# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Amy Yang, Chanthru Vimalasri, Yatindra Vegunta (Updated) |
| **Project Name** | NBA\_Career\_Prediction\_Week1 |
| **Date** | 15/11/2022 |
| **Deliverables** | data\_processing.ipynb  random\_forest.ipynb  Random Forest with Random Search  AUROC |

|  |  |
| --- | --- |
| 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?  **Goal:** To predict if a rookie player will last at least 5 years in the NBA league based on their current stats.  **Use:** The results will help the business direct their resources towards players with the greatest talent and potential.  **Impact of accurate/inaccurate results**: With accurate results, the NBA league can be more selective and they will be able to cultivate more successful players and teams. This can increase profits for the business by encouraging viewership, sponsorship deals, merchandise sales etc. and heighten the status of their ‘brand’. Inaccurate results could have an adverse impact on players' career paths. Those with the potential to do well may be weeded out from the league unnecessarily. |
| **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 insights we are seeking revolve around our goal of predicting if a rookie player will last at least 5 years in the NBA league based on their current stats. As this is our first experiment, we will prioritize exploratory analysis and data preparation to better understand and contextualize the data. Models with automatic hyperparameter tuning (e.g. Random Forest with Random Search) will be utilized to improve time efficiency over trial and error. This will allow us to establish clear baselines from quick yet accurate predictions that will fuel further investigation over the coming weeks. |
| **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.  By performing data exploration and cleansing, we expect that the unreasonable values or outliers are removed from the training data before the model is built. The improved data quality will help to fit a model with better prediction capability and provide a higher AUROC score.  We expect the Random Forest model to perform better than the baseline. It will likely overfit but hyperparameter tuning with Random Search should mitigate this. Regardless of our model’s performance, we hope to gain useful insight that will better steer our experimental process moving forward.  A Logistic Regression model was also on the Data to identify if the model would outperform the Random Forest. |

|  |  |
| --- | --- |
| 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 steps taken for data preparation are as follows:   * Load the dataset and display the first 5 rows * Present the summary of each column and dimension of the data frame * Display the descriptive statistics * Visualise the dataset   It is of importance to perform these steps as they allow for the initial data exploration to see what columns are included in the dataset and disclose data issues which may have great impact on feature selection and the performance of the prediction model. For example, the summary of each column suggests if there are any missing values in any column and their data types. The descriptive statistics highlight the potential outliers and unreasonable values of each feature and provide guidance on the data cleansing step. The visualisation provides further insight on the distribution of the target variable and the correlation between each feature. These graphs imply the existence of common data issues - imbalance data and collinearity. Some of the steps we decide not to do include extracting column names (df.columns), displaying data types (df.dtypes) as they are already included in the summary by using df.info().  The steps that may potentially be important for future experiments include the count of the target variable when it is equal to 0 or 1 respectively, which indicates the necessity of treating the imbalanced data with future steps or techniques such as SMOTE. The correlation matrix reveals the potential issue of collinearity in features which is worth further exploration in the coming weeks as we ascerain its impact on the model development. |
| **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  Based on the initial data exploration discussed in 2.a, the following features are removed from the original dataset:   * Id * 3P Made * 3PA * 3P% * BLK   As the Id column is only used to identify the order of the dataset and does not have any relevance to the target variable, we decide to remove this column to ensure only relevant features are used to fit the model. The rationale to remove the other features above (3P Made, 3PA, 3P% and BLK) is because the unreasonable values in these columns are significant compared to the total training data size. For example, there are 1,628 rows with negative values in the column of ‘3P Made’. If we remove all these records from the dataset, it means around 20% of data size will be lost and can not be used to build the model. Instead of reducing the data size, removing the whole column will be a better choice under this circumstance. Similar reasons can explain why the other features (3PA, 3P% and BLK) are removed as well.  Some features we would like to explore include MIN, PTS, FGM and FGA. According to the correlation matrix plot, these features are highly correlated. We would like to explore if removing some features among them will make a model have better performance. A similar experiment will be for features OREB, DREB and REB. Another feature that may be interesting to craft and explore is the total points for games which is equal to GP multiplied by PTS. |
| **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   * A Random Forest model was chosen. As this was our first opportunity to explore and analyze the data, we wanted a model that both would yield accurate results but still provide flexibility. Random Forest is especially useful because it can handle binary, categorical and numerical features. There is very little pre-processing that needs to be done and it can handle large datasets efficiently. * It is however prone to overfitting so automated hyperparameter tuning was performed via Random Search. This yielded the following ‘best\_parameters’ - max\_depth': 10, 'min\_samples\_leaf': 46, 'n\_estimators': 186. * Alternative classification models such as KNN were not chosen as they tend to be more computationally inefficient and manual with the value of K being selected through trial and error. With a dataset we are unfamiliar with, an automated form of tuning such as Random Search is much more time-efficient. * A drawback of using this Random Forest model is that it doesn’t clearly allow us to investigate the relationship between the features and target variable. Logistic Regression with L1 regularization may prove useful in future * A Logistic regression was also run to identify if it does any better than random forest, however the score was much worse than what was compared to random forest. * I would need to use a better way to use the Logistic regression model, where I think may not have done it accurately |

|  |  |
| --- | --- |
| 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.  Baseline AUROC: 0.50  Random Forest (with Default Hyperparameters) AUROC (Training Set): 1.00  Random Forest (with Default Hyperparameters) AUROC (Validation Set): 0.67  Random Forest (with Random Search) AUROC (Training Set): 0.78  Random Forest (with Random Search) AUROC (Validation Set): 0.71  Logistic regression Scores:  Accuracy: 0.84  Recall: 0.99  Precision: 0.84   |  |  |  |  |  | | --- | --- | --- | --- | --- | | CL Report: | precision | recall | f1-score | support | | 0 | 0.33 | 0.02 | 0.04 | 248 | | 1 | 0.85 | 0.99 | 0.91 | 1340 | | accuracy |  |  | 0.84 | 1588 | | macro avg | 0.59 | 0.51 | 0.48 | 1588 | | avg | 0.77 | 0.84 | 0.78 | 1588 | |
| **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)  Our baseline model performs very poorly and struggles to predict between classes within the imbalanced dataset. Our initial expectations were correct as our Random Forest models performed better than our Baseline AUROC. Without hyperparameter tuning however, it is clearly overfitting with a score of 1.00 for the training set. Whilst it can perfectly predict the probability of a rookie player lasting longer than 5 years in the NBA league with the given data, it performs significantly worse with unseen data as evidenced through the validation AUROC of 0.67. The implementation of this model would result in a misallocation of resources towards new players that may not have as much potential to succeed, affecting the business in terms of future revenue and reputational value.  Automatic hyperparameter tuning was performed via Random Search to reduce this overfitting. Overfitting has been reduced on the training set with an AUROC score of 0.78. The AUROC validation score has also been improved to 0.71, resulting in better predictive ability. |
| **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.   1. Issues encountered during data preparation and feature engineering were listed and discussed in sections 2.a. and 2.b. 2. Overfitting with Random Search. Countered with Hyperparater tuning - Random Search. 3. Collinearity - Points Per Game (PTS), Field Goals Made (FGM), Field Goals Attempts (FGA) are all highly correlated at around 97-99%. Similar case for (Free Throw Made) FTM and (Free Throw Made) FTA. This is understandable as the number of free throws made are most likely to be increased with more free throw attempts. Other features with high collinearity include (Offensive Rebounds) OREB, (Defensive Rebounds) DREB and (Rebounds) REB. Random Forest is not overly affected by collinearity between features but forms of mitigation will need to be explored as we consider other models for future experiments. 4. Imbalanced Data - The number of rows where target =1 or 0 is around 6:1. Thus, there is a potential issue for imbalanced target variables. Methods to combat this skew in distribution of data will need to be investigated for future experiments. 5. I could not spend enough time on building and exploring the model, due to the issue with Internet access, as raised with you earlier. |

|  |  |
| --- | --- |
| 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.  Our random forest model has performed well in comparison to the baseline. Hyperparameter tuning via Random Search proved effective in improving our AUROC score from 0.67 to 0.71 and significantly reducing overfitting with our training data from 1.0 to 0.78. This has provided useful insight about the tuning of hyperparameters and reinforced its significance for future experimentation.  With our current AUROC score, we can say our model is able to correctly predict between classes in the imbalanced dataset most of the time. Our current approach seems effective but to further improve upon this we will need to delve deeper to gain a more thorough understanding of the dataset and the relationship between the target variable and features. Improved domain knowledge may also allow us to hone our feature engineering and model selection to achieve better predictive ability.  I need to invest more time into these experiments and get up to speed with the upcoming experiments |
| **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 include:   1. Improving domain knowledge on NBA basketball rules to better understand the relationship between features and what may constitute unreasonable/invalid values. Including additional features such as GP (Games Played) x PTS (Points per Game) = Total Points Played may help improve the model’s predictive ability too. 2. Counteracting the imbalanced nature of the data by performing stratified sampling as opposed to the random sampling we did when splitting the data. We could also try resampling (down sampling/ up sampling), SMOTE or utilizing a SVC model that provides hyperparameters to adjust class weights. 3. Attempt Logistic Regression with L1 regularization with the purpose of better understanding the relationship between the target variable and features. 4. Improve the time-efficiency of our model process by creating functions for data processing and splitting our data. 5. Utilize accuracy in combination with AUROC to test our model’s performance. Useful insights can be derived from analyzing their respective scores (e.g. low AUROC, high accuracy will help direct our attention towards first resolving the imbalanced data issue) |