Project 4 Proposal

NBA Player Performance Analysis, Clustering and Prediction (Inspired by Movie Moneyball!)

Objective: Use regression analysis, PCA to reduce the dimensionality of NBA player performance metrics, K-Means Clustering to group similar players, Random Forest to predict player performance, and Neural Networks to enhance prediction accuracy and identify patterns in NBA player data.

Outcome:

- **Enhanced Player Profiling**: Identify distinct player styles and performance potential based on data-driven analysis.
- **Actionable Insights**: Provide valuable insights into player development, team strategy, and scouting decisions, enhancing the overall approach to managing and optimizing player performance.
- **Advanced Prediction**: Improve player performance predictions using Neural Networks, capturing patterns in player data that traditional models might miss.

Time period: 2007 to 2024 (Each year has a dataset with ~210 data points)

Dataset: NBA Playoff Player Statistics from NBA
- **Independent variable**
    - MIN (Minutes Played)
    - OFFRTG (Offensive Rating)
    - DEFRTG (Defensive Rating)
    - EFG% (Effective Field Goal Percentage)
    - TS% (True Shooting Percentage)
    - USG% (Usage Percentage)
    - AST% (Assist Percentage)
    - AST/TO (Assist to Turnover Ratio)
    - OREB% (Offensive Rebounding Percentage)
    - DREB% (Defensive Rebounding Percentage)
    - REB% (Total Rebounding Percentage)
    - TO RATIO (Turnover Ratio)
    - PACE (Pace)
    - AGE
- **Dependent variable**
    - PIE (Player Impact Estimate)

Detailed Steps:

1. **ETL**:
    - Collect data on player performance from NBA league including metrics such as points scored, assists, rebounds, tackles, goals, passing accuracy, etc.

- Clean and preprocess the data by handling missing values, standardizing numerical features, and encoding categorical data.

2. **Regression analysis to predict PIE**
3. **Dimensionality Reduction with PCA**:
    - Apply PCA to reduce the dimensionality of the performance metrics to focus on the most significant features that contribute to player performance.
    - Determine the optimal number of principal components by analyzing the explained variance ratio, retaining components that capture the majority of the variance.
4. **Player Clustering with K-Means**:
    - Use K-Means Clustering on the reduced data from PCA to group players into distinct clusters based on similar playing styles or performance metrics (e.g., scorers, defenders, playmakers).
    - Visualize the clusters using scatter plots with the principal components to interpret different player types.
5. **Performance Prediction with Random Forest**:
    - Use Random Forest to predict future player performance metrics based on historical data and identified clusters.
    - Feature importance analysis from Random Forest to highlight which metrics are most predictive of future performance.
6. **Enhanced Prediction with Neural Networks**:
    - **Neural Network Model**: Build a Neural Network model (e.g., Multi-Layer Perceptron) to capture complex, non-linear relationships in player performance data that might be missed by Random Forest.
    - **Network Architecture**: Use an architecture with input layers matching the number of PCA components, hidden layers tuned based on data complexity (e.g., 2-3 hidden layers with a varying number of neurons), and an output layer predicting player performance metrics.
    - **Activation Functions**: Use ReLU activation functions for hidden layers and an appropriate activation for the output layer based on the target variable (e.g., linear activation for regression).
    - **Training**: Train the Neural Network using backpropagation and optimize using techniques like Adam optimizer with mean squared error (MSE) as the loss function.
    - **Evaluation**: Compare the performance of the Neural Network with Random Forest using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score to evaluate prediction accuracy.
    - **Regulation:** L2 regulation to avoid model overfitting.
7. **Model Comparison and Insights**:
    - **Compare Models**: Compare the predictions of the Neural Network and Random Forest to see which model better captures player performance trends.
    - **Model Interpretation**: Use visualization tools like SHAP (SHapley Additive exPlanations) to interpret the Neural Network's predictions, identifying which features (or PCA components) most influence the results.

8. **Deploy Insights**:
   - Deploy the model insights to assist coaches, scouts, or sports analysts in decision-making processes like player trading, game strategy adjustments, and talent identification.
   - Use cluster-based player profiles to identify potential future stars or undervalued players.

Tools and Techniques:

- **Libraries**: Python libraries such as Scikit-learn (PCA, K-Means, Random Forest), TensorFlow or PyTorch (Neural Networks), Pandas (data manipulation), Matplotlib and Seaborn (visualization).
- **Visualization**: Use dimensionality reduction plots (e.g., 2D scatter plots of PCA components), cluster heatmaps, and SHAP plots to interpret model predictions, spider plot

Appendix

Slide presentation [link]

Report [link]

Glossary
GP
Games Played
W
Wins
L
Losses
MIN
Minutes Played
OFFRTG
Offensive Rating
DEFRTG
Defensive Rating
NETRTG
Net Rating
AST%
Assist Percentage
AST/TO
Assist to Turnover Ratio
AST RATIO
Assist Ratio
OREB%
Offensive Rebounding Percentage
DREB%

Glossary
Defensive Rebounding Percentage
REB%
Rebounding Percentage
TO RATIO
Turnover Ratio
EFG%
Effective Field Goal Percentage
TS%
True Shooting Percentage
USG%
Usage Percentage
PACE
Pace
PIE
Player Impact Estimate
POSS
Possessions