EXPERIMENT REPORT-1

Student Name	Ronik Karki
Project Name	Assignment 1 Part A
Date	20/08/2023
Deliverables	Notebook Name: karki_ronik-24886412-week1_kaggle_regression_model.ipynb

---

1. EXPERIMENT BACKGROUND	
1.a. Business Objective	The main goal of this project is to predict if a college basketball player will be drafted to join the NBA league based on his statistics for the current season. Since there are only two possible outputs, the result will be used to classify if the player will be drafted or not. As the data is highly imbalanced (i.e. class-0 has 99.04% more target values compared to class-1), using accuracy scores would be useless as the model which predicts all output as class-0 would have an accuracy of ~99.04%. Therefore the good metrics for this classification would be a high precision score, recall score, roc score, and f1 score for class-1 as we are more interested in finding how better the model performs in classifying and predicting the positive class. If the model misclassifies then the business will face two types of loss. 1. False Positive Error: In this, the model predicts that a member will be drafted but in reality, the member won't be selected. This error would increase the expectation of the viewers and hurt them as the player won't be drafted. 2. False Negative Error: In this error, the model predicts that a player will not be drafted. However, this would be more surprising to the spectators as the player who are not predicted to be selected will be drafted. Therefore, a good model should be able to minimize both errors in creating a reliable model.
1.b. Hypothesis	The experiment is the initial prediction model, therefore the hypothesis of this experiment is to create a model with only numerical features by replacing all the missing values with 0.

1.c. Experiment Objective	I am expecting the outcome of this model to have some strong predictions as numerical features have a decent correlation with the target variables. However, the missing data imputation and categorical feature addition can make the model more robust and precise for future experiments.

---

2. EXPERIMENT DETAILS	
2.a. Data Preparation	For the data preparation part, all of the features that were categorical were removed and only numerical features were selected. Among those numerical features, all the missing values were imputed by 0.
2.b. Feature Engineering	For this experiment, no feature engineering was applied. Only missing values were imputed with 0 values.
2.c. Modelling	For this experiment, again a simple Logistic Regression model was used. Here, I wanted to focus on how the simple model would perform without any special techniques applied. For the robustness of the model, stratified cross-validation was used to split the train and validation data for 10 splits. Stratified K-fold cross-validation was used as the dataset was highly imbalanced.

---

3. EXPERIMENT RESULTS

3.a. Technical Performance

The following outputs were obtained on a 10-fold cross-validation technique.

	Scores	Fluctuations
Precision	0.6524	0.0706
Recall	0.475	0.058
F1	0.5461	0.039
AUC	0.99489	0.0009

From the above results, we can see that only using numerical features is also good enough to perfectly distinguish between the positive class from the negative class focusing on the ROC-AUC score. With insignificant fluctuations, we can see that the

	model is robust enough to predict the players who will be drafted.
3.b. Business Impact	The model has decent precision and recall score but there is room for improvement. The model has the ability to correctly predict members who are being drafted the 65% of the time. It has learned from the features but it has only 47% ability to identify players who will be drafted. Using this model might have an ability to distinguish between two classes but increasing precision and recall would be recommended for business use. Using all features with feature engineering and feature selection with different models might improve the model performance
3.c. Encountered Issues	Using this model improves the scores of the model as it was able to provide some useful information to the model. However, the model still doesn't have high precision and recall scores. There are many unclean data mixed with different columns(i.e. Categorical features). Therefore they were removed for this experiment.

---

4. FUTURE EXPERIMENT	
4.a. Key Learning	Adding numerical features only also produces decent results. Imputing 0 might also work for some predictions as we are able to store other important information for that particular feature.
4.b. Suggestions / Recommendations	Although the model has high scores for AUC (0.99625), the model score for the competition is still in 8th place as of now. Therefore, adding more categorical features and selecting the best features might increase the model performance and might help in achieving the best score in the Kaggle competition. So, the next experiment should be included more features and feature engineering techniques with advanced cleaning in the unclean data.