**Final Document**

**Team 10 - creative**

|  |  |  |  |
| --- | --- | --- | --- |
| Team Members | IDs | | Emails |
| Abdelrahman Mohamed Ezzat | 1190158 | abdel.marzouk01@eng-st.cu.edu.eg | |
| Ali Mohamed Ali Hashish | 1190223 | ali.hashish01@eng-st.cu.edu.eg | |
| Farah Mohamed Abdelfatah | 1190176 | farah.mohammed01@eng-st.cu.edu.eg | |
| Ziad Ahmed Hamed | 4200002 | ziad.abdelhameeed01@eng-st.cu.edu.eg | |

**Brief Problem Description:**

Data analysis for FIFA players.

Showing how different factors affect the player­ price and overall score. This provides good feedback to the developers for future releases, where they would be able to determine what the users think of players, as some might be overrated (which suggests they need to be nerfed), or others who are underrated (and need to be buffed). This analysis can also be used by professional users when creating their teams, or when comparing different teams against each other. It can also be used to show how price correlates with other player attributes.

**Dataset:**

<https://www.kaggle.com/datasets/joebeachcapital/fifa-players/data>

We analyzed both male and female players.

**Project Pipeline:**

1. Data Cleaning
2. Preprocessing
3. Exploratory Data Analysis
4. Model Development (with and without map reduce)

**Analysis:**

1. Data preprocessing

* Handling nulls (dropping columns with high percentage of nulls – using KNN imputation to fill some null rows – filling some null cells with statistical approaches)
* Handling duplicates
* Changing categorical to numerical features (One-Hot encoding – target encoding – manipulating features – direct mapping)

1. Visualization

* Using box plots to determine data distribution and outliers
* Plotting histograms of top and bottom nationalities and clubs
* Analyzing weight vs height ratio against overall
* Analyzing left vs right footed player, and their weak foot
* Checking international reputation for each position
* Plotting club names histogram using international reputation
* Creating Hexaco graphs comparing 2 teams against each other based on positions, and providing some recommendations based on results
* Creating a correlation matrix to see the most influential features compared to the labels we want to predict

1. Extracting insights from data

* Most players earn much higher than average in value eur and wage eur, which is evident from the boxplots. We handle this through removing outliers using IQR.
* Learning the general distribution of feature values through the histograms, while also getting insights on the number of outliers through the boxplots. This data, in addition to the correlation matrix, was used to figure out which features would be most useful for our models.
* Comparing between the nations with most and least players, offering constructive proposals to increase the number of players in some nations, in order to increase the fanbase in said nations.
* Noticed that only players having a weight by height ratio between 0.35 and 0.5 have ratings above 90, which might be indicative of the ideal ratio for players.
* Implementing a thorough score measuring system for each position. This allows users to determine what their teams lack, and to intuitively compare between their teams and other contending teams. Furthermore, it offers recommendations for users on what their team lacks, and what needs to improve, and to which threshold.

1. Model

* Through the EDA, it was quickly noticed that some features have much greater values than others, numerically speaking. Thus, we start by normalizing the dataset, to prevent bias towards said features.
* We depend on the correlation matrix to determine the most suitable features based on the correlation to the target feature. Through trial and error, we noticed that taking 15 features offers high accuracy in a timely manner.
* For the model development, we experimented with a few models, some displayed good results, while others did not fare as well.

**Results and Evaluation:**

The following briefly showcases the models and their accuracies for value euro:

* + XGBoostRegressor: R2 score: 94.9% – MAE 0.0023 – MSE 4.68e-05
  + XGBRegressor using KFold cross validation: R2 score 94.5%
  + Support Vector Regression (SVR): Horrible results, with a negative R2 score after training for more than half an hour.
  + KNNRegressor: R2 score 89% – MAE 0.0027 – MSE 0.0001
  + KNN with handmade map reduce: R2 score 89% – MAE 0.001 – MSE 1.15e-05 (Comparable or better than built-in KNNRegressor)

As for the overall:

* XGBRegressor: R2 score 96.2% – MAE 0.018 – MSE 0.0006
* XGBRegressor using KFold cross validation: R2 score 97%
* XGBRegressor using KFold and PCA: R2 score 96.9%
* LinearRegression: R2 score 92.6% – MAE 0.027 – MSE 0.0012
* LinearRegression with 2 PCA features: R2 score 28.2% – MAE 0.08 – MSE 0.012
* LinearRegression with 8 PCA features: R2 score 87.3% – MAE 0.03 – MSE 0.002
* KNN with handmade map reduce: R2 score 90.36% – MAE 0.027 – MSE 0.0014

**Unsuccessful Trials:**

* Tried implementing a neural network
* SVR
* Using MrJob library for map reduce
* LinearRegression with 2 features after PCA

**Future Work:**

* Fixing the neural network trial
* Mixing different models with PCA