Data Science

A logo with a black background

AI-generated content may be incorrect.

A collection of logos of different teams

AI-generated content may be incorrect.

Project Protocol

**Background on the NBA**

The NBA (National Basketball Association) is the premier professional basketball league in the world. It consists of 30 teams, 29 in the United States and one in Canada, and is considered the pinnacle of a professional basketball career. Every year, hundreds of players from around the world aspire to join the NBA, but only a select few succeed. The high level of competition and the demanding physical, mental, and technical requirements make the league an exclusive and prestigious destination.

**Project Objective**

This project aims to evaluate whether a basketball player is likely to reach the NBA based on historical and statistical data. In an era where data analysis plays a crucial role in sports decision-making, this predictive tool can benefit various stakeholders:

* **Scouts and Recruiters** – Assist in identifying high-potential players early in their careers.
* **Professional Teams** – Assess the suitability of players before making long-term commitments and investments.
* **Recruitment Websites and Databases** – Enhance search engines and tailor recommendations for players and coaches.
* **Basketball Enthusiasts and Analysts** – Understand trends and developments in the careers of promising players.

Project Design & Methodology

**Data Journey**

**Data sources:**

The dataset used in this project originates from Kaggle and is named players\_stats\_by\_season\_full\_details.csv.

**Initial Data Analysis**

✅ **Total Records:** 53,949

✅ **Number of Columns (Features):** 34

✅ **Unique Players:** 14,582

✅ **Unique NBA Players:** 1,119

✅ **Players Appearing in More Than One Record:** 9,334

**Data Preparation Steps**

1. **Converting Data Types:**
   * Object-type columns were converted into appropriate formats (string, integer, float, etc.) to enable proper analysis.
2. **Handling Inconsistent Height and Weight Data:**
   * Players with inconsistent height and weight values were identified.
   * The most common value for each attribute was selected.
   * A new indicator column was created to mark whether the value was original or adjusted.
3. **Handling Inconsistent birth year:**
   * Players with inconsistent birth year values were identified.
   * If the birth year was missing, the birth date column was checked, and the birth year was extracted from there.
   * If no birth date was available, the missing data was supplemented using Wikipedia
   * A new indicator column was created to mark whether the value was original or adjusted.
4. **Understanding Player Development Pathway** 
   * Nationality – Determines whether a player followed the international route or the American system.
   * High School – Indicates if a player played in the U.S. high school system, which can impact their chances of receiving a college scholarship and being ranked in the draft.
   * Draft Round – Specifies in which round of the NBA draft a player was selected (first-round picks generally have higher expectations and opportunities).
   * Draft Pick – Represents the exact draft position, with lower numbers indicating earlier selection and potentially greater opportunities.
   * Draft Team – The NBA team that selected the player, which influences their development depending on the team's system and resources.

These attributes are primarily relevant to NBA players, as only those who were drafted have values for Draft Round, Draft Pick, and Draft Team. This distinction played a role in later decisions regarding feature selection for model training

1. **NBA Indicator Creation**
   * An additional column named NBA\_Indicator was created for each row, indicating whether the player has played in the NBA or not.
   * The value is assigned as 1 if the player has played in the NBA (i.e., League = NBA), otherwise 0.
   * If a player appears in multiple records, some of which indicate NBA participation and others do not, the NBA\_Indicator column reflects 1 for rows where League =NBA and 0 otherwise.

**Next Steps**

* The dataset was, prepared for Exploratory Data Analysis (EDA).

**Exploratory Data Analysis (EDA) Summary**

**1. Descriptive Statistics**

* The dataset contains 53,949 records with 34 columns.
* Unique Players: 14,582
* Unique NBA Players: 1,119
* Players Appearing in Multiple Records: 9,334

**2. Data Distribution & Skewness**

* Some features show skewed distributions, particularly:
  + Points (PTS)
  + Assists (AST)
  + Rebounds (REB)
  + Field Goals Made (FGM) & Attempted (FGA)
  + Three-Point Shots Made (3PM) & Attempted (3PA)
* The skewness indicates that a few players have significantly higher stats than the majority.

**3. Missing Values**

* Several columns contain missing values, particularly:
  + Draft-related attributes (only applicable to NBA players)
  + Height and Weight (adjusted based on the most common value)
  + Birth Year (handled by extracting from the birth date or external sources)

**4. Correlation Analysis**

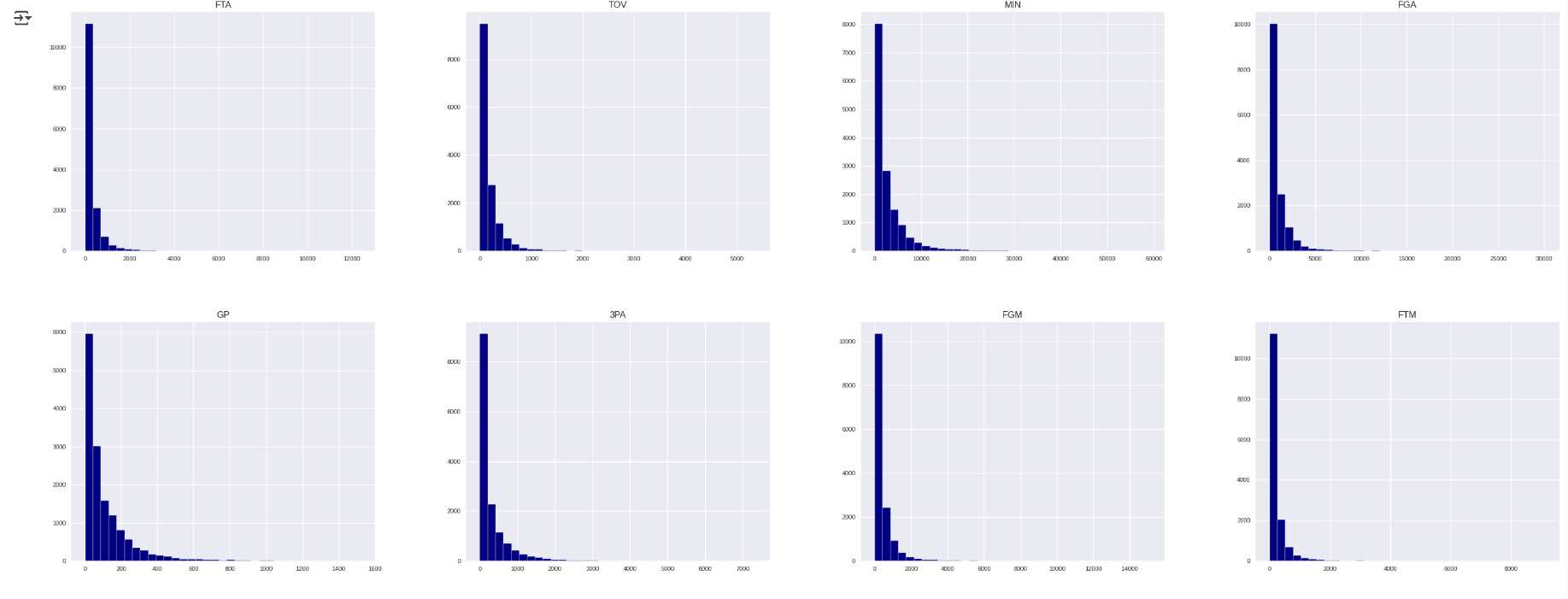
* High correlation detected among:
  + FGM and PTS (Field Goals Made strongly contributes to total points)
  + REB, ORB, DRB (Rebound metrics are highly interrelated)
  + FTM and FTA (Free Throws Made is dependent on attempts)
* Low correlation between Draft Pick and performance metrics, suggesting that being a top draft pick does not always lead to higher performance.

**5. Outliers**

* Outliers identified in key metrics such as:
  + Points per Game (PTS)
  + Minutes Played (MIN)
  + Assists (AST)
  + Turnovers (TOV)
* Outliers were detected using IQR (Interquartile Range) and visualized with boxplots.

**7. Next Steps**

* Further feature engineering to improve prediction accuracy.
* Consideration of normalization or transformation for skewed variables.
* Addressing outliers based on domain knowledge.



A red square with blue squares

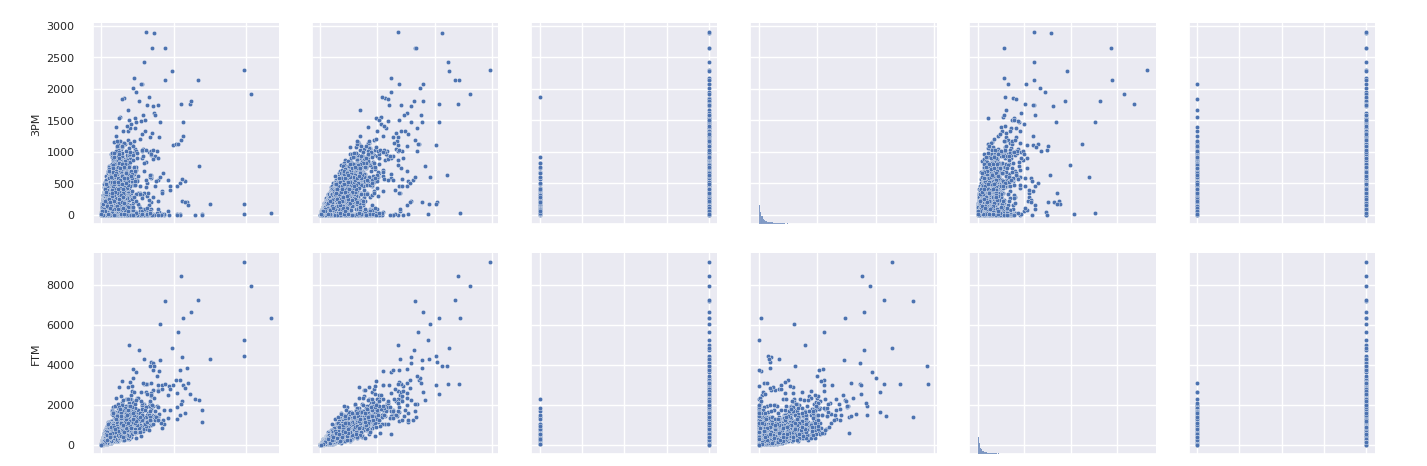
AI-generated content may be incorrect.

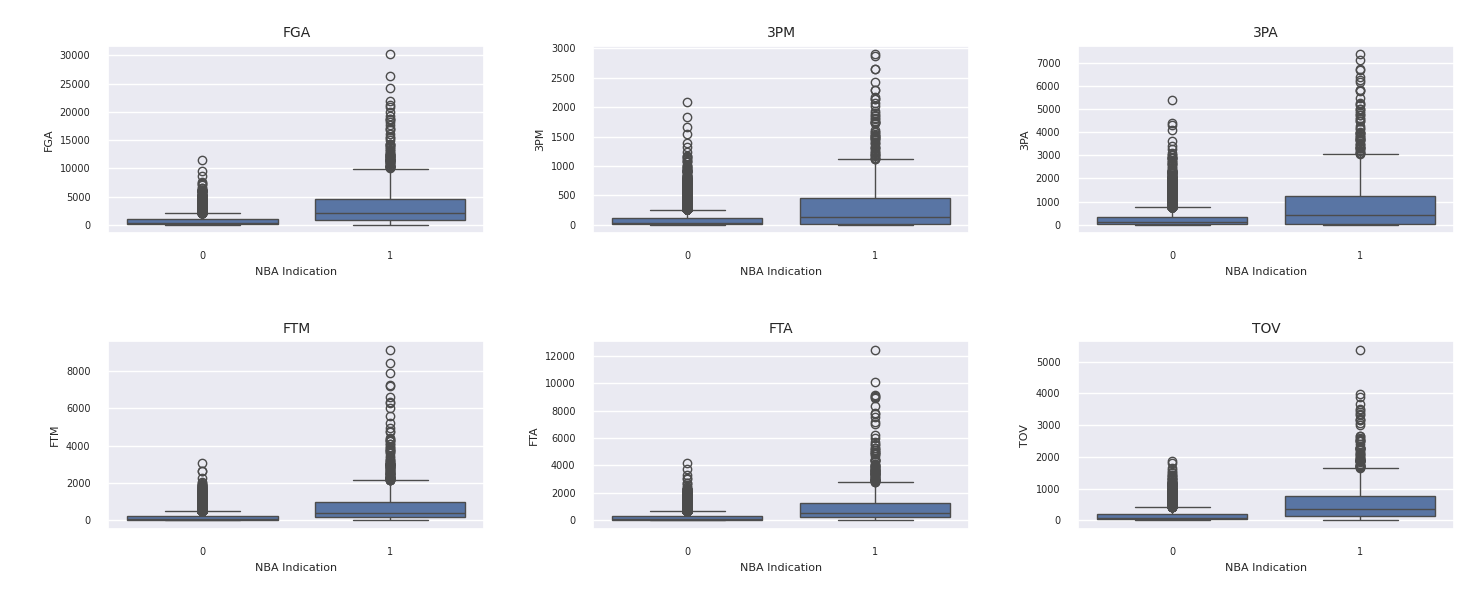
Possible linear relationships between variables (e.g., a potential positive correlation between FTM and 3PM).

Data dispersion – a wide distribution can be observed for some variables.

Anomalies (Outliers) – players with extreme values compared to the rest of the group.

Data accumulation – a high concentration of players with relatively low values in certain statistics.





FGA (Field Goal Attempts) – The number of field goal attempts a player has taken.

3PA (Three-Point Attempts) – The number of three-point shot attempts a player has taken.

FTM (Free Throws Made) – The number of successful free throws made by a player.

FTA (Free Throws Attempted) – The number of free throws a player has attempted.

TOV (Turnovers) – The number of times a player has lost possession of the ball to the opposing team.

Summery

* + The NBA\_Indicator column was determined by taking the maximum value among a player’s records. This ensures that if a player played in the NBA at least once, the indicator is **1**.
  + Other statistical measures were summed, as the data is aggregated in any case.
  + The player’s name was retained only for validation purposes (it will be removed before running the model).
  + birth\_year was retained.
  + birth\_month, birth\_date, height, and weight were removed as they were deemed irrelevant.
  + height\_cm and weight\_kg were kept.
  + Two dummy columns were created to indicate whether the data was original or imputed—though these were ultimately not used.
  + The columns nationality, high\_school, draft\_round, draft\_pick, and draft\_team were removed since they have a strong correlation with NBA players (as expected), making them redundant given the existing NBA\_Indicator column.

**Final Dataset Statistics:**

* From **53,949** records, we reduced the dataset to **14,582** unique players.
* Among them, **1,119** are NBA players.
* The processed data is flattened and ready for the next level outliers

**Summary of Outliers Analysis**

**Objective:** The goal of this notebook is to identify and handle outliers in the dataset using statistical methods. Outliers can significantly impact machine learning models and distort data analysis, making their detection and treatment crucial.

**Methodology:**

1. **Understanding Outliers:**
   * Outliers are extreme values that deviate from the rest of the data.
   * They can result from errors in data collection, rare events, or genuine variability in data distribution.
2. **Detection Methods:**
   * **Interquartile Range (IQR) Method:**
     + Calculated Q1 (25th percentile) and Q3 (75th percentile) for each numeric feature.
     + Defined outliers as values below **Q1 - 1.5 \* IQR** or above **Q3 + 1.5 \* IQR**.
   * **Boxplots:**
     + Visualized feature distributions to spot extreme values.
     + Confirmed anomalies in variables such as scoring statistics and physical attributes.
   * **Z-Score Method (Standardization):**
     + Used the formula: **Z = (X - Mean) / Standard Deviation**.
     + Values with **|Z| > 3** were marked as potential outliers.
3. **Handling Outliers:**
   * **Removal:**
     + Completely removed extreme outliers if they were due to errors.
   * **Capping (Winsorization):**
     + Replaced extreme values with the nearest acceptable limit within **Q1 - 1.5 \* IQR** and **Q3 + 1.5 \* IQR**.
   * **Flagging:**
     + Created a binary indicator column to mark potential outliers without removing them.
4. **Final Dataset Adjustments:**
   * Outliers were handled based on their impact on key features.
   * Data was cleaned while preserving necessary variability for accurate modeling.

**Key Findings:**

* Certain statistics such as points per game (PTS), rebounds (REB), and assists (AST) had extreme values that needed adjustment.
* Physical attributes like height and weight also showed some anomalies, but minimal correction was required.
* Most outliers were concentrated in a small subset of players with significantly higher performance metrics.

**Next Steps:**

* Validate the impact of outlier treatment on model performance.
* Compare different preprocessing techniques to ensure optimal results.
* Proceed with Exploratory Data Analysis (EDA) using the cleaned dataset.

**Summary of Imbalanced Data Analysis**

**Objective:** The goal of this notebook is to analyze and address the issue of imbalanced data, which occurs when certain classes in a dataset are underrepresented compared to others. This imbalance can lead to biased model predictions and reduced generalization performance.

**Methodology:**

1. **Understanding Imbalanced Data:**
   * Imbalanced datasets occur when one class significantly outweighs the others.
   * Standard machine learning algorithms often favor the majority class, leading to poor predictive performance on the minority class.
2. **Exploratory Data Analysis (EDA):**
   * **Class Distribution:**
     + Visualized the distribution of target classes to assess imbalance severity.
   * **Performance Metrics Evaluation:**
     + Accuracy alone is insufficient for imbalanced data.
     + Focused on Precision, Recall, and F1-Score.
3. **Handling Imbalanced Data:**
   * We tested multiple techniques, including Random Oversampling (ROS), Random Undersampling (RUS), and SMOTETomek.
   * SMOTE was selected based on the best balance between Precision and Recall.

* Best F1-Score (0.607): Ensures a good tradeoff between Precision and Recall.
* Higher Precision (0.483) than other techniques, reducing false positives
* Strong Recall (0.817), meaning more minority class instances were correctly classified
* Outperformed Random Oversampling (ROS) and Random Undersampling (RUS)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Technique** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| ROS | 0.908 | 0.448 | 0.839 | 0.584 |
| RUS | 0.911 | 0.455 | 0.821 | 0.586 |
| SMOTE | 0.919 | 0.483 | 0.817 | 0.607 |
| SMOTETomek | 0.917 | 0.477 | 0.835 | 0.607 |

* + Algorithmic Approaches:
    - Implemented cost-sensitive learning, adjusting model loss functions to penalize misclassification of the minority class.
    - Explored ensemble methods such as balanced random forests.
  + **Fine-Tuning:**
    - Hyperparameter optimization was applied to improve model performance while considering class imbalance.
    - Balanced class weighting was incorporated into model training.

1. **Final Model Performance:**
   * Evaluated models before and after applying balancing techniques.
   * Achieved improved recall and F1-score for the minority class without significantly compromising overall accuracy.

**Key Findings:**

* The dataset had a severe class imbalance, leading to biased model predictions.
* Using SMOTE and undersampling improved minority class representation.
* Cost-sensitive learning and class weighting significantly enhanced model performance.
* Fine-tuning hyperparameters further optimized results.

**Comparing Multiple Models on Imbalanced Data**

**Defines features (X) and target (y).**

**Splits the dataset into:**

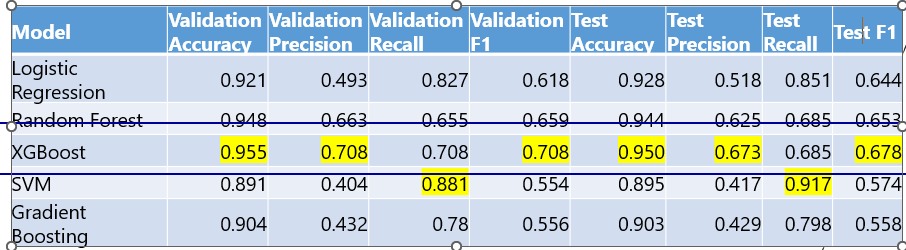
* **Training set (70%)**
* **Validation set (15%)**
* **Test set (15%)**

**Training & Evaluating Multiple Models**

* **Defines five different models:**
  + **Logistic Regression**
  + **Random Forest**
  + **XGBoost**
  + **Support Vector Machine (SVM)**
  + **Gradient Boosting**

**Loops through each model to:**

* **Train on the resampled dataset.**
* **Make predictions on validation and test sets.**
* **Compute evaluation metrics:**
  + **Accuracy**
  + **Precision**
  + **Recall**
  + **F1-score**

****

**Hyperparameter Tuning with GridSearchCV**

* Defines a **grid of hyperparameters** for **XGBoost**, including:
  + **Number of trees (n\_estimators)**
  + **Tree depth (max\_depth)**
  + **Learning rate (learning\_rate)**
  + **Subsampling ratio (subsample)**
  + **Feature sampling (colsample\_bytree)**
* Uses **GridSearchCV** with **5-fold cross-validation** to find the best combination.
* Evaluates models based on multiple metrics:
  + **Accuracy**
  + **Precision**
  + **Recall**
  + **F1-score**
* Selects the **best model** based on accuracy.

A close-up of a number

AI-generated content may be incorrect.