

ENGR 5775:NBA LineUp Prediction Project

Sweta Patel 100915164
Vallika Kasibhatla 100928820

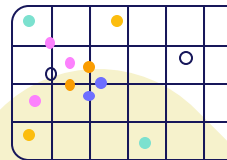
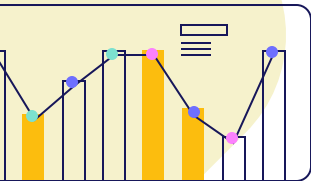


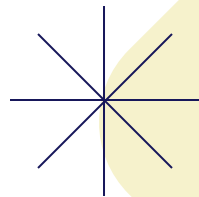
Table of Contents

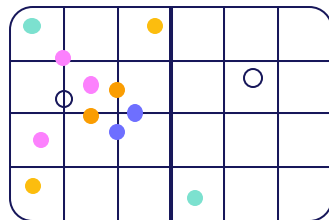
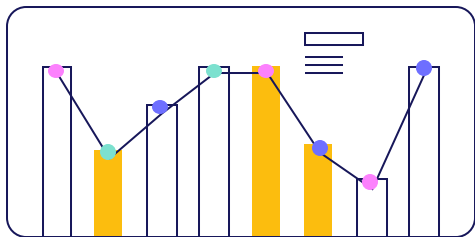
Part 1 (Base Case)

- 01. Problem Statement**
- 02. Dataset**
- 03. Preprocessing**
- 04. Feature Transformation**
- 05. Model Building**

Part 2 (Enhanced Case)

- 01. Feature Engineering**
- 02. Embeddings**
- 03. Model Building**
- 04. Evaluations**
- 05. Test Cases**





Part 1: Base Cases



Problem Statement

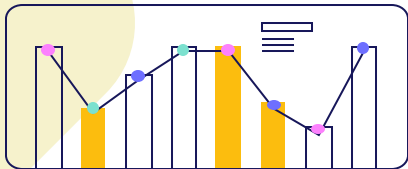
Problem 1

When provided five home team players names and five away team players names, predict the outcome of the game.

Problem 2

When four away team players and five home team players are given, predict the fifth player in the lineup, predict the fifth player in the lineup to enhance the likelihood of the away team winning.





Facts of the Dataset

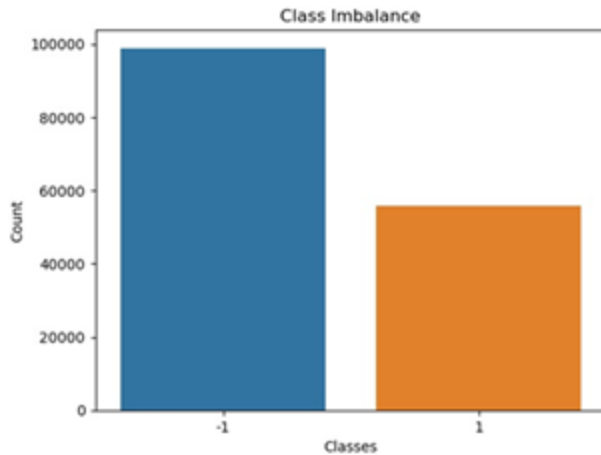
- There are 197718 states
- 2007-2012 seasons
- Around 813 unique players
- There are 31 unique teams. Short names: PHO, LAC, UTA, DAL, MEM, PHI, NOK, NJN, HOU, MIL, SAC, GSW, TOR, WAS, POR, NYK, MIA, SEA, CLE, ORL, MIN, SAS, ATL, CHI, BOS, IND, LAL, CHA, DET, DEN, NOH, OKC.
- Numeric Data for the team Statistics
- Target Variable: Outcome (-1,1)

Pre-Processing

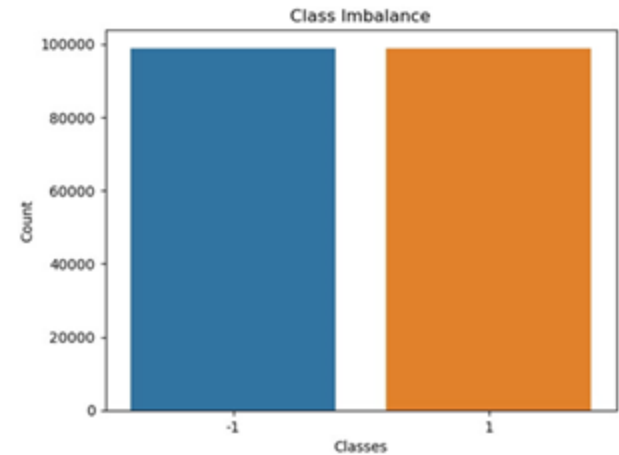
1. Missing Values

2. Scaling and Standardization

3. Class Imbalance

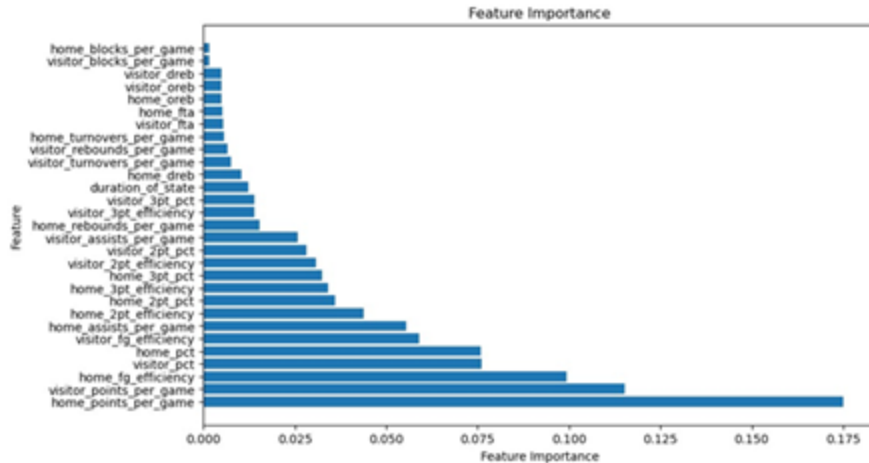


Up Sampling



Feature Selection and Transformation

Feature Importance



Feature Transformation

- ❑ Changed Class Label to 0 and 1

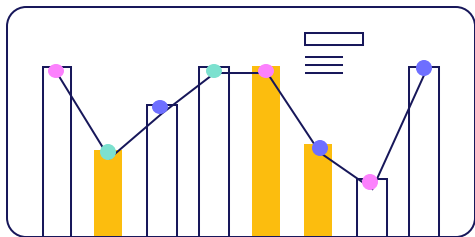
home_0	home_1	home_2	home_3	home_4
41	482	511	692	715
41	482	511	692	715
511	554	673	715	785
482	554	692	715	785
41	482	511	554	715

- ❑ Label Encoding: Each unique player will be mapped with a unique number
- ❑ Better to avoid sparsity

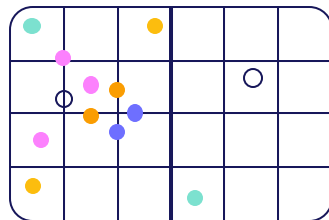
Model Building

Classification Models

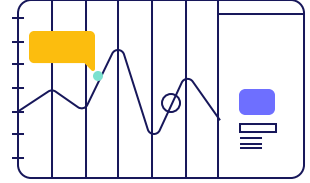
Model	Hyper parameters	Input	Accuracy	Issues
SVM	Kernel-linear	Label Encoding and Standardization	51%	More than simply encoding the players, talking about their performance with internal team members and how players play is necessary because there needs to be a differentiation factor. This can be the reason for low accuracy. It is difficult to define the decision boundary.
Logistic Regression	Max Iteration=1000		51%	
Neural Network	Dense Layer = Relu Sigmoid layer = last layer Learning Rate = 0.001 Loss=Binary-Cross entropy		52%	
Gradient Boosting	Learning Rate = 0.1		57%	



Part 2: Enhanced Case



Feature Engineering



- **Shooting the ball**

Effective Field Goal Percentage= (Field Goals Made) + 0.5*3P Field Goals Made)/((Field Goal Attempts)

- **Taking care of the ball**

Turnover Rate=Turnovers/ (Field Goal Attempts + 0.44*Free Throw Attempts + Turnovers)

- **Offensive rebounding**

Offensive Rebounding Percentage = (Offensive Rebounds)/ [(Offensive Rebounds) + (Opponent's Defensive Rebounds)]

- **Getting to the foul line**

Free Throw Rate= (Free Throws Made)/((Field Goals Attempted) or Free Throws Attempted/Field Goals Attempted



○ Embeddings

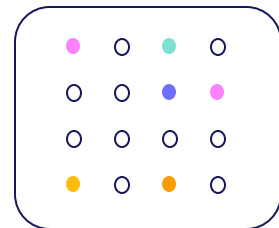
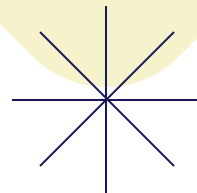
Embeddings: A common technique in data analysis.

Significance: Adds meaningful insights to text data.

Dimension Reduction: Aids in managing complex data.

Semantics Captured: Embeddings capture semantics through similarities.

Application: Similar combinations of players with statistics are closer in the embedding space.



Embeddings

Supervised Learning Approach: Utilized due to the presence of a target variable (labelled data).

Model Architecture: Neural network with three inputs: home players' names, away players' names, and statistics.

Flattening and Concatenation: Embeddings are flattened and concatenated with statistics before passing into a dense layer.

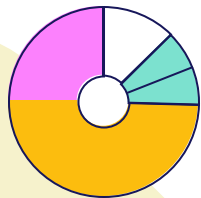
Output: Classification of the embedding as win or lose.

Feature Vector: The model generates a 32-dimensional feature vector for each player.

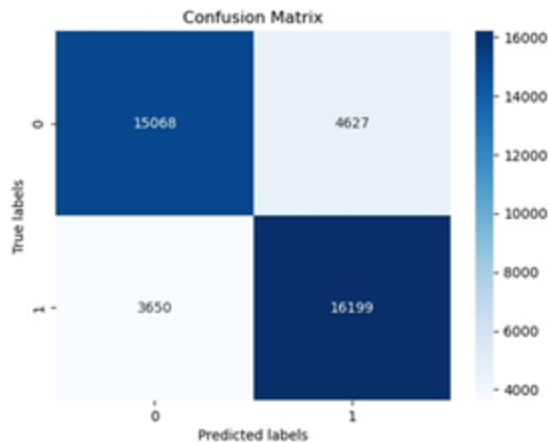


Model Building

Model	Hyperparameters	Input	Accuracy
SVM	Kernel="Linear"	Embeddings	59%
Gradient Boosting	Learning Rate=0.001	Embeddings	57%
Neural Network	Dense Layer= <u>Relu</u> , Sigmoid	Embeddings	68%
KNN	Euclidean Distance, Cos-Sin	Embeddings	68%
Random Forest	-	Embeddings	79%

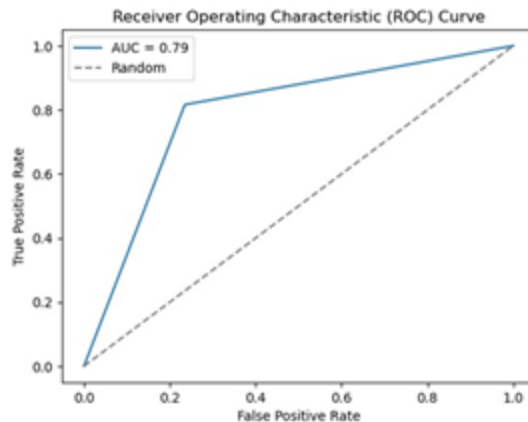


Evaluating Random Forest Model



Classification Report:

	precision	recall	f1-score
0.0	0.81	0.77	0.78
1.0	0.78	0.82	0.80
accuracy	0.79		









Test Case



Problem 1: Predicting Game Outcome

- **Objective:** Determine if a lineup of five home players and five away players will lead to a win or loss for the home team.
 - **Input:** Embedded values of the 10 players (home and away) after encoding.
 - **Output:** Target outcome: "1" for home team win, "0" for away team win.
- 
- 
- 
- 

Test Case

Problem 2: Player Selection for Away Team

- **Objective:** Identify one player for the away team that maximizes the chances of winning.
- **Input:** Information about home and away team players, along with the season year and team names.
- **Approach:** Given the input information, fetching the remaining away players for that season and then getting the corresponding embeddings for all the players including home and away team members and then passing those embeddings in the classification model defined for the problem 1 to predict the outcome of the game.
- **Output:** Target outcome: "0" for a successful lineup suggestion since it indicates the winning of away team, "1" otherwise. Getting the winning probability for every remaining player of the away team and suggesting the one with the highest one as an ideal player.

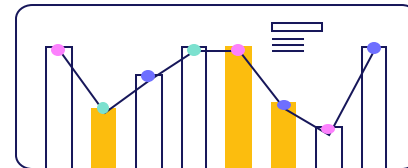
Test Case

Example Scenario:

- Home Team Players: 'Joe Johnson', 'Josh Childress', 'Marvin Williams', 'Salim Stoudamire', 'Tyronn Lue'.
- Away Team Players: 'Grant Hill', 'Leandro Barbosa', 'Raja Bell', 'Sean Marks'.
- Team Names: Home (ATL), Away (PHO).

```
Player : Shaquille O'Neal Probability : 0.6375
Player : Steve Nash Probability : 0.73
Player : Boris Diaw Probability : 0.8
Player : Brian Skinner Probability : 0.7659999999999999
Player : D.J. Strawberry Probability : 0.696
Player : Linton Johnson Probability : 0.7275
Player : Gordan Giricek Probability : 0.7875
Player : Marcus Banks Probability : 0.8
Player : Alando Tucker Probability : 0.74
Player : Shawn Marion Probability : 0.86
Player : Eric Piatkowski Probability : 0.7875
Player : Amar'e Stoudemire Probability : 0.7959999999999999
```

Conclusion



Dataset Overview: The NBA dataset, which ran from 2007 to 2012, included detailed game information such as player names, team statistics, and outcomes.

Predictive Tasks: We addressed two major tasks:

Away Team Win Prediction: Determine if the home team will win.

Key Player Identification: Predicting the 10th player is critical for a victory.

Challenges: Overcame data imbalances by upsampling techniques. Used feature engineering, label encoding, and four-factor statistics.

Model Evaluation: We investigated SVM, KNN, Neural Networks, Random Forest, and Gradient Boost.

Key findings:

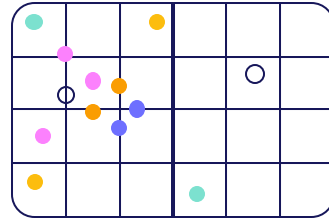
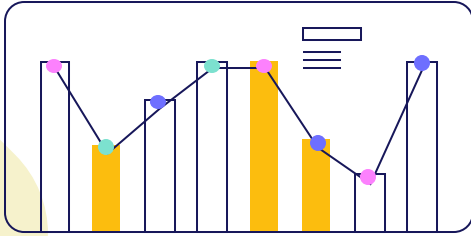
Embeddings Boost: Using embeddings to improve model performance.

Random Forest was the best-performing model, with 79% accuracy.

Implications and recommendations: Embeddings and Random Forest demonstrated promise for similar predicting tasks.

Video Recording

clideo.com



Thank You!

Any Questions?