Capstone Proposal

For this final project I have chosen a dataset on football players from the FIFA 2019 game. This data set contains information on players (categorical data) and player attributes (numerical data). Combined they give a full description of all players in the game.

Domain Background

I chose this dataset because it is something I am passionate about. I love football and gaming. The style of this project could be assessed as it was carried out for a real life team or analysis you could perform in the game to decide what players are worth buying and which are not.

Problem Statement

The problem I will assess here is weather or not a player is worth buying from another club and what wages the should be paid/the club would be willing to pay them. This is somewhat of a 'MoneyBall' style issue but instead of assessing the worth of the player to a specific club we will do it in general.

Datasets and Inputs

The dataset used was taken from Kaggle and contain information on 18,159 players. It contains 54 fields of information, including ID and Name which are removed straight away so really 52 features for me to use.

Solution Statement

Use machine learning techniques from the supervised learning section such as regression and support vector machines to create a model to determine the Value and Wages a player should have based on their remaining features.

Benchmark Model

I will compare the results of the testing dataset against the actual figures to see how they compare. Then use at least 2 different models to compare which one better predicts the data. Select the most appropriate model.

Evaluation Metrics

Due to the continuous nature of the fields I am to predict I will use the R-Squared value of the model to evaluate it. This is similar to the Boston Housing project evaluation.

Project Design

I will bring in the data from the csv file I have saved down and give it an initial inspection. The data must then be cleaned, which from an initial look at the csv will involve finding a way around null values and transforming some of the features. On inspection I saw variables such as high saved in s string format, eg 5'11'' so this will be cast to a numeric value that can be used by our model.
I plan to perform some form of PCA on the dataset reducing the dimensionality of the data. Also, assuming I will see some outliers in the dataset such as Cristiano Ronaldo I may decide to exclude players who earn more that a certain amount.

Once the data is ready to be used the remaining steps will be used to predict wages and value. So the y-predictor variables will be value and wage, modelled separately. Then all feature variables will be the same for each of them. Then I will perform a grid search for each model to optimize the parameters. Once the models are optimized they will be used to predict the Value and the Wages they are worth.

An R-Squared evaluation will be performed on different models used to choose the best one. Then run the model on some feature variables for random players to see if the results seem reasonable.