OLYMPIC MEDALS PREDICTIONS IN SUMMER GAMES USING DATA MINING METHODS

Dr. Vasilis Aggelis
PIRAEUS BANK SA
Syggrou Ave., 87, Athens, GREECE
AggelisV@piraeusbank.gr

ABSTRACT

After the closing ceremony of Athens 2004 Olympic Games, it is an evergreen topic the prediction of Olympic Medals with strong accuracy. Essential role in this scope has data mining. The introduction of data mining methods in the sport area can be considered of great assistance as to prediction and forecasting. In this paper, we present PREMOM, a predictive model for Olympic medals. The main factors in this model are the medals in all previous summer Olympic Games. It is demonstrated that the PREMOM (**Pre**dictive **M**odel for **O**lympic **M**edals) predictive model contributes to the more efficient forecasting for Olympic Medals.

KEYWORDS

Data Mining, Predictive Model, Data Preparation

1. INTRODUCTION

Data treatment is generally conducted using several data mining methods such as Linear Regression, Neural Networks and Decision Trees aiming at the development of patterns, rules, predictive models and finally forecasting. These methods produce interesting as well as useful results enhancing the more efficient knowledge management. However, not all kinds of results lead to rigid conclusions.

From this point of view the data miner and the judgment of the user are essential in evaluating the results and especially the predictive models efficiency. Therefore the co-operation between people expert in data mining and others with good knowledge of the data sets is important leading to proper evaluation of the predictive model.

In this paper the development and evaluation of the PREMOM predictive model is studied. For the present case study the software SPSS Base 12.0 was used. A general description of predictive models follows in section 2, while in section 3 the procedure of the data preparation is presented. Experimental results are discussed in section 4 while section 5 contains final conclusions and future work plans.

2. PREDICTIVE MODELS BASICS

A model is an abstract representation of a real-word process. A typical form of a model is Y=aX+b, where Y, X are variables and a, b are parameters. In a predictive model, one variable is expressed as a function of the others. This permits the value of the response variable to be predicted from given values of the others (the predictor variables). The response variable in general predictive models is often denoted by Y, and the p predictor variables by X1,…, Xp. The model will yield predictions, $\hat{y}$ = f(x1, …, xp; θ) where $\hat{y}$ is the prediction of the model and θ represents the parameters of the model structure. When Y is quantitative, this task of estimating a mapping from the p-dimensional X to Y is known as regression.

Prediction models in which the response variable is a linear function of the predictor variables, yields prediction:

$$\hat{Y} = a_0 + \sum_{j=1}^{p} a_j X_j$$

Where θ = {a0, …, ap}. We have used $\hat{Y}$ rather than simply Y on the left of the expression because it is a model, which has been constructed from the data. In other words, the values of $\hat{Y}$ are values predicted from the X, and not values actually observed.

An easy way to evaluate and compare predictive models in order to choose the best is the use statistical metrics such as R, R square and adjusted R square.

3. DATA PREPARATION

Scope of this study is the generation of a predictive model concerning the number of gold, silver and bronze Olympic medals in relation to previous medals in Olympic Summer Games. The significance of establishment of such a model is undoubtful as it helps in prediction for following Olympic events.

Response variable (Y) is the number of medals. Predictor variables (Xi) are medals in the very previous Games and country population.

Our dataset includes all countries and nations which have ever participated in Summer Olympic Games and all the medals since 1896. In order to have more precise results we don't take into account non existing participant nations which had won medals in previous Summer Games. These former "participants" are Australasia (1908-1912), United Team of Germany, West Germany, East Germany, Czechoslovakia, Soviet Union and Yugoslavia. Also mixed teams and their winning medals are not taken into consideration.

A sample of the above data set is shown in Table 1.

Table 1 – Sample of Data Set

NATION	GOLD	SILVER	BRONZE	GOLD in previous Games	SILVER in previous Games	BRONZE in previous Games	Year	Population
...	...	...	...	...	...	...	...	...
USA	35	39	29	40	24	33	2004	293027571
CHN	32	17	14	28	16	15	2004	1298847624
RUS	27	27	38	32	28	28	2004	143782338
AUS	17	16	16	16	25	17	2004	19913144
JPN	16	9	12	5	8	5	2004	127333002
ITA	10	11	11	13	8	13	2004	58057477
KOR	9	12	9	8	10	10	2004	48598175
GBR	9	9	12	11	10	7	2004	60270708
CUB	9	7	11	11	11	7	2004	11308764
ROM	8	5	6	11	6	8	2004	22355551
GRE	6	6	4	4	6	3	2004	10647529
...	...	...	...	...	...	...	...	...

In order to generate a prediction like, $\hat{Y} = a0 + \sum_{j=1}^{p} a_j X_j$, the stepwise linear regression method was used.

The Stepwise method of field selection builds the equation in steps, as the name implies. The initial model is the simplest model possible, with no input fields in the equation. At each step, input fields that have not yet been added to the model are evaluated, and if the best of those input fields adds significantly to the predictive power of the model, it is added. In addition, input fields that are currently in the model are reevaluated to determine if any of them can be removed without significantly detracting from the model. If so, they are removed. Then the process is repeated, and other fields are added or removed. When no more fields can be added to improve the model, and no more can be removed without detracting from the model, the final model is generated.

4. RESULTS

The stepwise method builds in four steps the following predictions of PREMOM model.

$\hat{GOLD}$ **= 0.481*PG + 0.183*PS + 0.257*PB - 0.051**

$\hat{SILVER}$ **= 0.211*PG + 0.181*PS + 0.452*PB + 0.347**

The same method builds in three steps the following prediction of PREMOM model.

$\hat{BRONZE}$ **= 0.141*PG + 0.227*PS + 0.509*PB + 0.838**

where PG is number of Gold Medals in previous Summer Games, PS is number of Silver Medals in previous one and PB is number of Bronze Medals in previous Olympic Games.

Predictor variable population is not figured in the models, but it joins in the model development and its accuracy

The degree to which two or more predictors (X variables) are related to the response (Y) variable is expressed in the correlation coefficient R, which is the square root of R-square. To interpret the direction of the relationship between variables, one should look at the signs (plus or minus) of the regression or parameters (θ). If a parameter is positive, then the relationship of this variable with the dependent variable is positive; otherwise in case the parameter is negative so is the relationship.

Table 2 – PREMOM model for Gold Medals

Model	**R**	**R square**	**Adjusted R square**
1	0.878	0.771	0.770
2	0.891	0.793	0.792
3	0.894	0.799	0.798
4	**0.896**	**0.803**	**0.802**

Table 3 – PREMOM model for Silver Medals

Model	**R**	**R square**	**Adjusted R square**
1	0.855	0.732	0.731
2	0.889	0.791	0.790
3	0.892	0.796	0.795
4	**0.895**	**0.800**	**0.799**

Table 4 – PREMOM model for Bronze Medals

Model	R	R square	Adjusted R square
1	0.836	0.699	0.699
2	0.859	0.738	0.737
3	**0.862**	**0.743**	**0.742**

As can be seen in Tables 2, 3 and 4 the value of R concerning the last step model is appropriate since it is close to 1.

R square is commonly used as measure of a model's goodness of fit. The range of R square value between 0.743 and 0.803 is considered satisfactory and indicates an acceptable model.

Adjusted R square measures the proportion of the variation in the response variable due to the predictor variables. Unlike R square, adjusted R square accounts for the degrees of freedom associated with the sums of the squares. Therefore, even though the residual sum of squares decreases or remains constant as new predictor variables are added, this is not the case for the residual variance. This is the reason, adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square.

If adjusted R square is significantly lower than R square, this normally means that some predictor variables are missing. The absence of these variables causes the improper measurement of the variation in the dependent variable.

The range of Adjusted R square value between 0.742 and 0.802 is almost the same with R square indicating therefore an acceptable model.

5. CONCLUSIONS AND FUTURE WORK

In this study, the development of the PREMOM predictive model concerning the number of Olympic Medals in relation to the previous winning medals is described while experimental results are also supplied. It is concluded that there exists a strong relation between the expected medals and the winning medals in the very previous Summer Games.

Another basic conclusion is that a data mining method such as predictive modeling, contributes in a better forecasting. The PREMOM model can be trained with new data and becomes the basis for further predictions.

Future plans employ the development of predictive models using other sources such as number of participants (men, women), number of events, economic nation parameters and others.

REFERENCES

Bernard A., and Busse M., 2004. Who Wins the Olympic Games: Economic Resources and Medal Totals. *Review of Economics and Statistics Journal*.

Draper N.R., and Smith H., 1998. *Applied Regression Analysis*. John Wiley & Sons, Inc.

Hand D., Mannila H., and Smyth P., 2001, *Principles of Data Mining*. The MIT Press.

Hong S., and Weiss S., 1999. Advances in Predictive Model Generation in Data Mining. *Proceedings of 1st International Workshop Machine Learning and Data Mining in Pattern Recognition.*

Hong S.J., and Weiss S., 2000. Advances in Predictive Model Generation for Data Mining. *Pattern Recognition Letters Journal.*

Johnson D., and Ali A., 2000. Coming to Play or Coming to Win: Participation and Success at the Olympic Games. *Wellesley College Working Paper.*