EXPERIMENT REPORT

Student Name : Yasaman Mohammadi
Project Name : advmla-2023-spring

Date:
18/08/2023
Deliverables:
notebook name: Mohammadi_Yasaman-24612626-week1_PolynominalLogisticRegression

1.EXPERIMENT BACKGROUND

1.a. Business Objective

This project aims to create a model that predicts whether a college basketball player will be drafted into the NBA based on their current season's statistics. As a result, NBA teams can make better draft selections and allocate resources more efficiently.

By making accurate predictions, draft decisions can be improved, players can be developed, and teams can perform better. Missed opportunities, wasted resources, and potential damage to the model's credibility can all be attributed to incorrect predictions. Draft choices, player development strategies, and fan engagement are all impacted by the model's success.

1.b. Hypothesis

Do college basketball player statistics provide a sufficient basis for predicting their likelihood of being drafted into the NBA?

> Accurate predictions can lead to improved draft selections, resulting in a better roster for the team and an overall improvement in performance.
>
> Data-driven predictions can enhance sports analysis, media coverage, and fan discussions. As a result, it can generate excitement and stimulate informed discussion about the potential outcomes of a draft.

1.c. Experiment Objective

The experiment is expected to result in a predictive model with a high AUROC score that accurately predicts whether a college basketball player will be drafted into the NBA based on their current season's statistics. AUROC scores should be significantly higher than random chance, indicating that the model can reliably distinguish between players who are likely to be drafted and those who are not.

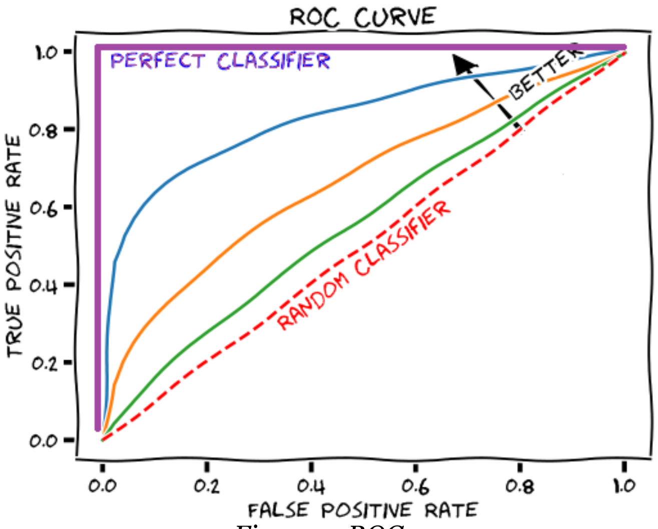

Figure 1- ROC curve

Different scenarios might happen based on AUROC curve performance. The area under the receiver operating characteristic (AUROC) is a performance metric for evaluating classification models. The larger the area under this curve, the better the performance.

2.EXPERIMENT DETAILS

2.a. Data Preparation

	the steps taken for preparing the data	Explanation
1	Data collection	Building a machine learning model begins with the collection of data from a variety of sources. In this particular case, the data was collected from a Kaggle competition.
2	Explore Dataset	A general understanding is derived from the data in this step.
3	Data cleaning	Data should be preprocessed to remove any irrelevant, missing, or corrupted data points and any discrepancies in the data.
4	Data visualization	Visual understanding of data distribution is gained by plotting different charts in both Python and Tableau.
5	Feature engineering	In feature engineering, new features are created, or existing features are transformed to improve the model's performance. This experiment used a label encoder and one hot encoding to transform categorical data into numerical data to consider their influence on the machine learning model.
6	Removing outliers	An outlier is a data point significantly different from the rest. Outliers can adversely affect machine learning models and they need to be removed. (identifying them by using box plots)

6	Splitting the dataset	The dataset was randomly split into training, validation, and testing. The training set is used to train the model, the validation set is used to tune the model's hyperparameters and prevent overfitting, and the test set is used to evaluate the final performance of the model. The test size parameter specifies the percentage of data allocated to the test set, in this case, 20%. The random state parameter is used to ensure the reproducibility of the split. For many datasets, the 80/20 split is a good rule of thumb since it provides enough data to train the model while still leaving a considerable amount for testing. The ratio may vary depending on the size and complexity of the dataset, as well as the specific problem being addressed. As part of this project, there was a test set without drafted sections and the objective was to predict a probability for the drafted variables at the end.
7	Scaling the dataset	The data is crucial to prevent sensitivity to some features so all features have the same scale. Since some machine learning algorithms are sensitive to the scale of input features, this is important..

Table 1- the steps taken for preparation of the data

Additional feature engineering can be conducted in future experiments.

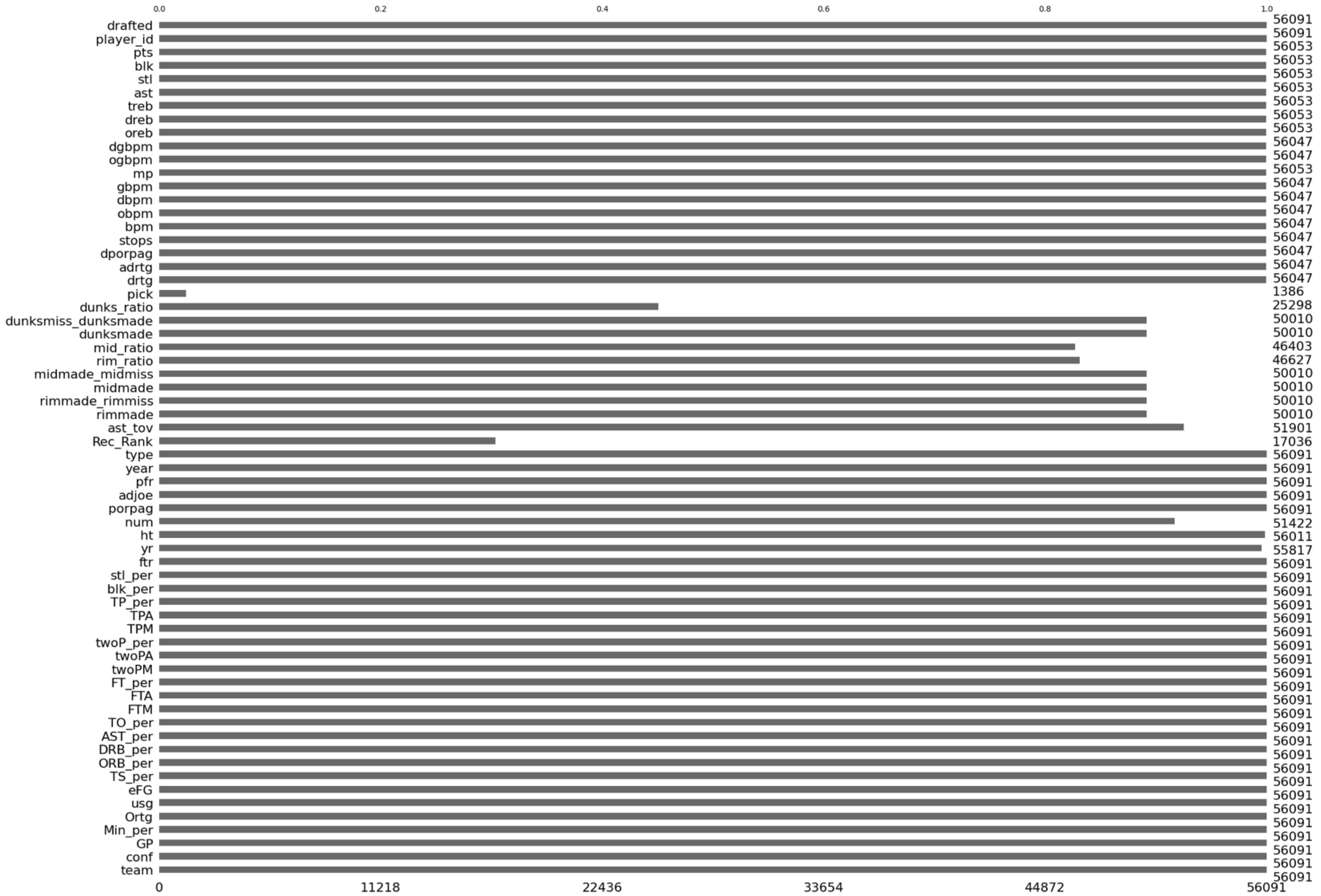

Figure2-The mnso bar chart for fiding missing values

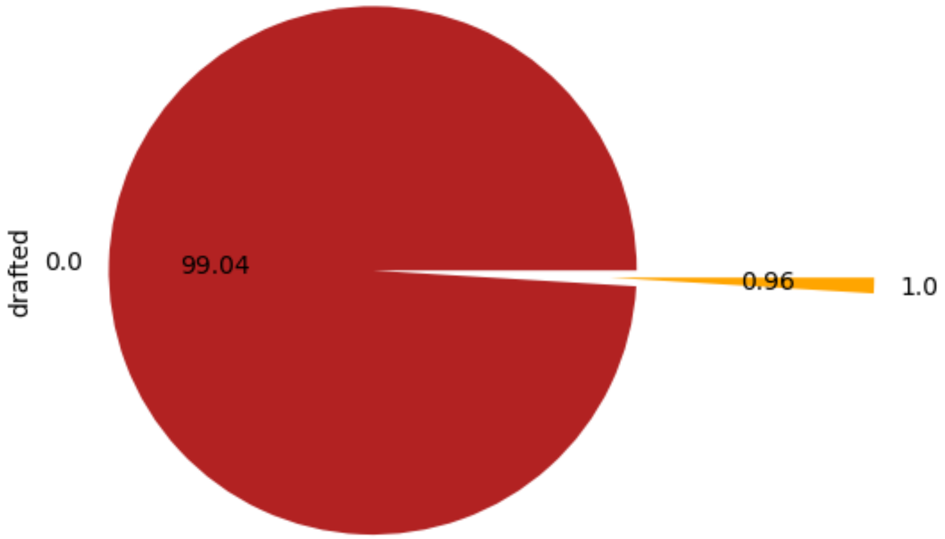

Figure 3-Pie chart of drafted distribution

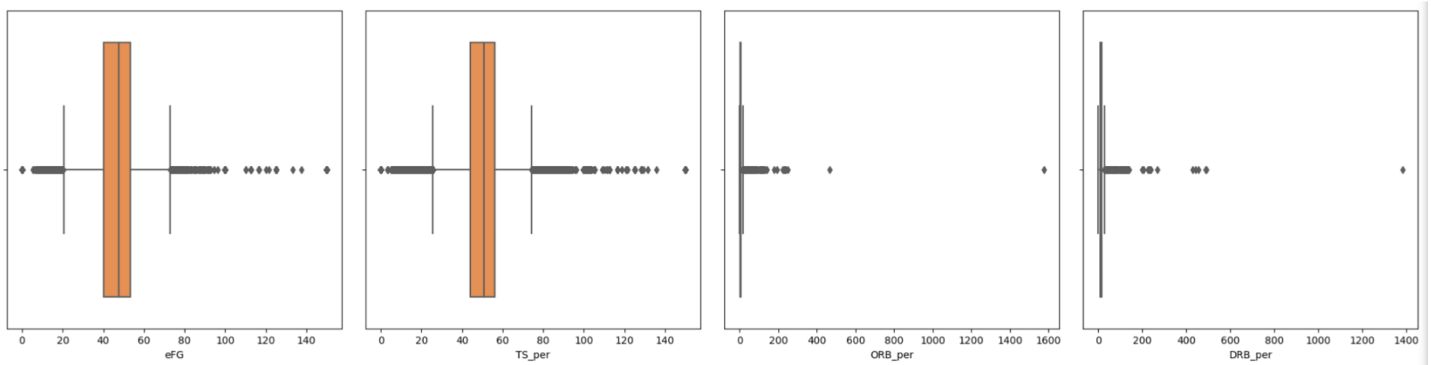

Figure 4 – Random box plots

2.b. Feature Engineering

In this experiment, label encoding and one hot encoding were used to convert the two categorical features to numerical features. This is since the polynomial logistic regression cannot use categorical features.

2.c. Modelling

This experiment used Polynomial Logistic Regression To model the relationship between the independent variables and the probability of a binary outcome using polynomial functions,
This extension of Logistic Regression allows fitting a nonlinear decision boundary by introducing polynomial terms.

	Hyperparameter	**Explanation**
1	C	This is the inverse of regularization strength; it must be a positive float. The smaller the value, the stronger the regularization. Values that used in this case was (100, 10, 1.0, 0.1, 0.01)
2	Penalty	Regularization (penalty) can sometimes be helpful. L1 and L2 was used is this case.
3	Solver	Not all the solvers support all the penalties in this case “liblinear” was used to support specified penalties.

Table 2-Hyperparameters tuning explanation

It was decided to use stratified k-fold cross-validation due to the imbalanced distribution of the target feature in the primary data. Therefore, training and test data in each fold will reflect the imbalanced distribution of the target feature in the primary data.

3. EXPERIMENT RESULTS

3.a. Technical Performance

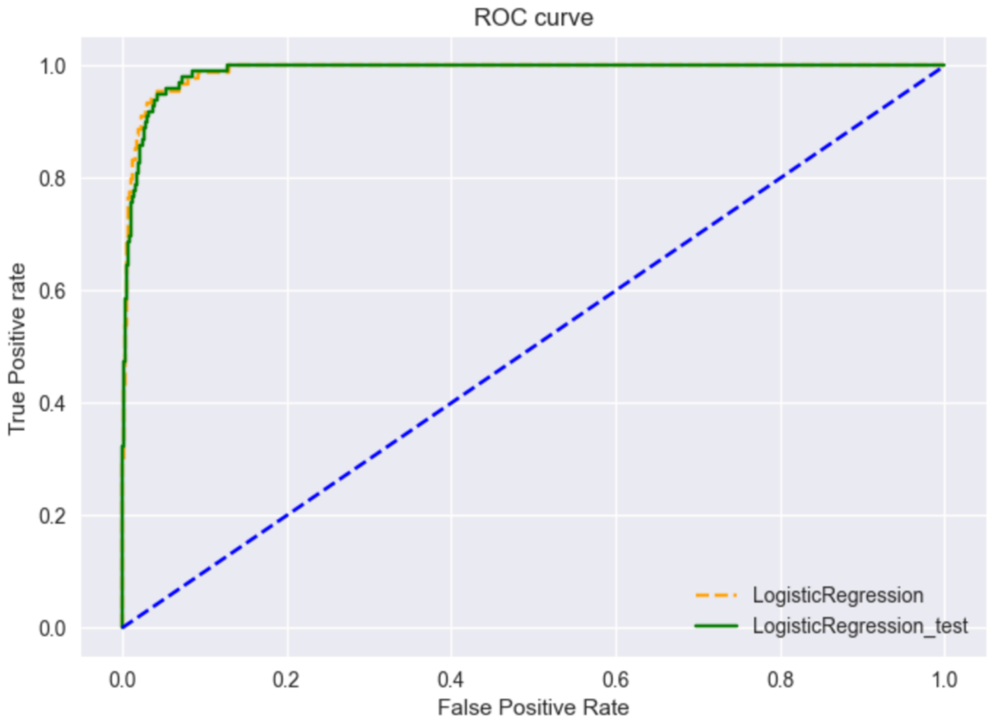

Figure 5-AUROC curve of Polynomial Logistic Regression model

The higher the AUC, the better the model will be at distinguishing between positive and negative classes. This ROC curve indicates good performance, but there is room for improvement.

It may be possible to achieve better results using other models or feature engineering.
The final score achieved on Test set by uploading the final result on Kaggle was 0.95891.

3.b. Business Impact

The success of the model affects several aspects of the NBA teams' operations, including draft choices, player development, resource allocation, fan engagement, and team performance. Errors in results can have a wide range of negative impacts, including wasted resources, missed opportunities, credibility damage, and long-term competitive disadvantages. For the model to provide meaningful value to NBA teams and stakeholders, accurate predictions are critical.

3.c. Encountered Issues

In the height column, there has been a problem with unique values. A few dates appear in the column. This value can be critically important for the basketball player election. However, as its values do not make sense, we have discarded them, but further investigation will be conducted in future experiments to determine whether or not these values are meaningful.

4.FUTURE EXPERIMENT

4.a. Key Learning

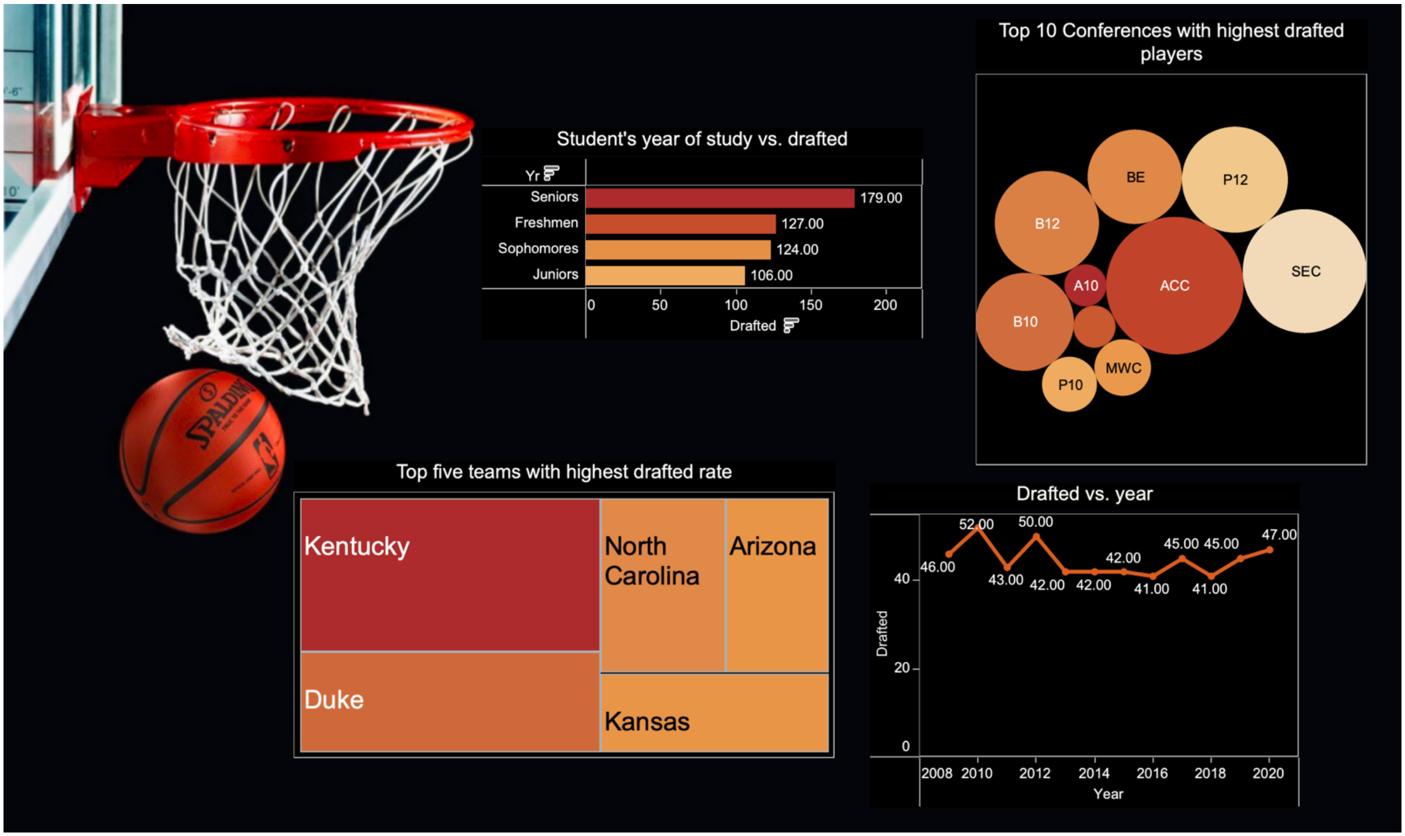

Figure 6- Tableau Dashboard

An interactive dashboard in Tableau has been created to visualize the correlation between drafted players' rates and teams, seasons, and conferences.

Further, we achieved a high score in this experiment using a simple machine-learning model. In conclusion, it is worth considering other models and feature engineering to achieve higher scores.

4.b. Suggestions / Recommendations

In future experiments, different machine learning models with additional feature engineering and hyperparameter tuning will be used.
In order to deploy the final solution into production, the user interface must be refined so that predictions are easy to access, ethical biases must be addressed proactively, and ongoing model adjustments should be monitored in real-time.

References:

1. **GlassBox Medicine.** (2019, February 23). *Measuring Performance: AUC & AUROC.* Retrieved from https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/#:~:text=The%20area%20under%20the%20receiver,use%20to%20evaluate%20classification%20models.