Premier League predictive learning algorithm (PLePA)

|  |  |
| --- | --- |
| **Contents** | |
| 1 | Project title |
| 1 | Glossary |
| 1 | Section 1 Preparation and planning |
| 9 | Section 2 Project work to date |
| 11 | Section 3 Review and reflection |
| 12 | Bibliography |
| 13 | Appendices |

|  |  |  |
| --- | --- | --- |
| Appendix | Name | Description |
| 1 | Project journal | A log of all the work done, broken down by each day. |
| 2 | Database choice | A small text describing why MySQL was chosen as the database choice. |
| 3 | Assumptions made | A start on assumptions made for the project. |
| 4 | Tutor conversation | Full log of conversation with tutor, this only covers emails. It doesn’t cover conversation during tutorials. |

|  |  |  |
| --- | --- | --- |
| Figure | Name | Description |
| 1 | Structured-case life cycle | An image created to show the structured-case life cycle which will be used by PLePA |
| 2 | Project schedule | This is a screenshot taken of PLePA’s schedule. |
| 3 | Database structure | The structure of the database tables and relationships between tables |

|  |  |  |
| --- | --- | --- |
| Tables | Name | Description |
| 1 | Evaluation criteria | The evaluation criteria for PLePA to know if it has been successful. |
| 2 | Resources list | Resources that will be required for PLePA. |
| 3 | Skills required | Skills identified as being required for PLePA. |
| 4 | Risk managements | Risk assessment for potential risks to PLePA. |
| 5 | Literature review | The review of literature used so far for PLePA. |

**Glossary**

Premier League – The highest league in English football.

**Section 1 Preparation and planning**

**1.1 Project description**

The project is to create a predictive algorithm for Premier League results. It will use historic data; it will use the last three years’ worth of data. This may be extended or reduced but that will depend on the ease of gathering and cleansing the data.

The result is aimed at Premier League fans and people interested in the prediction of football results. It’s also something I have an interest in solving, I do a bit of sports betting so this could prove yourself for this.

The aim is to produce four separate algorithms and determine which has the most accurate results. The goal is to have around a 75-80% pass rate, this will be a benefit to all football fans, fans who place bets and even possibly teams, that may be a bit far though.

If the project is unable to give good results after many iterations for each algorithm and tweaking the numbers, then it will serve as research for people who take on a similar project. There is no real issue if it not developed because it is more of a personal project and I am the only stakeholder.

There are different ways to achieve the results, the aim is to develop four separate predictive algorithms, this will give the opportunity to move on if one is not going well.

The stats data will be stored in a MySQL database and all the algorithm code will be written in Python. All code and documents will be stored in a GIT repository. A stretch goal would be to have an interface for user interaction or to pull the latest set of fixtures from a football website.

**1.2 Activities, tasks and subtasks**

* Define the goals and contents of my project.
* Research SDLC choices and decide on one for the project.
* Research difference between Oracle and MySQL and decide between the two.
* Think about how the database will be structured
* Look into Python modules which may be useful.
  + Distance calculation
* Set up base Python project with GIT version control
* Install database software
* Investigate similar studies for ideas. Document key findings.
* Find the best source for the Premier League statistics required.
* Investigate Machine Learning, ML, predictive algorithms and decide on 4 possibilities.
* Write TMA01
* Write TMA02
* Write TMA03
* Write EMA
* Reflect on progress to date, what went well/bad (Needs to be done multiple times in the project.
* Revaluate project after each TMA given feedback from tutor for TMA. Make sure project still makes sense.
* Conceptual Framework (CF) 1 – setting up the database and data
  + Gather data required.
  + Cleanse data.
  + Insert data into database.
* CF2-5 – work on the four algorithms, each one is a separate CF.
  + Plan how algorithm will work.
  + Do some more research on top of what has already been done.
  + Code the algorithm
  + Test the findings
  + Evaluate
  + Produce graphs and report to show successfulness of the algorithms
* CF6 – develop a feed from a sports website to pull in the latest fixtures
  + This will feed the predictions for the upcoming fixtures and display on screen or email to a user.
* CF7 – develop a user interface
  + This will allow users to select two clubs and will display the predicted results
  + This could be developed further to allow the user to select which algorithm to predict with.

**1.3 Life cycle choice**

The life cycle chosen to use is a structured-case life cycle; the reasoning behind the decision is below. Initially the iterative waterfall was chosen but after some discussion with my tutor I chose to move towards structured-case life cycle. I was not aware of the structured-case life cycle but after looking into it, I found it seemed perfect for my project.

The four main parts of a life cycle are analysis, design, implementation and evaluation.

***Classic* Waterfall** (The Open University, 2020)

A classic waterfall is a life cycle which follows an order and does not revisit previous parts of the life cycle. It follows analysis -> design -> implementation -> design

Benefits  
The main benefit of the waterfall method is that it is harder to deviate off track because it is a flow through one cycle. You do each task in a linear fashion, i.e. you do the analysis, then the design, then implement your design and evaluate at the end.

Disadvantages  
The big disadvantage is also because of the linear fashion of the life cycle, you do not revisit any from before. If there is an oversight which is discovered at the end, using this life cycle, you would not revisit and improve.

Potential usefulness in project  
This would be useful for keeping the project on track since it has a set deadline which is the length of the module, however, by not allowing iteration this would not work for my project.

Accept/reject  
Reject

***Iterative* Waterfall** (The Open University, 2020)

An iterative waterfall life cycle is like the classic waterfall but iterates over and over each part.

Benefits  
The biggest benefit of an iterative waterfall is that each of the four main parts are revisited over and over and can be revisited from any part, e.g. when you are at the evaluation stage of the project, you can return to the analysis, beginning, stage.

Disadvantages  
A disadvantage is that this can mean you might keep revisiting earlier stages of the life cycle and not finish in the timeline given.

Potential usefulness in project  
This would be useful for the project if the time is managed using a schedule to ensure that the targets are met.

Accept/reject  
Accept, but later declined based on feedback from tutor. It was suggested to investigate a Structured-case life cycle which was ideal for this project.

**Agile** (Victor Osetskyi, 2017)

An agile life cycle sets out work for a set period, usually two weeks, called a sprint. The progress made from each sprint is then factored in the following sprint.

Benefits  
With agile development, you may have long term goals, but you are mainly thinking in terms of every two weeks. You set yourself goals for each two-week period. This is very good for keeping you on track short term as you’re not allowed to deviate from the goals you’ve set yourself for those two weeks. If you surpass your goals, you may have work to do from the backlog.

Disadvantages  
A disadvantage for this project as it is very for collaborative work because you will all collaborate on how things have gone and review each sprint. It may be difficult for me to think in my mind after every two weeks what went well and what didn’t.

Potential usefulness in project  
This could be useful for the project, but I don’t think it would get all the benefits of an agile life cycle given that it’s a solo project. There is also a set deadline so working in sets of two weeks may put the project behind and wouldn’t be realised until too late.

Accept/reject  
Reject

**Structured-case** (J.M.Carroll et al., 2000)

A structured splits multiple parts of work into 4 phases: Plan, collected data, analyse and reflect. These are known as conceptual frameworks and each one is referred to as CF1, CF2, CF3 etc.

Benefits  
This organises each chunk of work into its own section which can be seen below in figure 1 and will be worked through chronologically which means you will not get distracted by starting to work on CF4 if you’re still on CF2 as an example.

Disadvantages  
The tasks which are later in the life cycle may never get touched because of the chronological ordering. If you have ten CFs but only manage to do eight, the final two will never get looked at.

Potential usefulness in project  
This will be very useful for the project because there will be three to four algorithms which will be worked through but each bit of work is separate so if only three are managed then that’s ok because the other work will not be affected by having not looked at the fourth algorithm.

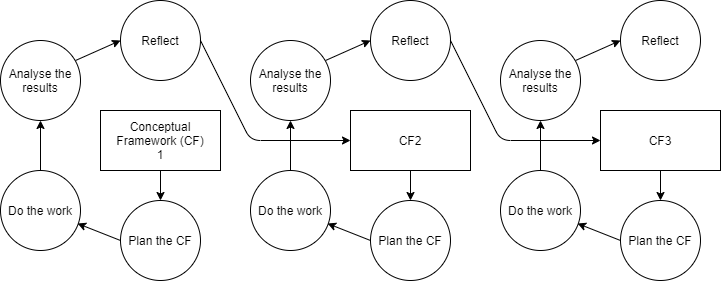
Accept/reject  
Accept  
 

Figure 1 (Blagg, 2020)

**1.4 Schedule**

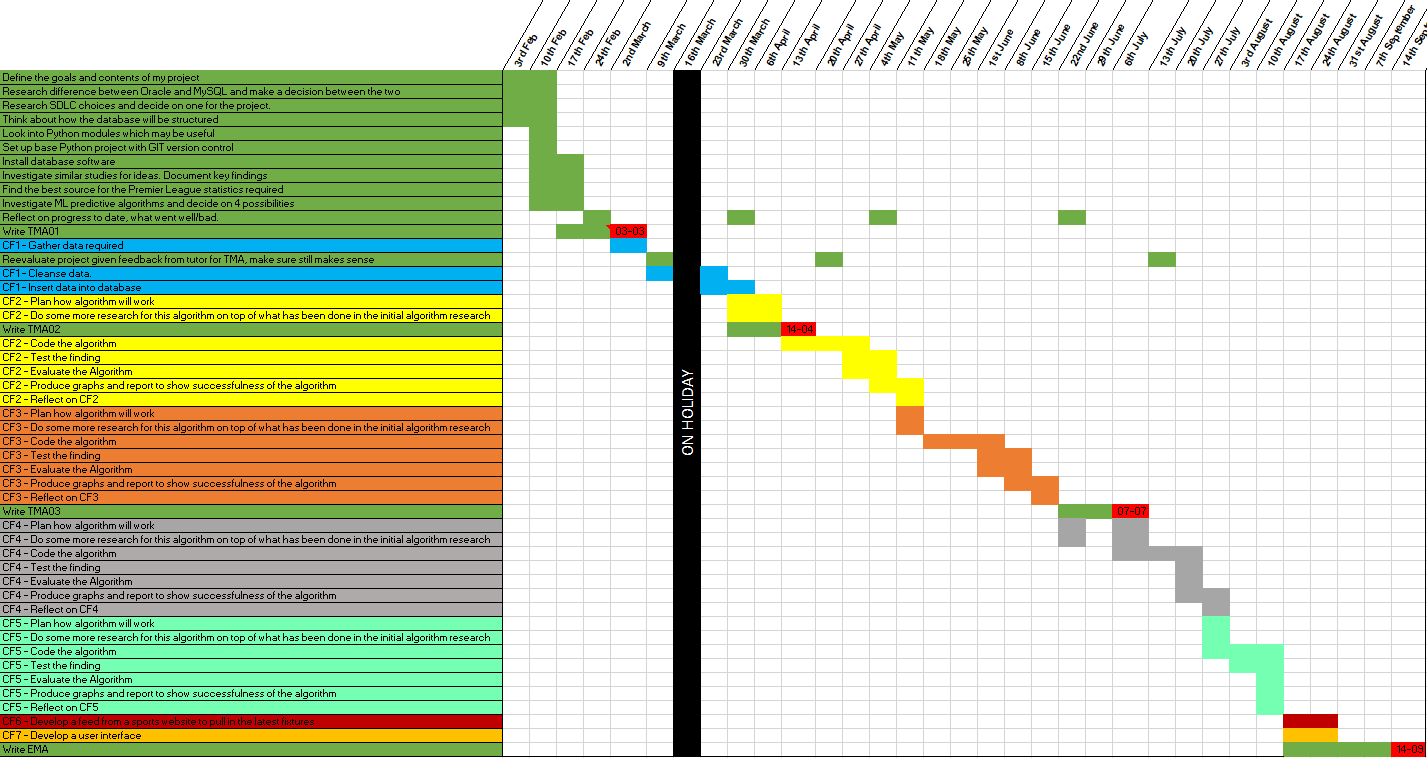
****

Figure 2 (Blagg, 2020)

**1.5 How will it be evaluated**

|  |  |  |  |
| --- | --- | --- | --- |
| Activity | Criteria | Evaluation | Progress |
| Premier league stats in the database. | Must have. | There will be a year’s worth of data in the database. | Decided on database design. |
| Should have. | Three years. |
| Nice to have. | Five years. |
| A working predictive algorithm to predict the results. | Must have. | Two algorithms will have been developed and can be used. | Decided on the four algorithms to be used. |
| Should have. | Three algorithms. |
| Nice to have. | Four algorithms. |
| An interface for user interaction. | Nice to have. | A user can load up an executable and select two clubs, they will then receive a prediction. | No progress. |
| Nice to have. | Further development could mean the user could select the algorithm to predict with and the results using the teams and the algorithm are displayed. | No progress. |
| A feed from a football website. | Nice to have. | The upcoming fixtures are pulled from a website and predicted using the chosen algorithm. | No progress. |

Table 1

**1.6 What will be needed to achieve end goals**

Resource list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resource | Why needed | When needed | Problems if not available | How to ensure availability |
| Time | Complete any task for the project. | Throughout project. | The less time available, the less tasks can be completed. | Cannot ensure time but can try to get ahead on tasks in case of any complications. |
| PC | To complete all tasks, requires pc availability. | Throughout project. | Whilst no PC is available, most tasks will be unable to be achieved. | I have a desktop and laptop. As a last resort, I can also use my work laptop. |
| Python | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | It is installed on all PCs available. |
| MySQL | Store the data used for algorithms. | Set up in CF1 but will be required from CF1 – CF7. | The algorithms will have no database to read from. | It is installed on all PCs available. |
| Premier League data | The project is using data from the Premier League for the algorithms. | Pre CF1, it will then be stored in the MySQL database. | No data to be used for the algorithms. | Once gathered and stored in the database, store a copy of the database. |

Table 2

Skill list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Skill | Why needed | When needed | Problems if not available | How to ensure availability |
| Python | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | Research unused additional Python modules which are required. |
| SQL | Write the code for the database. | During CF1. | Unable to create the database structure and insert the data. | Used in everyday work, fluent in SQL. |
| Time management | Ensure milestones are met. | Throughout project. | Milestones may be missed; project will be incomplete. | Research good time management ideas and seek advice from people with good time management. |
| Report writing | To write TMA1-3 and EMA. | For all assignments but mainly the EMA. | The quality of writing for TMAs and EMAs will be lacking. | Research and check previous module’s advice on report writing. |
| Researching | For several tasks to investigate best approach | Throughout project. | Project will suffer due to decisions being made without right information. | Look into ideas for researching and the best approaches to research. |

Table 3  
Risk management list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk involved | Priority – low to high | Overcome/manage risk | What to do if not managed | Review of risk |
| Planning on a house move. | Medium | Get ahead on project when possible, manage time during move. | Make up the time after the house move is complete. | Accept – try to manage time accordingly around busy times for house move. |
| Work travel, 2-4 days a month. | Low | Manage workload accordingly around work travel to get ahead. | Try to find a bit of extra time following the travel to catch up on project. | Accept – try to get ahead before work travel and catch up if needed afterwards. |
| Desktop and laptop failure | Low | Data stored at GitHub and locally on both machines. | Library work or purchase new laptop/desktop to replace. | Avoid – Unlikely that both desktop and laptop will fail. |
| Knowledge gap on new Python modules | High | Research the modules required, also have a Python expert at my work who I can consult with. | Lower the detail of the algorithms used. | Accept – start researching required modules early and learn how to use. |

Table 4

**Section 2 Project work to date**

**2.1 Literature review**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | Presentation | Relevance | Objectivity | Method | Provenance | Timeliness |
| TMA351 part 1-26 | Very easy to follow, the material was split into 26 good chunks. | It is mostly relevant because it uses Python to design two predictive algorithms, however, it uses MongoDB rather than a MySQL database. | There are two algorithms used for the reader to compare. | It is clear where the information has come from. | The author is reliable because it is material directly from an OU module. | The material has been updated in the last three years. |
| Machine Learning Algorithms in Python (Dataflair Team, September 2018 | Very nice flow to the information presented. Easy to understand. | Very relevant as it is about ML algorithms in Python, however, not all are predictive. | The purpose of the literature is to give information on several different algorithms so there are no conclusions made by the author. | The information uses direct code snippets of Python to further explain the information. | The author is reliable, I have used them for multiple work projects in the past when learning new material. | Material is just under a year and half old. |
| Structured-case: A methodological framework for building theory in information systems (J.M.Carroll et al., 2000) | It had a clear layout, but the wall of text made it difficult to read. | It was useful for researching the structured-case life cycle. | Clearly gives the strengths and weaknesses of the life cycle. | Clear where the research has taken place. | The two authors are both university professors. | Material is twenty years old but the framework is still used today. |
| Choosing a Lifecycle Model (The Open University, 2020) | Nicely presented, easy to read. | Very useful for looking at life cycles however the one I ended up using wasn’t given. | Multiple life cycles with pros and cons given. | Clear where information has come from. | OU is a reliable source. | Material presented this year by the OU so assumed it is up to date. |
| SDLC Models Explained: Agile, Waterfall, V-Shaped, Iterative, Spiral (Osetskyi 2017) | Very good layout, broken up by pictures, text and tables. | Useful for doing additional researches on SDLCs. | All life cycles are presented with advantages and disadvantages. | The information has come from Existek, a full-cycle software development company. | The author works for Existek and his written other articles I have read which have been informative. | Written about two and half years ago. |

Table 5

**2.2 Work done**

The name of the project has been determined as Premier League Predictive Algorithm (PLePA). The goals, stretch goals and the contents of the project has been decided upon, this changed a few times due to not being certain on the end goal that was wanted.   
The activities, tasks and subtasks have been confirmed and a schedule has been created with assignment due dates highlighted to keep the project on track.  
Some research was carried out on the differences between Oracle and MySQL to determine which was best for the project, it was mainly discussion with colleagues and experience already obtained. It was determined that MySQL was a better fit to PLePA because it is lightweight and quick to set up compared with Oracle.  
As presented in section 1.3, four life cycle choices were researched and a decision was made on the most appropriate for PLePA.  
The database software has been decided on (See appendix 2)  
The design of the database has been decided on, this has been improved upon multiple times after useful feedback provided from my tutor. The final version can be seen below in figure 3.

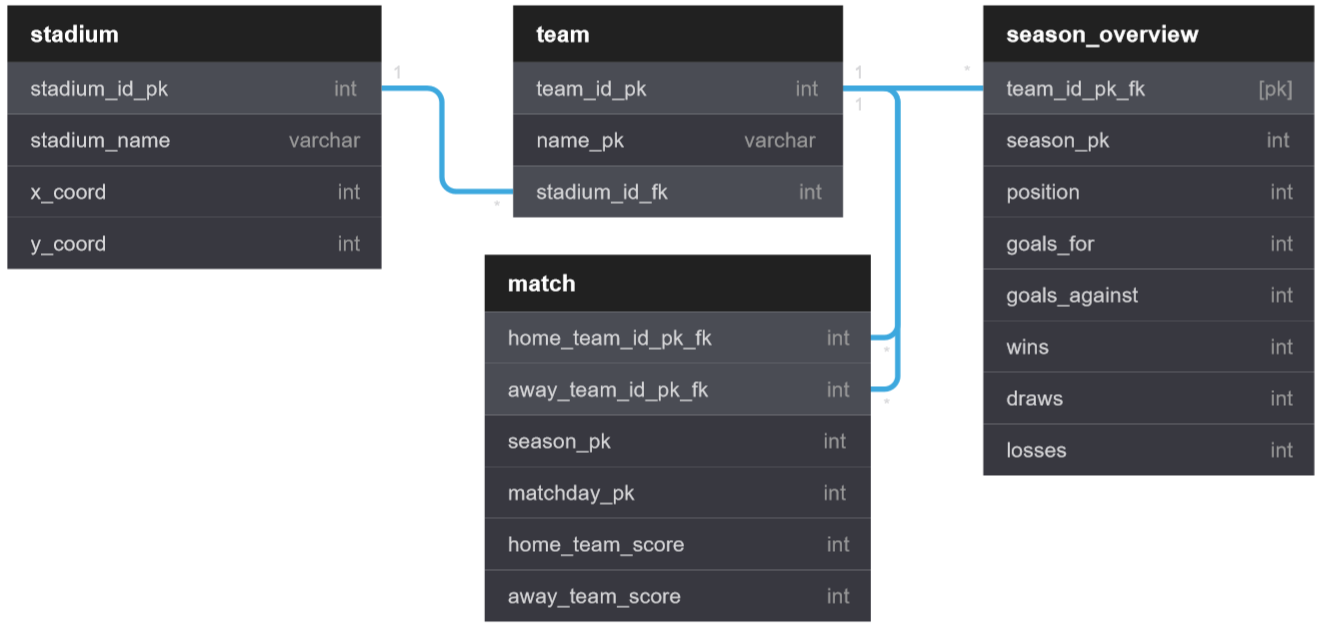


Figure 3 (Blagg, 2020)

A GIT repository has been set up for storage of all files, this is useful when working between a laptop when travelling and a desktop at home.  
MySQL has been installed and a test database has been set up and tested.  
Five predictive algorithms have been researched and the four which are going to be used for the project have been decided on.  
A task has been created for looking into Python modules which could be used, however, this proved difficult since it is uncertain yet what modules will be needed for the project. Time could be wasted if modules which are researched are not even considered.  
The location for where the stats data will be collected for the matches (Football-Data, February 2020) and past league finishes (FBREF, February 2020) has been established.  
A couple of similar project were investigated and gave some good insight and ideas for PLePA, it was useful to see the layout and structure to their report. It was also helpful to see the different paths taken for similar projects.  
The project journal has been kept up to date (See appendix 1)  
There has all been a start, albeit, very small on assumptions to be made. (See appendix 3)

**Section 3 Review and reflection**

**3.1 What has gone well**

The communication with the tutor has been very good, she has helped a lot with giving feedback on all parts of the project which have been provided (See appendix 4). Keeping to time for tasks has been mostly good and there have not be many real-life distractions so far. The research of similar projects and algorithms was very useful and proved to give good insight into what PLePA might look like.

**3.1 What has not been so good**

There was some confusion on what the conceptual model should look like and how the database should be structured, but after discussion with my tutor, this has now been resolved. There were a couple of tasks which were finished behind schedule.

**3.3 What has not been so good**

The database structure was rushed into so the beginning stage, the conceptual model, was skipped and relationships between columns were established first. The tasks got behind schedule due to time dedication to the TMA to ensure completion by the deadline date. This was because I am away from 28th February to 1st March.

**3.4 How and why have these affected work, plan and progress**

The overall work has not been affected because I am now up to date with all tasks, however, they were not done as initially scheduled. The plan and progress is now fine going into the work after TMA01.

**3.5 Unable to comment on**

It is hard to comment on whether the SDLC is currently suitable for the project due to not really utilising the CFs, this will come into play after TMA01 has been submitted. I’m optimistic that the strategy for completing the project will work, it allows for some leeway if unable to complete some CFs.

(3178 Words)

**Bibliography**

Dataflair Team (September 2018) *8 Machine Learning Algorithms in Python* [Online]. Available at <https://data-flair.training/blogs/machine-learning-algorithms-in-python/> (Accessed 16th February 2020)

J.M.Carroll, P.A.Swatman (2000) *Structured Case: A Methodological Framework for Building Theory in Information Systems Research*. Available at <https://pdfs.semanticscholar.org/caff/950b0efc8f23e17e7de171d8836c8dfa1044.pdf>(Accessed 17th February 2020)

The Open University (2020) *Choosing a Lifecyle Model*, *TM470* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1557365>(Accessed 2nd February 2020)

Victor Osetskyi (2017) *SDLC Models Explained: Agile, Waterfall, V-Shaped, Iterative, Spiral* [Online]. Available at <https://medium.com/existek/sdlc-models-explained-agile-waterfall-v-shaped-iterative-spiral-e3f012f390c5>(Accessed 14th February 2020)

The Open University (2019) ‘Data mining I: classification tasks’ *The k-nearest neighbours algorithm, TMA351* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1349974&section=3> (Accessed 24th February 2020)

The Open University (2019) ‘Data mining II: clustering tasks’ *Partitional clustering: the k-means algorithm, TM351* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1349975&section=2.2> (Accessed 24th February 2020)

Football-Data (February 2020) *Data-files: England* [Online]. Available at <http://www.football-data.co.uk/englandm.php> (Accessed 24th February)

FBREF (February 2020) *2019-2020 Premier League Stats* [Online]. Available at <https://fbref.com/en/comps/9/Premier-League-Stats> (Accessed 24th February)

Tuan Doan Nguyen (August 2018) *The Beautiful Game: Predicting the Premier League with a random model* [Online]. Available at <https://towardsdatascience.com/o-jogo-bonito-predicting-the-premier-league-with-a-random-model-1b02fa3a7e5a> (Accessed 24th February)

Nick Campanelli (May 2019) *Betting on the English Premier League* [Online]. Available at <https://towardsdatascience.com/betting-on-the-english-premier-league-making-money-with-machine-learning-fb6938760c64> (Accessed 24th February)

**Appendices**

**Appendix 1 – Project Journal**

**5th Jan**

Set up GitHub repository for storing the project. This will also allow me to work between laptop and desktop.

**1st Feb**

Drew up the first draft of database structure. This involved thinking about the table structure, but I have also written down things to think about where there are multiple possibilities.

**2nd Feb**

Created the first draft of my schedule.xlsx.

**6th Feb**

Created project journal.docx and added all previous completed tasks, as above.   
Tutorial which explained the project scope and gave a general overview of what is required.  
Added several blank docx files to serve as reminders for tasks that needed doing – assumptions made, database choice, database structure thoughts, lifecycle options, project journal, research data to be used , similar projects, tutor conversation  
Wrote up my project journal, wrote a bit on database choice and documented tutor conversation so far

**8th Feb**

Tutorial in the morning, this was useful to understand how to reflect on the project well and the requirements for documenting literature used.  
Created the database design in a pdf file.  
Installed MySQL on my desktop and laptop.  
Added a blank document for research algorithms.

**9th Feb**

Updated my project journal with entries from 6th and 8th Feb.  
Briefly evaluated some lifecycle options.  
Updates to my schedule to try and be more specific and less generic to TMA.  
Emailed tutor my progress.

**10th Feb**

Created a conceptual model and wrote up my ideas around the conceptual model.  
Updated my project journal with entries from 9th and 10th Feb.

**12th Feb**

Updated my lifecycle choices to have a bit more detail on what I know about the life cycles and why I chose the life cycle I did.  
Updated my schedule considering my tutor’s comments. It is now more specific to my task. May need a further revision once I identify the specific code I need to write for my project.  
Updated my project journal with the entry for 12th Feb.

**15th Feb**

Wrote up my first draft of my task list and project description.  
Tutorial in the morning.  
Updated my project journal with the entry for 15th Feb.

**16th Feb**

Created my references sheet and added first reference.  
Updated conceptual model based on feedback from tutorial.  
Updated schedule further  
Started my research on some algorithms and wrote up some descriptions.  
Updated my project journal with the entry for 16th Feb.

**17th Feb**

Updated my SDLC choice to a structured-case based on feedback from tutor which triggered my own research since I had not heard of that SDLC before.  
Updated my task list and schedule to reflect this new SDLC.  
Added some references to literature I looked at so far.  
Updated my project journal with the entry for 17th Feb.

**18th Feb**

Updated database design based on tutor’s comments.

**20th Feb**

Started work on TMA01.

**22nd Feb**

More work on TMA01.

**23rd Feb**

More work on TMA01 – completed section 1 and 2.1.  
Updated schedule with review section for each CF.  
Added life cycle image.

**24th Feb**

More work on TMA01.  
Researched into where I can get the data required for the database.  
Looked back on TM351 and did some research on KNN and K-Means.  
Updated journal for 18th Feb – 24th Feb.

**Appendix 2 – Database choice**

Database choice

The three options I had in mind were MongoDB, Oracle and MySQL since I have used these. I studied MongoDB as a NoSQL database in module TM351 and I did not find the syntax as intuitive and as easy to use as SQL, so I disregarded this option. The benefits of being able to import lots of JSON data quickly didn’t seem to be necessary for this project. This left me with a choice between Oracle and MySQL, I am a lot more experienced in Oracle since this is what I work with most in my job, however, we do use MySQL for a couple of applications so I am able to use this too. I decided to choose MySQL because it is a lot more lightweight and easier to set up than Oracle. The syntax is not too different and for the purposes of this project I think Oracle is a bit overkill.

**Appendix 3 – Assumptions made**

Assumptions made

Treat 1st place from Championship in previous season as 18th place. 2nd from Championship as 19th. Play-off winner as 20th.

**Appendix 4 – Tutor conversation**

Welcome - at last !! Ref: 209355730 5 messages j.a.tope@open.ac.uk <j.a.tope@open.ac.uk> Fri, Jan 31, 2020 at 7:16 AM To: blaggrob@gmail.com

Hi We are up and running after a few hiccups with the allocation system. I have put my first post of the course in the tutor group forum - the forum is the medium I shall a fair bit to communicate with you so please subscribe to it. I look forward to working with you this year Ju

Rob Blagg <blaggrob@gmail.com> Fri, Jan 31, 2020 at 8:34 AM To: j.a.tope@open.ac.uk

Good morning, Ju, Hope you're well. I have two project ideas, below, but I'm not sure if I'm thinking along the right lines. Would appreciate some feedback on which is best and pointers on what I could change. I'm very much looking to do a research project. Kind regards, Rob

Project 1 – Football result predictor Where the problem is based (home / work / social / etc) The problem I wanted to address is football statistics. Why it is a problem – issues Unable to see the likelihood of a team winning a game, I understand better companies must have something similar but this is not publicly available. I want to build a predictive machine learning model using historic football data with a range of stats such as form, distance travelled to the game, previous season league position etc. to predict the score of the upcoming game. What the solution should achieve – aims The solution should predict correctly the result of 75% (rough at this point) of games. It should be able to learn from the data that comes in each week and refine its algorithm. The level 3 course(s) on which your solution will be based and possibly L2 ones I want to base this on TM351, I also write a lot of python and do data investigation in my current job role so this will be useful for this project. Type of project (development / research / evaluation) The project is a research project.

Project 2 – Best gaming setting (I also have an interest in chess and thought the same could apply to that) Where the problem is based (home / work / social / etc) The problem I wanted to address is home, but also for some people it is work (eSports). Why it is a problem – issues You do not know the best environment to play video games in. This could include heart rate, time of day, whether it’s a weekend/weekday, how long after eating food, current mood (subjective) and other factors. What the solution should achieve – aims The solution would be to provide statistics on what effects gaming performance the most, what has the biggest impact to performance positively and negatively. The level 3 course(s) on which your solution will be based and possibly L2 ones I want to base this on TM351, I also write a lot of python and do data investigation in my current job role so this will be useful for this project. Type of project (development / research / evaluation) The project is a research project.

J.A.Tope <j.a.tope@open.ac.uk> Fri, Jan 31, 2020 at 10:16 AM To: Rob Blagg <blaggrob@gmail.com> Just getting over a tummy bug, thanks for asking!

Your project would seem to fall between a bit of development and a lot of 'research' so ﬁne, no probs. Just don't use the stats from my team this year as the Canaries sit rock boom - we are a great championship side but cannot aﬀord the premier league

Cheers Ju

Judith Tope AL: 00081839

From: Rob Blagg <blaggrob@gmail.com> Sent: 31 January 2020 08:34 To: J.A.Tope <j.a.tope@open.ac.uk> Subject: Re: Welcome - at last !! Ref: 209355730

Hi Ju, Glad to see you're on the better side of your tummy bug.

Yes, I didn't think of it that way but there will be plenty of development too I guess. Well, my team is Ipswich so I don't have a lot of room to talk after last season. Thanks for your feedback. Kind regards, Rob

J.A.Tope <j.a.tope@open.ac.uk> Fri, Jan 31, 2020 at 11:22 AM To: Rob Blagg <blaggrob@gmail.com> we'll meet again next year!

Cheers Ju

Judith Tope AL: 00081839

don't forget my offer ... Ref: 209500446 3 messages j.a.tope@open.ac.uk <j.a.tope@open.ac.uk> Mon, Feb 3, 2020 at 9:17 PM To: blaggrob@gmail.com

... of an introductory online tutorial into the course, project types, lifecycles , planning, resources, skills and risks Thursday 6 Feb 7pm tutor group online room Just join in but arrive a couple of minutes beforehand everyone welcome best wishes Ju -- The Open University is incorporated by Royal Charter (RC 000391), an exempt charity in England & Wales and a charity registered in Scotland (SC 038302). The Open University is authorised and regulated by the Financial Conduct Authority in relation to its secondary activity of credit broking.

Rob Blagg <blaggrob@gmail.com> Thu, Feb 6, 2020 at 8:02 PM To: "J.A.Tope" <j.a.tope@open.ac.uk>

Hi Judith, Thanks for the tutorial tonight, was very useful. Shame about the problems. Please can you send over the slides when you get a chance please? Kind regards, Rob [Quoted text hidden]

J.A.Tope <j.a.tope@open.ac.uk> Thu, Feb 6, 2020 at 8:03 PM To: Rob Blagg <blaggrob@gmail.com> recording up, slides attached, sorry for sound issues, thanks for attending

Cheers Ju

Judith Tope AL: 00081839

TM470 UPDATE ROBERT BLAGG 12 messages Rob Blagg <blaggrob@gmail.com> Fri, Feb 7, 2020 at 10:17 AM To: "J.A.Tope" <j.a.tope@open.ac.uk>

Hi Ju, I'll give you an update at the end of the weekend as I have plan to get quite a bit done by then. I just have a question. I'm using Git/GitHub for version control between my laptop and desktop. It's a lot better to use .txt files rather than .docx files with GitHub to track changes. Is it ok if my research and the write up on why I chose to do certain things are done in a .txt format rather than a .docx? Kind regards, Rob

J.A.Tope <j.a.tope@open.ac.uk> Fri, Feb 7, 2020 at 2:33 PM To: Rob Blagg <blaggrob@gmail.com> Rob

By all means use .txt in your environment BUT most level 3 modules expect doc/docx/rtf and this is partly for eTMA processing purposes but also to allow profession presentation - good presentation especially relevant for TM470 otherwise LO7 Communications will get penalised lots

it is also difficult to include tables and figures in a .txt file.

you can have a free copy of Office 365 as a student, which will also enable automatic wordcounting, spellchecking etc.

so, my answer to your question is - I would prefer that you convert your .txt to .docx for submission but I have no control over how you collate your work on your own machines

Cheers Ju

Judith Tope AL: 00081839

J.A.Tope <j.a.tope@open.ac.uk> Sat, Feb 8, 2020 at 10:52 AM To: Rob Blagg <blaggrob@gmail.com> slides and pdf as promised let me know if missing

Cheers Ju

Judith Tope AL: 00081839

2 attachments 02tut2020.pptx 559K sample style TMA01 2020.pdf 204K

Rob Blagg <blaggrob@gmail.com> Sun, Feb 9, 2020 at 4:04 PM To: "J.A.Tope" <j.a.tope@open.ac.uk>

Hi Ju, My weekly/fortnightly update. Progress made - Got a rough plan for my schedule, looked at database\_options, created my project journal and keeping it up to date, starting document assumptions made when I think of them. Set up MySQL on my desktop and laptop. I started creating blank documents that will be required. Got the first draft of my database table structure created. Decisions taken - I'm going to use MySQL for my database because it is lightweight and I'm more confident in using SQL database rather than a NoSQL database. I'm thinking about an incremental/iterative lifecycle but not finalized. What has gone well - I have made some good progress on getting some starting points for each area, the database design is coming along nicely. What has failed to go well - I have not made as much progress into looking at lifecycle choices. What you aim to do next - Per the schedule, this week I aim to finish identifying the initial goals and contents of projects, start looking into similar studies, look at where I can find the stats that are required for the Premier League, decide on a lifecycle choice, look at what learning algorithms I could use (I'm currently only familiar with KNN) Questions - Is my schedule too broad, does it need more detail? I've attached some documents, I would like to know if I'm on the right sort of track for documenting things.I have some more but didn't seem necessary to add all documents, let me know if you want them all. Kind regards, Rob [Quoted text hidden]

4 attachments

schedule.xlsx 24K project\_journal.docx 13K lifecycle\_options.docx 13K database\_design.pdf 174K

J.A.Tope <j.a.tope@open.ac.uk> Sun, Feb 9, 2020 at 4:47 PM To: Rob Blagg <blaggrob@gmail.com> Phew, lots of work done, try not to take umbrage at some of my comments here, just trying to be helpful

Journal good, keep it up

descriptions of life cycle choice a bit thin - check the example, albeit not amazing, in the slides for the areas to address for each, pros / cons / reason to accept or reject; include and acknowledge pix from OU document on lifecycles

schedule needs to show a little more relation to life cycle not to the TMAs otherwise slippages could cause havoc, TMAs are milestones not life cycle stages some of the tasks therein might be in wrong order eg in 02 you are reﬂecting before doing anything substantial; don't just throw tasks in; you will relate literature to project ALL the me not just in 03; I don't see any testing going on

database - keep table names in singular because each entry is only one stadium, one team etc etc convention can you not combine upcoming and match, null entries in score will show yet to be played I would distinguish more clearly home and away teams in the attribute names win, draw, loss should be calculable from data in match as would scores is matchday a date? you have season as a pk in two tables but no obvious link twixt the two; should there only be one season otherwise what happened with promotion and relegation, thence no need to store? no obvious indication whether a relationship is mandatory or optional you seem to have by-passed the conceptual model or is that in a separate document? sorry, I won't nit-pick any more on your database structure!

yet to develop? - chat about pros and cons of possible database software, similar to one for life cycle once that's beefed up

OK, so you have shouted expletives at the computer and my comments but I hope you can understand what I am trying to say and I hope you can appreciate I am trying to help

come back to me if I have caused more confusion than conﬁdence

Cheers Ju

Judith Tope AL: 00081839

From: Rob Blagg <blaggrob@gmail.com> Sent: 09 February 2020 16:04 To: J.A.Tope <j.a.tope@open.ac.uk> Subject: Re: TM470 UPDATE ROBERT BLAGG

Hi Ju, Thanks for the informative feedback. It is a good time to find out things that I'm not doing correctly so I can correct them before I'm too far along. Will probably give you another update at the end of this week to make sure I've understood your points correctly. Kind regards, Rob [Quoted text hidden]

J.A.Tope <j.a.tope@open.ac.uk> Mon, Feb 10, 2020 at 5:24 PM To: Rob Blagg <blaggrob@gmail.com> ok, look forward to it

Cheers Ju

Judith Tope AL: 00081839

From: Rob Blagg <blaggrob@gmail.com> Sent: 10 February 2020 14:42 [Quoted text hidden]

Hi Ju, Thanks for the slides and useful tutorial yesterday. Thought i'd give you an update. Let me know if I'm using too much of your time. I just want to make sure, whilst in the early stages, that I'm on the right track. Progress made - keeping my project journal up to date. Added my conceptual model, googling around it seems many people have a different opinion of what the conceptual model consists of, I've tried to base mine on your feedback mainly. Added a bit more detail to my lifecycle choices (will put this in tabular format for the TMA). Updated my schedule multiple times but still not sure if it's exactly how it should be. Wrote up my project description adn task list. Started researching algorithms, this is by no means complete. Decisions taken - I decided on an iterative waterfall life cycle, most other bits are research at the moment so no final decisions made. What has gone well - In general, I feel like I'm getting on well with the tasks, I think I'm happy with my choices made so far. What has failed to go well - Just the schedule, I'm not sure if I'm going in the right direction.

What you aim to do next - Research a bit more into the algorithms. Improve my schedule further, if necessary, based on your feedback. Research similar projects and where I can find the data required. Lots of research this week Questions - I would like some feedback on my schedule, task list and conceptual model, if possible, please. Kind regards, Rob [Quoted text hidden]

3 attachments database\_structure\_thoughts.docx 45K schedule.xlsx 19K task\_list.docx 15K

J.A.Tope <j.a.tope@open.ac.uk> Sun, Feb 16, 2020 at 1:38 PM To: Rob Blagg <blaggrob@gmail.com> database you seem to have gone overboard with PKs anyway here goes --

stadium only needs stadium id, the other attributes are not part of PK

team why do you need a team id anyway, surely name is unique and stadium id not necessary in conceptual model (it is FK)

match ﬁne

season-overview ﬁne

schedule you are still using TMAs as banner titles - do me a favour and remove those, leave colour coding as is are you not aiming to do any reﬂection apart from green group? you are reviewing project in dark yellow lot I admit but ... holidays? or have you nothing planned?

task list not totally consistent with schedule

I still have a minor thought that you should have a combined lifecycle of structured case and iterative where cf1 = plan, develop and test database cf2 = plan, research, code, test and evaluate algorithm 1 cf3 = alg 2 cf4 = alg 3 cf5 = alg 4 then you can write ﬁnal summary but if me rears its ugly head you can just drop cf5 but you are tackling the project in reasonable chunks AND it will have a more coherent schedule associated with it

Good luck

Cheers Ju

Judith Tope AL: 00081839

From: Rob Blagg <blaggrob@gmail.com> Sent: 16 February 2020 12:43 [Quoted text hidden]

Hi Ju, Thanks for your feedback. Database bits make sense now, I think. I will update the schedule and task list and try to keep consistent. Yes, having just looked at structured case life cycle, I agree, it is a good idea. I hadn't come across it before but it's perfect for this project. Kind regards, Rob [Quoted text hidden]

J.A.Tope <j.a.tope@open.ac.uk> Mon, Feb 17, 2020 at 9:46 AM To: Rob Blagg <blaggrob@gmail.com> phew! nice to have a student agree with me! it will mean a total revisit of task list and plan, sorry

Cheers Ju

Judith Tope AL: 00081839