A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal

2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED) | 978-1-6654-5389-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCED56140.2022.10010351

1st Noviyanti T M Sagala
Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
noviyanti.sagala@binus.edu

2nd Muhammad Amien Ibrahim
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia, 11480
muhammad.ibrahim1@binus.edu

***Abstract*— Predicting whether an athlete is likely to win a medal in the Olympic games is new. The studies on Olympic Games are mostly trying to predict the total medals of a nation possible to achieve or a country's performance by applying statistics approaches. Some works even expand the data utilized for medal predicting by including more years and predictor factors such as country host as well as increasing the level of data granularity. Machine learning, in particular boosting algorithms, has had a massive influence in improving the accuracy of prediction models. To accurately classify an athlete, three different machine learning approaches can be utilized. In this study, three separate boosting algorithms, namely Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost), and Category Boosting (CatBoost) are evaluated using Olympic historic dataset, first with default parameters, then with hyperparameters by applying Grid Search algorithm. Four different types of performance evaluation metrics were computed with 5-fold Cross-Validation (CV) approach. The best results were obtained with the XGBoost approach on hyperparameters, achieving an accuracy of above 90%, a precision of 96.8%, and a recall of 83.2%.**

Keywords*— *CatBoost, Grid Search Cross-Validation, LightGBM, Olympic Medal Prediction, XGBoost

I. INTRODUCTION

The Olympics are recognized as one of the most important events, providing a legitimate and shared platform for players from around the world to demonstrate their ability and talents. Predicting Olympic medal achievement has a long history in predicting research, since the 1970s [1]. One of many techniques that are applied on prediction research field is machine learning.

In recent times, machine learning (ML) has had a huge influence on many aspects of modern society. The promise of machine learning in discovering hidden patterns and improving predicting accuracy has not yet been fully realized in Olympic sport context [2]. Boosting, also known as "meta-algorithms," is a sequential or chronological process in which each consecutive model attempts to fix or fix prior model errors. The algorithms used for reducing bias and variance among the encountered variables and turn them into strong classifying variables [3].

Since the first use of Tobit model proposed by Ball in the early 2000s [4], the accuracy of such Olympic predictions has progressively increased over time. A few authors interested in predicting a country's Olympic performance have started using new predicting approaches while the other researchers in the area have studied more varied and large data sets to gradually increased the predictive models [5-6]. Predicting the medal outcomes using Ordinary Least Squares Regressions (OLS) as it produced results that were simple to understand [7]. Due to the incorporated exponential function penalized small, expected medal numbers, some authors adopted Poisson-model based to address this methodological issue [8-10].

Somewhat similarly, the authors have greatly expanded the data sets utilized for medal predicting in three ways. First, by including more years [11]; second, by increasing the granularity degree beyond country-specifics [11], different sports [6][12], and even on the level of individual athlete [13]; and third, by exploring at more predictor factors such as country host. It is found that that hosting the Games boosts the predicted number of medals and lasts until the subsequent Games [14].

This paper takes a different perspective and attempts to predict whether an athlete is likely to win a medal or not. This study makes several contributions to the literature by incorporating three ways mentioned above and machine learning techniques.

1) It implements three prominent boosting algorithms namely Light Gradient Boosting Machine (LightGBM), eXtreme Gradient Boosting (XGBoost) and Category Boosting (CatBoost) approaches
2) The GridSearch is applied on all algorithms stated in point 1 with the goal to improve the performance of them by using hyperparameters, then evaluated with 5-fold Cross Validation.
3) The granularity degree of data set is increased beyond continent-specifics, sport popularity, country host, and more years.

II. BOOSTING ALGORITHMS

Although there are many boosting algorithms beside XGBoost, LightGBM, and CatBoost, these three algorithms are applied in this study due to two main reasons: First, provide open-source software libraries for Python programming; Second, they are fast to train (in sklearn) and deliver precise results. To achieve the best results using XGBoost, LightGBM, and CatBoost, there are several

parameters must be configured in such a way that they are appropriate for the dataset being utilized. Some of these parameters include:

- Maximum depth of the tree (max depth). The deeper the tree, the more complicated the model obtained
- Optimal number of trees(n_estimators). The number of runs XGBoost will try to learn.
- Learning Rate denotes how fast the model learns. The learning rate must be set as low as possible.
- Minimum amount of data in one leaf (min_child_samples). To deal with over-fitting.
- Maximum number of leaves in one tree (num_leaves)
- L1 regularization(reg_alpha).
- The maximum number of trees that can be built. The lower the number the faster the training.
- The maximum depth to be used for the Decision Tree Algorithm (depth).

A. *XGBoost*

XGboost adopts gradient boosting framework. This algorithm is one of the most efficient ones to perform predictions and ranking when it comes to small to medium structured datasets because offers well-structured memory usage and uses simultaneous and distributed computation to guide abrupt learning [15]. The parameters need to be configured for maximizing the performance are max depth, n_estimators, and learning rate.

B. *LightGBM*

LightGBM uses XGBoost as a foundation. Therefore, it adopts not only gradient boosting framework but also decision tree algorithm to carry out classification, and other machine learning tasks. Among all boosting techniques, LightGBM has the greatest accuracy in classification problems [16]. The configured LightGBM parameters are like XGBoost.

C. *CatBoost*

This approach is notable for its data processing speed, which in some cases may be up to 60 times faster than lightGBM [17].

III. Data Exploration

This research paper will look into the enormous history of Olympic Games to see if it can predict whether or not an athlete will win a medal. The history of the Olympics is influenced by several variables. Before creating a predictive model, we need to undertake a thorough data examination to determine these factors. For this study, we used a previous Olympics history dataset [18]. The dataset keeps records of Olympics history from the earliest competition in 1896 to recent games in 2016. There are 206,165 rows and 15 columns in the dataset. Each row in the dataset represents the record of each athlete that participated in a particular branch of sport and indicate whether they won a medal or not.

An examination was conducted to look at the distribution of float data in the dataset such as athletes' age, height, and weight. Fig. 1 shows that the average age of the Olympic Athletes is 25 years and 7 months old with a median value of 24 years, making the distribution right skewed. Fig. 1 also shows the distribution of the height of the athletes which is around 175.37 centimeters with a slightly similar median value at 175 centimeters, which rather resembles a normal distribution. Similarly, the distribution of athletes' weight has the average value at 70.69 kilograms while the median is 70 kilograms.

Despite the histograms are able to provide insights regarding the physicality aspects of the athletes, these three float variables are the only numerical variables with missing values. There are about 3.5%, 22%, and 23% missing values for each of age, height, and weight variables respectively. In this study, the missing values are then imputed with the mean value taken from each of these three float variables. Furthermore, the medal variable also contains missing values. It is important to note that the medal variable only keep records of athletes having won a medal. As a result, an empty record will be given to those athletes that didn't win the competition. In this case, the missing values in the medal variable is imputed with a label of "DNW" which stands for "Did Not Win".

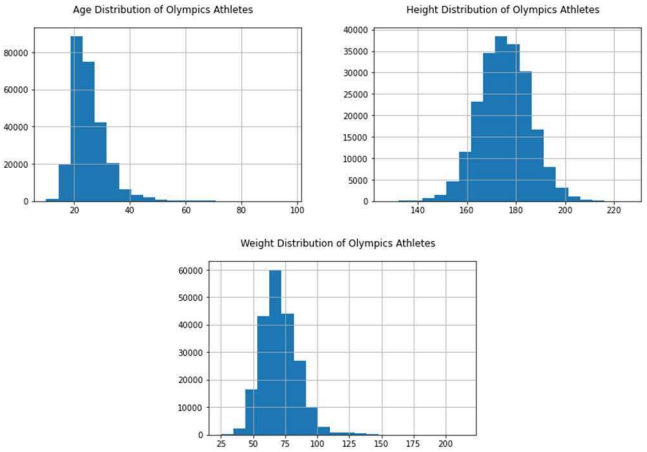

Fig.1 Age distribution of Olympics athletes.

Fig. 1 shows that the right tail is longer than left, while Height and Weight show that it is close to a normal distribution.

There are several variables that indicate location such as team, NOC, and city. The team variable represents the country or organization that they represent. Previously, there are certain cases where people that live in minor islands also participated in the Olympics, thus they represent an organization rather than a country. Currently, most of these minor islands have become part of countries known nowadays. However, as a categorical variable, the team variable is considered to contain a high variance as there are more than 600 different teams. Therefore, the team variable is grouped into continents although the continent data itself is not provided in the dataset. In this study, an attempt to group up the team variable into continents was done by manually collecting country name which resembles the team variable as well as collecting their respective continent. As a result,

TABLE II. RESULTS OF THE VARIOUS ALGORITHMS FOR ALL PERFORMANCE METRICS

Model	Parameter	Accuracy	Precision	Recall	ROC_AUC
CatBoost	Default	0,903	0,974	0,829	0,903
LightGBM		0,890	0,956	0,818	0,890
XGB		0,802	0,822	0,772	0,802
CatBoost	GridSearch CV 5 Folds	0,891	0,960	0,816	0,891
LightGBM		0,901	0,965	0,832	0,901
XGB		**0,902**	**0,968**	**0,832**	**0,902**
CatBoost	GridSearch CV 10 Folds	0,890	0,958	0,817	0,890
LightGBM		0,898	0,960	0,832	0,899
XGB		0,900	0,966	0,829	0,900

Table II shows the comparison of accuracy, precision, recall, and ROC-AUC between models with default parameters and tuned parameters.

the continent variable is added to the dataset which contain six different categories. Both North America and South America are merged into a single category called “Americas” due to their similar geographical positions. On the other hand, Antarctica and any team variable that are not categorized are included in a single category called “unknown”.

In terms of sport, there are 56 different sports at the Olympics. This categorical variable is considered to be having too many variations. Therefore, these sports are grouped into five categories based on their popularity [19][20].

Table I describes each group with its corresponding sports. Category A includes the type of sports that are the most popular among other sports at the Olympics while Category E is the least popular sports, which could be due to the recently added Olympic sports. In addition to this, we have removed a set of records containing some sports that are discontinued such as Tug-Of-War, Cross Country Skiing, Art Competitions, Lacrosse, Polo, Cricket, Racquets, Motorboating, Military Ski Patrol, Croquet, Jeu De Paume, Roque, Alpinism, Basque Pelota, and Aeronautics. These sports have been excluded from the model since they do not represent current data and so do not contribute to the prediction model.

TABLE I. SPORT CATEGORIES BASED ON POPULARITY

Category	Sport
A	athletics, aquatics, gymnastics
B	cycling, tennis, basketball, football, volleyball
C	archery, badminton, boxing, judo, rowing, shooting, table tennis, weightlifting
D	canoe/kayaking, equestrian, fencing, field hockey, handball, sailing, taekwondo, triathlon, wrestling
E	modern pentathlon, golf, rugby

Table I shows the list of sports according to their popularity. Category A indicates the most popular sports while Category E specifies the least popular sports.

The city variable represents the city where the Olympics was held geographically. However, there are as many as 42 cities in the dataset. In this study, the cities are grouped into more general categories. A closer look found that the Olympics were held three, two, or one time in each of the cities. Therefore, the city variable is categorized into “threetimes”, “twotimes”, and “onetime”.

In this study, several variables have been selected for modelling. The independent variables are age, height, weight, sex, season, city, sport, and continent. It is important to note that city, sport, and continent are derived variables where continent are obtained from manually gathered external data. Other variables such as event is not included since it represents a deeper classification of sport. In addition, variables that represent location such as team and noc are already represented by a more general variable such as continent. All the independent variables are then one-hot-encoded, increasing the size of the dataset horizontally. In contrast, the medal variable is selected as the dependent variable. There are two attributes in the medal variable such as “1” which represents the athlete winning any medal and “0” which represents that the athlete did not win any medal.

IV. MODELING & EVALUATION

Variables such as age, height, and weight, have different scales. To provide the same influence, these three variables are scaled using standardization technique. This is done by using StandardScaler from Scikit-learn. StandardScaler removes the mean and scale to unit variance to standardize their characteristics. As a result, the mean value is 0 and the standard deviation is 1.

The next stage in the process of data preparation for modelling is splitting the data into training and testing set. A train_test_split function from Scikit-learn is used to perform this step. As many as 80% of the data is assigned as training set while 20% is used as testing set.

A deeper examination revealed that the records with the target variable of "0" outnumbered the other attributes of "1". This seems to be logical since there are more athletes that did not win a medal than athletes that win a medal. In the dataset, the records show that there is only about 15% of athletes that won any medal. In order to prevent model misclassification, SMOTETomek is used to balance the data. SMOTETomek is a data balancing function from library that combines oversampling and undersampling technique using SMOTE.

Several traditional machine learning algorithms are utilized in the modelling stage. These machine learning algorithms were selected from a variety of libraries for boosting methods. The parameters used are those provided by the library as defaults. To evaluate and compare one model to another, accuracy, precision, recall, and ROC-AUC are utilized. Another set of the same experiment is conducted by training the same machine learning algorithms but with different parameters. This is done to see if the evaluation parameters would result higher performance. Table 2 shows the comparison between accuracy, precision, recall, and ROC-AUC for models with default parameters and models with tuned parameters. Overall, it can be seen that the models with tuned parameters achieved higher performance, although the increases are marginal. Despite having lower computational cost, models trained with 5-fold cross validation achieved slightly higher score for the evaluation metrics compared to models trained with 10-fold cross validation.

Furthermore, the tuned parameters used for the models are specified in Table III. The tuning process is conducted through GridSearchCV step using Scikit-learn library.

TABLE III. LIST OF HYPERPARAMETERS FOR EACH MODEL

Model	Parameter
CatBoost	CatBoostClassifier(depth=10, iterations=100, learning_rate=0.1, random_state = 42, verbose = False)
LightGBM	LGBMClassifier(learning_rate=0.2, max_depth=25, min_child_samples=5, num_leaves=200, reg_alpha=0.01, random_state = 42)
XGB	GradientBoostingClassifier(max_depth=9, n_estimators=200, random_state = 42)

Table III shows the tuned parameters obtained from performing GridSerachCV.

V. CONCLUSION

This study attempts to perform analysis and build a model that predicts whether an Olympic athlete is likely to win a medal. The dataset used comprised of records collected from 1896 to recent events in 2016. A combination of athletes' physical data, sports type, and location are used as independent variables while a binary variable containing information medal achievement is used as the dependent variable. The result shows that XGBoost performs best compared to other machine learning algorithms with an accuracy value of above 90%. There are certain aspects that could be improved such as modelling with more complex models which requires more computation and time. These are left for future works.

REFERENCES

[1] Ball DW. Olympic Games Competition: Structural Correlates of National Success. International Journal of Comparative Sociology. 1972; 13:186–200.

[2] Schlembach, Christoph & Schmidt, Sascha & Schreyer, Dominik & Wunderlich, Linus. (2020). Forecasting the Olympic Medal Distribution during a Pandemic: A Socio-Economic Machine Learning Model. SSRN Electronic Journal.

[3] N. T. M. Sagala and S. D. Permai, "Enhanced Churn Prediction Model with Boosted Trees Algorithms in The Banking Sector," 2021 International Conference on Data Science and Its Applications (ICoDSA), 2021, pp. 240-245, doi: 10.1109/ICoDSA53588.2021.9617503.

[4] Bernard AB, Busse MR. Who wins the Olympic Games: Economic resources and medal totals. Review of economics and statistics. 2004; 86:413–7. 3.

[5] Forrest D, Sanz I, Tena JD. Forecasting national team medal totals at the Summer Olympic Games. International Journal of Forecasting. 2010; 26:576–88. doi: 10.1016/j.ijforecast.2009.12.007.

[6] Vagenas G, Palaiothodorou D. Climatic origin is unrelated to national Olympic success and specialization: an analysis of six successive games (1996–2016) using 12 dissimilar sports categories. Sport in Society. 2019; 22:1961–74.

[7] Kuper, Gerard H. and Sterken, Elmer, Olympic Participation and Performance Since 1896 (January 2001).

[8] Lui, H.-K., Suen, W., 2008. Men, money, and medals: An econometric analysis of the Olympic Games. Pacific Economic Review 13, 1–16.

[9] Leeds Eva Marikova and Michael Leeds, (2012), Gold, Silver, and Bronze: Determining National Success in Men's and Women's Summer Olympic Events, Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik), 232, (3), 279-292.

[10] Blais-Morisset, P., Boucher, V., Fortin, B., 2017. The Impact of Public Investment in Sports on the Olympic Medals. Revue economique 68, 623–642

[11] Nicolas Scelles, Wladimir Andreff, Liliane Bonnal, Madeleine Andreff, Pascal Favard. Forecasting National Medal Totals at the Summer Olympic Games Reconsidered. Social Science Quarterly, Wiley, 2020, 101 (2), pp.697-711.

[12] Noland, M., Stahler, K., 2016a. Asian Participation and Performance at the Olympic Games. Asian Economic Policy Review 11, 70–90.

[13] Johnson, D.K.N., Ali, A., 2004. A tale of two seasons: participation and medal count at the Summer and Winter Olympic Games. Social Science Quarterly 85, 974–993.

[14] Vagenas, G., Vlachokyriakou, E., 2012. Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the "population-GDP" model revisited. Sport Management Review 15, 2, 211–217.

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

[16] Sun, Xiaolei, Mingxi Liu, and Zeqian Sima. "A novel cryptocurrency price trend forecasting model based on LightGBM." Finance Research Letters 32 (2020): 101084

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical Features.," 2018.

[18] Pradhan, Rahul & Agrawal, Kartik & Nag, Anubhav. (2021). Analyzing Evolution of the Olympics by Exploratory Data Analysis using R. IOP Conference Series: Materials Science and Engineering. 1099. 012058. 10.1088/1757-899X/1099/1/012058.

[19] "Athletics to share limelight as one of top Olympic sports". The Queensland Times. 31 May 2013. Retrieved 18 July 2013.

[20] "Winners Include Gymnastics, Swimming - and Wrestling - as IOC Announces New Funding Distribution Groupings". The Association of Summer Olympic International Federations. Retrieved 18 July 2013.