# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Aibarna Singh Basnet |
| **Project Name** | AT1 |
| **Date** | 18/08/2023 |
| **Deliverables** | basnet\_aibarna-24585717-week1\_randomforest\_feature\_selection.ipynb  https://github.com/aibarna96/adv\_mla\_at1 |

|  |  |
| --- | --- |
| 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?  The goal of this project is to develop a predictive model that classifies basketball players as potential draft picks for the upcoming season based on specific attributes. The intended business outcome is to aid basketball teams in optimizing their player recruitment process by identifying prospects with higher draft potential. The model's accurate predictions will assist teams in efficiently allocating resources, forming strategically balanced rosters, and gaining a competitive edge. However, incorrect predictions could lead to misallocation of resources and missed opportunities for team improvement. The project's success lies in providing actionable insights for effective player selection, thereby enhancing team performance and overall success. |
| **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,  We aim to investigate whether specific player attributes, such as scoring efficiency, defensive performance, and overall productivity, have a substantial impact on a basketball player's likelihood of being drafted in the next season. Our hypothesis suggests that these attributes play a crucial role in draft prospects. By exploring the correlations and predictive power of these attributes, we aim to uncover insights that could enhance player recruitment strategies. This investigation is worthwhile as accurate predictions could optimize resource allocation and inform recruitment decisions, ultimately contributing to team success. |
| **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. |

|  |  |
| --- | --- |
| 1. **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  For data preparation, we began by loading the training dataset and transforming column names to lowercase for consistency. We then selected relevant features with strong correlations to the target variable "drafted." As missing values were present, we filled them with zeros for simplicity. The data was split into features and the target variable, followed by a division into training and testing sets. To address class imbalance, we employed randomoversampler to oversample the minority class. We standardized the features using StandardScaler to ensure model effectiveness. While we treated missing values by filling with zeros, other advanced techniques like imputing could be explored for improved data quality. Additionally, feature engineering, interactions, or polynomial features might enhance model performance. The rationale for these steps is to ensure data is suitable for model training and that the model is provided with relevant information. |
| **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  In the process of feature generation for the NBA draft prediction model, several steps were executed. First, the SelectKBest method was used to choose the top five features most correlated with the target variable "drafted." This step aimed to streamline the model by focusing on the most pertinent player statistics. Then, feature standardization was performed to ensure all features contributed equally to the model. Balancing the dataset was another crucial step, accomplished using RandomOverSampler to address the class imbalance in the training data. No features were explicitly removed, but feature selection focused on the top five. Looking forward, future experiments should consider advanced feature engineering, outlier handling, categorical variable encoding, and hyperparameter tuning to enhance model performance further. These steps and considerations lay the foundation for refining the model in the future and improving NBA draft predictions. |
| **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  For this NBA draft prediction experiment, the primary model trained was a Random Forest Classifier. The choice of Random Forest was motivated by its ability to handle both classification tasks and feature importance evaluation. Random Forest can capture complex relationships within the data and is robust to overfitting.  Hyperparameter tuning was performed on the Random Forest model. The key hyperparameters tuned included `n\_estimators` (number of trees in the forest) and `max\_depth` (maximum depth of each tree). Values tested for `n\_estimators` ranged from 50 to 300, and for `max\_depth`, values from 5 to 30 were explored. This range allowed for the assessment of model performance under various tree and ensemble settings.  Other models, such as Support Vector Machines (SVM) and Logistic Regression, were not trained in this experiment. SVM might be considered in future experiments, as it can handle complex decision boundaries. Hyperparameter tuning and other ensemble methods, like Gradient Boosting and AdaBoost, could also be explored to potentially enhance model performance further. |

|  |  |
| --- | --- |
| 1. **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 relevant performance metric for this experiment is the Area Under the Receiver Operating Characteristic (AUROC) score, which measures the model's ability to distinguish between drafted and undrafted players. Unfortunately, the AUROC score is not provided in the results.  However, based on the classification report, the model performs exceptionally well in identifying undrafted players (class 0) but has room for improvement in identifying drafted players (class 1). The main underperforming cases are likely those where the model predicted a player to be drafted but they were not, resulting in lower precision for class 1. The root cause could be the class imbalance in the dataset, where there are significantly more undrafted players than drafted ones. Additionally, the features selected for prediction might not capture all the nuances that lead to a player being drafted, which could be explored further in future experiments. |
| **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)  The results of the experiment indicate that the model has a very high accuracy of approximately 99.37%. However, this high accuracy is primarily driven by the model's ability to correctly identify undrafted players, which make up the majority of the dataset. The model's performance in identifying drafted players, though, is less impressive, with a lower precision and recall.  In the context of the business objective, the impact of incorrect results is significant. Misclassifying a player as "drafted" when they are not could lead to false hopes and expectations for both the player and their team. Conversely, failing to identify a player who should be drafted could mean missed opportunities for NBA teams to acquire talent. Therefore, while the high accuracy is promising, it's crucial to improve the model's ability to correctly identify drafted players, as these cases have more substantial business implications. |
| **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 dealt with in future experiments.  Throughout the experiments, several challenges were encountered and addressed. These included handling class imbalance by oversampling, feature selection using SelectKBest, and the choice of model, RandomForestClassifier. However, hyperparameter tuning, exploring a broader range of models, and evaluating with additional metrics like precision-recall and F1-score remain future considerations. Data preprocessing was basic, missing more sophisticated techniques. Feature engineering and interpretable models were also unexplored. Deployment strategies, scalability, and ethical considerations for real-world use need further attention. These issues provide a roadmap for future experiments, emphasizing the need for more comprehensive approaches to improve model performance and applicability. |

|  |  |
| --- | --- |
| 1. **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.  The experiment's outcome provides valuable insights into the predictive capabilities of the model for identifying drafted basketball players. The experiment highlighted the challenges of achieving a balanced precision-recall trade-off, especially given the rarity of drafted players. The insights gained include the importance of refining the model's precision, investigating additional features, and exploring advanced techniques for handling class imbalance. Despite achieving a promising AUROC score and demonstrating potential for enhancing team recruitment, the limitations in precision call for further experimentation. The current approach is not a dead end; instead, it signals the need for iterative refinement. By conducting more experiments, testing alternative algorithms, and exploring diverse feature sets, the model's performance can likely be improved, offering more reliable predictions for player drafting and contributing to more effective decision-making in professional basketball team management. |
| **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.  Potential Next Steps and Experiments:  1. \*\*Feature Engineering Refinement:\*\* Explore additional player attributes that might have a strong impact on draft outcomes, such as player statistics, team performance, and player rankings. Expected Uplift: Moderate. This step could lead to improved model performance by incorporating more comprehensive player profiles.  2. \*\*Algorithm Selection and Hyperparameter Tuning:\*\* Experiment with different classification algorithms (e.g., gradient boosting, support vector machines) and perform hyperparameter tuning to find the best model configuration. Expected Uplift: Moderate. Selecting the most suitable algorithm can enhance predictive accuracy and reliability.  3. \*\*Ensemble Methods:\*\* Investigate ensemble methods, such as stacking or blending multiple models, to leverage the strengths of different algorithms. Expected Uplift: Moderate. Ensemble methods can potentially yield more accurate and robust predictions.  4. \*\*Advanced Sampling Techniques:\*\* Explore advanced sampling techniques beyond SMOTE to further address class imbalance. Expected Uplift: Moderate. Leveraging state-of-the-art sampling methods could lead to better balance in precision and recall.  5. \*\*Temporal Analysis:\*\* Perform temporal analysis to determine if draft criteria have evolved over time, necessitating model updates. Expected Uplift: Low to Moderate. This could offer insights into the changing dynamics of player drafting.  6. \*\*External Data Integration:\*\* Incorporate external data sources, such as player background information and injury history, to enrich the dataset. Expected Uplift: Moderate. Additional context might improve the model's predictive power.  7. \*\*Domain Expert Consultation:\*\* Collaborate with basketball experts to gain insights into domain-specific factors that contribute to draft decisions. Expected Uplift: Low to Moderate. Expert insights could inform feature selection and model development.  Ranking and Recommendations:  Based on the potential next steps, the ranking and recommendations are as follows:  1. Feature Engineering Refinement  2. Algorithm Selection and Hyperparameter Tuning  3. Ensemble Methods  4. Advanced Sampling Techniques  5. Temporal Analysis  6. External Data Integration  7. Domain Expert Consultation  If the experiment achieved the required outcome for the business (high precision, recall, and F1-score), the next steps would involve deploying the solution into production. This would entail the following steps:  1. \*\*Model Deployment:\*\* Choose a production environment (e.g., cloud platform) and deploy the trained model to make real-time predictions.  2. \*\*Monitoring and Maintenance:\*\* Implement monitoring mechanisms to track model performance over time. Regularly retrain the model with updated data to ensure its relevance.  3. \*\*Integration:\*\* Integrate the model's predictions into the team's recruitment process, providing decision-makers with actionable insights.  4. \*\*Feedback Loop:\*\* Establish a feedback loop where the model's predictions and actual outcomes are continuously compared. This loop informs model refinement and improvements.  5. \*\*Continuous Improvement:\*\* Based on feedback and performance, iterate on the model by incorporating new insights, features, and techniques to enhance its accuracy and value to the business. |