EXPERIMENT REPORT

Student Name	Sahil Kotak
Project Name	36120 Advanced Machine Learning Application Spring 2023 - Assignment 1
Date	08-09-2023
Deliverables	● kotak_sahil-24707592-week4_xgboost_model_3.ipynb ● XGBoost with HyperOpt

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 overarching objective remains unchanged: This project aims to predict which college basketball players are likely to be drafted into the NBA based on their statistics from the current season of college basketball. This prediction can be used by NBA teams to make informed decisions during the draft process, by sports analysts to provide insights into the draft prospects, and by the players themselves to gauge their likelihood of being drafted.

Accurate results will improve draft decisions, increase team performance, and more insightful draft analysis. Inaccurate results may lead to suboptimal draft decisions, decreased team performance, and misleading draft analysis. The stakes are high, as a team's success in the NBA can have significant financial implications, including higher ticket sales, increased merchandise sales, and more lucrative sponsorship deals. Therefore, developing a reliable and accurate predictive model for this purpose is crucial.

1.b. Hypothesis

Given the class imbalance in our dataset, we hypothesized that using the XGBoost model, which is inherently adept at handling such challenges, in conjunction with HyperOpt for hyperparameter tuning and SMOTE for oversampling, would enhance our model's performance.

Rationale:

XGBoost: Known for its efficiency and capability to handle large datasets with higher dimensionality, XGBoost can maximize computational speed and performance. It also inherently manages class imbalance by adjusting weights.

HyperOpt: Hyperparameter tuning is crucial for the optimal performance of machine learning models. HyperOpt automates this process, seeking the best hyperparameters for a given model and dataset, potentially boosting our model's performance.

SMOTE: With the observed class imbalance, there's a risk our model might be biased towards the majority class. By synthetically generating samples for the minority class using SMOTE (Synthetic Minority Over-sampling Technique), we aim to create a more balanced dataset, potentially leading to a more generalized model.

1.c. Experiment Objective

Our continual goal with these experiments is to enhance the predictive accuracy of our models. For this week, the focus is on determining if the combination of XGBoost, HyperOpt, and SMOTE can outperform our previous week's ensemble model and exceed an AUROC score of 0.99933.

Possible Scenarios:

- **Successful Outcome:** Our model, with the integration of XGBoost, HyperOpt, and SMOTE, surpasses the previous AUROC score of 0.99933 and demonstrates good generalization on the test set.

- **Partial Success:** While there's an improvement in the AUROC score, the model might be showing signs of overfitting, indicating the need for further adjustments or reconsideration of our oversampling technique.

- **No Improvement:** Despite the changes, the model's performance remains stagnant, suggesting that XGBoost with the given hyperparameters and SMOTE might not be the best fit for this particular dataset.

- **Degradation:** The model underperforms compared to previous models, indicating that the additions might be introducing more noise, or that the model complexity might need reconsideration.

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

Steps Taken:

- Retained Steps from Previous Weeks: Continued with the data preparation steps from Weeks 1, 2, and 3.
- Year Column: The 'year' column, which was previously dropped, was retained this week to allow the model to recognize player statistics trends over different years.
- Mapping Conferences to Regions: The 'conferences' were mapped to their respective 'regions', and a new feature 'region' was created to represent this information.

Rationale:

- Year Column: Including the 'year' column allows the model to capture temporal trends and changes in player statistics over time. This might provide the model with better context, especially if the performance metrics or playing style has evolved over the years.
- Mapping Conferences to Regions: Grouping conferences by region can help the model identify regional patterns in player performance, providing another

layer of context to the player statistics.

Steps Not Taken:

- Feature Elimination: Similar to last week, even though feature importance techniques were considered, we did not proceed with feature elimination.

2.b. Feature Engineering

Steps Taken:

- Retained Steps from Previous Weeks: Continued with the feature engineering steps from Weeks 1, 2, and 3.
- Anomaly Scores: Generated anomaly scores for each row in the dataset using the Isolation Forest algorithm, similar to last week.

Rationale:

- Anomaly Scores: Drafted players are outliers in the dataset, showcasing exceptional skills and performance. Anomaly scores help quantify this uniqueness, aiding the model in identifying potential draft picks.
- Features Removed: None.

Future Considerations:

- Interaction Features: As mentioned last week, there's potential in exploring interaction terms among features that show high correlation with the target variable in upcoming experiments.

2.c. Modelling

Models Trained:

XGBoost with HyperOpt and SMOTE: A sophisticated hyperparameter optimization setup was implemented using HyperOpt for the XGBoost model.

Hyperparameters Tuned:

XGBoost:

- Learning rate
- Number of estimators
- Maximum depth
- Minimum child weight
- Gamma
- Subsample ratio
- Colsample bytree
- Regularization alpha

Rationale:

- XGBoost: Known for its efficiency and capability to handle both linearity and non-linearity in data, XGBoost is a versatile model that can be adapted to various datasets.
- HyperOpt: Hyperparameter tuning plays a critical role in achieving optimal model performance. HyperOpt provides an automated way to search through the hyperparameter space, potentially leading to better results than traditional grid search or random search.

Models Not Trained:

- Linear Models and Simpler Models: Given the focus on XGBoost this week and its capability to capture complex relationships in the data, simpler models or linear models were not explored.

Highlight for Future Experiments:

The XGBoost model's performance with hyperparameter tuning via HyperOpt will set the stage for upcoming experiments. Depending on its performance, there might be further avenues to explore in terms of feature engineering, data preparation, or even other boosting techniques.

---

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

Models Trained:

XGBoost with HyperOpt and SMOTE: This model was the primary focus for the week. It utilized the XGBoost algorithm, combined with HyperOpt for hyperparameter optimization and SMOTE for oversampling, considering the class imbalance in our dataset.
Scores:

Validation Score: The XGBoost model with hyperparameter tuning achieved an impressive AUROC score of 0.99955 on the validation set.

Test Score (Kaggle): The model's performance was consistent when evaluated on the Kaggle test set, achieving an AUROC score of 0.99918.

Analysis:

The scores achieved this week surpass those from the previous week, especially with the stacking ensemble model. The minuscule difference between validation and test scores indicates that the model is generalizing well, and overfitting seems to be controlled.

However, the scores are very close to 1, which may raise concerns about whether the model is too tailored to this dataset or whether there's any data leakage. Future experiments might focus on ensuring robustness by using techniques like cross-validation or bootstrapping.

3.b. Business Impact

From a business perspective, the results of this week's experiments further emphasize the robustness and accuracy of our predictive capabilities. An AUROC score near perfection suggests the model can be an invaluable tool for NBA teams, analysts, and players.

However, as mentioned previously, while the model's statistical performance is

	outstanding, its practical application should be considered. The complexity of the XGBoost model, although justified by its performance, needs to be weighed against its interpretability and the computational resources required for real-world deployments.
3.c. Encountered Issues	● Hyperparameter Space Search: The HyperOpt algorithm searches for the best hyperparameters over a vast space, which can sometimes lead to convergence issues or longer training times. ● Solution: The search space was defined with feasible ranges for each hyperparameter, ensuring efficient convergence. Additionally, setting a maximum number of evaluations for HyperOpt helped in bounding the search time. ● Class Imbalance: The dataset's inherent class imbalance can bias the model towards the majority class. ● Solution: SMOTE was utilized to synthetically generate samples for the minority class, ensuring a balanced dataset for model training. **For Future Experiments:** ● Model Interpretability: Given the increasing complexity of the models being used, future experiments might focus on model interpretability solutions, such as SHAP or LIME, to understand the model's decision-making process better. ● Data Augmentation: Considering the high performance, we might look into more advanced data augmentation techniques, ensuring the model is learning genuine patterns rather than memorizing the dataset. ● External Datasets: Incorporating data from external sources could provide additional context and improve model performance. For instance, data on player injuries, personal achievements, or even social media presence could be influential. ● Regularization & Cross-Validation: As the scores are inching closer to perfection, it becomes essential to ensure that the model isn't overfitting. Techniques like regularization can be emphasized, and more extensive cross-validation strategies can be employed to validate model performance across various data splits.

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	**Power of Boosting:** The XGBoost model demonstrated superior performance, emphasizing the strength of boosting algorithms in handling both linear and non-linear patterns. Its ability to handle class imbalances, combined with HyperOpt for hyperparameter tuning, turned out to be a potent combination. **Consistency Across Datasets:** The minuscule difference between validation and Kaggle test scores indicates that the model has generalized well. This consistency is a good sign, ensuring that the model is likely to perform well on future, unseen data. **Approaching Model Limitations:** The model's scores are inching closer to perfection. While on one hand, this is an indicator of excellent performance, on the other hand, it raises questions about overfitting or the model being too tailored to the dataset. **Future Direction:** While the current approach is yielding great results, it might be

	approaching its performance ceiling. Thus, future experiments might need to focus on aspects other than just raw performance, such as interpretability, model robustness, and ensuring no data leakage.
4.b. Suggestions / Recommendations	**1. External Data Incorporation:** Rationale: Introducing external data sources can provide additional context. For instance, player injuries, personal achievements, or even social media presence might influence draft decisions. Expected Uplift: Moderate to High. New data dimensions can unlock previously unseen patterns and relationships. **2. Cross-Validation and Regularization:** Rationale: To ensure that the model isn't overfitting and to validate its performance over different data splits, techniques like k-fold cross-validation can be employed. Regularization can prevent overfitting, especially given the near-perfect scores. Expected Uplift: Moderate. Ensuring the model's robustness might not lead to significant performance gains but will make it more reliable. **3. Production Deployment:** Rationale: With scores nearing perfection, it might be time to consider deploying the model into a production environment, where real-time draft predictions can be made. This will involve steps like model serialization, setting up APIs for predictions, and integrating with existing systems. Expected Uplift: High (From a business perspective). Taking the model live will allow for real-world validation and provide tangible business value.