Advanced topics in machine learning

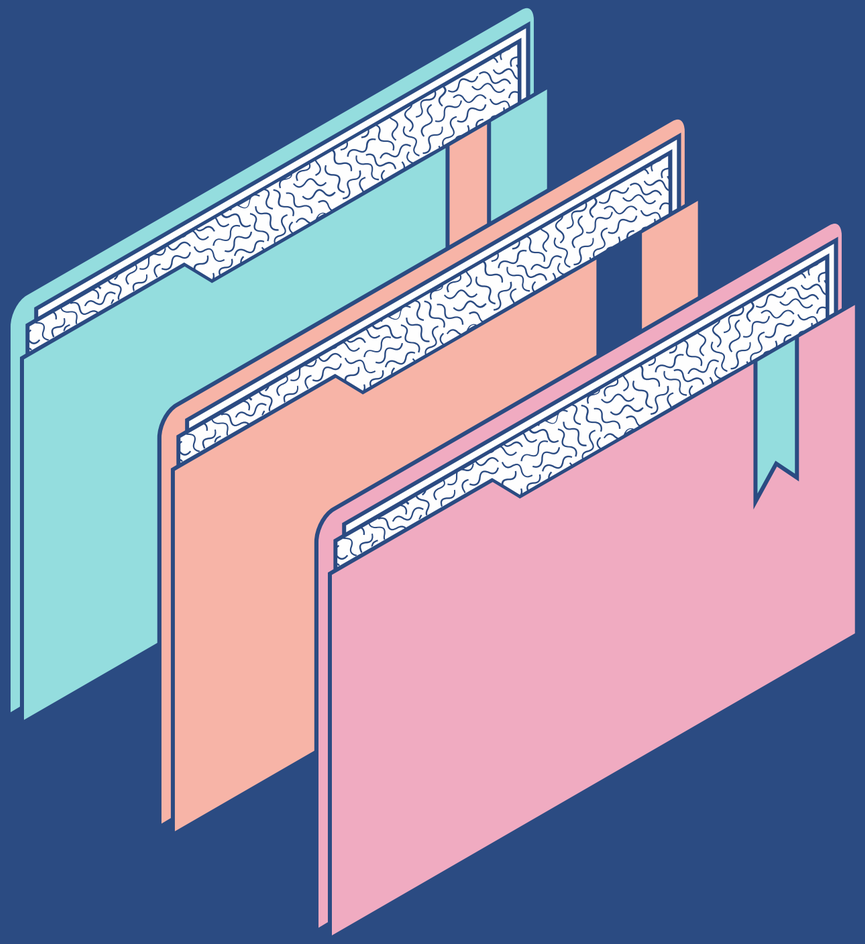
NBA All-Star Prediction

Lecturer:
Chen Hajaj

Liron Ohana

Problem Statement

Predicting who will make it to the All-Star team is not an easy task, as there are many factors influencing the voting process and the final selection.
Moreover, the criteria for selecting All-Stars may vary from year to year, depending on the format, rules, and trends of the game. The dynamic and evolving nature of the environment poses challenges to accurate prediction.



Original goals

- Develop predictive models to forecast NBA player performance based on historical data.
- Identify potential All-Star players using predictive modeling and feature engineering techniques.
- Evaluate the effectiveness of different machine learning algorithms in predicting player performance.
- Investigate the impact of feature engineering and clustering on predictive accuracy and model performance.

These goals aim to enhance the predictive capabilities of NBA players.

Methods and Techniques

- **Exploratory Data Analysis (EDA)**

Analyzed NBA player statistics to understand data distribution and identify patterns.

- **Feature Engineering**

Derived new features from existing data to enhance predictive modeling.

- **Machine Learning models**

Utilized various algorithms such as Logistic Regression, XGBoost, Random Forest, AdaBoost, Gradient Boosting, and LGBM for predictive modeling.

- **Clustering Technique**

Employed KMeans and hierarchical cluster for player segmentation and pattern recognition.

- **Outlier Detection**

Used Isolation Forest , LOF , Knn, DBSCAN, One Class SVM, Kmeans to identify and handle outliers in the dataset.

Dataset and Features

Our Data

Our dataset comprises NBA player statistics from various seasons, gathered through an API library and web scraping. It encompasses performance metrics like points, assists, rebounds, and field goal percentage. Preprocessing steps ensured data consistency, involving cleaning, normalization, and feature selection. Exploratory data analysis identified patterns, while feature engineering techniques like polynomial features and derived metrics enhanced predictive accuracy. The resulting dataset facilitated predictive modeling to forecast player performance and identify potential All-Stars.

	PLAYER_ID	SEASON_ID	TEAM_ID	TEAM_ABBREVIATION	PLAYER_AGE	GP	GS	MIN	FGM	FGA	...	REB	AST	STL	BLK	TOV	PF	PTS	full_name	Year	All_Star
0	76001	1990-91	1610612757	POR	23.0	43	0.0	290.0	55	116	...	89.0	12	4.0	12.0	22.0	39	135	Alaa Abdelnaby	1990	0
1	76001	1991-92	1610612757	POR	24.0	71	1.0	934.0	178	361	...	260.0	30	25.0	16.0	66.0	132	432	Alaa Abdelnaby	1991	0
2	76001	1992-93	1610612749	MIL	25.0	12	0.0	159.0	26	56	...	37.0	10	6.0	4.0	13.0	24	64	Alaa Abdelnaby	1992	0

3 rows × 29 columns

Main Featuers

1. PLAYER_AGE
2. GP (Games played)
3. GS (Games started)
4. MIN (Minutes played)
5. FG_PCT (Field goal percentage)
6. FG3_PCT (Three-point field goal percentage)
7. FT_PCT (Free throw percentage)
8. AST (Assists)
9. STL (Steals)
10. BLK (Blocks)
11. TOV (Turnovers)
12. PF (Personal fouls)
13. PTS (Points scored)
14. Year
15. All_Star (Target variable)

Methodology

1 STEP

Outlier Detection

Utilize visualization to pinpoint outliers and establish a threshold using 3 standard deviations. We employed this approach due to the abundance of atypical values in the dataset, aiming to determine whether deletion is warranted.

2 STEP

Time Series Sorting

Given the dynamic nature of the data, sorted it using time series, with past data for training and future data for testing

3 STEP

Model Evaluation

The evaluation of various models, parameter optimization based on performance, aims to select the most suitable model for predicting NBA player performance using our dataset.

4 STEP

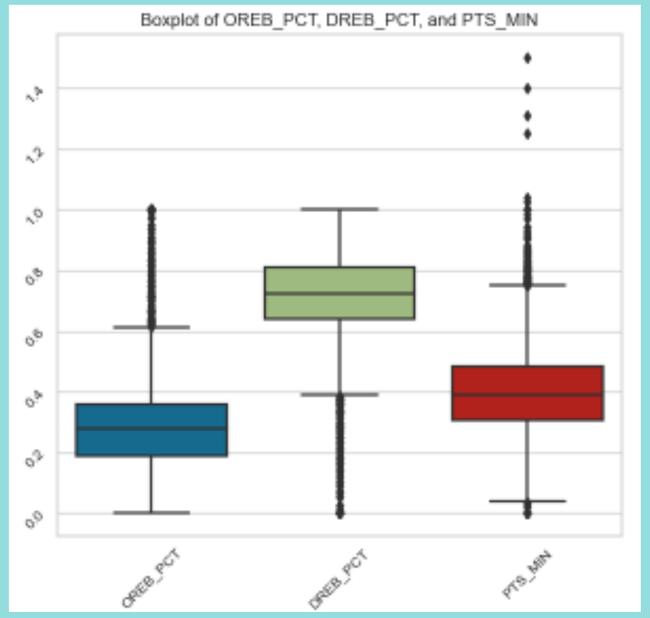
PCA

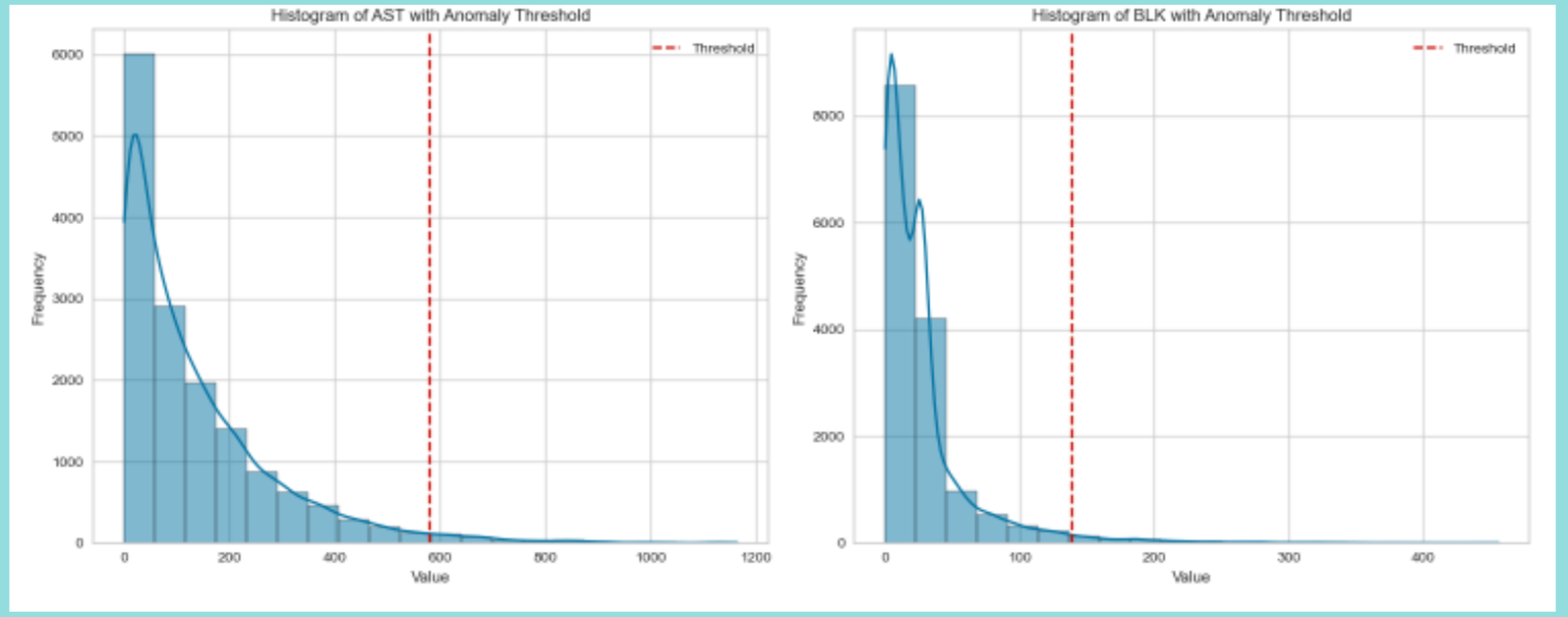
Utilize PCA for dimensionality reduction and dataset optimization. This step was conducted to prepare the data for clustering.

5 STEP

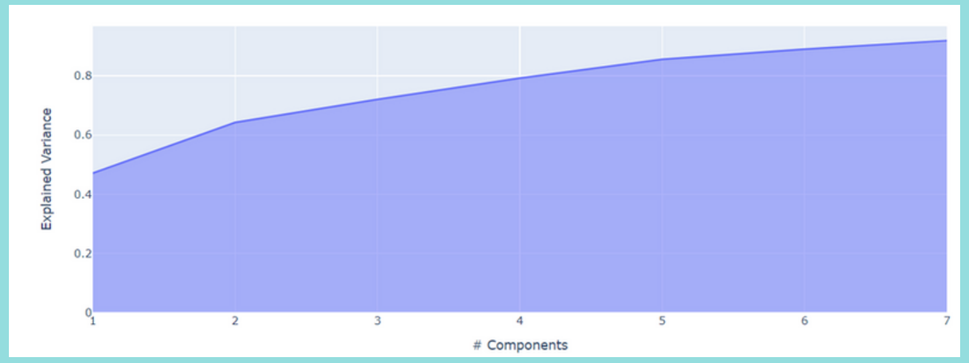
Clustering

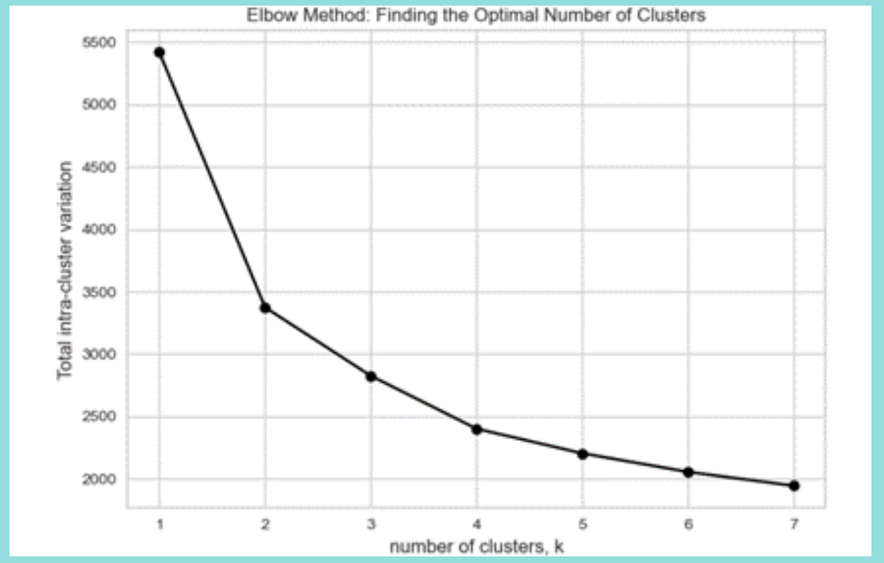
Utilize K-means and hierarchical clustering techniques. We applied these algorithms to assess their impact on prediction accuracy.


Boxplot of OREB_PCT, DREB_PCT, and PTS_MIN
OREB_PCT
DREB_PCT
PTS_MIN


Histogram of AST with Anomaly Threshold
Threshold
Frequency
Value
Histogram of BLK with Anomaly Threshold
Threshold
Frequency
Value

PCA


Explained Variance
Components

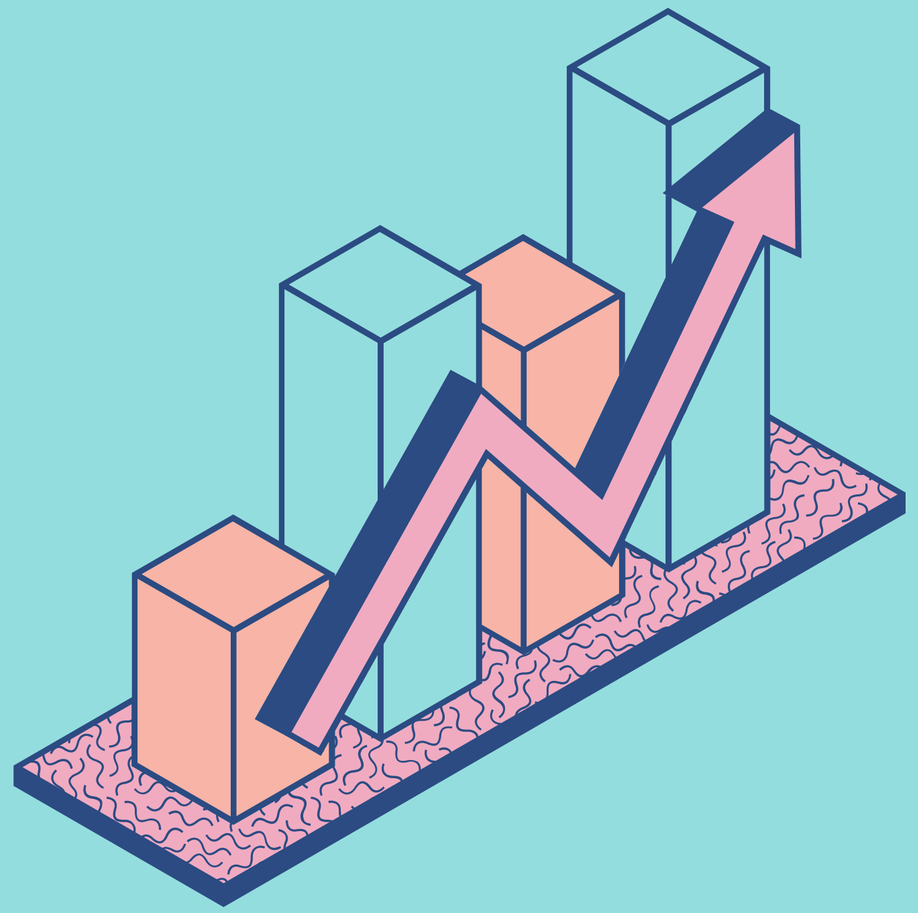

Elbow Method: Finding the Optimal Number of Clusters
Total intra-cluster variation
number of clusters, k

Experiment Description

In our experiment, we meticulously curated NBA player statistics and All-Star selections from historical seasons. After preprocessing the data to handle missing values and normalize features, we trained various machine learning models. We optimized model parameters and utilized PCA for dimensionality reduction, followed by clustering using KMEANS and hierarchical methods. Evaluation of model performance was conducted using metrics like accuracy, precision, recall, and F1-score, with careful attention to adjusting settings and configurations for robust experimentation and accurate assessment.

Evaluation Techniques

The metrics of accuracy, F1-score, recall, and precision are utilized to evaluate the performance of the model in predicting NBA All-Star selections. Accuracy provides a general overview of correct predictions, while F1-score, recall, and precision focus on the model's ability to correctly identify positive cases (All-Star players) and avoid false positives. These metrics are considered suitable for prediction evaluation due to their ability to provide a comprehensive assessment of the model's performance in binary classification tasks like predicting All-Star selections.



Results

THIS TABLE SUMMARIZES THE IMPACT OF USING KMEANS AND HIERARCHICAL CLUSTERING ON THE PERFORMANCE OF VARIOUS MACHINE LEARNING MODELS IN PREDICTING NBA ALL-STAR SELECTIONS.

MODEL	RANDOM FOREST	XGBOOST	GRADIENT BOOSTING	ADABOOST	LOGISTIC REGRESSION	LIGHTGBM
Kmeans	Slight decline	No significant change	No significant change	No significant change	Minor improvement	No significant change
Hierarchical	Slight decline	No significant change	No significant change	No significant change	Minor improvement	No significant change

Comparison of Techniques

KMEANS	HIERARCHICAL CLUSTERING
The inclusion of the cluster feature did not significantly affect the performance of the XGBoost, AdaBoost, Gradient Boosting, and LightGBM models. However, in the logistic regression model and Random Forest, the cluster feature had a minor effect, slightly improving the model's ability to identify All-Star players.	

Conclusion

In summary, our analysis suggests that integrating clustering algorithms to predict NBA All-Star players did not lead to significant enhancements in model performance due to the specific traits of our dataset. Challenges arose from numerous outlier data points, indicating true talent anomalies among players, as well as unbalanced data, which hindered achieving satisfactory accuracy levels. Consequently, relying on this model to evaluate players may not be worthwhile.