EXPERIMENT REPORT

Student Name	Sahil Kotak
Project Name	36120 Advanced Machine Learning Application Spring 2023 - Assignment 1
Date	18-08-2023
Deliverables	• kotak_sahil-24707592-week1_logistic_regression_model_2.ipynb • Logistic Regression Model

---

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	This project aims to predict which college basketball players are likely to be drafted into the NBA based on their statistics from the current season of college basketball. This prediction can be used by NBA teams to make informed decisions during the draft process, by sports analysts to provide insights into the draft prospects, and by the players themselves to gauge their likelihood of being drafted. Accurate results will improve draft decisions, increase team performance, and more insightful draft analysis. Inaccurate results may lead to suboptimal draft decisions, decreased team performance, and misleading draft analysis. The stakes are high, as a team's success in the NBA can have significant financial implications, including higher ticket sales, increased merchandise sales, and more lucrative sponsorship deals. Therefore, developing a reliable and accurate predictive model for this purpose is crucial.
1.b. Hypothesis	We aim to predict college basketball players' draft likelihood into the NBA using a logistic regression model. We hypothesize that higher performance metrics (points, steals, blocks) and offensive ratings (ORtg) increase draft chances. Players with more games and a higher percentage of team minutes are expected to have better draft prospects. The conference will significantly influence draft likelihood. The logistic regression model's performance will be evaluated based on its calibration and generalization to new data. We will gain insights into the factors influencing draft likelihood and the model's performance in predicting this outcome by testing these hypotheses. These insights will inform draft decisions and identify promising players.

1.c. Experiment Objective	This experiment aims to develop a logistic regression model to predict the likelihood of college basketball players being drafted into the NBA based on their season statistics. We aim for an AUCROC score of at least 0.85, indicating predictive solid accuracy. Possible outcomes include: ● **Successful Model:** AUCROC score above 0.85, enabling informed draft decisions and insights for players and analysts. ● **Moderate Success:** AUCROC score between 0.75 and 0.85, providing valuable insights but requiring further refinement. ● **Unsuccessful Model:** AUCROC score below 0.75, necessitating re-evaluation of features, model, or approach.

---

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	The main steps taken to run this experiment in terms of data preparation were:
	
	● **Data Cleaning:** We had to handle many missing values in the dataset. The approach taken was two-fold:
	  ○ Columns with a high percentage of missing values greater than 50% are dropped. This is because imputing such many missing values could introduce bias.
	  ○ For columns with a lower percentage of missing values, the missing values were imputed. For numerical columns, the median is used for imputation, and the mode is used for categorical columns.
	● **Data Splitting:** The data was split into a training set and a validation set, with an 80-20 split. This was done to evaluate the model's performance on unseen data.

2.b. Feature Engineering	For feature engineering, we encoded categorical variables and normalized numerical features. Categorical variables were one-hot encoded, and numerical features were left as they were since logistic regression is not sensitive to the scale of input features.

2.c. Modelling	The model trained for this experiment was Logistic Regression model. This was chosen due to its simplicity and efficiency, and because it is a good baseline model for binary classification problems, like this where we are predicting if a player will be drafted or not. The hyperparameters tuned for this model were: - Max_iter: This is the maximum number of iterations for the solver to converge. The value was set to 1000 to ensure that the model has enough iterations to converge. The model was used without any regularization, as the focus of this experiment was on feature engineering and preprocessing rather than model tuning.

---

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	From a technical perspective, the model achieved a high Area Under the Reciever Operating Characteristic (AUROC) score of 0.991 on the validation set. This score indicates that the model has a high ability to distinguish between players who will be drafted and those who won't. However, it's important to note that this high score might be due to overfitting, especially considering no regularization was used in the logistic regression model.
3.b. Business Impact	From a business perspective, the model's high AUROC score suggests that it could be a useful tool for predicting which players will be drafted. However, the potential overfitting issue mentioned above could limit its usefulness in practice. If the model is overfitting to the training data, it might not perform well on new data, leading to inaccurate predictions. Inaccurate results may lead to suboptimal draft decisions, decreased team performance, and misleading draft analysis. The stakes are high, as the success of a team in the NBA can have significant financial implications, including higher ticket sales, increased merchandise sales, and more lucrative sponsorship deals. Therefore, developing a reliable and accurate predictive model for this purpose is crucial.
3.c. Encountered Issues	During the experiment, several issues were encountered: - **Non-numeric values in numeric columns:** The 'num' column, which was assumed to be numeric, had non-numeric values. This could be due to poor data quality. - **Missing Values:** A large number of missing values were present in the dataset. Columns with more than 50% missing data were dropped, and the remaining missing values were imputed with the median. This approach might introduce bias if the missingness is not completely random. In future experiments, it might be worth exploring other methods of handling missing values, such as using machine learning-based imputation methods.

	- **Potential Overfitting:** The model achieved an AUROC score of 0.991 on the validation set, which might indicate overfitting. In future experiments, techniques like cross-validation could be used to assess the model's ability to generalize, and regularization could be applied to prevent overfitting.

---

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	From a technical perspective, this experiment demonstrated the importance of data preprocessing, including handling missing values, feature encoding and handling non-numeric values. The AUROC score achieved by the model suggests that the model was able to learn useful patterns from the data. However, the potential overfitting issues suggest that more work needs to be done. From a business perspective, the experiment showed that it's possible to build a model that can predict with high accuracy whether a player will be drafted or not.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, potential next steps could be" - **Regularization:** Applying regularization to the logistic regression model to prevent overfitting. This could involve tuning the regularization strength and trying different types of regularization (L1, L2 or ElasticNet). - **Feature Selection:** Applying feature selection techniques to identify the most important features and potentially improve the model's performance. This could involve using techniques like mutual information, chi-squared test, or recursive feature elimination. - **Model selection:** We could try different models such as Random Forest or XGBoost and compare their performance to the logistic regression model.