FAIRFIELD UNIVERSITY

**Predictive Modeling of Golf streaks using ShotLink® Data**

Thesis by

**Jeffrey Kramer**

Department of Software Engineering

Submitted in partial fulfillment of the requirements for the degree of

Master of Science, Software Engineering

Advisor\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee member\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Committee member\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

MSSE Program Director\_\_\_\_\_\_\_\_\_\_\_\_

## *ABSTRACT*

Golf is a difficult game. One can practice and play their whole life and never reach their potential with any consistency. Yet for everyone that is played, watched or followed golf there are times in which the quality of play markedly exceeds the norm. This period of time, which we shall refer to as a ‘streak’, is what this thesis will provide insight into. The ShotLink system has generously made available shot-by-shot data of PGA TOUR players which we shall use to build a predictive model, scoring engine, and player based system to illuminate specific criteria golfers can apply to get on and stay on streaks of high quality play.

***TABLE OF CONTENTS***

Abstract1

Table of Contents2

List of Figures4

Chapter 1: Introduction5

1.1 Problem Statement6

1.2 Motivation7

1.3 Contributions10

1.4 Thesis Outline10

Chapter 2: Background11

Chapter 3: Methodology13

3.1 Dataset14 3.1.1 Events14

3.1.2 Rounds16

3.1.3 Holes17

3.1.4 Shots18

3.1.5 Courses19

3.2 Datamining and Analyzing19

Chapter 4: System Design20

4.1 High Level Overview20

4.1.1 Client Server Based System20

4.2 Server Components21

4.2.1 Restful Controller23

4.2.2 Scoring Engine23

4.2.3 Data Classification Model26

4.2.4 Preprocessing27

4.2.5 Training Set27

4.2.6 Model Fitting28

4.3 Technology28

4.4 User Interface29

Chapter 5: Tests and Results31

Chapter 6: Conclusions and Future Work32

6.1 Conclusions32

6.2 Future Work / Timelines33

6.2 Potential Publishers34

References35

***LIST OF FIGURES:***

Figure 1: Course History vs Performance6

Figure 2: Full Architecture10

Figure 3: Server Components22

Figure 4: Restful Controller23

Figure 5: Scoring Engine24

Figure 6: Data Classification and Model Generation26

Figure 7: Preprocessing27

Figure 8: Training Set27

Figure 9: Model Fitting28

Figure 10: UI Input29

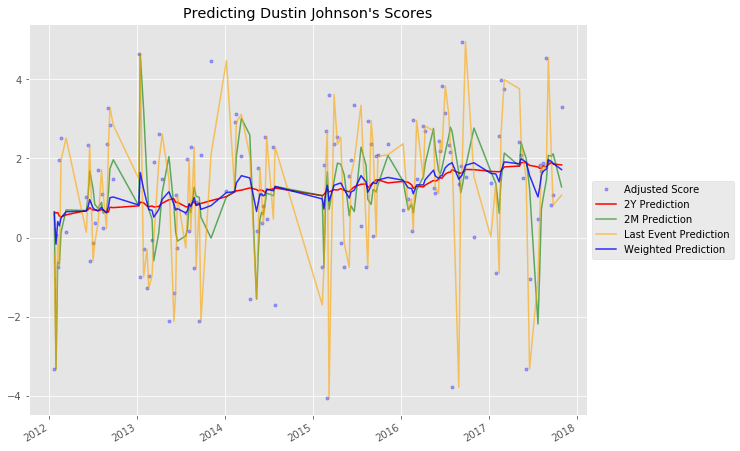
Figure 11: UI Results30

## *CHAPTER 1: INTRODUCTION*

Predictive modeling is the process of creating, testing and validating a computer science model[[1]](#footnote-1) made up of a number of predictors which are variables that are likely to influence future results to best predict the probability of an outcome. A number of modeling methods from machine learning, artificial intelligence, and statistics are available as predictive analysis tools for this task. The model is chosen on the basis of testing, validation and evaluation to best anticipate the probability of an outcome in a given set of input data. [1]

As an illustration of the potential of applying predictive models to the game of golf consider the following study on golf course history vs player performance:

In the graph below we plot a few different scoring averages calculated over different historical time horizons. The goal here is to evaluate different ways of predicting a player’s scores. [3]



Golf is a sport which allows an unlimited number of options and decisions for the player to make on every shot. Is a shot going to be aggressive or conservative? Is it better to fly your ball to the green or bounce it on? Should you hit the back of the cup with your putt or lag it up close? Understanding which shot decisions give the best chance for improved play and lower scores gives the golfer a competitive advantage.

This thesis focuses on using predictive analytics to understand performance streaks in golf and the contributing factors. A predictive model, scoring engine, and a player based software system have been designed and developed based on ShotLink data to provide players, coaches and trainers insight based on data sets reflecting their play into what characteristics are pertinent surrounding streaks enabling them to play their best more often.

This study is different than typical research into skilled performance streaks that have applied euphemisms such as ‘success breeds success’ [11] or ‘hot hands’ [10] to the phenomenon of sustained above average performance.

Our results have been generated through a rigorous program of data mining and analytics [1]; a well-defined scoring mechanism and custom feedback based on individual data submissions.

***1.1 PROBLEM STATEMENT***

Players want to play their best all the time yet golf by its nature yields inconsistent results. It is common to shoot over, under, and even par on different holes or to shoot different scores on the same hole played multiple days in a row, yet seemingly at random players get on ‘streaks’ – periods in which the quality of their play consistently exceeds the norm. The ability to understand the contributing factors behind streaks can provide a competitive advantage in the game of golf.

In this thesis we propose to generate prototypes of a data mining model, a scoring engine[2] and a player based system which will provide golfers opportunities to understand how their play aligns with streaks and what they may modify to improve.

For this thesis we have used a dataset known as ShotLink®: a proprietary system owned by the PGA Tour that captures multiple data points for every golf course, event, round, and shot struck during golf competition since 2006. Using ShotLink data we performed experimentation with some of the state-of-the-art data mining algorithms [1] to identify patterns, classify information and produce our base data mining model.

Our solution will also include a prototype of a scoring engine [2]. The scoring engine will use our model to make predictions on new data by running the new data through the model and assigning values to specified data points. The quantification provided by the scoring engine is critical in understanding and revealing the impact of data on the model.

The scoring engine and data mining model will be leveraged to create a prototype of a player based system which generates tangible customized results. The player based system will take as input a set of data points from the user through a questionnaire interface, run those data points through the base model, apply the scoring engine to the results and produce a report revealing the impact of given data points and recommendations on how to engage and sustain streaks. There is potential to expand the user based system to accept a complete data set with which a dedicated model can be produced for more granular results.

***1.2 MOTIVATION***

To the best of our knowledge there is no similar or recent research on how to get on or sustain streaks in the game of golf.

Just in the last twelve months over 25 million people have played golf in the United States alone [4], yet there’s no clear path to improvement and in fact most people do not get better in golf over time.

The average 18-hole score for the average golfer remains at about 100, as it has for decades, according to the National Golf Foundation, an industry research and consulting service. [5]

To improve in golf some advocate more practice such as Ben Hogan, one of the best ball strikers of all time. Mr. Hogan couldn’t get enough practice, claiming: *"There isn’t enough daylight in any one day to practice all the shots you need to."*



*Ben Hogan hitting a 1 iron, the most difficult club in the bag, at the US Open at Marion golf club*

While others don’t practice at all:

“What you have heard about me is true. I don’t practice…That’s when I decided that practice doesn’t do me any good, my swing doesn’t get any better when I go home and practice for a week.” - Bruce Lietzke 13 PGA and 7 Champions wins. From Links Letter 2005 annual edition



*Multiple PGA Tour tournament winner Bruce Lietzkie’s hobbies don’t include golf practice*

The research of this thesis is work in sports analytics to understand the true contributing factors of what causes good play and how to reproduce those conditions more consistently to ultimately play at a higher level more often and more consistently.

Sports analytics is bigger today than ever,

Data can be shared and used on an extremely granular level, enhancing the experience of professional sports for all parties involved. Instead of relying on intuition, experience and anecdotes, sports participants and enthusiasts can examine data that tells the real story to help with every aspect of the game – from player recruitment to fan engagement [6].

And the application of Big Data in Sports Analytics is really just in its infancy:

Big data is more of a long game in some contexts, setting the stage for the future rather than the present. In the ultra-competitive Olympics, where training is a full-time job, athletes are looking for any edge they can get along the way. Coaches are beginning to collect training data on young athletes, hoping to gain some insight into what makes top athletes succeed in the Olympics, and which factors play the biggest role. This data is also helping Olympic competitors learn more about the challenges they face during training, such as the clash between strength training and endurance training for rowers. [6].

The research here takes a scientific approach to understanding the true contributing factors of what causes good play and how to reproduce those conditions more consistently to ultimately play at a higher level more often and more consistently.

***1.3 CONTRIBUTIONS***

Major contributions ofthesis include:

1. A mining model based on ShotLink data purposed to quantify attributes associated with streaks
2. A scoring engine quantifying user provided data points resulting in tangible metrics applied to the mining model
3. A player based system that based on a well-defined set of inputs generates analysis and predictions for the impact of the data run through the mining model & scoring engine.

***1.4 THESIS OUTLINE***

The structure of this thesis is as follows:

1. *Introduction* provides an introduction into the game of golf and the motivation behind our research. It explains what streaks are, our data set, how predictive models can help, and include the contributions of our work.
2. *Background* provides a general overview into golf scoring and its intrinsic inconsistencies. It identifies streaks of various types and provide illustrations of streaks of different types.
3. *Methodology* describes the high level approach for mining, modeling, and scoring the data with focus on producing a method that generates predictive results
4. *System Design* provides detail about system requirements, architecture, and implementation of the software.
5. *Tests and Results* demonstrates a range of system experiments with proposed predictions based on various data points and examines the results
6. *Conclusions* provides a summary and review of our work. It also includes recommendations on further advancements of this research.

***CHAPTER 2: BACKGROUND***

In the field of Sports Analytics Major League Baseball Oakland Athletics General Manager Billy Bean is credited with bringing predictive analysis and modeling to the forefront by applying sabermetrics: *‘the search for objective knowledge about baseball’*

Sabermetrics is a science of sport. It is the empirical analysis of baseball through statistics, used to predict the performance of players, giving teams a winning edge. With the help of sabermetrics, teams can: Forecast results by making predictions based on previous data; analyze on-field performance by recording and evaluating important aspects of play; assist in decision-making by offering objective insights into players’ performance, matchups, and scouting prospects. Who’s a better pitcher? Where should our outfielders play? Which player is a better value for our team? Who was the greatest second baseman of all time? Thanks to sabermetrics, all of these questions can now have objective answers [9]

The game of golf is a sport played over 18 individual ‘holes’ wherein each hole provides a predetermined number of strokes that should be required called ‘par’ which players measure their quality of play against. Typically players will score better on some holes than others but on average will end up with nearly the same total score for each set of 18 holes (round) they play.



*Arnold Palmer, known as ‘The King’, one of the most successful and popular golfers ever*

Elegantly summarized by Winston Churchill:

Golf is a game [whose](https://www.definitions.net/definition/whose) aim is to hit a very [small](https://www.definitions.net/definition/small) ball into an even [smaller](https://www.definitions.net/definition/smaller) hole, with [weapons](https://www.definitions.net/definition/weapons) singularly ill- [designed](https://www.definitions.net/definition/designed) for the purpose. [8]



*Jean Van De Velde in the Barry Burn, 1999 Open Championship, Carnoustie.*

During those atypical periods in which everything magically works a player is on a ‘streak’. Streaks come in different shapes, sizes and forms – consider the following examples [9]

1. Putting:

5 holes in a row without a 3 putt:

Driving:

36 consecutive fairways in a row off the tee

Short Game:

29 consecutive greens in regulation

17 Up & down in a row out of the sand

Scoring:

16 rounds below par in a row

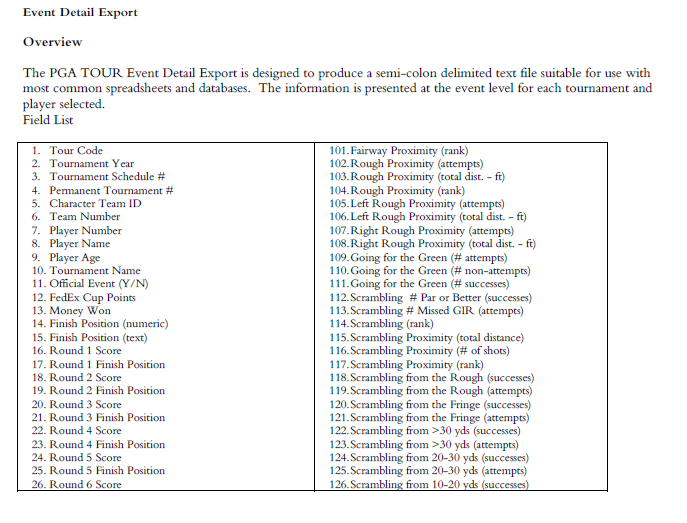
***CHAPTER 3: METHODOLOGY***

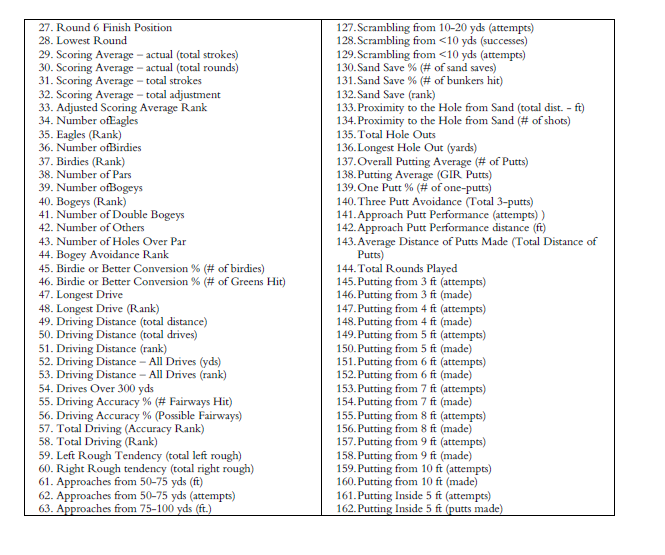
*A*pproach for mining and analyzing the data with special interest in the granularity of the data and how we are applying it to streaks.

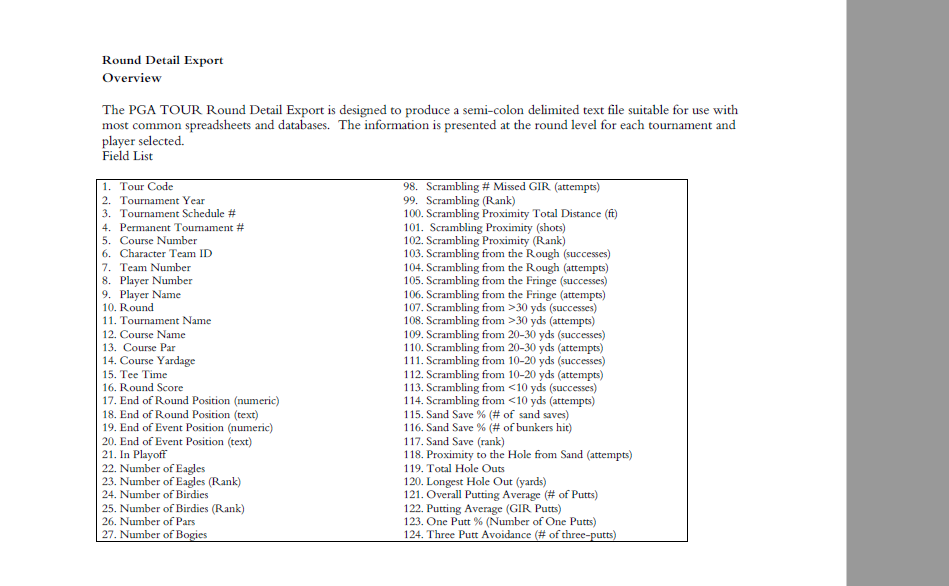
***3.1 DATASET***

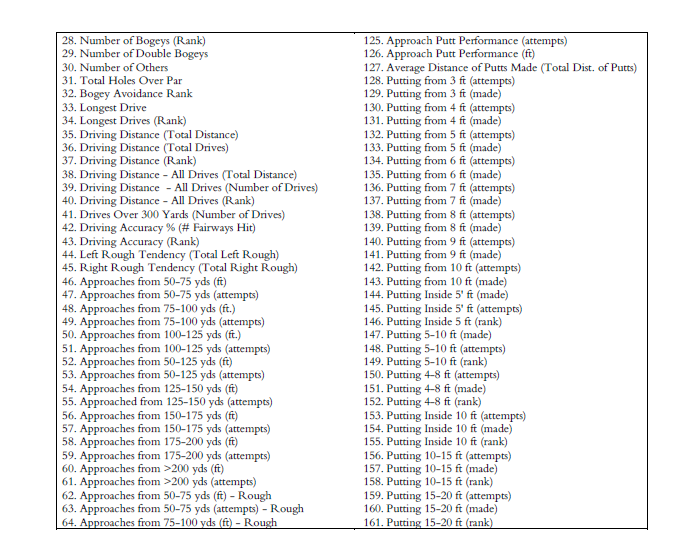
The dataset for this thesis is ShotLink®: a proprietary system owned by the PGA Tour that captures multiple data points for every golf course, event, round, and shot struck during golf competition since 2006. ShotLink provides 5 distinct datasets described at the attribute level as follows:

***3.1.1 EVENTS:***

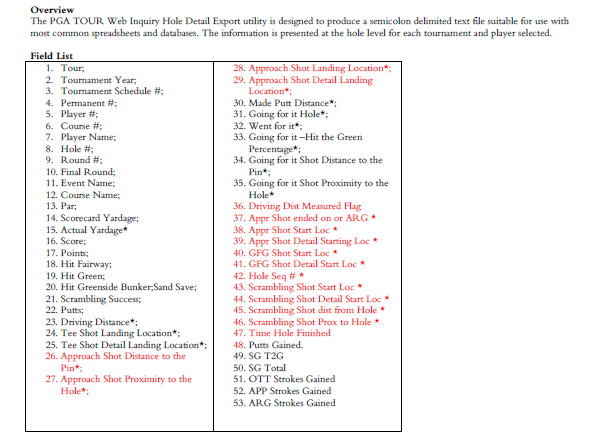




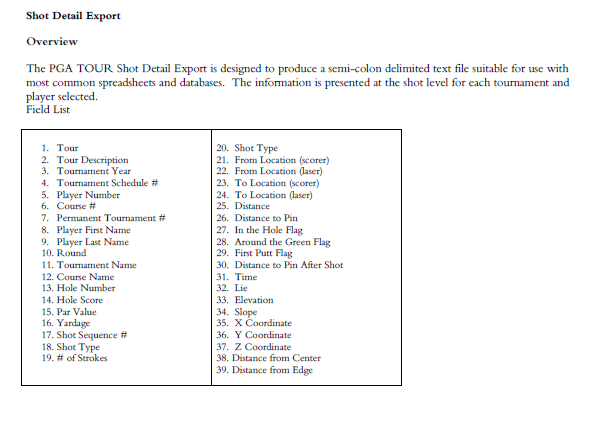
***3.1.2 ROUNDS***



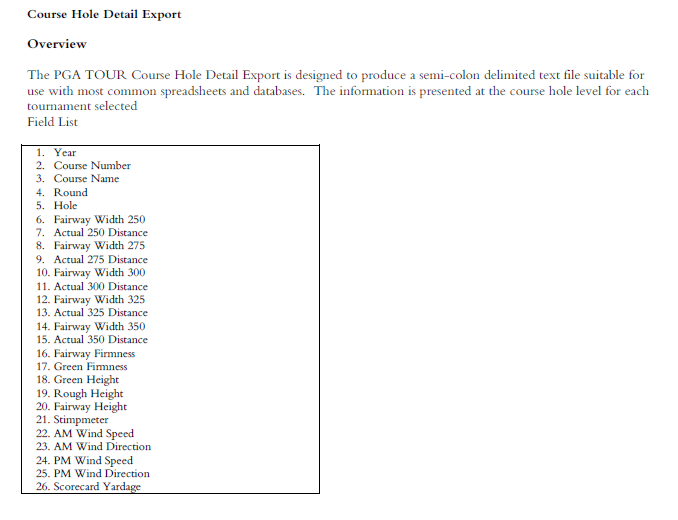
***3.1.3 HOLES***

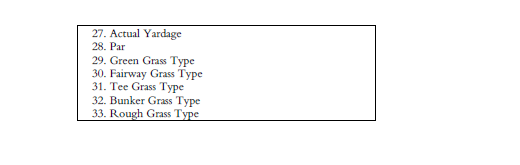


***3.1.4 SHOTS***



***3.1.5 EVENTS***



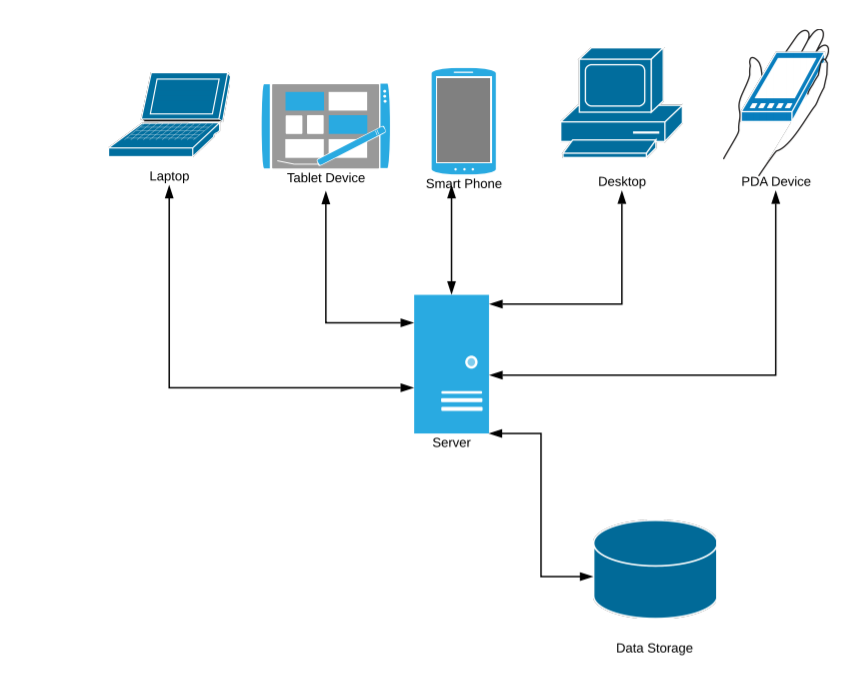


***3.2 DATA MINING AND ANALYZING***

1. Data Mining
   1. Build database
   2. Generate data mining model
   3. Explore mining algorithms [1] to identify patterns, classify information and produce our base data mining model
2. Build scoring engine
   1. Use mining model to make predictions on new data by running the new data through the model and assigning values to specified data points.
3. Design and develop player based system
   1. Take as input a set of data points from the user through a questionnaire interface
   2. Run those data points through the base model,
   3. Apply the scoring engine to the results and produce a report revealing:
      * 1. The impact of given data points
        2. Recommendations on how to engage and sustain streaks.
   4. Potential to expand the user based system to accept a complete data set with which a dedicated model can be produced for more granular results.

***CHAPTER 4: SYSTEM DESIGN***

* 1. ***HIGH LEVEL OVERVIEW***
     1. Client server based system



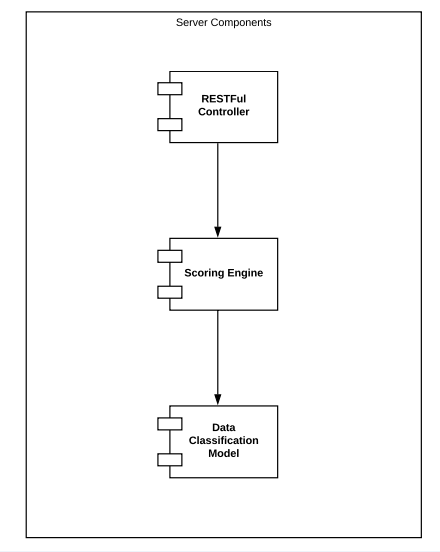
* 1. ***SERVER COMPONENTS***

Major server side components broken into:

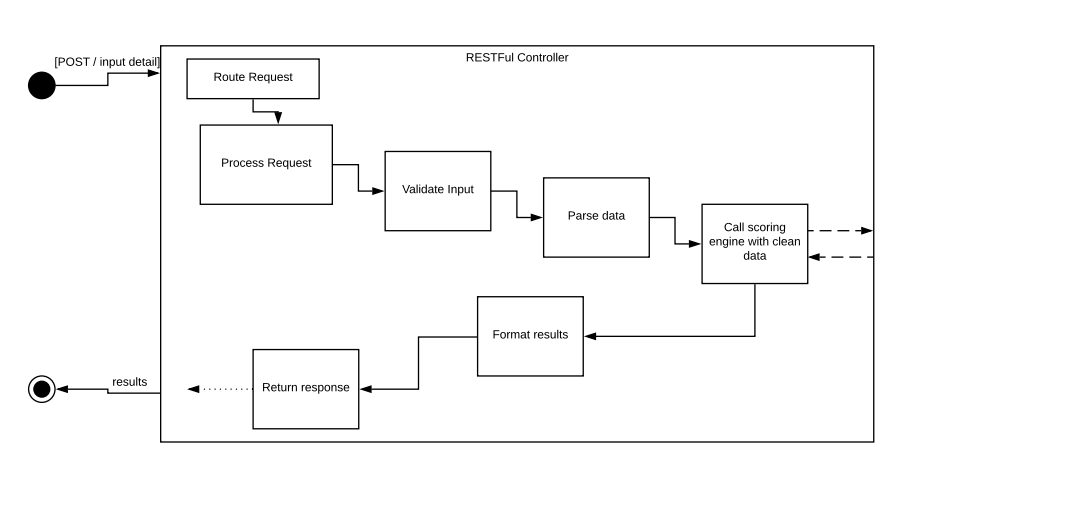
4.2.1 Restful Controller

4.2.2 Scoring Engine

4.2.3 Data Classification Model



***4.2.1 RESTFUL CONTROLLER***

******

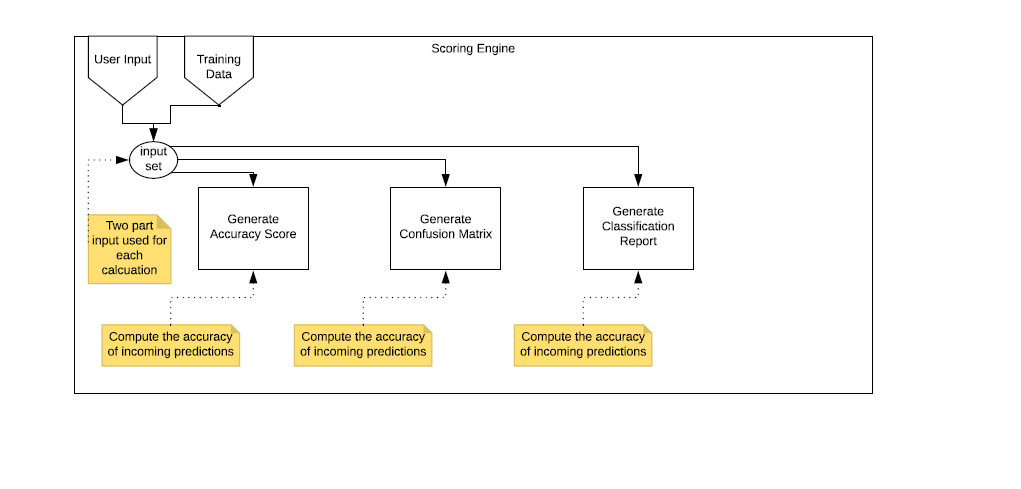
The restful controller is the entry point to the server and acts as a traffic controller. Based on the incoming request type program flow is routed such that the incoming data may be processed as intended.

Execution occurs through server side components, results are compiled and returned to the calling requestor.

***4.2.2 SCORING ENGINE***

The scoring engine requires two data sets:

* 1. Training set
     + 1. Set of examples used to fit the weights of a classifier
  2. Test set
     + 1. Data built up of user input used to compare against expected values

******

(Typical run)

-----------------------

KNearestNeighbors results:

-----------------------

Accuracy: 0.776354679803

Confusion Matrix: [[362 106]

[121 426]]

Classification Report: precision recall f1-score support

0.0 0.75 0.77 0.76 468

1.0 0.80 0.78 0.79 547

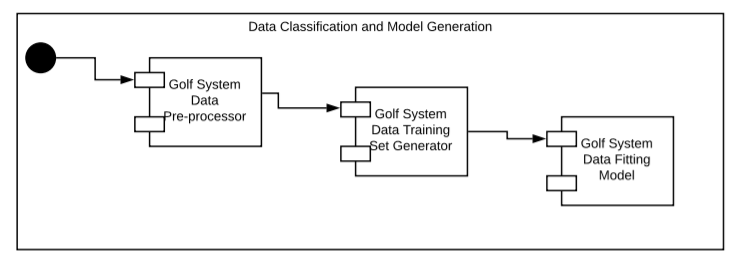
avg / total 0.78 0.78 0.78 1015

The scoring engine generates metrics identifying accuracy and performance of predictions

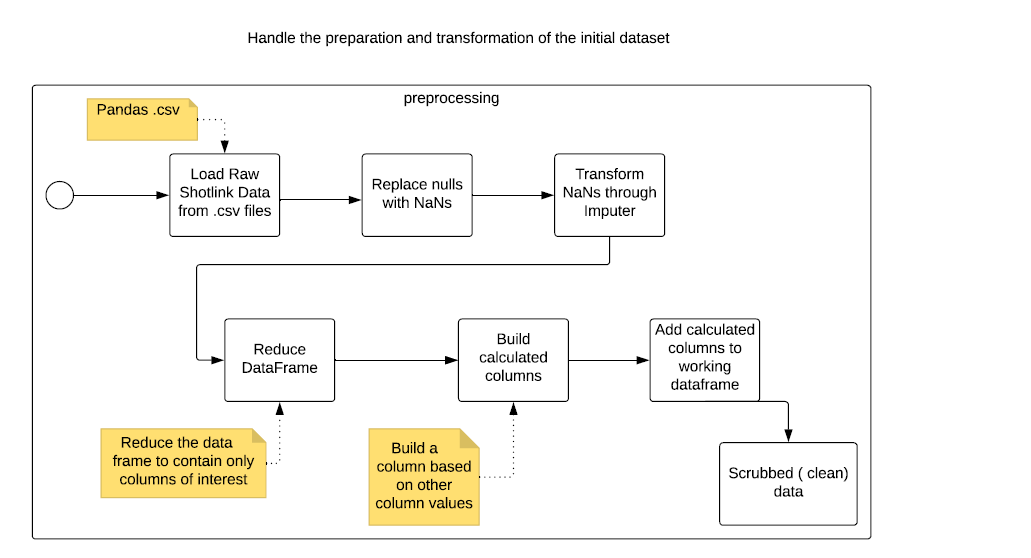
* 1. Accuracy Score:
     + 1. Given a set of data points from repeated measurements of the same quantity, the set can be said to be accurate if their average is close to the true value of the quantity being measured. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y\_true.
  2. Confusion Matrix:
     + 1. A table providing a visualization of the performance of the chosen algorithm such that each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. A confusion matrix C is such that C_{i, j} is equal to the number of observations known to be in group i but predicted to be in group j.
  3. Classification Report:
     + 1. A text report showing the main classification metrics. A table providing a visualization of the performance off1-score gives you the harmonic mean of precision and recall. The scores corresponding to every class will tell you the accuracy of the classifier in classifying the data points in that particular class compared to all other classes; Support is the number of samples of the true response to that class.

**4.2.3 DATA CLASSIFICATION MODEL GENERATION**

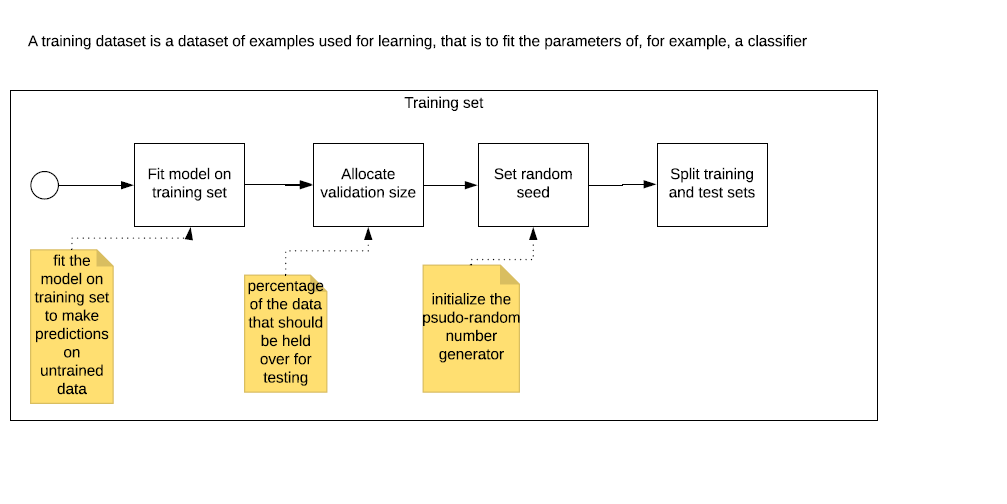
1. Building the Data classification model is a multistep process, requiring:
   1. Preprocessing
   2. Training set generation
   3. A ‘Fitting’ of the Model

******

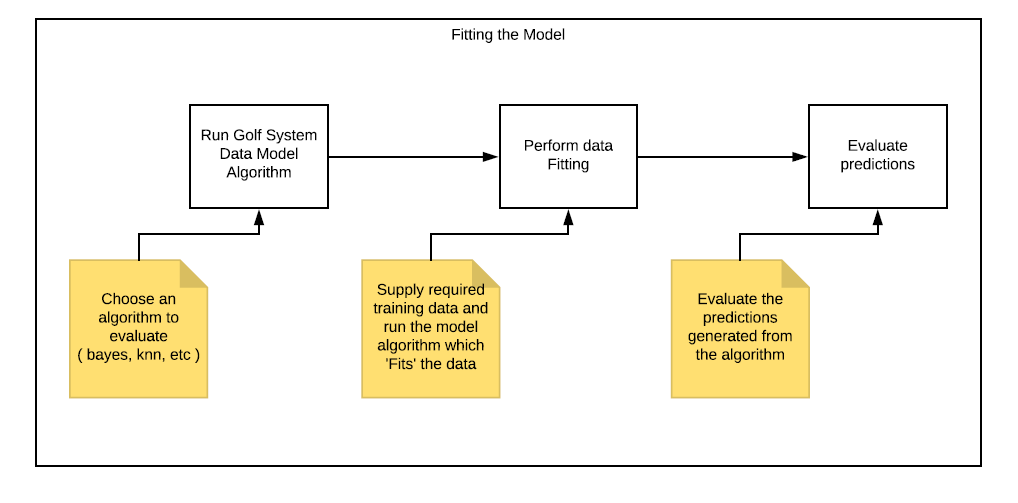
***4.2.4 PRE-PROCESSING:***

******

***4.2.5 TRAINING SET:***



***4.2.6 MODEL FITTING:***

******

***4.3 TECHNOLOGY:***

Data analysis in this thesis uses the Python programming language exclusively with special interest in the Pandas, NumPy, Sklearn and Matplotlib libraries.

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. [Pandas](http://pandas.pydata.org/) is one of those packages, and makes importing and analyzing data much easier. Pandas builds on packages like [NumPy](http://www.numpy.org/) and [Matplotlib](http://matplotlib.org/) to give you a single, convenient, place to do most of your data analysis and visualization work

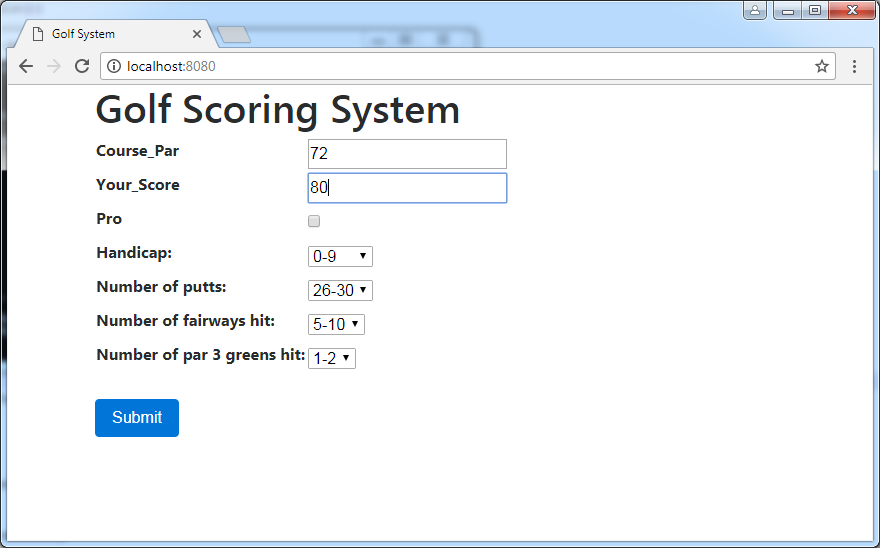
* + - * 1. **Pandas** is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](https://www.python.org/) programming language
        2. ***NumPy***, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.
        3. ***SKLearn*** features various classification, regression and clustering algorithms which interoperate with numpy
        4. ***Matplotlib*** is a plotting library for Python. It is used along with NumPy to provide an environment that is an effective open

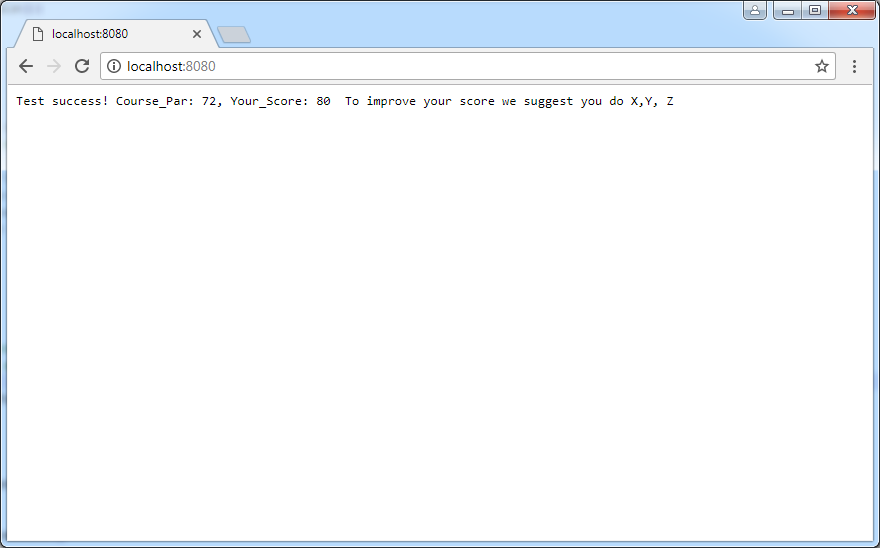
source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython

***4.4 USER INTERFACE***

The User Interface is as follows:

1. Browser based thin client accessing server via http
2. Responsive web page(s) supporting access through all typical devices
3. Series of screens dedicated to collecting groups of user input
4. Response/result page displaying report & recommendations
5. Optimized to manage profile & other individual details locally





***CHAPTER 5. TESTS AND RESULTS***

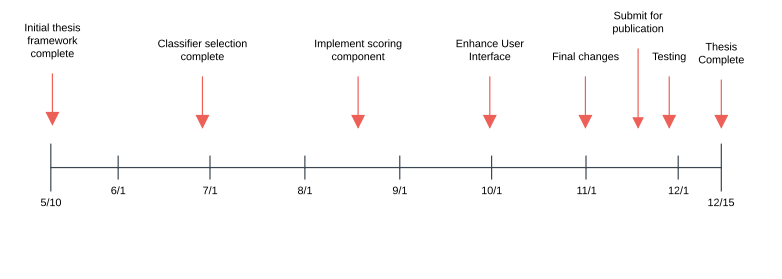
1. Attribute fitting tests
   1. Attributes used
2. Classifier and model fitting tests
   1. Classifiers used
3. Scoring Engine tests
   1. Results of different weights

***CHAPTER 6: CONCLUSIONS***

**6.1 CONCLUSIONS**

…work in progress

***6.2 FUTURE WORK / TIMELINE***



***6.3 POTENTIAL PUBLISHERS***

1. <https://www.ieee.org/publications/authors/publishing-benefits/index.html>
2. <https://journalsuggester.springer.com/select-journal?journalId=13688>
3. http://journals.theired.org/en-ijserm.html

*REFERENCES:*

1. “What is Predictive Modeling” Retrieved from: “<http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf>”
2. “Does a players course history predict performance?” December 27 2017. Retrieved from “http://www.datagolf.ca”
3. “What is Predictive Modeling” Retrieved from: “<http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf>”
4. “Number of people who played golf within the last 12 months in the United States from spring 2008 to spring 2017 (in millions)” Retrieved from: “https://www.statista.com/statistics/227420/number-of-golfers-usa”
5. Bill Pennington “Golf Club Prices Are Up; Scores Are Not Down” May 2005. Retrieved from: “https://www.nytimes.com/2005/05/24/sports/golf/golf-club-prices-are-up-scores-are-not-down.html”
6. Ryan Ayers, “How Big Data is Revolutionizing Sports” January 2018. Retrieved from: “http://dataconomy.com/2018/01/big-data-revolutionizing-favorite-sports-teams”
7. “Sabermetrics: Baseball Analytics and the Science of Winning [Infographic]” October 2015. Retrieved from: “https://onlinebusiness.syr.edu/blog/sabermetrics-baseball-analytics-the-science-of-winning”
8. [GEORGE PEPER](http://www.golf.com/author/george-peper-1)  https://www.goodreads.com/quotes/252141-golf-is-a-game-whose-aim-is-to-hit-a Wednesday, August 13, 2014
9. https://www.pgatour.com/stats/
10. J[oshua Miller](https://theconversation.com/profiles/joshua-miller-347516), [Adam Sanjurjo](https://theconversation.com/profiles/adam-sanjurjo-347532) “Momentum isn’t magic – vindicating the hot hand with the mathematics of streaks” Retrieved from <https://theconversation.com/momentum-isnt-magic-vindicating-the-hot-hand-with-the-mathematics-of-streaks-74786> March 26, 2017 10.38pm EDT
11. David Gliden, Stephanie Gray Wilson. Streaks in skilled performance [https://labs.la.utexas.edu/gilden/files/2016/04/StreaksPsychonomicBR1995.pdf 1995](https://labs.la.utexas.edu/gilden/files/2016/04/StreaksPsychonomicBR1995.pdf%201995).
12. “Top 10 algorithms in data mining” July 2007. <http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf>
13. “Scoring your customers” http://www.thearling.com/text/scoring/scoring.htm

1. *A formal construction which is described theoretically in the language of computer science: the language of algorithms and data structures* [↑](#footnote-ref-1)