EXPERIMENT REPORT

Student Name	Shi Wu
Project Name	Part3- Stacking (LightGBM+Random Forest)
Date	28/08/2023
Deliverables	https://github.com/frodorocky/adv_mla_ass1/tree/third_experiment

---

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	Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results? This notebook will use binary classification model to predict whether a college basketball player will be drafted into the NBA based solely on season statistics. This model helps NBA teams decide who to watch in person, plan for the draft, and see if their own ratings match the predictions. A good model helps teams scout smarter, but a bad one might lead them to miss out on talent or waste resources.
1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: College basketball players with higher season statistics are more likely to be drafted into the NBA. Question: Do college basketball players' season statistics directly influence their chances of being drafted into the NBA? We aim to determine the correlation between a college player's season performance and their probability of being drafted into the NBA. Understanding the link between college statistics and draft probability can greatly optimize the processes and decisions related to scouting, investing, and planning in the NBA.

1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The primary expected outcome is to determine a quantifiable relationship between college basketball players' season statistics and their likelihood of being drafted into the NBA. It's difficult to identify a specific objective given the wide range and abundance of factors that affect the draught. But let's imagine that based on a player's season stats, we want to be able to forecast with 80% accuracy whether they would be selected. Possible scenarios resulting from this experiment include: ● High Correlation: The model might find that players' season statistics have a strong correlation with their draft status. In this scenario, NBA teams can significantly rely on statistical data to make their draft decisions. ● Moderate Correlation: The model could indicate that while statistics play a role in draft decisions, other factors (like a player's physical attributes, potential, or team needs) also carry significant weight. In this case, teams would use the model's predictions as supplementary data in their overall scouting and decision-making processes. ● Low or No Correlation: The model may determine that there's little to no correlation between season statistics and draft status. This would mean that while statistics are informative, many other aspects influence draft decisions, and teams should not heavily rely on stats alone. ● Overfitting or Inconsistencies: The model might show a high accuracy during training but perform poorly on new, unseen data. This suggests that while it might have learned patterns from the historical data, it may not be generalized enough for real-world predictions. ● Variable Importance: The model might reveal that only certain statistics (e.g., points per game or efficiency ratings) are predominantly influencing draft

---

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	Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments 1. The dataset is loaded and explored to understand the data structure, missing values, and basic statistics. 2. We dropped those variables with more than 50% missing values, replaced all the missing value with 'unknown' for categorical features. 3. And we imputed missing values in the numeric features with the mean value of the respective feature. 4. We dropped one variable with single values
2.b. Feature Engineering	Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments 1. OrdinalEncoder was applied to all discrete variables. 2. We add the following features: avg_points_per_game(pts/GP), Ortg_diff_from_mean(ortg-ortg_mean), usg_Ortg_interaction(usg*Ortg), Ortg_square(ortg^2), pts_rolling_avg 3. We dropped 12 features with correlation greater than 0.95. 4. We keep the features that cumulative importance of 0.99

2.c. Modelling	Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

During the third experiment we kept the same preprocessing or feature engineering as previous experiment, but we employed a Stacking Ensemble technique that combines predictions from two primary models: LightGBM and Random Forest, used a logistic regression model as the meta-learner.

For LightGBM:

- We're tuning n_estimators to determine the number of boosting rounds, ranging from 100 to 1000.
- The learning_rate is being optimized among values 0.01, 0.05, and 0.1 to control the boosting step size.
- num_leaves is the number of leaves in a tree, and we're considering 31, 62, and 93 as potential values.
- We're experimenting with max_depth to constrain the depth of the trees.
- min_child_samples specifies the minimum samples required in a leaf, with values ranging between 20 and 40.
- We're also tuning L1 (reg_alpha) and L2 (reg_lambda) regularization.

For Random Forest:

- We're adjusting n_estimators, which denotes the number of trees in the forest, from 100 to 1000.
- We're experimenting with the maximum depth of the trees using max_depth.
- min_samples_split and min_samples_leaf are being tuned to determine the minimum samples required to split an internal node and for a leaf node, respectively.

The meta-learner (Logistic Regression):

- The inverse of regularization strength C is being optimized among values ranging from 0.001 to 10.
- We're considering both L1 and L2 penalties through the penalty parameter.
- The algorithm to be used in the optimization problem is being determined by the solver parameter, with potential choices being 'liblinear' and 'lbfgs'.

For hyperparameter tuning, instead of an exhaustive grid search, we utilized Random Search. This method samples a fixed number of parameter settings from the specified

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	Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes. The result AUC is 0.98483, which means the stacked model has a slightly lower testing AUC compared to the individual LightGBM models. This suggests that stacking did not provide a benefit in this scenario, at least in terms of AUC.
3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) With the insights from the model, teams can now make more informed draft decisions using clear metrics, leading to optimal player selections. These insights also enable coaches to tailor their training regimes, focusing on areas pinpointed by the model for enhanced player development. Additionally, scouts benefit by using the model to zero in on players who shine in specific areas, streamlining and refining their scouting processes. However, incorrect results may lead to the evaluation of players based on a wrong system, which may lead to the team assigning an unreasonable development plan and wasting resources eventually.
3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be deal with in future experiments. 1. With stacking, where multiple models are involved, explaining predictions becomes more complex. For instance, if we're stacking a random forest and a gradient boosting model and using a linear regression as a meta-model, understanding a prediction would require considering the outputs from all these models. 2. Stacking involves training multiple models, which demands more computational resources and time. So we used random search instead of grid search to find

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	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. We used feature importance to interpret the result of model, for LightGBM we obtained the most important features: dporpag, dunksmade, gbpm, porpag. For random forest we obtained the most important features: dporpag, portage, gbpm, bpm. ● dporpag & porpag: These seem like metrics related to a player's performance or some aspect of the game. Without knowing their exact definitions, we might infer that they represent some kind of efficiency or productivity metric. If these features are deemed important, it means that a player's efficiency or productivity in the game has a strong influence on whether they get drafted or not. ● dunksmade: This indicates how many dunks a player has made. Dunks could be seen as a display of athleticism and skill. If this is an important feature, it suggests that players who demonstrate higher athleticism (through dunks) have a higher likelihood of getting drafted. ● gbpm & bpm: these might represent some metrics related to a player's overall contribution or effectiveness in the game. If these are important, they indicate that a player's overall effectiveness or value in a game is a crucial factor in the drafting decision.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. In this experiment we employed a Stacking Ensemble technique that combines predictions from two primary models: LightGBM (LGBM) and Random Forest (RF), used a logistic regression model as the meta-learner. We obtained an model with AUC 0.98488. Since there is no improvement on ensemble tree models, thus we may try train deep learning model regardless of insufficient samples.