eProject C<7>: Predicting the stats of NBA All-Stars for the upcoming seasons.

Gregor Nepste,
Tamur Talviku

Task 1

Link to Github:
https://github.com/gnepste/Sissejuhatus-Andmeteadusese-NBA-statistics-Project-

Task 2 - business understanding

Backround

The objective of this project is to predict key NBA player stat lines for the upcoming season using machine learning algorithms, for example a regression model. This involves using historical player data, such as points per game, rebounds, assists, field goal percentage, etc., to forecast how individual players might perform in the upcoming season.

Business goals

The primary goal is to provide valuable insights for sports analysts and team managers who heavily rely on player performance predictions for strategic decision-making. This could help the analysts and the front offices of an NBA team in making informed choices about player rotations and strategies and in assessing team dynamics.

Business success criteria

The success of this project will be measured by the accuracy of the predictions made by the regression model. The model's ability to consistently forecast player performances close to their actual stat lines will be a key metric.

Inventory of resources

- NBA player data from previous seasons
- Machine learning algorithms
- Expertise in data preprocessing, feature engineering, and model evaluation

Requirements, assumptions, and constraints

- Requirement: Availability and quality of historical player data
- Assumption: Past performance is indicative of future performance

- Constraint: Unforeseen changes in player health, team dynamics, or game strategies could impact predictions

Risks and Contingencies

Risks: Regression line not handling the complexity of data, data scarcity for certain players, model performance affected by outlier data
Contingencies: Regular model evaluation, robust feature selection, and considering alternative data sources if required

Terminology

- PPG - points per game
- RPG - rebounds per game
- APG - assists per game
- FG% - field goal percentage

Costs and Benefits

Costs: time and computational resource
Benefits: insightful analytics for analysts, strategic guidance for team managers

Data-Mining Goals

- Develop and train a regression model using historical NBA player data
- Evaluate and fine-tune the model's performance to achieve accurate predictions

Data-Mining Success Criteria

- Model accuracy in predicting upcoming season stat lines with a low margin of error

Task 3

Gathering Data: The project requires historical NBA player statistics, focusing on points, rebounds, assists, and field goal percentages. Data sources include RealGM, Basketball_reference, official NBA sites and also Kaggle pages. This will help us with ensuring that we have access to several past seasons. Criteria for selection prioritize data completeness and recency.

- **Data requirements:** CSV file
- **Data availability:** As can be seen from above, there is plenty of data available for doing this research. Since it is not decided yet on which source/sources to use exactly (it may depend on the method we are going to use), exact data will not be pointed out
- **Selection criteria:** Data from 2020-2023, must consist of popular players in NBA, must have all the pointed out parameters.

Describing Data: The collected data is structured, with each row corresponding to a player's seasonal performance and columns detailing various statistics. Information on player names, teams, and positions is also included.

Exploring Data: Initial exploration will involve calculating basic statistics, such as averages and variance, and examining the relationships between different performance metrics. Visual tools like histograms and scatter plots will be utilized to identify patterns and outliers.

Verifying Data Quality: Quality checks will ensure data accuracy, addressing missing values and correcting any inconsistencies. Data types will be reviewed for correctness, and any obvious errors will be cleaned to prepare for the modeling phase.

This data understanding step ensures that we have a well-defined and high-quality dataset, crucial for building a reliable predictive model for NBA player performance.

Task 4

Project Plan:

NB all the tasks will be split evenly between the teammates!

Collect data (2 hours total): We'll get player stats from the past three years.

Clean data (2.5 hours total): Any errors in the data will be fixed. Also parameters that are not needed will be taken out in order to avoid any errors.

Build the models (8 hours total): The main task where we create our prediction tool. We will also conduct several other methods that will improve the poster.

Test the model (4 hours total): Testing the models in order to find out whether something needs adjusting or the output is not logical. Will also do several re-runs in order to find out if the output changes.

Write the report and design a poster (5 hours total): We'll all put in 2.5 hours to write up what we did and what we found.