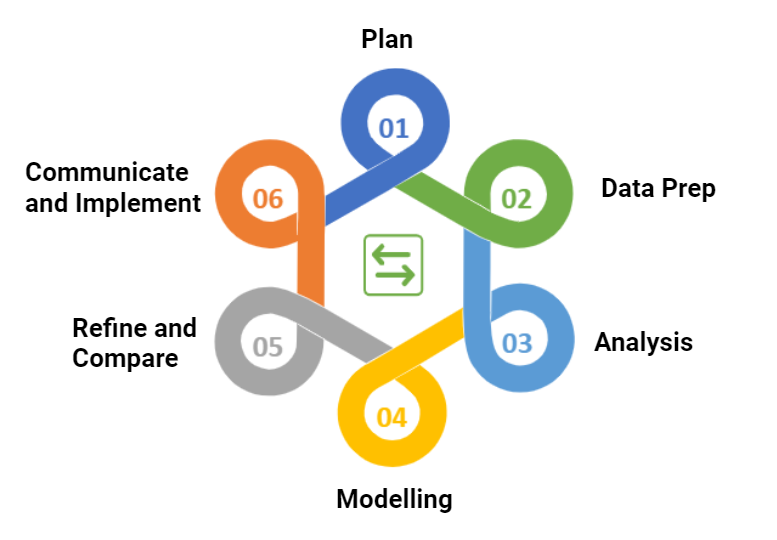
Assignment 1

*Part 1*

**Project Background**

The project I have chosen to use for this assignment is one that I did whilst at university, Predicting Football Match Outcomes Using Ridge Regression. This project investigated the use of the statistical method of ridge regression, specifically to analyse sports data and predict sport match outcomes.

**The Data Analytics Life Cycle**



**Stage 1: Plan**

Hypothesis: Ridge Regression is an effect tool in predicting sports match outcomes.

The hypothesis being investigated is a confirmatory one since this type of research has been done before, so we are just trying to confirm and develop pre-existing knowledge. This research is being done as part of my final year thesis for university and the intention behind this research is to identify if it is possible to do significantly better than the bookmakers. Previous academic papers can be used to help research. It is imagined that whilst initially it will only be a 6 months project, it will become continuous out of curiosity. Due to this there will be no plans for destruction as it will be continuously improved/updated.

Sports analytics is a collection of statistics that outline a team or individuals’ performance in any given sport. This can provide predictive performance of an individual or a team, in how well they might perform over the forthcoming season. Due to this they are commonly used by bookmakers to provide accurate betting odds. These betting odds produce information on the underlying probabilities of sporting results and are a good base level for new predictive models to compare to. Hence accuracy of the system will be measure by comparing its predictive performance against existing benchmarks as well as bookmaker's odds, and robustness can be measured by the ranking's sensitive to small changes in match outcomes.

The plan to model this research is for it to be done in R studio as there is a couple of pre-made packages there which will aid the use of the Multinomial Ridge Regression Model. A penalised extension of the multinomial logistic regression model will be used on the data sets. This will be accomplished by adding in a ridge penalty term to deal with multicollinearity in the data. The ridge penalty shrinks the coefficients of correlated predictors towards each other, hence meaning they are closer to the 'true' values. The addition of the ridge penalty will help improve predictive performance of the model due to this.

The model is expected to become more accurate over time, due to the collection of more data to test against. Hence, initially the seasons investigated will be the English Football League: Conference League 2018/19 and 2019/20 Seasons. At the time of completing this research, the 2019/20 Season was the current season, so we were using the 2018/19 season as a base model to predict the current games. All this data was collected from a website ([www.football-data.co.uk](http://www.football-data.co.uk)) as they had a wide variety or different variables from each match over a long time period (at least the past 5 years). Due to it being collected from a website, it is important that this data is checked for any inaccuracies (such as missing data or data which is known to be incorrect). This will be achieved through a different R script.

The model will include new predictor variables that are considered for every match to help improve accuracy of prediction. Come variables that could be explored are:

* Location Advantage
* Probability of a Draw
* Time Since Last Game
* Winning Streak
* Margin of Victory
* Number of Shot Opportunities
* Total Goals Conceded in the Season
* Player Injuries
* Incentive to Win
* Weather
* Number of Red and Yellow Cards in the Season

The results from this research will be presented back in predominantly a report form. A presentation outlining the method and key findings will also be given to relevant stakeholders.

**Stage 2: Data Prep**

First data needs to be collected. This is done from a website ([www.football-data.co.uk](http://www.football-data.co.uk)) and is downloaded into an Excel document (initially data is in a .xlsx). The data which is initially collected is from the English Football League: Conference League 2018/19 and 2019/20 Seasons. It includes the following variables for each match:

* Full Time Match Result
* Date
* Home Team
* Away Team
* Number of Red Cards for Each Team
* Number of Yellow Cards for Each Team
* Bet365 Home, Draw and Away Odds
* VC Bet Home, Draw and Away Odds
* William Hill Home, Draw and Away Odds
* PS Betting Home, Draw and Away Odds

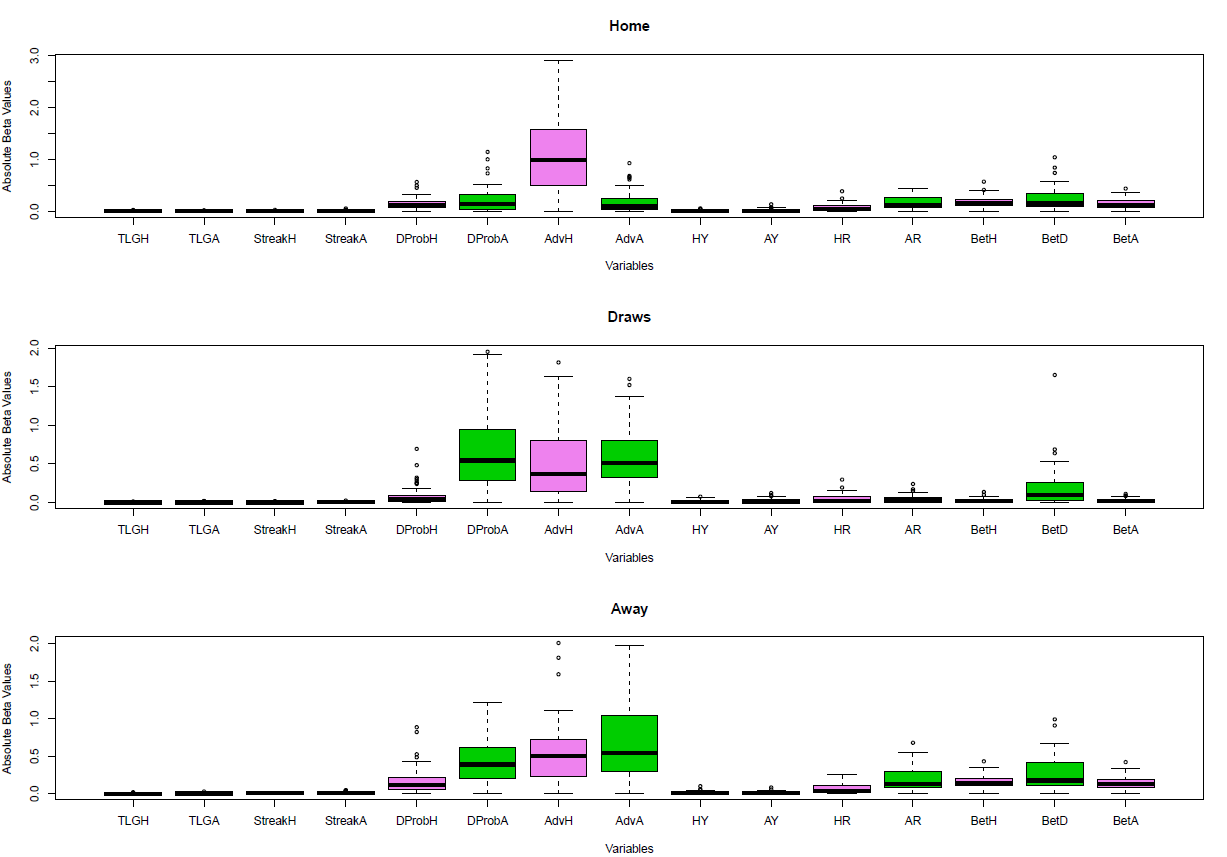
This data is then to be quickly checked over manually to make sure that all relevant data has been collected and downloaded has been completed correctly and successfully. Each column of information is made sure to be given a relevant name that will be easily identifiable. The Excel spreadsheet is then transformed into a .csv file to make sure it is easily connected to R Studio and saved on my local hard drive.

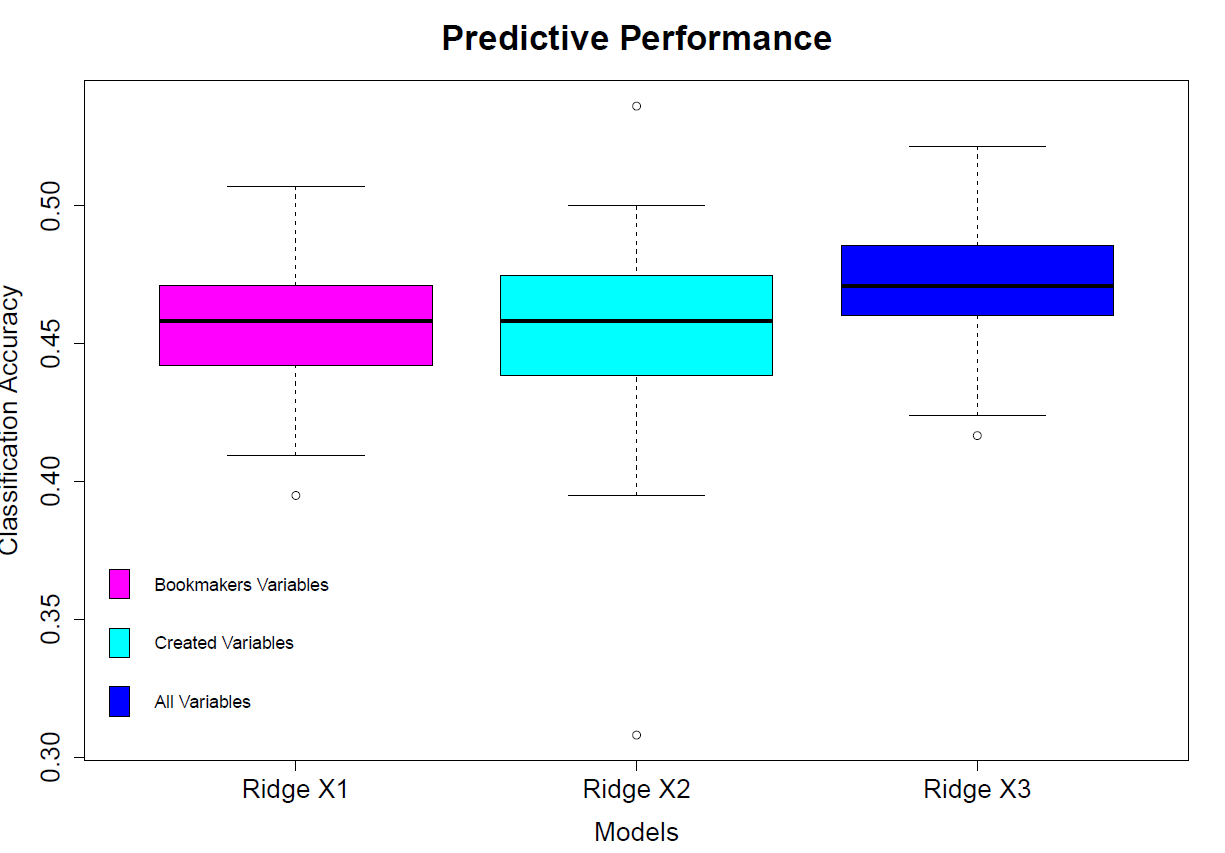
Due to the being collected from a website, it is important that this data is checked for any inaccuracies (such as missing data or data which is known to be incorrect). This will be achieved using an R script. The purpose of this R script will be to identify any missing values and to also make sure all fields are correctly inputted. For example, make sure team names are all spelt correctly, and the correct type of data appears in a specific field (e.g., a field which is expected to only contain a number doesn’t contain a string).

Any null values will be investigated and if it proves possible to fill them in, then they will be filled with the correct value. If they can’t be filled in, that match will be flagged, and model excluded from the dataset as to not introduce bias.

Another problem with the datasets is that teams are not consistent each year, due to promotion/relegation in the league. Therefore, it is important to exclude those teams matches from the model as to not impose bias.

**Stage 3: Analysis**



Overall, for the derived model, prediction accuracy was greater than the bookmakers model for the 2018/19 data set but was slightly lower for the 2019/20 data set. Further to this, prediction accuracy seemed to heavily rely on team advantage.

**Stage 4: Modelling**

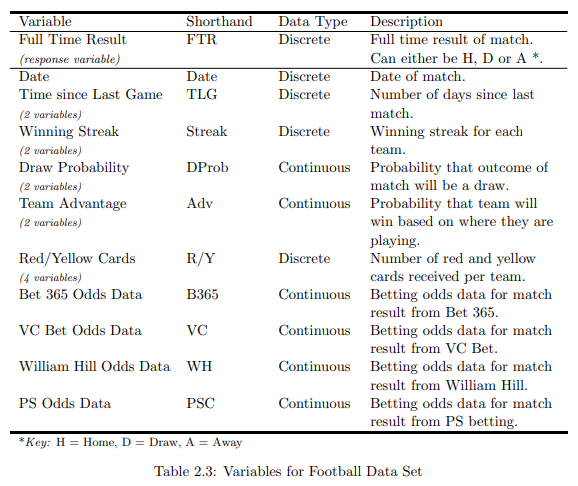
Regression analysis is a statistical tool used to model the relationship between a dependent variable and explanatory variables of a specific data set. Ridge regression is a technique that is applied when it is not possible to apply simple linear regression due to model complexity and the possibility of over-fitting, which is caused by the data suffering from multicollinearity.

As a football data set will be used, the response data of the model will be considered to be the outcome of the match, i.e., if the home team won; if it was a draw; or if the away team won. This means that the response variable is nominal, which suggests the use of multinomial logistic regression. This is a classification method that can be used to predict the probabilities of the different outcomes of the response variable, which is a useful tool for predicting sporting events.

A penalised extension of the multinomial logistic regression model will be used on the data. This is because there is no closed form solution to the multinomial logistic regression model, which is accomplished by adding in a ridge penalty term to deal with the multicollinearity of the data. One way to accomplish this is by the implementation of the package glmnet. glmnet fits a generalized linear model via penalized maximum likelihood. It is computed for a grid of values for the regularization parameter lambda.

**Stage 5: Refine and Compare**

The refinement stage saw the creation of other potentially useful variables to help improve prediction accuracy:



**Stage 6: Communicate and Implement**

The aim of this project was to create a good base prediction model for football matches that can accurately predict future results through the statistical method of ridge regression. This project demonstrated that ridge regression can act as a good base prediction model for sporting events and that the result is mainly due to the key variables, team advantage and draw probability.

**Key Takeaways**

In conclusion, the model performed reasonably well compared to the bookmakers model. In general, it was slightly better. Investigating specific factors could help to boost prediction accuracy for football matches. However, doing significantly better than the bookmakers on football matches does not seem capable due to the large amount of data the bookmakers have access to. To do significantly better than the bookmakers it would be worthwhile looking at sports which are less frequently bet on, such as Netball and Climbing. This is because the bookmakers spend less time collecting data for these sporting events as it is not a large source of their income, so a new predictive model has the potential to do significantly better than the bookmakers.

*Part 2*

**Email**

To: DXC Leadership Team

From: Olivia Watson

Subject: New DXC Social Engagement App

Attachments: Project Brief

Dear Leadership Team,

With this move to Virtual First something that I think would be worth exploring is creating a space, say an app, where you could go to interact with colleagues. Specifically, I was thinking an app where you could go to find colleagues in your local area to meet for a coffee or something. You would input your name, your company email, a rough location of where you are (your city/town for example, not your actual postcode) and provide a travel radius. Then you could search for people in the same area who might want to meet up.

I believe this would allow for some face-to-face social interaction that normally an office would provide and provide a networking opportunity outside your immediate work group.

Please look at the attached Project Brief for further information about this.

Many thanks,

Liv

**Olivia Watson**  
UKI Graduate Analytics  
  
M +44 7586 449631  
[olivia.watson@dxc.com](mailto:olivia.watson@dxc.com)  
  
**DXC Technology**

[dxc.technology](http://www.dxc.technology/) / [Twitter](https://twitter.com/dxctechnology) / [Facebook](https://www.facebook.com/DXCTechnology) / [LinkedIn](https://www.linkedin.com/company/dxctechnology)

Annual Leave: None

**Project Brief**

**Situation**

The New DXC 'Virtual First' Transformation initiative is going to have a direct impact on all our working lives, from the way we conduct business, collaborate with one another, the impact on our home working environment and individually as human beings. It is therefore important that we look to design, develop, and make available a mobile application that keeps all colleagues in touch with each other, to reduce the sense of isolation felt and maintain high levels of mental health and well-being.

Initial data gathered from apprentices and grads suggests that this move to Virtual First has made people feel more isolated, they are finding it harder to collaborate with colleagues and feel like it’s all work and no play.

The aim is to design and create a truly engaging app that all employees will benefit from and is agile and nimble to be 'upgraded' as and when the need arises.

**Relevance**

The current Global Pandemic has seen the majority of companies shutting their office doors and transferring their employees to working from home. Due to the amount of time that has been spent working from home and identifying that employees work productivity has gone up since this change, DXC made the choice to go Virtual First. The idea behind this being that DXC would be the industry front runners for this. However, being the industry front runner means we need to pave the way and create things to support employees with this change. This Social Engagement App will help to provide a smooth transition to Virtual First.

An app to combat social isolation, where you could go to find colleagues in your local area to meet for a coffee or something. You would input your name, your company email, a rough location of where you are (your city/town for example, not your actual postcode) and provide a travel radius. Then you could search for people in the same area who might want to meet up.

This type of app is very advantageous as it will be helping DXC employees to grow their professional networks by connecting colleagues and creating a community feel. It means people can move anywhere they want in the country and still be able to find a network of colleagues without travelling into an office. It will also help to improve employee wellbeing by combatting isolation and introduce and encourage a virtual social aspect to work. Furthermore, data is showing that young professionals (a few older) are not even considering applying to work at DXC due to it having gone Virtual First. It is suggested from the data that this is due to lack of social opportuninites from a lack of office space. This app should help to combat that.

This type of app is a platform for the future. This app could be developed to eventually allow for collaboration between businesses. Multiple businesses could end up on this platform and employees individual networks could grow and this might see the expansion of clients being brought in to work with DXC.

**Use Cases**

Use Case 1: Sketch

Outline initial ideas of what the app should look like. Generate a wide range of ideas and designs to see what people would want. Then consolidate ideas into a few good full suggestions to complete market research on.

Use Case 2: Market Research

Take ideas to fellow employees in DXC to see if this is something they would want, need, use and if they have any other feedback. Consolidate all this data and analyse it to identify next steps.

Use Case 3: Design

Sketch the first computer designs of the app, and gain feedback from others to see what areas need improvement.

Use Case 4: Development

Write the code behind the app, making sure all components are there and it has been debugged correctly.

Use Case 5: Testing

Run multiple tests on the app to make sure all components are working correctly. Analyse data and then update anything that is needed.

Use Case 6: Deployment

Deploy the app and monitor analytics – downloads, feedback, crashes, bugs, etc.

**Stakeholders**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Role | Responsible | Accountable | Consulted | Informed |
| Lee Joy | Talent Management | ✓ | ✓ | ✓ | ✓ |
| Astghik Stepanyan | People Manager Early Careers |  | ✓ | ✓ | ✓ |
| Cameron Shields | Technical Graduate |  |  | ✓ | ✓ |
| Justin Fallon | Head of Virtual First Transition |  |  | ✓ | ✓ |

**Challenges**

There are multiple challenges that project like this must combat. For example, the time frame to get this completed by is influenced by a multitude of factors – e.g., size of team, specification requirements, input from stakeholders, etc. Another is the technology requirements for the app – what platforms is it needed on, due to it being DXC are there any specific things that are needed to be included (say multi-factor authentication), etc. Furthermore, since the app will be storing personal information, it is imperative that GDPR rules are followed, and data security is incredibly high. An additional challenge is COVID 19. This app is designed to be a social interaction app, allowing people to meet up with fellow colleagues. However, current lockdown rules forbid this from happening so it needs to be investigated and considered the effects of an app like this if we were to go into another lockdown.