DATA ANALYTICS GROUP PROJECT

Module Code: B8IT160

Names: Jack Maguire; Sean Carroll; Shane Crowley

Student Numbers: 20026112; 20024157; 20024425

Introduction/Motivation

This project will conduct data analysis focused on addressing three key research questions using the “EA FC Players Data” (Ghimire, 2023) dataset. The dataset encompasses comprehensive information and statistics about various football players on the game EA FC 24. The motivation behind selecting this project topic stems from our collective familiarity with the game and our understanding of how player attributes directly influence gameplay.

As a group, we recognised the potential for machine learning models to yield insights that can significantly enhance the gaming experience, particularly within the ‘*Career Mode’* game mode. Accurate predictions regarding a players potential can profoundly impact strategic decisions within the game, providing players with a competitive edge.

Ultimately, our goal is to leverage our developed models to effectively address our research questions and directly impact the decisions made by players engaging with the game.

The following 3 research questions were formulated:

1. “How can EAFC24 users optimise player recruitment and development strategies based on age, overall rating, and potential?”
2. “How do different attributes of a footballer influence their overall rating?”
3. “What attributes exhibit the greatest significance in determining a player’s market value?”

Dataset Overview

The “EA FC Players Data” (Ghimire, 2023) dataset was obtained from Kaggle.com, providing a comprehensive repository for analysis. Comprising 18,332 instances and 76 features, it offers a rich source of information for our research endeavors. Prior to analysis, the dataset underwent cleaning procedures in Excel, where particular attention was given to addressing challenges associated with mixed data types observed in 8 of the features. To make the dataset more efficient and focus our analysis, 25 features were deemed irrelevant and consequently removed. These features dropped included ‘version,’ ‘full_name,’ ‘name,’ ‘description,’ ‘image,’ ‘positions,’ ‘work_rate,’ ‘body_type,’ ‘real_face,’ ‘release_clause,’ ‘specialities,’ ‘club_name,’ ‘club_league_name,’ ‘club_logo,’ ‘club_kit_number,’ ‘club_joined,’ ‘country_id,’ ‘country_name,’ ‘country_league_id,’ ‘country_league_name,’ ‘country_flag,’ ‘country_rating,’ ‘country_position,’ ‘country_kit_number,’ and ‘play_styles.’ In accordance with Jayan’s (2024) “ethical considerations in data collection,” ethical considerations were maintained throughout the data collection and analysis processes, ensuring the responsible and respectful use of the information contained within the dataset.

Methodology

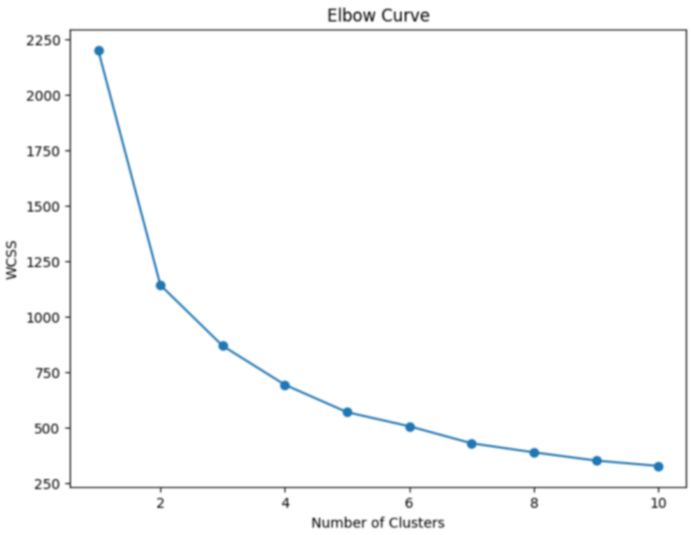
The methodology employed for addressing the research questions commenced with a comprehensive examination of the dataset as a whole. Prior to the data cleaning process, careful consideration was given to the significance of features and instances, facilitating a clear understanding of the dataset's composition and potential implications for analysis. Each research question was then individually addressed, with a dedicated model developed to effectively tackle the specific query posed. The overarching aim was to construct a minimum of two predictive models, comprising two regression models aimed at prediction, alongside a clustering model designed to uncover underlying patterns within the data. Throughout the model development process, evaluation and discussion were undertaken for each model, ensuring thorough examination of their efficacy and relevance to the research questions developed. Recommendations were formulated based on the insights gleaned from the models, providing actionable guidance for optimising player recruitment and development strategies in the context of EA FC24.

Analysis

Methods and Results: Research Question 1

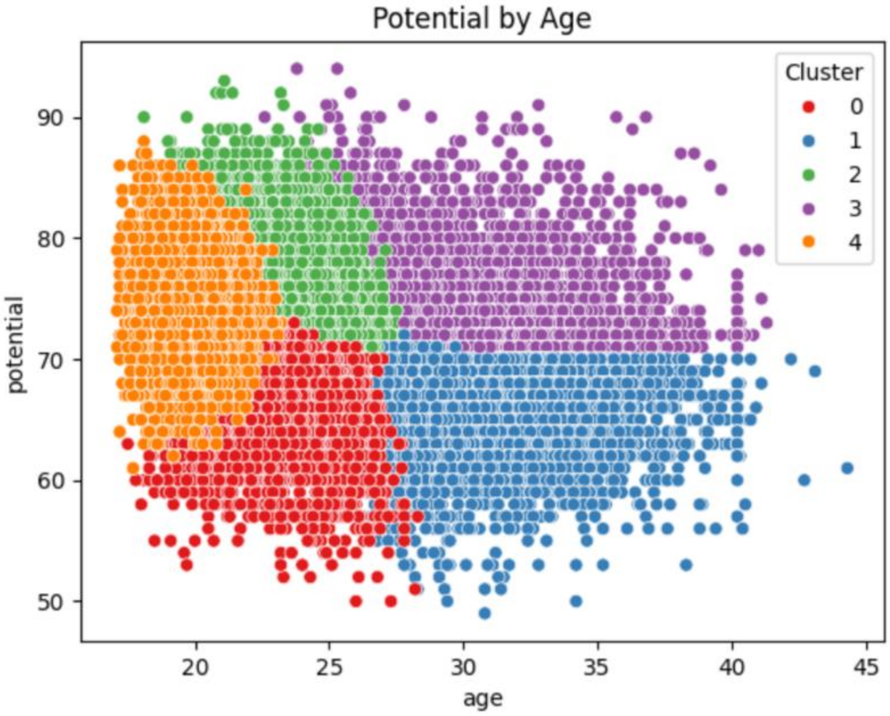
In addressing research question 1, “How can EAFC24 users optimise player recruitment and development strategies based on age, overall rating, and potential?” K-Means Clustering emerged as the selected model due to its computational efficiency and suitability for large datasets (Dublin Business School, 2024). The initial steps involved loading and preprocessing the dataset. Preprocessing steps included the creation of a new feature termed 'growth', calculated as the player's potential minus their current overall rating. To streamline the dataset, 48 irrelevant features were dropped, focusing attention on key attributes such as ‘overall_rating,’ ‘potential,’ ‘value,’ ‘age,’ and ‘growth.’ Standardisation of the data was then carried out using the StandardScaler function to ensure a mean of 0 and variance of 1, facilitating effective analysis and interpretation of the results.

Before applying clustering algorithms, it was essential to determine the optimal number of clusters for the dataset. The elbow method was used to determine the optimal number of clusters.

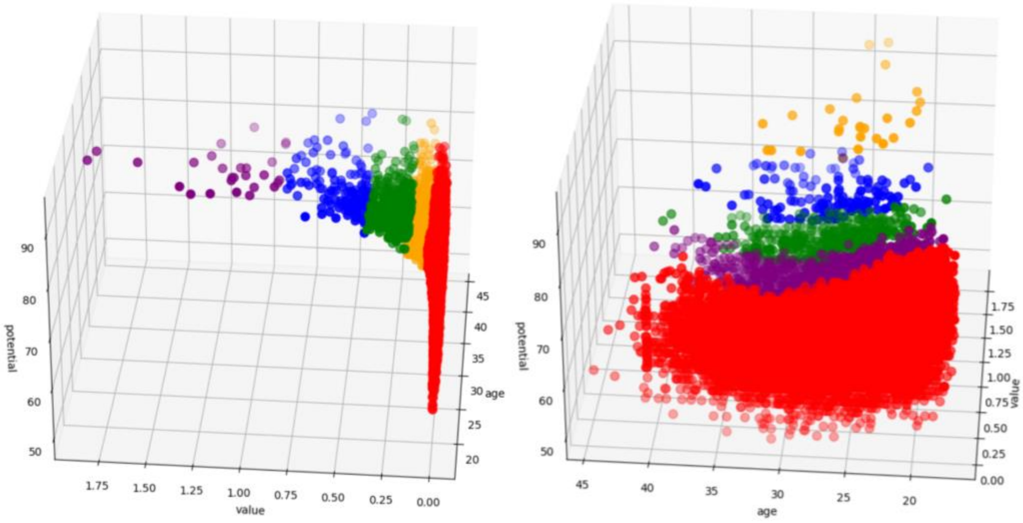


The elbow curve shows the sum of squared distances between each data point and its assigned cluster centre (also known as within-cluster sum of squares or WCSS) for different values of k (number of clusters). We can see that the curve starts to flatten out at k=5, so 5 was chosen as the optimal number of clusters for the dataset.

The K-Means clustering algorithm was then applied to the scaled data with 5 clusters and can be seen in the table below. The cluster labels were added to the original dataframe and visualised the clusters using a scatterplot. The scatterplot shows the distribution of players based on their age and potential, with different colours representing different clusters.



The 3D scatterplots below show front and end elevations of the data and how the potential and age of the players interacts with the value of the players to give a more holistic view of the clustered data.



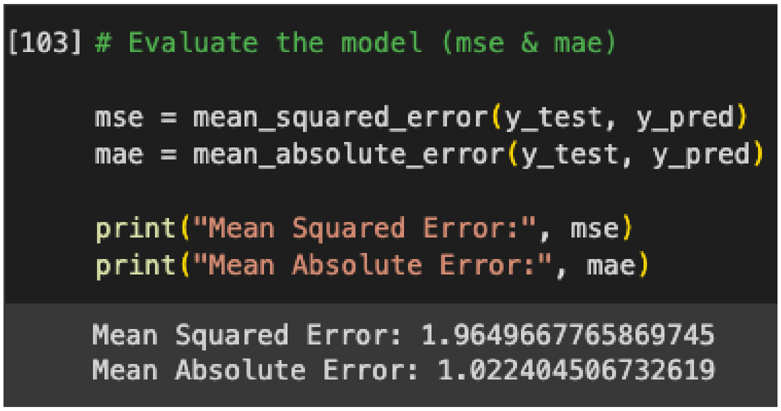
Methods and Results: Research Question 2

In addressing research question 2: “How do different attributes of a footballer influence their overall rating?” Random forest was the chosen regression model to predict the overall rating of a footballer based on different attributes. This was chosen due to the non-linear relationship between the attributes and the overall rating of a player. Once the dataset was uploaded to Google Colab, data preprocessing took place. This included dropping irrelevant columns including ‘age’, ‘wage’ and ‘potential’ as well as many others. These are irrelevant columns as they have no influence on the players overall rating based on their footballing attributes.

The relevant columns were used in the model prediction which included, ‘long_shots’ and ‘aggression’. After being balanced, the data was then trained and tested using the Random Forest model.

A total of 29 attributes were used for the model, predicting the target column ‘overall_rating’.

The output of the model produced both the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). MSE calculates the average squared difference between the actual and predicted values of the target variable (‘overall_rating’). MAE calculates the average of the absolute differences. The results were as follows: MSE = 1.96, MAE = 1.0. This means the model predictions are off, on average, by 1.



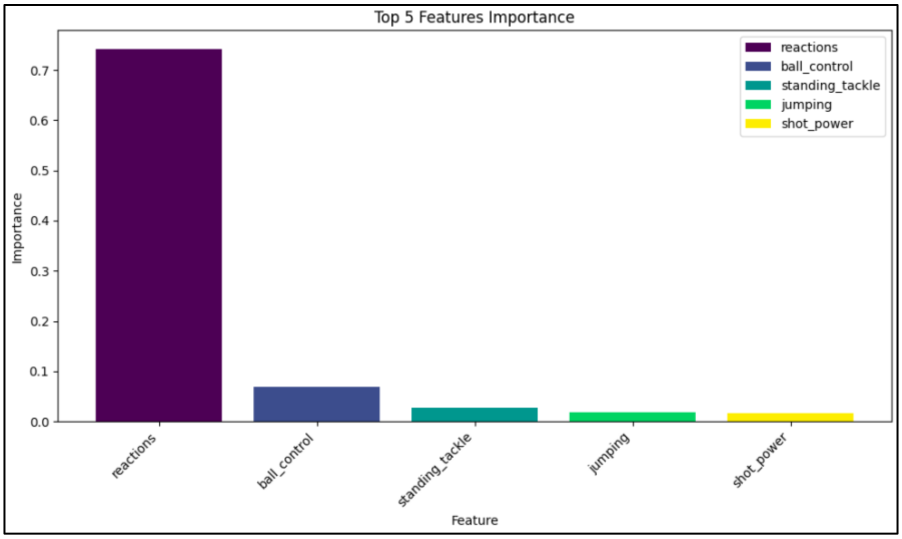
```
[103] # Evaluate the model (mse & mae)

      mse = mean_squared_error(y_test, y_pred)
      mae = mean_absolute_error(y_test, y_pred)

      print("Mean Squared Error:", mse)
      print("Mean Absolute Error:", mae)

      Mean Squared Error: 1.9649667765869745
      Mean Absolute Error: 1.022404506732619
```

A sample dataset was created as an example to give to the model to predict an overall rating. One row, of each attribute with numbers ranging from 50-90, was given to the model to predict an overall rating. The model predicted that the player would be rated 77.83 or 78.



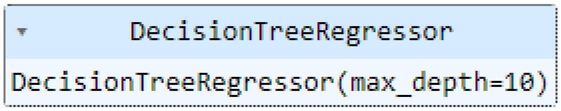
It is also important to analyse which attribute contributed the most to the overall rating. Using the ‘features_importances_’ function, ‘Reactions’ was the attribute which contributed to the overall rating the most. While it was ‘Agility’ that had the least effectiveness on the overall rating of a player.

Methods and Results: Research Question 3

In addressing the third research question of predicting EA FC player market values, a Decision Tree model was employed due to its suitability for capturing non-linear relationships within the dataset. The dataset, containing player information, was loaded and explored, revealing no missing values. The columns dropped included categorical columns such as “Preferred Foot” and columns of no relevance to the target variable such as “Player ID”. After selecting relevant numeric columns, the target variable ('value') was defined. Subsequently, the data was divided into training (70%) and testing (30%) sets.

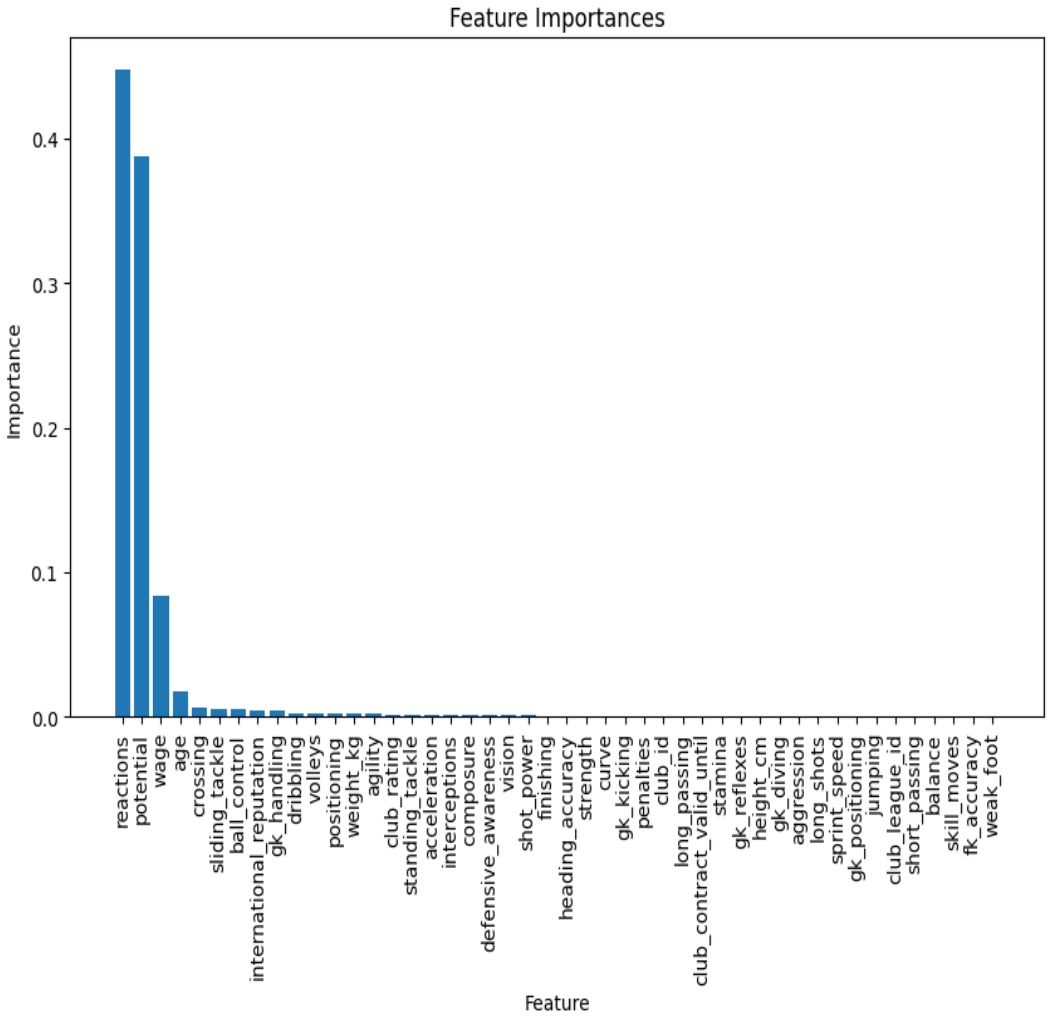
A Decision Tree Regressor was instantiated with a maximum depth of 10 fitted on the training data. The model's performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and R-squared (R2) score. Together, these metrics offer a comprehensive assessment of a regression model's accuracy, precision, and ability to explain variability in the target variable, making them well-suited for model evaluation in regression tasks. Additionally, the model was fine-tuned by varying the maximum depth of the decision tree and observing train and test errors across different depths.

Observations revealed that models with a tree maximum depth exceeding 14 showed errors scoring 0. However, given the risk of overfitting associated with excessively deep trees, it was decided to maintain the maximum depth at 10 as shown below:

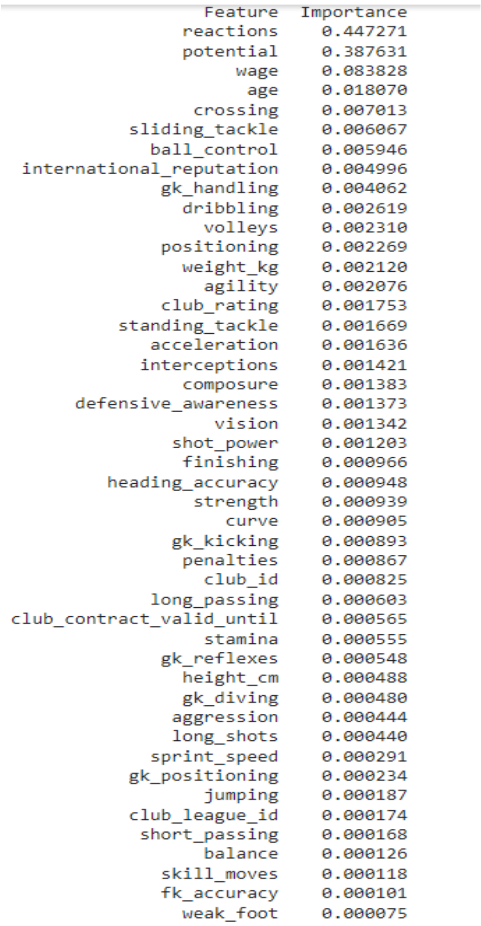


```
DecisionTreeRegressor
DecisionTreeRegressor(max_depth=10)
```

Feature importances were extracted from the trained model to identify the attributes contributing the most to predicting player values. The importance values were visualised using bar plots, providing insights into the relative significance of each feature.



A DataFrame was initialised named feature_importance_df to store feature importances. It sorted the DataFrame by importance in descending order. The importance values indicated each feature's contribution to the model's decision-making process, with higher values signifying greater influence.



```
                  Feature  Importance
                reactions    0.447271
                potential    0.387631
                     wage    0.083828
                      age    0.018070
                 crossing    0.007013
           sliding_tackle    0.006067
             ball_control    0.005946
 international_reputation    0.004996
              gk_handling    0.004062
                dribbling    0.002619
                  volleys    0.002310
              positioning    0.002269
                weight_kg    0.002120
                  agility    0.002076
              club_rating    0.001753
          standing_tackle    0.001669
             acceleration    0.001636
            interceptions    0.001421
                composure    0.001383
      defensive_awareness    0.001373
                   vision    0.001342
               shot_power    0.001203
                finishing    0.000966
         heading_accuracy    0.000948
                 strength    0.000939
                    curve    0.000905
               gk_kicking    0.000893
                penalties    0.000867
                  club_id    0.000825
             long_passing    0.000603
club_contract_valid_until    0.000565
                  stamina    0.000555
              gk_reflexes    0.000548
                height_cm    0.000488
                gk_diving    0.000480
               aggression    0.000444
               long_shots    0.000440
             sprint_speed    0.000291
           gk_positioning    0.000234
                  jumping    0.000187
           club_league_id    0.000174
            short_passing    0.000168
                  balance    0.000126
              skill_moves    0.000118
              fk_accuracy    0.000101
                weak_foot    0.000075
```

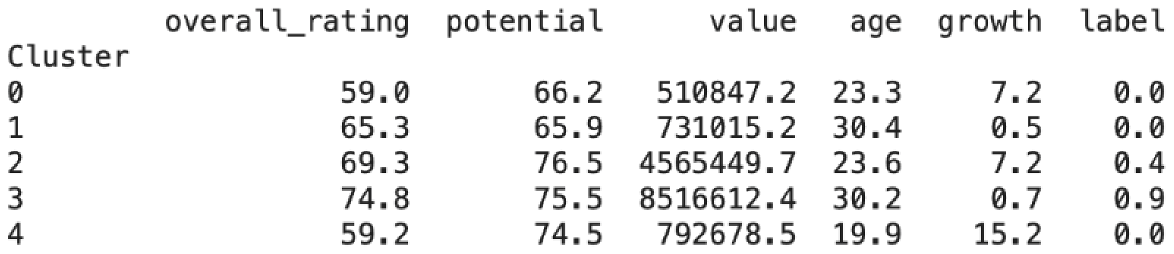
Overall, the methodology involved data preparation, model training and evaluation, model tuning, and interpretation of feature importances to predict FIFA player market values effectively using decision trees.

Discussion

Discussion: Research Question 1

"How can EAFC24 users optimise player recruitment and development strategies based on age, overall rating, and potential?"

Following the creation of a clustering model for the players, the clusters were used to make recommendations of players to users based on their age, overall_rating, value, growth and desired player potential. The average values of each feature were calculated for each cluster and can be seen below. This gives a summary of each cluster that can be used to aid in making these recommendations.



```
         overall_rating  potential      value   age  growth  label
Cluster
0                  59.0       66.2   510847.2  23.3     7.2    0.0
1                  65.3       65.9   731015.2  30.4     0.5    0.0
2                  69.3       76.5  4565449.7  23.6     7.2    0.4
3                  74.8       75.5  8516612.4  30.2     0.7    0.9
4                  59.2       74.5   792678.5  19.9    15.2    0.0
```

The player recommendations are as follows;

- **Cluster 0 - Investment in Young Talent**: Cluster 0 consists of relatively young players with high potential and a significant growth trajectory (average growth of 7.2). Users may consider investing in these players as they have the potential for significant improvement and long-term value appreciation.
- **Cluster 1 - Balanced Squad Development:** Cluster 1 comprises players with moderate potential and growth (average growth of 0.5). Users could focus on maintaining a balance

between experienced players and those with potential for growth to ensure a balanced squad.

- **Cluster 2 - High Potential Investments:** Cluster 2 consists of players with high potential and market value, indicating high demand and value in the transfer market. Users may consider strategic investments in these players to strengthen and enhance their squads.
- **Cluster 3 - Experienced Players with Stable Performance:** Cluster 3 includes experienced players with stable performance and minimal growth (average growth of 0.7). Users may prioritise retaining these players in their club for their consistent performance in game.
- **Cluster 4 - Youth Development Focus:** Cluster 4 represents young players with exceptional potential and significant growth (average growth of 15.2). Users could focus on nurturing and developing these promising talents to maximise their potential and long-term value to the team.

Discussion: Research Question 2

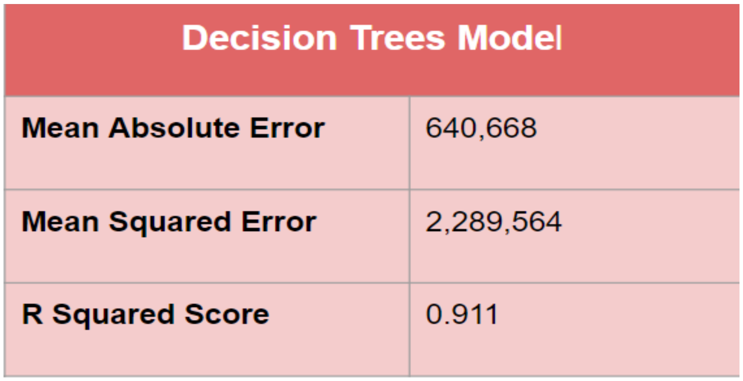
"How do different attributes of a footballer influence their overall rating?"

With the Mean Absolute Error coming in at 1.0 shows the model performed well overall. This means that the rating that the model predicts, is on average 1.0 away from the real overall rating based on the attributes used. It was interesting to see which attributes were most influential in predicting overall churn. Some attributes were lower on the list than expected, for example, 'Balance'. If there were more attributes, would this influence the effect of other attributes? In regards to Random Forest as a model, a large number of trees can make the algorithm too slow and ineffective for real-time predictions. Overall, the model performed well and could be relied upon when predicting an overall rating based on players' different attributes.

Discussion: Research Question 3

"What attributes exhibit the greatest significance in determining a player's market value?"

The decision trees model yielded the following results: the mean absolute error was $640,668, indicating an average deviation of $640,668 from the actual predictions. While this may seem substantial, considering players are valued in millions, it reflects reasonable accuracy. The mean squared error was $2,289,564, and the R-squared score was 0.911, implying that 91% of the variance could be explained by the model's features.



Decision Trees Model	
Mean Absolute Error	640,668
Mean Squared Error	2,289,564
R Squared Score	0.911

Upon analysing the importance values of features, several expected correlations emerged. Notably, player value showed strong correlations with factors like wage, age, potential, and crossing, aligning with conventional football wisdom. Surprisingly, reactions emerged as the most influential feature, resonating with its significance in determining overall player ratings, as observed in research question 2.

Conversely, features such as short passing, balance, skill moves, free kick accuracy, and weak foot scored low in importance. It's worth noting that the model's predictions reflect EA FC gaming environment values, which may not precisely mirror real-world market valuations.

Furthermore, biases inherent in EA FC player prices, such as nationality, league, injury proneness, or the identity of the buyer, must be acknowledged. Validation of the model's accuracy on unseen data, including new EA FC versions, is crucial to ensure its reliability.

These findings not only offer insights for players curious about market values but also suggest avenues for further research into EA FC player valuation. Future studies may explore advanced modelling techniques to account for the identified biases and enhance prediction accuracy. Such insights could prove valuable, especially in modes like career mode, where understanding player values can provide a competitive edge.

Conclusion

In conclusion, our assignment leveraged three distinct machine learning models to analyse and predict outcomes from the dataset, effectively addressing our research questions. Through this process, we acquired valuable insights into data preprocessing techniques and the training of machine learning models. By applying these models to a dataset of personal interest, we deepened our understanding and improved our skills in practical data analysis.

Overall, we are content with the outcomes achieved in response to the research questions, highlighting the practical utility of machine learning models in extracting meaningful insights from data

References

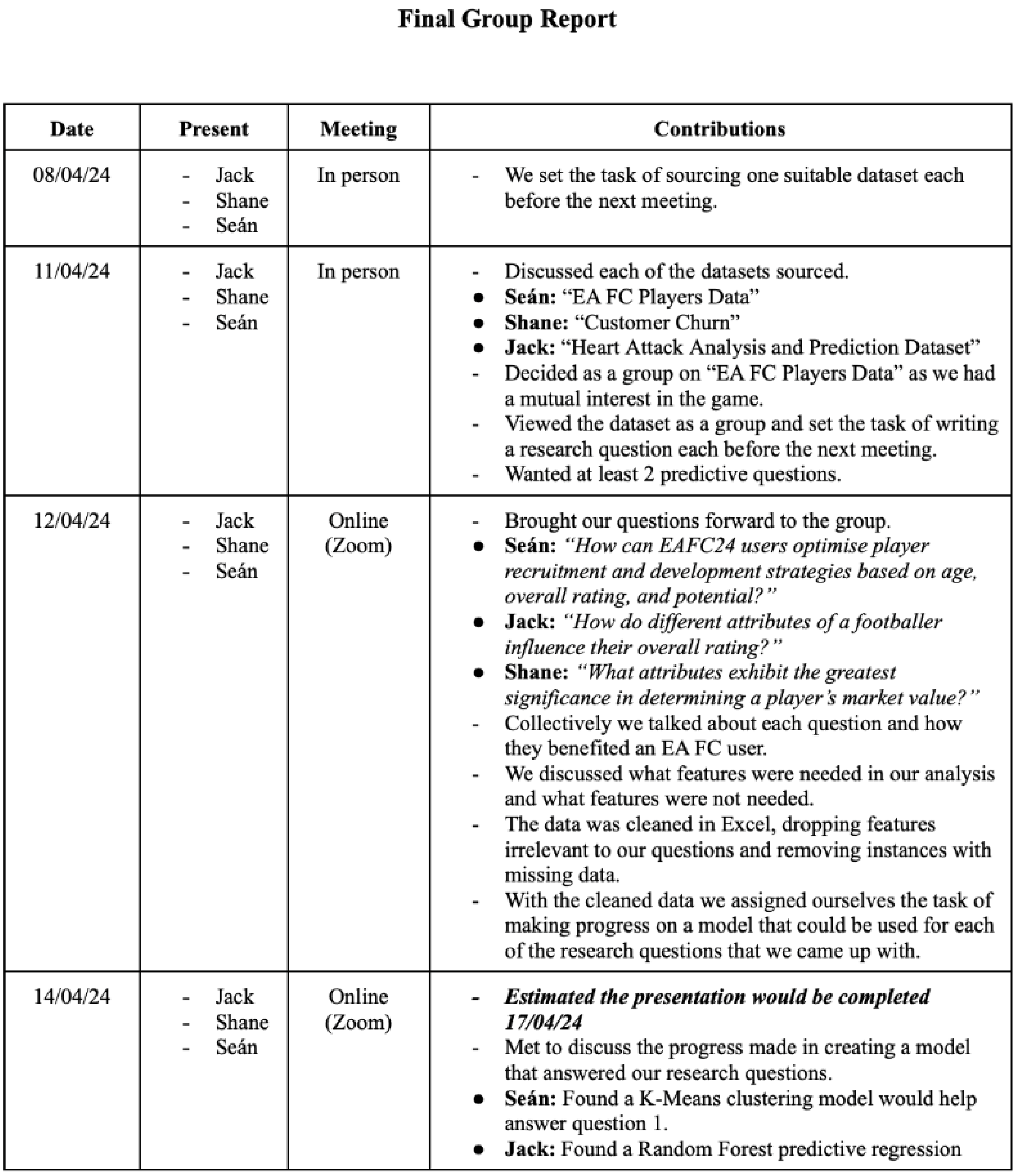
Dublin Business School (2024) *'Unsupervised Machine Learning'* [Lecture Slides]. Available at: https://elearning.dbs.ie/pluginfile.php/2112125/mod_resource/content/1/Unsupervised%20Machine%20Learning.pdf (Accessed: 23rd April 2024)

Ghimire, P. (2023) 'EA FC Players Data' [Dataset]. Available at: https://www.kaggle.com/datasets/ghimireprashant/fifa-players-data (Accessed: 1st April 2024)

Jayan, J. (2024) 'A Comprehensive Guide to Ethical Data Collection and Its Importance', Importance of Ethical Data Collection, 10 January 2024, Available at: https://www.promptcloud.com/blog/importance-of-ethical-data-collection/#:~:text=Data%20ethics%20is%20a%20field,analyzed%2C%20shared%2C%20and%20used. (Accessed: 24th April 2024)

Appendix

Final Group Report



Final Group Report

Date	Present	Meeting	Contributions
08/04/24	- Jack - Shane - Seán	In person	- We set the task of sourcing one suitable dataset each before the next meeting.
11/04/24	- Jack - Shane - Seán	In person	- Discussed each of the datasets sourced. ● **Seán:** “EA FC Players Data” ● **Shane:** “Customer Churn” ● **Jack:** “Heart Attack Analysis and Prediction Dataset” - Decided as a group on “EA FC Players Data” as we had a mutual interest in the game. - Viewed the dataset as a group and set the task of writing a research question each before the next meeting. - Wanted at least 2 predictive questions.
12/04/24	- Jack - Shane - Seán	Online (Zoom)	- Brought our questions forward to the group. ● **Seán:** *“How can EAFC24 users optimise player recruitment and development strategies based on age, overall rating, and potential?”* ● **Jack:** *“How do different attributes of a footballer influence their overall rating?”* ● **Shane:** *“What attributes exhibit the greatest significance in determining a player’s market value?”* - Collectively we talked about each question and how they benefited an EA FC user. - We discussed what features were needed in our analysis and what features were not needed. - The data was cleaned in Excel, dropping features irrelevant to our questions and removing instances with missing data. - With the cleaned data we assigned ourselves the task of making progress on a model that could be used for each of the research questions that we came up with.
14/04/24	- Jack - Shane - Seán	Online (Zoom)	- ***Estimated the presentation would be completed 17/04/24*** - Met to discuss the progress made in creating a model that answered our research questions. ● **Seán:** Found a K-Means clustering model would help answer question 1. ● **Jack:** Found a Random Forest predictive regression

			model would help answer question 2. ● **Shane:** Found a Decision Tree predictive model would help answer question 3. - Viewed each others' code so far and discussed the adjustment of values in relation to each of the predictive models - Set the task of completing the code for each model as well as starting the presentation.
15/04/24	- Jack - Shane - Seán	In person	- Went through the completed models and collectively wrote recommendations for the user based on each model. - Discussed the presentation and the roles each of us would take in the presentation. ● **Jack:** Motivation; Research Questions; Methodology: Research Question 2; Analysis: Research Question 2; Recommendations: Research Question 2 ● **Seán:** Dataset Overview; Methodology; Methodology: Research Question 1; Analysis: Research Question 1; Recommendations: Research Question 1 ● **Shane:** Methodology: Research Question 3; Analysis: Research Question 3; Recommendations: Research Question 3; Conclusion
18/04/24	- Jack - Shane - Seán	In person	- Carried out the presentation. - Discussed how the presentation went and how feedback could be incorporated into the report. - Assigned the sections each of us would write in the report. ● **Jack:** Introduction/Motivation; Analysis (Methods and Results: Research Question 2); Discussion: Research Question 2 ● **Seán:** Dataset Overview; Methodology; Analysis (Methods and Results: Research Question 1); Discussion: Research Question 1 ● **Shane:** Analysis (Methods and Results: Research Question 3); Discussion: Research Question 3; Conclusion - Set the task of writing the report. - ***Report estimated to be completed 24/04/24***
24/04/24	- Jack - Shane - Seán	Online (Zoom)	- We read through the report together and edited the report together to ensure fluency when reading. - Completed the report. - ***Report to be submitted 24/04/24***