EXPERIMENT REPORT

Student Name	Vishal Raj
Project Name	NBA Draft Prediction
Date	17th August 2023
Deliverables	Vishal_Raj-14227627-week1_RF1stExp .ipynb Random Forest

---

1. EXPERIMENT BACKGROUND	
1.a. Business Objective	The primary goal of this project is to predict whether a college basketball player will be drafted to join the NBA league based on his statistics for the current season. This prediction is of paramount importance for various stakeholders: • **NBA Teams**: Accurate predictions can guide team managers and coaches in their draft decisions, helping them identify potential star players or hidden gems. • **Sports Commentators and Analysts**: Having a data-driven model can provide commentators with insights to enrich their discussions and analyses during draft events. • **Players and Their Coaches**: For college players aspiring to join the NBA, understanding their draft likelihood can help them focus on areas of improvement. • **Fans**: NBA fans are always eager to know about the next big player. Predictions can enhance their engagement and discussions around draft events. The impact of accurate predictions is multifaceted. It could mean the difference between securing a future star player or missing out on one for the NBA teams. For players, it could provide clarity on their career trajectory. However, incorrect predictions could lead to misinformed decisions, potential missed opportunities for teams, and misplaced expectations for players.
1.b. Hypothesis	Given the statistics of a college basketball player for a particular season, it is hypothesised that certain performance metrics can be indicative of a player's likelihood of being drafted into the NBA. Historically, players with higher points, assists, rebounds, and other key metrics have been observed to have a higher chance of being drafted. This hypothesis is grounded in the belief that NBA teams look for players who have demonstrated exceptional skills and performance during their college years. By analysing these statistics, we aim to uncover patterns and relationships that can predict a player's transition from college basketball to the NBA.

1.c. Experiment Objective	This experiment aims to build a predictive model to determine the probability of a player being drafted accurately. The expected outcome is an AUROC score that is significantly higher than random guessing (0.5). By achieving this, we hope to provide a tool to assist various stakeholders, from sports analysts to NBA teams, in making informed decisions about potential draft picks. The possible scenarios resulting from this experiment include: • **Highly Accurate Model**: This would validate our hypothesis and provide a reliable tool for predicting draft outcomes. • **Moderately Accurate Model**: While imperfect, this would still offer valuable insights and could be improved with further experimentation. **Inaccurate Model**: This would indicate that other factors, possibly outside of the provided statistics, play a significant role in draft decisions.

---

2. EXPERIMENT DETAILS	
2.a. Data Preparation	Data preparation is crucial in any machine learning project, ensuring that the data fed into the model is clean, relevant, and structured appropriately. For this experiment: • **Data Loading**: Multiple datasets were sourced, including **train.csv**, **test.csv**, **sample_submission.csv**, and **metadata.csv**. Each dataset provided unique insights into player statistics and draft outcomes. • **Handling Missing Values**: The presence of missing values can skew predictions and reduce the reliability of a model. The **pick** column, which had a high percentage of missing values, was dropped to maintain data integrity. Other missing values were addressed using appropriate imputation techniques like using median and mode, ensuring no data point was left out due to incomplete information. • **Data Splitting**: The data was divided into training and validation sets to evaluate the model's performance. This ensures that the model is tested on unseen data, objectively assessing its predictive capabilities.
2.b. Feature Engineering	Feature engineering enhances the predictive power of the data by creating new features or transforming existing ones: • **Feature Selection**: Key performance metrics, such as **pts**, **ast**, **treb**, and **eFG**, were chosen to be plotted as histogram, relevance to a player's performance and likelihood of being drafted. Sports analysts often highlight these features when discussing a player's prowess. In future experiments, they might be specifically selected for model training. • **Class Imbalance**: The data revealed an imbalance in the distribution of the target variable **drafted**. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. This ensures the model is not biased towards the majority class and can make accurate predictions for both classes. • **Feature Scaling**: Standard scaling was applied to the features to ensure they all contribute equally to the model's predictions. This is especially important for models like Random Forest, which can be sensitive to the scale of input

	features.
2.c. Modelling	The choice of model and its configuration plays a pivotal role in the success of the experiment: • **Model Selection**: A Random Forest Classifier was chosen for its versatility and ability to capture complex relationships in the data. Its ensemble nature aggregates predictions from multiple decision trees and offers robustness against overfitting. • **Hyperparameter Tuning**: To extract the maximum performance from the Random Forest model, hyperparameters were fine-tuned using GridSearchCV. This exhaustive search tested various combinations of parameters, identifying the optimal configuration for the highest ROC-AUC score. Parameters like **bootstrap**, **max_depth**, and **n_estimators** were among those tuned to enhance the model's predictive power.

---

3. EXPERIMENT RESULTS	
3.a. Technical Performance	The model's technical performance was evaluated using the ROC-AUC score, a metric that measures the ability of the model to distinguish between the positive and negative classes: • **Initial Model Performance**: Without hyperparameter tuning, the initial model achieved an impressive ROC-AUC score of 0.9999818147225285. This score indicates that the model has a near-perfect ability to differentiate between players who will be drafted and those who won't. • **Optimized Model Performance**: After hyperparameter tuning, the model's performance improved marginally, achieving a ROC-AUC score of 0.9999825194526576. The best parameters identified were **bootstrap**: False, **max_depth**: None, **min_samples_leaf**: 1, **min_samples_split**: 2, and **n_estimators**: 200. • **Performance Analysis**: The exceptionally high scores suggest that the model has effectively captured the underlying patterns in the data. However, it's essential to be cautious and ensure the model is not overfitting the training data.
3.b. Business Impact	From a business perspective, the model's predictions can have significant implications: • **Draft Decisions**: NBA teams can leverage the model's predictions to inform their draft decisions, potentially identifying star players or hidden talents. • **Player Development**: College players and their coaches can use the model's insights to focus on areas of improvement, increasing their chances of being drafted. • **Potential Risks**: While the model's predictions are highly accurate, it's crucial to consider the potential risks of relying solely on the model. Incorrect predictions could lead to missed opportunities for teams or misplaced

	expectations for players.
3.c. Encountered Issues	During the experimentation phase, several challenges and issues were encountered: • **Data Imbalance**: The target variable, **drafted**, had an imbalanced distribution. This could have led the model to be biased towards the majority class. The issue was addressed using the SMOTE technique. • **Feature Selection**: Deciding which features to include in the model was challenging. While some features were obvious choices due to their relevance to player performance, others required deeper analysis and experimentation. • **Model Complexity**: The high ROC-AUC scores raised concerns about potential overfitting. Various regularisation techniques and model configurations were tested to ensure the model's generalisation capability.

---

4. FUTURE EXPERIMENT	
4.a. Key Learning	The experiment provided several valuable insights and learnings: • **Model Performance**: The exceptionally high ROC-AUC scores achieved by the model indicate that player statistics from their college season are strong predictors for their likelihood of being drafted into the NBA. • **Data Imbalance**: Addressing class imbalance is crucial for ensuring the model is not biased towards the majority class. Techniques like SMOTE can be instrumental in handling such imbalances. • **Feature Importance**: Not all features contribute equally to the model's predictions. Understanding the importance of each feature can guide future data collection and feature engineering efforts. • **Overfitting Concerns**: While high performance is desirable, ensuring that the model does not overfit the training data is essential. Regularisation techniques and cross-validation can help in assessing and improving the model's generalisation capability.
4.b. Suggestions / Recommendations	Based on the results and learnings from the experiment, the following recommendations are made for future work: • **Optimized Coding**: To make the codebase more maintainable and efficient, consider refactoring critical parts of the code into modular classes and functions. This improves readability and allows for easier testing and validation of individual components. • **Ensemble Classification**: While the Random Forest model is itself an ensemble method, combining predictions from multiple models can often lead to better performance. Techniques like stacking or blending can be explored, combining predictions from various models to produce a final prediction. This can help in capturing diverse patterns and relationships in the data. • **Feature Engineering**: Explore additional features or transformations that might improve the model's predictive power. For instance, player performance trends over multiple seasons could provide more context.

	● **Model Exploration**: Beyond ensemble methods, other models like Gradient Boosting or Neural Networks could be explored to see if they offer any performance improvements. ● **Deployment**: Given the high performance of the model, steps should be taken to deploy it as a tool for NBA teams, sports analysts, and other stakeholders. This could be in the form of a web application or API. ● **Feedback Loop**: Once the model is used, a continuous feedback loop can be established to improve it. As actual draft decisions are made, this data can be used to further train and refine the model.