Project Intermediate Report

Harshitha Manduva
Sachin Boban
Sai Varun Addanki

Project Sub-tasks

1) Suggest the potential cost of a player in the transfer market. This could be useful for an agent as well as a team Manager.

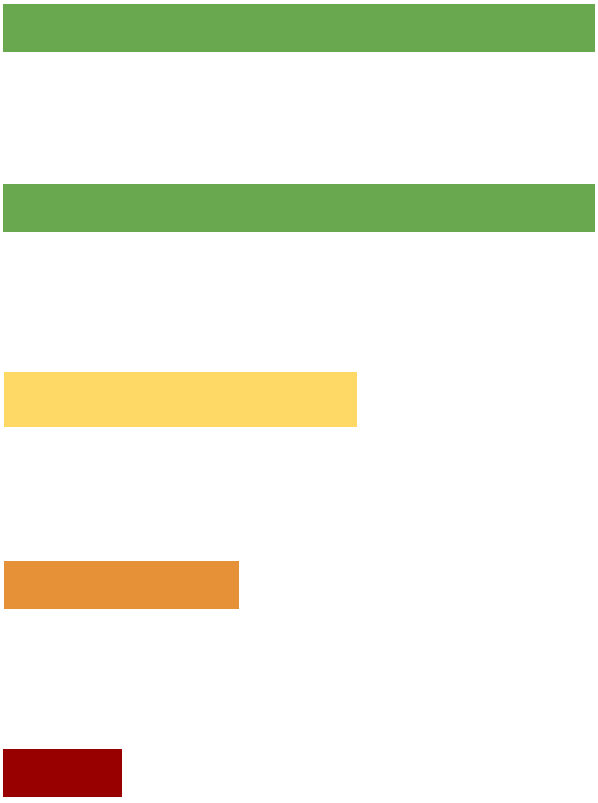
2) Injuries are part of the game. It would be interesting to see if we can suggest the best replacement for an injured player from a pool of available players.

3) Every player has a preferred position, but Managers might have a different plan. It would be interesting to see what attribute a player must improve to fit in the plan.

4) Explore the data on top players and mine the different attributes that contribute majorly to their success. Use this to determine if a new player has the potential to be top rated.

5) Given a set of players and their rating at different positions, we would like to suggest a team formation to the team Manager that would best suit the team.

Progress



Sub-task 1 - Suggest the potential cost of a player in the transfer market. This could be useful for an agent as well as a team Manager.

What progress you have made towards your proposed goal?

In order to first understand what features of a player contribute to his cost/Value, we built the correlation matrix. We picked the top 15 features with which the value of a player correlates and used that to do a regression on our data.

If you tried some basic approaches: what worked well and what did not?
We first used linear regression on the data but root means squared error was too high, of the order of 10^7.
We then used polynomial regression with a degree of 4, which reduced the root means squared error to the order of 10^5
This is a reasonable amount of error, given that the average value of a player in the dataset is again of the order of 10^7

What could be done to improve the basic approaches?
We can use a better algorithm for the regression but polynomial regression seems to do the trick for us now.

What experiments have you run and are you planning to run to demonstrate the effectiveness?
Results :-
Linear regression:
Root means squared Error - 6381240104.69675
Set of features regressed on - 'Overall','International Reputation', 'Potential'
Polynomial regression:
Root means squared Error - 536743.7611382051
Set of features regressed on - 'Age', 'Overall', 'Potential', 'International Reputation', 'Skill Moves', 'ShortPassing', 'LongPassing', 'BallControl', 'Vision'

Sub-task 2 - Injuries are part of the game. It would be interesting to see if we can suggest the best replacement for an injured player from a pool of available players.

What progress you have made towards your proposed goal?
We first built a conceptual model of what we mean by replacing a player. When a player gets injured or has to be replaced, we need to find someone who could play like him. This is again determined by the ability attributes of a player.

If you tried some basic approaches: what worked well and what did not?
Given an injured player, we compute the square of the pairwise Euclidean distance between this player and the rest of the players in the team to find the one that is closest. The parameters for this distance are only the ability attributes of a player. This technique worked well for us as we used some real examples to test this.

What could be done to improve the basic approaches?
We can use a better metric for the distances but Euclidean distance seems to do the trick for us now. But can experiment various other distances.

What experiments have you run and are you planning to run to demonstrate the effectiveness?
Results:-

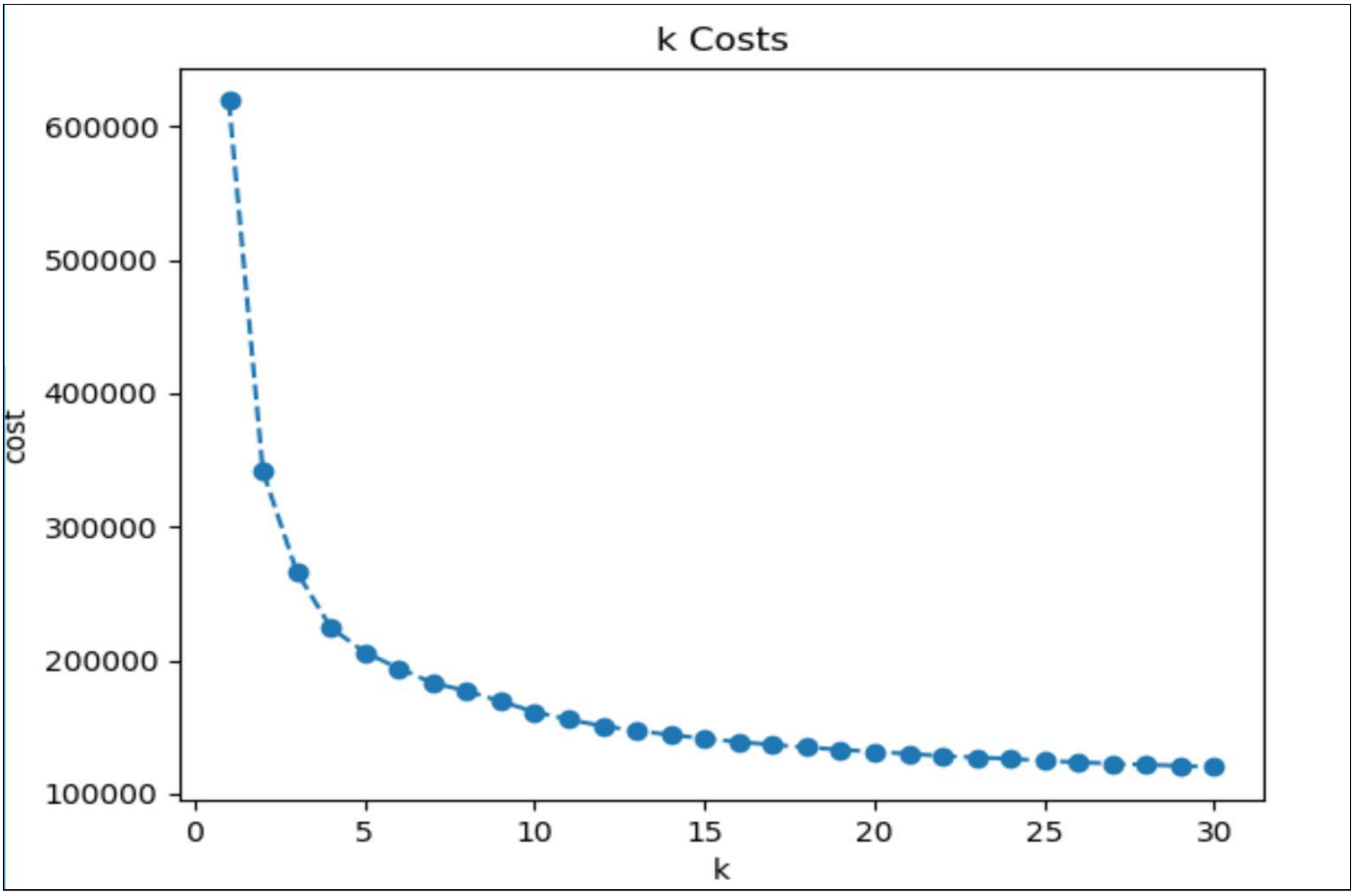
Test 1:-
Player to be replaced - L. Suarez
Player found - P. Coutinho
Team - Barcelona
Verdict - This is a perfectly valid replacement as this happened in an actual match when L. Suarez was red carded and P. Coutinho was moved up front from his usual winger position. Note that a substitute striker wasn't picked in either the real-life example or in our algorithm.

Test 2:-
Player to be replaced - R. Lukaku
Player found - M. Rashford
Team - Manchester United
Verdict - This is a perfectly valid replacement as this happened in an actual match when R. Lukaku was injured for a match M. Rashford was put in as a replacement striker from his usual winger position. Note that a substitute striker wasn't picked in either the real-life example or in our algorithm. M. Rashford continues to be the second choice striker for that position.

Test 3:-
Player to be replaced - A. Kepa
Player found - W. Caballero
Team - Chelsea
Verdict - This is a perfectly valid replacement as this happened in an actual match when A. Kepa was tired after playing a total of 120 minutes. The goalkeeper was (almost) replaced by the substitute goalkeeper W. Caballero by the team coach although the substitution could not go through. Note that a substitute goalkeeper was picked in both the real-life example and in our algorithm. W. Caballero continues to be the second choice goalkeeper for that position.

Sub-task 3 - Every player has a preferred position, but Managers might have a different plan. It would be interesting to see what attribute a player must improve to fit in the plan.

What progress you have made towards your proposed goal?
In order to know what attributes are needed for a player to improve, we first need to find the players who have the best of those abilities. To achieve this, we clustered the data based on ability attributes of a player. We decided to make 7 clusters based on the elbow of the below graph. This is an intermediate step towards the goal. We calculated the k costs for cluster ranging from 1 to 30 but observed an elbow at k = 7.



If you tried some basic approaches: what worked well and what did not?
Our approach of using k means to cluster the data worked well. As the clusters fairly correspond to the positions as well as the overall ratings of the player. We tried K means with K means ++ as well.

What could be done to improve the basic approaches?
Experiment with other types of clustering techniques and look for a metric that determines if the clusters obtained are efficient.