120 years of Olympic History

6103 Data Mining Project Report -- Group 1

Jason Witry,   Sarvesh Bhagat,   Jing Li

1. Introduction

Usain Bolt winning the 2008 100 meter race. Who could have predicted it? Given that he lost in 2004, the only defeat in his career, you would be hard-pressed to have said in 2007 that he would take the gold in Beijing. What happened in between? The answer is that Bolt trained meticulously. He even said so himself, "I really enjoy what I do, and I know the hard work pays off. On the track, it's all about staying number one." It's all about the hard work. If you kept up with Bolt's training regime, his diet, hydration habits, and other training indicators, you might be able to have made a convincing prediction that Bolt would shatter world records in 2008. However, how would you collect this data? Would you collect the information from each athlete to compare? Imagine the enormous resources and time required to collect this data, and this method quickly becomes infeasible.

Let's think a little differently; every Olympic Games, data is collected on the athletes, such as height, weight, age, team, country of origin, and more. Since this data is already available and has been for over a century, we do not have to wait to collect training data for all future Olympic Games to get a decent sample size. This is an enormous advantage, so only one question remains; Can we use athlete characteristics to predict Olympic medal winners as a proxy for training data and skill level?

To answer this, our group work on the data analysis on dataset of "120 years of Olympic history: athletes and results". This data spans over a century of Olympians, recording data as shown in Section 2.1. After collecting the data, we will use Data Preprocessing and Mining techniques to clean, impute and transform the data to see how well we can predict medal winners. We will also compare our results after preprocessing the data to a model run on the uncleaned dataset, to see if preprocessing the data affects our results. Finally, we will create an interactive GUI to deploy our models. Figure 1 provides a brief overview of our workflow.

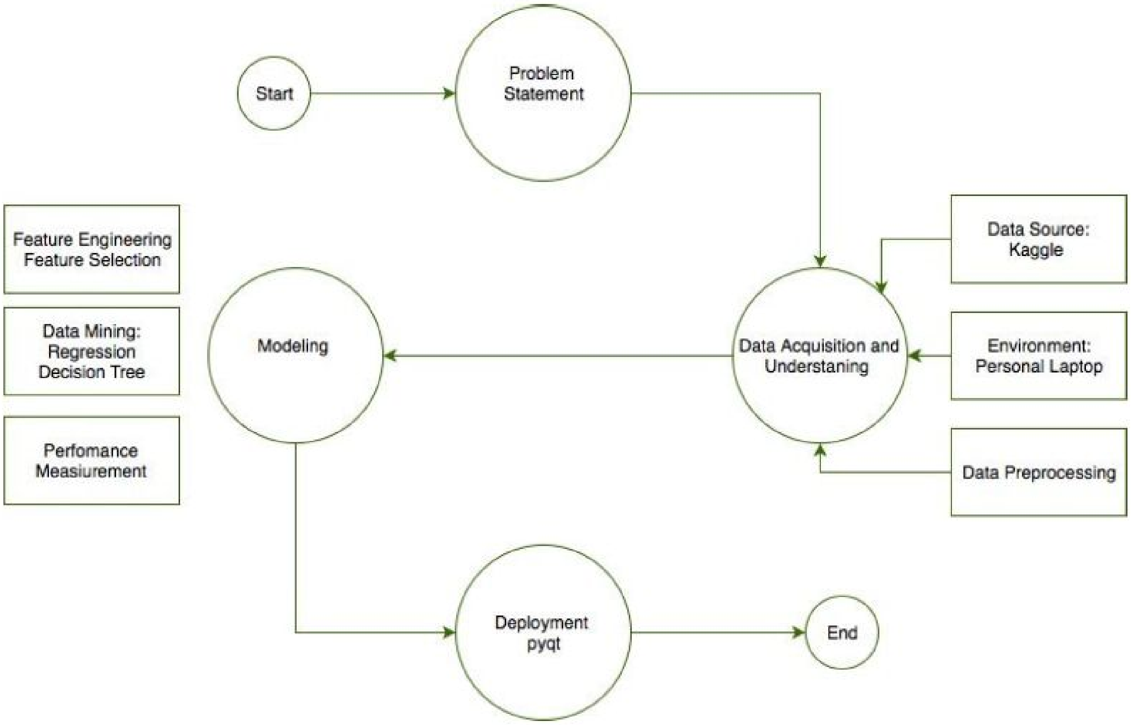

Start
Problem Statement
Data Source: Kaggle
Environment: Personal Laptop
Data Preprocessing
Data Acquisition and Understaning
Modeling
Feature Engineering Feature Selection
Data Mining: Regression Decision Tree
Perfomance Measiurement
Deployment pyqt
End

Figure 1: project workflow

Summer Olympics are the Olympic Games which are held in the summer season. The Winter Olympics, as the name suggests, are held particularly in the winter season. The games and sporting events organized at the Summer Olympics and Winter Olympic are totally different from each other. It is fair to say that Summer Olympic is the most popular competition and the most widely followed sporting events in the world. Both Summer Olympic and Winter Olympic were held in the same years until the year 1992, then International Olympic committee made the decision to organize them at a gap of every two years. Since then, the next Winter Olympics was held in the year 1994, the next Summer Olympics was held in the year 1996. Based on the difference between those two, we will do the data insight on 2 subsets summer and winter separately.

2. Data Set

2.1 Overview

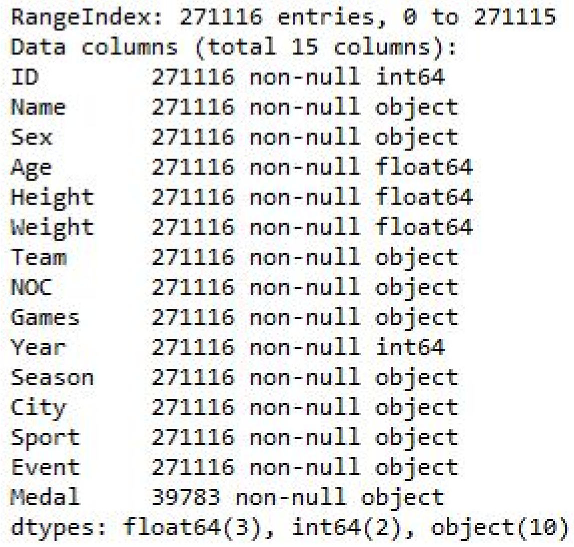
The main dataset we use is historical data of modern Olympic Games, including summer Olympic Games and winter Olympic Games, from Athens 1896 to Rio 2016. The data spans 120 years of Olympics and includes athlete id, name, gender, age, height, weight, team, NOC (National Olympic Committee 3 letter country code), as well as the name, year, season and city of the games. It also records the sport, event, and medal received. The file contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event. The columns are:

- ID - Unique number for each athlete
- Name - Athlete's Full Name
- Sex - M or F
- Age - Integer
- Height - In centimeters
- Weight - In kilograms
- Team - Team name
- NOC - National Olympic Committee
- Games - Year and season
- Year - Integer
- Season - Summer or Winter
- City - Host city
- Sport - Sport
- Event - Event
- Medal - Gold, Silver, Bronze, or NA

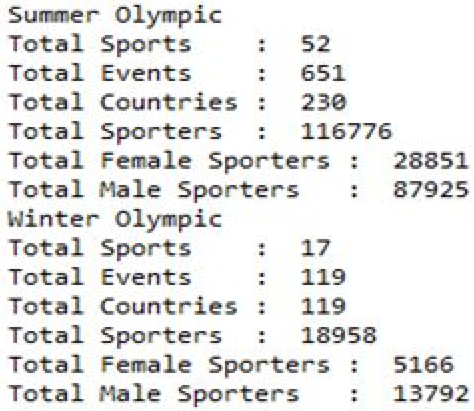
The Summer and Winter Games have not always been staggered. The Winter and Summer Games were held in the same year up until 1992. After that, the Winter Games occur on a four year cycle starting with 1994, then Summer Games in 1996, then Winter in 1998, and so on.

For preprocessing purpose, we also use a dataset of average height and weight in different countries to fill up missing value of body size in main dataset.

2.2 General Information



```
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
ID        271116 non-null int64
Name      271116 non-null object
Sex       271116 non-null object
Age       271116 non-null float64
Height    271116 non-null float64
Weight    271116 non-null float64
Team      271116 non-null object
NOC       271116 non-null object
Games     271116 non-null object
Year      271116 non-null int64
Season    271116 non-null object
City      271116 non-null object
Sport     271116 non-null object
Event     271116 non-null object
Medal     39783 non-null object
dtypes: float64(3), int64(2), object(10)
```



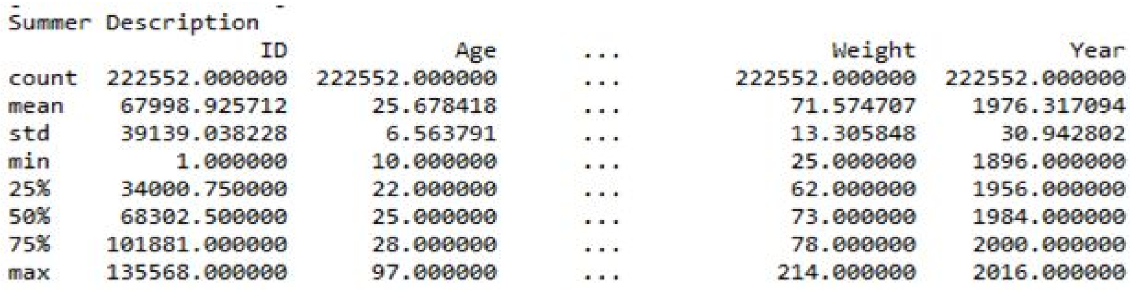
```
Summer Olympic
Total Sports     :  52
Total Events     :  651
Total Countries :  230
Total Sporters   :  116776
Total Female Sporters :  28851
Total Male Sporters   :  87925
Winter Olympic
Total Sports     :  17
Total Events     :  119
Total Countries :  119
Total Sporters   :  18958
Total Female Sporters :  5166
Total Male Sporters   :  13792
```

By using the code df1.info(), we get a detailed list of every column in dataset, including number of rows and data size. There are 271116 rows and 15 columns in total.

Separating the total dataset into 2 subsets, summer and winter, we could do analysis on both subsets. There are more sports in Summer Olympic than in Winter Olympic. The athletes participated in Summer Olympic is ten times more than the one for Winter Olympic. For both games, the number of male athletes was more than 2 times than female.

2.3 Descriptive Statistics

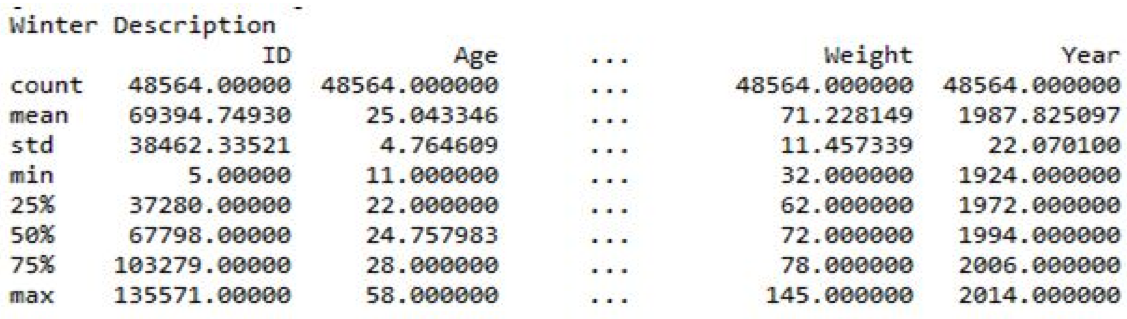
Summer



```
Summer Description
                  ID            Age      ...          Weight           Year
count  222552.000000  222552.000000      ...   222552.000000  222552.000000
mean    67998.925712      25.678418      ...       71.574707    1976.317094
std     39139.038228       6.563791      ...       13.305848      30.942802
min         1.000000      10.000000      ...       25.000000    1896.000000
25%     34000.750000      22.000000      ...       62.000000    1956.000000
50%     68302.500000      25.000000      ...       73.000000    1984.000000
75%    101881.000000      28.000000      ...       78.000000    2000.000000
max    135568.000000      97.000000      ...      214.000000    2016.000000
```

We get descriptive statistics table by code df1.describe. We have records as early as 1896, the oldest athletes was 97-year-old and the youngest athletes was only 10-year-old. A lot of interesting things could be found here.

Winter



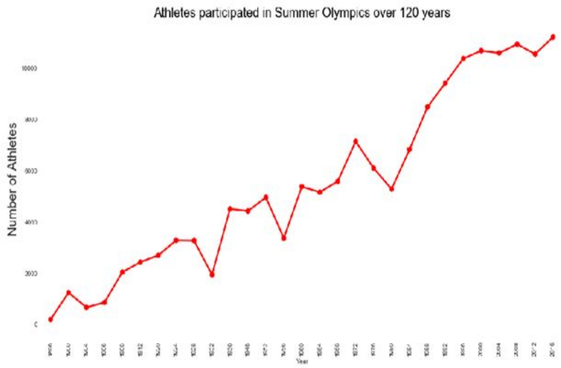
```
Winter Description
               ID           Age    ...          Weight          Year
count   48564.00000  48564.000000  ...    48564.000000  48564.000000
mean    69394.74930     25.043346  ...       71.228149   1987.825097
std     38462.33521      4.764609  ...       11.457339     22.070100
min         5.00000     11.000000  ...       32.000000   1924.000000
25%     37280.00000     22.000000  ...       62.000000   1972.000000
50%     67798.00000     24.757983  ...       72.000000   1994.000000
75%    103279.00000     28.000000  ...       78.000000   2006.000000
max    135571.00000     58.000000  ...      145.000000   2014.000000
```

Different from Summer Olympic, the oldest athletes ever participated in Winter Olympic was 58-year-old. And it looks like all the athletes were in great figures since the heaviest athletes was only 145 pounds.

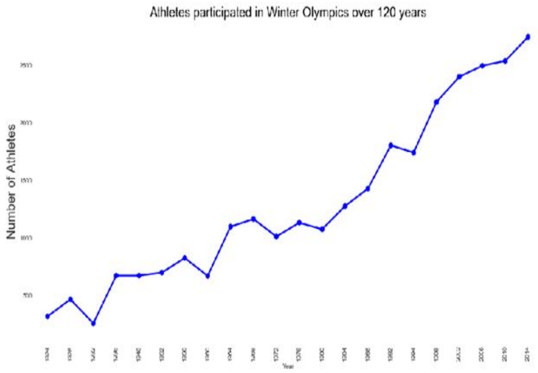
2.4 Participation Analysis

2.4.1 Participation over Years

Summer



Winter

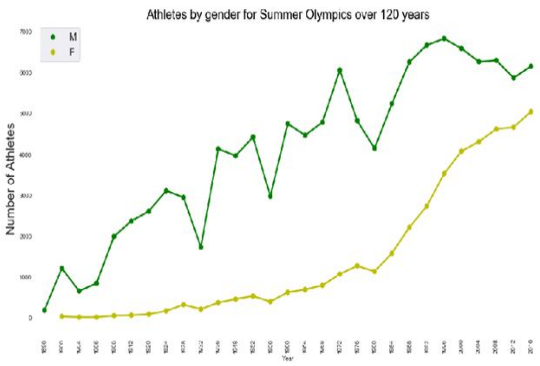


Over the past 120 years, the number of athletes participated in Summer Olympics has been growing fast. Because of the wars and boycott, there were some spikes in certain years.

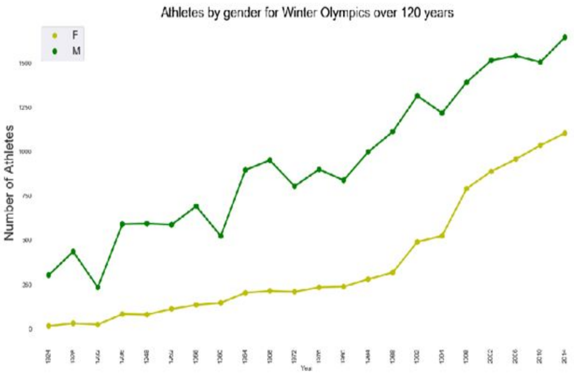
Compared to Summer Olympic, the number of athletes participated in Winter Olympics has been growing more smoothly.

2.4.2 Participation by gender over Years

Summer



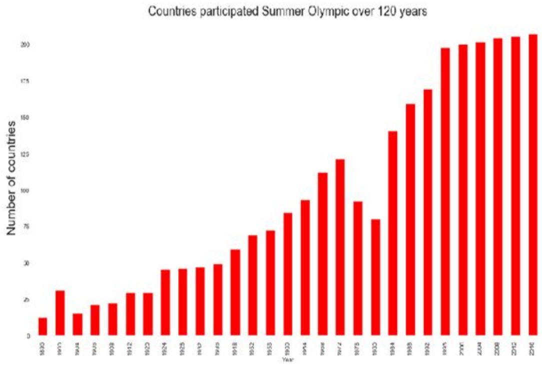
Winter



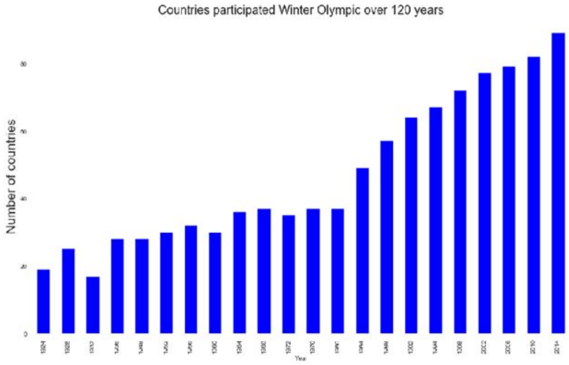
In every Olympic year, no matter it was Summer Olympic or Winter Olympic, the number of male athletes was always bigger than the number of females. Since year 1984, more and more female athletes participated in the Olympic games. It is nice to see that the gap between male and female athletes has getting smaller and smaller, especially in the past five Summer Olympic games.It is a great sign of civilization.

2.4.3 Participation by country over Years

Summer



Winter

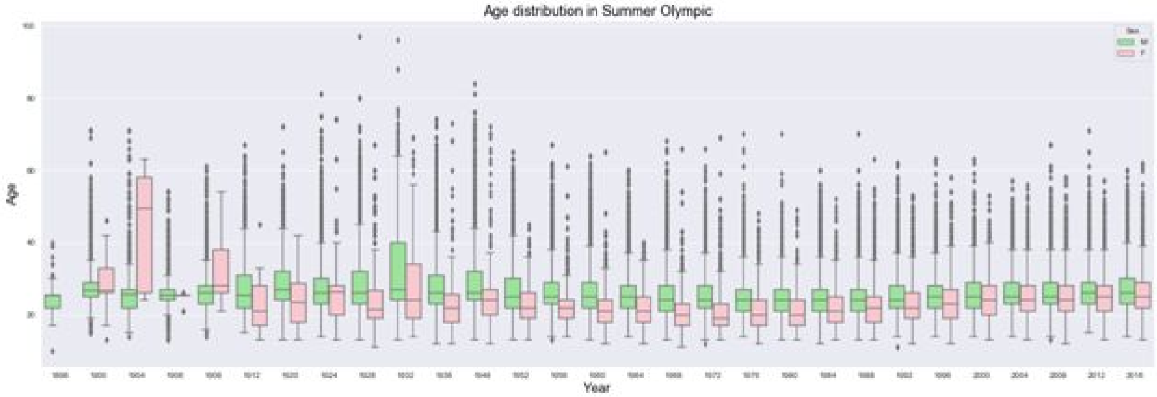


The number of participated countries is growing for both Summer and Winter Olympic Games over the past century. There was big decrease in both 1976 and 1980 because of the boycott.

2.5 Athletes Analysis

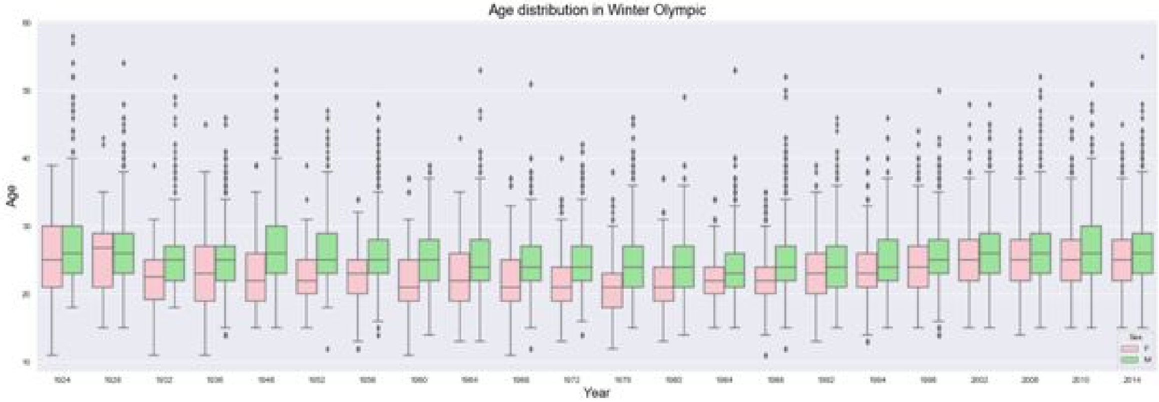
2.5.1 Age Distribution

Summer



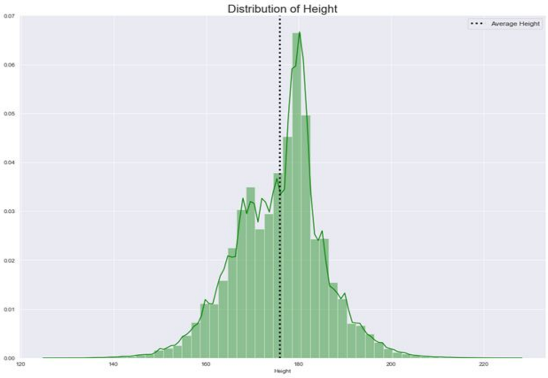
It is interesting that the age ranges were way wider in early last century for both male and female athletes. It is not hard imagine as the Olympic getting more and more popular, most countries would pick the best of best to participate, age limitation must be set for most of the situations.In every Olympic year, the male average age was older than the female average age.

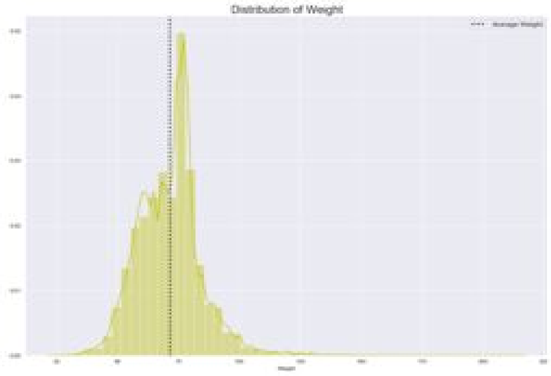
Winter



Compared to Summer Olympic, the age range in Winter Olympic is smaller in general. Just like summer Olympic, the male average age was older than female average age in every winter Olympic year. We see that for both Summer and Winter Games over time, the fluctuations in the distributions settle to being fairly constant. This provides evidence for dropping years that are below a threshold in order to reduce the noise in our dataset.

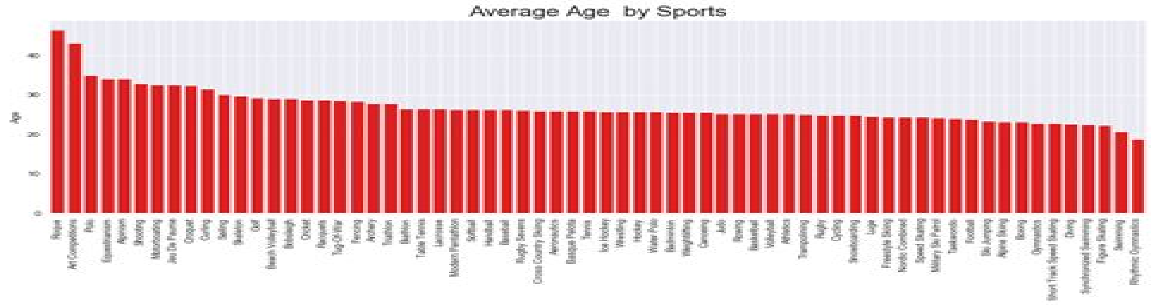
2.5.2 Height and Weight Distribution of Athletes

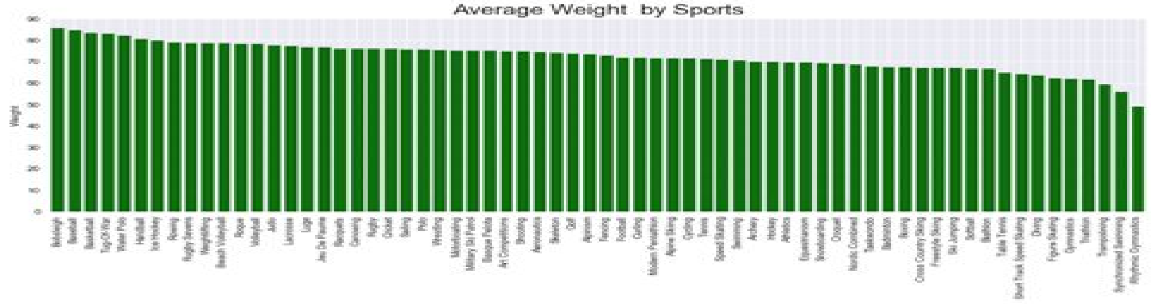


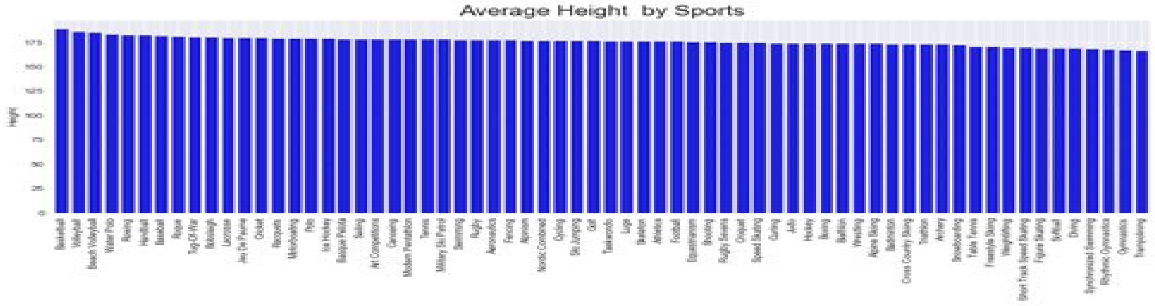


The average weight of all athletes is around 70 kilograms and the average height is around 176 centimeters. The distributions are nearly Gaussian, but are more skewed than we suspect. This may indicate that these variables will perform better in a non-parametric model that does not assume distributions for variables.

2.5.3 Average of Attributes by Sports



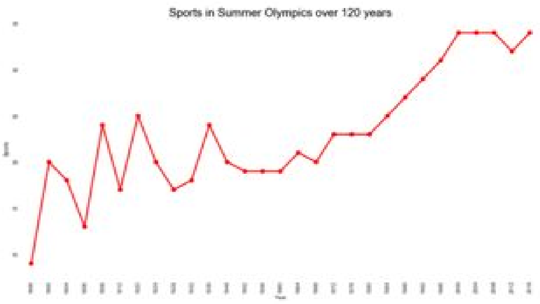


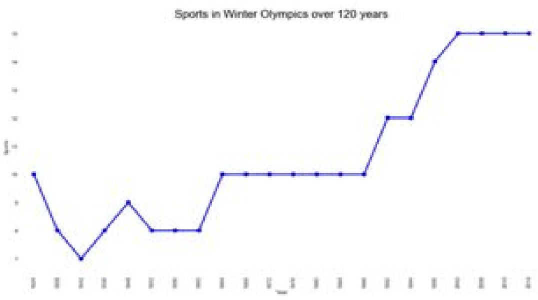


The average weight of athletes for sport Tug-Of-War is the heaviest. Like our common sense, the sport with highest average height is basketball. The average age of athletes for sports Art Competition and Equestrianism are way higher than others. This may indicate that a relationship between an athletes physical characteristics and the Sport or Event will be important. This makes sense, as in general, taller athletes perform better in Basketball than shorter athletes simply be design of the sport.

2.6 Sports Analysis

2.6.1 Sports over Years



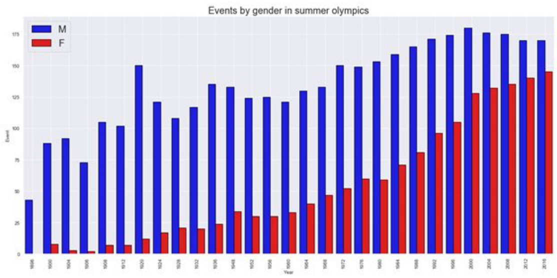


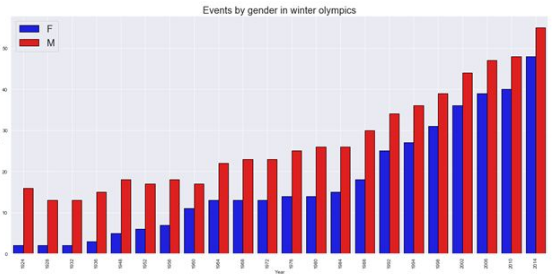
The events include all categories for Sport for both male and female athletes. More and more sports were added as Olympic Games. However, some Sports were also removed. This provides evidence that more recent data will be more important to our predictions, as old data may contain Events that are no longer played, or Events may have been played in the past that are not played anymore, such as Tug of War.

2.6.2 Events by gender over years

Summer

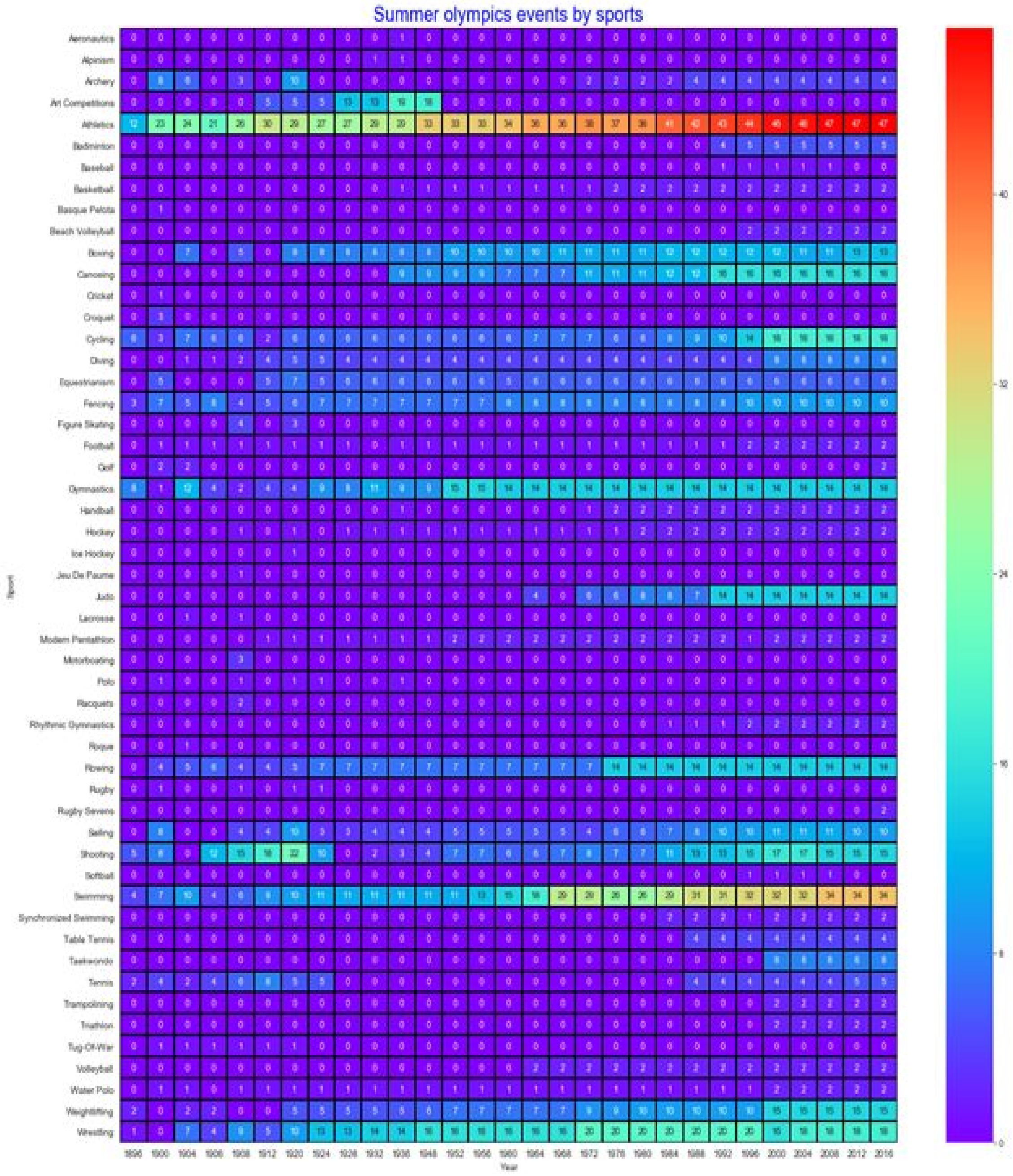
Winter



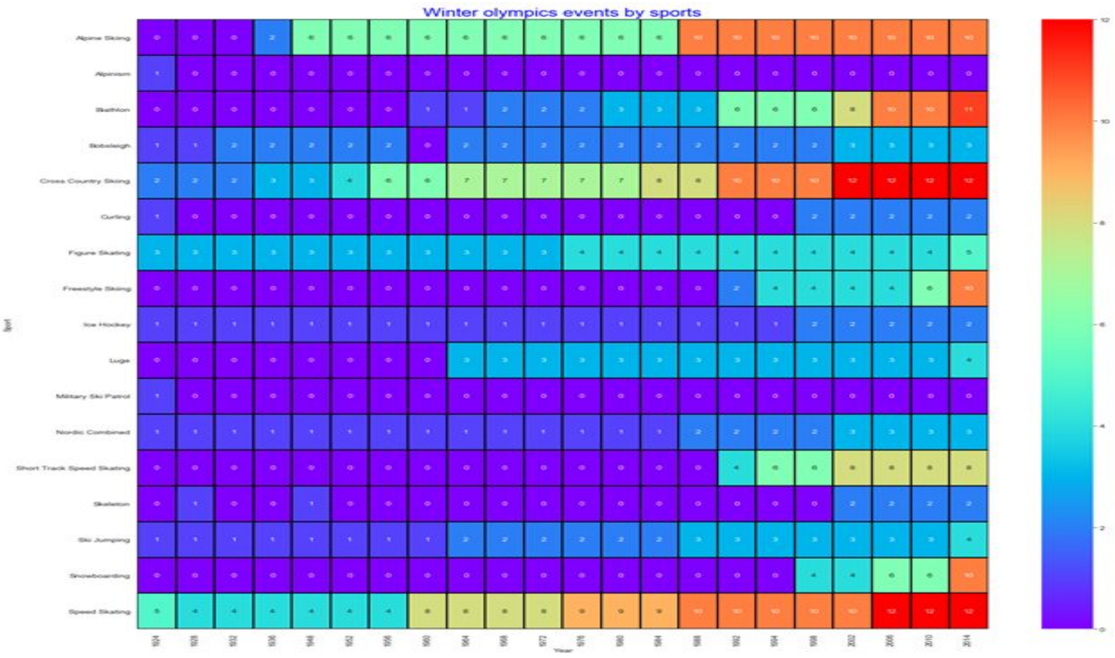


The events of women in the Olympics has been increasing since their first participation in 1900 for both Summer and Winter Olympics, and the gap between women events and man events has been getting smaller and smaller. Since Events are split by Sex, it may be that using more recent years will create more relevant predictions. For example, if we used only data with fewer amounts of women participants, the sample size for women may be too small to make a good prediction.

2.6.3 Event by sport



The numbers in each cell represents total number of events for every sport contested over the past years. The top 3 sports holding the most events in Summer Olympics are athletics, swimming and wrestling.

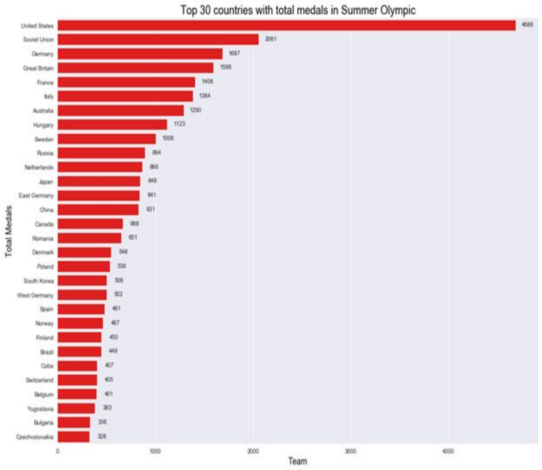


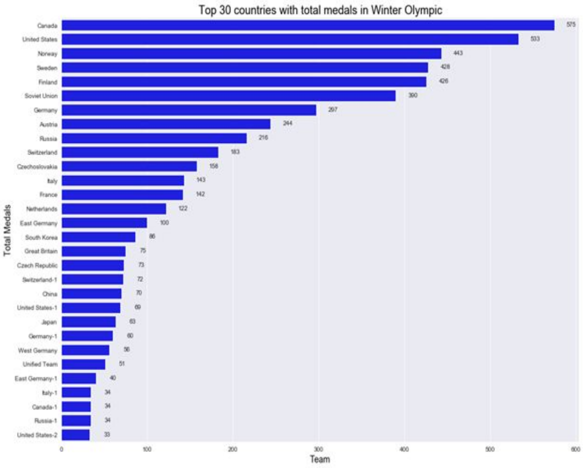
In Winter Olympic games, skating like speed skating, cross country skating, alpine skiing hold the most events in general. This indicates that the Sport will not be as important in predicting as the Event, as different sports contain different events. In order to reduce the noise in our model from this, we will remove the Sport column from analysis.

2.7 Medal Analysis

2.7.1 Top 30 countries with the maximum # of medals

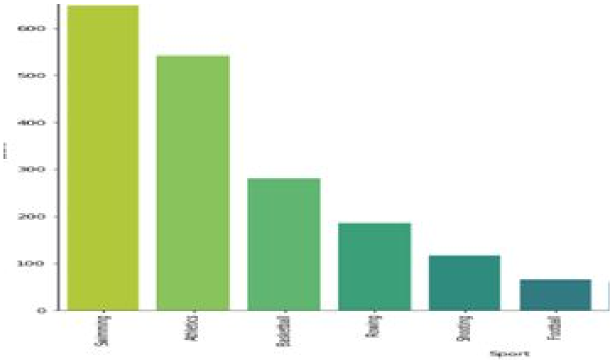
Summer ***Winter***





The top 3 countries with the most medals in Summer Olympic are United States, Soviet Union and Germany. The top 3 countries with the most medals in Winter Olympic are Canada, United States and Norway. From this analysis, we can conclude that the country of origin may be a good indicator of medal winners, as it seems that it is more likely that medal winners will come from the United States rather than Norway for example.

2.7.2 Medal by Sport for USA

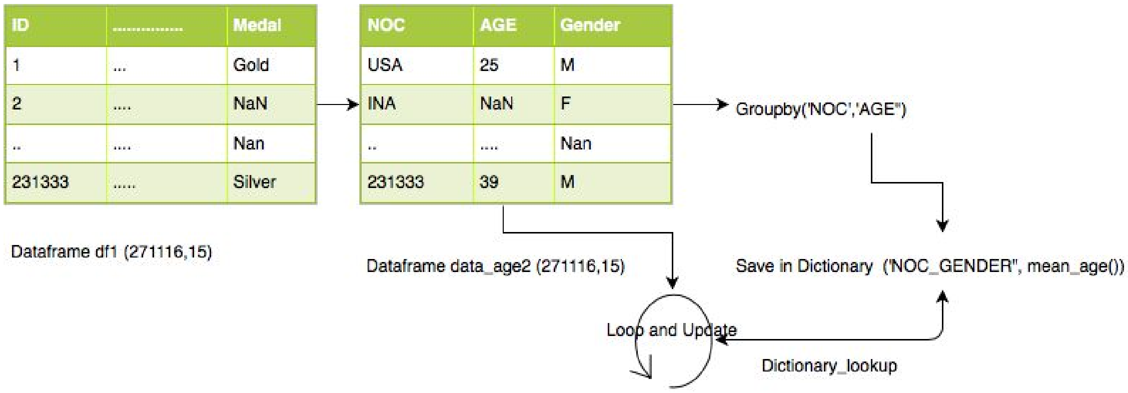


The top 3 sports USA won maximum gold from are swimming, athletics and basketball in order.

From the simple descriptive statistics table and all the plots, a lot of theory get confirmed. For example, male athletes in general have longer sport career and the number of male athletes is bigger than females. We get an overall understanding of the data set.

In sum, our data exploration shows that dropping the year before a threshold value could improve our models by eliminating noise from our dataset. We can see interesting combinations of variables that could be important, such as athlete physical characteristics and Event. It will likely improve our model to drop the Sports column, as it seems to generalize the Events. Country of Origin was also shown to be potentially important, as number of medal winners varies per country. It also may be that certain countries excel at specific events, such as the dominance of the Soviet Union in Hockey during the 1970's.

3. Preprocessing



ID	....	Medal
1	...	Gold
2	....	NaN
..	....	Nan
231333	.....	Silver

Dataframe df1 (271116,15)

NOC	AGE	Gender
USA	25	M
INA	NaN	F
..	....	Nan
231333	39	M

Dataframe data_age2 (271116,15)

Mean age, height and weight were calculated based on mean value of column grouped by country and gender. This was done so that age, height or weight based on person's origin

These values were used to update the initial Dataframe.

After this step, there were still some values missing:

Age: 13
Height 108
weight : 257

These missing values were updated using average values of from the new updated Dataframe.

After this step, there was no missing value in the data.

Age Distribution:

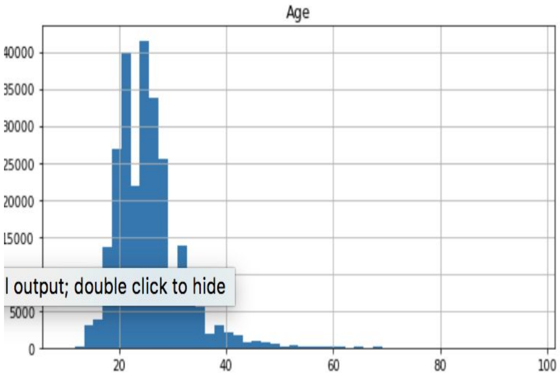


fig 1. Age distribution before Processing

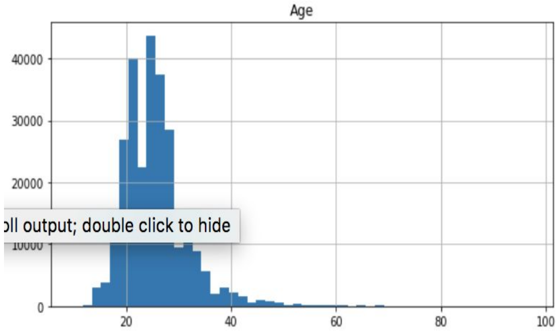


fig 2. Age distribution after Processing.

Height Distribution pre and post processing of data.

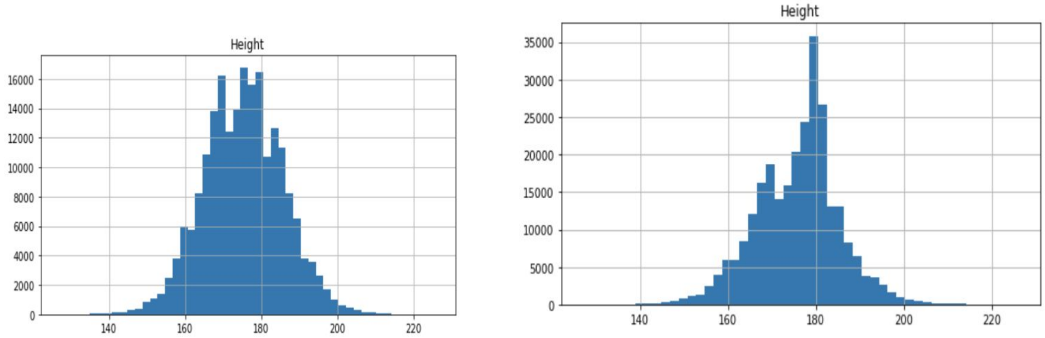


fig3. Pre and post processing height distribution

Weight Distribution:

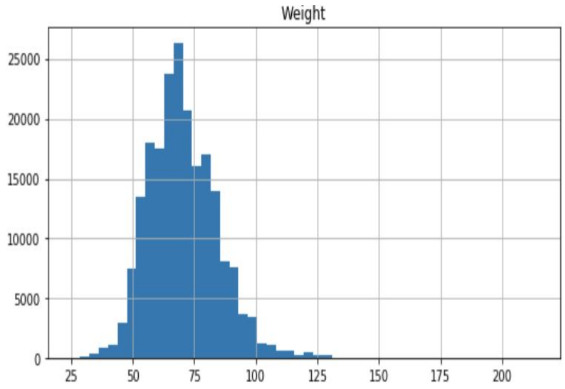


fig 4. weight distribution before data processing

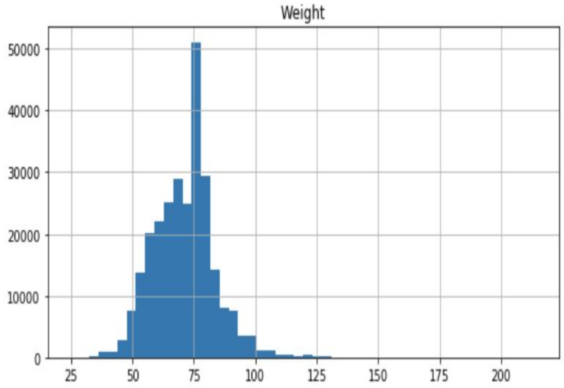


fig 5. weight distribution after data processing

4. Modeling

Before modelling, the data need to be encoded since this data is mostly categorical.

Below code, was used to encode the data using LabelEncoder.

```
from sklearn.preprocessing import LabelEncoder

# LabelEncoder
le = LabelEncoder()
# apply "le.fit_transform"
df_encoded = df1.apply(le.fit_transform)
```

Next step, to find correlation between the features.

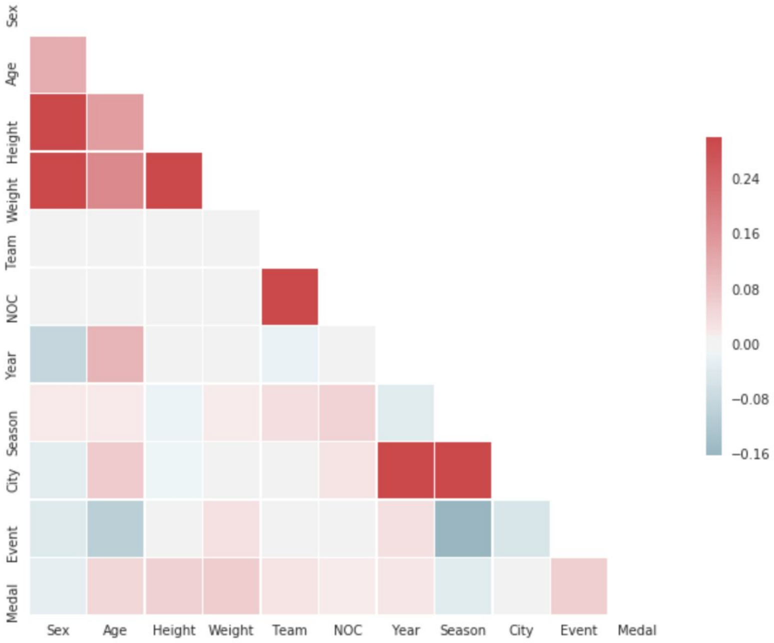


fig 6. Correlation matrix for feature.

4.1. Feature Selection

features like ID, Name, Game and sport. Feature 'Game' provides repeated information,

Feature 'sports' was removed because it provides a broad category of an event in olympics.

df1 = df1.drop(['ID', 'Name','Games','Sport'], axis=1)

Below features were used for modeling:

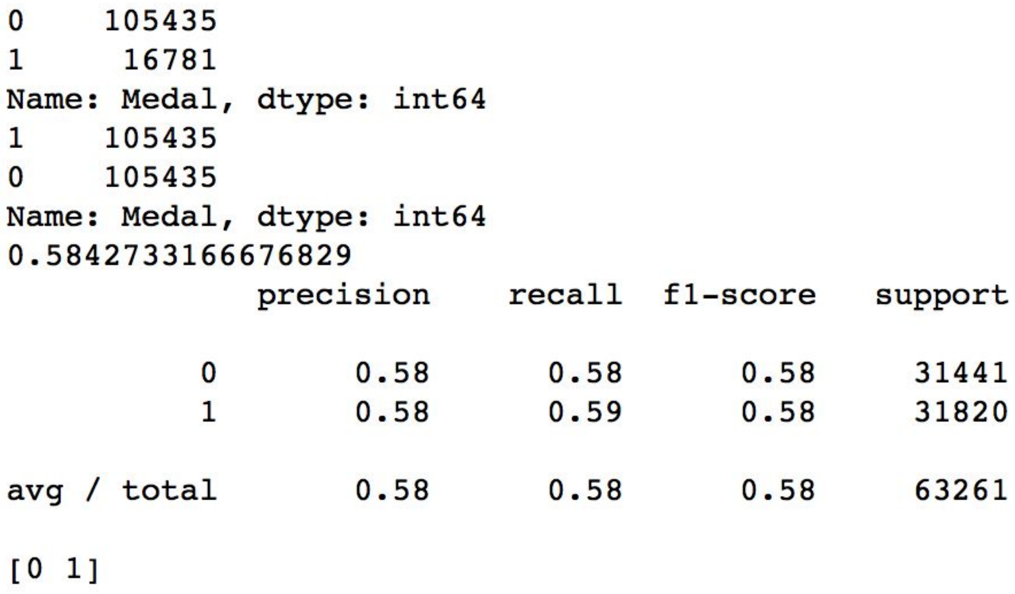
['Sex','Age','Height','Weight','Team','NOC','Year','Season','City','Event']

4.2. Data Mining

For the first run, data was fit to logistic regression. We chose logistic regression because the target variable is a binary categorical variable, and some of our predictors are floats, which lends itself to a logistic regression. Due to unbalanced data, only result with target 0 were predicted. To eliminate this problem, Upsampling and Downsampling techniques were used.

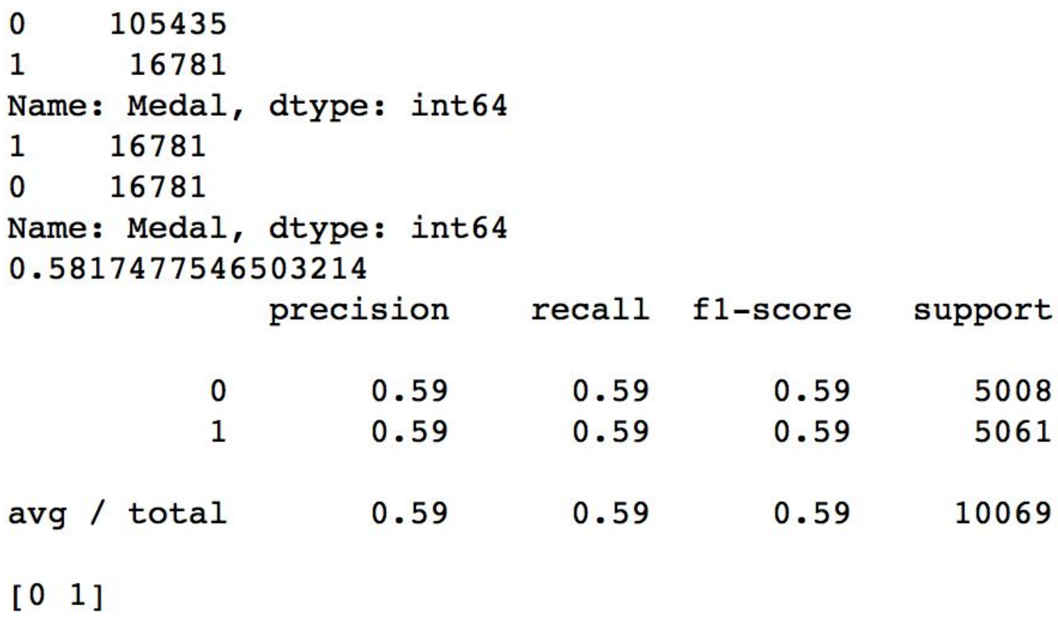
4.2.1. Logistic Regression

With upsampled data, logistic regression was able to predict both target values of 0 and . But the accuracy of model with this upsampled data is only 0.58. The same result was achieved with downsampled data as well. fig 7 and fig 8 show the results.



```
0    105435
1     16781
Name: Medal, dtype: int64
1    105435
0    105435
Name: Medal, dtype: int64
0.5842733166676829
             precision    recall  f1-score   support

          0       0.58      0.58      0.58     31441
          1       0.58      0.59      0.58     31820

avg / total       0.58      0.58      0.58     63261

[0 1]
```

fig 7. Logistic Regression Result with upsampled data.

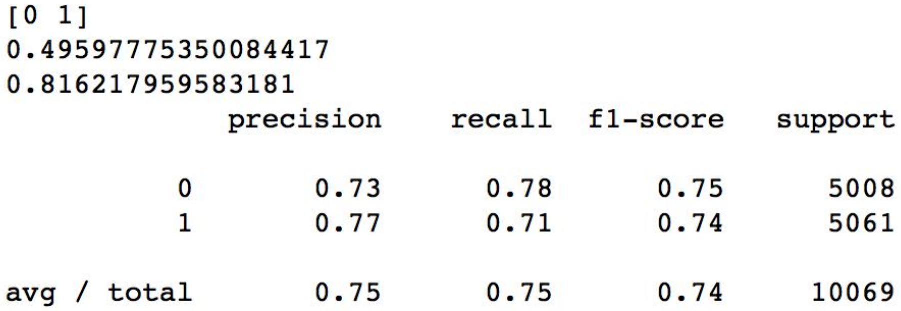


```
0    105435
1     16781
Name: Medal, dtype: int64
1    16781
0    16781
Name: Medal, dtype: int64
0.5817477546503214
             precision    recall  f1-score   support

          0       0.59      0.59      0.59      5008
          1       0.59      0.59      0.59      5061

avg / total       0.59      0.59      0.59     10069

[0 1]
```

fig 8. Logistic Regression Result with downsampled data.

4.2.2.. Random Forest classifier

When the upsampled or downsampled data was fitted to Random Forest Classifier model, a high accuracy was achieved. fig 9. provides the details of the result.

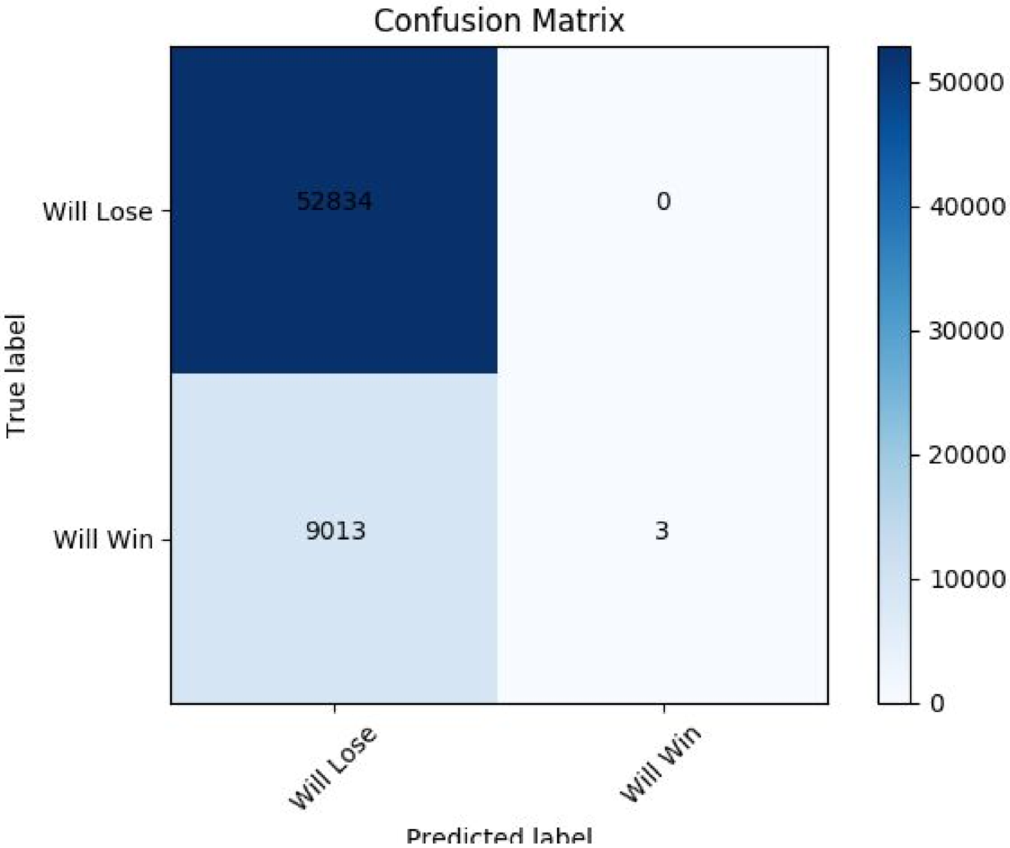


```
[0 1]
0.49597775350084417
0.816217959583181
             precision    recall  f1-score   support

          0       0.73      0.78      0.75      5008
          1       0.77      0.71      0.74      5061

avg / total       0.75      0.75      0.74     10069
```

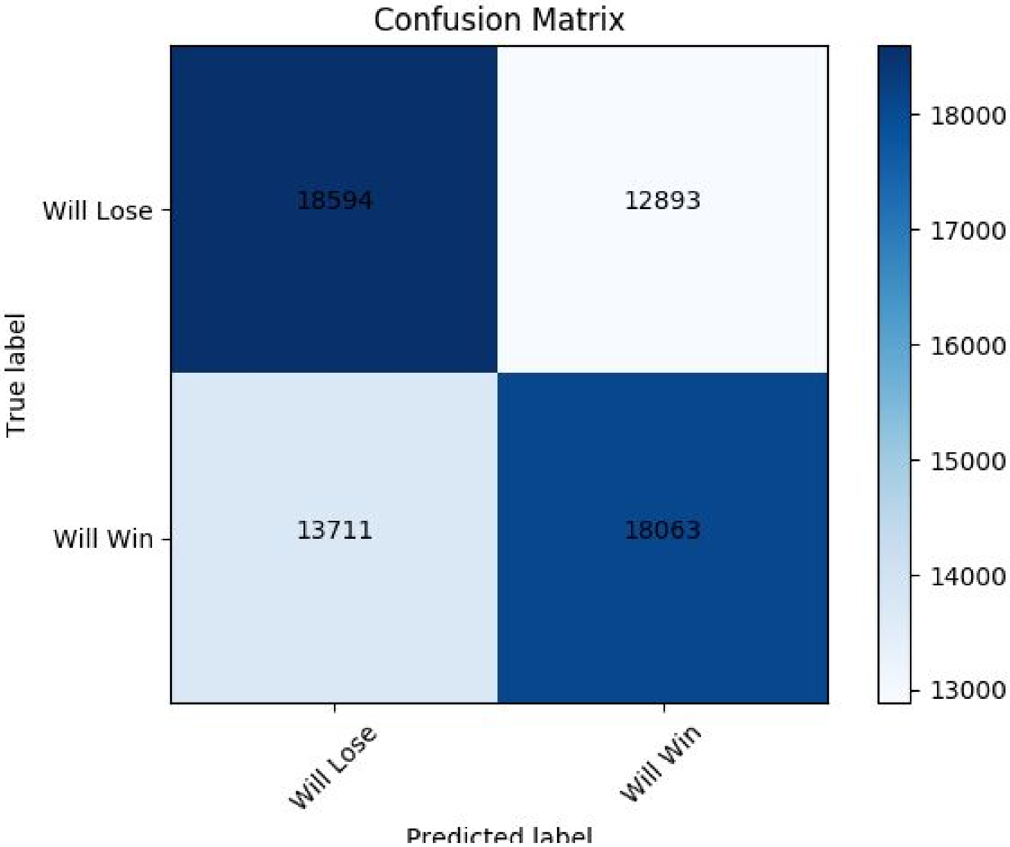
fig 9. Random Forest Classifier Result with upsampled data.

5. Results and Evaluation

We can see clearly that preprocessing the data improves the model significantly. Without preprocessing, the models suffer from an imbalanced dataset and many missing values. In the case of Logistic Regression, the confusion matrix without preprocessing is shown in the figure below.

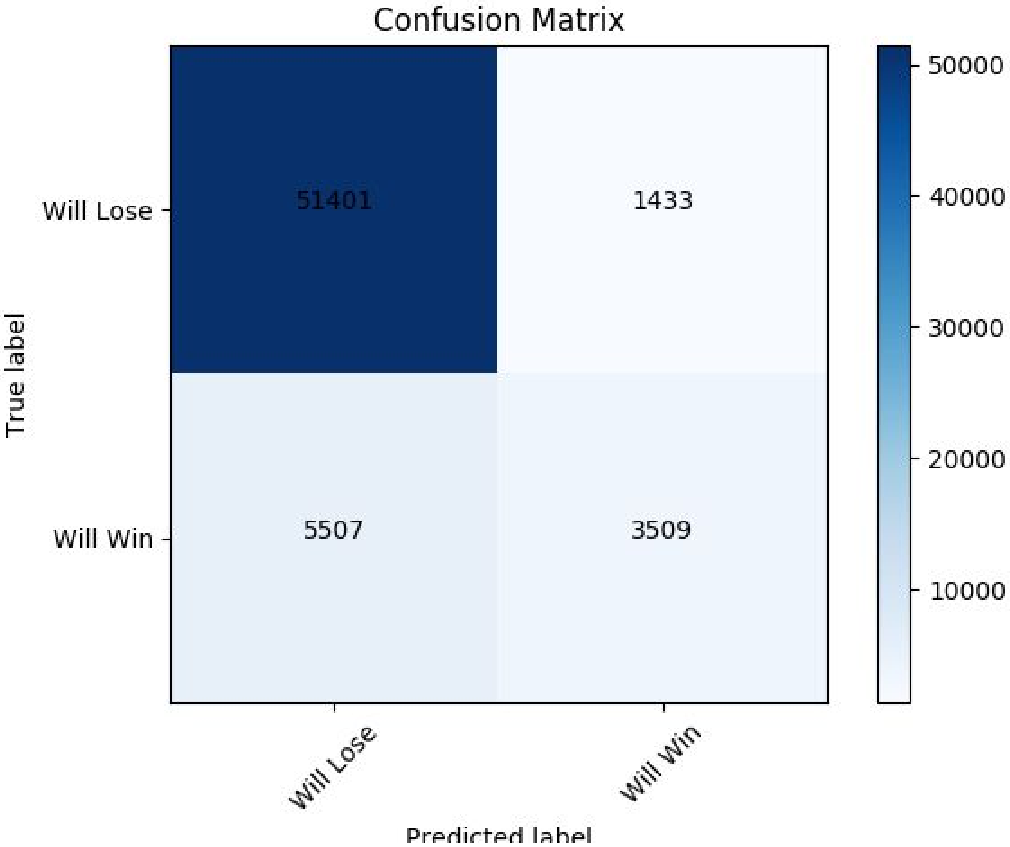


We can see that the model rarely predicts medal winners correctly, and the recall for medal winners is near zero. The precision for medal winners is 1, normally a perfect score, but the confusion matrix shows that we rarely predict someone is a medal winner, so this is not a reliable conclusion. However, once we preprocess the dataset, we find that the predictions improve significantly. In the case of oversampling in addition to mean imputation, standardization, and cutting out years before 1990, logistic regression returns the following confusion matrix.



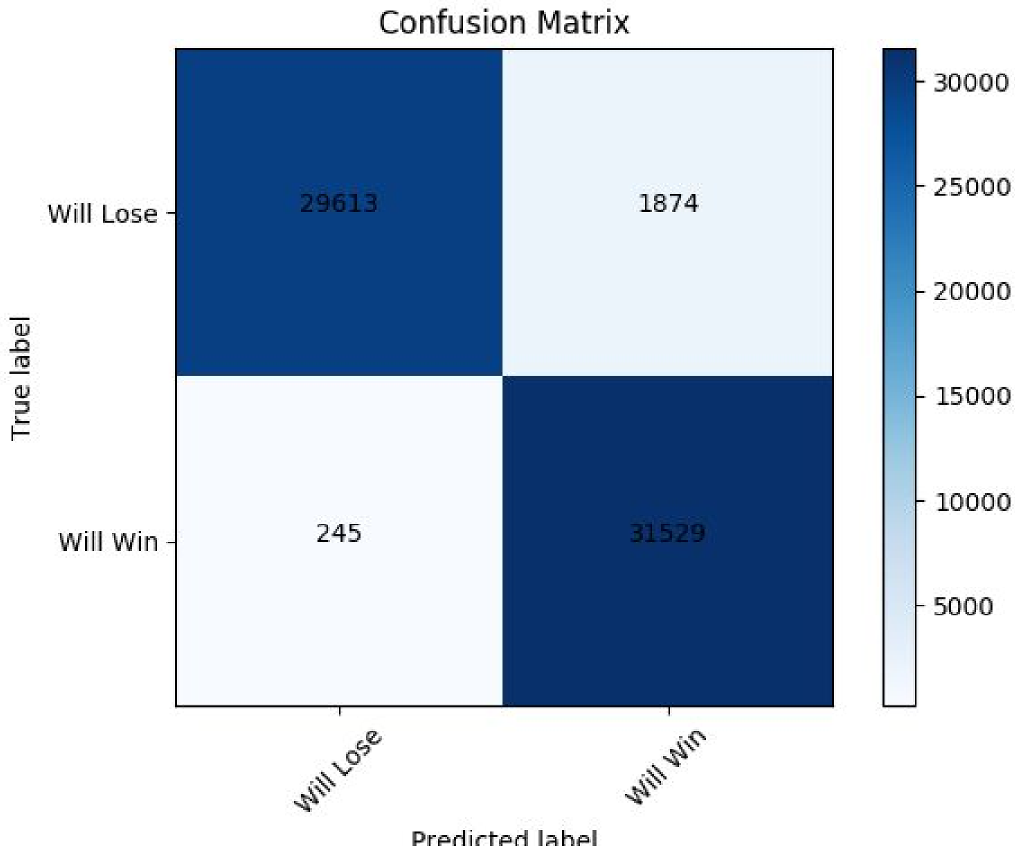
While the recall for winners drops significantly, this is a more reliable prediction, as we do not suffer from the class imbalance. The model also benefits from less noise due to years that are cut off.

For a Random Forest model using 10 predictors, we see the following results with no preprocessing.



The precision and recall for the non-medal winners is high, as expected from our imbalanced dataset. Similar to the logistic regression model, the precision and recall are lower. However, the Random Forest model seems to perform better on the dataset than Logistic Regression in this case, as we end up with more predictions for medal winners.

A Random Forest with preprocessing returns the following results.



We see a drastically improved precision and recall for each class prediction, above 0.9 for each. This is the best model we found for our data.

7. Summary and Conclusion

In conclusion, a Random Forest is the best model for our dataset, along with the preprocessing steps of considering years after 1990, mean imputation, and upsampling our non-medal winners. The Random Forest performs better than Logistic Regression due to the fact that our dataset is primarily categorical variables. Random Forests handle categorical variables in a much more natural way than Logistic Regression. Considering years after 1990 not only removes noise in the age, height and weight distributions, it also makes our predictions more relevant for future data like the 2020 Olympics. Imputing the missing values with the means prevents the loss of potentially relevant information, and our method using the Country of Origin and Sex is better than simply using the dataset mean. Finally, upsampling our dataset balances the classes, which is an important step for the Random Forest model, especially when it comes to pruning the trees. We attempted to mitigate this by setting the max depth rather than the minimum sample pruning techniques, which could lead to challenges of its own. Interestingly, letting the Random Forest algorithm run to completion without pruning does not seem to cause overfitting in our dataset. This suggests that our splits seem to divide the dataset well, and there are not many nodes with significant class imbalances in them.

As a result, we can predict Olympic medal winners with the data collected each Olympic Games, and yield a meaningful result that returns a high accuracy, precision and recall for our classes. It seems that there are significant relationships outside of training regimens that make a champion. However, given that training regimens result in different physiques for an athlete, the training regimen likely influences the height and weight relationship of an athlete. Also, younger athletes are in general stronger and more fit than older athletes, which can be important for physical events. Conversely, older athletes may perform better in more skilled based events, such as Archery. The age in this case likely reflects years perfecting the skill, which is an important feature for many events. In summary, we were able to accomplish the task of predicting Olympic champions using data that is readily available and cheap to gather, meaning future predictions of Olympic winners likely do not need to gather data on training regimen, years perfecting the skill, and other difficult information to gather.

Future steps will be to apply our model to the 2020 Olympic winners and test how well we predict them. We also may be able to extend our model to predict which medals athletes will win (bronze, silver or gold) using the Random Forest. The Random Forest can naturally handle the ordinality of this new target.

8. References

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#noc_regions.csv

https://wiki.python.org/moin/PyQt4

https://scikit-learn.org/stable/

https://www.worlddata.info/average-bodyheight.php

https://en.wikipedia.org/wiki/List_of_average_human_height_worldwide

http://www.wecare4eyes.com/averageemployeeheights.htm

9. Appendix

For this project, personal laptops will be used. If more computation power is required then the project will use a cloud platform like Amazon Web Services or Google public cloud.

● Machine Configuration: I5 Intel chip and 64 GB RAM

● SDK: Pycharm Community Version

● Python: 3.6

○ Numpy and pandas

○ MatplotLib

○ Sci-kit

○ PyQt4