PGA Tour Analysis

Which aspects of a pro golfer’s game leads to the most success?

The Data

Brad Klassen has scraped all PGA tour statistics from 2010-2018 from their official website and published them in a csv on Kaggle. The dataset contains:

- 3053 golfers
- 2081 variables
- 528 statistics

Narrowing the Scope

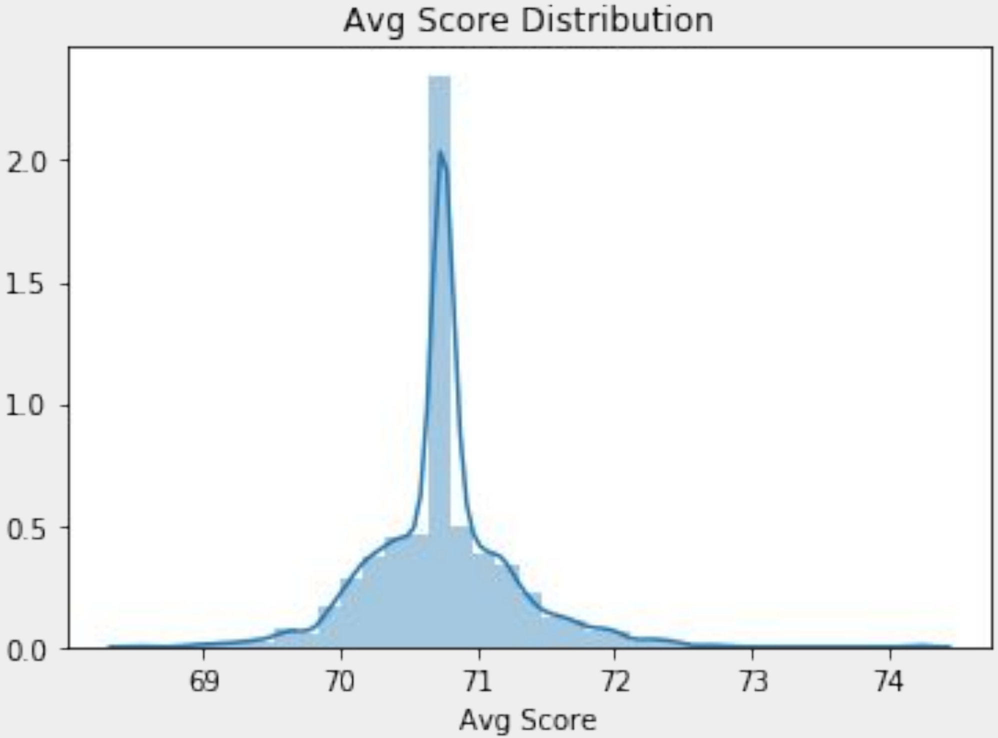
Narrowing the Scope

- Which aspects of a golfer’s game leads to the most success?
- Top 200 golfers by money earned
- Filled nulls with the mean

Exploring the Target Variable:
Average Score

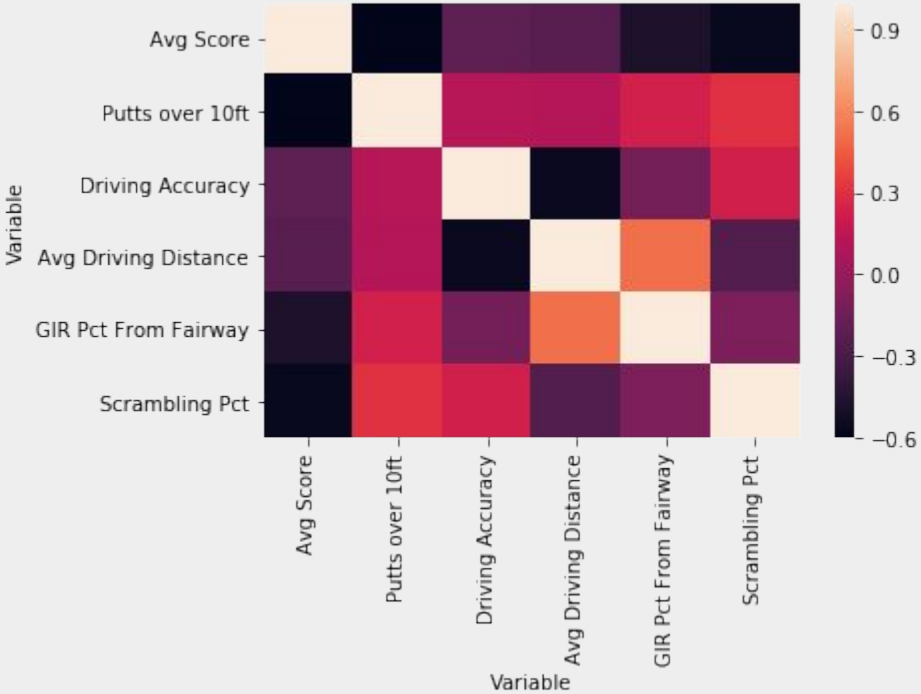
Average Score

Count	1711
Mean	70.75
Min	68.54
Max	74.22
Std	0.5



Breaking Down the Problem:
Feature Selection

Feature Selection: Aspects of a Golf Game



Feature Correlations to Avg Score

Putts over 10ft	-0.60
Scrambling Pct	-0.56
GIR pct from Fairway	-0.48
Avg Driving Distance	-0.23
Driving Accuracy	-0.21

Feature Selection: Aspects of a Golf Game

- Avg number of putts per round was not a good feature
- Putts over 10’ differentiates the pros
- Driving distance and accuracy are surprisingly low
- Some collinearity among the features because each golf shot affects the next one

Feature Correlations to Avg Score

Putts over 10ft	-0.60
Scrambling Pct	-0.56
GIR pct from Fairway	-0.48
Avg Driving Distance	-0.23
Driving Accuracy	-0.21

Choosing the Model

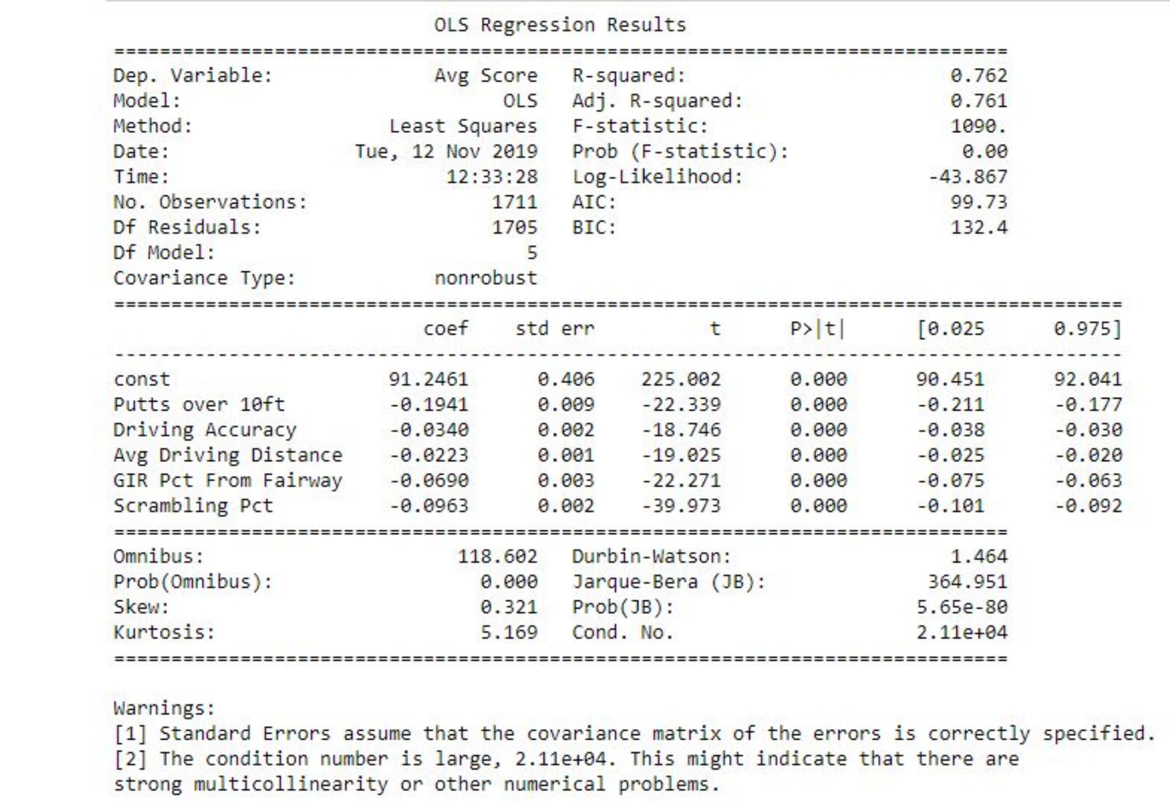
Why Linear Regression?

- All continuous variables
- Explanatory power is more important
- Small set of features
- We can still make useful predictions

Evaluating Model Performance

Evaluating the Model

- ❏ 76% of the variance of our target variable can be explained by our features
- ❏ Confident we’re not overfitting
- ❏ F-statistic p value is close to zero
- ❏ Collinearity among the features is an issue as each golf shot affects the next one.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Avg Score   R-squared:                       0.762
Model:                            OLS   Adj. R-squared:                  0.761
Method:                 Least Squares   F-statistic:                     1090.
Date:                Tue, 12 Nov 2019   Prob (F-statistic):               0.00
Time:                        12:33:28   Log-Likelihood:                -43.867
No. Observations:                1711   AIC:                             99.73
Df Residuals:                    1705   BIC:                             132.4
Df Model:                           5
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   91.2461      0.406    225.002      0.000      90.451      92.041
Putts over 10ft         -0.1941      0.009    -22.339      0.000      -0.211      -0.177
Driving Accuracy        -0.0340      0.002    -18.746      0.000      -0.038      -0.030
Avg Driving Distance    -0.0223      0.001    -19.025      0.000      -0.025      -0.020
GIR Pct From Fairway    -0.0690      0.003    -22.271      0.000      -0.075      -0.063
Scrambling Pct          -0.0963      0.002    -39.973      0.000      -0.101      -0.092
==============================================================================
Omnibus:                      118.602   Durbin-Watson:                   1.464
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              364.951
Skew:                           0.321   Prob(JB):                     5.65e-80
Kurtosis:                       5.169   Cond. No.                     2.11e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.11e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Interpreting the Coefficients for Practical Use

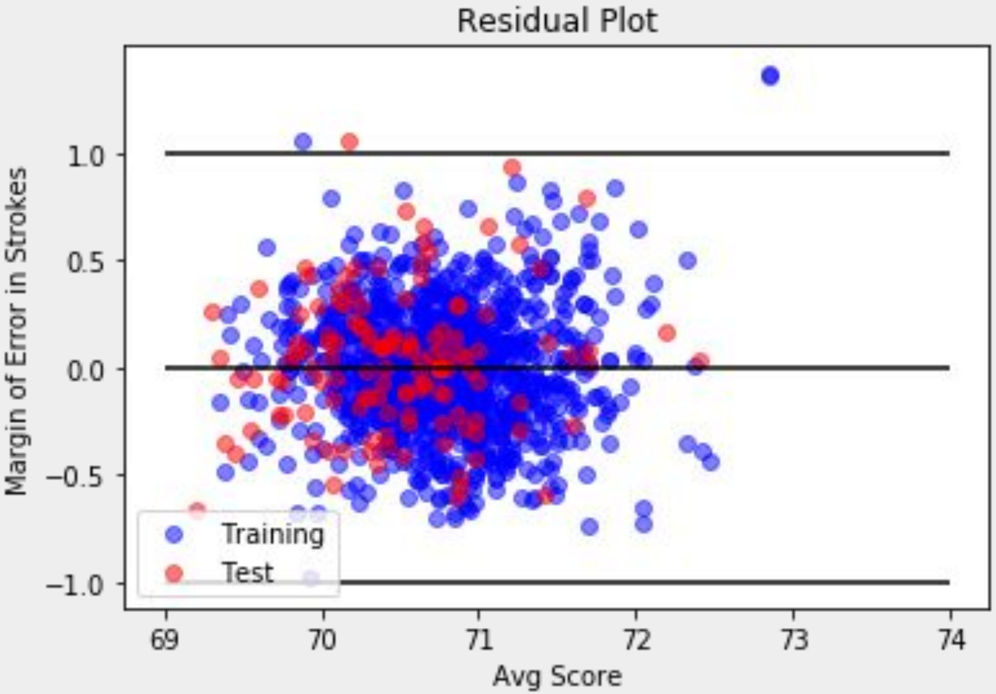
In order to increase avg score by 1 stroke do one of the following:

- Increase **driving accuracy** by about **30%**
    - Very Difficult, 7.2 standard deviations
- Increase **avg driving distance** by about **50 yds**
    - Very Difficult, 6.6 standard deviations
- Increase **GIR pct** from fairway by about **17%**
    - Very Difficult, 7.3 standard deviations
- Increase **scrambling pct** by about **10%**
    - Much more reasonable, 3.6 standard deviations
- Increase **putts made over 10’** by **1 putt per round**
    - Seems obvious, but this is by far the best opportunity to improve versus the rest of the field, less than 1 standard deviation

Making Predictions

Making Predictions

- **Training Set:**
  - 2010 - 2017 data
  - R-squared = 0.754
- **Test Set**:
  - 2018 data
  - R-squared = 0.786
  - Mean absolute error = 0.172



Conclusions

Conclusions

- Practicing long putting is the best use of time
- Driving distance and accuracy are not good indicators of a successful golfer
- Golf shots have an inherent collinearity
- The model is a good start for sports betting predictions
    - Is it more useful than simply using avg score itself?
    - Could narrow it to these features by specific golf course, weather conditions, etc.