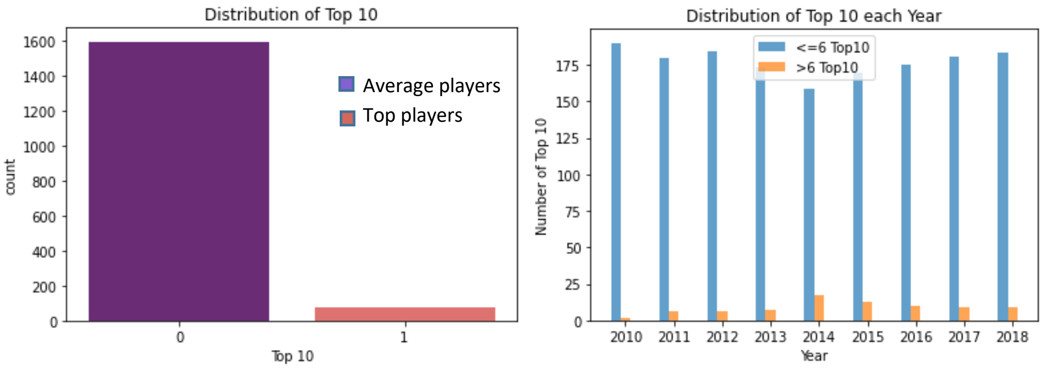


Figure 4: Distribution of top 10

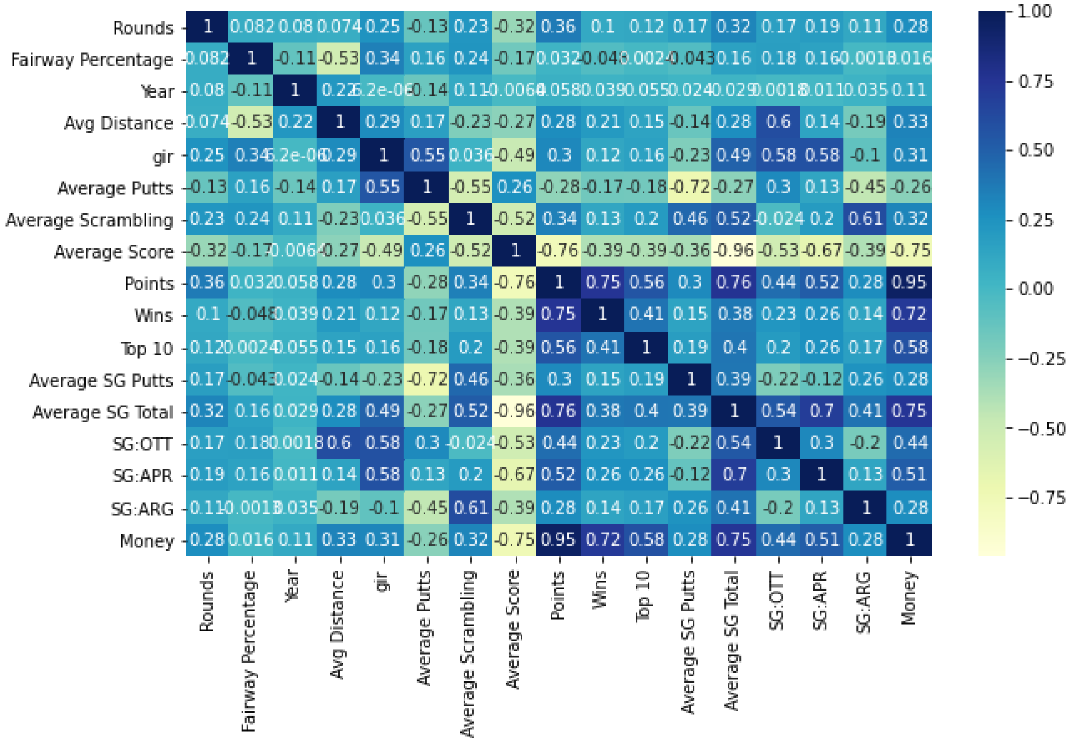
FEATURE SELECTION:

Correlation Matrix:

It is defined as the covariance between two variables divided by the product of the standard deviations of the two variables.

$$\boldsymbol{\rho(X,Y) = \frac{COV\ (X,Y)}{\sigma_X \sigma_Y}}$$

The value of ***ρ*** lies between -1 and +1. Values near +1 indicate strong positive relation and -1 represents strong negative relation.



	Rounds	Fairway Percentage	Year	Avg Distance	gir	Average Putts	Average Scrambling	Average Score	Points	Wins	Top 10	Average SG Putts	Average SG Total	SG:OTT	SG:APR	SG:ARG	Money
Rounds	1	0.082	0.08	0.074	0.25	-0.13	0.23	-0.32	0.36	0.1	0.12	0.17	0.32	0.17	0.19	0.11	0.28
Fairway Percentage	0.082	1	-0.11	-0.53	0.34	0.16	0.24	-0.17	0.032	-0.048	0.0024	-0.043	0.16	0.18	0.16	-0.0018	0.016
Year	0.08	-0.11	1	0.22	6.2e-06	-0.14	0.11	-0.0064	0.058	0.039	0.055	0.024	0.029	0.0018	0.011	0.035	0.11
Avg Distance	0.074	-0.53	0.22	1	0.29	0.17	-0.23	-0.27	0.28	0.21	0.15	-0.14	0.28	0.6	0.14	-0.19	0.33
gir	0.25	0.34	6.2e-06	0.29	1	0.55	0.036	-0.49	0.3	0.12	0.16	-0.23	0.49	0.58	0.58	-0.1	0.31
Average Putts	-0.13	0.16	-0.14	0.17	0.55	1	-0.55	0.26	-0.28	-0.17	-0.18	-0.72	-0.27	0.3	0.13	-0.45	-0.26
Average Scrambling	0.23	0.24	0.11	-0.23	0.036	-0.55	1	-0.52	0.34	0.13	0.2	0.46	0.52	-0.024	0.2	0.61	0.32
Average Score	-0.32	-0.17	-0.0064	-0.27	-0.49	0.26	-0.52	1	-0.76	-0.39	-0.39	-0.36	-0.96	-0.53	-0.67	-0.39	-0.75
Points	0.36	0.032	0.058	0.28	0.3	-0.28	0.34	-0.76	1	0.75	0.56	0.3	0.76	0.44	0.52	0.28	0.95
Wins	0.1	-0.048	0.039	0.21	0.12	-0.17	0.13	-0.39	0.75	1	0.41	0.15	0.38	0.23	0.26	0.14	0.72
Top 10	0.12	0.0024	0.055	0.15	0.16	-0.18	0.2	-0.39	0.56	0.41	1	0.19	0.4	0.2	0.26	0.17	0.58
Average SG Putts	0.17	-0.043	0.024	-0.14	-0.23	-0.72	0.46	-0.36	0.3	0.15	0.19	1	0.39	-0.22	-0.12	0.26	0.28
Average SG Total	0.32	0.16	0.029	0.28	0.49	-0.27	0.52	-0.96	0.76	0.38	0.4	0.39	1	0.54	0.7	0.41	0.75
SG:OTT	0.17	0.18	0.0018	0.6	0.58	0.3	-0.024	-0.53	0.44	0.23	0.2	-0.22	0.54	1	0.3	-0.2	0.44
SG:APR	0.19	0.16	0.011	0.14	0.58	0.13	0.2	-0.67	0.52	0.26	0.26	-0.12	0.7	0.3	1	0.13	0.51
SG:ARG	0.11	-0.0013	0.035	-0.19	-0.1	-0.45	0.61	-0.39	0.28	0.14	0.17	0.26	0.41	-0.2	0.13	1	0.28
Money	0.28	0.016	0.11	0.33	0.31	-0.26	0.32	-0.75	0.95	0.72	0.58	0.28	0.75	0.44	0.51	0.28	1

Figure 5: Correlation Matrix

- From the above matrix we will find values which are greater than ±0.5 to remove variables which are highly correlated as it will generate multicollinearity.
- We could see that the combinations, Money-Points, points-wins, points-money, have high correlation with one another hence must be treated or removed.

- Correlation with Dependent Variable (‘Top 10’): From the table to the right, we can observe the correlation strength of each variable with the dependent variable (‘Top 10’).
- We have eliminated the highly correlated attributes and least correlated attributes with respect to the “Top 10”
- This brings us to the following attributes: Average SG Total, Average score, SG-APR, SG-OTT, Average Scrambling, Average SG Putts, Average putts, GIR, Average distance.

Feature	Top 10
Top 10	1
Money	0.581556
Points	0.560281
Wins	0.407319
Average SG Total	0.396365
Average Score	0.393548
SG: APR	0.258632
SG: OTT	0.201581
Average Scrambling	0.196535
Average SG Putts	0.185977
Average Putts	0.180111
SG: ARG	0.168018
Gir	0.162325
Avg Distance	0.148674
Rounds	0.116192
Year	0.054823
Fairway Percentage	0.002379

Table 5: Correlation with Top 10

Variance Inflation Factor:

A variance inflation factor is basically a tool to help identify the degree of multicollinearity. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables or inputs.

VIF scores for our dataset variables:

- We further carried out the VIF calculation as a part of attribute selection and elimination of multicollinearity. From the table on right, we have discarded the features with high VIF and kept the remaining. The final list of features selected are : Average scrambling, average distance, GIR, average score, average putts.

feature	VIF
Average SG Putts	5.609118
SG:OTT	6.131054
SG:APR	7.568245
Average SG Total	19.29994
Average Scrambling	744.0964
Avg Distance	1845.826
Gir	2509.048
Average Score	16347.82
Average Putts	17410.79

Table 6: VIF

Variable Selection:

Ultimately, we have reached at the point where we could decide on which variables needs to be selected for model building.

- So using the above pre-processing techniques we conclude the list to be

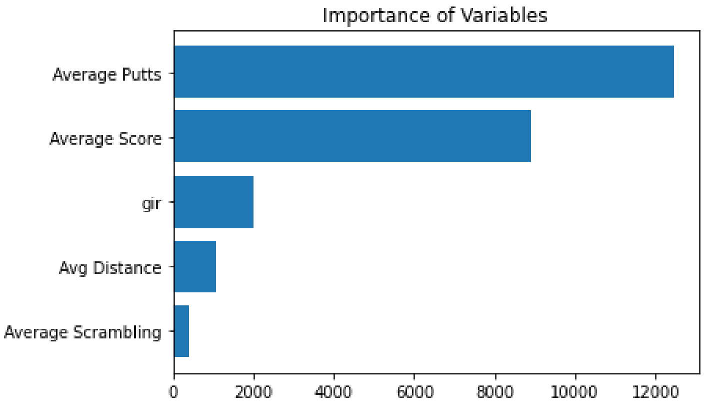


Figure 6: Variable importance through VIF

- Average putts
- Average score
- Greens in Regulation
- Avg. Distance
- Average Scrambling

ANALYSIS

Our logistic regression model which gave more accuracy than Lasso, ridge, OLS has arrived at the Mean Squared error of 0.0286 which is small and considered to be good, leads to a score of 0.957. Thus using the model, we are 96% sure to predict the top 10 players of the tournament.

- To check the bracket of score we have used cross-validation matrix on score. Cross validation matrix runs my model in broken chunks of data and gives the accuracy scores on each chunk. We got the bracket of 0.94 and 0.96. That shows our score is in the good range.

- Equation model is reached with the help of coefficient for each variable as well as the intercept:

$$\boxed{Y = -5.388144 - 0.00904975(\text{Avg Distance}) + 1.25189086 * (\text{gir}) - 1.48484432(\text{Average Putts}) - 0.27642216\ (\text{Average Scrambling}) + 1.86061881\ (\text{Average Score})}$$

Confusion Matrix:

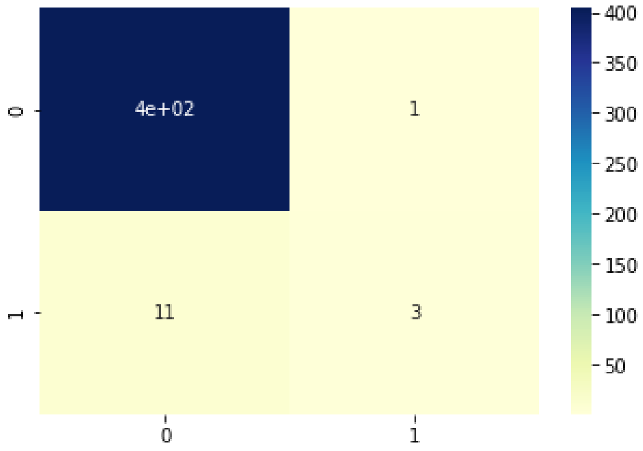


Figure 7: Confusion matrix

This represents-

- True Negative – 404
- False Positive – 1
- False Negative – 11
- True positive – 3

By seeing this we can say that we have correctly predicted 407(404+11) and 4 (1+3) wrongly predicted. Thus the accuracy score is 0.95 which is the score for the model.

Classification Report:

Using the classification report we have got more detailed analysis:

- Precision (91%), this means out of all values of quality we were able to predict 91% times correctly.

- Recall (96%), this means 96% of the actual values are predicted correctly.
- F1 score is 97% w
- which is the average of precision and recall.

ROC Curve and AUC:

- Using True Positive and False positive rate we have the ROC curve on different classification threshold.
- AUC is 0.84 for this model that means overall we are able to classify and differentiate the quality very well.

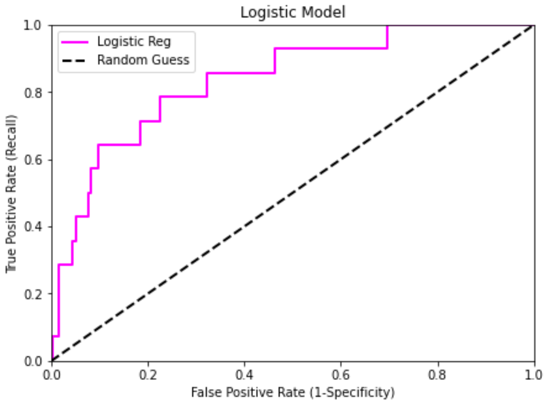


Figure 8: ROC Curve

- Below figure explains how the Precision of the model varies as the recall increases(True positive rate). This is because it overshadows the true negative and leads to declining precision.

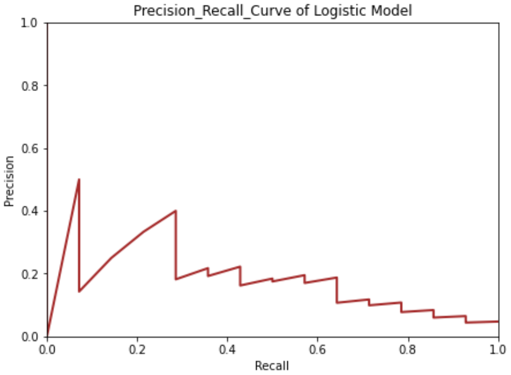


Figure 9: Precision vs Recall

INFERENCE

We have collected data for top 150 player for 2021 to predict the top 10 player for the US PGA Tournament (20th May -24th May, 2021). This is one of the most prestigious golf tournament. We have used our logistic regression model to predict the top 10 players for the tournament. This model can be extended to predict top 30 or top 50 players in the tournament.

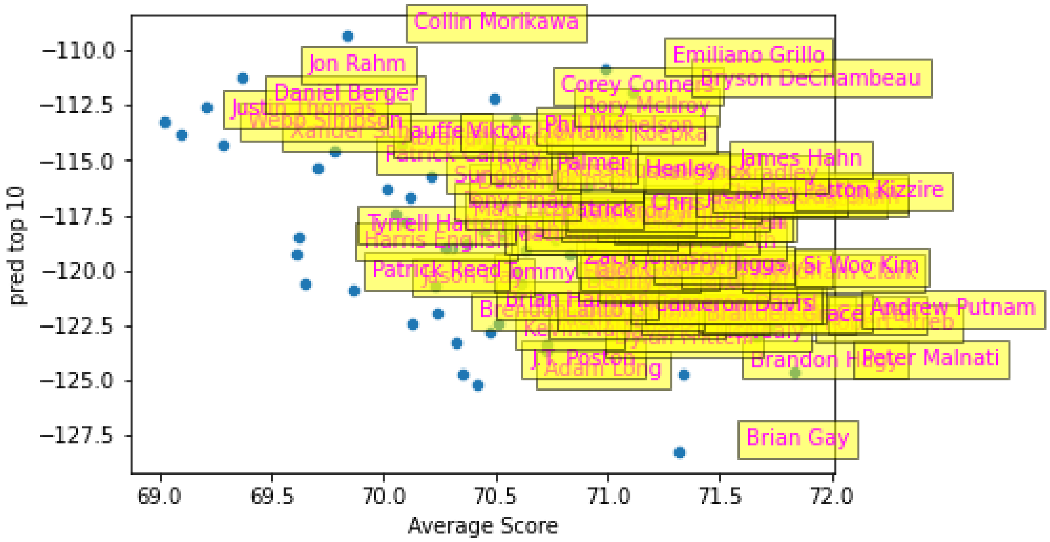


Figure 10: Avg. Score v/s Chance of winning

These are the top 10 player prediction for US PGA Tournament.

Brooks Koepka

Xander Schauffele

Viktor Hovland

Phil Michelson

Webb Simpson

Justin Thomas

Rory McIlroy

Daniel Berger

Corey Conners

Bryson DeChambeau

Jon Rahm

RECOMMENDATION

- Finally, we are ready with the equation using the variables through which we can easily predict the performance of a golfer in a tournament.
- This model can be made more useful if we are able to add the cost of false positive and true negative predictions. So we are able to optimize the cost and improve the values of one among them which is less costly using different classification thresholds.
- This model is specific to a PGA tournament data, thus to make it more universally applicable it needs to incorporate data from on European tour and Asian Tour played different region of world. Lot of player plays on multiple tours so it will help us to be more realistic in our prediction.
- Model is solely dependent on the different parameters of data on players but if we are able to incorporate course difficulty, weather and some parameter of luck, it would have been more realistic and acceptable.

REFERENCES

[1] Leahy, B., 2014. Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data.

[2] Menard, S., 2002. Applied logistic regression analysis (Vol. 106). Sage.

[3] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression. New York: Springer-Verlag.

[4] Peng, C.Y.J., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.

.

[5] Narkhede, S., 2018. Understanding auc-roc curve. Towards Data Science, 26, pp.220-227.

[6] Lewis, K.P., 2004. How important is the statistical approach for analyzing categorical data? A critique using artificial nests. Oikos, 104(2), pp.305-315.

[7] www.pgatour.com

[8] www.espnstar.com

[9] https://golfanalytics.wordpress.com/?s=avg+score

[10] https://datagolf.com/predictive-model-methodology/

[11] https://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares