National College of Ireland

BSc in Computing

2017/2018

Daniel Murran

X13114336

X13114336@student.ncirl.ie

Supervisor: Simon Caton

Ahead of the Curve,

Analysis of the NFL Draft

Technical Report

Contents

1 Table of Figures

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: Daniel Murran
Student ID: x13114336
Supervisor: Simon Caton

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: Daniel Murran Date:13/5/18

Executive Summary

The National Football League (NFL) draft is the primary gateway to the professional ranks of American football. Any college football player worth their salt dreams of being picked number one in the first round of the NFL draft.

This project attempts to predict the likelihood of whether a player entering the draft will be selected in the first round. Previous statistics from the National College Athletic Association (NCAA), NFL combine results and previous draft results will be analysed following the Knowledge Discovery and Data Mining methodology. Data will be targeted, selected, transformed and modelled to reveal knowledgeable patterns which will aid in the prediction.

Who, apart from avid fans, cares about this? Well, the NFL is a multi-billion-dollar industry made up of 32 teams and the individuals who choose team membership are the General Managers (GMs) of these teams. These GMs are faced with the decision at the start of each season of who to pick for their team; should they retain the players they have or choose the hot new prospect?

Techniques such as data visualization and principal component analysis were used to explore the datasets. Regression was used to develop models for the prediction, packages such as Glmnet and XGBoost provided smooth model development. The data was then visualized in a user-friendly dashboard using Tableau.

2 Introduction

American Football is America's favourite past-time; last year's Super Bowl on CBS drew 111.9 million viewers, making it the third-most-watched broadcast in U.S history.[1] The draft allows the thirty-two teams in the NFL to gather and pick the most aspiring talent from the amateur scene. The draft is comprised of 7 rounds and 256 picks which allow professional teams to reinforce their rosters with the best new talent available.

To adequately assess draft prospects, GMs of professional teams analyse the college careers of each player who enters the draft and their performance at the NFL combine. The NFL combine is often called the most significant interview on the planet; it is a showcase of college football talent that takes place over a week prior to the draft. During the combine, amateur players are assessed in tests of speed, strength and intelligence.

For many teams, the draft is the most prominent day of the year on the NFL calendar; teams who may have struggled during the previous season potentially have the chance to pick the next superstar quarterback. Under the rules of the draft, the team which finishes last in the previous season gets the first overall pick in the draft. With this comes significant pressure on teams to select the right player who will not only help them win games but will fit into their team's play style. As such, there is immense significance for teams to make the right choices in the draft.

---

[1] http://fortune.com/2017/02/06/super-bowl-111-million-viewers/ [accessed 1st January 2018]

2.1 *Aims*

The overall aim of this study is to find out what is the likelihood that a player entering the NFL draft will be picked during the first round. Figure 1 shows the path an amateur football player follows to progress to a professional career.

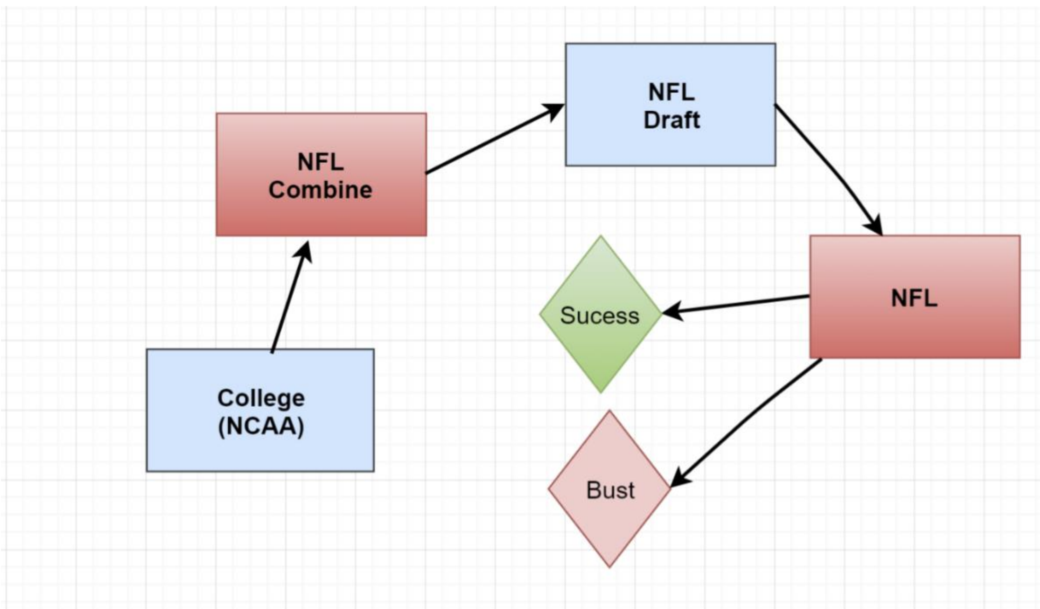

Figure 1: Path of a college football player

This goal will be achieved by analysing data from the years 2000 to 2018 of the NFL Draft, NFL Combine and NCAA College Football. Data was collected on individual players, their performance over their college years, their performance during the combine and their subsequent placement in the NFL draft.

A machine learning model will then be built to predict the player placement.

The collected data will be integrated into a visualization dashboard on Tableau for a user-friendly customer appeal.

2.2 *Technologies*

R Studio: R Studio is an open-source development environment for the programming language R. R made up the backbone of the technologies for this project. R packages were used throughout the lifecycle of the project to scrap, interact, visualize and model the data.

MySQL: MySQL is a relational database management system. A MySQL database was used to store the data for this project securely. The R package RMySQL was used to interact with the database through the R Studio integrated development environment.

Tableau: Tableau is a visualization dashboard software which was used to visualize the data for customer testing as a visualization dashboard.

SPSS: SPSS is an IBM software package that is used for statistical analysis. SPSS was used to test both regression models.

3 Background & Literature Review

"*The goal is to turn data into information, and information into insight*." – Carly Fiorina, former executive, president, and chair of Hewlett-Packard Co.[2]

The value that data holds is more important than ever; the NFL is no exception, being a multi-billion-dollar industry, teams are seeing the benefit of investing in data now more than ever.

In 2012, the Baltimore Ravens hired the first official data analysts in the NFL; one analyst focused on scouting new talent and the second analyst evaluated current players' performance. [3]

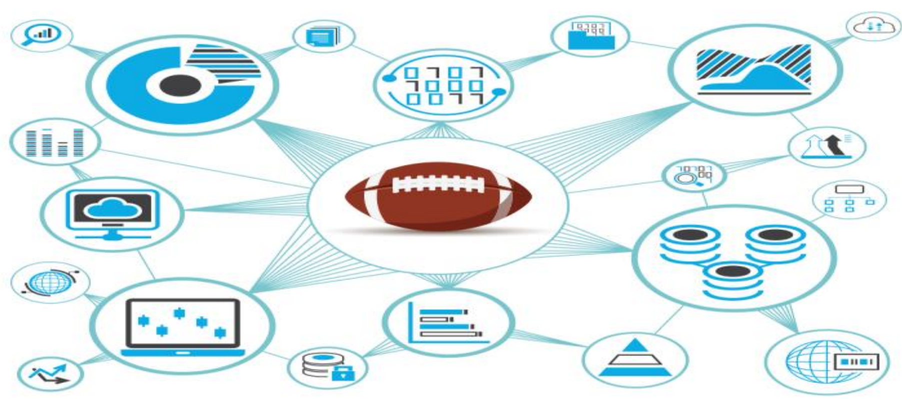

Figure 2: Data Analysis in NFL[4]

In 2016, the Cleveland Browns hired data analyst Paul DePodesta as their chief strategy officer. The Browns had failed to turn their early draft picks throughout

---

[2] http://www.hp.com/hpinfo/execteam/speeches/fiorina/04openworld.html [accessed 16th January 2018]

[3] https://www.ibmjournal.com/guides-and-journals/how-data-analytics-is-transforming-the-nfl [accessed 13th March 2018]

[4] https://www.ibmjournal.com/guides-and-journals/how-data-analytics-is-transforming-the-nfl [accessed 11th March 2018]

the years into successful professionals. Paul was hired to aid in the selection of players during the draft. Actor Jonah Hill portrays Paul DePodesta in the Oscar-nominated film Moneyball. This film details the true story of how Paul and his colleagues in the Oakland Athletics baseball team used sophisticated statistical approaches to build a competitive baseball team. [5]

Marc Fridson developed a tool that analyzed thirty years of the NFL draft and player outcomes. Using R studio with the library Shiny, Marc developed a tool that visualized information on NFL teams' draft selections throughout the years[6].

Bill Lotter wrote a blog post on the Harvard sports analysis collection titled "the NFL combine actually matters". The post reviewed combine data in correlation with career performance by analyzing all positions of an NFL team. Using a linear regression model, the study found significant success in predicting NFL outcomes for all positions.[7]

Brad Jones attempted to predict quarterback draft selection using random forests and elastic net regression. The study found the elastic net regression to be more accurate with 60% accuracy rating. The study used previous quarterback data from the combine to build its prediction model.[8]

---

[5] https://www.si.com/mmqb/2016/01/06/nfl-cleveland-browns-paul-depodesta-jimmy-haslam [accessed 11th May 2018]

[6] https://nycdatascience.com/blog/student-works/nfl-draft-30-years-outcome-analysis/ [accessed 12th May 2018]

[7] http://harvardsportsanalysis.org/2015/02/the-combine-actually-matters-part-2/[accessed 12th May 2018]

[8] https://www.dailynorseman.com/2014/7/19/5918071/from-boom-to-bust-building-a-predictive-quarterback-model [accessed 12th May 2018]

Adam McCann wrote a paper on using machine learning algorithms to predict success in the draft. The study focused on analyzing the Quarterback position. The study used decision trees machine learning methods, paired with data on quarterbacks from 2000 – 2016, to predict the success of new quarterbacks entering the draft. Utilizing a confusion matrix to analyze the model, the study found a predictive value of 76% precision. ("Reducing NFL Draft Risk_FINAL.pdf," n.d.)

Wesley Olmsted wrote a paper that compared three different machine learning approaches to predict NFL draft picks and career success. Regression, Neural networks and K-Means were used on quarterback data from 1990 to 2013. The study could predict the draft round for QBs (quarterbacks) with a 35.3% success rate, with an average error of 1.47. The comparison of machine learning approaches used showed neural networks and k-means predicted more effectively than regression did. ("final.pdf," n.d.)

Gary McKenzie wrote a paper on improving the accuracy of the draft. The study used the MGA-SS (Multilayer Genetic Algorithm with Singular Selection) algorithm and Random Forests to predict success in the NFL draft. The study found that the MGA-SS algorithm outperformed the random forest approach by 89%. ("research.pdf," n.d.)

Julian Wolfson focused a study on college quarterback draft selection for the NFL since 1997, based on statistics such as games played and points scored during their college careers. The study used logistic regression to build a model for the prediction. The conclusions reached from the author of the study where that college and combine statistics can be reliable for predictive of success in the NFL. (Wolfson, 2018)

Amrit Dhar wrote a paper on the success of drafting NFL wide receivers. Amrit used data on previous NFL Combines, NFL wide receiver's statistics and NCAA wide receiver statistics. The study built a recursive partitioning regression tree model. The study concluded that college performance and NFL combine data provided better results than NFL wide receiver data. The model received an R^2 value of 0.35 which the author believed the result implied that there was a significant amount of variance between draft rankings and NFL performance. (Dhar, 2018)

All papers reviewed for this study came to similar conclusions. The relationship between machine learning and the NFL draft is at an early stage, but there is still value to be gained from this approach.

4 Methodology

The Knowledge Discovery and Data Mining (KDD) methodology approach was followed for this study. KDD is a series of processes used by researchers in databases, machine learning, artificial intelligence and many other domains. As seen in figure 3 below, the KDD process allows a step-by-step method of, at a high level, turning data into knowledge. ("KDD.pdf," n.d.)

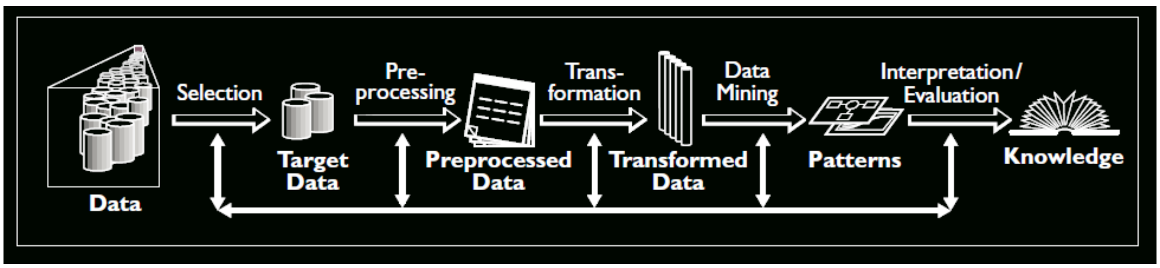

Figure 3: KDD Methodology[9]

Prior to starting the KDD process, it is crucial to develop an understanding of the application domain that is being studied and the goals of the study. Throughout the KDD process loops and repetition of processes occur; processes are correlated to the user's requirement.

4.1 *Data Selection*

The Data selection step of the KDD process targets the required datasets which are needed to fulfil the requirements of the study. To accomplish the desired outcome of this study, data was required on the NFL draft, the combine and the NCAA. [10]

[9] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html [accessed 15th December 2018]
[10] https://www.sports-reference.com/ [accessed 11th May 2018]

The data selected for this study was scraped from the Pro Football Reference site which maintains historical data for the NFL. Data was scraped for the NFL draft, and the NFL combine for the years 2000-2018. Data for the NCAA college football was scraped from Sports-Reference, a domain holding statistics for college football. [11]

4.2 *Pre-Processing*

The pre-processing step of the KDD process involves cleaning the data. This step can include techniques such as removal of outliers, dealing with missing data and accounting for noise in the data.

For this study, the R package mice were used to impute missing values where applicable. The datasets for college statistics, combine and draft data were all manipulated and cleaned. This process allowed all data to be merged into one dataset for the study which in turn allowed the data to be set up for machine learning algorithms in test and training set formats.

4.3 *Transformation*

The transformation step involves data projection and reduction; this can include representing the data statistically or visually depending on the goal of the study.

Principal Component Analysis (PCA) was used on the combine data to establish whether there was a possible correlation between variables. Data visualization techniques were used to perform exploratory analysis of the datasets.

---

[11] https://www.pro-football-reference.com/ [accessed 11th May 2018]

4.4 *Data Mining*

The data mining process involves the selection of appropriate machine learning algorithms which will be performed on the transformed dataset. Furthermore, this step includes choosing parameters from that dataset which will be included in the machine learning model and the searching and identification of patterns in the data. Two types of regression were used to fulfil the requirements of the study. The package Glmnet allowed the implementation of regularized regression and the XGBoost package provided the tools to implement a gradient boosting model.

4.5 *Interpretation / Evaluation*

The interpretation and evaluation step of the KDD process can involve returning to previous steps to reiterate processes, the removal of redundancies, and the evaluation of the overall knowledge gained from the study.

This step of the study involved visualizing the relevant models to view the receiver operating characteristics (ROC) curve and compute the area under the curve (AUC). The models were then tested to make an actual prediction for the first round of the NFL draft for 2016, 2017 and 2018.

5 System

5.1 *Design and Architecture*

The architecture of the project is shown in figure 4. The functionality of the study was carried out in R Studio which is an integrated development environment (IDE) for the R engine. The data is stored in a MySQL database which will be integrated through the RMySQL Library in Studio. The project incorporates many libraries to fulfil the needs and the requirements of the study, some of the more central libraries are viewable in figure 4. The libraries XGboost and Glmnet are used for the regression aspect of the project. Dplyr was used throughout many phases of the project to manipulate and interact with the data. Tidyverse has many libraries in its package and proved valuable especially during the data exploration phase of the project. FactoMiner allowed principal component analysis to be carried out on the data and the test of that package permitted the scripts of the project to be unit tested.

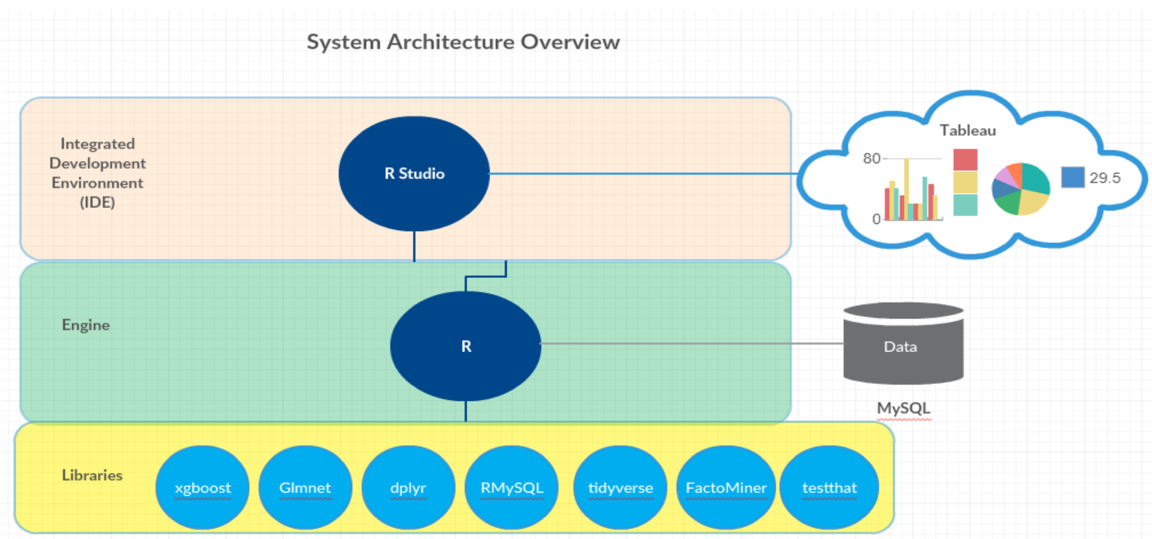

Figure 4 System Architecture Overview

5.2 *Implementation*

5.2.1 Data Selection

The first phase of the project was to target and select the data needed to fulfil the study. The domains *sports-reference* and *pro-football-reference* provide sabermetric and basic statistics on all American sports. The project required data on the draft, the combine and NCAA college statistics. Data was collected on 9,922 players who played college football from 2000-2018. The library rvest was utilized to gather the required data in conjunction with RCurl which allows the function to fetch URIs. The libraries tidyr, dplyr and stringr aided in manipulating, tidying and removing whitespace from the data. In figure 5 we can see the function to extract the Urls, and figure 6 shows a snippet of code from the combine data scraping function. The feather library allowed the use of feather files. Feather files offer a high-speed file format for storing data frames. [12][13]

```
extract.urls <- function(tds) { results <- c()
for(td in tds) { children <- html_children(td)
if (length(children) == 0) { results <- c(results, NA)
} else{results <- c(results, (html_attr(html_children(td), 'href')))  } }results}
```

Figure 5: Urls extracted

[12] https://www.r-bloggers.com/rvest-easy-web-scraping-with-r/ [accessed 11th May 2018]
[13] https://www.r-bloggers.com/webscraping-using-readlines-and-rcurl/ [accessed 9th April 2018]

```
combine.table <- data_frame(year = 2000:2018) %>%
  group_by(year) %>% do({url <- paste('http://www.pro-football-reference.com/draft/', .$year, '-combine.htm', sep ='') html.table <- read_html(url) %>%
```

Figure 6: Combine scraped data and merge into a table

5.2.2 Data Preparation

The data preparation step used libraries tidyr, dplyr and stringr to interact, clean and manipulate the data. The datasets for combine, draft and college data were joined to make one dataset for machine learning and exploratory purposes. Some rows shared the same player multiple times; this was filtered out using the mutate and filter functions, see figure 7.[14]

```
group_by(key) %>%
mutate(appearance = row_number()) %>%
filter(appearance == 1) %>%
select(-appearance) %>%
ungroup
```

Figure 7: Filtering multiple rows with the same player

[14] https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/ [accessed 11th May 2018]

The library mice was used to impute missing scores for the combine data and any other missing data throughout the data preparation process. In figure 8 we can see the library mice used to impute data.

```
training1b <- complete(mice(training1a %>% select(-key, -carav)))
training1b$key <- training1a$key
training1b$carav <- training1a$carav  ungroup
```

Figure 8: Mice Imputation of data

Throughout the data preparation process, the data frames were converted back and forth from long format to wide format to facilitate the inclusion of all datasets.

The data was then prepared for machine learning purposes as seen in figure 9. The training data was set up with 75% of the data where the test had the other 25%. Outcome variables were then declared for pick and first.round to aid in the machine learning process. The data was then exported as a .csv file for convenience purposes.

```
N <- nrow(training)
train.set <- (rbinom(N, 1, prob = 0.75) == 1 & training$year < 2018)
test.set <- (!train.set & training$year < 2018)
holdout.set <- !(test.set | train.set)
```

Figure 9: Training & Test Set Preparation

5.2.3 Data Transformation

The transformation step of the project was made up of exploratory analysis and principal component analysis (PCA).

5.2.3.1 Exploratory Analysis

Exploratory data analysis aims to: -

- Produce questions regarding your data;
- Examine visuals and transformations for feedback from your data;
- Provide insights to aid in the understanding of the data.[15]

The explore.r data file examines and explores many of the variables from the dataset. The combine data was the focused sample for the exploration shown in this section.

Figure 10 and 11 shows a correlation matrix, and a correlation plot for the variables pick, bench, forty, three-cone and vertical. The pick variable is where a player is selected in the draft, and the other variables in this matrix are all measures of tests in the combine. The plot and matrix show low negative correlation from combine variables and pick. It makes sense that if a player is only strong at one aspect of the combine that this would not correlate to a high pick in the draft. The correlation plot reveals a high positive correlation between the forty and the threecone, both tests are tests of speed, so this correlation statistic makes sense. On the other hand, both the threecone and the forty (speed tests) show a high negative correlation with the vertical variable, which is a jumping test. This means that if someone is very fast, they might not be able to jump very high. The library corrplot was used to visualize the correlation matrix as a correlation plot.

[15] http://r4ds.had.co.nz/exploratory-data-analysis.html [accessed 11th May 2018]

	pick	bench	forty	threecone	vertical
pick	1.00000000	-0.01802133	-0.03422789	-0.03398015	0.01612007
bench	-0.01802133	1.00000000	0.41838871	0.42801488	-0.23351545
forty	-0.03422789	0.41838871	1.00000000	0.79593644	-0.71638781
threecone	-0.03398015	0.42801488	0.79593644	1.00000000	-0.61377958
vertical	0.01612007	-0.23351545	-0.71638781	-0.61377958	1.00000000

Figure 10: Correlation Matrix

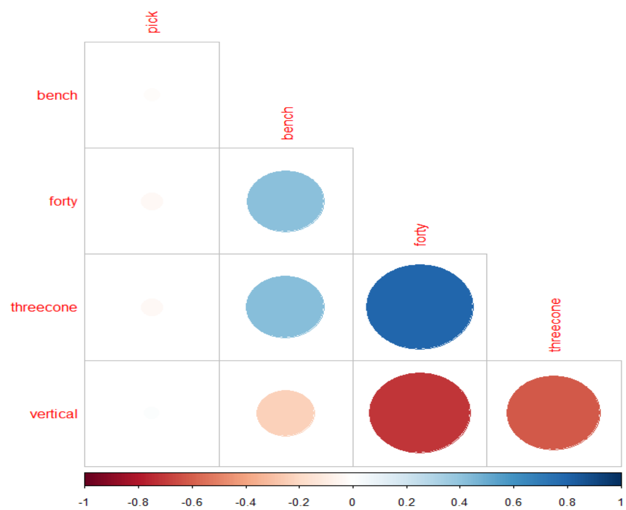

Figure 11: Correlation Plot

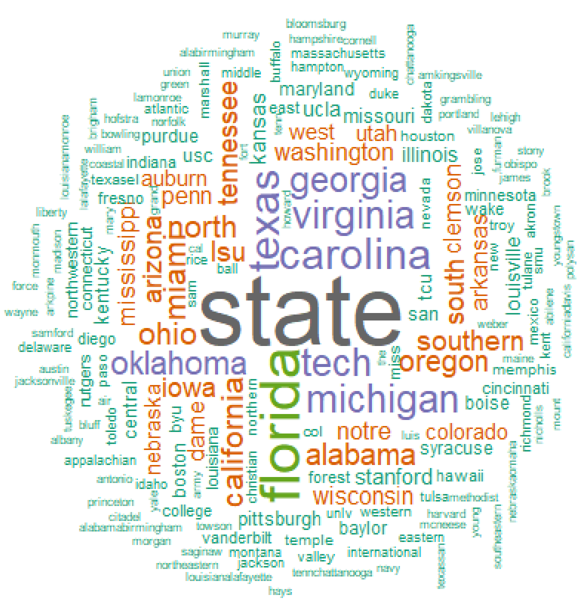

Figure 12: Word Cloud

In figure 12 a word cloud is used to visualize the most common colleges in the data set. This analysis provides an insight into which colleges produce the best athletes. We can see from the word cloud that some well-known NCAA college football team's standout at the top such as Oklahoma, Michigan and Florida. The word clouds were developed using the word cloud library.

A histogram was used to visualize the shuttle variable which is a measurement from the NFL combine. We can see the distribution of all measurements for the shuttle from 2000-2018. Visually the distribution looks to be normal, if not slightly skewed to the left. This shows times of 4.0-5.5 seconds with the majority of measurements being between 4.0 and 4.5. The tidyverse package which contains ggplot2 was used to create the shuttle visual.

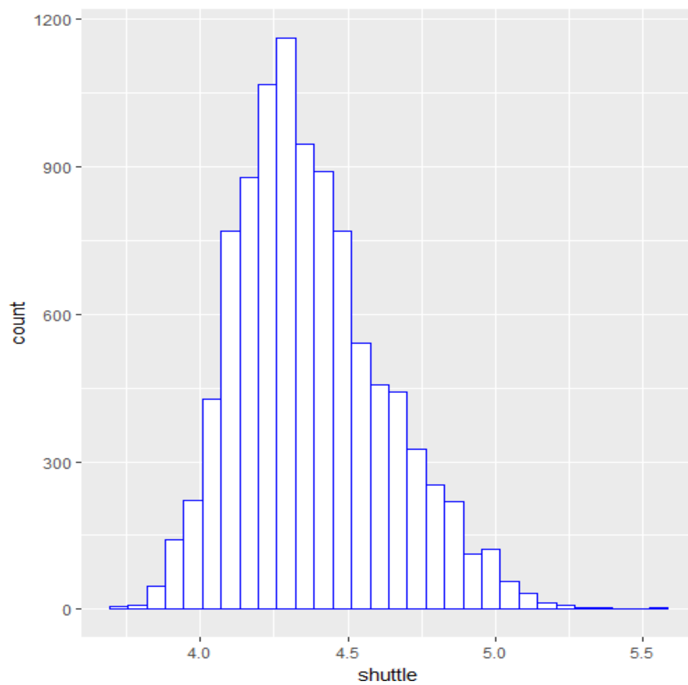

Figure 13: Histogram for the shuttle

The last sample visualization is a scatterplot, seen in figure 14, which shows the shuttle variable compared with the weight variable. We can see the weight on the y-axis and shuttle and x-axis. The points on the graph are placements of players who took part in the shuttle from the dataset. The points are coloured with different shades of blue which represent the age of the player. The plot shows that usually the lighter the weight of the player the faster the shuttle is done, and it seems from the chart that age is not a factor in the speed of the shuttle. The tidyverse package which contains ggplot2 was used to create this visual.

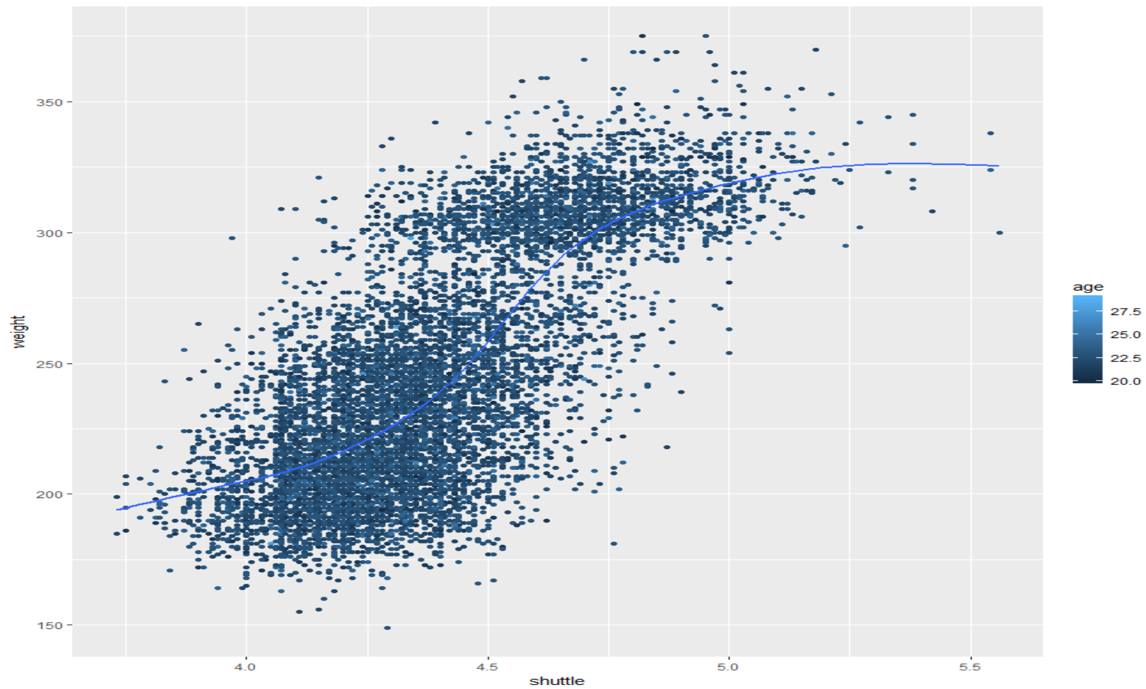

Figure 14: Scatterplot for the shuttle with weight and age

5.2.3.2 Principal Component Analysis

Principal component analysis (PCA) transforms a set of variables that may be correlated to identify which variable maximizes variance. PCA can aid in the model building process to increase the scalability and power of a model. PCA was carried out in R studio using the FactoMiner library.16 Nine variables related to the combine were analyzed using FactoMiner. In figure 15 we can see an individual's factor map and a variables factor map. The variables factor map shows the variables projected on a plane accompanied by two principal components. We can see that the first component has 50.1% of variances and the

[16] https://www.r-bloggers.com/pca-course-using-factominer/ [accessed 11th May 2018]

second has 12.3% which results in the principal components explaining 62.4% of the variance. We can see the variables threecone, shuttle and forty correlates highly with the first component and the variables pick and age correlate with the second component.

The individuals factor map shows the scores of all individuals on the two components, where players are identified by numbers and, given that there are over 9000 players, it may be difficult to differentiate on this plot.

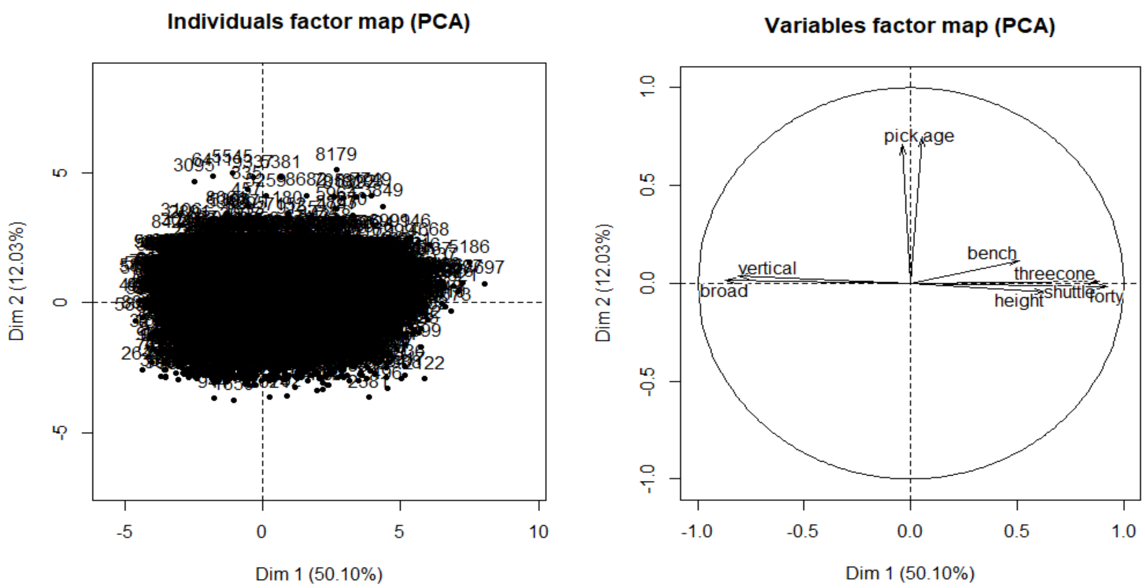

Figure 15: Individuals factor map & Variables factor map

A function was developed to visualize a scree plot to check what eigenvalues are greater than one, as seen in figure 16. This can give insight into what variables to retain.

```
Scree.Plot <- function(R,main="Scree Plot",sub=NULL){
  roots <- eigen(R)$values
  x <- 1:dim(R)[1]
  plot(x,roots,type="b",col='blue',ylab="Eigenvalue",
     xlab="Component Number",main=main,sub=sub)
  abline(h=1,lty=2,col="red")}
```

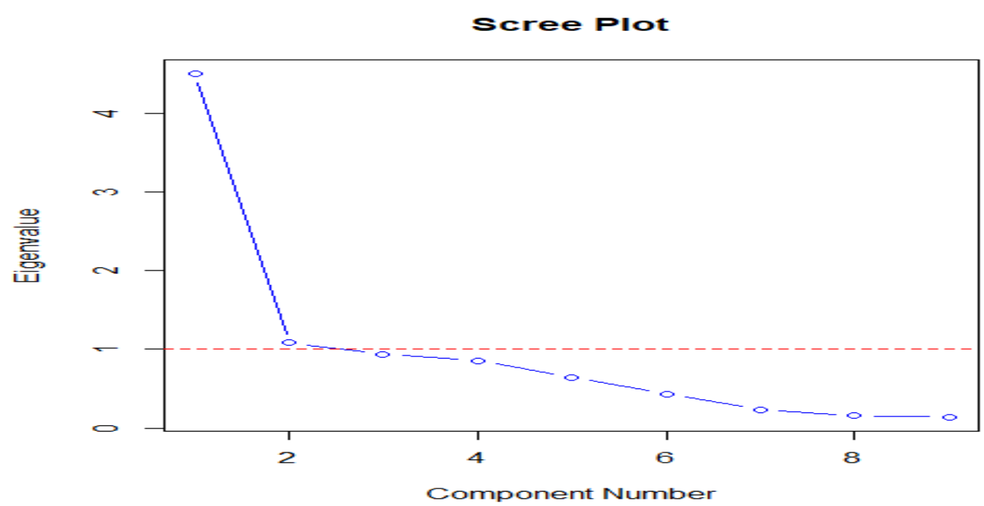

Figure 16: Scree Plot & Code Snippet

We can look at the correlations and p values for each variable for both dimensions. In figure 17 we can see the relationships between each dimension and the p values. This can aid in assigning variables to specific components.

```
$Dim.1
$Dim.1$quanti
          correlation      p.value
forty      0.92679557 0.000000e+00
threecone  0.88520489 0.000000e+00
shuttle    0.88048184 0.000000e+00
height     0.63549759 0.000000e+00
bench      0.51336321 0.000000e+00
age        0.05406985 7.080348e-08
pick      -0.04224199 2.564566e-05
vertical  -0.81204900 0.000000e+00
broad     -0.87156749 0.000000e+00
```

```
$Dim.2
$Dim.2$quanti
          correlation      p.value
age        0.74890475 0.000000e+00
pick       0.70967919 0.000000e+00
bench      0.11741701 8.436973e-32
vertical   0.03896583 1.034553e-04
height    -0.04367858 1.347054e-05
```

Figure 17: Dimensions Correlations and p values

5.2.4 Database Interactions

The library RMySQL provided seamless access to a local MySQL database. In figure 18 we can see the script that connects to the MySQL database. [17]

```
con <- dbConnect(MySQL(),
        user="root", password="password",dbname="NFLDraft", host="localhost")
on.exit(dbDisconnect(con))
```

Figure 18: Database connection script

In figure 19 we can see the script that creates the table in the database and exports the dataset from R studio to the database.

```
dbWriteTable(con, "draftdata", draftdata)
```

Figure 19: Database connection script

[17] https://www.r-bloggers.com/accessing-mysql-through-r/ [accessed 26th April 2018]

5.2.5 Data Mining with Regression

5.2.5.1 Glmnet Regularized Regression

The library Glmnet provided the tools for the implementation of a regularized regression model. Glmnet fits models by the method penalized maximum probability. The idea behind this method is by adding variables to a model it increases the fit to the data. Regularized regression works well with many interactions (see the model matrix in figure 20).

```
sparsemx <- sparse.model.matrix(~  + (1 + factor(pos)) * (1 +
                    factor(short_college) +age + height + weight +forty + bench + vertical +
                   threecone + broad + shuttle +games + seasons +completions + attempts +
                    pass_yards + pass_ints + pass_tds + rec_yards + rec_td + receptions +
                   rush_att + rush_yds + rush_td +solo_tackes + tackles + loss_tackles + ast_tackles +
                   fum_forced + fum_rec + fum_tds + fum_yds +sacks + int + int_td + int_yards + pd +
                   punt_returns + punt_return_td + punt_return_yards +
                   kick_returns + kick_return_td + kick_return_yards),training)
```

Figure 20: Matrix for model

The model is then built with the train set data, the first-round outcome variable, as this is what we want to predict.[18] The library ROCR was used to plot the ROC curve for the sample training data. Figure 21 shows the ROC curve for the model. The code snippet in figure 21 was used to compute the AUC of the ROC curve which resulted in 0.73.[19]

---

[18] https://www.r-bloggers.com/using-sparse-matrices-in-r/ [accessed 20th April 2018]
[19] https://www.r-bloggers.com/a-small-introduction-to-the-rocr-package/ [accessed 11th May 2018]

```
predictions <- prediction(training$sparse.fr.hat[test.set], first.round[test.set])
perform <- performance(predictions, 'tpr', 'fpr')
plot(perform)
performance(predictions, 'auc')
```

Figure 21: Code snippet for ROC curve & AUC

5.2.6 XGBoost Gradient Boosting Regression

The XGBoost library provides gradient boosting frameworks for R. Gradient boosting is a regression technique used to convert weak predictions to strong leaners by minimising bias and variance.[20] To assemble the model, the sparse matrix was used, see code snippet in figure 22.

```
Xmodel <- model.matrix(~ 0 +factor(pos) + year +sparse.fr.hat +
              age + height + weight +forty + bench + vertical +
              threecone + broad + shuttle +games + seasons +
              completions + attempts +pass_yards + pass_ints + pass_tds +
              rec_yards + rec_td + receptions +rush_att + rush_yds + rush_td +
              solo_tackes + tackles + loss_tackles + ast_tackles +
              fum_forced + fum_rec + fum_tds + fum_yds +
              sacks + int + int_td + int_yards + pd +
              punt_returns + punt_return_td + punt_return_yards +
              kick_returns + kick_return_td + kick_return_yards
              ,training)
```

Figure 22: Matrix for dense boosting model

[20] https://www.r-bloggers.com/an-introduction-to-xgboost-r-package/ [accessed 11th May 2018]

```
tuning <- expand.grid(depth = c(3, 4, 5, 6),rounds = c(50, 100, 150, 200, 250)) %>%
  group_by(depth, rounds) %>%do({m <- xgboost(data = Xmodel[train.set,],
              label = as.numeric(training$pick[train.set] <= 32),max.depth = .$depth,
              nround =.$rounds, print.every.n = 50, objective = 'binary:logistic')
    yhat <- predict(m, newdata = Xmodel)data_frame(test.set = test.set, yhat = yhat,
             label = as.numeric(training$pick <= 32)) })
```

Figure 23: Tuning the model

Figure 23 shows the hyperparameter tuning of the model through a grid searching technique which aids in improving the model's performance.[21]

The gradient boosting method returned the best AUC of 0.81 as a result. Both modelling techniques are assessed on their performance in the evaluation section.

[21] https://www.r-bloggers.com/introduction-to-xgboost-r-package/ [accessed 30th March 2018]

5.3 *Evaluation*

The two-machine learning regression modelling approaches returned a mean AUC of 0.73 for Glmnet, and 0.83 for XGBoost from 50 model runs each. These model runs allowed further statistical testing to be performed in SPSS, see statistical testing section. The library ROCR was used to evaluate the Glmnet model. A sample ROC curve generated from this library for the model is seen in figure 24.

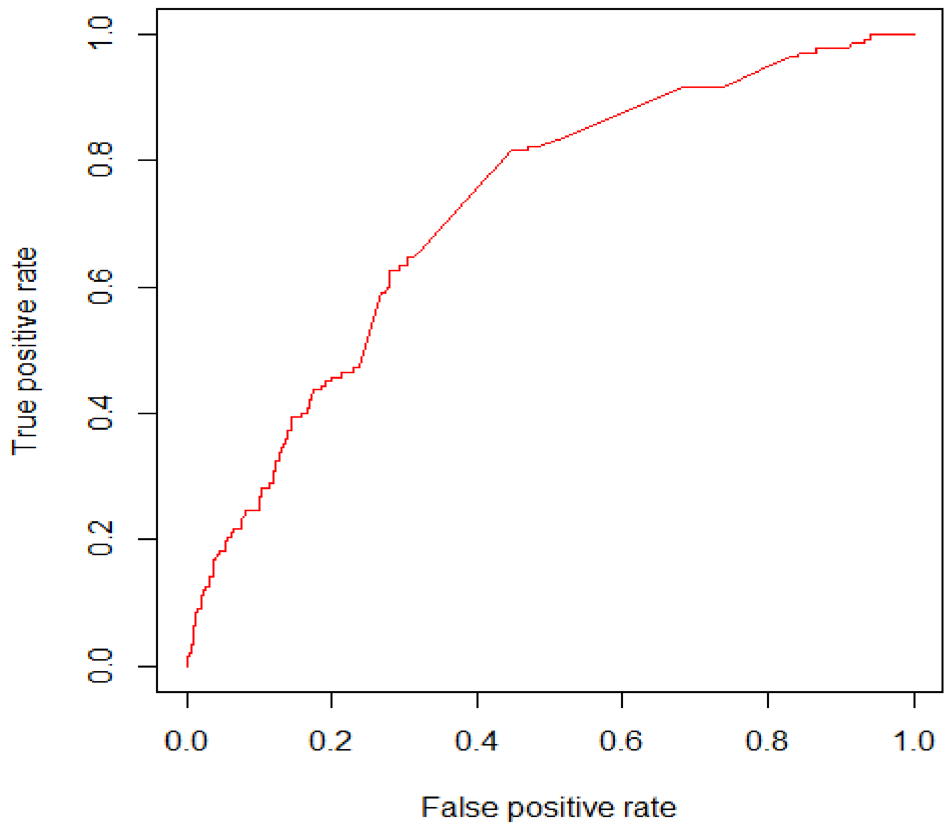

Figure 24: Glmnet ROC Curve

The ROC curve allows the computation of the AUC of the models; it shows the true positive rate and the false positive rate.

The evaluation process of the XGBoost model was carried out using the inbuilt AUC performance function of the library. This was done by tuning the model grid and producing twenty AUCs as seen in figure 25. The top AUC is selected from the tuning process as seen in figure 25.

```
AUC <- tuning %>% ungroup %>%  filter(test.set) %>%group_by(depth, rounds) %>%
  do({ auc <- performance(prediction(.$yhat, .$label), "auc")@y.values[[1]]
    data_frame(auc = auc)}) %>%  ungroup %>% arrange(-auc)
AUC
bestAUC <- AUC %>% head(1)
bestAUC
```

Figure 25: Tuning the best performance AUC of the XGBoost Model

The library tictoc was used to measure the run times for both modelling approaches. The results of this evaluation resulting in the glmnet modelling returning 61.74 seconds elapsed and the XGBoost model returning 16.65 seconds elapsed.[22] This result shows that the XGBoost model clearly performs at a higher speed - see the output of tictoc tests in the visuals section of the appendix.

[22] https://www.r-bloggers.com/timing-in-r/ [accessed 11th May 2018]

Another approach to evaluate our model is to predict first-round draft picks for a certain year. The XGBoost model performed more effectively in the evaluation process, so this model was used to make the predictions. The library knitr was used to visualize the draft table.[23] We can see the predictions for the first-round picks of 2018, 2017 and 2016 in figure 26. Using machine learning methods, the model could predict 14/32 first round draft picks for 2018, 21/32 draft picks for 2017 and 22/32 for the 2016 draft with five predicted first rounders being selected in the second round. Variables such as what positions teams needed going into the draft, injuries and off the field issues would have a strong influence on the official picks of the draft. These variables paired with expert opinion and machine learning models could prove to become invaluable to the NFL draft process.

1	257	Saquon Barkley	Penn St.	RB
2	257	Denzel Ward	Ohio St.	CB
3	257	Kentavius Street	North Carolina St.	DE
4	257	Jamarcus King	South Carolina	CB
5	257	Bradley Chubb	North Carolina St.	DE
6	257	Damon Webb	Ohio St.	S
7	257	Jerome Baker	Ohio St.	OLB
8	257	Minkah Fitzpatrick	Alabama	DB
9	257	Troy Apke	Penn St.	S
10	257	Quenton Meeks	Stanford	DB
11	257	Sam Hubbard	Ohio St.	DE
12	257	Marcus Davenport	Texas-San Antonio	DE
13	257	Ronald Jones	USC	RB
14	257	Chris Campbell	Penn St.	CB
15	257	Lamar Jackson	Louisville	QB
16	257	Auden Tate	Florida St.	WR
17	257	Tremaine Edmunds	Virginia Tech	LB
18	257	Rasheem Green	USC	DE
19	257	J.T. Barrett	Ohio St.	QB
20	257	Sam Darnold	USC	QB
21	257	Dominick Sanders	Georgia	S
22	257	Brian Allen	Michigan St.	C
23	257	Holton Hill	Texas	CB
24	257	Foley Fatukasi	Connecticut	DT
25	257	Chad Thomas	Miami	DE
26	257	Jojo Wicker	Arizona State	DE
27	257	Karaun White	West Virginia	WR
28	257	Billy Price	Ohio St.	C
29	257	Jalyn Holmes	Ohio St.	DE
30	257	Deontay Burnett	USC	WR
31	257	Mike White	Western Kentucky	QB
32	257	Josh Jackson	Iowa	CB

1	Marshon Lattimore	Ohio St.	CB
2	Jamal Adams	LSU	S
3	Malik Hooker	Ohio St.	S
4	Evan Engram	Mississippi	TE
5	Ryan Ramczyk	Wisconsin	T
6	T.J. Watt	Wisconsin	OLB
7	Haason Reddick	Temple	LB
8	Gareon Conley	Ohio St.	CB
9	Patrick Mahomes	Texas Tech	QB
10	David Sharpe	Florida	T
11	Cam Robinson	Alabama	T
12	Roderick Johnson	Florida St.	T
13	Takkarist McKinley	UCLA	DE
14	DeShone Kizer	Notre Dame	QB
15	Evan Engram	Mississippi	TE
16	Charles Harris	Missouri	OLB
17	JuJu Smith-Schuster	USC	WR
18	Jarrad Davis	Florida	LB
19	Curtis Samuel	Ohio St.	WR
20	Garett Bolles	Utah	T
21	Sam Tevi	Utah	T
22	Derek Barnett	Tennessee	DE
23	Reuben Foster	Alabama	LB
24	Solomon Thomas	Stanford	DE
25	Ryan Glasgow	Michigan	DT
26	Malachi Dupre	LSU	WR
27	Marquel Lee	Wake Forest	LB
28	Myles Garrett	Texas A&M	DE
29	Adrian Colbert	Miami (FL)	DB
30	John Ross	Washington	WR
31	Jonathan Allen	Alabama	DE
32	Mike Williams	Clemson	WR

[23] https://www.r-bloggers.com/knitr-in-a-knutshell-tutorial/ [accessed 11th May 2018]

1	Joey Bosa	Ohio St.	DE
2	Darron Lee	Ohio St.	OLB
3	Eli Apple	Ohio St.	CB
4	Laremy Tunsil	Mississippi	T
5	Jalen Ramsey	Florida St.	CB
6	Ezekiel Elliott	Ohio St.	RB
7	Taylor Decker	Ohio St.	T
8	Ronnie Stanley	Notre Dame	T
9	Sheldon Rankins	Louisville	DT
10	Kenny Clark	UCLA	DT
11	DeForest Buckner	Oregon	DE
12	Emmanuel Ogbah	Oklahoma St.	DE
13	Artie Burns	Miami (FL)	CB
14	Carson Wentz	North Dakota St.	QB
15	Vernon Hargreaves	Florida	CB
16	Keanu Neal	Florida	SS
17	Vernon Butler	Louisiana Tech	DT
18	Jared Goff	California	QB
19	Shaq Lawson	Clemson	DE
20	Joshua Garnett	Stanford	G
21	Jack Conklin	Michigan St.	T
22	Carl Nassib	Penn St.	DE
23	Christian Hackenberg	Penn St.	QB
24	Will Fuller	Notre Dame	WR
25	Caleb Benenoch	UCLA	T
26	Corey Coleman	Baylor	WR
27	Michael Thomas	Ohio St.	WR
28	William Jackson	Houston	CB
29	Dean Lowry	Northwestern	DE
30	Su'a Cravens	USC	OLB
31	Andrew Billings	Baylor	NT
32	Tyler Boyd	Pittsburgh	WR

Figure 26: 2018, 2017 & 2016 first round predictions

6 Graphical User Interface (GUI) Layout

The data visualization tool Tableau was used to interact with the project's dataset in R studio. The library Rserve provided a seamless connection from R to Tableau.[24] Once the data was loaded into Tableau, visuals were developed. The result of utilizing Tableau produced a user-friendly interactive dashboard as seen in figure 27.

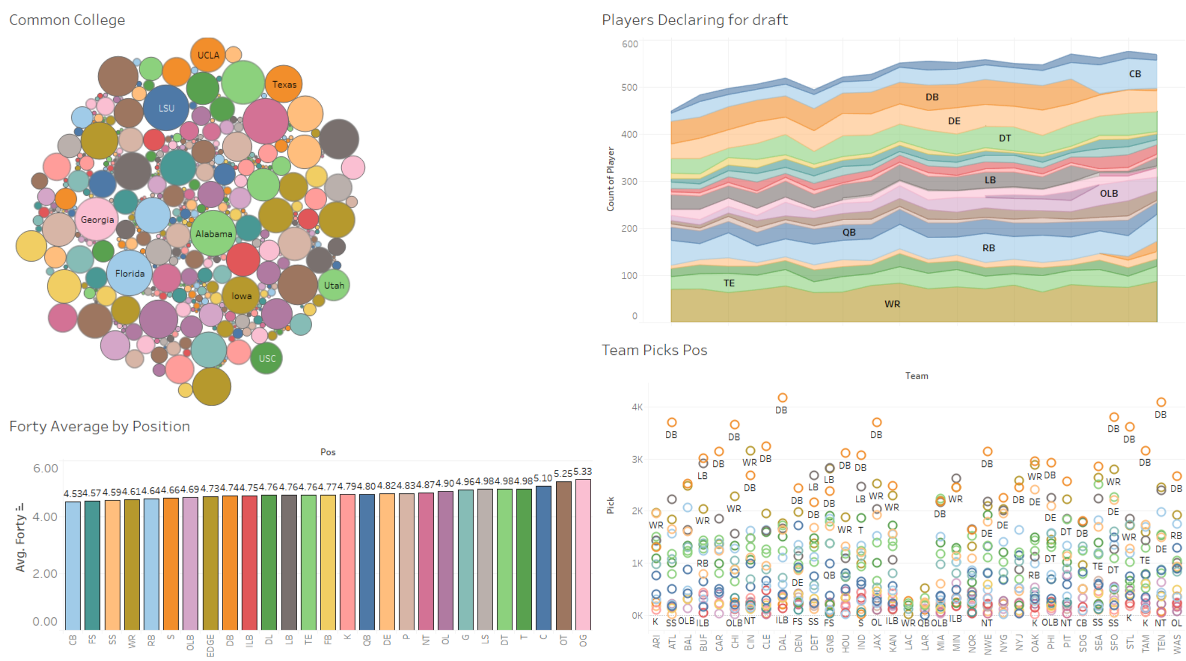

Figure 27: Tableau Dashboard

The dashboard shows visuals on the most common colleges attended by the players who enter the draft, players who declare for the draft by position throughout the years, the forty-yard dash by position and the position picked by NFL teams throughout the draft from 2000-2017. In figure 28 we can see the

[24] https://www.r-bloggers.com/dream-team-combining-tableau-and-r/ [accessed 11th May 2018]

interactive aspect of the dashboard. By clicking on any area of the dashboard, the dashboard updates dynamically. Figure 28 shows the position of defensive end (DE) picked in 2017, the colleges that defensive ends declared for the draft in 2017, the times of defensive ends in the forty-yard dash for that year and the number of defensive ends that NFL teams drafted from 2000-2017.

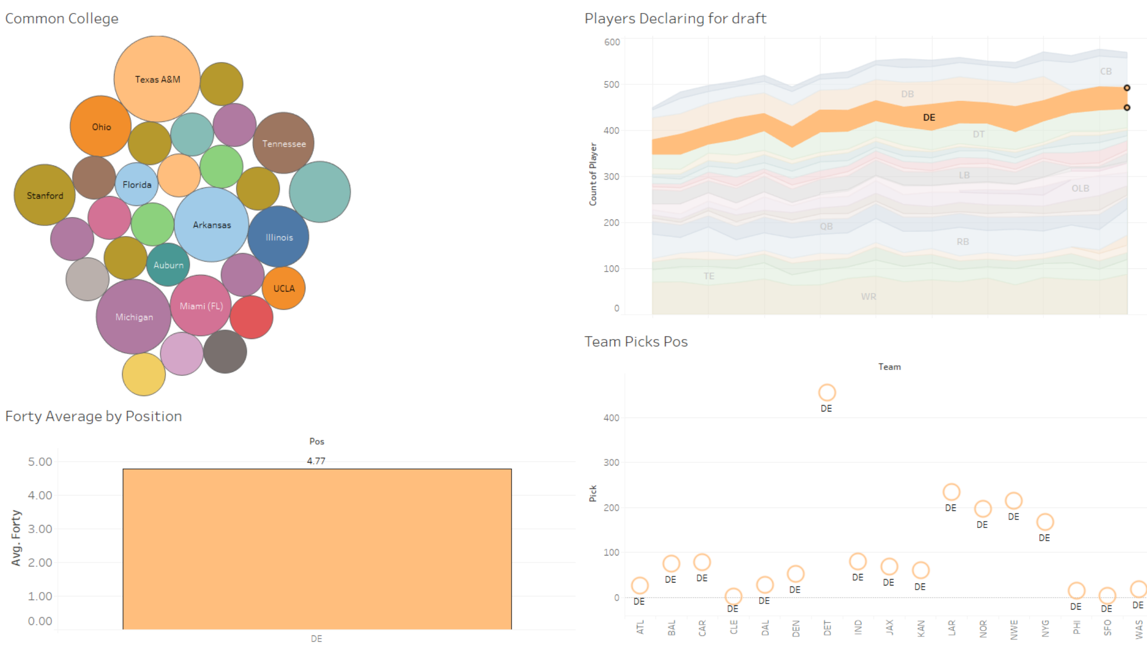

Figure 28: Tableau Dashboard Interaction

Tableau has allowed the project data to be represented to the untrained eye in a customer friendly way.

6.1 *Testing*

6.1.1 Unit Testing

Unit testing was carried out in R Studio. Unit testing is a widely used software testing method that tests small parts of an application; this can help with application robustness. The prevention of bugs increases the quality of the code and simplifies future changes to the code. A tests.r file contains tests for each R file of the project. The library testthat was employed to run these unit tests.[25] Unit tests are methods that validate code and can help maintain integrity. See code snippet in figure 29 below to see how the test is carried out, and the output returned when the test is passed.

[25] https://www.r-bloggers.com/unit-testing-with-r/ [accessed 11th May 2018]

```
library(testthat)
source("C:/Users/Daniel/Dropbox/sproject/dataprep.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/utils.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/glmnet.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/xgboost.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/explore.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/database.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/PCA.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary")
source("C:/Users/Daniel/Dropbox/sproject/tableau.R")
test_results <- test_dir("C:/Users/Daniel/Dropbox/sproject", reporter="summary"
```

```
== DONE ========================================================================
```

Figure 29: Unit test code and output of the successful test

If a unit test fails, the testthat library will output an error, for example in figure 30 the error is given regarding the database connection class. This error was fixed by amending the connection handle in the database file.

```
Error in .local(dbObj, ...) :
  internal error in RS_DBI_getConnection: corrupt connection handle
```

Figure 30: Unit test output of the unsuccessful test

6.1.2 Statistical Testing

The two regression model approaches were run 50 times each. This allowed a mean to be computed for the models. Statistical testing was applied to see if there was a difference between both model groups during the 50 runs. The data was recorded using Microsoft Excel and exported in SPSS to conduct statically testing. A Mann-Whitney U test was conducted on the two datasets. A Mann-Whitney U test compares differences between two independent groups for non-normal data. As our datasets are at a small sample size of 50 each, it is good practice to use a non-parametric test like the Mann-Whitney U test.[26]

The hypothesis for our test are: -

- *Null* – H0: the distribution of scores for the two groups are equal;
- *Alternative* – HA: the distribution of scores for the two groups are not equal.

This test was carried out using a significance level of 0.05. Figure 26 shows the descriptive statistics, showing a mean for Glmnet of .73 and XGBoost if .82. in figure 32 we can see the result of our Mann-Whitney U test which shows the distribution of AUC is the same across both categories of models. We are given the decision to reject the null hypothesis and accept the alternative hypothesis that the distribution of scores for the two groups are not equal.

[26] https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php [accessed 11th May 2018]

Group Statistics

	MODEL	N	Mean	Std. Deviation	Std. Error Mean
AUC	Glmnent	50	.730522	.0220099	.0031127
	XGBoost	50	.824918	.0250160	.0035378

Figure 26: Descriptive Statistics for Models

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of AUC is the same across categories of MODEL.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 31: Test result from SPSS

6.1.3 Customer testing

The customer testing for this project was carried out by displaying the Tableau dashboard and the results of the predictions for the NFL draft for 2016-2018. Each customer was also presented with the abstract and conclusion of the project to provide a better understanding of the study at hand.

Ten customers, who are involved in an Irish NFL Podcast and fantasy football league, took part in the testing. After interacting with the dashboard, being shown the results of the predictions and reading the abstract - each customer was asked to partake in a five-question survey. Each question uses the five-point scale Likert answering method.[27]

The results of the survey are as follows:

1. How satisfied are you with the look and feel of this dashboard?

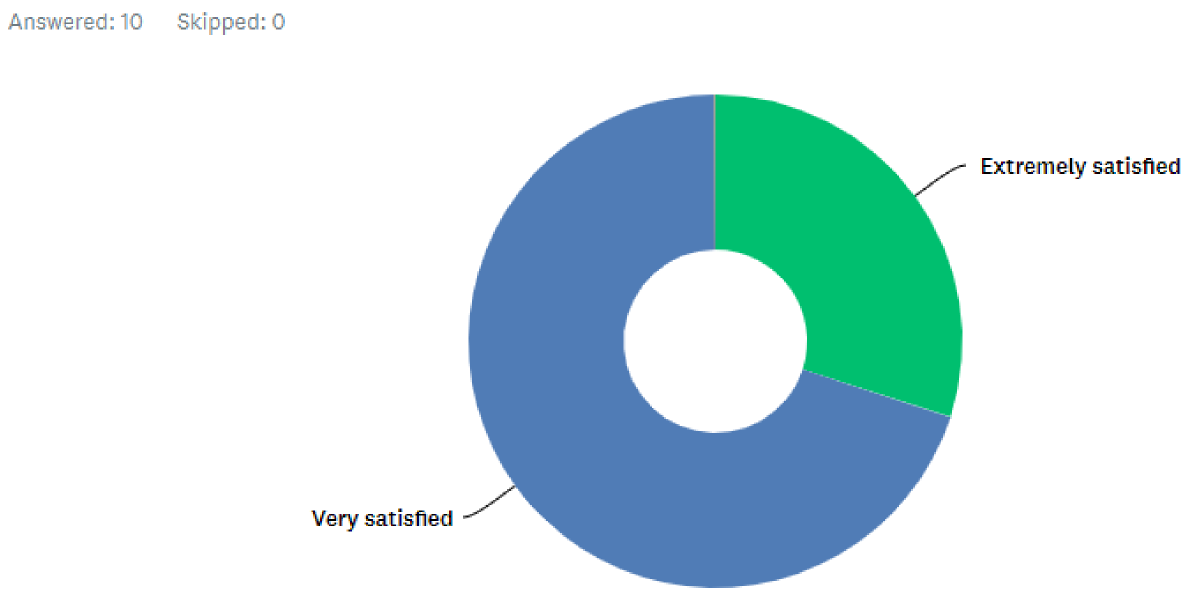

Figure 32: Question 1

[27] https://www.surveymonkey.com/mp/likert-scale/ [accessed 11th May 2018]

2. **How satisfied are you that combine and college statistics play a role in draft selection?**

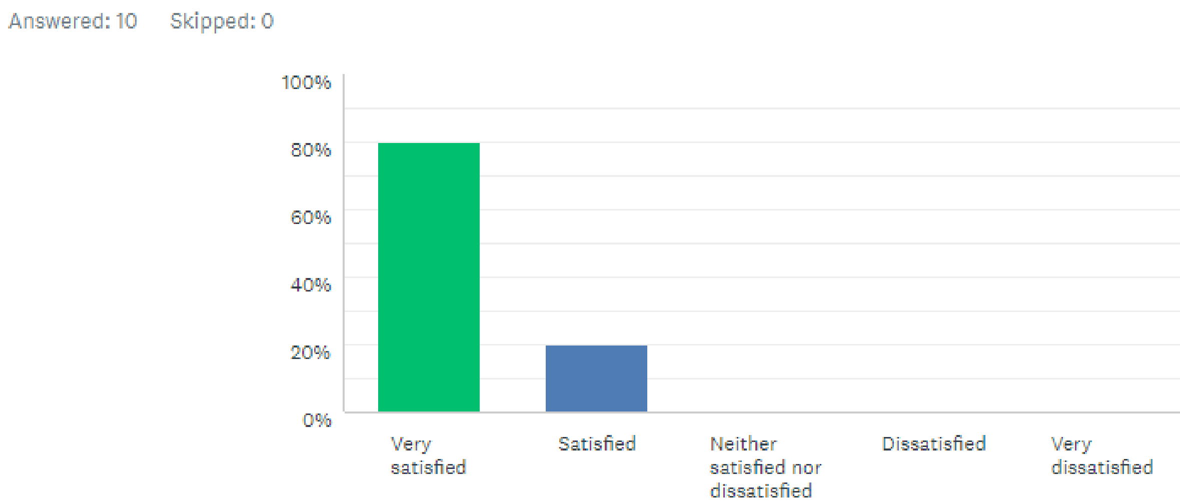

Figure 33: Question 2

3. **How satisfied are you with the machine learning predictions?**

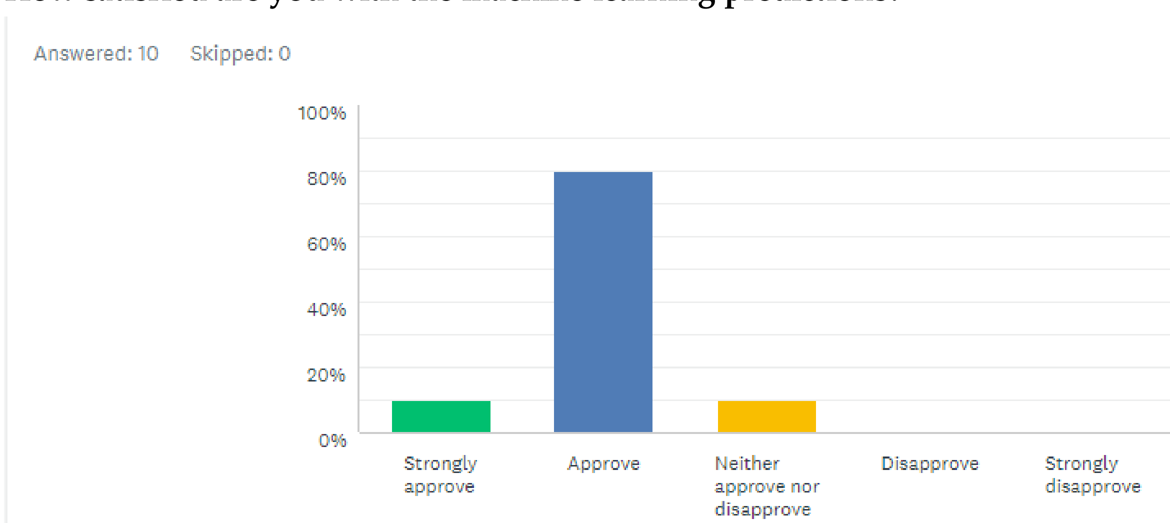

Figure 34: Question 3

4. **Do you feel machine learning has a place in the NFL Draft process?**

Answered: 10 Skipped: 0

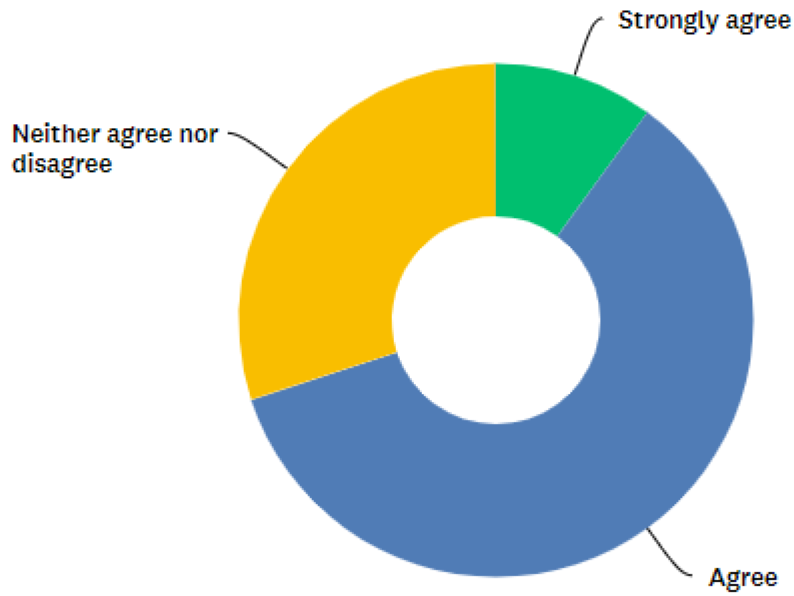

Figure 35: Question 4

5. Would you be interested in an interactive version of this dashboard?

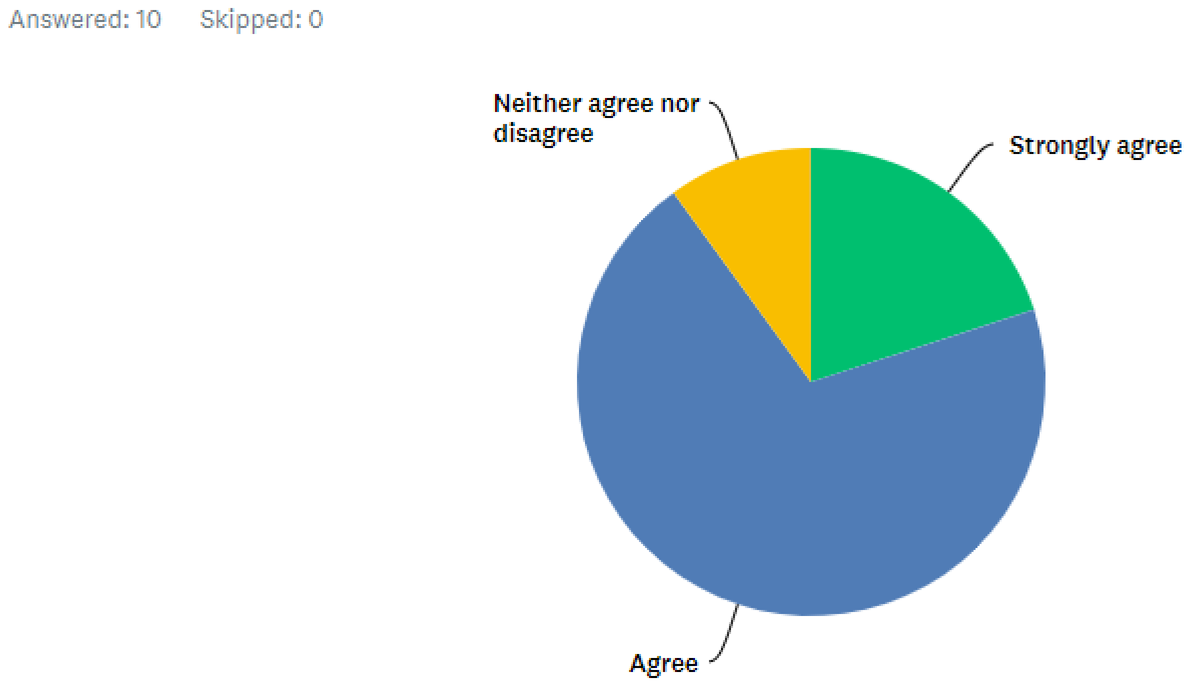

Figure 36: Question 5

The overall results were analysed in Microsoft Excel using the data analysis tool pack. The results returned the bar chart shown in figure 37.

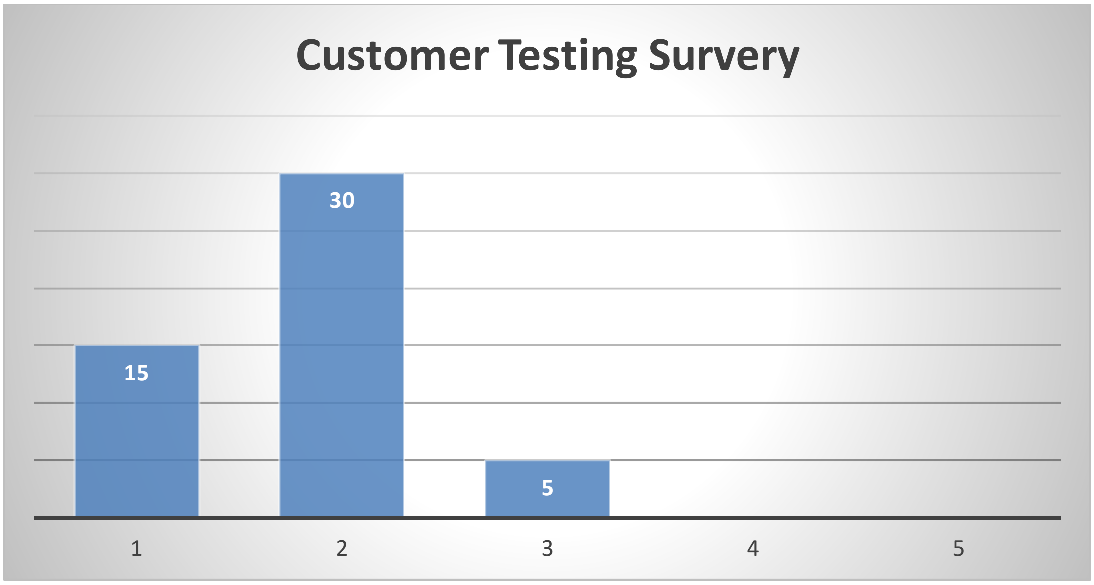

Figure 37: Bar chart on overall result based on the five-point scale

The results are based on the Likert five-point scale where the 1s in the diagram would fall under strongly agree or extremely satisfied, and the 5s would fall under strongly disagree and very dissatisfied.

The overall results of the survey returned positive feedback based on the project.

6.2 Exploration Plan

Opportunity

Problem Summary

The aim of this study is to predict the first round of the NFL Draft. The research needs data on NCAA college football players, NFL Combine Statistics and previous NFL Draft results.

Solution Summary

The solution to this problem is carried out by using machine learning algorithms for prediction. The data and results are integrated into a user-friendly dashboard.

Market

The target market for this project is fans and consumers of the NFL and NFL general managers.

Competition

Currently, NFL draft predictions are made by NFL experts in a mock draft format. There are no machine learning draft applications currently available.

6.3 Impact Summary

An impact summary is aimed to encourage researchers to think about potential recipients who could benefit from their project.[28] The impact summary is made up of who the project will bring benefit to and how. A pathway to the impact

[28] https://www.gla.ac.uk/media/media_426471_en.pdf [accessed 11th May 2018]

statement follows an impact summary; this statement details the activities and strategies the project should develop to deliver on the impact summary.

Impact Summary

This project is aimed at benefiting consumers, fans of the NFL and GMs of the NFL. The project will help consumers and fans by providing a comprehensive analysis of the NFL Draft, Combine and NCAA College statistics. Fans will be able to view visuals and graphs of their favourite college player on their college career, combine performance and view if those statistics have warranted a prediction for a first-round pick in the NFL draft. GMs on NFL teams will benefit from this project by having up-to-date analysis of the NFL Draft based on college and combine performance and previous drafts. Based on the results of the predictions carried out in this study these stats alone forecast the first round of the NFL draft.

Pathways to Impact

To fulfil the pathways to impact of this project customer testing took place at the end of the project lifecycle. These results are viewable in the customer testing section of the project. Customers were giving an interactive dashboard of statistics and charts developed from the data gained from this study. Customers were also shown the abstract of this study, the conclusion and the predictions for 2016 – 2018 of the draft. Customers tested the dashboard and took part in a questionnaire to provide feedback on the project. The feedback received was overall positive.

7 Conclusions & Further Development

The initial goal of the study was to predict the likelihood of a player being selected in the first round of the draft. As seen from the results, predictions have been made on three years of the NFL draft with success.

The results gathered from the predictions make it clear that machine learning methods may aid NFL draft player selection process. The diversity of the results from the 2016 draft, 22/32, 2017 draft, 21/32, and the 2018 draft, 14/32, show that each year has multiple factors involved in the selection process during the draft.

The project faced some challenges throughout its life cycle. Working with mixed types of data can always be problematic. Packages such as stringr helped deal with this issue at an early stage of the project. Another challenge that was anticipated from an early stage was that not all factors of player selection are performance based. The reality is that draft picks are not solely based on player performance in college or the combine. Other factors are also relevant such as maturity level, injury history and the teams' positional needs. These "other factors" complicate what would otherwise be a simple process.

Of the two machine learning techniques, the gradient boosting method produced superior results. The XGBoost model borrowed techniques from the sparse model during the assembly process which makes it evident that both modelling methods have their advantages and disadvantages. It is clear from the overall project results that data analysis has a place in the NFL; even though it might only have one foot in the door now - it is a growing industry. GMs who are the overseers when selecting players for their teams might need to add machine learning skills to their arsenal in future years to stay ahead of the curve.

Some further development aspects might include the following options: -

Development of more machine learning techniques to assess and analyze in comparison to current methods. From the literature review carried out, it is clear that many machine learning approaches can be used on the draft. The literature reviews delineated above focused on one NFL position to predict draft outcome. This study predicts all players entering the draft for all positions. To apply some of the machine learning techniques seen in the literature reviews such as neural networks, in conjunction with this study might provide further insights into the NFL draft.

Building an interactive application with a front end where fans and customers can view predictions based on specific input factors may have commercial viability. The interactive tableau dashboard received extremely positive feedback from the customer testing. To use the algorithms of this paper to develop a user dashboard where the user could input specific options could be a productive option for future development of this study.

There are always opportunities to make improvements to a dataset. The dataset for this project could be improved by obtaining NFL scout reports and by using text mining techniques which could improve the overall predictive powers of the model.

8 References

1 King, C. (2018). [online] Fisherpub.sjfc.edu. Available at: https://fisherpub.sjfc.edu/cgi/viewcontent.cgi?referer=https://www.google.ie/&httpsredir=1&article=1064&context=sport_undergrad [Accessed 3 May 2018].

2 McKenzie, G. (2018). [online] Mospace.umsystem.edu. Available at: https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/47027/research.pdf?sequence=2&isAllowed=y [Accessed 3 May 2018].

3 McCann, A. (2018). *Reducing NFL Draft Risk_FINAL.pdf*. [online] Google Docs. Available at: https://drive.google.com/file/d/0B7xv5XOVMuVzbXVJREhtSXQzU0E/view [Accessed 3 May 2018].

4 Fayyad, U. (2018). [online] Shawndra.pbworks.com. Available at: http://shawndra.pbworks.com/f/The KDD process for extracting useful knowledge from volumes of data.pdf [Accessed 3 May 2018].

5 Wolfson, J. (2018). *THE QUARTERBACK PREDICTION PROBLEM: FORECASTING THE PERFORMANCE OF COLLEGE QUARTERBACKS SELECTED IN THE NFL DRAFT*. [online] Biostat.umn.edu. Available at: http://www.biostat.umn.edu/ftp/pub/2010/rr2010-022.pdf [Accessed 12 May 2018].

6 Dhar, A. (2018). *Drafting NFL Wide Receivers: Hit or Miss?*. [online] Stat.berkeley.edu. Available at: https://www.stat.berkeley.edu/~aldous/157/Old_Projects/Amrit_Dhar.pdf [Accessed 12 May 2018].

9 Appendix

9.1 *Definitions, Acronyms, and Abbreviations*

NFL: National Football League

PCA: Principal Component Analysis

NCAA: National Collegiate Athletic Association

KDD: Knowledge Discovery and Data mining, a type of methodology used for the project.

API: Application Program Interface, a set of routines, protocols, and tools for building software applications.

MYSQL: My Structured Query Language

GUI: Graphical User Interface, this is the visual aspect of the project, and in my case a dashboard like tableau

QB: Quarterback

AUC: Area under the curve, used to view the performance of a model

ROC: Relative Operating Characteristics, used in model evaluation

9.2 *Visuals*

```
> library(tictoc)
> tic()
> model <- cv.glmnet(sparsemx[train.set,],
+                    first.round[train.set],
+                    alpha = 0.05,
+                    family = 'binomial')
>
> training$sparse.fr.hat <- predict(model, newx = sparsemx, type = 're
sponse')[,1]
> toc()
61.74 sec elapsed
```

Figure 38: TicToc timer test on glmnet Model

```
> tic()
> tuning <- expand.grid(depth = c(3, 4, 5, 6),
+                            rounds = c(50, 100, 150, 200, 250)) %>%
+   group_by(depth, rounds) %>%
+   do({
+     m <- xgboost(data = Xmodel[train.set,],
+                  label = as.numeric(training$pick[train.set] <= 32),

+                  max.depth = .$depth,
+                  nround =.$rounds,
+                  print.every.n = 50,
+                  objective = 'binary:logistic')
+     yhat <- predict(m, newdata = Xmodel)
+     data_frame(test.set = test.set, yhat = yhat,
+                label = as.numeric(training$pick <= 32))
+   })
> toc()
16.65 sec elapsed
```
Figure 39: TicToc timer test on XGBoost Model

9.3 *Technical Details*

The computer specifications I am performing my project on are:

- Intel Core i7-7700HQ Quad Core Processor
- 32GB DDR4-2400MHz
- 1TB Solid State Drive

9.4 *Requirements*

9.4.1 Functional requirements

For the functional requirements of this project, I will preview use cases that are needed to make the project practical. In my case, I am the primary user, and I will show how I am accessing the data and how I plan on following the KDD methodology.

The below use cases are the most critical functional requirements for accessing the data that is needed to make my project viable. I will provide the techniques and methods used to achieve these use cases in detail below.

9.4.2 Use Case

9.4.3 Requirement 1 Web Scrape Draft Data

Description & Priority

This is the highest priority of use cases for the project to be successful as the draft data scraped from pro-football-reference will build the foundation of my project.

Use Case

Scope

The scope of this use case is to retrieve data from pro-football-reference, transform and evaluate the data and gain knowledge from such data.

Description

This use case describes the process of web scraping data from pro-football-reference.

Use Case Diagram

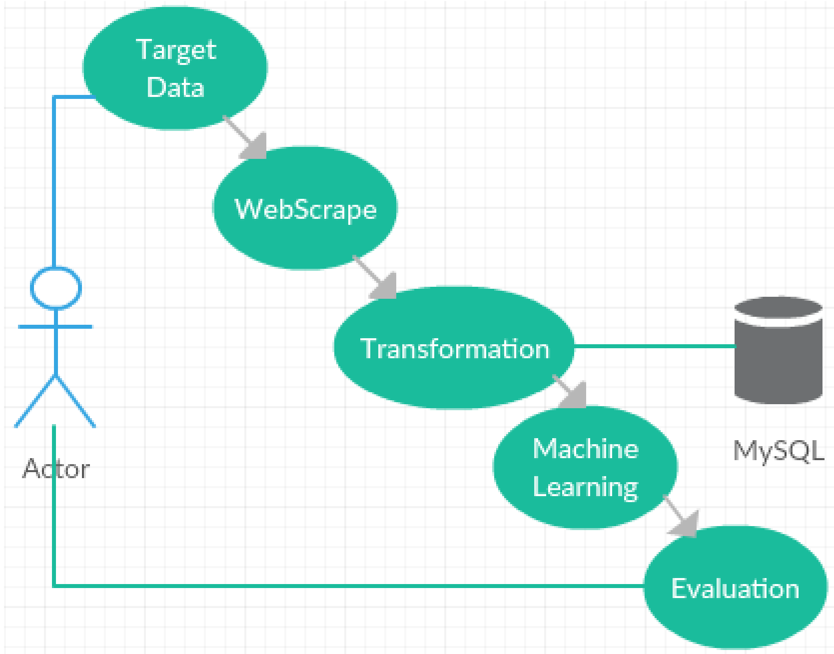

Figure 40: pro-football-reference Data Use Case

Flow Description

Pre-condition

The system is in initialisation mode when the script is ready to be run in R Studio.

Activation

The use case starts when a user runs the script to retrieve data from pro-football-reference.

Main flow

1. The system identifies the script has started
2. The web scraping process has begun
3. Pro-football-references responds with the data
4. The data is collected
5. The data is then transformed
6. The data is sent to the database so it can be evaluated for patterns
7. Knowledge is gained from the data

Exceptional flow

E1: Failed to pull Data

1. The system identifies the script has started
2. The web scraping process has begun
3. Pro-football-reference domain has not responded and there a connection error

Termination

The system presents the data to be evaluated.

Postcondition

The script is ready to be rerun for the next attempt to scrape data.

9.4.4 Requirement Web Scrape Combine Data

Description & Priority

This is also a high priority use case for the project to be successful as the combine data scraped from pro-football-reference will complement the analysis of the draft data already collected.

Use Case

Scope

The scope of this use case is to retrieve data from pro-football-reference, transform and evaluate the data and gain knowledge from such data.

Description

This use case describes the process of web scraping data from pro-football-reference.

Use Case Diagram

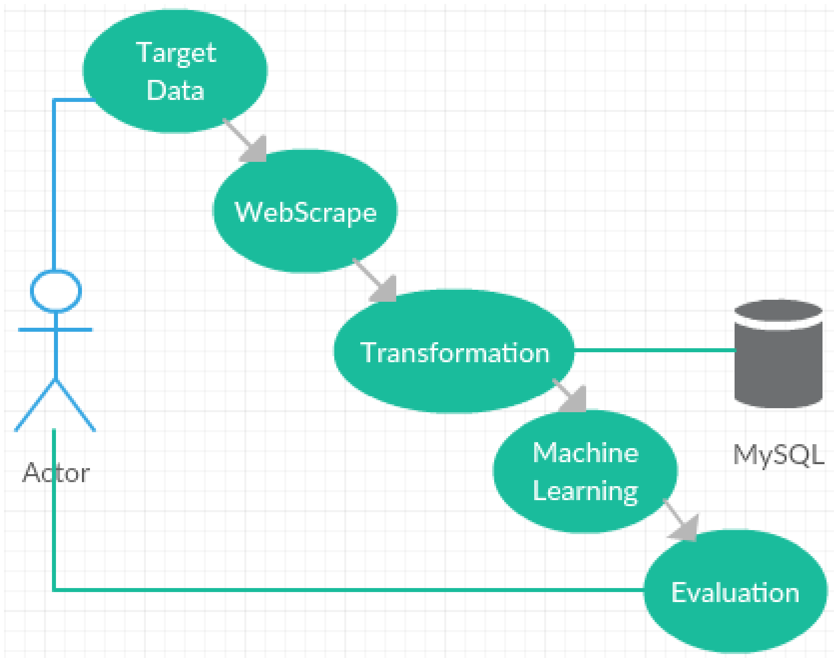

Figure 41: pro-football-reference Data Use Case

Flow Description

Pre-condition

The system is in initialisation mode when the script is ready to be run in R Studio.

Activation

The use case starts when a user runs the script to retrieve data from pro-football-reference

Main flow

1. The system identifies the script has started
2. The web scraping process has begun
3. Pro-football-references responds with the data
4. The data is collected
5. The data is then transformed
6. The data is sent to the database so it can be evaluated for patterns
7. Knowledge is gained from the data

Exceptional flow

E1: Failed to pull Data

1. The system identifies the script has started
2. The web scraping process has begun
3. Pro-football-reference domain has not responded and there a connection error

Termination

The system presents the data to be evaluated.

Postcondition

The script is ready to be rerun for the next attempt to scrape data.

9.4.5 Requirement Web Scrape College Data

Description & Priority

This is also a high priority use case for the project to be successful as the NCAA College data scraped from pro-football-reference will complement the analysis of the draft data already collected.

Use Case

Scope

The scope of this use case is to retrieve data from pro-football-reference, transform and evaluate the data and gain knowledge from such data.

Description

This use case describes the process of web scraping data from sports-reference

Use Case Diagram

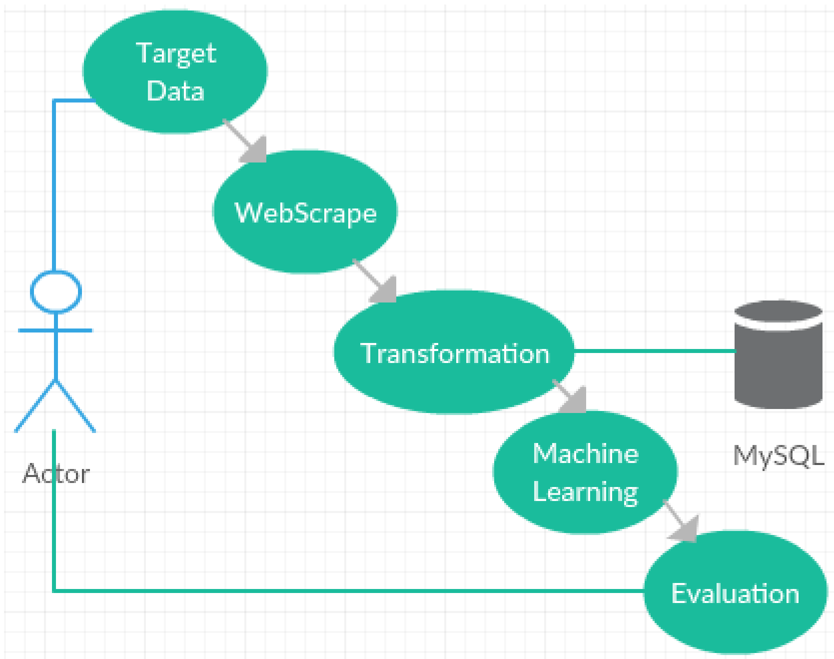

Figure 42: sports-reference Data Use Case

Flow Description

Pre-condition

The system is in initialisation mode when the script is ready to be run in R Studio.

Activation

The use case starts when a user runs the script to retrieve data from pro-football-reference

Main flow

1. The system identifies the script has begun
2. The web scraping process has started
3. Pro-football-references responds with the data
4. The data is collected
5. The data is then transformed
6. The data is sent to the database so it can be evaluated for patterns
7. Knowledge is gained from the data

Exceptional flow

E1: Failed to pull Data

The system identifies the script has started

The web scraping process has begun
Pro-football-reference domain has not responded and there a connection error

Termination

The system presents the data to be evaluated.

Postcondition

The script is ready to be rerun for the next attempt to scrape data.

9.4.6 Requirement Database Creation

Description & Priority

This requirement will be utilized at the start the project. Whether the data is being stored on a hard drive or a database such as MySQL this requirement is of high importance.

Use Case

Scope

The scope of this use case is to create a safe and secure storage system that will be maintained during the project.

Use Case Diagram

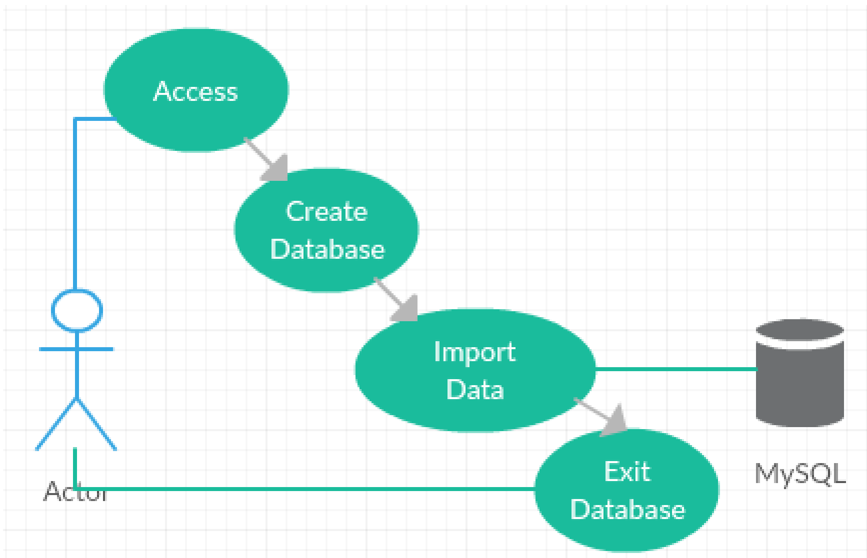

Figure 43: Database Creation Use Case

Flow Description

Pre-condition

The database is accessible through R studio, Tableau or SQL Workbench.

Activation

This use case starts when the storage is accessed.

Main flow

1. The user starts the database application
2. The user creates some form of storage for the data
3. The user imports the dataset
4. The application is exited

Exceptional flow

E1: Failed to start the application

1. The system fails to start
2. The system is restarted

3. The use case starts from the beginning of the main flow

Termination

The database application is closed.

Postcondition

The data is stored securely and safe.

9.4.7 Requirement Data Visualization

Description & Priority

This requirement will be at the end of the project when visualising the data on a dashboard software such as Tableau.

Use Case

Scope

The scope of this use case is for the system user to create and visualize the data.

Use Case Diagram

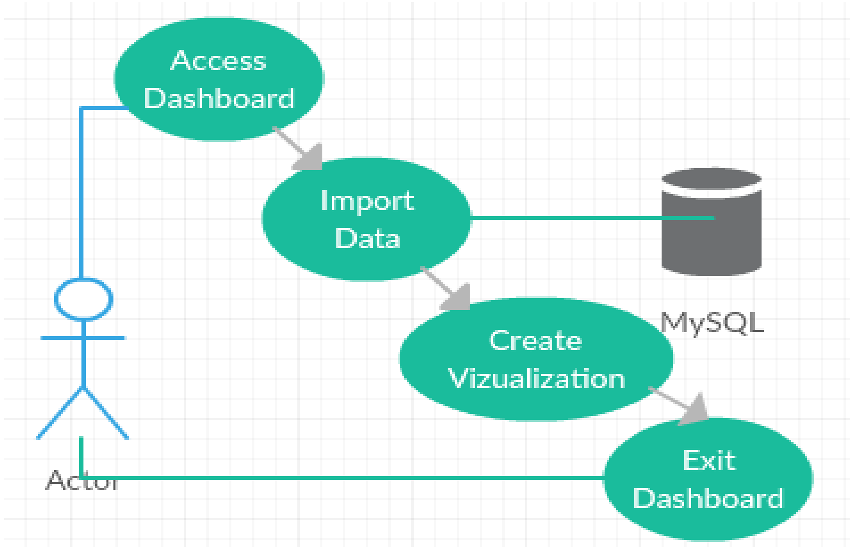

Figure 44: Dashboard Creation Use Case

Flow Description

Pre-condition

Dashboard software is accessible.

Activation

This use case starts when the dashboard software is accessed.

Main flow

1. The user starts dashboard software
2. The user imports the data from the database
3. The user creates the visualizations from the data
4. The application is exited

Exceptional flow

E1: Failed to start the application

1. The system fails to start
2. The system is restarted
3. The use case starts from the beginning of the main flow
4.

Termination

The dashboard application is closed.

Postcondition

The data visualization is stored safely.

9.4.8 Data Requirements

The data requirement is vital to the study as without data the study cannot be carried out. The study aims to acquire exceptional data of high quality and integrity that lacks redundancies.

9.4.9 Performance/Response time requirement

The performance and response time requirement for the project will be related to the access of data. When using R to scrape or pull data from certain websites or APIs, it will be dependent how the website or API responds. Once the data is acquired, the performance then relies on the machine or database it is stored on and the integrity of the data itself.

9.4.10 Availability Requirement

The availability requirement of the project will be related to the availability of the datasets. Once the data is taken from sources, it will be available from the database or the hard drive it is stored on.

9.4.11 Recover requirement

The recover requirement of the project is of great importance. As this is a project focused on data it is vital that the data and files are backed up in case of disaster. All files will be backed up on Dropbox, and the scripts and data will be stored on Bitbucket.

9.4.12 Security requirement

The project will be stored on a fingerprint identification accessed laptop, and the files and scripts as stated in the recover requirement will be on password protected sites such as Bitbucket and Dropbox.

9.4.13 Reliability requirement

The web scraping technique is widely used in the data analytics community with packages that are regularly updated.

9.4.14 Maintainability requirement

The project is due for final assessment in May; however, as I have a keen interest in the NFL I intend to keep the project going as a hobby of interest.

9.4.15 Extendibility requirement

There is a proper scope for extendibility for this project; please see the System Evolution section.

9.4.16 Reusability requirement

The scripts used to access and manipulate the data will be reusable in other projects as I intend to use good programming practice to make this requirement viable by using comments.

9.5 *Project Plan*

Below is a Gantt chart which sets out a guideline for the project to ensure that milestones and deliverables of the projects are met.

For the first semester of college, I plan on laying the foundation work for my project. My goal is to have completed the required research on datasets, techniques and methods and merge this with the training and skills I gain from the college modules of my first semester and focus on the application and technical part of my project in my second semester.

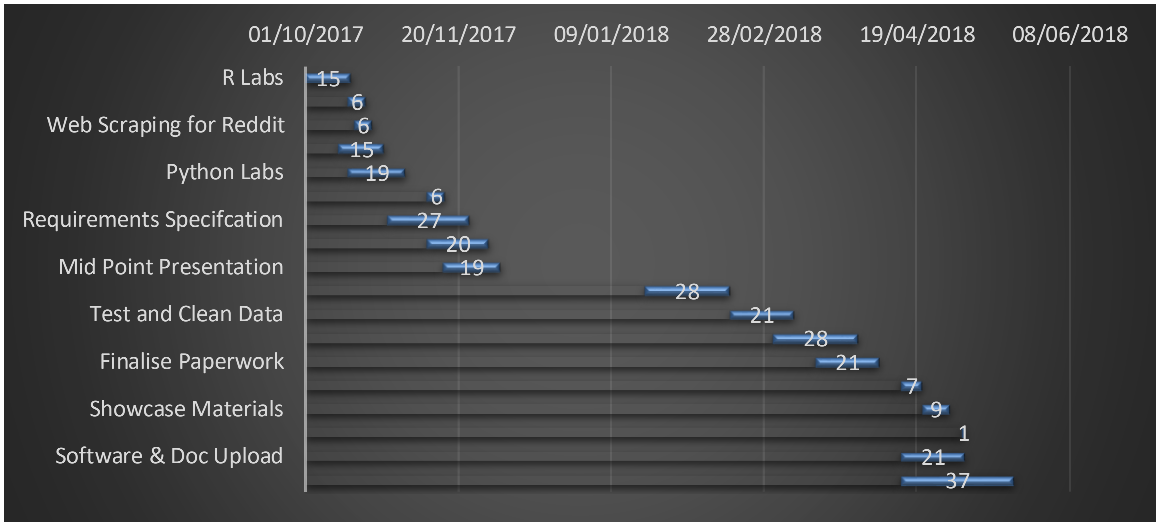

Figure 45: Gantt chart project plan

9.6 *Monthly Journals*

9.6.1 September

September was my first month back to college after my nine-month placement with an enterprise cloud data management company called Informatica. When I started my work placement back in January 2017, I did not have any idea what specialization I would pick for my final year in college - let alone a project.

I worked as a global customer support engineer during my time with Informatica. This involved dealing with databases, virtual machines and, in particular, I supported Informatics newest product which was their Cloud data integration product. I received massive insight into the world of data with Informatica and this led, in a significant way, to my decision to pick data analytics for my final year specialization. I spoke to my manager, and the data scientist at Informatica and their advice gave me great insight and reassured me on my decision to pick data analytics.

After deciding on my final year specialization, I decided to do more research on what exactly "data analytics" is composed of and brainstormed ideas for the project.

I went to the previous fourth year's project showcase day in May, which gave me an excellent overview of what technologies a lot of the projects were made up of. My primary concern, after researching data analytics projects, was how and where to get the datasets. I asked some of the final year students at the showcase and I researched on sites such as Reddit and Quora.

After researching how I would approach a project, I started learning the fundamental skills that would make up the backbone of my project. I started two courses on Udemy for RStudio and Python; I knew that there would be a module for these languages during the semester, but I thought if I could get a head start and, even more importantly, an understanding of these languages it would benefit my project.

Now onto the project idea - I always had a fondness for the quote "*Choose a Job You Love, and You Will Never Have To Work a Day in Your Life*", and with that being said I decided to do my project on the National Football League of America (NFL). More precisely my project title is 'An Analysis of How Fans React to Major Events During the NFL Season' (which has since changed!).

The NFL is the most watched sport in the United States. On average, it attracts four times more viewers than its closest rival, the National Basketball Association(NBA). I have my passion for the NFL. My brother and I started following the sport in the 90s. It became a Sunday tradition to watch the games at 6 pm. I then started playing American Football with the Dublin Dragons at the age of nineteen. The Dragons were one of only eight teams in Ireland at the time.

At this stage I have my idea for my project, we have now started our classes for the new term, and we are learning the fundamentals of what our projects will be built on. Over the first two weeks of college, we had our software project class that went through all the requirements and grading for the project. We were shown the deadlines and given examples on all of the resources needed for the project. I spent a lot of time over these two weeks finalizing my idea and preparing for my pitch which was due for week 3. I asked my lecturers' advice on my project and lookedfor critiques on my thoughts. After getting feedback, I finalized my pitch for the Monday of week 3.

My project was approved and during my pitch I received some excellent feedback relating to accessing the data and methods on how to get the data. I spent some time during the week researching these methods.

9.6.2 October

Since my project idea was approved, I commenced working on my project proposal which was a marked deliverable for my project. This week I also met with my project supervisor who would advise and assist me over the coming year. I met my supervisor Simon on the Friday of this week. During our meeting, he went through the feedback I had received from my project pitch. The feedback was mostly positive, which was good.

During my meeting, I went over my thoughts and ideas on my project with Simon. He gave me some great feedback on my project and some great additions that I could include in my project. He also helped me set a plan in motion for research on the skills and concepts I would need for my project.

Over the month I continued to work on my R programming skills as this would be the primary language I would be using to develop my project. We had a CA at the end of the month for Data Application Development which tested our R programming skills, so this was an excellent time to focus a lot of my attention on developing my skills in this area.

I also set aside some time during the month to look at past projects which were done in the Data Analytics Master's Degree. This gave me a good insight into what is done at that level and some good research ideas. I also went through a lot of research papers and articles that my supervisor Simon recommended; this also gave me a deeper understanding of Data Analytics.

9.6.3 November

During November, I spent a lot of time working on my requirements specification for my project. I continued to research my idea and similar projects that have been done in the NFL. I decided to slightly adjust my project at the end of November. I ran my new idea by my supervisor and the reasons I had for changing, and he gave me good feedback regarding this.

The change in my project idea did not complicate my work load too much. I had to research the data slightly differently, and I wanted to keep the social media aspect of my original idea for my new project. The only things I had to change slightly were the project proposal and the requirements. My idea went from a social media sentiment analysis of the NFL to an analysis of the NFL Draft and players who are selected in the draft. I felt the scope of my new idea widened significantly and gave me room to expand on my idea and develop a more complex overall project. I worked on my technical report over the last week of November and prepared my data and presentation slides for my midpoint presentation that would take place in the first week of December.

9.6.4 December

I had my midpoint presentation during the first week of December. The mid-point presentation aims to show that you have thoughtfully researched your project and started the early stages of implementating the project. For my midpoint, I showcased the initial stage of the methodology my project followed: the KDD methodology. The early stages of the knowledge discovery and data mining methodology that I showcased were the target data sets I had to choose for the data selection stage, the cleaning of the data for the prepossessing stage and some visualization for the data mining stage. I used R studio to carry out all these methods and showcased some charts and visuals of data at an exploratory level to gain more knowledge on the data for myself.

After receiving positive results from my midpoint, I spent the remainder of December finalizing projects for other modules and studying for my semester one exams in January.

9.6.5 January

During the start of January, I had my first semester exams. We started semester two at the end of January and our modules for the semester were advanced business data analysis and data and web mining. Both modules were of assistance to my project and furthered my data analytical skills. I also began weekly meetings with my supervisor, Simon. These meeting provided me with excellent opportunities to receive feedback on my project and run any queries or issues I was having with my project by Simon.

9.6.6 February

February came and went abruptly. The new modules contained advanced data analysis techniques which required a lot of study. I continued my weekly meetings with my supervisor and chipping away at my project. During the month, I got to review the related literature to my project idea.

9.6.7 March

March was the month I broke the back of my project. As the project was based on the NFL draft, I needed to wait for the NFL combine to finish for the 2018 season which took place during the start of March. The combine is a technical interview for players who intend to enter the NFL draft using 40-yard sprints and bench presses. Once the combine ended, I went about gathering the data from the event. I now had the data I needed from college statistics, NFL combine and previous NFL statistics. I set about transforming and cleaning the final dataset over the coming days. I then started the exploration of the data. Furthermore, during March I set up unit testing for my project and developed two regression-based prediction models to predict the draft. I continued my weekly meetings with my supervisor, showing him my work to date and getting guidance and feedback.

9.6.8 April

Thankfully the work I completed in March on the project paid off as April was a hectic month for my other classes. I had a group project for my data and web mining module, and at the end of April, I had my final exams ever in my degree course. The group project provided me with some goods skills I could apply to my project: while I used classification for the group project some of the packages were very useful and allowed me to see different methods I could combine into my software project.

Once I had finished my group project I had two weeks till my exams; I set about creating a timetable for the next two weeks to study for the exams and a checklist for my software project. If I could do a small bit each day, I would manage the workload, and there would be no need to cram, as they say, slow and steady wins the race! I continued my weekly meeting with my supervisor which allowed me to tie up any loose ends with regard to the marking scheme of the project.

9.6.9 May

I finished my exams at the end of April and had 16 days remaining till my project was due. I set about going over everything I had done on the project to date and writing down what was outstanding. I worked over the next two weeks on tidying up everything with the project on a functionality front and starting the write-up. The visualization software Tableau was learned to make a user-friendly dashboard for my project. I conducted anonymous in-person testing with the dashboard to receive feedback on the project and the visuals. I met with my supervisor twice over the remaining two weeks before upload. I would have my final presentation of the project at the end of May and our showcase of projects a few days after. The whole lifecycle of the project was a fantastic experience that spanned a full academic year.