OnSports - Problem Statement

**Submission type**

:

File Upload

**Due Date**

:

Feb 06, 10:00 AM

**Total Score**

:

60

**Description**

**Context**

Fantasy sports are online gaming platforms where participants draft and manage virtual teams of real professional sports players. Based on the performance of the players in the real world, players are allotted points in the fantasy sports platform every match. The objective is to create the best possible team with a fixed budget to score maximum fantasy points, and users compete against each other over an entire sports league or season. Some of these fantasy sports require actual financial investments for participation, with the chances of winning monetary rewards as well as free matchday tickets on a periodic basis.

The fantasy sports market has seen tremendous growth over the past few years, with a valuation of $18.6 billion in 2019. The football (soccer) segment led in terms of market share in 2019, with over 8 million participants worldwide, and is expected to retain its dominance over the next couple of years. Digitalization is one of the primary factors driving the growth of the fantasy sports market as it allows participants the opportunity to compete on a global level and test their skills. With an increase in smartphone usage and availability of fantasy sports apps, this market is expected to witness a globe surge and reach a $48.6 billion valuation by 2027.

**Objective**

OnSports is a fantasy sports platform that has fantasy leagues for many different sports and has witnessed an increasing number of participants globally over the past 5 years. For each player, a price is set at the start, and the price keeps changing over time based on the performance of the players in the real world. With the new English Premier League season about to start, they have collected data of the past season and want to analyze it to determine the price of each player for the start of the new season. OnSports have hired you as a data scientist and asked you to conduct a cluster analysis to identify players of different potentials of each player based on previous season performance. This will help them understand the patterns in player performances and fantasy returns and decide the exact price to be set for each player for the upcoming football season.

**Data Description**

The data comprises player stats like the number of goals scored, goals created, minutes played, fantasy points scored in the previous season, etc. The detailed data dictionary is given below.

* Player\_Name: Name of the player
* Club: Club in which the player plays
* Position: Position in which the player plays
* Goals\_Scored: Number of goals scored by the player in the previous season
* Assists: Number of passes made by the player leading to goals in the previous season
* Total\_Points: Total number of fantasy points scored by the player in the previous season
* Minutes: Number of minutes played by the player in the previous season
* Goals\_Conceded: Number of goals conceded by the player in the previous season
* Creativity: A score, computed using a range of stats, that assesses player performance in terms of producing goalscoring opportunities for other players
* Influence: A score, computed using a range of stats, that evaluates a player's impact on a match, taking into account actions that could directly or indirectly affect the match outcome
* Threat: A score, computed using a range of stats, that gauges players who are most likely to score goals
* Bonus: Total bonus points received (The three best performing players in each match receive additional bonus points based on a score computed using a range of stats. 3 points are awarded to the highest-scoring player, 2 to the second-best, and 1 to the third.)  
  Clean\_Sheets: Number of matches without conceding a goal in the previous season

**Best Practices for Notebook**

* The notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is preferable to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).

**Best Practices for Presentation**

* The presentation should be made keeping in mind that the audience will be a business leader like CMO, COO, CFO, or CEO.
* The key points in the presentation should be the following:
  + Business overview of the problem and solution approach
  + Key findings and insights which can drive business decisions
  + Model overview and performance summary
  + Business recommendations
* Focus on explaining the takeaways in an easy-to-understand manner.
* The inclusion of the potential benefits of implementing the solution will give you the edge.
* Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
* The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file.

**Submission Guidelines**

1. There are two parts to the submission:
   1. A well commented Jupyter notebook [format - .html]
   2. A presentation as you would present to the top management/business leaders [format - .pdf]
2. Any assignment found copied/plagiarized with other groups will not be graded and awarded zero marks.
3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation.
4. Submission will not be evaluated if
   1. it is submitted post-deadline, or,
   2. more than 2 files are submitted.

Happy Learning!

**Scoring guide (Rubric) - OnSports**

| **Criteria** | **Points** |
| --- | --- |
| **Exploratory Data Analysis**  - Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis - Insights based on EDA | 10 |
| **Data preprocessing**  - Duplicate value check - Missing value treatment - Outlier check - Feature engineering - Feature scaling | 6 |
| **Applying K-means clustering**  - Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out appropriate number of clusters - Cluster Profiling | 12 |
| **Applying Hierarchical clustering**  - Apply Hierarchical clustering with different linkage methods - Plot dendrograms for each linkage method - Check cophenetic correlation for each linkage method - Figure out appropriate number of clusters - Cluster Profiling | 12 |
| **K-means vs Hierarchical Clustering**  Compare clusters obtained from K-means and Hierarchical clustering techniques | 4 |
| **Actionable Insights & Recommendations**  Key takeaways for the business | 6 |
| **Presentation - Overall quality**  - Structure, flow, and visual appeal - All key insights and recommendations covered | 6 |
| **Notebook - Overall Quality**  - Structure and flow - Well-commented code |  |