Assignment 1

Kaggle Competition

Ronik Karki
Student ID: 24886412

08/09/2023

36120 - Advanced Machine Learning Application
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Executive Summary

The Kaggle competition was created for the first assignment of the advanced machine learning course. This competition used the dataset of the NBA draft and the main objective of this project is to predict if the player will get selected for the upcoming season in the NBA league. It is one of the most followed sports events and many spectators would love to guess who would be selected.

The main objective of this competition is to achieve a higher Receiver Operating Characteristic (ROC) score. To predict the players who will get selected, a machine-learning model was created. This model had a very high ROC score of 0.99922 for the Linear Discriminant Analysis model. It shows that the model is good enough to predict the players and can be used for deployment.

This report follows the CRISP-DM methodology. A **CR**oss **I**ndustry **S**tandard **P**rocess for Data Mining (CRISP-DM) is a data science process in which the model follows six sequential steps: Business understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Hotz, 2022).

2. Business Understanding

a. Business Use Cases

This project is useful for universities that invest in sports activities. They can look after the chances of their student being selected for the NBA. In this way, they can set up the values for the players who will get selected for the NBA and get high profits from those players by signing a contract. Apart from this, fans and commentators can also benefit from this project to guess who will be selected.

There are many statistical data of a player and it is really difficult for one to predict the players who will be selected. There are many important factors that are required for a player to be selected and in order to solve this complexity, machine learning algorithms are required. Machine learning models are robust and have enough power to figure out the pattern and accurately predict the players.

b. Key Objectives

The main objective of this project is to correctly predict if the player gets selected or not. Firstly, for businesses like university sports teams, if they get a model that would predict players that would be selected for the NBA, then these players can be given a contract that would make the NBA teams pay the amount for buying these players and it would help the sports team get profits. To achieve this profit, a machine learning model should have a good ability to distinguish between the players who get selected and players who don’t and not miss players who will be selected. This means the model should have a higher ROC score and recall score.

In the context of the spectators and commentators, they would like to have correct predictions so that they can start supporting their players. It would turn out to be sad if the players are predicted to be selected but in reality, they won’t. Therefore a good model would have the ability to correctly identify as well as predict those players who would get selected. This means the model should have a higher precision score and recall score.

The expected output of this model is to predict if the player is selected or not selected so it is a classification problem and would require classification machine learning algorithms for predictions.

3. Data Understanding

The dataset was downloaded from assignment 1 of the Kaggle competition for Advanced MLAA coursework. This dataset consists of information about the players for the selection of NBA. The dataset is split into two parts. One for the training and another for the testing part. There are 56,091 rows for the training data and 4970 rows for the test data. As we are using this data for the prediction of players who are selected, we can see from the graph below (as shown in fig 1) that only a few players are selected for the NBA. This means that the dataset is highly imbalanced for this project.

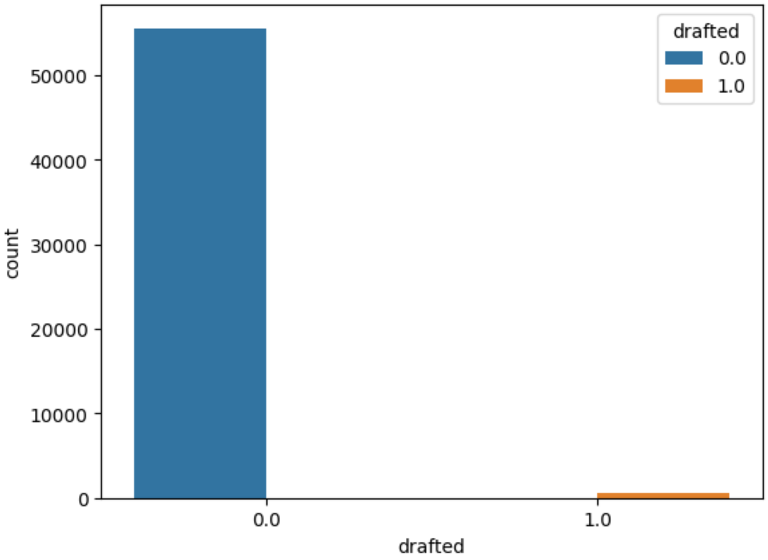

Fig 1: Imbalance Target Variable counts

Out of 56k players only 536 actually make it into the team. Since the dataset is highly imbalanced (as shown in Fig 1), metrics like accuracy won't be reliable. A naive model that predicts all the targets as class 0 would still be likely to achieve ~99% accuracy. Therefore, metrics like precision, recall, and ROC score are used for this project.

Furthermore, there are many features available in the dataset. There are 63 columns including the player id and the target variable which is drafted in the dataset. However, there are highly missing values that have very noisy values in the dataset. For example, the 'pick' column has 54,705 values missing out of 56,091 values. Therefore, these types of missing values were imputed by applying feature engineering techniques. Furthermore, the height column had a date as an input value and had mixed data types. They were also cleaned using different mapping strategy and then it was converted to centimeters before passing to a model.

4. Data Preparation

For the preparation of this model, many columns were cleaned and preprocessed before applying for modeling.

For experiment 1, all of the missing values were replaced with 0 to create a baseline model. Furthermore, only features that had numerical data type columns were selected.

Moving on to experiment 2, the height column was converted to centimeters. Initially, it had a date format, it was mapped to feet and inches. Then it was converted to centimeters. The jersey number of players had unclear values in the data and they were removed and imputed with the highest frequency which is “number 1”. 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. Conference names had upper and lower case differences. Converting all to uppercase before encoding as a feature. All of the missing values were then replaced by the mean values of the respective columns of train data. The “Team” column was dropped as it had some new values on the test data which was unseen and it was removed.

In experiments 2, 3, and 4, Both ‘one hot encoding’ and ‘Label encoding’ were used for the feature engineering of categorical data. Furthermore, in the fourth experiment, feature engineering was applied, for creating a new feature from combined features. For example, features that were combined like ‘dunksmiss_dunksmade’, ‘rimmade_rimmiss’, and ‘midmade_midmiss’ were used for gaining the values of a miss for the respective fields and they were transformed into the new feature. Highly missing variable like ‘pick’ was converted to categorical feature (1 if there is a value and 0 if the value is missing) as imputing any values would make the model biased according to the value imputed. All the remaining features were then replaced with the mean values of train data. Multicollinearity was detected and these columns were removed from the data as it creates more complexity to the model.

5. Modeling

There are many machine learning models used for this project and different models were used for different experiments. The objective of this model is to solve classification problems. So all experiments used classification models. For all of the experiments and models, the training dataset was split into test and train using a cross-validation model. It is used for calculating the performance of the model. A stratified K-fold cross-validation model was used as the dataset was highly imbalanced. Each of the models was run 10 times within these splits and the scores are the combination of all 10 folds.

a. Approach 1

For the first experiment, A simple logistic regression model was used and it didn’t have any parameters. It was used for creating a baseline model. For this model, only numerical features were selected and fit into the model for prediction.

b. Approach 2

For the second experiment, label encoding was used and feature selection was done before using the model. Two techniques were used for selecting features (Brownlee, 2019):

1. Using the chi2 test for features that have input as categorical and output as categorical, and features that had a p-value less than 0.05 were selected.
2. Using Pearson correlation and selecting features that are greater than 0.1 and less than -0.1

Gradient Boosting model was used for this model as the features were more complicated than the previous one and this model had a better ROC score than the simple logistic regression model. None of the parameters were included in the model.

c. Approach 3

For the third experiment, one hot encoding was used and models like Gradient boosting and XGBOOST models were used. No other parameters were added to the model

d. Approach 4

For the last experiment, many different models were used in expectation of improvement in the model scores. Features that had multicollinearity were dropped and only remaining features were passed in the model. The following models were used for this experiment:

1. Discriminant Analysis models: Linear discriminant Analysis model and Quadratic Discriminant analysis model

2. Linear model: A weighted logistic regression model was used for handling the imbalance dataset.

3. Ensemble model: XGB, LightGBM, Gradient boosting model, and Ada boost model

4. Naive Bayes model: Gaussian naive Bayes model

XGBoost model was further tuned using the hyperparameter tuning method. After tuning the model, the following hyperparameters were selected: 'eval_metric': 'auc', 'max_depth': 5, 'min_child_weight': 1, 'gamma': 0.3, 'subsample': 1.0, 'colsample_bytree': 0.0, 'alpha': 1, 'lambda': 1.1, 'seed': 54

6. Evaluation

a. Evaluation Metrics

The evaluation metrics used were ROC score, Recall score, and Precision score. Firstly, the ROC score was used for the Kaggle competition. This score helps us identify the models' distinguishing ability between the two classes. Recall metrics are used for identifying the players who will be selected and precision is used for correctly predicting the players who were identified correctly. For the spectators and commentators Precision and Recall are important whereas for the competition and for businesses like universities ROC score is important.

b. Results and Analysis

The following table consists of the best scores achieved during each of the different experiments.

Splits	**Training**		**Testing**	
Model	**Scores**	**Fluctuations**	**Score**	**Fluctuations**
Precision	0.65626	0.0090	0.65084	0.08516
Recall	0.48693	0.01763	0.47211	0.03850
F1	0.55894	0.01370	0.54267	0.02244
AUC	0.99536	0.0001	0.99505	0.00087

Table 1: LR model

Splits	**Training**		**Testing**	
Model	**Scores**	**Fluctuations**	**Score**	**Fluctuations**
Precision	1.0	0.0	0.75694	0.07813
Recall	1.0	0.0	0.70887	0.05751
F1	1.0	0.0	0.72966	0.04963
AUC	1.0	0.0	0.99799	0.00045

Table 2: Gradient boosting model

Splits	Training		Testing	
Model	**Scores**	**Fluctuations**	**Score**	**Fluctuations**
Precision	0.38674	0.00316	0.38832	0.030004
Recall	1.0	0.0	1.0	0
F1	0.55776	0.00329	0.55881	0.03053
AUC	0.99726	0.00006	0.99723	0.00075

Table 3 : LDA model

Splits	Training		Testing	
Model	**Scores**	**Fluctuations**	**Score**	**Fluctuations**
Precision	1	0	0.75934	0.037269
Recall	1	0	0.71086	0.03743
F1	1	0	0.73326	0.019890
AUC	1	0	0.99804	0.00019

Table 4: XGBOOST model

From all of the scores we can see that, XGBOOST has the highest score of 0.99804 ROC score and has higher precision and recall score than any other model. But looking at the training scores, the xgb model seems to overfit. The LDA model seems to be the most stable model with high high-performing ROC score and is perfectly fitting with a model that has very less fluctuations across different folds. This is the reason why this model was rigid enough to predict highly in the in-kaggle competition. Logistic regression seems to be the weakest model as it might not have the ability to deal with the higher complexity features and the gradient boosting model performs well with slightly overfit.

c. Business Impact and Benefits

From the two of the best models, we can see that either one can be used according to the business conditions. For businesses like university sports clubs and competitions, LDA model can be used as it has a very high recall score and high ROC score. This means that the model is capable enough to perfectly distinguish between the players who will be selected or not and perfectly identify them. This model perfectly captures the players who will be selected. However, it also predicts players who won't be selected. Which leads to disappointment for the spectators and commentators. Therefore, they can use the XGBOOST model as it has a higher ability to correctly classify players as well as predict them correctly 76% of the time. This model is overfit but it performs consistently across the different folds and proves its reliability.

d. Data Privacy and Ethical Concerns

The dataset consists of player information like height, club name, and school year which might be sensitive, but the information of the player is being masked as a player_id which ensures that this step prevents the data privacy of a player.

7. Deployment

- Explain the process of deploying the trained model.
- Discuss any integration steps or considerations for real-world implementation.
- Address any challenges or considerations related to deployment.

Instructions: Explain the process of deploying the trained model, including any integration steps or considerations for real-world implementation. Discuss any challenges or considerations related to the deployment process and provide recommendations or suggestions for future deployment efforts.

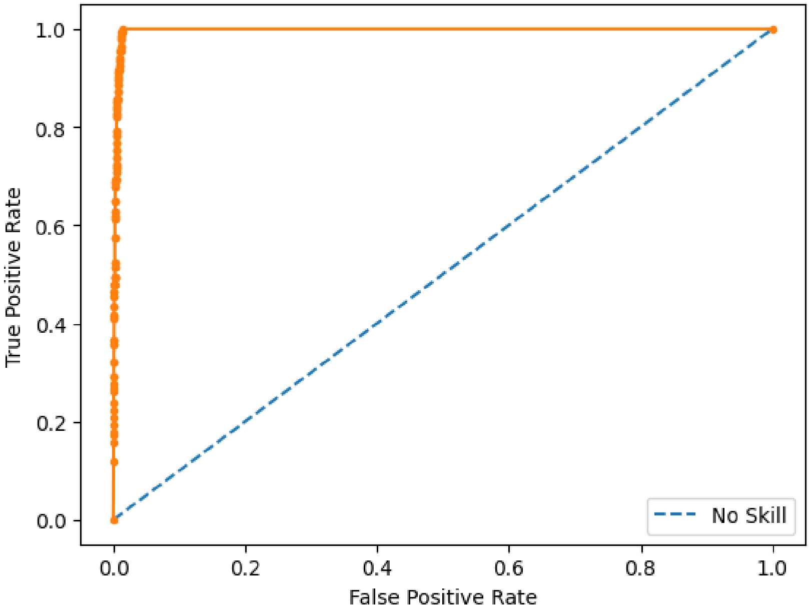

Fig 2: AUROC curve

From the above figure (see Fig 2), we can see from the Area Under the ROC curve that the model has the perfect ability to distinguish between the players who will be selected and those not selected. This model can be deployed for predicting the players who will be selected for the NBA league. However, there are some challenges in the model, If any new values are seen in the model then we might face an issue in predicting. For example, the categorical values are encoded with the known values. If any unknown values are present in the future then the model deployment will be failed and throws an error. Therefore, handling these issues can be a bigger challenge in the future. Furthermore, numerical columns might consist of many unclean values which also creates a problem. This should also be addressed in future projects. The model needs to be updated continuously. For example, we can monitor the performance of this model after a month and retrain the model again so that it achieves the standard.

8. References

Brownlee, J. (2019, November 26). How to Choose a Feature Selection Method For Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Hotz, N. (2022, November 13). CRISP-DM. Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/