Project 1: Feature Engineering

CMPE 462, Machine Learning, Spring 2018

Instructor: Ali Taylan Cemgil
TA: Rıza Özçelik

Due: Thursday, March 8, 2018, 23:55

1 Description

Sergen really loves playing bets, for pleasure only, especially on soccer. Though he has been betting on single games for years, he is looking for new adventures now. He aims to predict league standings at the end of the season, even before the season starts. To do so, he has requested help from his friend, Ibrahim, to provide the necessary background expertise. Since Ibrahim is also interested in playing bets he already has a dataset of the past seasons of the Turkish Soccer League and he delivered it to Sergen.
Sergen is a man of investment science and heard what machine learning and artificial intelligence is capable of, thus he has hired you as a machine learning expert to analyze the dataset Ibrahim has provided. He expects you to extract the important features from the dataset and present them in a visual and well documented way to make him correctly guess the ranking in a season as accurate as possible.

2 Input & Output

You are already provided with a notebook that contains some functions. **Do not modify the last cell apart from the outfile name**. Instead make your code compatible[1]. This means you are obliged to use Linear Regression module that is already implemented in scikit-learn and your test performance will be evaluated by the metric provided. However, you are free to use any metric you desire during the feature selection part.
You are expected to modify the above-mentioned notebook that to display your work and submit it. Your notebook should contain but not limited to:

- visual representation(histogram, box plots etc.) of dataset to help understanding the underlying structures,
- features designed to express the teams' points at the end of the season. Note that there are no best features, you are expected to provide reasonable features and ideally also a discussion of the thought process and justification. Hence, provide comparisons of different feature sets with ups and downs,

[1] If that is not possible, contact rizaozcelik96@gmail.com

- explanation of the features that expresses why you have thought that this feature could be a sign of teams' success. Interpret if your reasoning is parallel with the model results,
- a final feature set that you think that best explains the teams' success with reasons that you selected them,
- student IDs of the team members, not the names.

3 Grading Criteria

Criterion	Value
Data Analysis & Visualization	25 pts
Creativity of the features	25 pts
Success of the final features	20 pts
Feature reasoning and comparisons	20 pts
Obeying project rules	10 pts

4 Submission Details

You are supposed to use the GitHub system provided to you for all projects. No other type of submission will be accepted. Also pay attention to the following points:

- Your notebook name should consists of your student IDs. Each team will consist of three members.
- Your are expected to use python and its libraries efficiently. To exemplify, do not iterate over a data frame cell by cell.
- Make sure your notebook contain necessary comments and explanations.