Statistical Learning Project
Technical Appendix

Davide Ghiotto 1236660 — Darko Ivanovski 1243085

January 2021

Introduction

This is the **Technical Appendix** for the *Statistical Learning Project*.
This document contains a list of theoretical concepts that were used inside the project along with all the model formulae.

1 Multiple Linear Regression

In our project we used extensively the *Multiple Linear Regression* instead of the *Simple Linear Regression* because we had a lot of predictors, each contributing to our model.
The generalization from *Simple* to *Multiple Linear Regression* is given by the following formula:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{12} + ... + \beta_p x_{ip} + \epsilon_i \tag{1}$$

In particular, if we use matrix notation, the model can be written as:

$$Y = X\beta + \epsilon \tag{2}$$

And the β coefficients are calculated minimizing the residual sum of squares:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

where y is the value of the response variable.

2 Adjusted R^2

The adjusted R^2, denoted by $\bar{R}^2$ is a modified version of R^2 that adjust for the number of explanatory terms in a model relative to the number of data points. Unlike R^2, $\bar{R}^2$ increases only when the increase in R^2 (due to the inclusion of a new explanatory variable) is more than one would expect to see by chance.
To compute $\bar{R}^2$ the following formula is used:

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right) \tag{4}$$

3 Backward Step-wise Selection

In order to apply variable selection in our models we used *Backward Step-wise Selection*.
It works as following:

- let M_p be our full model containing all p predictors
- For $k = p, p-1, ..., 1$:
  - Consider all k models that contains all but one of the predictors in M_k, for a total of $k-1$ predictors
  - Choose the best among these k models, and call it M_{k-1}
- Select a single best model from $M_0, ..., M_p$ using cross-validated prediction error, AIC, BIC, or Adjusted R^2

4 Linear Discriminant Analysis

As a comparison to our classification model we implemented a *Linear Discriminant Analysis (LDA)* algorithm to our training data and used to predict directly a categorical class.
The class assignment process of the algorithm works as follows:

$$\delta_j(x) = x\frac{\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2} + log(\pi_j) \tag{5}$$

where:

- $\delta_j(x)$: discriminant score for x to belong to class j
- $1 < j < |G|$
- G : the set of distinct classes that x can assume

Once that for every x all the $\delta_j(x)$ are computed, the highest is selected and that is the qualitative prediction for the data point x.

5 Methods Formulae

5.1 Model 1

$$\begin{aligned} Y_i = \beta_0 + \beta_1 x_{i,a} + \beta_2 x_{i,h} + \beta_3 x_{i,o} + \beta_4 x_{i,ce} + \sum_{l \in L} \beta_l x_{i,l} + \\ \beta_{10} x_{i,ga(y)} + \beta_{11} x_{i,go(y)} + \beta_{12} x_{i,mi(y)} + \beta_{13} x_{i,as(y)} + \\ \beta_{14} x_{i,y(y)} + \beta_{15} x_{i,o(y)} + \beta_{16} x_{i,r(y)} + \epsilon_i \end{aligned} \tag{6}$$

- Y_i : market value
- $x_{i,a}$: age
- $x_{i,h}$: height
- $x_{i,o}$: offensive (dummy variable)
- $x_{i,ce}$: contract expires
- $x_{i,l}$: current league (6-1 dummy variables), for $l \in L = \{Leagues\}$
- $x_{i,ga(y)}$: games 20/21
- $x_{i,go(y)}$: goals 20/21
- $x_{i,mi(y)}$: minutes 20/21
- $x_{i,as(y)}$: assists 20/21
- $x_{i,y(y)}$: yellow player 20/21 (dummy variable)
- $x_{i,o(y)}$: orange player 20/21 (dummy variable)
- $x_{i,r(y)}$: red player 20/21 (dummy variable)
- $(y) := 2021$.
- Parameters = 1+4+(6-1)+7 = 17

5.1.1 Model 1.1

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_{i,a} + \beta_2 x_{i,h} + \beta_3 x_{i,o} + \beta_4 x_{i,ce} + \sum_{l \in L} \beta_l x_{i,l} + \\ \beta_{10} x_{i,ga(y)} + \beta_{11} x_{i,go(y)} + \beta_{12} x_{i,mi(y)} + \beta_{13} x_{i,as(y)} + \\ \beta_{14} x_{i,y(y)} + \beta_{15} x_{i,o(y)} + \beta_{16} x_{i,r(y)} + \epsilon_i \end{aligned} \tag{7}$$

where here we just log-transformed the response variable Y_i to $log(Y_i)$.

5.2 Model 2

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_{i,a} + \beta_2 x_{i,h} + \beta_3 x_{i,o} + \beta_4 x_{i,ce} + \sum_{l \in L} \beta_l x_{i,l} + \\ \sum_{y \in Y} [\ \beta_j x_{i,ga(y)} + \beta_{j+1} x_{i,go(y)} + \beta_{j+2} x_{i,mi(y)} + \beta_{j+3} x_{i,as(y)} + \\ \beta_{j+4} x_{i,y(y)} + \beta_{j+5} x_{i,o(y)} + \beta_{j+6} x_{i,r(y)}\ ] + \epsilon_i \end{aligned} \tag{8}$$

- $y \in Y = \{2021, 2020\}$
- $j = (2021 - y) \times 7 + 10$ (counter for the β inside the second sum)
- Parameters = 1+4+(6-1)+7x2 = 24

5.3 Model 3

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_{i,a} + \beta_2 x_{i,h} + \beta_3 x_{i,o} + \beta_4 x_{i,ce} + \sum_{l \in L} \beta_l x_{i,l} + \\ \sum_{y \in Y} [\ \beta_j x_{i,ga(y)} + \beta_{j+1} x_{i,go(y)} + \beta_{j+2} x_{i,mi(y)} + \beta_{j+3} x_{i,as(y)} + \\ \beta_{j+4} x_{i,y(y)} + \beta_{j+5} x_{i,o(y)} + \beta_{j+6} x_{i,r(y)}\ ] + \epsilon_i \end{aligned} \tag{9}$$

- $y \in Y = \{2021, 2020, 2019\}$
- $j = (2021 - y) \times 7 + 10$ (counter for the β inside the second sum)
- Parameters =1+4+(6-1)+7x3 = 31

5.4 Model 4

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_{i,a} + \beta_2 x_{i,h} + \beta_3 x_{i,o} + \beta_4 x_{i,ce} + \sum_{l \in L} \beta_l x_{i,l} + \\ \sum_{y \in Y} [\ \beta_j x_{i,ga(y)} + \beta_{j+1} x_{i,go(y)} + \beta_{j+2} x_{i,mi(y)} + \beta_{j+3} x_{i,as(y)} + \\ \beta_{j+4} x_{i,y(y)} + \beta_{j+5} x_{i,o(y)} + \beta_{j+6} x_{i,r(y)}\ ] + \epsilon_i \end{aligned} \tag{10}$$

- $y \in Y = \{2021, 2020, 2019, 2018\}$
- $j = (2021 - y) \times 7 + 10$ (counter for the β inside the second sum)
- Parameters = 1+4+(6-1)+7x4= 38

5.5 Model 5

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_a + \beta_2 x_h + \beta_3 x_o + \beta_4 x_{ce} + \sum_{l \in L} \beta_l x_l + \\ \beta_{10} x_{ga} + \beta_{11} x_{go} + \beta_{12} x_{mi} + \beta_{13} x_{as} + \\ \beta_{14} x_y + \beta_{15} x_o + \beta_{16} x_r + \epsilon_i \end{aligned} \tag{11}$$

- $x_{i,a}$: age
- $x_{i,h}$: height
- $x_{i,o}$: offensive (dummy variable)
- $x_{i,ce}$: contract expires
- $x_{i,l}$: current league (6-1 dummy variables), for $l \in L = \{Leagues\}$
- $x_{i,ga}$: total games
- $x_{i,go}$: total goals
- $x_{i,mi}$: total minutes
- $x_{i,as}$: total assists
- $x_{i,y}$: total yellows
- $x_{i,o}$: total oranges
- $x_{i,r}$: total reds
- Parameters = 1+4+(6-1)+7 = 17

5.6 Final Reduced Model

$$\begin{aligned} log(Y_i) = \beta_0 + \beta_1 x_a + \beta_2 x_h + \beta_3 x_o + \beta_4 x_{ce} + \sum_{l \in L} \beta_l x_l + \\ \beta_{10} x_{ga} + \beta_{11} x_{go} + \beta_{12} x_{as} + \beta_{13} x_y + \epsilon_i \end{aligned} \tag{12}$$

- Parameters = 1+4+(6-1)+4 = 14