EXPERIMENT REPORT

Student Name	Monali Patil
Project Name	AT1 - Kaggle Competition
Date	18/08/2023
Deliverables	Notebook name: Patil_Monali-14370946-week1_LR.ipynb Model name: Logistic Regression Classifier GitHub Repo: https://github.com/MonaliPatil19/adv_mla_assignment1

---

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: The project involves building a binary classifier model that predicts whether a college basketball player will be drafted into the NBA league based on their statistics records. The dataset provided contains a wide range of features that illuminate players' performance during their college basketball season comprising 64 players' performance related attributes.

Application: This event captures the attention of sports commentators, fans, and enthusiasts who eagerly track the careers of college players and speculate about their chances of being selected by NBA teams. So, the model results will be used to assist sports commentators, fans, and teams in understanding the likelihood of a player's selection.

Impact: The model's accurate predictions can provide valuable insights for both players and teams, aiding decision-making during the NBA draft process. Additionally, this model will offer valuable insights to sports commentators, fans, and scouts, aiding them in predicting the potential NBA draft prospects of individual players. Incorrect predictions, on the other hand, might lead to missed opportunities for both players and teams, potentially impacting the overall performance and composition of NBA teams.

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, Hypothesis: To test whether the statistical attributes of college basketball players from their college season can effectively predict whether they will be drafted into the NBA league. The key question would be, can a machine learning model accurately predict whether a college basketball player will be selected in the NBA draft based on their performance metrics? In this experiment, employing logistic regression classifier, the aims is to assess the predictive power of these player statistics and evaluate whether they hold valuable insights into a player's likelihood of getting drafted. Rationale: Considering this hypothesis is worthwhile because the NBA draft is a significant event in a basketball player's career, and accurately predicting draft selection can have multiple benefits. For teams, it can provide a data-driven approach to identify potential talents and make informed decisions during the draft process. For players, it offers insights into the factors that influence their selection, potentially guiding their training and career decisions. Additionally, for fans and commentators, accurate predictions can enhance engagement and discussions around player performances and draft prospects. The use of statistical attributes also aligns with the growing trend of data-driven decision-making in sports, making this experiment relevant and valuable.
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 expected outcome of this experiment would be the development of a binary classifier model utilizing Logistic Regression Classifier algorithm that can accurately predict whether a college basketball player will be drafted to the NBA league based on their college season statistics. The model would be trained on a dataset of college players information and their performance statistics, and it would use Logistic Regression Classifier algorithm to identify patterns and make accurate predictions. The goal of this experiment would be to identify potential player for an annual event aimed at identifying the players to be drafted in the NBA league utilising Logistic Regression Classifier algorithm. The potential scenarios arising from this experiment include the following situations: • The model accurately predicts players who will be drafted. This provides teams with valuable insights for selecting promising talents, potentially improving their performance and team success. • The model inaccurately predicts players who will be drafted. Teams may miss out on potential talents, impacting their competitiveness. False positives rate can lead to misallocation of resources. • If the model performs moderately, there's an opportunity to fine-tune it. However, the model's marginal performance would be similar to random guessing. • The model is not accurate enough to be useful for predicting players position, and the experiment is unsuccessful. In this case, the NBA team may need to explore other methods for identifying potential players for their annual event.

2. 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 players' statistical data was explored, cleaned, and prepared through the subsequent activities for the binary classification algorithm to be employed.

• Data Understanding

1] Loading Data: Imported data from the given CSV file into the panda's data frames to use and create a binary classification model.

2] Exploring Data: Examined and studied players' information using different pandas functions to comprehend and uncover patterns, aiming to identify prospective players for NBA draft selection.

* Additionally, to ensure the quality of the data to be utilized by the model analysis.
- Handing missing/null values.
- Eliminating identifiers.
- Duplicate records.
- Data distribution of various features.
- Accessing if imbalance targets classes.

df.head(): Checking initial records of the dataset.
df.shape(): Verifying the dimension of the dataset.
df.columns: Identifying attributes name.
df.info(): Assessing the summary information of the attributes of the dataset.
df.describe(include='all'): Examining statistical summary information for all variables of the dataset across different data types.
df.isnull().sum(): Examining whether there are any null values in the dataset.
df.duplicated().sum(): Identifying whether there are any duplicate values in the dataset.

• Data Preparation

3] Handling Missing Values

A significant number of 33 attributes in the training dataset and 24 in the testing dataset possess missing values, necessitating addressing to ensure data readiness for modelling as machine learning algorithms are incapable of processing features with missing values.

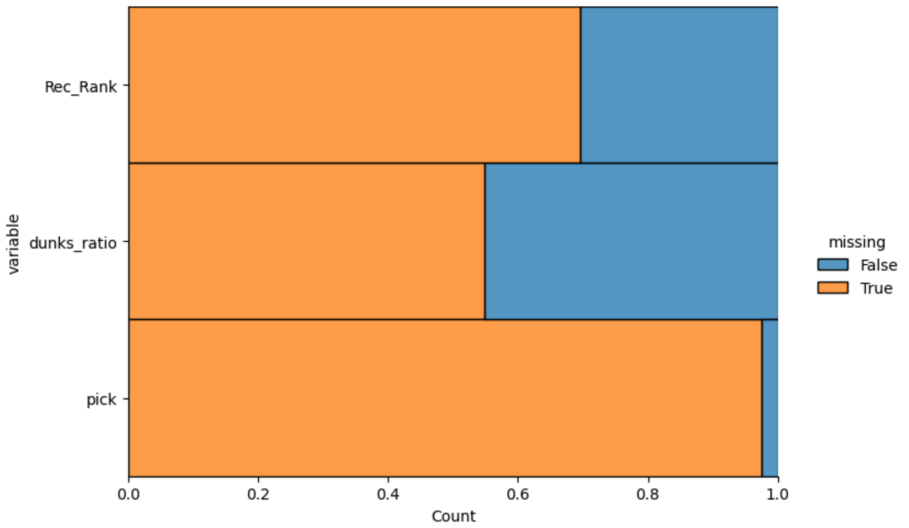

Figure 1: Attributes with majority of missing values.

* Among the features with missing data, the following three attributes exhibit notably high levels of missing values, exceeding 50%.

- Rec_Rank: Training -> 69.55% (39055/56091) and Testing -> 71.15% (3536/4970)
- dunks_ratio: Training -> 54.89% (30793/56091) and Testing -> 54.64% (2717/4970)
- pick: Training -> 97.64% (54705/56091) and Testing -> 98.98% (4921/4970)

* Therefore, it is practical to exclude these features in order to prevent potential biases in the model arising from imputation.

* Given that the remaining features have missing values comprising less than 2%, it is a reasonable approach to fill these missing values using the mean for numerical attributes and the mode for categorical attributes.

4] Verifying and Removing Identifiers

* Attributes 'player_id' and 'num' are unique identifiers for each basketball player, and its inclusion in the analysis can lead to overfitting, where the model fits to these specific values rather than the underlying generalized patterns in the sportsman's records.

* Additionally, 'team' and 'conf' attributes representing the names of teams and conferences, are overly specific and do not contribute to the model's general learning.

* Therefore, removed 'player_id', 'num', 'team' and 'conf' features from the datasets.

5] Analysing correlation between predictors and target variable

* Performing correlation analysis to determine which features would be most valuable for building a predictive model for the NBA draft prediction task.

* Correlation using heatmap, calculates the correlation between each feature and the target variable 'drafted' to understand how much they influence the target. Features with high correlations to the target variable are potentially valuable and are selected as predictors.

6] Accessing if imbalance targets classes

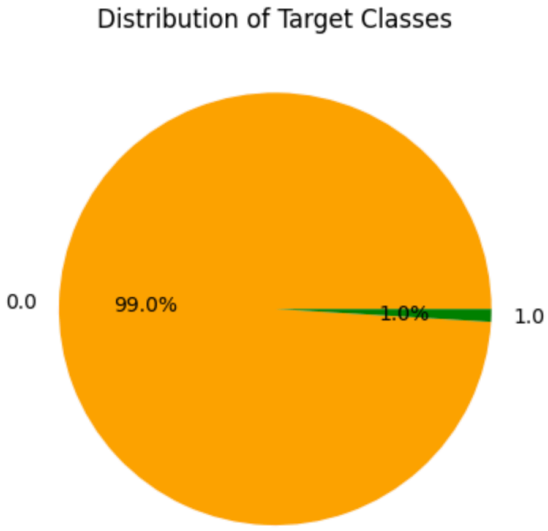

Figure 2: Target classes distribution.

* The pie chart above illustrates a significant class imbalance within the dataset. The majority of observations are attributed to a single target class, representing players who have not been drafted denoted by the value 0.

* Performed Oversampling with SMOTE (Synthetic Minority Over-sampling Technique) method to address class imbalance by generating synthetic observations for the minority class.

* It works by creating synthetic samples that are similar to existing observations in the minority class without duplicating the existing observations, rather it generates new observations by interpolating between existing ones.

7] Features Scaling

* Using feature scaling prevents the algorithm from prioritizing high-value features over other more informative ones. It ensures uniformity in feature values, enabling the algorithm to learn generalized patterns from all features for accurate player identification and predictions.

* Employed StandardScaler method because it maintains the features data distribution's shape and retains outliers by scaling data using the mean of 0 and standard deviation of 1 across the entire dataset, rather than for individual data points.

8] Hyperparameters Tuning

* Employed Hyperparameter Tuning to determine optimal values for the hyperparameters of the Logistic Regression Classifier algorithm to further improve the performance of the model.

* Utilizing the Random Search technique for hyperparameter tuning, which introduces randomness by selecting values from the search space. This can lead to combinations of hyperparameter values that minimize errors.

The below are essential measures that could hold significance for any future classification experiments.

1. In order to avoid the model from overfitting on specific data points and to facilitate its learning of generalized patterns, it is crucial to remove any identification attributes.

	2. Thoroughly assessing and managing missing values is essential to prevent introducing biases into the model. 3. Choosing the appropriate feature scaling method among MinMaxScaler, MaxAbsScaler, or StandardScaler should depend on the presence of natural data ordering and considerations like outlier preservation. 4. In classification tasks, particularly when dealing with imbalanced data, it's crucial to maintain the same target variable's class frequencies in each set accurately.
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 Feature engineering was not performed and kindly refer to 4] Verifying and Removing Identifiers from above section 2.a for the rationale behind eliminating certain 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 target variable is binary, containing 0 signifying players who are not selected, and 1 representing drafted players. With our goal to predict outcomes within these two classes, a Binary Classification model is appropriate, aligning with the objective of categorizing players into these distinct groups. In this experiment, utilised the Logistic Regression Classifier algorithm, which is well-suited for binary classification, to accurately determine whether the player will be selected for the NBA league. Hyperparameters Selected. Hyperparameters Name / Values Tested penalty: elasticnet l1_ratio: 0.5 / 1 (l1) class_weight: balanced / None solver: saga / lbfgs C: 4.281332398719396 * Began with the Logistic Regression algorithm's default hyperparameters and evaluated its performance. The above hyperparameters values are used to enhance the model performance. * Employed the hyperparameters class_weight='balanced' and NONE to modify class weights according to their representation in the training set, aiming to attain optimal model performance.

	* The hyperparameter C represents the regularization strength or the inverse of regularization strength. The value of C controls the trade-off between fitting the training data well (lower value of C) and keeping the model's coefficients small to avoid overfitting (higher value of C). * The choice of solver='saga' and 'lbfgs' hyperparameter determines the algorithm utilized for model optimization, particularly suitable for extensive datasets or those with numerous features. * Used class_weight='balanced' and solver='saga' to help address issues related to class imbalance and model optimization, respectively, which can ultimately lead to better performance of the model. Due to time constraints, could not to explore additional hyperparameters like fit_intercept, intercept_scaling for the Logistic Regression Classifier algorithm.

---

3. 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 Performance: 0.5 * In the 1st experiment, the logistic regression algorithm with default hyperparameters had the AUROC performance score of 0.96015 for the training dataset and 0.96011 for the validation dataset, indicating that the model is good at identifying players to be drafted in the NBA league. * Applying regularization techniques, the models from the 2nd and 3rd experiments has slightly dropped to the same AUROC performance score of 0.9583 for the training dataset and 0.9579 for the validation dataset, demonstrating that the model may slightly miss some potential players who are likely to be drafted in the NBA league in comparison to the 1st experiment's model. * With Hyper-tuned parameters in the 4th experiment, the model performance has considerably increased to the AUROC sore of 0.9881 illustrating that the model has generalized well enough to accurately detect potential players that will be drafted on the unseen data.
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 AUROC score of 0.98 on training and validation and 0.97 on the testing dataset informs that the model predicts players who will be drafted almost consistently across all the datasets. This provides teams with valuable insights for selecting promising talents, potentially improving their performance and team success. However, the model could be further fine-tuned to avoid missed opportunities.

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. Upon the installation of pyenv, an extra Python version was added, due to which encountered difficulty in unifying the version for all learning subjects. However, after going through the first lecture material on this matter, the issue was resolved. It was unclear to decide which features to choose as predictors. So, I tried to analyse the metadata, actual data and further tried to familiarise myself with domain knowledge related to the NBA league and the basketball game before utilizing heatmap correlation.

---

4. 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 logistic regression models, one with default hyperparameters showed slightly better performance than the others with regularization techniques, indicating that regularization did not sufficiently generalize and show relatively lesser AUROC performance. So, it would be worth exploring other binary classification algorithms that perform better than the current 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. As part of the rest learning process, I would investigate other binary classification algorithms, by conducting experiments and evaluating performance for selection of the most appropriate model for accurately identifying potential players who will be drafted in the NBA league in the annual event.