# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Arthur Su |
| **Email** | arthur@infinitydigital.com.au |
| **Project Name** | NBA rookie prediction |
| **Date** | 29/11/2024 |
| **Deliverables** | https://github.com/arthurso2023/ADSI\_Assignment.git |

|  |  |
| --- | --- |
| 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 primary business objective of this project is to develop a robust machine learning model capable of accurately predicting the long-term viability of NBA rookie players. By analyzing a comprehensive dataset of rookie season statistics, the model aims to identify key performance indicators that correlate with a successful 5-year or longer career in the NBA. This predictive capability will provide valuable insights to NBA teams, enabling them to make more informed decisions during the draft process, player evaluation, and contract negotiations. Ultimately, this model will contribute to improving team performance and long-term 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,  The hypotheses is that there is a statistically significant relationship between a rookie player's statistical performance and their long-term career longevity in the NBA. This knowledge can be valuable for teams in making informed decisions during the draft process and player development. |
| **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.  **Expected Outcomes:**  A successful machine learning model for predicting NBA rookie longevity  **Possible Scenarios:**   1. **High Accuracy Model:**    * **Positive Impact:** Teams can confidently identify potential stars and allocate resources accordingly.    * **Negative Impact:** Overreliance on the model might lead to overlooking other important factors, such as intangibles or unforeseen circumstances. 2. **Moderate Accuracy Model:**    * **Positive Impact:** The model can still provide valuable insights, especially when combined with traditional scouting methods.    * **Negative Impact:** Incorrect predictions could lead to missed opportunities or misallocated resources. 3. **Low Accuracy Model:**    * **Positive Impact:** The model might still be useful for identifying outliers or extreme cases.    * **Negative Impact:** Relying heavily on the model could lead to poor decision-making and decreased team performance. |

|  |  |
| --- | --- |
| 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  The following data preparation has be conducted on the dataset.   * Check for duplicated records * Check for missing value in records * Remove rows containing negative values (in theory there should be no negative attempts or achieves.) * Create a balanced dataset using under-sampling method (the original dataset contains 83% players who has career length >=5 years) |
| **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  Two methods are used to select the features:   * Manually select feature by comparing the mean values and according to my knowledge of basketball games (which is not good at all). * Check feature importance after training the ML models and then filter out the less important features. |
| **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  The experiment involved training machine learning models on three different datasets: the original dataset, a cleared dataset, and a balanced dataset. While the first two datasets yielded high accuracy for Random Forest and ExtraTree models (over 83%), a closer examination revealed a significant issue: the models were unable to accurately predict players with less than 5 years of career. This problem arose because the first two datasets were heavily biased towards players with over 5 years of career, leading to skewed model performance.  Therefore, I chose models trained based on the balanced dataset. While some models had accuracy lower than the baseline accuracy of the balanced dataset (less than 50%), Random Forest (64.6%) and ExtraTree (63.9%) models demonstrated the best performance, achieving an accuracy of over 60%. |

|  |  |
| --- | --- |
| 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 final machine learning models, primarily Random Forest and ExtraTree, while achieving accuracies over 60% on the balanced dataset, did not reach a level of high accuracy. The challenge was exacerbated by the difficulty in identifying key features from the data due to limitations in data understanding and domain knowledge. The lack of information regarding data collection and calculation methods further hindered the analysis. Additionally, the absence of deep basketball knowledge limited the ability to perform effective feature engineering. It is concluded that a strong foundation in domain knowledge is crucial for selecting appropriate features and algorithms to build a robust predictive model. |
| **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)  A moderately accurate model can still provide some value to NBA teams:   * **Identifying Potential Stars:** The model can help identify promising rookies who may have a long and successful career. * **Allocating Resources:** Teams can use the model to allocate resources (e.g., coaching staff, training facilities) more effectively. * **Making Informed Decisions:** The model can provide additional insights to supplement traditional scouting methods.   However, relying solely on the model's predictions can be risky. It's essential to combine the model's insights with human judgment and domain expertise. |
| **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.  The experiment encountered several challenges that hindered the development of a highly accurate predictive model for NBA rookie longevity.  **Time Constraints:**   * Limited time for thorough data exploration, feature engineering, and model experimentation.   **Data Quality and Understanding:**   * Lack of clarity regarding data collection and calculation methods, making it difficult to assess data quality and identify potential biases. * Insufficient domain knowledge to effectively interpret and preprocess the data.   **Model Complexity and Algorithm Limitations:**   * The need to balance model complexity with the limited dataset size. * Constraints on the range of machine learning algorithms that could be explored due to computational limitations and time constraints. |

|  |  |
| --- | --- |
| 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.  To improve future results, it is essential to allocate sufficient time for data exploration, feature engineering, and model selection, as well as to collaborate with domain experts to gain deeper insights into the data. |
| **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.  To improve the model's performance and impact, I suggest the following:   * Collect more data: A larger and more diverse dataset can help the model learn more complex patterns. * Incorporate domain expertise: Collaborate with basketball analysts to identify relevant features and interpret the model's results. * Experiment with different algorithms: Explore other machine learning algorithms, such as neural networks or XGBoost, to potentially improve performance. * Regularly retrain the model: As new data becomes available, the model should be retrained to maintain its accuracy. |