Capstone Project Proposal

1. **What is the problem you want to solve?**

To identify the best players for a specific playing position from the European Soccer Database

1. **Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?**

The client, a famous European Football Club wants to have a list of players to consider during transfer season. In case they need to make a transfer, they will bring in a replacement. A major part of player selection depends upon which position the player will play at. Depending on the need of the team, the managers will consider the up and coming players who will suit their need. The client requires us to build an algorithm that will predict the best 20 players for each specific position

1. **What data are you going to use for this? How will you acquire this data?**

The data we will use for this problem is sourced from several websites such as

http://football-data.mx-api.enetscores.com/ : scores, lineup, team formation and events

http://www.football-data.co.uk/ : betting odds

http://sofifa.com/ : players and teams attributes from EA Sports FIFA games. FIFA series and all FIFA assets property of EA Sports; by Hugomathien and made available on Kaggle at https://www.kaggle.com/hugomathien/soccer

1. **In brief, outline your approach to solving this problem (knowing that this might change later).**

**\*\*Cleaning:\*\***

1. From the available database, we use the tables 'players' and 'player\_attributes' and identify each players preferred position of playing using their X, Y playing coordinates. We create a table of player\_positions to make merging of datasets easier

2. After merging all the available datasets, we keep only the features we need to predict the performance for a player. We use the feature 'overall\_rating' given by FIFA and we calculate the age of each player for the same rating date. This is to understand whether age of the player has impact on player performance. Alternatively, we can also collect rating for each player on the latest date available for the player and use that to correlate age and overall\_rating

3. We check this data for missing and incorrect values and clean it

**\*\*Descriptive Analytics:\*\***

1. Test the correlation of age with overall rating and fit a model to predict rating for each player position

2. Also test additional data from player attributes dataset like player strength, height, weight etc. to check for correlation with score (overall\_rating)

**\*\*Modeling:\*\***

1. For each position, check assumptions, hypotheses and test

2. Fit a suitable model on training data and test on testing data

1. **What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.**

Deliverables will be a iPython Notebook consisting of the code along with the data story and a slide deck for just the data story of the project