

A prediction model for which Country will win the highest number of “Gold” in 2016.

Antarlina Sen MS in Business Analytics, SAS® and OSU Data Mining Certificate, Oklahoma State University; Gaurang Margaj MS in Business Analytics, SAS® and OSU Data Mining Certificate, Oklahoma State University

ABSTRACT

Olympic Games was started to establish a healthy friendship between countries. Given its history and value, <http://www.olympic.org/>, www.iaaf.org and <http://www.rio2016.com> and other website contain records of 300+ events where 10,000+ athletes (men and women) from 204 countries compete. This data is available from 1896 to 2012. The goal of this paper is to predict which country will score the maximum number “Golds” in the Rio Olympics of 2016 using past trends of countries; performances and the statistics of players listed for 2016. Before building this prediction model, exploration and descriptive analytics will be carried out to answer questions such as following. Does the current form of a player impact his/her performance in the Olympics? Has a specific country performed better with respect to the others in the last two years in a particular sport? If a player wins one medal in one Olympics, what are the chances he/she will win in a “Gold” in the next one? Does a more experienced player have a better chance of winning? Insights generated from answering these questions will be used to build the predictive model.

INTRODUCTION

Olympics is one of the most talked about competitions in the whole world. More than often, people watch the Olympics for sentimental reasons. In the United States alone, the viewership for the Olympics is 39.8 million. There is always a sentiment analysis (big or small) associated with how people choose their favorite player in any sport. Most of which is based on past performances. This paper puts these past performances into good use! The main challenge in any Data Mining Project is data collection and cleaning. Once the data is collected and sorted, the model building becomes fairly simple. To have a reliable source for that purpose is always an added advantage. But the process can be tedious. We collected most of our historical data from Olympics.org. Depending on the average ranking of each sport based on its viewership in the United States, we picked the top ten most popular sports for our analysis.



DATA PREPARATION AND CLEANING

Olympics data has been collected from Olympics.org and rio2016.org. The Current table contains Year of the Event, The city where the game took place, the sport, the discipline, the event, the athlete, the country to which the athlete belongs, the medal he won. Since the variable for analysis is the medal count, three dummy categorical variables were made one for each type of medal.

For Team Sports, we have cleaned more than one medal count for a winning country that corresponded to each individual player. In order to count the medal for a winning country, we considered one medal per sport instead of one medal per player.

For specific player names, we cleaned occurrences of name ambiguities. Certain countries that played the Olympics before 2016, and are not playing in the 2016 Olympics were removed.

The Final Data Set contains Year, City, Event, Sport, Discipline, Athlete, Country, Medal, Gold, Silver, and Bronze.

The input Predictor Variables identified are Sport, Discipline, and Country. (All Categorical Variables).

The target Variable is Gold which is a dummy created from Medal. When the Country has won a gold medal, we have set it to one

	SAS Variable Format	DB2 Data Type
YEAR	DATE9.	Date
CITY	\$CHAR21.	Character
SPORT	\$CHAR23.	Character
DISCIPLINE	\$CHAR23.	Character
EVENT	\$CHAR23.	Character
ATHELETE	\$CHAR23.	Character
COUNTRY	\$CHAR23.	Character
GENDER	\$CHAR5.	Character
MEDAL	\$CHAR6.	Character
GOLD	BEST12.	Binary
SILVER	BEST12.	Binary
BRONZE	BEST12.	Binary

DESCRIPTIVE ANALYTICS

Over the last few Olympics, the United States and China have been the countries that have won the maximum number of medals collectively across all sports, events and disciplines. If we want to analyse the sports in which they have won the maximum number of medals, it gives us a fair idea of the sports that dominate the count of medals for these two countries.

These graphs shows the top two countries, that are United States and China that have frequently won over the last 10 Olympic games for 10 randomly picked Sports. It shows the number of medals won by United States and China in those sports

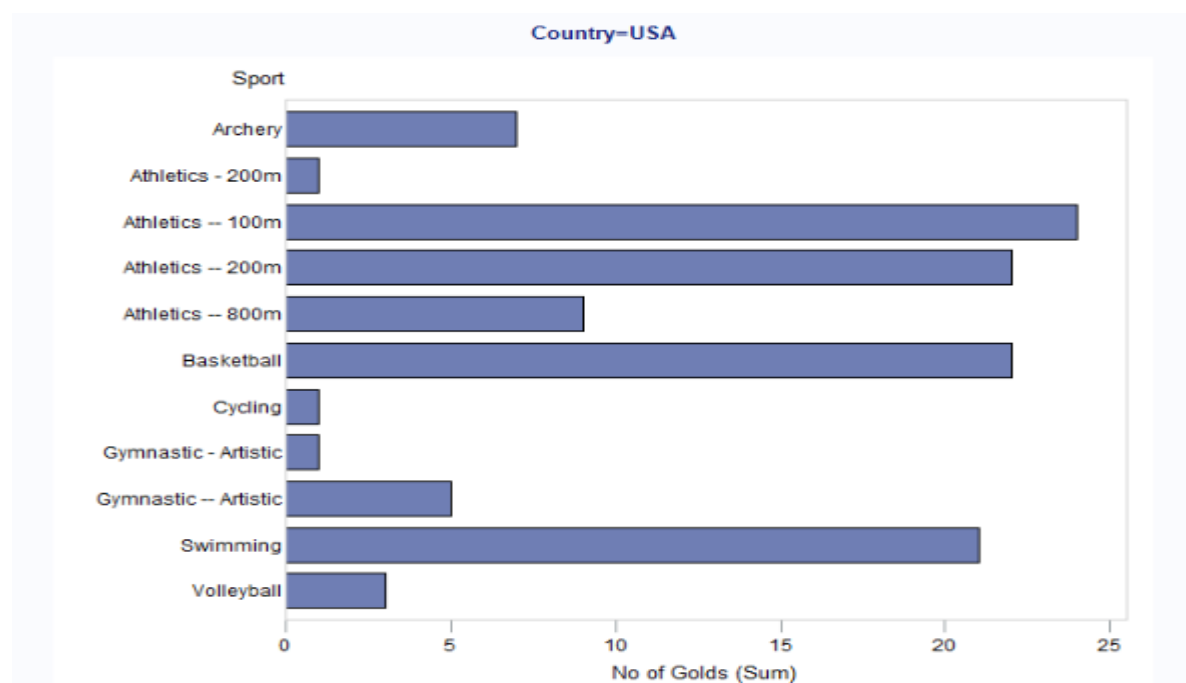


Fig: Average Count of Medals for USA for the last 10 Olympics.

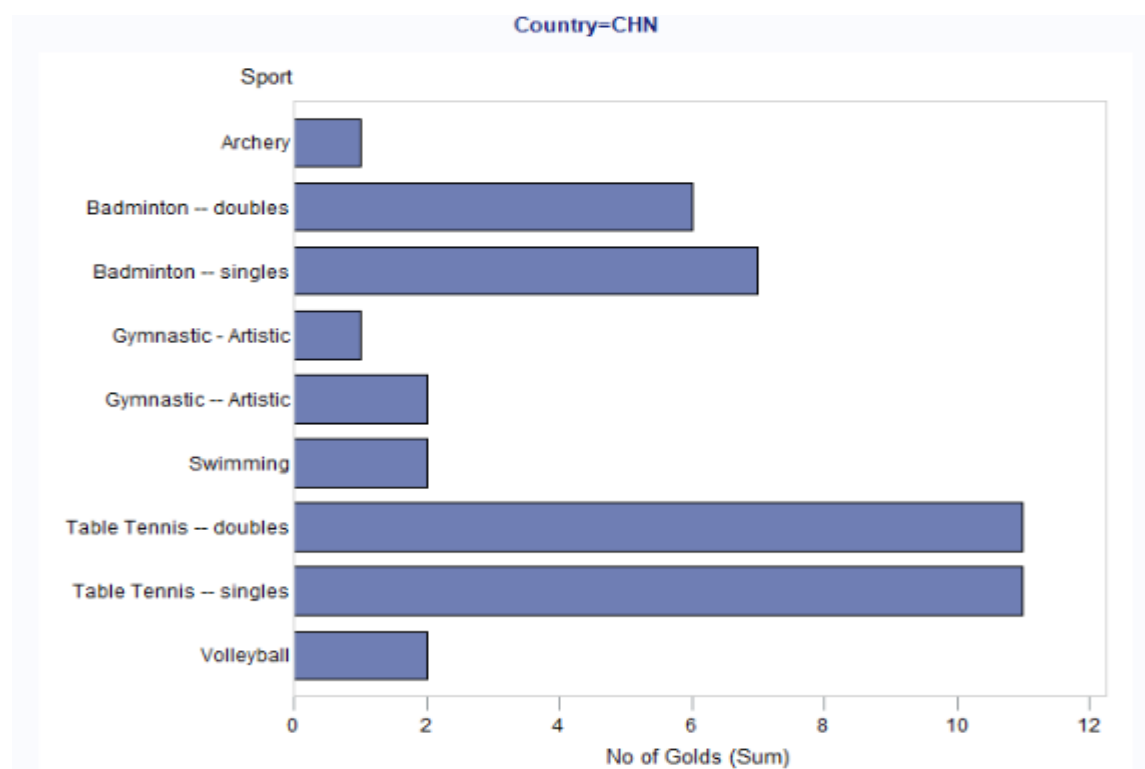
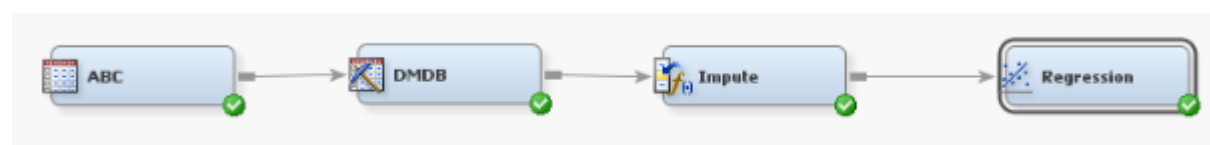


Fig: Average Count of Medals for China for the last 10 Olympics.

MODEL BUILDING AND ANALYSIS

After we were able to create our final dataset, we went into model construction. We were unable to perform time series analysis because of the extent of categories under each sport. Our best bet for this analysis was logistic regression analysis because of the nature of our response variable. Keeping in mind that we want to predict which country will predict the maximum number of gold medals in the Rio Olympics of 2016, we went ahead with Logistic Regression.

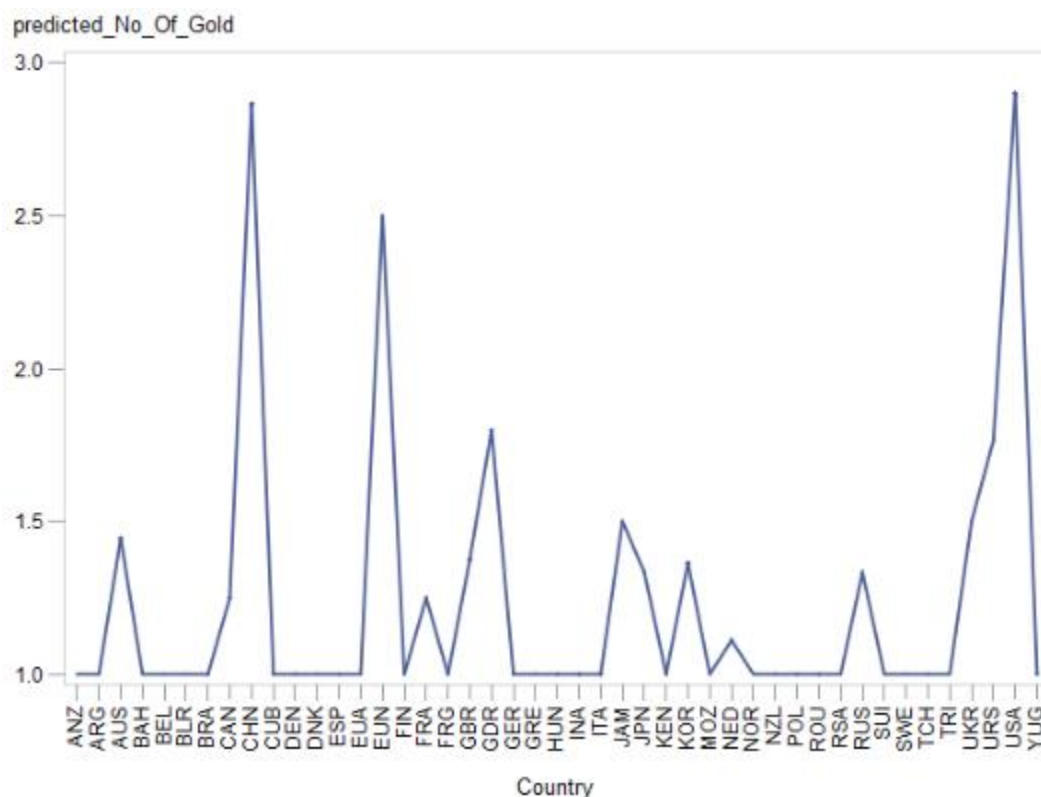
We used the DMDB Node to optimize the performance of the nodes and get basic summary statistics in order to be sure that we were on the right track.



Based on the Average Square Error of 0.22, we went ahead with this model for our prediction. The Mean Square Error of our Model is 0.22 which shows that the model could be used.

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	885.05
ASE	Average Squared Error	0.22
AVERR	Average Error Function	0.73
DFE	Degrees of Freedom for Error	800.00
DFM	Model Degrees of Freedom	2.00
DFT	Total Degrees of Freedom	802.00
DIV	Divisor for ASE	1203.00
ERR	Error Function	881.05
FPE	Final Prediction Error	0.22
MAX	Maximum Absolute Error	0.67
MSE	Mean Square Error	0.22

The prediction model clearly shows that USA and CHINA are undoubtedly the countries with the maximum number of gold medals for RIO2016. This is not a surprise given the history of these two countries. Certain other countries like EUN, GDR and JAM also are predicted to perform well in the Olympics of 2016.



CONCLUSION

The graph shows the predicted value for number of ‘Golds’ versus the countries that will be playing in the Rio Olympics of 2016. In the Next Olympics of 2016, United States and China will have a close battle for the highest Gold Medal winner with each country having equal chances of making it to the top of the scoreboard. Unified Team of Germany and Britain will compete for the third position. Based on our research on this trend of why these countries are making it to the top of the list for years in a row, we have found that they spend a bulk amount on the infrastructure for each sport. China and USA together have more 10000 training centers and spend more than 5 Million Dollars on training Athletes excluding their state salary. There are other Demographic and Physical Factors involved like the height of the player, where they grew up, wind and other geographical possibilities that players use to their advantage.

This paper has raised certain questions, why is United States beating other countries in the overall medal count for so many years at a stretch. There are other questions like why Jamaican Players run so fast, why players less than 24 years of age are consistently winning in Swimming and Gymnastics, etc.

REFERENCES

Modeling Olympics Performance <https://www.pwc.com/gx/en/research-insights/pdf/pwc-olympic-medals-paper2012-final.pdf> by <http://www.pwc.co.uk/>.

<http://www.sloansportsconference.com/activities/research-papers/> by MIT Sloan's Sports Analytics Conference.

<https://channels.theinnovationenterprise.com/articles/analytics-at-the-olympics-an-exercise-in-venture-capital> by Enovation Enterprise.

ACKNOWLEDGMENTS

We thank Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS® and OSU Data Mining Certificate program - Oklahoma State University and Dr. Miriam Mcgaugh, Professor, Department of Marketing for their support throughout the research.

CONTACT INFORMATION

Please Contact the author at:

Name: Antarlina Sen
Enterprise: Oklahoma State University
Address: 127 N Duck Street, Apt F31
City, State ZIP: Stillwater, OK - 74075
Work Phone: 4057620152
E-mail: Antarlina.sen@okstate.edu

Name: Gaurang Margaj
Enterprise: Oklahoma State University
Address: 127 N Duck Street, Apt F32
City, State ZIP: Stillwater, OK - 74075
Work Phone: 9189340292
E-mail: gaurang.margaj@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.