EXPERIMENT REPORT

Student Name	Kritika Dhawale
Project Name	AT1 - Kaggle Competition Week 2
Date	25th August 2023
Deliverables	dhawale_kritika-24587661-week2_logistic_reg.ipynb `*log_reg_week2.joblib*` `*scaler.joblib*` *+ preprocessed data files*
Github Link	adv_mla_at1

---

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	- The goal is to provide NBA teams, sports commentators, and fans with a reliable predictive model that evaluates the likelihood of college basketball players being drafted into the NBA based on their current season's performance statistics. - The model aims to provide valuable insights into player potential by leveraging machine learning, assisting teams in making informed draught decisions and fostering a data-driven approach to talent selection.
1.b. Hypothesis	- Does the current season performance metrics of college basketball players have a discernible impact on their likelihood of being drafted into the NBA? - It is expected that by analysing these statistics, a predictive model that effectively distinguishes between players who are drafted and those who are not will be developed. - The hypothesis assumes that the model's predictive power will outperform random chance, as demonstrated by an AUROC score greater than 0.5. - Furthermore, the hypothesis predicts that certain key metrics will emerge as important factors in the draught selection process.

1.c. Experiment Objective	- A selective list of factors from given 63 features in the dataset that have a constructive relationship with the target "**drafted"**. - Create and test a logistic regression model that fits the selected data accurately with a meaningful level of predictive accuracy, as measured by the AUROC score, so that we can predict the players that can be drafted in the NBA on certain given factors. - Possible scenarios: 1. The model achieves an AUROC score (>0.50) and differentiates (could be/could not be exceptionally predictive) between drafted and undrafted players. Key player performance metrics that influence draught selection are identified, providing teams and analysts with actionable insights. 2. The model's AUROC score remains near 0.5, indicating difficulties in accurately predicting expected outcomes. If the model is not accurate enough, it cannot be relied on for decision-making for business aspects. 3. There might be several other factors that affect the probability of being drafted, so we can't state a fact based on this given dataset. 4. Several **ethical concerns** can be raised such as bias and discrimination that can have an impact on certain groups of individuals or society as a whole. The model may perpetuate biases if certain groups of players are systematically favored or disadvantaged.

---

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 ***Includes feature engineering**	**Data Cleaning:** - Removed 32162 duplicates from the data (repeated player IDs) - Removed high missing values column and unnecessary features. - Filled Nan values with 0. - Replaced missing values in `ast_tov` column with median after removing outliers. - Repeated the same steps for test set. **Feature Engineering:** - `ht` column was in wrong format (date). Ex: 05-Jun meant 6'5. Month represents foot and date inches. Converted date format to numeric handling all exceptional cases through a function `convert_height_numeric`. Calculated height in inches and saved in a new column `ht_inch`. - Dropped the original height column `ht` **Data Preprocessing:** - Separated numerical and categorical columns to perform standard scaling on numerical features. There was only one categorical column left i.e `type`. - After visualizing type, there was only one value "all" in the column. So dropped the `type` column, no need to one hot encode. - Repeated the same steps for test set. **Data Split:** - Split the train set into train and val with a ratio of 30% (used stratify to create

	unbiased split and shuffled before split) - Shape check
2.b. Modelling	1. Logistic Regression (Compare with a baseline algorithm) **Hyperparameters:** As the data was large, had to maximize the number of iterations (default was 100). Set it to 2000. - AUROC on validation set. **Models that I didn't use:** 1. Random Forest Classifier (Model that can handle non-linear relationships) a. Tried normal processed features X_train to fit the model - AUROC on validation set. b. Transformed X_train into polynomial features of degree 2 and fit the model. - AUROC on val set. 2. Gaussian Process Classifier A kernel that can model complex non-linear relationships. Training time - 282.8275 seconds was high - AUROC on val set. **Reason to not use:** score not as good as logistic regression **Future models that can be used:** Analyze feature importance through random forest feature importance method to understand which statistics contribute most significantly to predicting draft selection. Furthermore, we might consider employing more sophisticated models like Multinomal or Polynomial Logistic Regression as there are a lot of features involved.

---

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	Logistic regression model - AUROC score on the val set - **0.9907** Reason: There are very less samples of players being drafted into the NBA. Visualizations show that only 0.9% of players are drafted out of total train set samples. This results in an imbalance dataset giving us a straight line in ROC curve. Implementing oversampling techniques might help. **Note:** There's a slight improvement from last experiment which can be a result of feature engineering and handling missing data properly.
3.b. Business Impact	This could be presented to our stakeholders as a developing model. For example: Proving a hypothesis that players with a high number of points and from sophomore year have a high chance of being drafted into the NBA could result in enhanced decision-making for teams. It can also be helpful in assisting business professionals in stimulating discussions and engaging debates about draft prospects and potential team selections in fans. To have a powerful impact on them, we aim to achieve an AUROC score > 0.99 on the test set on kaggle.

3.c. Encountered Issues	1. Was not able to remove specific indices while dealing with outliers in the "ast_tov" column. Solved: reset the dataframe index after dropping duplicates 2. While converting height from date to numeric formate, there were attribute errors and some difficult/exceptional cases to be solved. 3. Nan values after converting height. Solved: I was returning “None” if the height value is null/ the date/month is empty. Changed it to **0.**

---

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	- Data Processing: We learned how to handle missing values and outliers. Scaling numerical values and imputing missing values are essential for data quality and model performance. Feature engineering improved the model performance. Model Selection: Logistic regression was our baseline model. It performed well, also proved that complex models are not needed as there seems to be linear relationships in the data. Model Evaluation: The AUROC metric clearly measured the model's predictive power. - There are still chances for more research and experimentation, though. For instance, we should balance the data with respect to the target variable with SMOTE or other oversampling techniques. We could also look into additional variables for ex: teamwork, personality, and coachability that influence decisions about a player being drafted. Overall, our current strategy has delivered insightful information, but there is still room for investigation and experimentation.
4.b. Suggestions / Recommendations	1. Data balancing was something that was lacking in this experiment and surely has some improvement scope. 2. Feature engineering can be investigated to produce more pertinent features that will assist the model to perform better by better capturing the variation in the data. Deployment Scope: Advanced modelling, feature engineering, and data augmentation should improve predictive accuracy if this model needs to be deployed. Model interpretability, class imbalance, privacy, and ethics are also crucial. 1. Ensuring data quality: This can be achieved by doing data cleansing when required and running routine data quality checks. 2. Optimising the model: Refining and optimizing the model created originally.. 3. Deploying the pipeline: A user-friendly interface, continuous monitoring, and retraining are essential for team members. With the model complementing their expertise, analysts, scouts, and decision-makers must collaborate. 4. After the pipeline has been installed, it is essential to constantly check on its operation and take care of any problems as soon as they appear. To match dynamic player performance and team strategies, the model should be deployed iteratively with user feedback and continuous improvement.

Appendix

1. McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

2. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

3. Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, … Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

4. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

5. Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

6. Van Rossum, G. (2020). The Python Library Reference, release 3.8.2. Python Software Foundation.

7. Joblib Development Team (2023) Joblib/joblib: Computing with python functions., GitHub. Available at: https://github.com/joblib/joblib.