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.

Similar Projects:

1. FIFA Player Data - https://www.kaggle.com/aishwarya1992/fifa-data-analysis-player-value-prediction
2. MoneyBall Model
    a. https://towardsdatascience.com/linear-regression-moneyball-part-1-b93b3b9f5b53
    b. https://towardsdatascience.com/linear-regression-moneyball-part-2-175a9dc72e89
3.

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. I have often thought of developing a model like this for a career mode in FIFA or for choosing my fantasy football team throughout the years and now I have the perfect reason to get it started.

Problem Statement

The problem I will assess here is what value should be placed on a player and how much should they earn. These are 2 separate problems and will need 2 models, however, since they are highly related I have decided to run them in parallel and see if I can find a model that can be used to predict both Value and Wage. 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. So we will not be attempting to assess how many extra wins a player if both would add to our club, just a model that can find a fair value fot Value and Wage.

Datasets and Inputs

I searched online for interesting datasets. Considered using google search trends but after checking out some of the most popular datasets on Kaggle I found this one and thought it seemed to be a good fit.The dataset used was taken from Kaggle and contains 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.

There will be 2 models and therefore 2 target variables. First target will be Value, second will be Wage. The same features will be used for predicting the two.

https://www.kaggle.com/karangadiya/fifa19

Solution Statement

Use machine learning techniques from the supervised learning section such as regression and support vector machines to create 2 different models that can determine the Value and Wages respectively a player should have based on their remaining features.

Benchmark Model

For our benchmark I thought using a simple average of players Value or Wage. This model will not be accurate due to high variance but will allow me to see what a lowest acceptable model evaluation should be better than. If my model can not perform better than a simple average I will know that it has no value.

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. All categorical data will be one-hot-encoded so the number of feature variables will increase a lot. 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.