Midterm Project

Supervised Learning - AY 2022-2023 - 15-27 November 2022

This project aims to analyze a dataset representing beach volleyball match scores. We want to build a model which will be able to predict the strength of the team based on the team player's individual strength. The provided dataset includes the following information:

- `Score`: Team Score calculated at the end of the match.
- `Player_Score_0`: Score of the player 0 before playing the match
- `Player_Score_1`: Score of the player 1 before playing the match
- `Player_Score_2`: Score of the player 2 before playing the match
- `Player_Score_3`: Score of the player 3 before playing the match
- `Player_Score_4`: Score of the player 4 before playing the match
- `Player_Score_5`: Score of the player 5 before playing the match
- `Player_Score_6`: Score of the player 6 before playing the match
- `Performance`: Performance descriptive rating from a random observer of the match
- `Country`: Country in which the match took place
- `Players_Injured`: Players Injured during the match

In order to build the desired predictive model, develop the following tasks and answer the following questions.

Questions and Tasks

1. Load and explore the dataset
   (a) How many numerical features are there? How many categorical features?
   (b) Verify if there are missing values in the dataset and handle them
   (c) Justify the choices you make for handling the missing values
2. Prepare the dataset for a Linear Regression task.
   (a) Verify the features values distribution of the numerical variables?

(b) Is features transformation necessary for the numerical variables? Let's take into account that we are preparing the dataset for a Linear Regression task, with the goal of building a "Score" predictive model. If transformation is necessary, after justifying your choices, do proceed as described.

(c) Verify the presence of outliers and eventually handle them. Justify your choices.

(d) Is encoding necessary for the categorical variables? If yes, which kind of encoding? Specify your choices, justify them and perform categorical data encoding, if necessary.

(e) Increase the dimensionality of the dataset introducing Polynomial Features – degree = 3 (continuous variables)

(f) Eventually include any other transformation which might be necessary/appropriate and justify your choices.

3. Features Selection

(a) Perform One Way ANOVA and test the relationship between variable `Country` and `Score`. Eventually, consider the possibility to remove the feature. Justify your choice.

(b) Perform Features Selection and visualize the features which have been selected. Select one appropriate methodology for features selection and justify your choice.

4. Linear Regression

(a) Train a Multiple Linear Regression model, using the Sklearn implementation of Linear Regression to find the best $\boldsymbol{\theta}$ vector. Use all the transformed features, excluding the derived polynomial features. Evaluate the model with the learned $\boldsymbol{\theta}$ on the test set.

(b) Use all the transformed features, excluding the derived polynomial features, to identify the best values of $\boldsymbol{\theta}$ by means of a Batch Gradient Descent procedure. Identify the best values of $\boldsymbol{\eta}$ (starting with an initial value of $\boldsymbol{\eta}$ = 0.1). Evaluate the model with the trained $\boldsymbol{\theta}$ on the test set. Plot the train and the test error for increasing number of iterations of the Gradient Descent procedure (with the best value of $\boldsymbol{\eta}$). Provide a comment of the plot.

(c) Use the complete set of features, including the derived polynomial features. Train a Multiple Linear Regression model, using the Sklearn implementation of Linear Regression to find the best $\boldsymbol{\theta}$ vector. Evaluate the model with the learned $\boldsymbol{\theta}$ on the test set. Plot the train and the test error for increasing the size of the train-set (with the best value of $\boldsymbol{\eta}$). Provide a comment of the plot.

(d) Use the complete set of features, including the derived polynomial features. Train a Ridge Regression model identifying the best value of the learning rate $\boldsymbol{\alpha}$ that allows the model to achieve the best generalization performances. Evaluate the model.

(e) Use the complete set of features, including the derived polynomial features. Train a Linear Regression model with Lasso regularization. Comment on the importance of each feature given the related trained parameter value of the trained model. Also, verify the number of features selected (related coefficient $\boldsymbol{\theta}$ different from zero) with different values of α.

(f) Use the subset of features selected in the Feature Selection task (question *3b*). Train a Multiple Linear Regression model using the Sklearn implementation of Linear Regression to find the best $\boldsymbol{\theta}$ vector. Evaluate the model.

(g) Create a table with the evaluation results obtained from all the models above on both the train and test sets.

(h) Compare and discuss the results obtained above.