EXPERIMENT REPORT

Student Name	Sahil Kotak
Project Name	36120 Advanced Machine Learning Application Spring 2023 - Assignment 1
Date	25-08-2023
Deliverables	• kotak_sahil-24707592-week2_ensemble_model_2.ipynb • Logistic Regression & SVM Stacking Model

---

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	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	**Hypothesis and Objective:** The central hypothesis we aimed to test was that given the potential linear nature of the dataset, simpler models like logistic regression can perform on par with more complex models like XGBoost. Given that, an ensemble of linear models, either through averaging or stacking, could enhance predictive performance. **Rationale:** Exploring this hypothesis is worthwhile for several reasons. Firstly, simpler models like logistic regression are more interpretable, enabling better understanding and trust in the model's decisions. If their performance is comparable to complex models, they might be preferred for their transparency. Secondly, ensemble methods can leverage the strengths of multiple models, potentially leading to improved prediction scores and robustness. Investigating the effectiveness of ensembling linear models can guide whether the added complexity is justified by a notable increase in performance. By conducting this experiment, we aim to identify an optimal balance between model complexity, interpretability, and predictive power.

1.c. Experiment Objective	The experiment aimed to assess the performance of logistic regression, SVM, and ensemble methods predicting the “drafted” column. Given the dataset's potential linear characteristics, we hypothesized that simpler models might yield results close to complex algorithms like XGBoost. **Expected Outcome:** We anticipated: Similar or better performance using logistic regression and intricate models. Enhanced AUROC scores via ensemble techniques, possibly outdoing standalone models. **Possible Scenarios:** ● **Scenario A:** Logistic regression and SVM match the performance of XGBoost, underscoring the dataset's linear patterns. ● **Scenario B:** Ensembling substantially improves results, highlighting the benefit of model combination. ● **Scenario C:** Simpler models lag behind XGBoost, suggesting non-linear dataset patterns. ● **Scenario D:** Ensembles offer minimal gains, questioning the utility of the added complexity for this dataset.

---

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	Along with the steps from Week 1 experiments, the following additional steps were taken: ● **Height Transformation:** The ‘ht’ column had a formatting issue which had turned the raw data into “Jun-00”, “Jul-02”, etc values. We took steps to fix the formatting issues and convert the height into inches. ● **Handling Missing Values:** Missing data in columns like 'midmade_midmiss', and 'dunksmade' were imputed using median values from the training dataset. ● **Year Mapping:** The 'yr' column's string labels were mapped to numerical values (e.g., 'Fr' to 1). ● **Dropping Unnecessary Columns:** Columns such as 'num', 'pick', and 'Rec_Rank' were dropped. **Rationale:** ● Transforming heights into a uniform numerical format enables better analysis and model training. Handling missing values ensures models won't fail during training or prediction due to NaN values. Mapping 'yr' maintains ordinality. **Important for Future Experiments:** ● Consideration of advanced feature engineering, and potential binning of certain columns for categorization.

2.b. Feature Engineering	For feature engineering, we encoded categorical variables and normalized numerical features. Categorical variables were one-hot encoded, and numerical features were scaled using Standard Scaler. No additional features were created for this experiment. For the future experiments, we would want to focus on this aspect more.
2.c. Modelling	**Models Trained:** ● **XGBoost:** Based on the baseline results of the last week, we wanted to see how the dataset would perform on a complex algorithm like XGBoost. ● **Logistic Regression & SVM:** Given the dataset's potential linear nature, these simpler models were chosen. They are easily interpretable and can perform comparably to complex models on linearly separable data. ● **Ensemble Techniques:** We combined logistic regression and SVM predictions via averaging and stacking to harness the strengths of multiple models. **Hyperparameters Tuned:** ● XGBoost using RandomizedSearchCV: The best configuration included a learning rate of 0.05, depth of 10, 500 trees, subsample of 0.7, gamma of 0.25, and 90% columns sampled by tree, ensuring a balanced and efficient model. **Highlight for Future Experiments:** The superior AUROC performance of the Stacking Ensemble model compared to the previous week's model and the XGBoost model reaffirms our hypothesis about the dataset's predominantly linear nature. This insight is crucial for directing future modeling strategies and experiments.

---

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	**Scores:** ● SVM: 0.9906 ● Ensemble (Averaging): 0.9913 ● Stacking: 0.9913 ● XGBoost: 0.9903 **Analysis:** All models performed exceptionally well, with AUROC scores hovering close to 1. The Stacking model slightly outperformed the rest, though the difference between the Ensemble and SVM models was marginal. XGBoost, while still yielding a high score, was the least performant among the tested models.

	Underperforming Observations & Potential Causes: **XGBoost's Slight Underperformance:** Given that XGBoost is a complex tree-based model, it might be capturing minor non-linear patterns. However, our hypothesis leans towards the data's linear nature, which might give simpler models like SVM a slight edge. **Marginal Difference between Ensemble and Stacking:** This indicates that while stacking does add a meta-learner layer, the benefit over simple averaging in this dataset's context might be minimal.
3.b. Business Impact	Referencing the previous week's insights, the experiments in the current week continue to underscore the model's potential as a powerful predictive tool for NBA draft decisions. The high AUROC scores, especially from the Stacking and Ensemble models, further instill confidence in our predictive capabilities. However, business implications go beyond just model performance. The minute differences in AUROC between the models, although statistically significant, need to be weighed against practical considerations. For instance, the added complexity of the Stacking model over the SVM might not justify its marginally better performance in a real-world setting.

3.c. Encountered Issues	• **Complexity vs. Performance Trade-off**: Advanced models like XGBoost and Stacking were introduced, but the increment in AUROC, compared to the simpler SVM, was marginal.
	  ○ **Implication:** This reiterates the challenge of ensuring we're not adding unnecessary complexity without a justified improvement in performance. Future experiments should always weigh the benefits of marginal performance gains against model complexity and computational costs.
	• **Data Transformation Issues**: The 'ht' column presented varied formats, requiring custom transformation to a consistent numeric format.
	  ○ **Solution**: A dedicated function was designed to handle these variations and convert heights to inches. Continuous validation of this transformation is essential as datasets evolve.
	
	For Future Experiments:
	
	• **Hyperparameter Exploration:** While we tuned the XGBoost model, there's scope in the future to explore an even broader set of hyperparameters or use more advanced tuning methods.

---

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	**Linear Nature of Data:** The consistently high performance of the linear SVM model strongly suggests that our data exhibits a linear trend. The limited benefit seen from more complex models like XGBoost or Stacking Ensemble further corroborates this observation. **Trade-off between Complexity and Performance:** The minimal performance

	increment achieved with complex models compared to simpler models like SVM indicates that adding complexity doesn't always translate to significant performance improvements. It's crucial to evaluate the trade-off between model interpretability, computational costs, and performance gains. **Ensemble Benefits:** The ensemble method, even simple averaging, provided a slight boost in performance. This indicates the potential benefits of combining model predictions to harness diverse learning patterns. **Future Direction:** Given the insights, it's worth iterating on the current approach, especially around ensemble models, to harness the strengths of individual models.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, potential that we want to explore: **Refine Ensemble Techniques:** - Rationale: While simple averaging provided a boost, more sophisticated ensemble methods or different weight allocations could further enhance performance. - Expected Uplift: Moderate. Combining models might harness diverse strengths, potentially boosting accuracy by a few points. **Feature Engineering and Selection:** - **Rationale:** Re-evaluating and engineering features might help in capturing underlying patterns better. Feature selection can also help in reducing overfitting. - **Expected Uplift**: Moderate to High. Effective feature selection can significantly improve model performance and reduce overfitting.