EXPERIMENT REPORT - 2

Student Name	Ronik Karki
Project Name	Assignment 1 week 2
Date	25/08/2023
Deliverables	Notebook Name: karki_ronik-24886412-week2_gradient_boosting.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 whether 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 expectations 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 players 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 based on an attempt to improve the scores obtained from the previous experiment. The hypothesis for this experiment is that the model will obtain better scores if we add more features (i.e. categorical features) and impute the missing continuous variable values with mean rather than replacing them with 0. Furthermore, boosting model techniques might have better scores than simple logistic regression models.

1.c. Experiment Objective	The expectation from this experiment is that there will be many more features once we include categorical features. Therefore, different types of feature selection techniques will be used for selecting the best feature and the scores might be improved and more robust predictions might be expected at the end.

2. EXPERIMENT DETAILS

2.a. Data Preparation	For the data preparation part, the categorical features were very unclean. The following cleaning was applied on both the test and train data set. 1. Heights had been converted to date initially. Therefore, it was mapped and converted to centimeters resulting in a numerical feature. 2. The number of players had noise, which was removed and replaced with 1 as number 1 had the highest usage. 3. Values in years had many unclear values which were removed and replaced with 'Jr' as it had the highest mode which is used for imputing categorical values. 4. The conference name seems to have an upper and lower case difference. Converting all to uppercase before encoding as a feature. 5. Feature name ‘type’ was dropped as it had the same values for all points.
2.b. Feature Engineering	For this experiment, two types of feature engineering techniques were used: 1. Data imputation: a. All numerical features that had missing values were replaced with their mean values. b. For the test data set all numerical features that had missing values were replaced with mean of train data features to prevent data leakage. c. For categorical features like number and year, the missing values were replaced with the highest frequency or mode of the feature. 2. Label encoding: a. All of the categorical features were converted to integers by applying label encoding techniques. 3. Feature transformation: a. Height was transformed from the categorical feature to the numerical feature by converting it to centimeters from a date format.

2.c. Modelling

Feature selection techniques were used for selecting the highly correlated and significant features. Two types of feature selection techniques were used.

1. Pearson correlation which had greater than or equal to 0.1 and features that had a correlation score less than or equal to -0.1
2. Chi2 test for the feature that had both input and output as a categorical feature. If the p-value was less than or equal to 0.05, then the feature was selected.

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. It was tested in the logistic regression model first and then the gradient boosting model was used to observe the difference between models and their performances.

3. EXPERIMENT RESULTS

3.a. Technical Performance

Two of the models (i.e. logistic regression model and gradient boosting model) were run for this experiment.

1. Logistic Regression Model:

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

	Scores	**Fluctuations**
Precision	0.63227	0.12311
Recall	0.29105	0.06659
F1	0.39573	0.07840
AUC	0.98502	0.00414

2. Gradient Boosting Model:

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

	Scores	**Fluctuations**
Precision	0.75114	0.08837
Recall	0.67882	0.06174
F1	0.71039	0.05641
AUC	0.99771	0.0006

From the above results, we can clearly see that the Gradient boosting model outperformed the logistic regression model. It may be due to the way gradient boosting

	has a better ability to absorb non-linear relationship data compared to the logistic regression model. It might also be the outliers and missing data handling, which can be investigated in the future experiment. Furthermore, Using mean values imputation instead of 0, the score decreased for the logistic regression model. Meanwhile, imputing mean improved the efficiency of the gradient boosting model without any hyperparameter tuning.
3.b. Business Impact	The gradient boosting model has robust prediction performance. Compared to the previous model, this model is much better in correctly identifying the players who are to be drafted in the squad and also has higher preciseness to predict the players who will be selected. With a precision score of 75% and recall of 67%, the model is decent enough for the business to get a highly accurate model for player prediction. The fluctuations around the 10 folds are less and this model can be deployed. However, it can be improved for even more rigid performance.
3.c. Encountered Issues	There were some issues while converting the team feature in the data set. The train data didn’t have information or unseen values presented in the test dataset which raised an issue while converting to the labels using a label encoder. Using train data for encoding, it didn’t have sufficient information about new teams and couldn’t encode. It can be focused on new experiments to overcome the issue.

4. FUTURE EXPERIMENT	
4.a. Key Learning	Adding categorical features and imputing numerical features with mean didn’t turn out to increase the model performance for the logistic regression model. It might be due to the complexity of the features and its outliers. However, the features were super useful for the gradient boosting model and it was far better than the logistic regression model.
4.b. Suggestions / Recommendations	The Gradient boosting model has a much-improved performance than the logistic regression model and also has a higher ROC score which is required for the kaggle competition. Although, this model is good for deployment, it is still in the 7th rank in Kaggle competition. Therefore, there is still room for improvement in this model. Hyperparameter tuning and testing on different models might perform better as the feature sets are getting more complex. Furthermore, One hot encoding feature and feature selection might increase more features and reduce complexity at the same time for improvement in the scores.