6103 FINAL PROJECT PROPOSAL

Jason Witry, Sarvesh Bhagat, Jing Li

Table of Contents

Project Overview

Problem Statement

Predict whether or not Olympians will win medal based on the historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. Out of the various features provided in data set, project will use some or all to build a model which will predict if a Olympian will win a medal or not. Figure 1 provides a brief overview of our workflow.

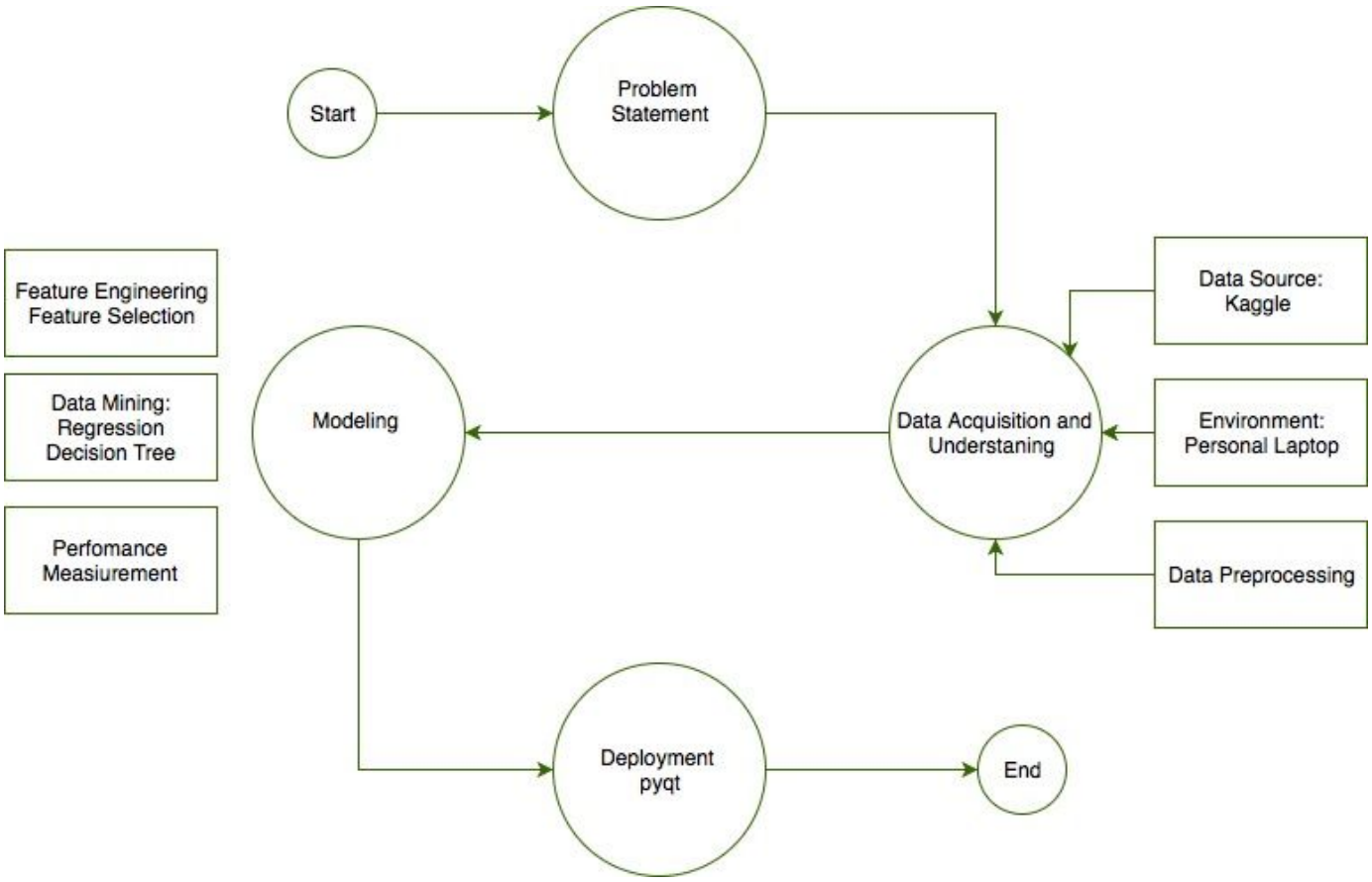

Figure 1: project workflow

Data Acquisition and Understanding:

Data Source

We will use the ”120 years of Olympic history: athletes and results” dataset from Kaggle to build our model. 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 dataset will need to be cleaned, with redundant features and rows removed, instance selection on events that have persisted for 120 years, and removal of those events that are specific to only a few Olympic Games. We will impute missing values in age, height and weight, as well as adjust the medal into a boolean (received medal or did not). Our data is on several different scales, and will have to be adjusted for our logistic regression model. Finally, the medal boolean will have to be balanced using (INSERT METHOD TO BALANCE).

Environment

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

Data Preprocessing:

- Data Cleaning: correcting bad data, filtering incorrect data out, reducing the unnecessary detail, detecting discrepancies and dirty data.
- Data Transformation: converting and consolidating data. smoothing, the feature construction, aggregation or summarization of data, normalization, discretization and generalization.
- Data Integration: merging data from two data sets, “athlete_events” and “noc_regions”
- Data Normalization: attributes would be expressed in the same measurement units and use a common scale or range
- Missing Data Imputation: adding reasonable estimates of suitable data values.
- Noise Identification: detecting random errors or variances and applying correction-based process.

Modeling

Feature Engineering and Feature Selection

Project will try to achieve reduction of the data set by removing irrelevant or redundant features. The goal will be to find minimum set of features such as the resulting probability distribution of the data output attributes is as close as possible to the original distribution obtained using all attributes.

By searching in different directions like Sequential Forward Generation (SFG) and Sequential Backward Generation (SBG) and using different strategies like Exhaustive Search, Heuristic Search and Nondeterministic Search, FS can be explored in many perspectives. we will use selection criteria like Information Measures, Distance Measures, Dependence Measures, Consistency Measures and Accuracy Measures to select the best subset of attributes.

Data Mining Algorithm

To predict whether an athlete will medal or not, we will use both a logistic regression and decision tree/random forest. We will use a standard logistic regression, but will prune the trees to limit bias in our result.

Performance Measurement

To measure our performance, we will compare our result to random classification (50% accuracy). We will use a confusion matrix to obtain the classification accuracy, precision, recall and Cohen Kappa score of our model. From the precision and recall, we will be able to evaluate how well we balanced our classes. The Cohen kappa score and classification accuracy will allow us to compare our model with preprocessing to a model on the raw data and evaluate potential improvement from cleaning.

Deployment

Project will use PyQt4 GUI development toolkit for display result of model.

Reference Materials

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/

To accomplish the above, we will be using the Python framework with modules NumPy, Pandas, MatPlotLib and Sci-Kit Learn. NumPy and Pandas are the standard for Python when it comes to manipulating arrays and datasets. MatPlotLib is Python's plotting module, which we will use to create our data visualizations. Sci-Kit Learn is a Python module dedicated to data processing and the implementation of data mining algorithms, and we will use it to build our predictive model. We will wrap the software in a user-friendly GUI using PyQt4.

Project Schedule

Date	Deadline	Final Project Schedule	Notes
4/2/2019		Group meeting	After class
4/3/2019 -- 4/6/2019	Proposal due by April 6th	Group Proposal	
4/7/2019 -- 4/8/2019		Data Preparation	
4/9/2019	Quiz 4	Group meeting	After class
4/10/2019 -- 4/15/2019		Data analysis: Preprocessing, modeling, testing	
4/16/2019	Final Exam	Group meeting	After class
4/17/2019 -- 4/22/2019		write final report, prepare slides, representation rehearsal	
4/23/2019	*Final Presentation*	Group meeting: Presentation rehearsal	Before Class