Report

Kaggle Competition

By: Ratana Sovann

Github	Experimentation Repo: link to repo Custom Package Repo (group): link to repo

36100 – Data Science for Innovation
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Executive Summary

The objective of the project is to develop binary classification models to predict whether a college basketball player will be drafted into the NBA. By accurately forecasting draft status, the model aids NBA teams, coaches, and data analysts in evaluating player potential. The project aimed to leverage advanced ML models to enhance the accuracy of draft predictions using AUROC as the assessing metric.

The project addresses the challenge of predicting NBA draft outcomes (only 0.1% of players are drafted), a critical decision-making process for teams aiming to scout talent from college basketball programs. The project focuses on historical player data, including game stats and physical attributes to overcome this challenge.

Our team of 4 have built multiple machine learning models including Logistic Regression, KNN, XGBoost, Random Forest and Gradient Boosting out of which the Gradient Boosting algorithm provided the best results. The following report only showcase the best model.

2. Business Understanding

a. Business Use Cases

The objective of this project is to predict whether a college basketball player will drafted into the NBA based on their statistics. Some of the business use cases are stated below:

● **NBA Scouting and Recruitment**: NBA invests heavily in scouting and using this model will enhance their recruitment on top of human judgement. The model may serve as positive or negative indicators of drafting a given player.

● **Player Development Program:** Colleges can utilize this model to determine which player has greater potential in making it to the drafts allowing them to focus more on the players with higher probability of success through custom training programs.

● **Player Contract Negotiations:** This model will allow agents and player representatives to secure better contracts from teams by emphasizing the player's capabilities compared to the rest of the talent pool.

Challenges and Opportunities:

CHALLENGES	OPPORTUNITIES
• **Imbalanced Data:** Working with a target that is highly skewed towards undrafted players can lead to misleading accuracy. • **External Factors:** Draft decisions are influenced by factors beyond stats, such as team needs and player marketability, making the predictive task more complex.	• **Scaleability:** The model can be scaled beyond NBA to analyse players in other leagues offering a broader perspective on international talent. • **Insights into Key Metrics:** The model can reveal which statistics are most predictive of draft success, offering deeper insights into player performance that can be used for future training and development.

Relevance of Machine Learning:

Machine learning is crucial for this situation since it will be very difficult to analyse such a large pool of talent with so many features through manual analysis. Most importantly relationships between the features with the target variable might not be linear in nature so determining potential players manually might be misleading. Moreover, using complex machine learning models will allow the stakeholders to deal with such an imbalanced dataset which emphasizes the model's ability to discriminate between classes.

b. Key Objectives

The project aims to build multiple classification models using the CRISP-DM Framework and agile methodologies to find which college basketball players are likely to be drafted

into the NBA. The metric against which we will be assessing the models is AUROC which we will try to maximise while minimising the level of overfitting. The goal is to create a robust and generalised machine learning model to handle a highly imbalanced dataset. Additionally, the aim is to get the highest score on Kaggle to climb up the leaderboard.

Stakeholders	Role	Influence	Requirement	Engagement Strategy
NBA Teams	Decision-makers on player drafts	High	Accurate predictions on potential draftees based on player stats	Regular updates on model accuracy and insights
Sports Analysts and Data Scientists	Develop and refine the model	High	Access to player data, performance metrics, and tools for analysis	Continuous communication with stakeholders to refine model requirements, test results, and implement feedback
College Basketball Coaches	Provide insights into player performance	Medium	Insights into how player stats reflect draft potential	Present model outcomes with interpretable explanations
Players	Subjects of the predictions	Low	Understanding how their stats impact draft potential	Offer transparency on key performance indicators

Fig: Stakeholder Analysis

The stakeholder analysis shows the stakeholder requirements and the corresponding strategies using the model to meet these requirements. By analysing player statistics, the model identifies KPIs that influence draft outcomes providing valuable insights to coaches, analysts and players allowing targeted improvements and strategic planning. The use of AUROC ensures that the model's predictions are accurate and reliable even with imbalanced data.

3. Data Understanding

This dataset is provided as part of a Kaggle competition without specifying the source acquired. It contains many years of NBA statistics for 20377 players in the training set and 4970 players in the testing set.

Feature Name	Data Type	Missing Data (Abs / %)
team	object	0 / 0.0%
conf	object	0 / 0.0%
GP	int64	0 / 0.0%
Min_per	float64	0 / 0.0%
Otrg	float64	0 / 0.0%
usg	float64	0 / 0.0%
eFG	float64	0 / 0.0%
TS_per	float64	0 / 0.0%
ORB_per	float64	0 / 0.0%
DRB_per	float64	0 / 0.0%
AST_per	float64	0 / 0.0%
TO_per	float64	0 / 0.0%
FTM	int64	0 / 0.0%
FTA	int64	0 / 0.0%
FT_per	float64	0 / 0.0%
twoPM	int64	0 / 0.0%
twoPA	int64	0 / 0.0%
twoP_per	float64	0 / 0.0%
TPM	int64	0 / 0.0%
TPA	int64	0 / 0.0%
TP_per	float64	0 / 0.0%
blk_per	float64	0 / 0.0%
stl_per	float64	0 / 0.0%
ftr	float64	0 / 0.0%
yr	object	292 / 0.6%
ht	object	64 / 0.1%
num	object	4,679 / 10.0%
porpag	float64	0 / 0.0%
adj_oe	float64	0 / 0.0%
pfr	float64	0 / 0.0%
year	int64	0 / 0.0%
type	object	0 / 0.0%
Rec_Rank	float64	32,281 / 69.2%
ast_tov	float64	3,467 / 7.4%
rimmade	float64	5,758 / 12.4%
rimmade_rimmiss	float64	5,758 / 12.4%
midmade	float64	5,758 / 12.4%
midmade_midmiss	float64	5,758 / 12.4%
rim_ratio	float64	8,499 / 18.2%
player_id	object	0 / 0.0%

ast	float64	36 / 0.1%
mid_ratio	float64	8,568 / 18.4%
dunksmade	float64	5,758 / 12.4%
dunksmiss_dunksmade	float64	5,758 / 12.4%
dunks_ratio	float64	26,218 / 56.2%
pick	float64	45,411 / 97.4%
drtg	float64	42 / 0.1%
adrtg	float64	42 / 0.1%
dporpag	float64	42 / 0.1%
stops	float64	42 / 0.1%
bpm	float64	42 / 0.1%
obpm	float64	42 / 0.1%
dbpm	float64	42 / 0.1%
gbpm	float64	42 / 0.1%
mp	float64	36 / 0.1%
ogbpm	float64	36 / 0.1%
dgbpm	float64	36 / 0.1%
oreb	float64	36 / 0.1%
dreb	float64	36 / 0.1%
treb	float64	36 / 0.1%

Table 1. Data Before Cleaning

The Data is heavily imbalanced with drafted = 1 as the minority class (around 0.96% of the dataset)

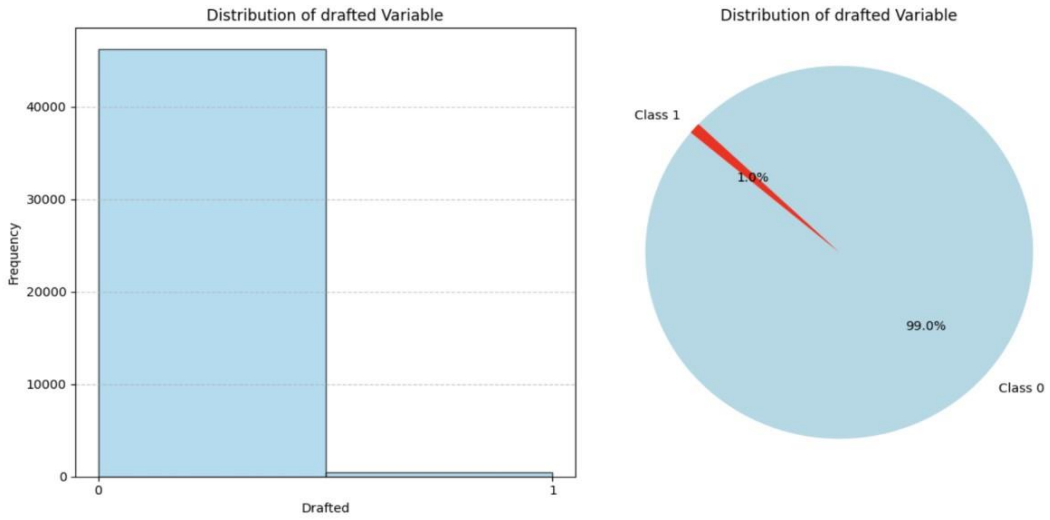

Fig. Class distribution of the target variable 'drafted'

The ‘ht’ (heights) feature was miss input as dates (7-Jun corresponds to 6”7’). Functions were defined to convert it into the correct format and subsequently centimeters (integers) for modelling.

ht	
7-Jun	4611
8-Jun	4554
4-Jun	4408
5-Jun	4355
6-Jun	4274
3-Jun	4228
2-Jun	3874
9-Jun	3285
1-Jun	2974
Jun-00	2523

Fig. Miss-specified height

Most of the numerical features are heavily skewed and highly correlated with one another as shown below.

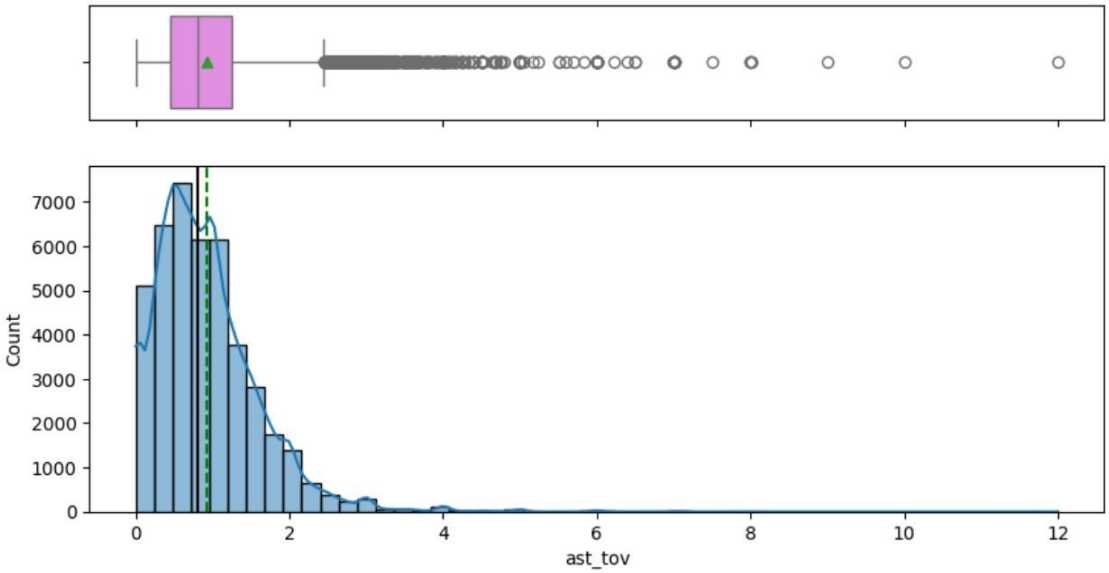

Fig. Distribution of ast_tov feature

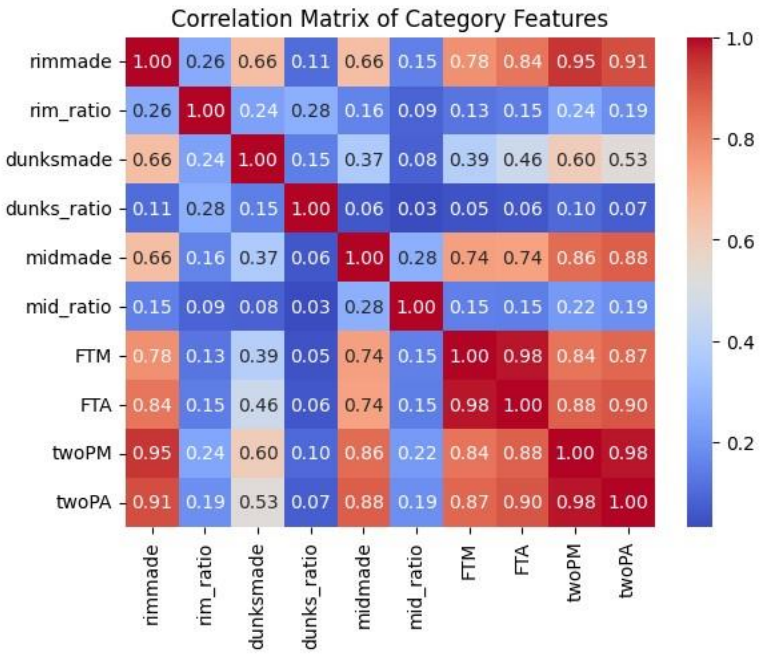

Fig. Numerical features correlation matrix

The skewness of features and correlations should be kept in mind when assessing model performance and results.

4. Modelling

4.1. Logistic Regression & XGBoost

4.1.1. Data Preparation

Through many iterations of experiments, we concluded that more effort should be allocated to the data preprocessing stage to enhance the models' prediction further. The following flowchart results in the most optimal training and testing sets for modelling.

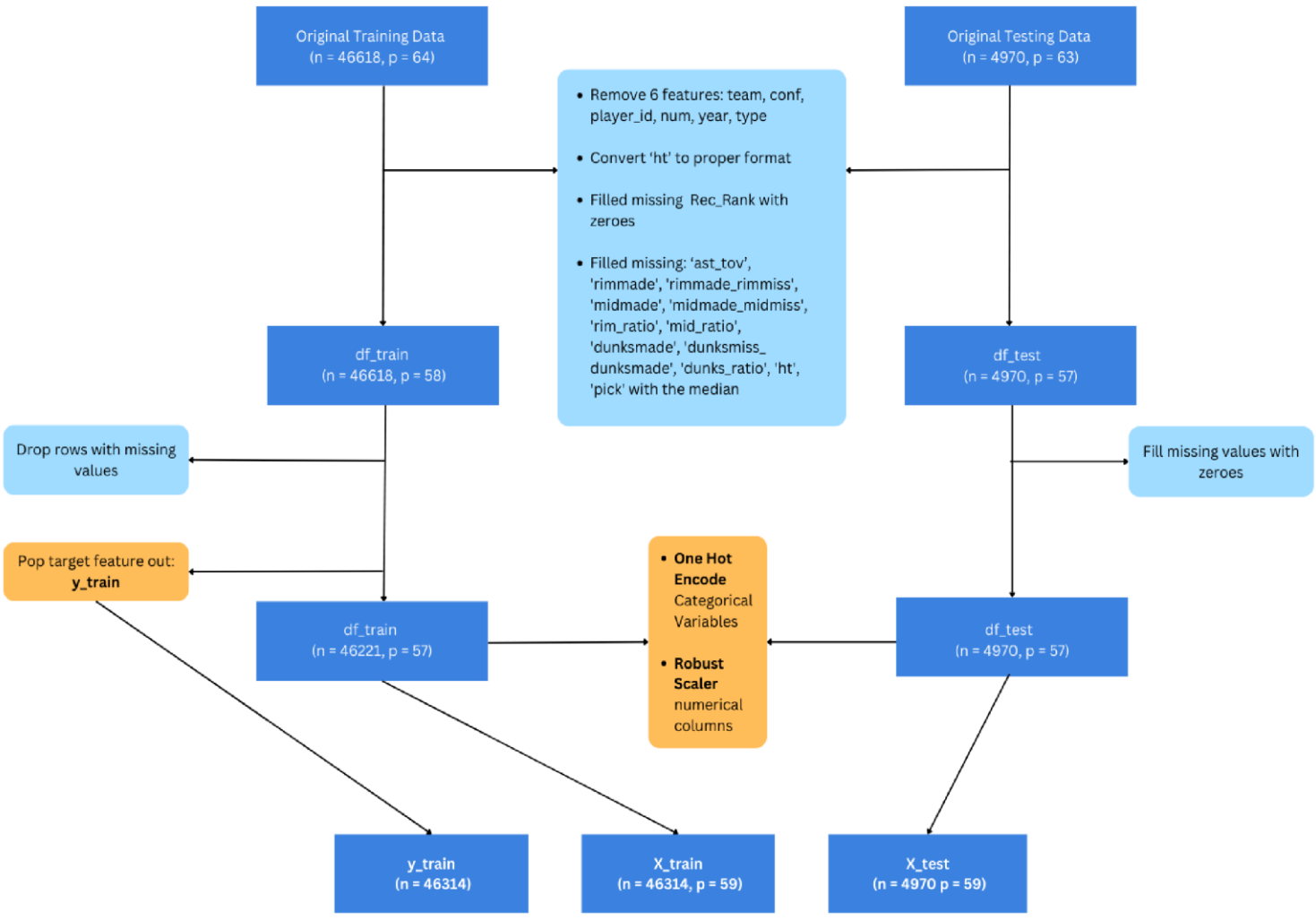

Fig: Data Cleaning flow chart

Justifications for dropping features:

- For ‘num’ (Player’s Jersey number) and ‘year’: in the context of the NBA draft, a player's jersey number and which year they play do not link to the likelihood of getting drafted.
- ‘player_id’ is just an identifier for the players
- Utilizing the CRISP-DM Framework, we discovered that including ‘team’ and ‘conf’ worsens the model prediction during the modelling stage, hence they are dropped. Filling missing ‘Rec_rank’ with zeros and other missing numeric variables with the median also led to better results.

Further scrutiny of the datasets reveals that most of the numerical features are very skewed. Hence, many different scaling methods were trialed and tested. Robust Scaler results in the best outcome.

4.1.2. Modelling

ML ALGORITHM	RATIONALE	HYPERPARAMETERS TUNED
Logistic Regression	Simplistic, fast, and produces relatively good results	C (inverse regularization strength) were used. A smaller C indicates stronger regularization. C = 0.1, 0.5, 1, 2, 5, 10 were used. The best combination was c=10 and penalty = L2
XGBoost	XGBoost has strong performance with an imbalanced dataset where the hyperparameters can be adjusted to scale positive weights and customize the loss function. It can determine the features importance and has built-in regularization to reduce overfitting. It is a very popular algorithm in Kaggle competitions	The following hyperparameters were tuned: • n_estimators = 100, 150, 200 • learning_rate = 0.1, 0.2, 0.01 • max_depth = 3,4,6 • subsample = 1, 0.1 • cosample_bytree = 0.5, 0.7 The best hyperparameters were: • n_estimators = 150, • learning_rate = 0.01 • max_depth = 4, • subsample = 1 • colsample_bytree = 0.7 • gamma = 1

4.1.3. Achieved Results

MODEL		AUROC	Kaggle Score
Logistic Regression	Training	0.9926	0.99082
	Validation	0.9932	
XGBoost	Training	0.9987	0.99848
	Validation	0.9977	

5. Evaluation

a. Business Impact and Benefits

Our best model achieved a 0.99 AUC score, indicating that it is highly capable of distinguishing between players who will and won't be drafted. This level of accuracy can significantly improve the scouting process, reducing the risk of overlooking promising talents.

A consistent and accurate drafting model could save time and scouting resources by filtering highly likely candidates. Additionally, fewer missed draft picks save millions in contract and player development costs.

An accurate probability of getting drafted will also lead to a more interesting discussion among fans, players and sports commentators boosting media engagement with more fan interactions. Hence, the NBA community as a whole will be elevated.

b. Risks and Incorrect Predictions

Negative Impacts of Incorrect Predictions

- **Draft Decisions**: In the case of false positives, where the model predicts that the player should be drafted when they should not be, it will impact the resources allocated for making the drafting decisions such as scouting effort and team evaluation. In the opposite case of false negatives, there will be a loss of not recruiting potential players into the team leading to missed opportunities.

- **Resource Allocation**: Incorrect predictions made by the model can indeed lead to suboptimal results, which can significantly impact budget planning and resource allocation. Such as the travel costs, evaluation costs and coach efforts.

- **Strategy**: Basketball involves utilising strategy not only on the court but off the court as well. Making the correct trades during the off-season makes a huge impact on the team's overall outcome. If such decisions are made based on incorrect predictions it can make the team lose even before the match starts.

By understanding these potential impacts and adopting strategies to address the risks, teams can effectively manage the challenges posed by model inaccuracies and make more informed decisions during the draft process.

c. Data Privacy and Ethical Concerns

There are several important privacy and ethical concerns to consider regarding this model. This includes how the sensitive data of the players are handled, work around potential data bias and deploying the model ethically in the real world.

Data Privacy Implications

- **Player identifiable information:** The dataset contains several sensitive personal data including team, conf, pick, year, and ht which could be used to identify players raising concerns about data privacy. It should be ensured that these data can only be accessed by authorised bodies.
- **Data Security:** Data breach and unauthorised access might happen so the data must be encrypted with secure storage methods.

Ethical Concerns

- **Bias in data:** The dataset may reflect biases such as overrepresentation from prominent teams and conferences which could skew the predictions and cause unfavourable predictions towards players from smaller teams and conferences.
- **Impact on Players' Careers:** Teams that heavily rely on the model to recruit players might put many players in a disadvantageous position by overlooking intangible traits such as leadership qualities, teamwork, etc.

Steps to ensure Data Privacy and Mitigate Ethical Concerns

- **Anonymisation and Data Minimisation:** Remove or anonymize identifiable information (e.g., player_id) before sharing the dataset or deploying the model to limit the risk of data misuse.
- **Fair and Ethical Model Deployment:** Ensure that the model is deployed in such a way that it compliments human decision-making rather than completely ignoring it. The model should be a tool to help the teams in the draft process to get an initial

assessment of the players alongside running other qualitative analyses. This will prevent players from being unfairly judged.

6. Deployment

a. Deployment Process, Integration Steps and Considerations

The deployment stages are shown below which starts off with training and exporting the model which is stored in a cloud server and is used for making real-time predictions using a REST API which will allow stakeholders to input player statistics and get predictions. Batch processing can be conducted to allow large predictions to be made all at once. This can be for an entire school or conference to evaluate all the players.

As the system may handle multiple predictions at once, ensuring the deployment at scale can handle these large predictions. Containerisation tools like Docker may come in handy in these cases.

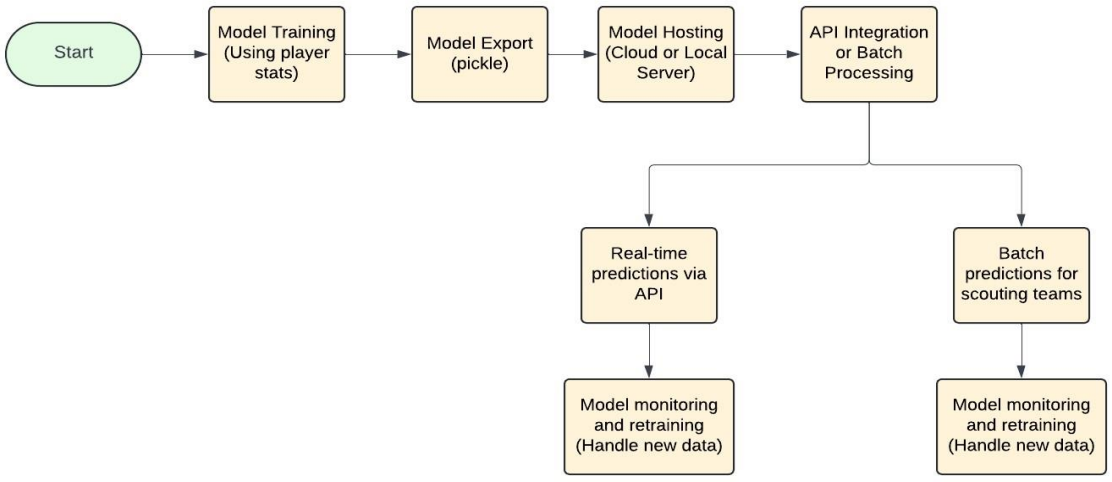

Fig: Model Deployment Flowchart

b. Challenges and Considerations for Deployment:

- **Model performance with real-world data:** The model performance on actual, unseen players might differ from those in the training set. Continuous monitoring of the model is essential in the long run to retain accuracy of the model.
- **Latency:** The predictions need to be made quickly for real-time applications and therefore optimising the model is integral to reduce the lag.
- **Model Versioning:** It is important to keep track of all the models since NBA teams might want to use previous versions of the model to make predictions across different seasons.

7. Conclusion

The high AUC score suggests that the project has met its primary goal of developing a highperforming predictive model. Once the model is deployed, developing a real-time version of the model could allow teams to evaluate players as the season progresses. Future work should look to expand the model to other sports beyond the NBA.

In developing the right model, the team underwent many iterations of the CRISP-DM methodologies. We share the key learning from each stage of the output, continue to improve data cleaning and feature qualities and test out many different preprocessing steps to reach the optimal model results.

Pushing the code to Github, and designing custom packages under the cookiecutter template was another learning topic throughout the project. The initial set-up of our own repositories and working environment created some delays in the progression of the project, but the processes became more efficient as we progressed through each stage. Overall the Kaggle competition project has provided a realistic hands-on experience that elevates our skills as data scientists.