

Assignment

1

Kaggle Competition

[SimonUTS24661225/ADV_MLA_AT1: This is assignment 1 in Advanced Machine Learning and Algorithm \(github.com\)](#)

Simon Lim

StudentID:

24661225

<08/09/2023>

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

Table of Contents

1. Executive Summary	2
2. Business Understanding	3
a. Business Use Cases	3
3. Data Understanding	4
4. Data Preparation	5
5. Modeling	6
a. Approach 1	6
b. Approach 2	6
c. Approach 3	6
6. Evaluation	8
a. Evaluation Metrics	8
b. Results and Analysis	8
c. Business Impact and Benefits	8
d. Data Privacy and Ethical Concerns	9
7. Deployment	10
8. Conclusion	11
9. References	12

1. Executive Summary

The NBA draft is an annual event in which American teams select players from their American colleges to join their rosters. The process of moving to NBA league is very crucial for the progress of basketball player's career. In this regard, the objective of the project is to build potential models that correctly predict if a random college basketball player will be drafted to join the NBA league based on players' statistics information for the current session. There are two datasets, training and testing datasets. Training dataset is used to train models and testing dataset is used to assess the performance of models. The performance of models is assessed based on AUROC scores on testing dataset, measuring the probability of predictions whether players can be drafted by NBA league or not. This project conducts three different experiments, in which different classification models are trained and their performance are assessed based on AUROC score. The models used in experiments include logistic regression, LightGBM and SVM models. Prior to training models, the process of data preparation, cleaning, feature engineering, data split and data scaling have also been performed. Consequently, the results of experiments were all successful, as all models showed very high AUROC scores, around 0.98 and 0.99, indicating that models can correctly predict random players, who will be drafted to join the NBA league.

■ ■ ■

2. Business Understanding

a. Business Use Cases

The NBA draft is an annual event in which American teams select players from their American colleges to join their rosters. The process of moving to NBA league is very crucial for the progress of basketball player's career. Many fans from different teams are curious about potential college players being drafted by NBA teams. Furthermore, NBA teams can scout potential players based on results of machine learning algorithms. In terms of this context, machine learning algorithms can be potentially helpful by enabling to build classification models that predict if a random college basketball player will be drafted to join the NBA league based on players' statistics information for the current session.

b. Key Objectives

The main goal of the project is to build binary classification models that can correctly predict whether random players can be drafted by NBA teams based on various statistic information for the current session. Classification models would allow us to categorize players into one of two classes, in which one is associated with players being drafted and the other is with players not being drafted. The results of models can help NBA teams, fans and commentators in finding potential basketball players, who will be likely to be drafted by NBA league and show great performances within teams. The project will find potential future NBA players by using various statistics information for each player. That information would include numeric values of performances, such as effective field goal percentage, defensive rebound percentage, turnover percentage, block percentage and throw attempts etc. Machine learning models will be trained based on those true statistics and predict whether players can be drafted by NBA league or not.



3. Data Understanding

There are training dataset and testing dataset, each containing 56091 and 4970 rows. Most of data features involve numeric values, representing various statistics of players for the current session. Statistics information include numeric values, such as effective field goal percentage, defensive rebound percentage, turnover percentage, block percentage and throw attempts etc. Those statistics values are main features that will be used to train models in future stage. On the other hand, there are some limitations in those datasets. First, there are three object columns 'ht', 'yr' and 'num' and some values in those columns are ambiguous and difficult to be transformed into numeric values. For example, there are '23B', '4A' and '31/24' values and such values are hard to be interpreted. Furthermore, there are three columns, 'pick', 'Rec_Rank' and 'dunks_ratio', which contain 71%, 55% and 99% missing values in each column. Those columns will disturb the ground truth of machine learning models in later stage.



4. Data Preparation

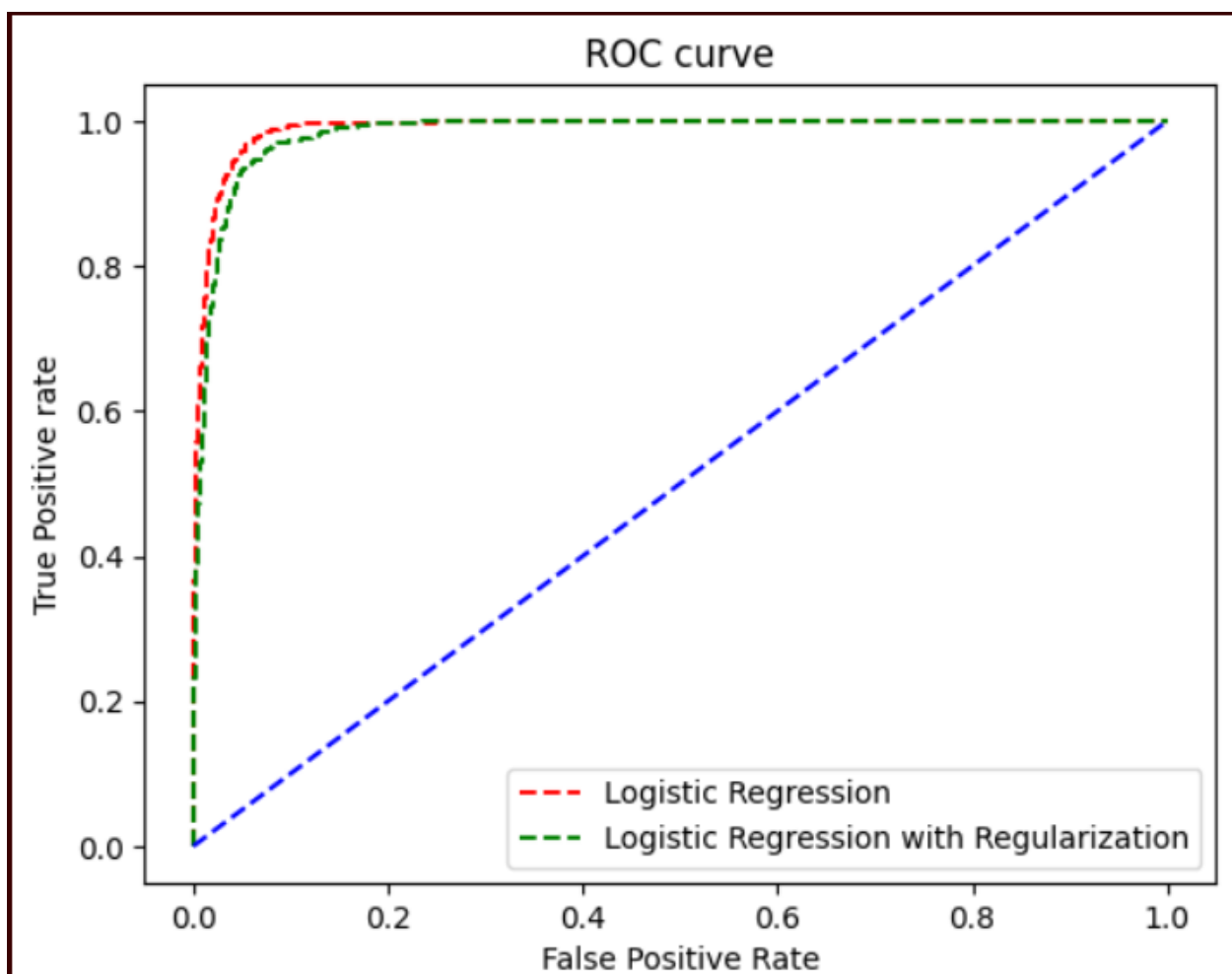
- There were a lot of missing values in a couple of columns, almost 24 columns. There were three columns (e.g., 'pick', 'dunks_ratio', 'Rec_Rank'), in which there were more than 50% of missing values in both training and testing datasets. I decided to delete these three columns rather than dropping missing values, in order to enhance the ground truth of the project (i.e., including all rows of data). The other columns were filled with means of the columns.
- There were three object columns ('ht', 'num' and 'yr'), in which string values in 'num' and 'yr' were replaced with appropriate numeric values and data types were changed to integer or float. Unlike 'num' and 'yr', 'ht' column was unable to be converted to numeric values due to their date format (ambiguous values).
- While other missing values in training dataset were dropped, few missing values in testing dataset were filled with means of values in columns to keep the ground truth of machine learning models.
- EDA was performed with histogram and boxplot to check distributions and outliers of data. There were many outliers in each column, and therefore decided to normalize data.
- In training dataset, numeric data were split into X_train, containing all numeric columns except 'drafted', target variable and y_train containing 'drafted' column. For testing dataset, X_test was only created, containing all numeric columns and there was no y_test, as 'drafted' column was not given in testing dataset. (In this project, instead of obtaining AUROC score for testing dataset through target variable, AUROC score for testing dataset will be obtained by submitting probabilities of predictions for data on Kaggle).
- 'StandardScaler' was used to remove outliers in data.
- Baseline model was tested before training models to track progress.



5. Modeling

a. Approach 1

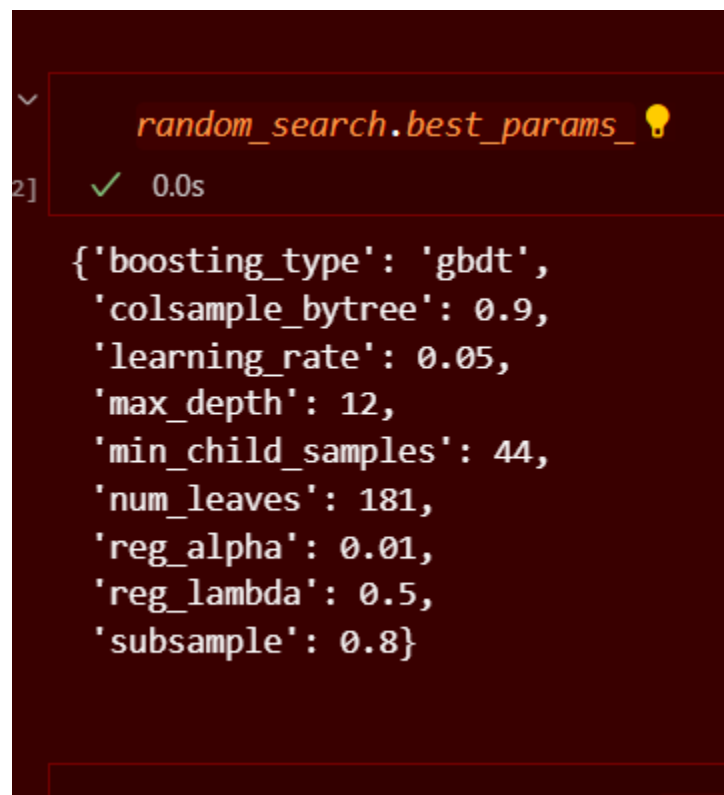
In Experiment 1, two kinds of logistic regression models were used, one with default model and the other with regularizations (penalty= 'elasticnet', l1_ratio=1, solver='saga'). The rationale behind choosing logistic regression is that the target variable is binary and is linearly separable. Logistic regression is powerful in binary classification when data is linearly separable. Furthermore, logistic regression models do not purely take any other assumptions into account except categorizing data into two classes based on statistics information. As the purpose of the project is to simply distinguish players between drafted class and non-drafted class, logistic regression models are one of suitable models in this project. Logistic regression default model and regularization model were trained, and two training models were visualized, and probability scores were obtained on testing dataset.



b. Approach 2

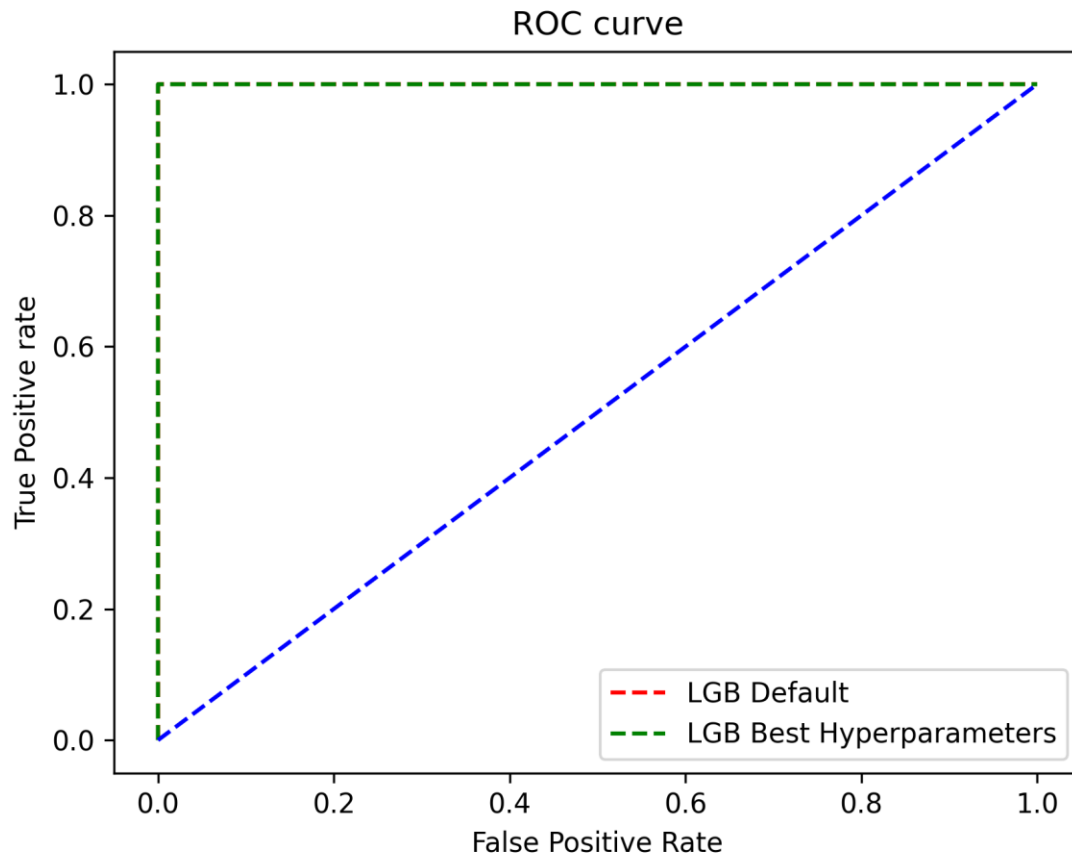
In Experiment 2, LightGBM is main algorithm that was used in the second experiment. There are two LGBM models in this experiment, including one with default model and the other with optimal hyperparameters, which were obtained using 'RandomSearchCV' technique. The rationale behind LGBM is that LGBM algorithm is fast and accurate and particularly well-suited for large datasets. As this dataset contains over 50000 data on the training dataset, LGBM can be a powerful model in distinguishing potential NBA league players.

In order to optimize the LGBM model, 'RandomSearchCV' was employed for tuning hyperparameters. In a range of parameters for LGBM, best hyperparameters were obtained as follow.

A screenshot of a Jupyter Notebook cell. The cell's prompt is '2]'. The output is a dictionary of best hyperparameters for LightGBM, displayed in a light blue box with a lightbulb icon. The parameters are: 'boosting_type': 'gbdt', 'colsample_bytree': 0.9, 'learning_rate': 0.05, 'max_depth': 12, 'min_child_samples': 44, 'num_leaves': 181, 'reg_alpha': 0.01, 'reg_lambda': 0.5, and 'subsample': 0.8. The execution time is shown as 0.0s with a green checkmark.

```
random_search.best_params_💡  
✓ 0.0s  
{'boosting_type': 'gbdt',  
 'colsample_bytree': 0.9,  
 'learning_rate': 0.05,  
 'max_depth': 12,  
 'min_child_samples': 44,  
 'num_leaves': 181,  
 'reg_alpha': 0.01,  
 'reg_lambda': 0.5,  
 'subsample': 0.8}
```

LGBM default model and best hyperparameter model were trained, and two training models were visualized, and probability scores were obtained on testing dataset.



c. Approach 3

Support Vector Machine (SVM) is a main algorithm used in Experiment 3. There are two kind of SVM models in this experiment, including one with default model and the other with optimal hyperparameters, which were obtained using 'RandomSearchCV' technique. The rationale behind SVM models is that SVM models can effectively classify data by finding the optimal hyperlane, which exactly separates data points of two different classes.

In order to optimize the LGBM model, 'RandomSearchCV' was employed for tuning hyperparameters. In a range of parameters for LGBM, best hyperparameters were obtained as follow.

```
param_grid = {
    'C': np.logspace(-3, 3, 7),
    'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
    'gamma': ['scale', 'auto'] + list(np.logspace(-3, 3, 7)),
    'probability': [True]
}

svc_best = svm.SVC()

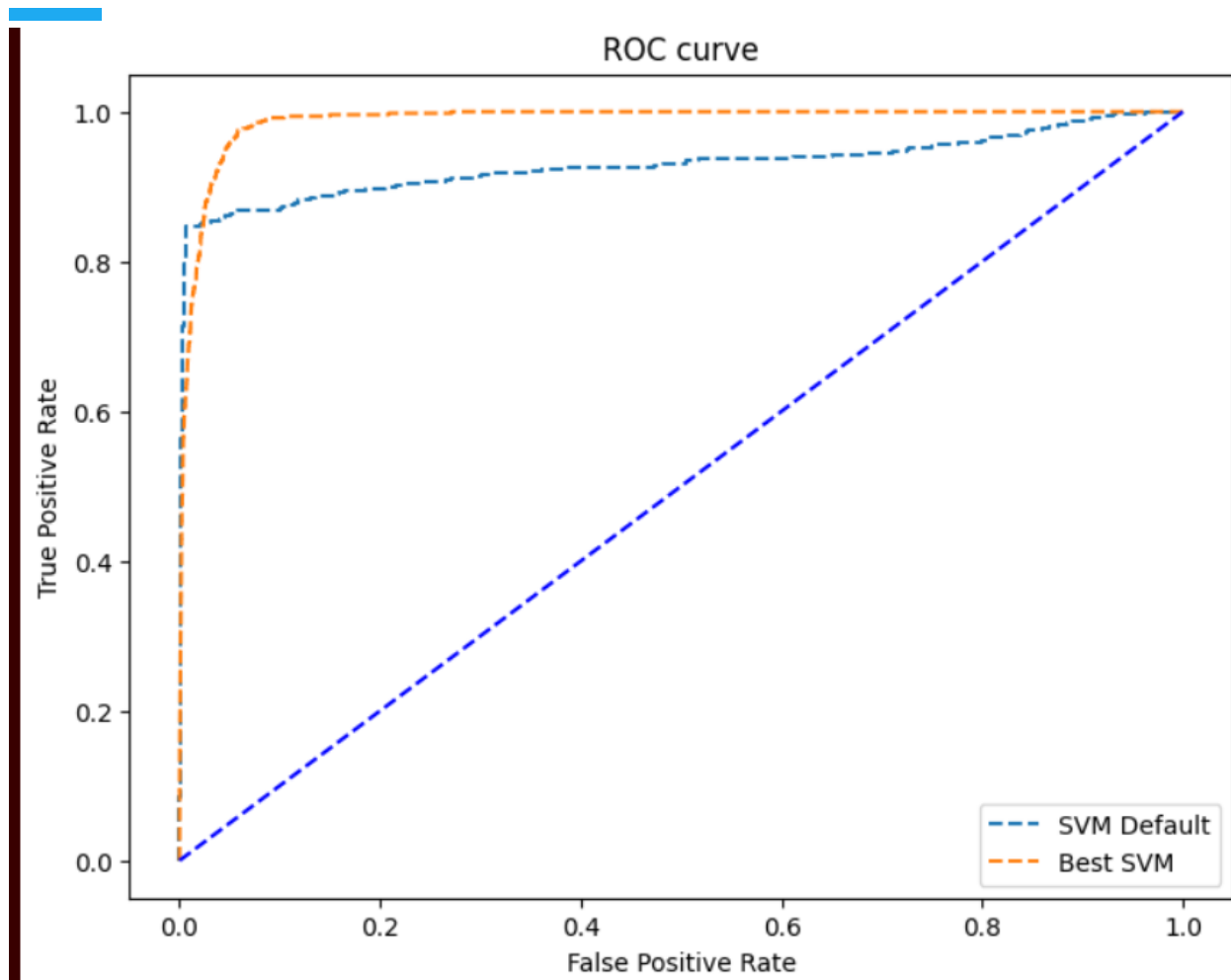
random_search = hyperparameter_tuning_randomized_search(svc_best, param_grid, X_train_std_scaled, y_train)

Best Parameters:
{'probability': True, 'kernel': 'linear', 'gamma': 1000.0, 'C': 0.1}
Best roc_auc Score: 0.9872198682069715
```

[+ Code](#) [+ Markdown](#)

In Experiment 3, function of 'randomSearchCV' was used to simplify code. Also, for SVM models, I manually set probability score as True, as SVM model does not take probability score into account.

SVM default model and best hyperparameter model were trained, and two training models were visualized, and probability scores were obtained on testing dataset.



■ ■ ■

6. Evaluation

a. Evaluation Metrics

For performance metric, AUROC score was used as a main metric. The purpose of the project is to build models that can correctly distinguish random players based on statistics of players between positive and negative classes (whether drafted or not into NBA league). Given that two classes of data were very imbalanced, AUROC score can successfully measure performance of model's probability of predictions whether players belong to positive or negative classes. Therefore, AUROC score will be a suitable metric for this project.

b. Results and Analysis

The main metric in this project is AUROC score and therefore only AUROC score was evaluated in this project.

Experiment 1 (Logistic Regression).

For logistic regression models, default logistic regression model showed a greater performance than logistic regression with regularizations (penalty= 'elasticnet', l1_ratio=1, solver='saga'). The default model obtained 0.98243 AUROC score on testing dataset. However, there was a lack of searching optimal hyperparameters for logistic regression model. Through Experiment 1, I realized that finding optimal hyperparameters would be crucial steps to improve the performance of models. Nevertheless, logistic regression model obtained a reliable and high AUROC score, which can correctly distinguish random players between drafted and non-drafted classes.

Experiment 2 (LGBM models).


For LGBM models, LGBM model with optimal hyperparameters showed a greater performance than LGBM default model. For Experiment 2 and Experiment 3, 'RandomSearchCV' enabled to find optimal hyperparameters, which successfully resulted in greater performances than default models. AUROC score of default LGBM model was 0.98366 and AUROC score of best hyperparameters model was 0.98392. Overall, LGBM models showed slightly better performance than logistic regression models.

Experiment 3 (SVM models).

For SVM models, SVM model with optimal hyperparameters also showed a greater performance than default model. 'RandomSearchCV' was also used to find optimal hyperparameters in this experiment. SVM model with optimal hyperparameters obtained 0.98938, while SVM default model obtained 0.86241. Overall, SVM models with tuned hyperparameters showed the greatest performance in this project but default model relatively showed a poorer performance than other models. Again, it is assured that tuning optimal hyperparameters are essential step to improve the performance of models.

c. Business Impact and Benefits

In consequence, the models obtained very high AUROC scores, indicating that models can correctly predict if players will be drafted to join the NBA league in the future. It also means that models can successfully distinguish players between drafted and non-drafted classes. By using



the final model, stakeholders, such as teams, fans and commentators will be able to find potential basketball players, who are likely to be drafted by NBA league.

d. Data Privacy and Ethical Concerns

In terms of data privacy implications of the project, it is important to ensure that players are anonymized to protect the privacy of players. Also, statistics information should not be shared personally. Furthermore, it is also crucial to maintain transparency in the process of building models. Most importantly, when models are used by stakeholders in the future, they should make informed decisions-making based on the predictions of models. In terms of the data security, collected and stored data should be protected, and appropriate security measure should be conducted to prevent data breaches and unauthorized access.

■ ■ ■

7. Deployment

- In this project, the ground truth of models has been emphasized and hence all data on testing dataset were included. Keeping the ground truth of models also indicates that models can provide true answer to the objective of the project.
- Subsequently, cleaning data and feature engineering were performed on training and testing dataset to optimize the productivity of models.
- Also, models' artefacts were saved and kept in separate file (using Joblib) to reload and use the models directly when needed.
- Data quality was not good, as dataset contained a few ambiguous values and lots of missing values. It was crucial to fill missing values rather than dropping them (For ground truth).
- Github was a great environment for access to users, to save codes, models and other relevant data and keep tracks of models, monitoring and handling models. Also, by setting github private, it also protects both the model and data.



8. Conclusion

In conclusion, the project aimed at predicting if random college players can be drafted by NBA league. Three experiments have been conducted with different models and models successfully predict players who are likely to be drafted by NBA league. AUROC score was suitable as a performance metric in this project, as it measures the probability of players being drafted by NBA league. The models can be used in real-life application for stakeholders to obtain the list of players, who can potentially be drafted by the NBA league.

There are some important key learnings and steps that can be considered for future projects.

- Deciding main performance metric in advance is very helpful in building optimal models.
- Rather than removing missing values, filling out missing values can potentially enhance the performance of models (in case the proportion of models is not large).
- Feature engineering is an essential stage in machine learning, that can influence models and outcomes of the projects.
- Optimal hyperparameters can be found using techniques, such as 'RandomSearchCV' or 'GridSearchCV'.
- If procedures (cleaning, feature engineering) in experiments are same, using custom packages or different functions can help to identify trends and patterns.
- Saving models in separate folder enables to re-load models in the future.



9. References

ADVMLA-2023-spring (no date) *Kaggle*. Available at: <https://www.kaggle.com/competitions/advmla-2023-spring> (Accessed: 08 September 2023).

Bhandari, A. (2023a) *Guide to AUC ROC curve in machine learning : What is specificity?*, *Analytics Vidhya*. Available at: <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/> (Accessed: 08 September 2023).



Note: The CRISP-DM steps (Cross-Industry Standard Process for Data Mining) provide a framework for structuring the project report, but feel free to adapt the template to match the specific requirements and guidelines of your project or organization.