**Project Design Document Template**

Divide your design document into sections that address each bullet point in paragraph or visual form,

being as comprehensive as possible. You have the liberty to reorganize these sections as makes sense,

as this generic template may not be perfect for every idea.

**Discovery**

● Idea: -

A back-testing app that attempts to model EPL results by using past results and other related predictive criteria in an attempt to record a profit with the bookmakers.

1. Identify / Clean / Structure Relevant Datasets:

<https://www.kaggle.com> EPL Results 1993-2018, CSV file with the following columns:

* Division (irrelevant)
* Match Date
* Home Team
* Away Team
* Full Time Home Goals Total
* Full Time Away Goals Total
* Full Time Result (A,H or D)
* Half Time Home Goals Total
* Half Time Away Goals Total
* Half Time Result (A,H or D)
* Season

<https://www.football-data.org> free API data re competitions, scores, fixtures and league tables

<https://www.api-football.com> API data (free id =< 50 a day) re countries, seasons, leagues, standings, teams, livescore, head2head, events, line-ups, players&coaches, transfers, odds, statistics

<http://www.football-data.co.uk/data.php> CSV files for betting and odds data

<https://wheresmykeyboard.com/2016/02/list-football-api-epl/>

* <https://datahub.io/sports-data/english-premier-league#python>
* Manager Histories, win/loss/draw % … locate
* Squad Value/Avg Player value … locate
* Player rankings … locate
* Historical betting odds for all EPL games. The following resource is specifically maintained for informing betting strategies and contains the odds from many bookmakers <http://www.football-data.co.uk/englandm.php>

1. Predictive Criteria (ideas)

Analyse the EPL Results dataset above in conjunction with the other datasets above to id the credibility of the following possible predictive criteria,

* Home advantage
* Club Size/Investment Level: Analyse player spending levels over previous seasons, eg. avg cost of a squad player
* Club Performance History: all time win %
* Recent Club Performance History: all time win % vs win % in last 5 years
* Manager Performance History: Manager career win/loss/draw record, plus more recent record (say last 2/3 seasons)
* Upcoming opponent head to head record, win %
* Team Confidence/Mentality/Form: Eg. last 5/10 games win %, avg chances created/conceded per game in last 5/10 games)
* Player Availability: Total player ranking points of available players for upcoming match, assumes access to a player ranking dataset … may be tricky to source
* Current Success Levels: finishing league position re previous season

1. Data Modelling

Identify a method of modelling to use to predict results … explore.

This is a short dive into Data Science, so I need to do a bit of research here,

1. Simulate Bets

* Either find the best odds from the betting odds dataset for each game or select one bookmaker and use their odds
* For each game if our model predicts a win, simulate a bet on that game for a to be determined amount with the selected odds, and calculate profit/loss
* Retrospectively simulate bets on previous seasons to see if a profit can be made

1. Retrospective modelling/bet simulation Vs real-world upcoming match predictions and betting strategies

The project can be viewed in 2 parts:

* data modelling followed by bet simulation on historical/static datasets to obtain fictional profits/losses.
* using current (and historical) data to inform a real-world betting strategy, run via a web-app with current data-feeds.

**1) Finalise Data**

Sourced ([www.kaggle.com](http://www.kaggle.com)) :

Date

HomeTeam

AwayTeam

FTR

FTHG

FTAG

Season

Derived:

H2H Win %

AllTime Win %

LastFiveYears Win %

LastSeasonsFinishingPosition

FormOverLastFiveGames

Avg Goals scored / conceded last 5 games

Don’t Have But Would Like – Come back to later time permitting:

Nice To Have:

ManagerCareerWin %

ManagerLastFiveYearWin %

MatchDaySquadValuation

Chances created / conceded per match

**2) Establish criteria to feed into model**

i) Home advantage

ii) Lasts season’s finishing position in league

iii) All time head to head win ratio Vs opponent

iv) All time win %

v) Last 5 Years Win %

vi) Form - points taken over last 5 games

vii) Avg goals conceded & scored last 5 games

**3) Ad-Hoc Exploratory Data Analysis**

Perform some data analysis to investigate the validity of some of the above criteria

eg. do teams gain more points playing at home versus away over a sample dataset

of say 3 seasons ?

**4) Explore Some Data Models To Achieve Highest Precision**

Explore some models to identify one that gives a decent level of precision

Try:

Binary Logistic Regression

Multinomial Logistic Regression

Random Forest Classifier

Training Set = every season from 94/95 to 16/17

Testing Set = 17/18

Target = variable with 3 classes, win, loss & draw

Load the identified criteria into the model

Summary = select the model with the highest precision

**To Do List :**

* Download data & load
* Organise data in Jupyter NB
* Start exploratory data analysis - Jupyter
* Explore diff models to see which give the best precision
* Start to identify what front-end will look like

**5) Bet Simulation**

* Source historical betting odds ([www.football-data.co.uk/englandm.php](http://www.football-data.co.uk/englandm.php))
* Simulate bets and report outcomes by using model predictions allied to the historical odds for those matches
* Did the model produce a profit ?

**Front-End:**

* Offer a back testing strategy to determine if any money would be made using a given model betting a notional amount per game for a selected season
* Provide decent graphics - investigate
* Functionality:

i) Able to change the length of training & testing sets

ii) Choose different combinations of identified criteria, is it possible to weight the criteria

iii) Possibly change the criteria, ie. form = points take over last 5 games, change to 3 games, or however many

* Adapted to predict upcoming matches

i) Need up to date feeds - APIs

● User Personas- Who is using your project? What is your target audience, and why would they bother to start using your project?

○ In describing your problem/solution, be sure to mention any industry/technical

background as well. Assume the reader only knows the code side of things.

● User Stories- walk through ALL the functionality of your project in paragraph form, as each user

would go through it (e.g. start with login, and go from there to a main menu, etc.)

○ Reference appropriate wireframes/visual mockups as needed

○ Create a separate user story for each user persona (e.g. admins and users would have

different user stories)

**Design**

● Wireframes

○ Include visual representations of every single front-end component of your application as

it would appear to the user (computer-created, not hand-drawn)

● Entity Relationship Diagram

○ Identify each of your project’s components, and how they will relate/communicate with

each other. This should be portrayed visually, with arrows showing how each of the

components are linked.

○ Include any database schema (tables, columns, foreign keys, if applicable), as well as

what sort of data storage your project is using (traditional relational database,

nonrelational database, blockchain/distributed storage, etc.).

● List the proposed technologies/languages/frameworks/libraries you plan to use for each specific

component

**Development/Deployment**

● Describe what the final project would look like (If you were developing it full-time with unlimited

time). If that is outside the scope of what you could do in Phase 3, describe what paired-down

version you plan to achieve in only 3-4 weeks.

● Describe how you plan to deploy and host your project once it has been developed.