Capstone Project Report

Hongbin Lin

2020/06/24

Background and Objectives:

Soccer is the most popular sport in many countries. With the development of soccer, the data is playing a much more important role and even becomes the key for a soccer team to succeed. So, in this capstone project, the goal is to analyze the data and create models that can help some national teams, clubs, some soccer relevant industries and even soccer players themselves. Based on their specific attributes, I can predict their overall rating score and help them to identify the best position that they are more likely to perform well and become successful in their further career.

Data Source:

In this capstone project, FIFA 20 complete player dataset will be used for analysis and modelling. The data is collected from the real players by EA sports which is one of the most professional sports games companies, so the dataset is reliable. This dataset is a CSV file that contains around 18,000 observations and 104 features. It organizes the player positions with the role in the club and in the national team. The player attributes are some statistics as attacking, skills, defence, mentality and so on. Player personal data is like nationality, club, date of birth, wage, salary, etc.

Data Cleaning:

The original dataset has some duplicated and useless features as well as some null values. In order to have the cleaned dataset, I firstly dropped the duplicated and useless features. Then I would fill some null values. Within this dataset, there is a huge difference in attributes between goalkeepers and other players. The goalkeepers just have the specific goal keeper's attributes. In other words, other players have all kinds of attributes except the goalkeeper's specific attributes. Thus, I decided to assume that those null values could be 0. Finally, I also assume that the first option of position is the player's best position.

Data Visualization:

In order to have a better understanding and explore more insights into this dataset, I create some visualizations. The figures show the top 10 clubs and top 10 national teams that have the highest average scores. Then, there also have two graphs to

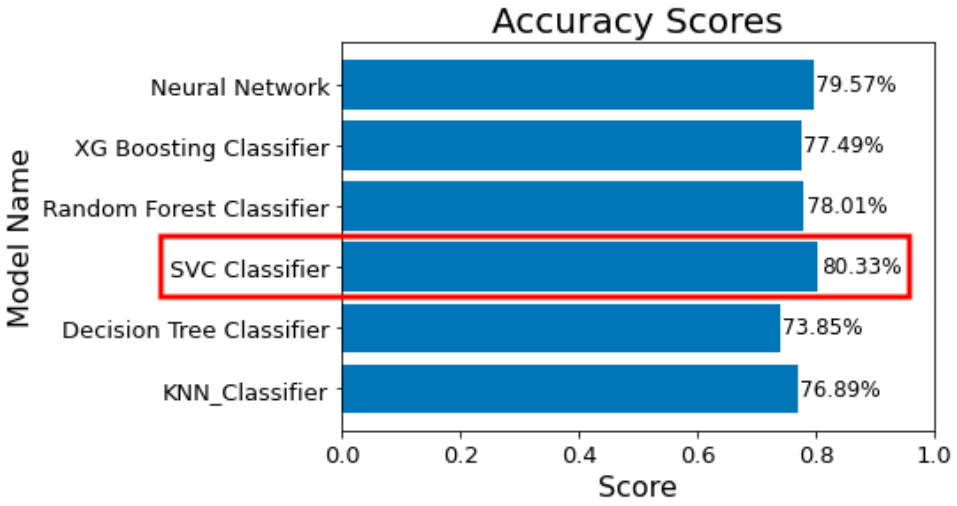
display the player's average value and wage based on seven different kinds of positions. (Figures are shown on the Jupyter Notebook)

Preprocessing and Feature Engineering:

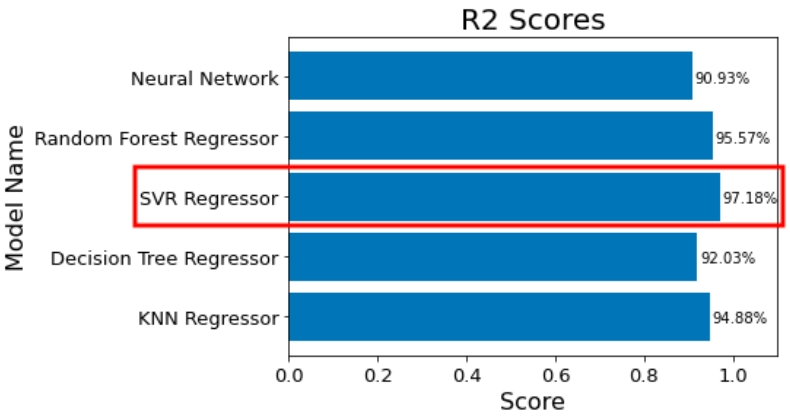
By using the standard scaler, I can make sure that all the features are on the same scale. And I also use the PCA to transform the originally dependent variables to the independent variables and also reduce the 104 features to 10 components. Although feature reduction would decrease the precision of the models, however, if the decreasing of precision is in an acceptable range and we can save lots of run time, it would be pretty worth.

Modelling and Finding:

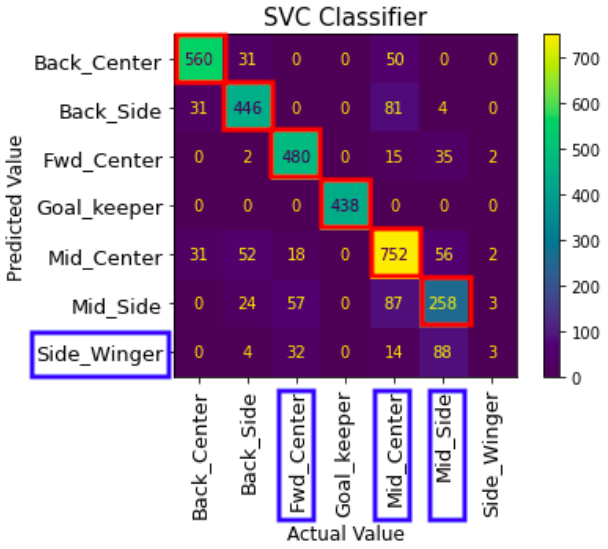
For modelling, I used the pipeline to deal with the leakage problems between all those estimators. And by using the Grid Search CV, I can tune the hyperparameters to find out the best combinations which can give me the most predictive models. After that, I use six different classifiers to predict the player's best position. And as a result, the SVC classifier returns us with the highest accuracy score around 80%.



And I also predict the player's overall score by using five different regressors and the SVR is the most predictive model with around 97 R2-score.



After that, I also create a confusion matrix to show the result of predicting the player's best position by using the most predictive model, SVC classifier. As we can see, the predictions are pretty precise except for the side winger. The side wingers are more likely misclassified into the forward-center, mid-side player and mid-center player. It is because the requirement of becoming a good side winger is pretty high. They need to have more comprehensive and strong abilities so that they can switch to other positions and handle their job during the game.



Future Potential direction:

For my future potential direction, I will continue to track and analyze the famous soccer players to see what is the turning point of their career. Including the turning point that they tend to reach the top of their career and also the turning point where they start to fall down. And I also want to collect more data and explore personality descriptions to see the key personalities for soccer players to succeed in their careers.