EXPERIMENT REPORT

Student Name	Saumya Bhutani
Project Name	AT 1 Part B
Date	25/08/2023
Deliverables	<SaumyaBhutani_AT1_PartB_14360820> <AdaBoost Automated Hyperparameter Tuning> <GitHub Link: https://github.com/bhutanisaumya/Adv_MLA.git>

---

1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	The project aims to create a predictive model to forecast whether a college basketball player will get drafted into the NBA based on their performance matrices. Using a machine learning model to predict which players are more likely to be drafted, the business can optimise their efforts for NBA teams by helping them allocate resources effectively for scouting, assisting analysts in player evaluation, aiding player agents in advising their clients, engaging fans and media with data-driven discussions, and guides strategic draft decisions. The project's success is measured by how well the model can predict draft outcomes, evaluated through the AUROC metric. In this experiment, I aim to determine how well an AdaBoost with Automated Hyperparameter Tuning can predict the drafting probabilities using multiple independent features. Impact of accurate result: If the model accurately identifies drafting of players then the business can optimize the accurate results from the predictive model positively to influence draft strategies, player decisions, scouting efficiency, and fan engagement. Accurate results will increase the model's reliability, and the model is more likely to be used to make better predictions. Impact of incorrect result: If the model incorrectly identifies players less likely to be drafted, then the incorrect results could lead to suboptimal decisions, misallocation of resources, and potential disappointment among stakeholders. The model will not be reliable and cannot be used to make predictions. Businesses' goal of predicting the players to be drafted in NBA will remain unfulfilled as the model predicts the incorrect result.

1.b. Hypothesis	Hypothesis: The AdaBoost model can effectively predict the players to drafted in NBA(target) by analysing various independent features such as player performance metrics, technicalities of player's game etc. The hypothesis I want to test is whether specific performance statistics of college basketball players can be used to predict their likelihood of them being drafted into the NBA. Specifically, I want to explore if metrics such as scoring efficiency, rebounding, assists, defensive metrics, and overall playing time, along with other metrics, significantly impact a player's draft status. Considering this hypothesis is worthwhile because it can potentially bring data-driven insights to the NBA draft process, benefiting teams, players, scouts, and fans. It aligns with the industry's increasing reliance on analytics and addresses the complexities of selecting and predicting the success of future NBA players. The insights gained from this project can contribute to more informed decisions and a deeper understanding of the factors influencing the transition to professional basketball.
1.c. Experiment Objective	I expect the model to give results in favour of my Hypothesis. However, it may have the following possible outcomes: 1. Positive outcome: The AdaBoost classifier can identify significant predictors. The model's roc au score is high. The model can be used by the business to predict potential NBA players. 2. Negative outcome: The AdaBoost classifier cannot identify significant predictors. The model's roc au score is low. The business cannot use the model to predict potential NBA players. 3. Inconclusive outcome: The AdaBoost classifier identifies some significant predictors, but the model's roc au score is low or insignificant enough to be actionable.

---

2. EXPERIMENT DETAILS
Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation	

2.a. Data Preparation

For this experiment, as I am using AdaBoost model,

I performed the following data preparation steps:

1. Checked for null values across all the features. It was identified that :

   a. The features ‘pick’ and 'Rec_Rank' have a high percentage of missing values of more than 97% and 69%, respectively. Therefore, I did not consider these features for my analysis.

   b. The columns ‘ast_tov’, ‘rimmrade’, ‘rimmade_rimmiss’, ‘mid_made’, ‘midmade_midmiss’, ‘dunksmade’, ‘dunksmiss_dunksmade’, 'drtg', 'adrtg', 'dporpag', 'stops', 'bpm', 'obpm', 'dbpm', 'gbpm', 'mp', 'ogbpm', 'dgbpm', 'oreb', 'dreb', 'treb', 'ast', 'stl', 'blk' and 'pts' having missing values were substituted with their mean values.

   c. For the columns having ratios, the missing values were calculated using the below formulas.
      - rim_ratio = rimmade / rimmade_rimmiss
      - mid_ratio = midmade / midmade_midmiss
      - dunks_ratio = dunksmade / dunksmiss_dunksmade

   d. The columns 'yr' and 'num' contain a lot of irregularities. Since the data is unreliable. Thus I have decided to remove these columns.

2. Data splitting
   I separated the dependent feature (‘drafted’) into y. I have split the training set in 80:20 into the training and validation set. I have specifically split this ratio so there is enough training data. I will be using the validation set to calculate roc_auc_score.

   The X_feaures contains player performance metrics, technicalities of the player and draft attributes. Thus, I have considered them to be relevant for the analysis.

   X_features: 'Min_per', 'usg', 'eFG', 'TS_per', 'ORB_per', 'DRB_per', 'AST_per', 'FT_per', ''twoP_per', 'TP_per', 'porpag', 'ast_tov', 'rim_ratio', 'mid_ratio', 'dunks_ratio', 'dporpag', 'stops', 'bpm', 'obpm', 'dbpm', 'gbpm', 'ogbpm', 'dgbpm', 'oreb', 'dreb', 'treb', 'ast', 'stl', 'blk', 'pts', 'ht'

3. Resampling of Training data
   The dependent feature is highly imbalanced, wherein 0 is more than 99%, and 1 is less than 1%. Class 1 represents if the player is drafted in NBA, and class 0 represents the player is not drafted. Data resampling is important here as there is a significant class imbalance, as one class represents only a small fraction of the dataset. Thus, the model can be biased towards the majority class, making it difficult to predict the minority class accurately.
   Resampling the Minority class,
   I have performed oversampling of minority class(1) using SMOTE(Synthetic Minority Over-sampling Technique) in the training dataset only.

2.b. Feature Engineering	Dropped features: The features ‘pick’ and 'Rec_Rank' have a high percentage of missing values of more than 97% and 69%, respectively. Therefore, I did not consider these features for my analysis. The columns 'yr' and 'num' contain a lot of irregularities. Since the data is unreliable. Thus I have decided to remove these columns Categorical features: I have not considered categorical features such as ‘team’, ‘conf’. Handled ‘ht’ column: The datatype was object and the values were represented in date format. eg. 2-June-2023. I converted them in int datatype (feet.inches)as described below: 2/06/2023 —> 2062023 —> 6022023 (M/d/Y) —-> 6.02 (removed the year i.e., M/d) Which implies to 6 Feet 2 Inches Feature scaling: I have performed feature scaling on all independent features (training, validation and testing sets) as It ensures that all features contribute to the model fairly and that the algorithms operate effectively regardless of the original scale of the data. It improves model performance, convergence, and the ability to make meaningful predictions by removing any distortions caused by feature scales.
2.c. Modelling	An AdaBoost model was harnessed to forecast the probability of a college basketball player's selection for the NBA draft, leveraging their statistical performance metrics. AdaBoost is a well-regarded ensemble learning method recognized for its notable proficiency in addressing intricate data associations. The selection of AdaBoost as the central modeling technique stemmed from its resilience and its capacity to navigate the intricacies inherent in the data. XGBoost has been used for modelling before. It did perform well but to further improve te performance I have used another boosting algorithm.Models such as Linear Regression and Naive Bayes were deemed less suitable for this specific project due to the dataset's inherent complexities and the nuanced relationships expected in predicting NBA draft outcomes. Automated Hyperparameterization using the Grid Search learning_rate = 0.5, n_estimators=200

3. EXPERIMENT RESULTS	
Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.	
3.a. Technical Performance	Automated Hyperparameterization generated using: Grid Search learning_rate = 0.5, n_estimators=200 ROC AUC score on validation set: 0.9335 The ROC AUC score of 0.9335 for the AdaBoost model on the validation set indicates strong discriminative ability and predictive performance, suggesting that the model effectively distinguishes between classes. In practical terms, based on their features and statistics, the AdaBoost model effectively separates the drafted players from the non-drafted players. The predictions on the probability of players being drafted or not have been made for the X_test_scaled for each player_id.
3.b. Business Impact	The predictive model achieved a strong ROC AUC score of 0.9335, indicating its effectiveness in distinguishing between drafted and non-drafted college basketball players based on their performance and technical statistics. However, incorrect predictions can have varying impacts. False positives could lead to missed opportunities for players, while false negatives might overlook talented players. Balancing both error types is crucial for credibility and ensuring deserving players aren't overlooked. The model's results influence team selections, player aspirations, fan expectations, and team performance, underscoring the need for continuous refinement and evaluation of the model's predictions.
3.c. Encountered Issues	Throughout the experiment, several challenges were encountered and addressed, while some remain open for future consideration: Feature Scaling: Variability in feature scales was managed using Min-Max scaling for consistency. Missing Data: Missing values were imputed using relevant methods, like median imputation for ratio features. Imputing Nan values with 0 Feature Engineering for ‘ft’ column Hyperparameter Tuning: Time-consuming hyperparameter optimization was handled through systematic GridSearchCV. Model Selection: AdaBoost

---

4. FUTURE EXPERIMENT

Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	The experiment provided valuable insights into predicting NBA draft outcomes using machine learning. The models, particularly AdaBoost, demonstrated strong predictive power with high ROC AUC scores. Key player statistics showed significant correlations with draft outcomes. Hyperparameter tuning enhanced model performance, and addressing missing data improved data quality. The approach shows promise and warrants further experimentation. Exploring ensemble and other boosting methods, class imbalance handling, and advanced feature engineering could enhance results. Model interpretability and real-world deployment are areas for future focus. The current approach is not a dead end. Continued refinement and adaptation are essential for practical applications.
4.b. Suggestions / Recommendations	Steps to be performed in future: Ensemble Models: Experiment with ensemble methods like stacking or blending different models to further enhance predictive accuracy. Planning to use lightGBM. Expected Uplift: Moderate. Feature Engineering: Explore more advanced feature engineering techniques, leveraging domain knowledge to create novel features that capture player performance nuances. Expected Uplift: Moderate to High. Model Interpretability: using various metrics for feature importance and decision-making. Expected Uplift: Moderate. Validation on Unseen Data: Test the model's performance on unseen data from subsequent NBA draft years to assess generalisation. Expected Uplift: Moderate. Refining the feature engineering, which would enhance the predictions of the model. lightGBM may be used for modelling, which will likely improve the model's performance further. Deployment Steps: Model Evaluation: Ensure the model meets validation and unseen data performance thresholds. Interpretability: Implement techniques to understand feature importance and decision logic. Infrastructure Setup: Prepare the deployment environment with the necessary packages. Scalability Test: Assess the model's scalability for real-time or batch predictions. Model Deployment: Choose a deployment approach (API, microservices, cloud). Monitoring & Maintenance: Continuously monitor, retrain, and update data. Feedback Loop: Gather insights and iterate on model improvements with stakeholders. Deploy if the model achieves the required outcomes and shows robustness in testing. Maintain regular monitoring and updates to align with dynamic NBA draft trends.

GitHub Link: https://github.com/bhutanisaumya/Adv_MLA.git