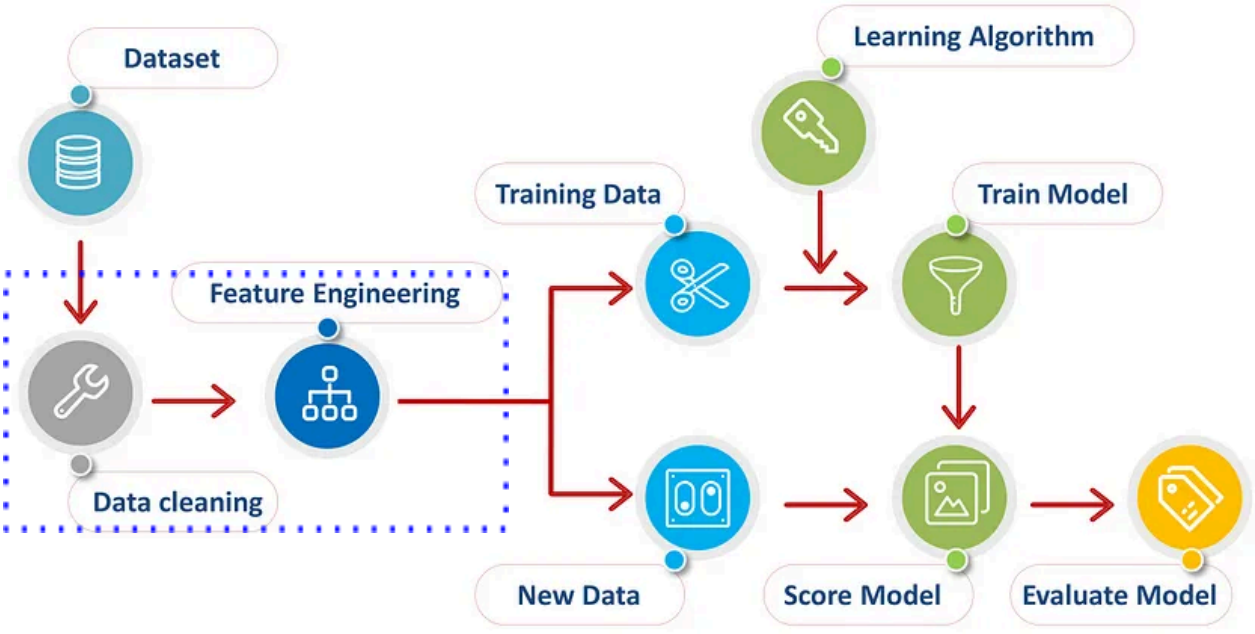
Ahmedabad University

ML Project Weekly Report

Coure Name: Machine Learning (CSE 523)
Week Number: 5 (3rd March - 9th March 2024)
Group Name: White
Instructor's Name: Mehul S Raval
Project no.: 7

Project 7 | Athlete Statistics Visualization and Prediction

Introduction



Ready with appropriate research work and literature review regarding the project's requirements, we made a flow structure which commenced with Data Cleaning followed by Feature Engineering.

The data that was collected was unusual and noisy; null values were in majority to a good extent for a group of features. Since the predictability of the machine learning model/s that would be applied on the data also depends on the quality of the data (*Garbage In = Garbage Out*), it is essential (and strictly mandatory!) for us to perform Data Preprocessing to help make out important assumptions and expressions of features and modalities, besides enhancing the accuracy of the machine learning model/s.

Weekly Activities

1. Data Preprocessing

We followed the following steps for the task of preprocessing the provided data files(.csv), using Python:

- Importing essential libraries for required data operations
- Importing the datasets into the data structures compatible with the Python environment
- Management of the null values
- Encoding categorical data
- Splitting the dataset into Test and Training datasets
- Feature Scaling

The first two steps were easily implemented by going through the syntax of the programming language. Before proceeding to the third step, we studied and observed the data through rough glances to get an idea of the structure it possesses. As we found out that null values were present to a great extent, we employed Data Imputation.

1.1 Data Imputation

In Data Imputation, we retain the majority of the information by substituting null values with a different value. This needs to be done since it would be impractical to keep on removing data from a dataset every time, besides the reduction in the size of the dataset under consideration (which would further raise questions about bias and impairing analysis).

Our research for techniques under this domain gave valuable insight to the technique named MICE (Multiple Imputation by Chained Equation) which is itself based on the iterative nature of Machine Learning algorithms. We decided to go on with this technique in light of the learnings of our current curriculum. Here is a short insight regarding the methodology:

Fill in missing values from random draws of non-missing data
For each iteration
  For each variable v with missing values
    Optional: subset data where v was originally nonmissing
    Train model $v \sim \boldsymbol{X}$ where $\boldsymbol{X}$ are the other variables in the dataset
    Do one of:
      1) Replace missing values with predictions from model
      2) Replace missing values using mean matching
  End
End

1.2 Correlation-based Feature Importance

Feature selection is important to identify the most relevant variables contributing to a model's performance, ensuring that the model does not pick up patterns that seem to appear while testing but does not represent closeness anywhere near to the ground truth.

We decided to implement this as it selects subsets of features that are highly correlated with the target variable but have low correlation with each other. The underlying principle consists of a subset of features that can provide the maximum amount of information about the target variable while minimizing redundancy among the features.

1.3 Quartile Ranging for RSI categorization

RSI Quartile Range used to categorize athletes into 4 groups/classes.

These techniques were employed in Python using a Python notebook on Google Colab.

Challenges Faced

Noise in Data:

Upon observing great volumes of rows and columns filled with N.A., we were convinced that a great deal of effective measures were needed to take care of the missing features on top of unraveling the black box of the ground truth we were dealing with.

Missing information introduces three main challenges:

- Introducing a significant Degree of Bias
- Increased Difficulty in effective Data Processing and Analysis
- Accuracy and Efficiency is hindered

Furthermore, dismissing instances with missing value/s leads to addition of more bias or impair the generalization of the results.

Technique Selection based on Complexity and Performance:
The first time we faced the need of data cleaning and implemented it was during our first year in the course CSD102: Introduction to Data Science. However, the techniques then employed were equivalent to something that barely scratches the surface for what the problem definition of the current ML Project requires from us. Of course we cannot simply substitute the missing values by the mean/median of the entire column!

Thus, we stressed our brains and researched for techniques that were of equivalent level of performance and closer to the accuracy of the work submitted in the research papers provided to us. The Python library named Imputer, commonly used for substituting missing data with the mean/median of an entire column, led to us finally landing on the technique of MICE Imputation, and surely this high jump was equivalent to those employed by the basketball players about whom we are studying!

Learnings

- We believed that the predictability testing and training of the machine learning model/s would be of the most significant complexity level to perform but upon commencement of data preprocessing, we understood that first-hand dealing with datasets is also important and of much complexity given the fact that it determines the degree of useful analysis and therefore, the performance of the prediction.

- One of the key learnings included the lesson of how to deal with missing data without using the traditional simple mean/median methods but to go with imputing missing values by predicting them using other features from the dataset. Initially, we believed that the use of ML algorithms would be only in the training and testing phases and we did not even think about the fact that we can apply the iterative nature of these algorithms in the data preprocessing in itself.

Conclusion

A significant deal of data preprocessing has been conducted to make room for further training of the data. Deep observations and study, we believe, would be still required for dimensional reduction and then proceed towards the modeling phases of the project using Random Forest and XGBoost, both of which are based on the working of multiple decision trees. Our next target aim also includes to design a paper prototype of the dashboard in order to gain insights of the features whose data visualization is required to display.