# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Shakif Sattar Ani |
| **Project Name** | shakif.sattar@gmail.com |
| **Date** | 03/12/2024 |
| **Deliverables** | https://github.com/Shakif2024gH/adv\_dsi\_lab\_1/tree/assignment/assignment/notebooks |

|  |  |
| --- | --- |
| 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 predict if an NBA rookie will remain active in the NBA after five years. Accurate predictions can provide valuable insights for teams regarding talent acquisition, player investments, and long-term team development strategies. Accurate predictions can help reduce risks in drafting decisions, while incorrect results could lead to suboptimal roster management and financial loss. |
| **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 hypothesis is that certain player statistics, such as games played (GP), minutes played (MIN), and scoring efficiency (PTS per MIN), are good predictors of an NBA rookie's long-term career prospects. This hypothesis is based on the assumption that player efficiency and experience correlate strongly with career longevity. |
| **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.  The objective of this experiment is to develop a machine learning model capable of predicting whether a player will stay in the NBA for at least five years, based on their rookie-year statistics. The desired outcome is a model with moderate to high performance metrics (e.g., AUC score above 0.70), capable of reliably distinguishing between players who will and will not have long careers in the league. |

|  |  |
| --- | --- |
| 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 dataset was cleaned to handle missing values, and duplicates were removed. Feature scaling was performed using MinMaxScaler to bring all numeric features into the 0-1 range. This scaling ensures that features are comparable in magnitude, which is crucial for many machine learning models. SMOTE (Synthetic Minority Over-sampling Technique) was applied to address class imbalance, where the target class had a significant bias towards players remaining active in the NBA. |
| **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  A new feature, 'PTS per MIN' (Points per Minute), was created to capture scoring efficiency. This feature helps the model evaluate players based on their effectiveness in scoring during their playing time, which is a more meaningful indicator than total points. Existing features such as 'GP' and 'MIN' were retained, as they are critical to understanding player activity and experience. |
| **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  Two models were trained: Random Forest and Logistic Regression.   * Random Forest was chosen for its robustness and ability to handle complex interactions. Grid Search optimised parameters like n\_estimators, max\_depth, min\_samples\_split, and min\_samples\_leaf. * Logistic Regression was used as a simpler, interpretable baseline. Grid Search identified the best hyperparameters: C=100, penalty='l2', and solver='saga'.   A baseline model was also trained using a DummyClassifier with a 'most\_frequent' strategy to provide a reference point for model performance. |

|  |  |
| --- | --- |
| 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 Random Forest model achieved an AUC score of **0.7025**, which suggests a reasonable ability to distinguish between players who will and will not be active in the NBA after five years. The training accuracy was **0.8645 (86.45%)**, and the validation accuracy was **0.8337 (83.37%)**, indicating that the model generalises well without significant overfitting. However, the confusion matrix showed a high number of false positives compared to true negatives, indicating that the model struggles to correctly classify players who will not stay in the league.  The Logistic Regression model slightly outperformed it with an AUC of **0.7170** but faced significant challenges in detecting unsuccessful players, showing high recall for successful players but poor precision and recall for the negative class. Both models demonstrated limitations in accurately identifying players unlikely to succeed in the NBA. |
| **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)  Both models moderately meet the business objective of predicting NBA player success, but their limitations pose risks. The Random Forest model **(AUC: 0.7025)** and Logistic Regression model **(AUC: 0.7170)** struggle with high false positives, leading to overestimations of player success. These errors could result in costly investments in underperforming players. Additionally, challenges in detecting unsuccessful players may cause missed opportunities to identify risky candidates, impacting drafting efficiency and financial outcomes. Improvements are necessary to better align model predictions with business needs. |
| **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.   * **Class Imbalance**: The dataset was highly imbalanced, which affected the model's performance. SMOTE was applied to address this, but further experimentation may be needed to improve model accuracy. * **False Positives**: The model had difficulty correctly identifying players who would not remain active, leading to a high number of false positives. Adjusting class weights or trying alternative models may help address this issue. * **Hyperparameter Optimisation**: While Grid Search was utilised, it may not have fully explored the parameter space. Future experiments could apply more advanced optimisation techniques, such as Bayesian optimisation. |

|  |  |
| --- | --- |
| 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 demonstrated that player efficiency metrics like GP, MIN, and PTS per MIN are important predictors of career longevity. However, the challenge of class imbalance and false positives highlights the need for further refinement of the model. Additional data or advanced techniques like ensemble methods could improve performance.  Logistic Regression outperformed Random Forest (AUC: 0.7170 vs. 0.7025), but both models struggled with class imbalance and high false positives. Further experimentation with resampling techniques, ensemble methods, or additional data is needed to improve performance. |
| **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.   * **Hyperparameter Tuning**: Consider using RandomizedSearchCV to explore a broader range of hyperparameters without the high computational cost of Grid Search. * **Alternative Models**: Experiment with other algorithms, e.g., as Decision Tree or Kmeans, which may handle imbalanced datasets better. * **Feature Selection**: Investigate further feature engineering and selection to improve model performance, focusing on features that reduce false positives. * **Domain knowledge**: Retrieve data from experts from the industry to assist in narrowing down relevant features and add where possible. |