Streamline Model Development - Demo

Lizzy Huang

2018-11-08

Contents

Chapter 1

Introduction

This is a demo about how to streamline model development process with synchronized documentation using **R Markdown**. Here we only show a sample process where we have a general logistic regression model that will apply to 2 different subgroups of the data, corresponding to 2 different "products". The ability of **bookdown** using separate `.Rmd` files for different sections of the documentation provides more flexibility for collaboration.

Chapter 2

Data

Assuming that part of the data extraction part has been completed, we start directly with the data analysis part.

In this example, we use the data sets from the **Kaggle European Soccer Database** to estimate the likelihood of winning, losing and getting a draw of soccer games in the Premier League based on several related variables. There are in total 608 games in 8 seasons, with 304 home games and 304 away games. The data set has been pre-processed for this demo.

2.1 Variables Summary

In this model, we use the following variables extracted from the **match events**

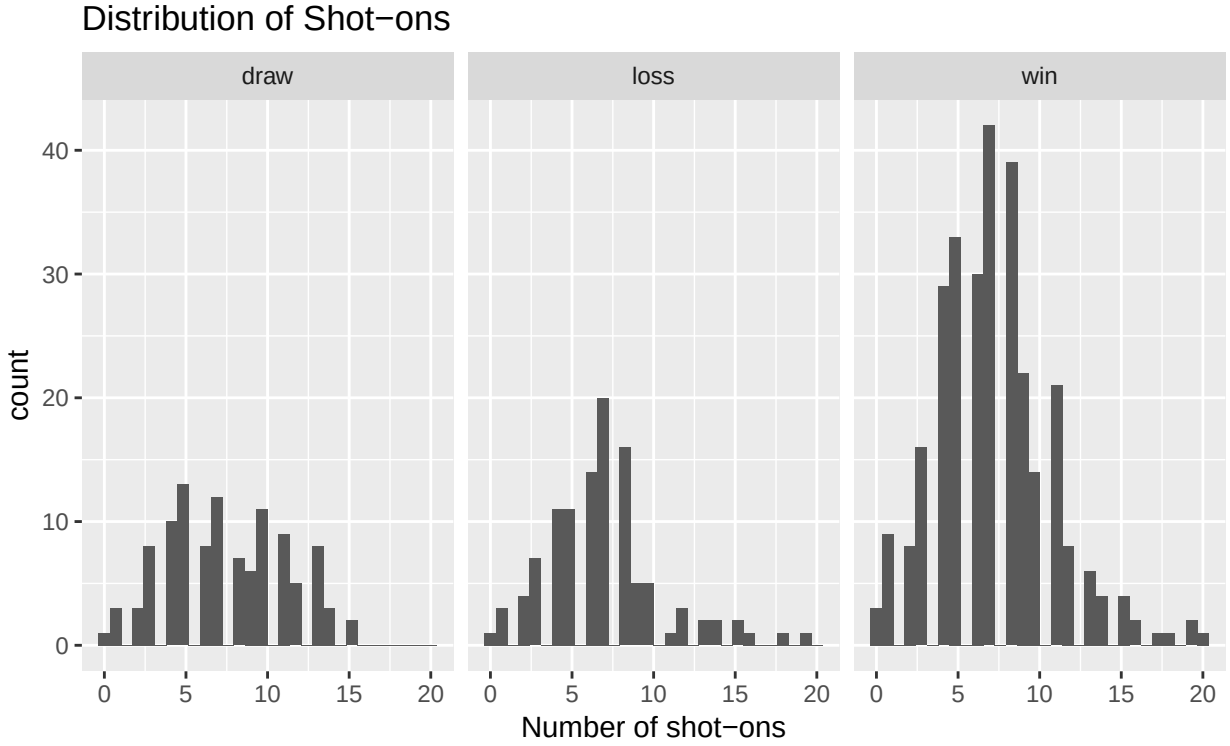
Variables	Definitions
`match id`	unique ID for every match
`team_long_name`	the name of the team
`season`	match season, from 2008/2009 to 2015/2016
`team_goal`	the number of goals the team scored in the match
`opponent_goal`	the number of goals the opponent team scored in the match
`game_type`	identify whether it's a home game or an away game
`result`	match results, either win, loss, or draw
`team_foul`, `opponent foul`	the number of fouls the team / opponent team committed in the match
`team_rcard`, `team_ycard`, `opponent_rcard`, `opponent_ycard`	the numbers of red/yellow cards that the team/opponent team received
`team_cross`, `opponent_cross`	the number of crosses of the team / opponent team made
`team_corner`, `opponent_corner`	the number of corners the team / opponent team received
`team_shotoff`, `opponent_shotoff`	the number of shot-offs the team / opponent team made
`team_shoton`, `opponent_shoton`	the number of shot-ons the team / opponent team made
`team_pos`, `opponent_pos`	the percentage of possession the team / opponent team had. The two adds up to be 100

2.2 Training Data and Testing Data

We select seasons 2008/2009 - 2014/2015 as the training data, and the last season 2015/2016 as the testing data. We exclude observations which have missing values in one or more columns. This leaves us 514 games in the training set and 72 games in the testing set.

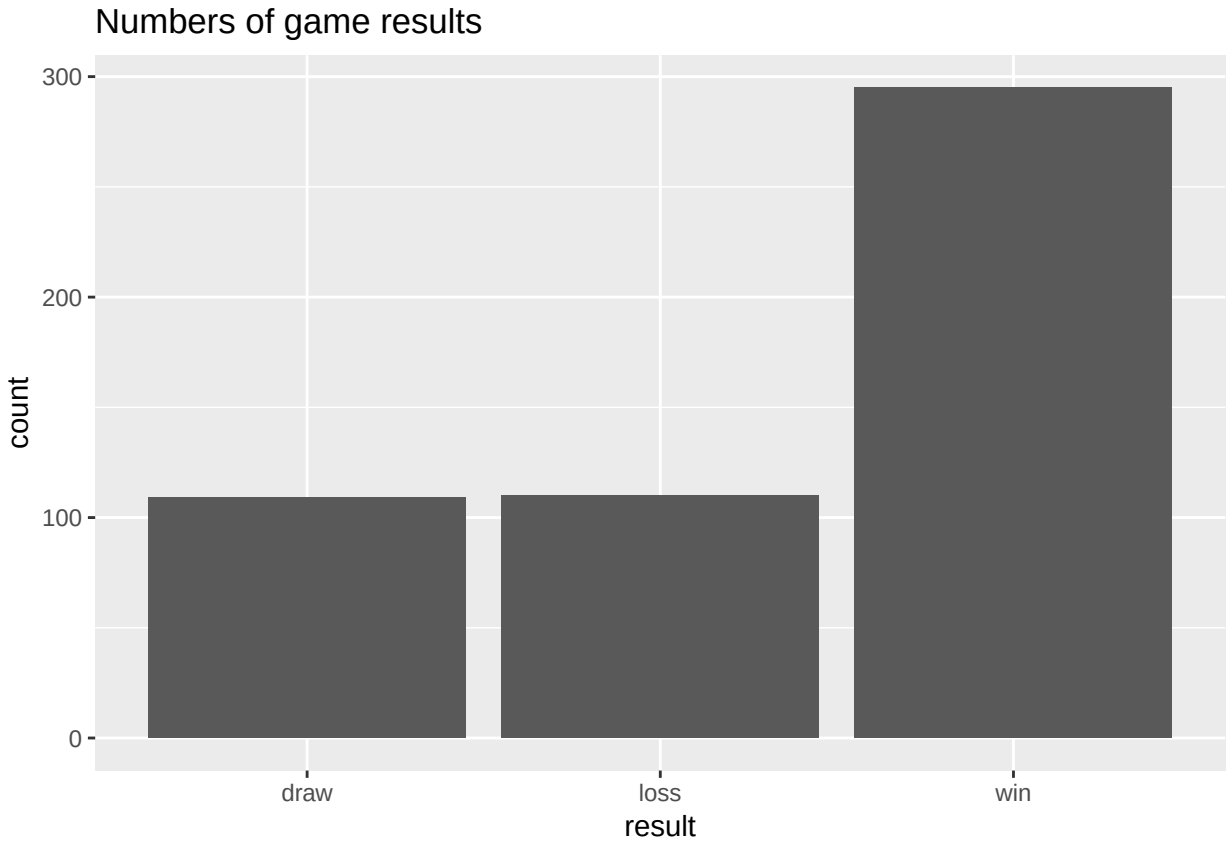
2.2.1 Input Variable Analysis

Here, we summarise the distributions of some of the input variables from the training data set, say, the distribution of `team_shoton` based on the results.



2.2.2 Output Variable - `result`

We show the counts of the 3 results: draw, loss, win.


Numbers of game results
300
200
100
0
count
draw
loss
win
result

Chapter 3

Model

3.1 Methodology

We use the multinomial logistic regression model for estimation. The response variable `result` has 3 levels: win, loss, and draw. We use

- `game_type`,
- `team_shoton`, `opponent_shoton`, `team_shotoff`, `opponent_shotoff`,
- `team_corner`, `opponent_corner`,
- `team_cross`, `opponent_cross`,
- `team_pos`, `opponent_pos`

as the explanatory variables to predict the results.

To avoid overfitting, we apply the penalized LASSO model with hyperparameter `lambda` (λ).

3.2 Case 1: Manchester United

In this section, we focus on building the estimation model for **Manchester United**.

3.2.1 Model Result

The 10-fold cross validation from LASSO algorithm shows that

$$\text{score(draw)} = 0.0086 * \text{away game}$$

$$\text{score(loss)} = 0.0052 * \text{team crosses} - 0.0090 * \text{opponent crosses}$$

$$\text{score(win)} = -0.040 * \text{team fouls} - 0.0046 * \text{team crosses} + 0.036 * \text{team possession} - 0.5123 * \text{away game}$$

Using penalized hyperparameter $\lambda = 0.07$, we get the that the model is 51.4% correct.

3.3 Case 2: Liverpool

In this section, we focus on building the estimation model for **Liverpool**.

3.3.1 Model Result

The 10-fold cross validation from LASSO algorithm shows that

$$\text{score(draw)} = 0$$

$$\text{score(loss)} = -0.001 * \text{team cross} + 0.254 * \text{away game}$$

$$\text{score(win)} = -0.017 * \text{team cross} + 0.018 * \text{team possession}$$

Using penalized hyperparameter $\lambda = 0.0967407$, we get the that the model is 51.4% correct.