

Assignment 1

Kaggle Competition

Kritika Dhawale
Student ID: 24587661
September 6, 2023

36120 - Advanced Machine Learning Application
Master of Data Science and Innovation
University of Technology Sydney

Table of Contents

1. Executive Summary	2
2. Business Understanding	3
a. Business Use Cases	3
3. Data Understanding	5
4. Data Preparation	10
5. Modeling	12
a. Approach 1	12
b. Approach 2	12
c. Approach 3	12
6. Evaluation	13
a. Evaluation Metrics	13
b. Results and Analysis	13
c. Business Impact and Benefits	14
d. Data Privacy and Ethical Concerns	14
7. Deployment	16
8. Conclusion	17
9. References	18

1. Executive Summary

A machine learning experiment was carried out for this project in order to predict NBA draft picks as part of a Kaggle competition. This project holds significance in the domain of sports analytics, as it can provide valuable insights for player selection and team management in professional basketball. The primary goal was to develop a model that could correctly classify players as "drafted" or "not drafted" based on performance metrics and other relevant characteristics.

In the context of professional basketball, the process of player selection, particularly in the annual draft, plays a pivotal role in team building and long-term success. Teams spend a lot of money and time evaluating and selecting players who have the potential to make a big difference on their rosters. To enhance this selection process, the project sought to develop a machine learning model that could predict whether a player is likely to be drafted or not based on various player attributes, including statistics, physical characteristics, and performance metrics.

The developed outcomes include the development of a classification model that predicts the likelihood of college players being drafted and proves to be quite useful in meeting the project's business objectives. The model's accuracy was assessed using the AUROC score, which was found to be greater than 0.9, indicating that the model's predictive power outperformed random chance. These results suggest that the developed models can assist basketball teams and talent scouts in the player selection process by providing data-driven insights into a player's draft prospects.

In summary, this project offers a useful resource for the basketball industry by offering a data-driven method of player evaluation and selection.



2. Business Understanding

a. Business Use Cases

The project is primarily applied in professional basketball for player selection, team management, and talent scouting. There are several business use cases and scenarios:

- The annual player draft is a crucial professional basketball decision. Teams seek talented players to help them succeed. The project's predictive model can help teams evaluate draft prospects and make data-driven decisions.
- Player Evaluation: After the draft picks, teams must evaluate players for trades, signings, and roster changes. The model helps trade negotiations and free-agent signings by objectively evaluating players' performance metrics.
- Talent Scouting: The model helps talent scouts and recruiters find promising players from different leagues and regions. This increases teams' roster options, potentially revealing hidden gems.
- Player analytics can optimize player rotations, game plans, and tactics for teams. Data-driven decisions can help coaches match teams' strengths and weaknesses.

The project was driven by various challenges and opportunities:

- Basketball data is complex, including player statistics, physical attributes, and performance metrics. Complexity requires advanced machine learning techniques for management and analysis.
- Competitive Advantage: In the highly competitive sports industry, an edge is essential. The project's predictive model helps teams make better player-related decisions, giving them an edge over competitors.
- Talent Identification: Assessing players from diverse backgrounds and leagues is difficult. Machine learning can reveal patterns and attributes that traditional scouting may miss.
- The project supports the growing trend of data-driven sports decision-making. Teams and organizations are increasingly using analytics to make strategic and tactical decisions, and the project contributes.
- Scalability: The model can be applied to college and professional basketball leagues to streamline talent evaluation.



b. Key Objectives

1. Player Selection Improvement:

To create a predictive model that predicts professional basketball draftees to improve player selection. Basketball teams and talent scouts need a more data-driven and efficient draught prospect assessment. -> The project creates a machine-learning model that predicts draft outcomes using player attributes and performance metrics.

2. Data-Driven Decisions:

To enable data-driven player selection, team management, and talent scouting. Analytics and data-driven insights give teams and organizations an edge. -> Machine learning algorithms analyze historical player data to provide objective insights to help teams make decisions.

3. Versatility and Scalability:

Develop a scalable, adaptable model for basketball leagues and age groups. Scalability is needed to accommodate diverse league and age pools. -> The machine learning model is adaptable and scalable, making it suitable for basketball applications.

Stakeholder Needs:

- To build competitive basketball teams, teams need accurate player assessments for draft picks, trades, and roster management.
- Talent Scouts and Recruiters: Talent scouts look for promising players from diverse backgrounds and leagues.
- Coaches: Data-driven insights help coaches create strategies and tactics that match their team's strengths and weaknesses.

Meeting Stakeholder Needs:

The project uses historical player data to predict draft outcomes with a machine learning model to satisfy stakeholders. It facilitates stakeholder collaboration and data-driven decision-making by providing a common framework for player evaluation. The model's scalability makes it useful for talent assessment and team management across basketball leagues and age groups. The project uses data-driven player selection and management to give basketball teams an edge.



3. Data Understanding

a. Data Overview

The dataset used for the project was obtained from Kaggle and contained information on college basketball players' performance statistics, physical attributes, and other relevant features. The data was collected from various sources, including the NCAA website, ESPN, and other basketball-related websites.

The dataset contained 64 variables with 56091 data samples, including the player's ID, team, number of games played, height, year of study, and various performance statistics such as points per game, rebounds per game, assists per game, and other relevant metrics. The dataset also included a binary variable indicating whether the player was drafted or not. The figure below shows a sample of the first 5 rows of the train set.

GP	Min_per	Ortg	usg	eFG	TS_per	ORB_per	DRB_per	...	dgbpm	oreb	dreb	treb	ast	stl	blk	pts	player_id	drafted
26	29.5	97.3	16.6	42.5	44.43	1.6	4.6	...	-1.941150	0.1923	0.6154	0.8077	1.1923	0.3462	0.0385	3.8846	7be2aead-da4e-4d13-a74b-4c1e692e2368	0.0
34	60.9	108.3	14.9	52.4	54.48	3.8	6.3	...	-0.247934	0.6765	1.2647	1.9412	1.8235	0.4118	0.2353	5.9412	61de55d9-1582-4ea4-b593-44f6aa6524a6	0.0
27	72.0	96.2	21.8	45.7	47.98	2.1	8.0	...	-0.883163	0.6296	2.3333	2.9630	1.9630	0.4815	0.0000	12.1852	efdc4cfc-9dd0-4bf8-acef-7273e4d5b655	0.0
30	44.5	97.7	16.0	53.6	53.69	4.1	9.4	...	-0.393459	0.7000	1.4333	2.1333	1.1000	0.5667	0.1333	4.9333	14f05660-bb3c-4868-b3dd-09bcd64279d	0.0
33	56.2	96.5	22.0	52.8	54.31	8.3	18.6	...	-0.668318	1.4242	3.3030	4.7273	0.8485	0.4545	0.3333	7.5758	a58db52f-fbba-4e7b-83d0-371efcfed039	0.0

Fig. 1 The first 5 samples of the Train set

b. Data Limitations:

- **Data Completeness:** The dataset contains missing values or incomplete records, which can impact the quality of analyses and modeling.
- **Data Accuracy:** Data accuracy is critical in sports analytics. Errors in player statistics or attributes can lead to inaccurate conclusions.
- **Temporal Variability:** Player performance can change significantly over time, and the dataset may not capture recent developments in player careers.

c. Features and their Significance

The dataset contains many variables with different levels of significance:

1. Player Statistics: Number of games played, TRB - Total Rebounds, AST - Assists, STL - Steals, BLK - Blocks, PTS - points, and shooting statistics (dunks made and dunks missed) are the key performance metrics. These statistics are essential for evaluating a player's play.
2. Height and Career stage (college year) are crucial for assessing a player's physical readiness and potential.
3. Recruiting rank: What the player was ranked as a recruit coming out of high school can help in choosing top-ranked players.
4. Draft Status: The target variable "drafted," is crucial because it indicates whether a player was drafted in an NBA league or not. This variable controls project predictive modeling.

Overall, the dataset provided a comprehensive set of features that were relevant to the project's objectives.

d. EDA

An exploratory data analysis was conducted to understand the data better. The analysis revealed that the dataset contained missing values, duplicates, outliers, and other data quality issues that needed to be addressed.

1. Plot the missing value count plot for each variable: It shows that the pick and Recruiting rank have the highest number of missing values, followed by dunks ratio (ratio between dunks made and dunks missed). This visualization led to the handling of missing values.

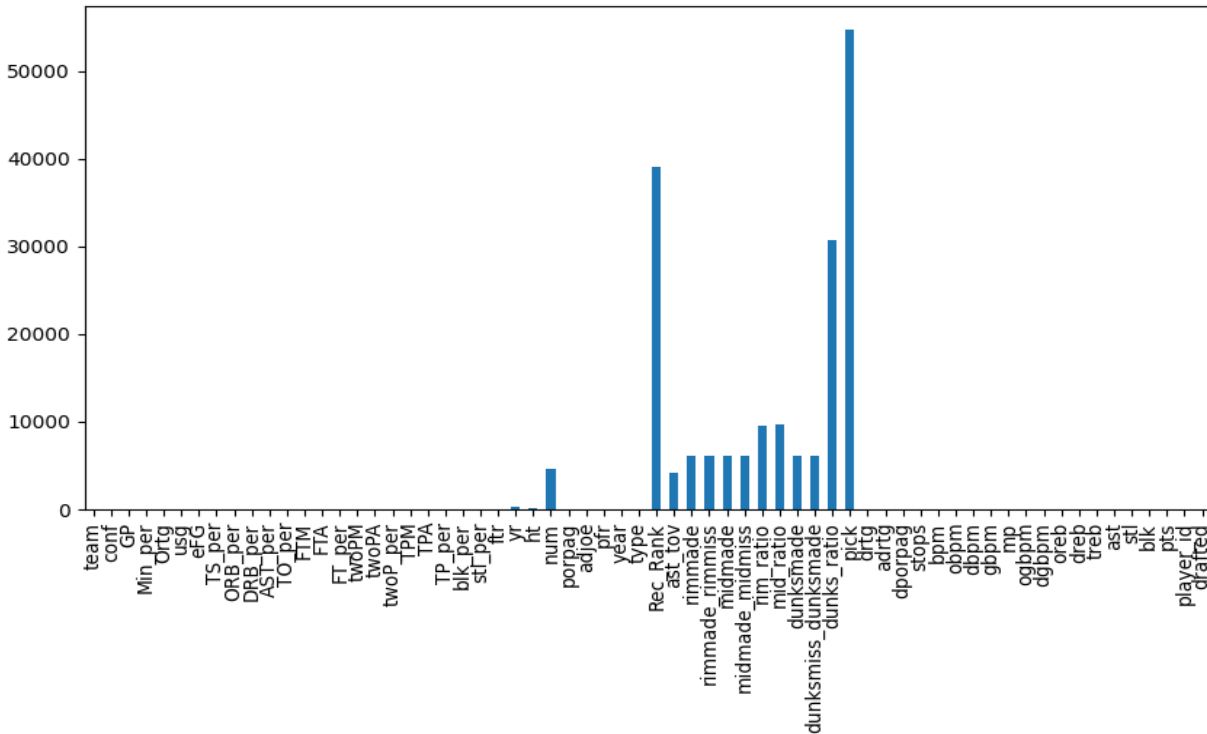


Fig. 2 Count plot for all missing values in each column of the training dataset

2. See the distribution of the target variable: Only ~0.9% of the players have been drafted into NBA league out of the total data samples given. This leaves room for data oversampling and splitting data into train/val correctly.

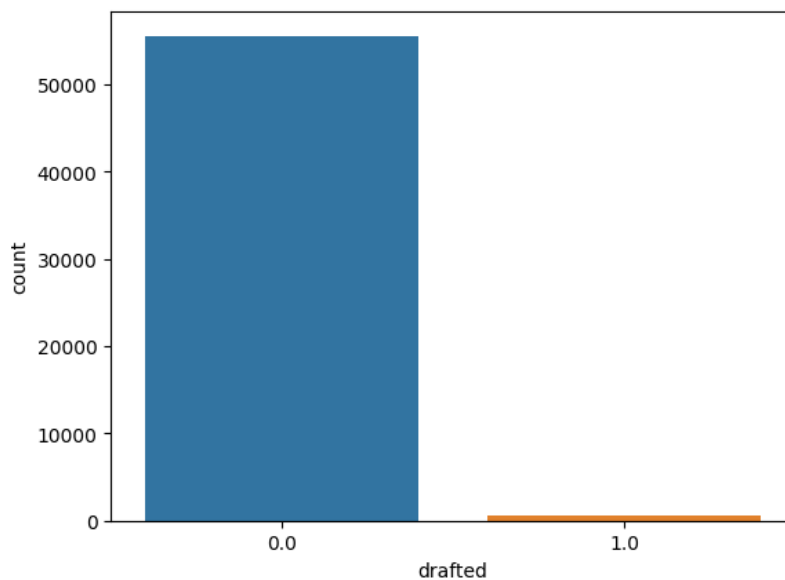


Fig. 3 Distribution graph of “drafted” variable

3. Check for any duplicate player IDs: It was found that there were 32162 duplicates in the “player_id” column.
4. Height of players: It was seen from the data file that the height values are given in date format, as shown in the table below. At first, we thought it might be incorrect data that is put in the height column but later realized the format of the data was wrong.

ht
2-Jun
4-Jun
8-Jun
1-Jun
Jun-00

Table 1. The height variable in the dataset

5. Checking for outliers in the “ast_tov” column: It showed that there are many outliers in the column that need to be removed.

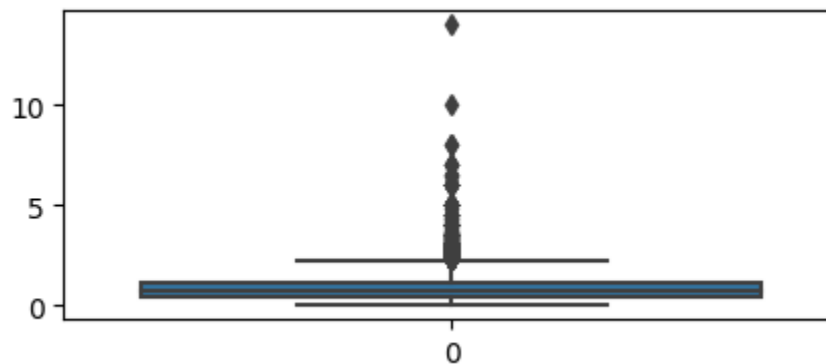


Fig. 4 Box plot for ast_tov column values

6. Distribution of the “ast_tov” column: To handle the missing values in the column, we saw that it was a right-skewed distribution, so the median should be used to replace the missing values.

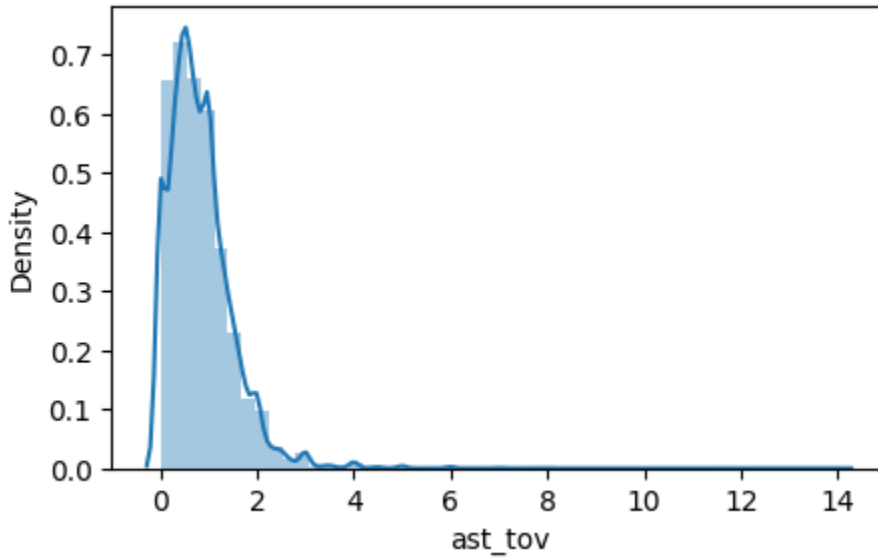


Fig. 5 Distribution of the ast_tov column

7. See the categories of Types of metrics displayed: It was noticed that all the players had the same type “all”. This means that the type column does not have any importance in predictive modelling, hence should be removed.

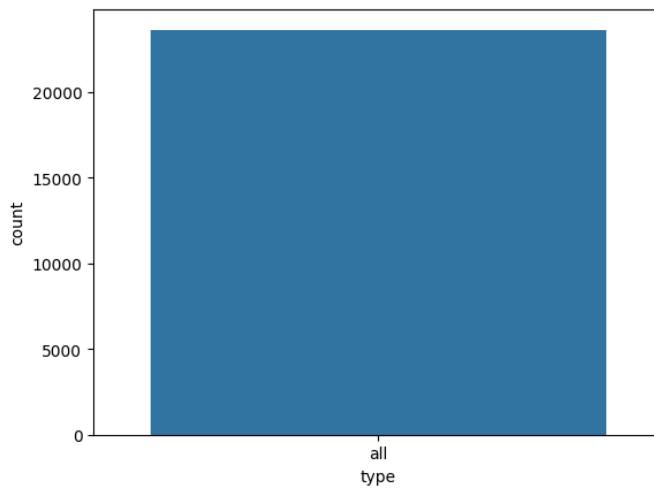


Fig. 6 Count plot for column type

4. Data Preparation

We performed different data preprocessing methods to see which features were relevant and needed engineering to achieve a high-performance score in data modelling.

a. Approach 1

Data Cleaning:

1. Removed 32162 duplicates from the data (repeated player IDs)
2. Removed high missing value columns and unnecessary features: From the metadata, we saw that the player number (num) is not imp and "ht" (height of the player) is incorrect. Similarly, columns with a very high percentage of missing values were removed.
3. Replaced Nan values with 0.
4. Repeated the same steps for test set.

Data Preprocessing:

1. Separated numerical and categorical columns to perform standard scaling and one hot encoding, respectively.
2. Repeated the same steps for test set.
3. Data Split: Split the train set into train and val with a ratio of 30% (used stratify to create unbiased split)

b. Approach 2

Data Cleaning:

1. Removed duplicates
2. Dropped columns that have more than 15% missing values and unnecessary features.
3. Filled Nan values with 0.
4. Replaced missing values in `ast_tov` column with median after removing outliers.

Feature Engineering: `ht` column was in the wrong format (date). Ex: 05-Jun meant 6'5. Month represents foot and date inches. Converted date format to numeric handling all exceptional cases through a function `convert_height_numeric`. Calculated height in inches and saved in a new column `ht_inch`. Dropped the original height column `ht`

Data Preprocessing:

1. Separated numerical and categorical columns to perform standard scaling on numerical features. There was only one categorical column left i.e `type`.
2. After visualizing the type column, there was only one value; “all”. Dropped the `type` column, so there was no need to use one hot encoding.
3. Repeated the same steps for test set.
4. Data Split

c. Approach 3

Here, we repeated the data preparation steps used in Approach 2, but with no numerical features dropped or any outliers removed. All the numeric variables added value to the probability of a player being drafted, while the categorical variables did not have much impact (This was experimented in week 2 and week 3, and the results of the latter were better). On top of this, we applied the [Synthetic Minority Oversampling Technique](#) (SMOTE) to balance the data. SMOTE does not simply reproduce the minority class but synthesizes the samples to produce similar data values. With this, we were successful in oversampling the “drafted” samples.



5. Modeling

a. Approach 1

We used [Logistic Regression](#) to start off with a basic algorithm. It is a simple and interpretable model, and its strong performance could suggest that the data might have a relatively linear decision boundary.

Hyperparameters: As the data was large, we had to maximize the number of iterations (default was 100). Set it to 2000.

b. Approach 2

We tried three models to test the type of relationship between the features and the target variable.

1. [Logistic Regression](#) (works well when there is a linear relationship)
2. [Random Forest Classifier](#) (can capture both linear and non-linear relationships in the data, controls over-fitting)
3. [Gaussian Process Classifier](#) (a kernel that can model complex non-linear relationships). In this method, the training time was high (4.7126 min)

c. Approach 3

On the balanced dataset, three models - [Logistic Regression](#), [Random Forest](#) and [Adaboost](#) were trained to see if upsampling has really improved the results. Reason to use these models were to see if the data exhibits linear or non-linear relationships and does ensembling help in improving the performance.

[Adaboost](#) fits whole data and copies of subsets with weight adjustments for incorrectly classified instances to learn better in future (Represented by learning rate).

- **Hyperparameters:** changed the learning rate to 0.05 (so that the model slowly moves towards the optimal weight and do not miss it by taking a large jump, default = 1)



6. Evaluation

a. Evaluation Metrics

AUC-ROC (Area Under the Receiver Operating Characteristic Curve):

Why Chosen: AUC-ROC is a powerful metric for binary classification models. It measures the area under the ROC curve, which represents the trade-off between true positive rate (recall) and false positive rate across different thresholds.

Relevance to Project Goals: AUC-ROC provides insights into how well the model can distinguish between drafted and non-drafted players across various decision thresholds. It is relevant because it assesses the model's discrimination capability; also, it was given in the assessment description to use this metrics.

We evaluated the model on Validation set and Test set (through kaggle).

b. Results and Analysis

The best results were obtained from week 3 experiments mainly due to balancing the drafted players data using SMOTE, and we will discuss those results.

Part A: Logistic regression model

- Val set - 0.9964
- Test set - 0.9952

Part B: Random Forest model

- **Val set - 0.99997**
- **Test set - 0.99967**

Part C: Adaboost model

- Val set - 0.9977
- Test set - 0.9981

The **random forest** model performs exceptionally well on both the validation and test sets, hence is picked as the best model. It achieves extremely high AUCROC score almost equal to 1 (fig 7), which suggests that it's capturing complex relationships in the data. Nevertheless, it is important to check if the model's performance is consistent across different unseen datasets or if it's mainly memorizing the training data.

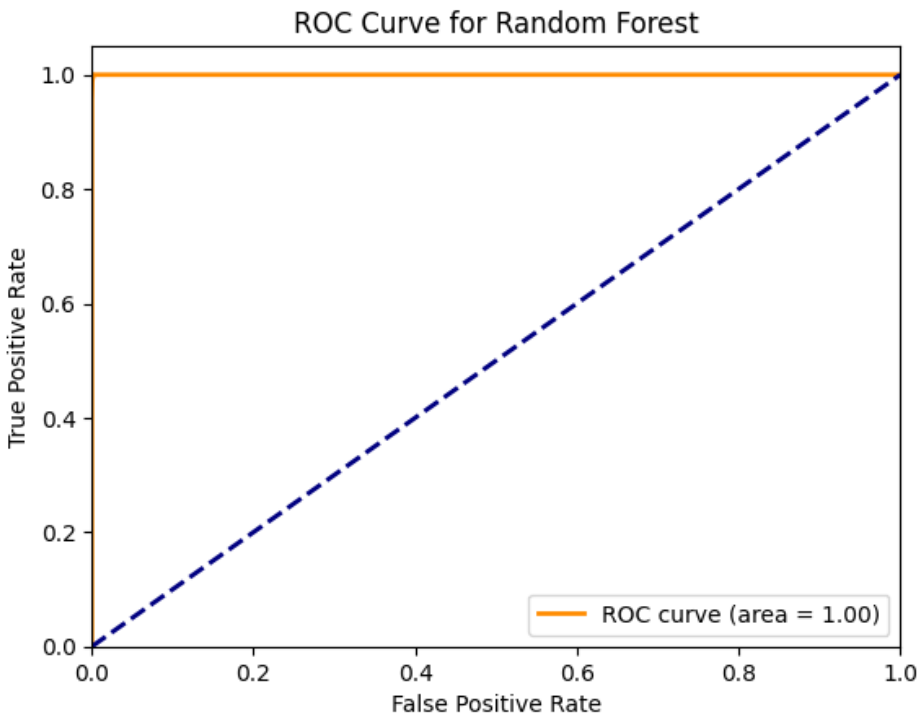



Fig. 7 Area under the ROC curve for the Random forest model on validation set

c. Business Impact and Benefits

This could be presented to our stakeholders as we have achieved astonishing results on both the sets. The kaggle leaderboard also shows that this is the [best result](#) amongst all other teams. This final predictive model has the potential to revolutionize player selection in professional basketball, offering significant benefits in terms of accuracy, efficiency, and strategic advantage. It addresses the challenges associated with identifying talent in a highly competitive and dynamic environment, ultimately contributing to the success and sustainability of basketball teams.

d. Data Privacy and Ethical Concerns

This project maintains a strong focus on data privacy and ethical concerns. It takes care with sensitive player data, anonymizing individual identities by player IDs. These safeguards protect players' privacy while still allowing for effective evaluation.



Several ethical concerns can be raised such as bias and discrimination that can have an impact on certain groups of individuals or society as a whole if data privacy is not maintained properly. The model may perpetuate biases if certain groups of players are systematically favored or disadvantaged.

To uphold data privacy, data encryption and strict access controls are used to prevent unauthorised access to sensitive information. Data retention policies follow regulatory guidelines, ensuring that data is only kept for as long as is necessary. Furthermore, the project is subjected to ethics reviews, which evaluate potential biases and fairness concerns and address them thoroughly.



7. Deployment

While deploying the trained random forest model, we will need to follow several steps:

1. Save the trained model in a deployable format
2. Containerize models with Docker for cross-platform execution.
3. Develop a RESTful API for remote prediction.
4. Scalability: Load balance and containerize high-volume usage.
5. Preprocessing: Transform data consistently during deployment.
6. Monitor and log model performance, data statistics, and errors.

Challenges and Considerations:

1. Manage data drift with monitoring and retraining.
2. Maintain multiple model versions for backward compatibility.
3. Secure the model with access controls and authentication.
4. Ethics: Check for biases and follow rules.
5. Legal Compliance: Follow data protection laws.
6. Provide thorough usage documentation.



8. Conclusion

This project was a success in terms of leveraging machine learning to improve player selection in the sports domain. The developed models, which included logistic regression, random forest, Adaboost, etc. performed exceptionally well in predicting player draught outcomes. They provide valuable insights for talent scouting due to their high validation and test set accuracy scores.

Not only did the project address the initial challenges and opportunities in player selection, but it also contributed to data-driven decision-making in the sports industry. It demonstrated how machine learning can be used to improve team performance and resource utilisation by optimising player recruitment.

Future research should concentrate on continuous model monitoring in order to address data drift and ensure model fairness. Furthermore, we can definitely look into feature importances using random forest technique. We could also look into additional variables for ex: teamwork, personality, and coachability that influence decisions about a player being drafted.

Overall, this project establishes a solid foundation for incorporating machine learning into sports management and scouting processes.



9. References

- [1] McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
- [2] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- [3] Waskom, M., Botvinnik, Olga, Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, ... Qalieh, Adel. (2017). *mwaskom/seaborn: v0.8.1* (September 2017). Zenodo. <https://doi.org/10.5281/zenodo.883859>
- [4] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.
- [5] Pedregosa, F., Varoquaux, Gaël, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
- [6] Van Rossum, G. (2020). The Python Library Reference, release 3.8.2. Python Software Foundation.
- [7] Joblib Development Team (2023) *Joblib/joblib: Computing with python functions.*, GitHub. Available at: <https://github.com/joblib/joblib>.

