Assignment 1  
Kaggle Competition

short line

Thirada Tiamklang

Student ID: 14337188  
8 September 2023

36120 - Advanced Machine Learning Application

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[1. Executive Summary 2](#_Toc145036047)

[2. Business Understanding 3](#_Toc145036048)

[a. Business Use Cases 3](#_Toc145036049)

[3. Data Understanding 5](#_Toc145036050)

[4. Data Preparation 9](#_Toc145036051)

[5. Modeling 11](#_Toc145036052)

[a. Experiment A 11](#_Toc145036053)

[b. Experiment B 11](#_Toc145036054)

[c. Experiment C 12](#_Toc145036055)

[6. Evaluation 13](#_Toc145036056)

[a. Evaluation Metrics 13](#_Toc145036057)

[b. Results and Analysis 13](#_Toc145036058)

[c. Business Impact and Benefits 14](#_Toc145036059)

[d. Data Privacy and Ethical Concerns 14](#_Toc145036060)

[7. Deployment 15](#_Toc145036061)

[8. Conclusion 16](#_Toc145036062)

[9. References 17](#_Toc145036063)

# 

# Executive Summary

The primary objective of this project is to predict whether a college basketball player will be drafted to join the NBA league based on their statistics for the current season. This prediction task holds significant value for both sports commentators and fans, as it provides insights into the future prospects of aspiring players. We assess the model's performance using the AUROC (Area Under Receiver Operating Characteristic) score, a key benchmark for classification models. A higher AUROC score signifies a model's improved ability to accurately classify positive and negative examples, enhancing prediction accuracy.

However, we encountered some challenges during the experiments. Theraw dataset presented challenges, primarily stemming from missing values (NA), which could potentially impact model training and results. Additionally, the choice of features and hyperparameters significantly influenced the AUROC scores, leading to a structured approach involving three experiments labeled A, B, and C. Each successive experiment builds upon the insights gained from the previous one.

After rigorous experimentation, the AdaBoost classifier model emerged as the top performer. It achieved an impressive AUROC score of 0.9886 on the testing set. This result was achieved by leveraging all available features and employing grid search to optimize hyperparameters, including learning\_rate and n\_estimators. The predictive power of this model will be harnessed for assessing the probability of player drafting in the test dataset.

This project underscores the value of machine learning in providing valuable insights into the world of basketball, benefiting analysts, enthusiasts, and the broader sports community.

short dash

# Business Understanding

## Business Use Cases

**Project Scope:** This project applies machine learning algorithms to address specific business use cases related to predicting the likelihood of a college basketball player being drafted into the NBA based on their statistics for the current season.

**Challenges and Opportunities:**

**- Data Quality:** The raw dataset posed challenges due to missing values (NA), potentially affecting the accuracy of model training and predictions. Addressing these data quality issues was a key motivation.

**- Predictive Insights:** The project offers valuable insights for various stakeholders, including sports commentators and fans, who are eager to assess the potential of college players and make predictions regarding their NBA draft prospects.

**- AUROC Score Improvement:** The project aims to improve the AUROC score, which measures the model's ability to distinguish positive and negative examples effectively. A higher AUROC score signifies enhanced prediction accuracy.

1. Key Objectives

**Project Goals:**

The primary objective is to predict the probability of a college basketball player being drafted into the NBA, leveraging machine learning algorithms. This prediction is crucial for assessing players' prospects.

**Stakeholder Requirements:**

Commentators require accurate predictions to provide informed insights and analysis during games and broadcasts. Also, fans are keen to follow the careers of college players and make informed speculations about their future in the NBA.

**Addressing Stakeholder Requirements:**

**- Model Training:** To address these requirements, a machine learning model is trained using a labeled training dataset that includes the 'drafted' feature.

**- Model Comparison and Improvement:** The project focuses on comparing and improving the AUROC scores of various machines learning algorithms, including the Adaboost classifier, Random Forest classifier, and Polynomial logistic regression. Different hyperparameters and feature selections are explored to enhance model performance.

**- Prediction:** After model training and selection, the best-performing model is utilized to predict the probability of drafting for players in the test dataset. This prediction serves as a valuable tool for sports commentators and fans to assess the potential of college basketball players in joining the NBA.

This project aligns the capabilities of machine learning algorithms with the needs of sports commentators and fans, providing accurate predictions and enhancing the overall understanding of player prospects in the NBA draft.

short dash

# Data Understanding

The 'NBA Draft' dataset, sourced from [Kaggle,](https://www.kaggle.com/competitions/advmla-2023-spring/overview) comprises data on college basketball players' NBA draft prospects, as determined by their statistics for the current season. This dataset is divided into two subsets: the train dataset and the test dataset.

**Train Dataset:** The training dataset consists of 56,091 rows and 64 columns, including the target variable 'drafted,' which indicates whether a player was drafted into the NBA.

**Test Dataset:** The test dataset, used for predictions by the trained model, contains 4,970 rows and 63 columns (excluding the 'drafted' target variable).

The dataset encompasses 7 categorical features and 57 numeric features, each contributing to the predictive power of the model. Below, you will find a list of these features and their respective descriptions.



**Figure 1** The table of the dataset description

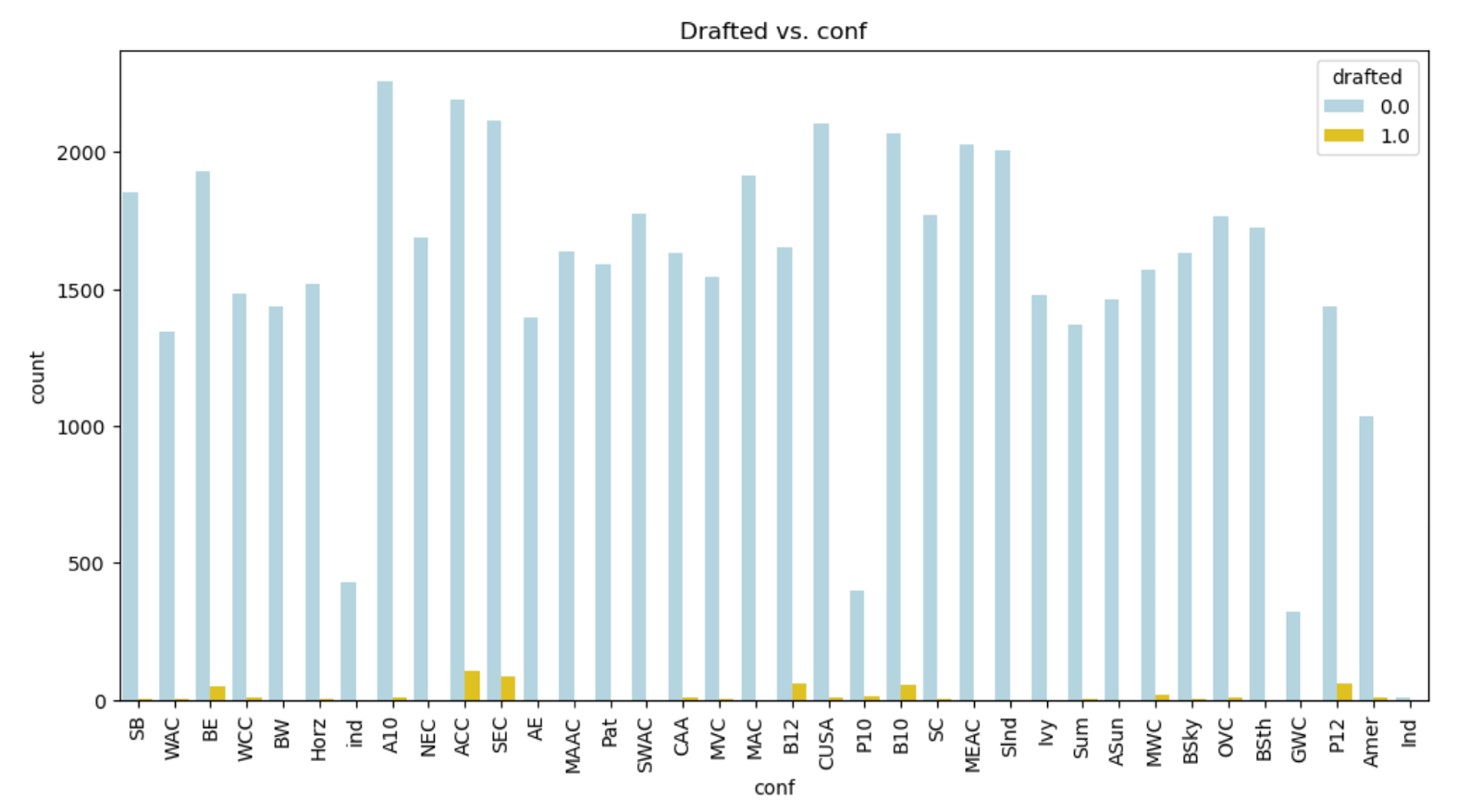
In preparation for this project, an exploratory data analysis (EDA) was conducted to gain insights into the dataset's characteristics. EDA allows for the visualization and understanding of trends, distributions, and potential correlations among features. This process aids in selecting relevant variables and optimizing the model-building process.

A graph of a bar

Description automatically generated with medium confidence

**Figure 2** The bar chart indicates whether a player was drafted into the NBA or not.

It can be seen obviously that the target variable, drafted, is imbalanced. In this case, we have a significantly higher number of players who are not drafted compared to those that are drafted. This can lead to a bias in the model towards predicting that a player is not drafted. We may use the Over-sampling Technique (SMOTE) to address this issue.



**Figure 3** The bar chart of Name of conference vs drafted.

It is evident that the majority of players in the conference were not drafted. However, there are some significant yellow bars in the chart, indicating drafted players in the ACC and SEC.

short dash

# Data Preparation

**Pre-processing data:**

1. Load two dataset by reading CSV. file. The train.csv dataset used for training the model and the test.csv dataset which is not contain the drafted feature will be used for predicting probability in the final result.

2. Clean NA:

**- Check NA**

A screenshot of a computer code

Description automatically generated

After sorting all the features, 'pick,' 'Rec\_rank,' and 'dunks\_ratio' have a high number of missing values, which is above 10k. We will drop these variables since they could potentially lead to inaccurate outcomes. However, dunks\_ratio seems to be significant for the player to be drafted. There are some patterns to be relevant to this feature.

**- Missing value pattern**

**A graph of a heatmap

Description automatically generated**

**Figure 4** The missing value Heatmap

It's evident that 'rimmade,' 'rimmade\_rimmiss,' 'midmade,' 'midmade\_midmiss,' 'dunksmade,' and 'dunksmiss\_dunksmade' share the same number of missing values. This is because the data contained in these features starts from the season year 2010.

**- Clean data**

First, we will drop features with more than 10,000 missing values and then replace missing values with zero in columns with missing data for the year 2009. We will also drop irrelevant features (e.g., 'num' and 'ht') with respect to the target variable and impute missing numeric values with the mean and missing categorical values with the mode.

**Feature engineering:**

1. Label encoding the categorical features including yr, type, team, player\_id, and conf.

2. Scale data with StandardScaler

3. using Synthetic Minority Over-sampling Technique (SMOTE) since the ‘drafted’ is imbalanced. The new balance is 55,555 for each drafted and not drafted.

4. Split data into train, validation, and test set with ratio 80:20.

short dash

# Modeling

With binary classification as the target variable, we will train the first dataset containing the target variable using a different models, hyperparameters, and features selection. This will allow us to assess the AUROC scores of the training, validation, and test sets. Subsequently, we will employ the best trained model to predict the probability of being drafted for the second (test) dataset, which does not contain the target variable. This project divided into three experiments, including A, B, and C.

## Experiment A

In this experiment, we will train the first dataset containing the target variable using a **logistic regression classifier with polynomial features**. This will enable us to assess the AUROC scores for the training, validation, and test sets in comparison to the baseline models.

The hyperparameters were selected using a random forest classifier and the Variance Inflation Factor (VIF). The final feature set, denoted as 'X,' comprises four features: 'adjoe,' 'rimmade,' 'dunks\_ratio,' and 'adrtg.' Additionally, we addressed the issue of imbalanced data by employing SMOTE during the feature engineering process.

## Experiment B

In this experiment, we will train the dataset using an **AdaBoost classifier with hyperparameter tuning.** The goal is to improve the AUROC scores on the training, validation, and test sets by exploring various hyperparameters.

To facilitate a comparison with the best model from the previous experiment, we will train the AdaBoost classifier on the same feature selection and dataset split used in the previous experiment.

## Experiment C

In this experiment, we will train and compare the models of **Polynomial Logistic Regression, AdaBoost, and Random Forest classifiers using grid search**. The objective is to enhance the AUROC scores on the validation and test sets by experimenting with different models.

To improve the AUROC score, we will explore a new feature selection, which includes all the features from the dataset after dropping and cleaning NA values.

The hyperparameters used after grid search are as follows:

1. LogisticRegression(random\_state=42)

2. AdaBoostClassifier(learning\_rate=0.1, n\_estimators=200, random\_state=42)

3.RandomForestClassifier(max\_depth=10,min\_samples\_leaf=2,min\_samples\_split=5, random\_state=42)

short dash

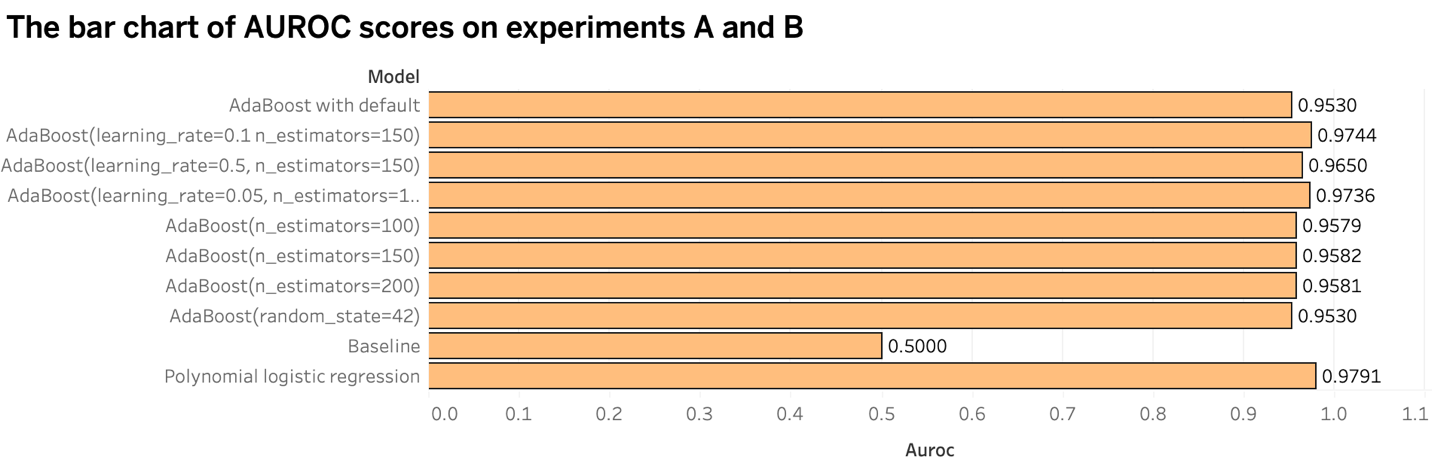
# Evaluation

## Evaluation Metrics

**AUROC Score**

The AUROC measures model accuracy in predicting NBA draft prospects, making it a key metric for our project's success. Its flexibility, balanced assessment, and ability to facilitate model comparison make it a suitable and relevant metric for evaluating the project's success in achieving this goal.

## Results and Analysis



The bar chart above represents the AUROC scores of three models: the baseline, Polynomial Logistic Regression, and Adaboost with hyperparameter tuning. It is evident that **Polynomial Logistic Regression** achieves the highest score at 0.9791487117335488.

A graph with a number of text

Description automatically generated with medium confidence

The bar chart above displays the AUROC scores for the AdaBoost classifier, Polynomial Logistic Regression, and Random Forest classifier using grid search. Notably, Random Forest and AdaBoost classifiers exhibit nearly identical AUROC scores of 0.98. However, **the AdaBoost classifier** outperforms the Random Forest model significantly, achieving a score of 0.9885671963371937 on the testing set, whereas the Random Forest scores 0.9866246970105036. This demonstrates the superiority of boosting over bagging in this context.

In conclusion, training **the AdaBoost classifier with hyperparameters learning\_rate=0.1, n\_estimators=200, and random\_state=42,** using all available features, yielded the highest AUROC score. We will utilize this trained model to predict the probability of being drafted in the test dataset.

## Business Impact and Benefits

The AdaBoost classifier, with its optimized parameters, significantly improves the accuracy of predicting NBA draft prospects. This positive impact is evident through the availability of the result\_C.csv file, which allows sports commentators and fans to assess a player's likelihood of being drafted into the NBA, enhancing their engagement and predictions. The model's superiority over bagging techniques reinforces its value in this context.

## Data Privacy and Ethical Concerns

1. Data Privacy Implications:

**Sensitive Information:** The dataset used for this project likely contains sensitive information about individuals, particularly college basketball players. This includes their performance statistics and potentially other personal details. Sharing or using this information without proper consent or anonymization could raise data privacy concerns.

2. Ethical Concerns:

**Informed Consent:** Ethical concerns arise if the data used in this project was collected without proper informed consent from the individuals involved. If players' data was used without their knowledge or consent, it raises ethical questions regarding data usage.

**Fair and Unbiased Models:** There is an ethical responsibility to ensure that the predictive models are fair and unbiased. Any biases present in the data or model can result in unfair treatment or discrimination.

3. Data Privacy and Ethical Considerations:

**Data Anonymization:** Steps should be taken to anonymize the data, removing any personally identifiable information to protect the privacy of individuals.

**Consent and Transparency:** If possible, ensure that the data used for this project was collected with proper consent and transparent information sharing practices. Inform individuals about the purpose of data collection and how their data will be used.

short dash

# Deployment

**Deployment Process:**

The deployment of the best-performing models has been streamlined and is available through the '.joblib' files, which are located in the 'models/' directory. You can access these files in the GitHub repository [here.](https://github.com/thirada2799/Thirada_AML_AT1/tree/main/Thirada_adv_mla_at1)

A screenshot of a computer

Description automatically generated

**Real-World Implementation:**

To apply the model in real-world scenarios, follow these steps:

*1. Predicting on Test Data:* To predict the probability of college basketball players being drafted into the NBA using the test dataset, refer to the 'result\_C.csv' file. It contains the predictions based on the trained models.

*2. Predicting on New Data:* If you have new or current data and wish to make predictions, you can use the provided script (e.g., ['predict.py'](https://github.com/thirada2799/Thirada_AML_AT1/blob/main/Thirada_adv_mla_at1/src/models/predict_model.py)). This script is designed to load the pre-trained models and is ready to predict the likelihood of college players being drafted into the NBA.

**Recommendations:**

1. Regularly monitor model performance in the real-world context and be prepared to fine-tune or update the model as needed.

2. Document the deployment process, including instructions for users, as provided in the README.md file in the GitHub repository, to facilitate ease of use and understanding.

3. Encourage user feedback to identify any issues or improvements in the deployment process and the predictions made by the model.

short dash

# Conclusion

In summary, this project has yielded valuable insights and outcomes for predicting the likelihood of college basketball players being drafted into the NBA. Key findings include:

- The AdaBoost model, configured with 'learning\_rate=0.1,' 'n\_estimators=200,' and 'random\_state=42,' demonstrated outstanding performance on the testing set, achieving an AUROC score of 0.9886.

- This model's accuracy in predicting draft prospects makes it a valuable tool for sports broadcasters and fans seeking real-time insights into current season results.

The project has succeeded in achieving its goals, particularly in building a predictive model that enhances our ability to assess NBA draft prospects accurately. It meets the requirements of stakeholders, contributing to more informed predictions in the world of basketball.

**Future Work and Recommendations:**

- Model Updates: Continual updates to the model using new data can further enhance accuracy.

- Real-Time Deployment: Consider deploying the model in real-time applications for immediate predictions during the NBA season.

- Enhanced User Experience: Improve user interfaces and accessibility for sports broadcasters and enthusiasts.

Overall, the success of this project paves the way for more accurate predictions and deeper engagement in the world of college basketball and the NBA. The continuous evolution of this model promises even more accurate assessments in the future.

short dash

# References

So, A. (2023). *36120\_AdvMLA-Lab1\_Exercise2-Solutions.ipynb*. <https://colab.research.google.com/drive/15OZMUMwUBoAmtrfuzJaEhF1Ta8XkCmQZ?authuser=1#scrollTo=Pw_LqGuGC9Oz>

So, A. (2023). *36120\_AdvMLA-Lab1\_Exercise3-Solutions.ipynb*. <https://colab.research.google.com/drive/1sHbkg8n7cU_GSm4AB7oKKIzFjJvobBxK?authuser=1#scrollTo=Bc10AnKyW23U>

short dash