Predicting End of Year Golf Points

By: Kolton Fowler, Akshat Johari, Prakhar Bansal, Shreya Bhootda

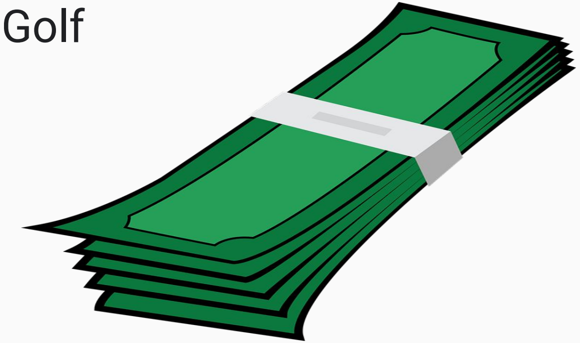
Introduction of Dataset

- Our data set contains ≈ 250 golfers and their playing statistics for each year from 2010-2018.
- We will be using these statistics in order to predict the end of year points in 2018 golf season.
- Our measurement statistics account for shot distance, shot accuracy, strokes taken, and average score per round.

Problem & Expected Impact

Intended Audience

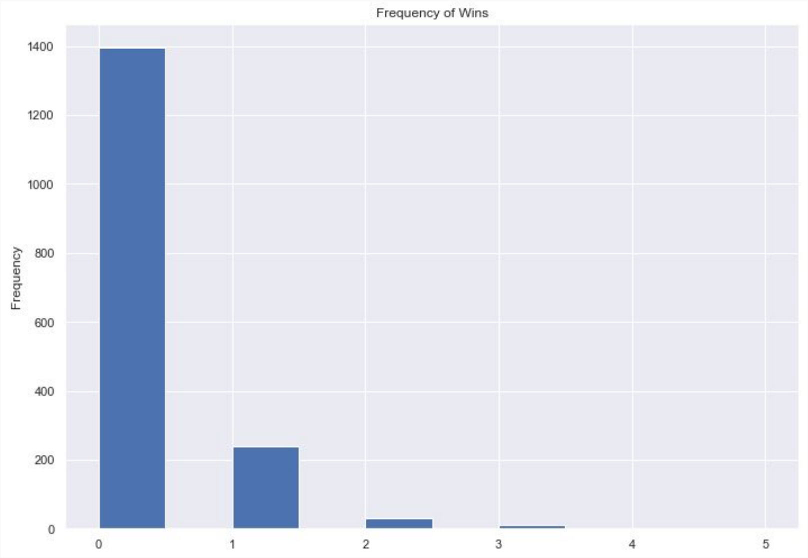
- Golfers
- Sports Bettors
- Fans of Golf



Impact

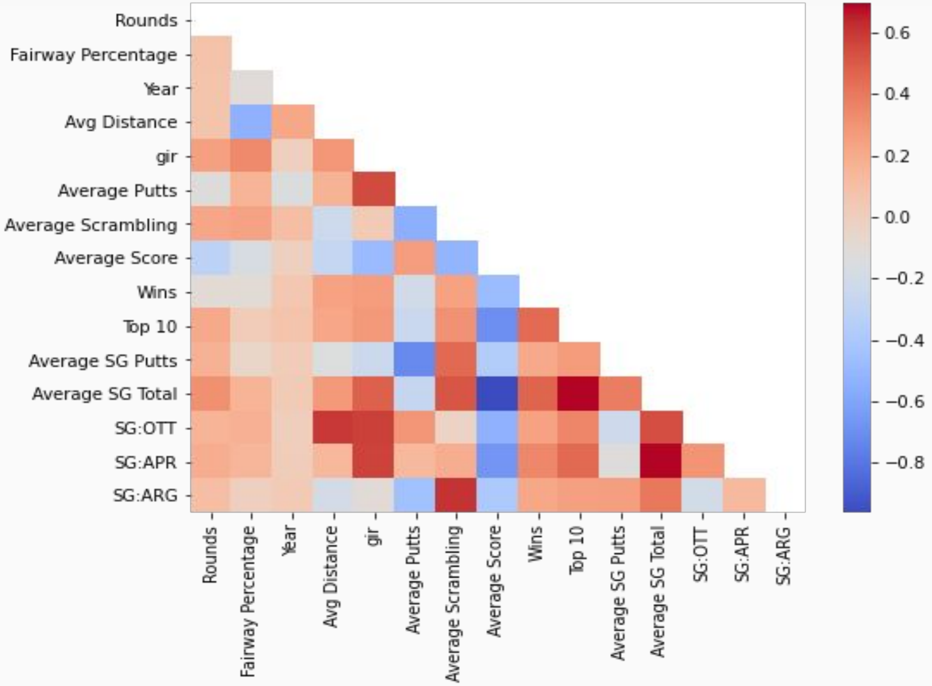
- Better predict Golfer’s points and rank
- Can help marketing/sport agencies make decisions regarding their investments.
- Can help players understand the key contributors to their points

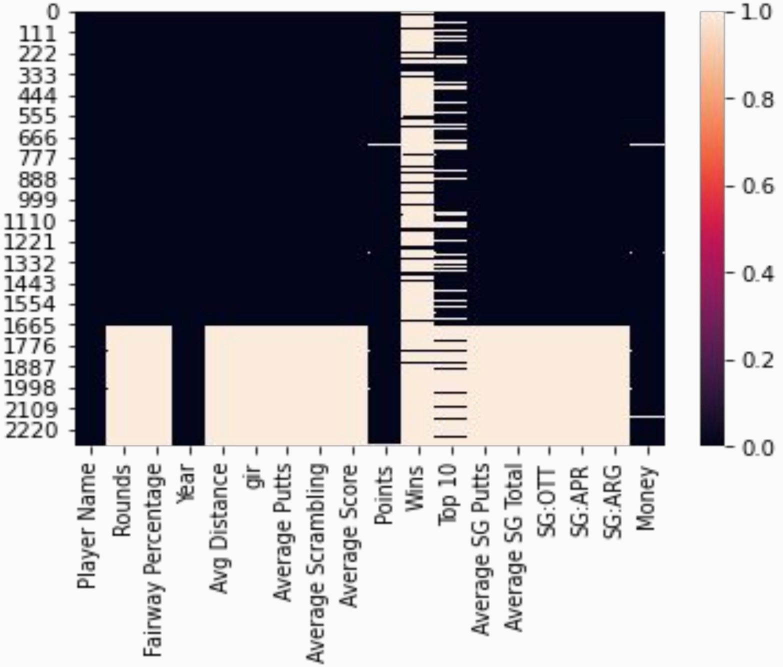
Exploratory Data Analysis/Cleaning



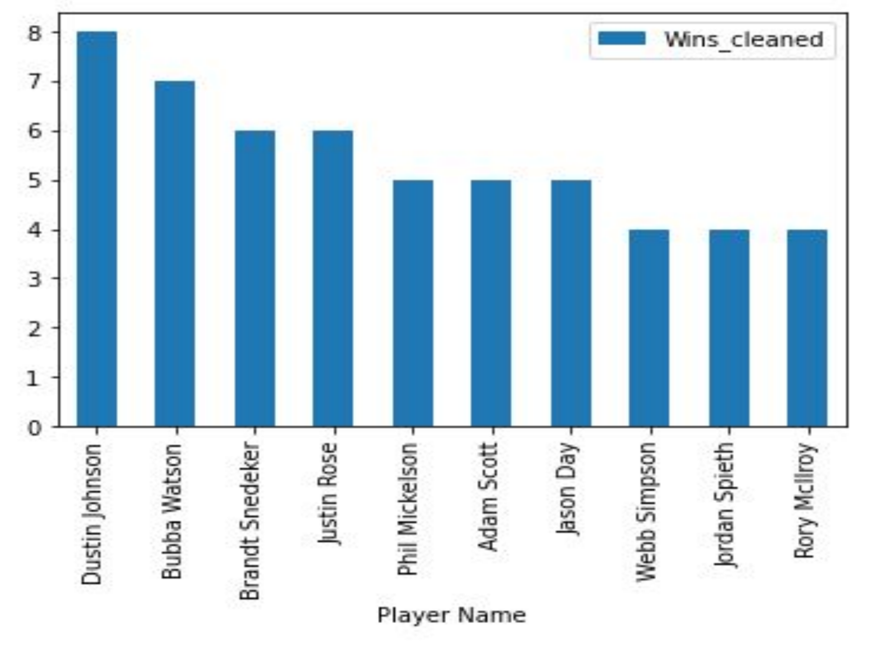
Dataset Characteristics	
Number of Variables	18
Number of Observations	2312
Missing Cells	25.3%
Duplicate Rows	0
Categorical Variables	4
Numerical Variables	14

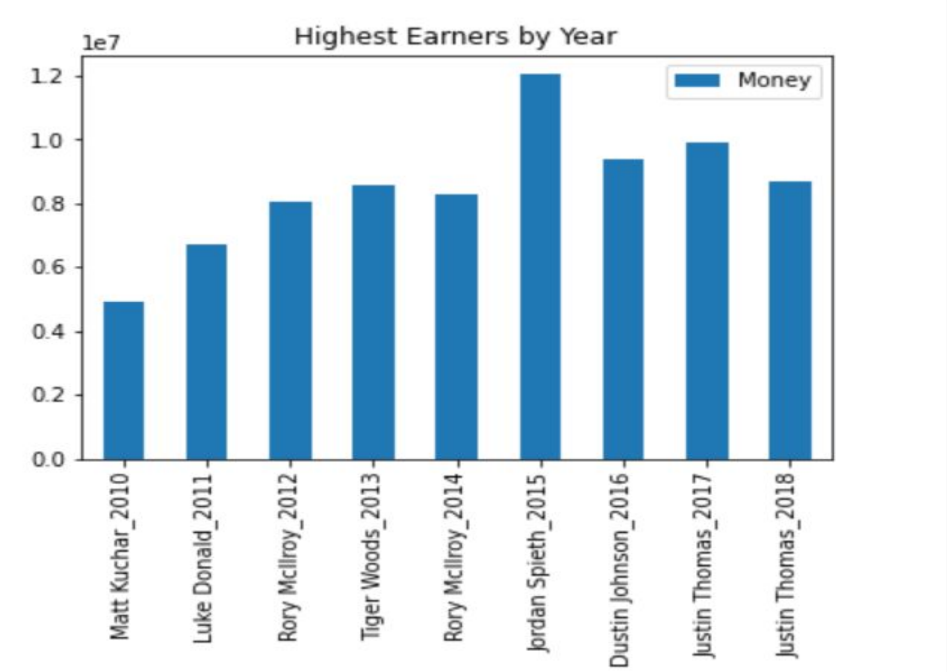
Exploratory Data Analysis/Cleaning



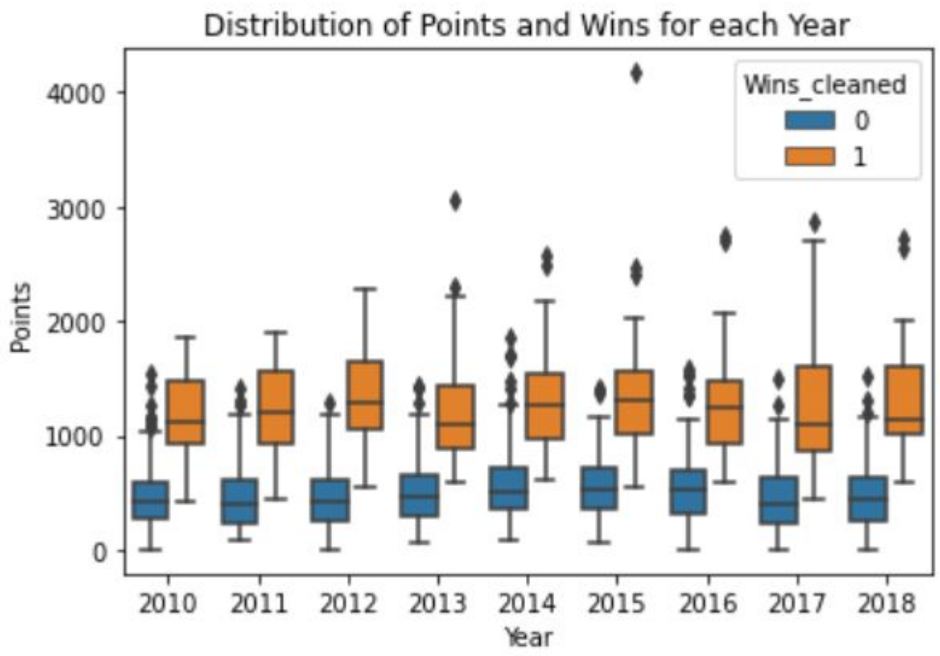


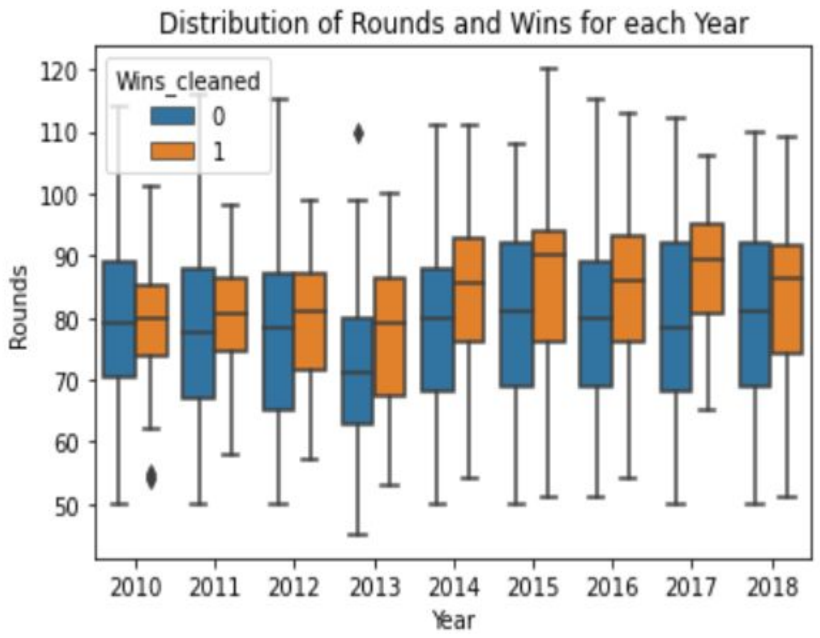
Variable Relationships



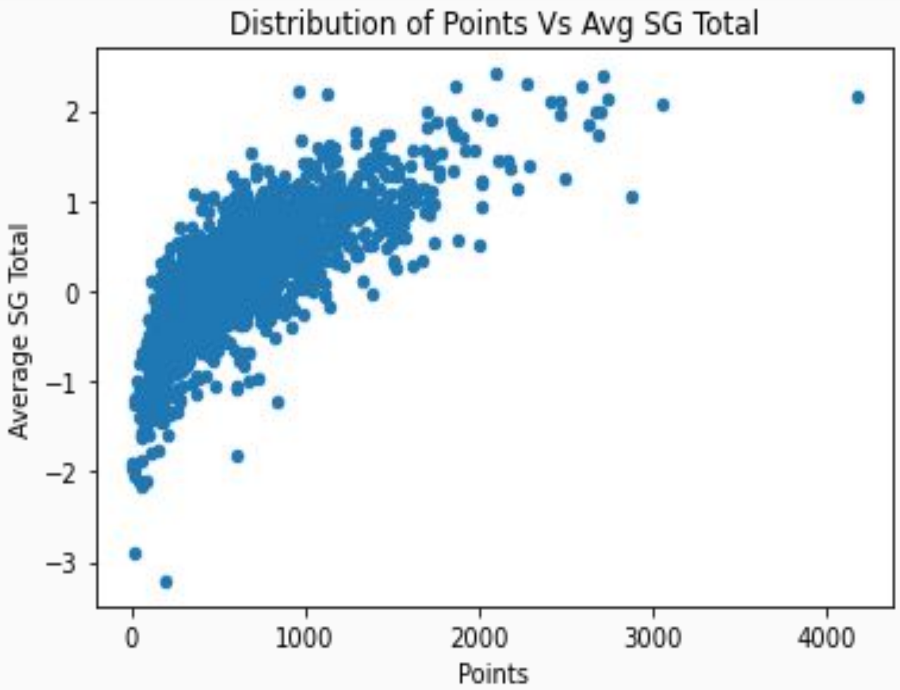


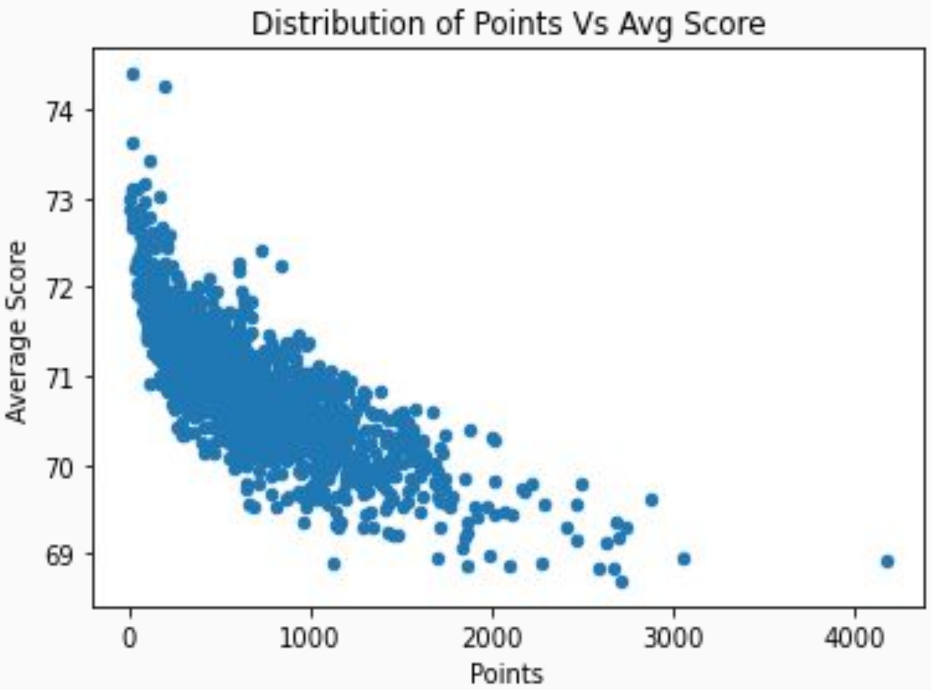
Exploratory Data Analysis/Cleaning





Exploratory Data Analysis/Cleaning





Multiple Linear Regression model

Variable	P-value (0.01)
Average Distance	0.424
Fairway Percentage	0.026
Greens in Regulation	0.038
Average SG Total	0.109
Average SG Putting	0.045
SG on Approach Shots	0.014
SG Around the Green	0.028

Without the Interaction Terms:

- R-Squared: 0.6433
- RMSE: 282.866

With the Interaction Terms:

- R-Squared: 0.7195
- RMSE: 239.982

Shows the true effect of some X's is dependent on other X's

Random Forest

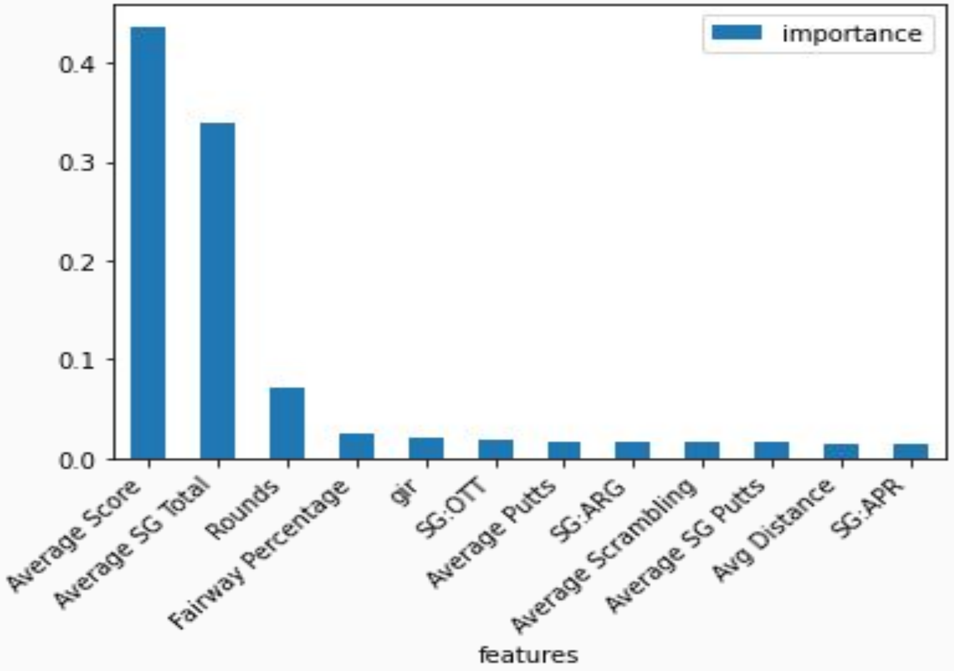
- Random Forest to identify the non-linear relationships between the predictors and points.
- Hyperparameter tuning using Randomized Search CV.
- Grid Search CV (3 folds) with values concentrated around hyperparameters identified by random search.

```
In [175]:  rf_random.best_params_

Out[175]: {'n_estimators': 1200,
           'min_samples_split': 2,
           'min_samples_leaf': 2,
           'max_features': 'sqrt',
           'max_depth': 20,
           'bootstrap': True}
```

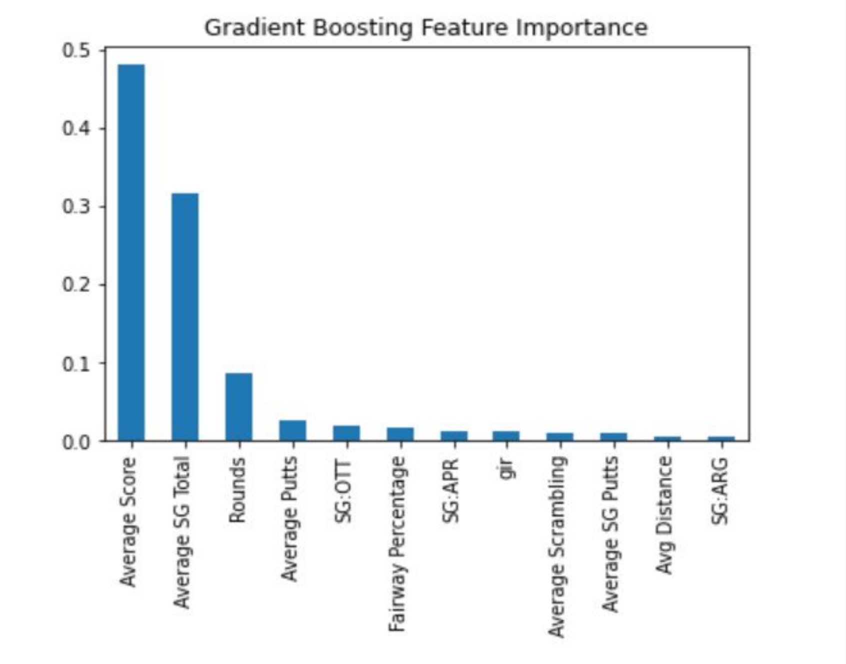
Random Forest

- GridSearch evaluated 240 fits
- RMSE - 239.383
- R2 - 73.986%
- Features Importance :



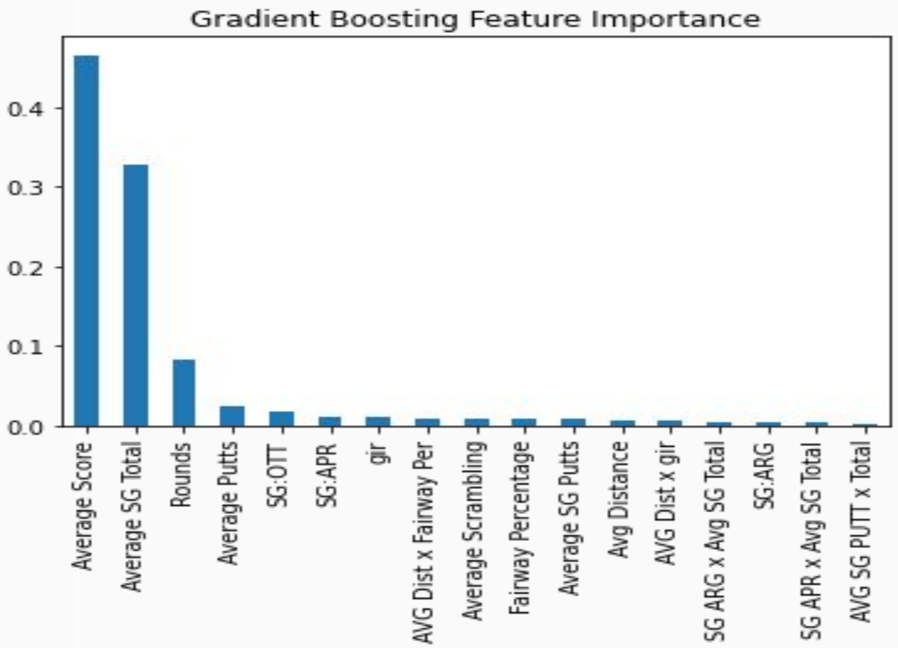
Gradient Boosting

- Hyperparameters Tuned:
  - n_estimators - 100, 200, 500, 1000
  - max_depth - 3, 5, 7, 10
  - min_sample_leaf - 1, 2, 4
- 5-fold Cross Validation to select best parameters
  - n_estimators - 100
  - max_depth - 3,
  - min_sample_leaf - 1
- Test R2 - 76.6%
- Test RMSE - 226.68



Gradient Boosting with Interaction features

- Hyperparameters Tuned:
  - n_estimators - 100, 200, 500, 1000
  - max_depth - 3, 5, 7, 10
  - min_sample_leaf - 1, 2, 4
- 5-fold Cross Validation to select best parameters
  - n_estimators - 100
  - max_depth - 3,
  - min_sample_leaf - 1
- Test R2 - **77.34%**
- Test RMSE - **223.38**



Conclusion

- Gradient Boosting model is the most accurate predictive model for the data set.
- “Rounds”, “Average SG Total” and “Average Score” have the most impact on a player’s points.
- Future scope : Statistics broken down by tournament would increase the prediction accuracy of our model.

Model	RMSE	R Square
Linear Regression	282.86	.6433
Linear Regression with Interactions	239.98	.7195
Random Forest - CV	239.38	.7398
Gradient Boosting - CV	226.68	.7667
Gradient Boosting - CV with Interactions	223.38	**.7734**

Questions?