Project Report: the *football_player_sale* network

Gianluca Di Tuccio and Lorenzo Orsini

In the domain of the football transfer market, it usually happens that a club receives offers for one or more players: in these situations, the management decides if it is better to sell the player according to various criteria. For example, a club that has financial problems will be more inclined to sell the player concerning a club that has a stable situation. Another factor that influences the decision is the overall value of the player, which itself depends on many other factors.
In this project, we have built the Bayesian Network football_player_sale that suggests whether it is convenient for a club to accept an offer for its player, according to some criteria that we'll define in the next paragraphs. The description of the Network and the CPDs are explained respectively in Chapters 1 and 2 and also in the program ***get_value_from_dataset***. In addition, we computed the probability of some queries (can be interactively selected by the user) with the variable elimination technique and also with the likelihood weighting: for 100000 samples generated, we have reported the best value found (the one with the lowest error) and the number of samples required for this value; all these computations are described in Chapter 3. Finally, we have shown some results in Chapter 4. All the code developed is available at https://github.com/DitucSpa/BayesianNetworkProject.git.

1. Our Bayesian Network: *football_player_sale*

We have created the Bayesian Network illustrated in *Figure 1*, which is called ***footbal_player_sale.*** To develop this network we have used the pgmpy library on Python. First of all, we need to define the variables of this network. As we said before, the network has been implemented to suggest the sale of a football player: for this reason, we have introduced the variable ***Sell***, whose Boolean values are "*yes*" or "*no*". Now, we need to show which variables the node depends on. To simplify the network, we have only considered two variables: ***club_economic_situation*** and ***player_alternative***. The node club_economic_situation is a Boolean variable with two values "*stable*" and "*debts*": intuitively, a club that needs to pay off its debts will be more inclined to sell a player than one with a stable economic situation. In addition, the node player_alternative takes into account if the replacement for the player comes from the market (and so the club needs to invest the money earned with the offer) or from the youth academy (for free): this variable has indeed three values, *"market"*, "*youth_academy*" and *"not_exist"* (to take in account if there isn't an alternative for the player). Then, we need to define also which variables the node player_alternative depends on. We have assumed that it depends on the market and youth academy alternative: so we have introduced the two Boolean variables ***market_alternative*** and ***youth_academy_alternative*** with the values "*yes*" or "*no*", to state if the respective alternative is available. These two variables are also influenced by the overall of the player because it is more difficult to find a replacement for a top player than a bad one. For this reason, we introduce the node ***player_overall***, with the three values "*top_player*", "*normal_player*" and "*bad_player*". In addition, we consider that only the age of the player and his performance in the last year influence it. So, we have the node ***age***, with the three values "*young*" (under 23), "*mature*" (24 - 30 years old) and *"old"* (over 31 years old), and the node ***last_year_performance***, with the two values "*above_expectations*" and "*under_expectations*".

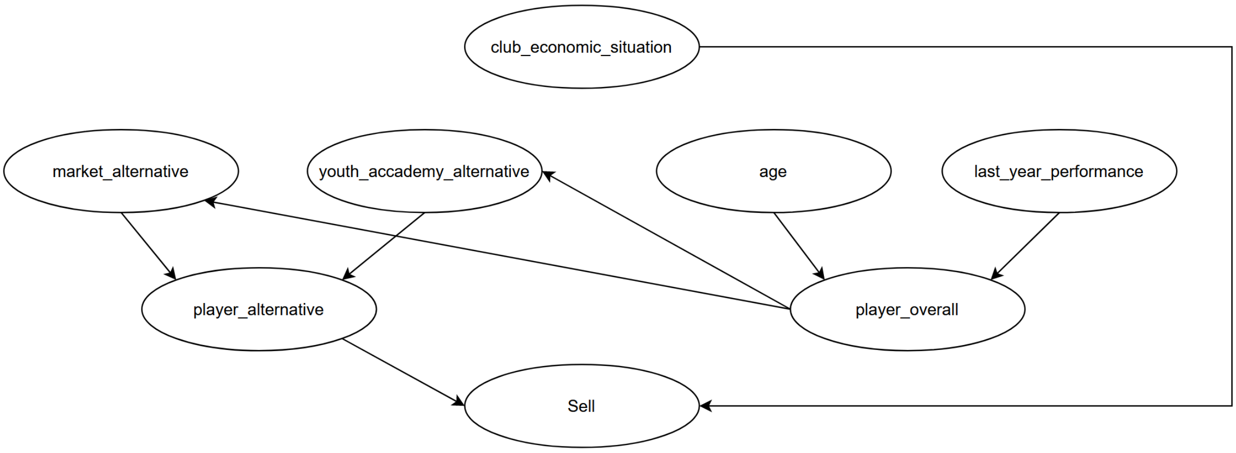
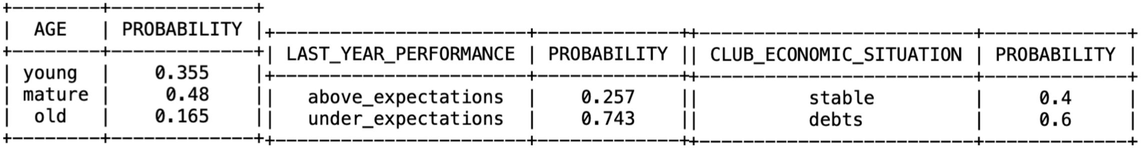


Figure 1: the football_player_sale network

2. CPDs

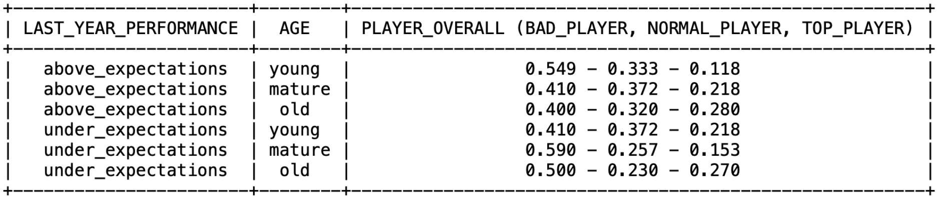
Regarding the CPDs, we have used the information available from the Serie A clubs for the 2019/20 season. Starting from the variable ***club_economic_situation***, we have looked at the balances reported from each club's website [1]: the results show that 40% of the clubs have stable economic situations, while the others have many debts (*figure 2.1*). Then, we have looked to the dataset available on the website *fbref* [2] to see the ages of all the Serie A players and their performances during the 2019/20 season, to obtain the CPDs of the variables ***age*** and ***last_year_performance***, which are shown in *figure 2.1* and in the program ***get_value_from_dataset***. Simply, for the performance we evaluate the number of presences, goals (or clean sheet) and assists of the previous year, using the age groups filter.



```
+--------+-------------+
|  AGE   | PROBABILITY |+-----------------------+-------------++--------------------------+-------------+
+--------+-------------+| LAST_YEAR_PERFORMANCE | PROBABILITY || CLUB_ECONOMIC_SITUATION  | PROBABILITY |
| young  |    0.355    |+-----------------------+-------------++--------------------------+-------------+
| mature |     0.48    ||   above_expectations  |    0.257    ||          stable          |     0.4     |
|  old   |    0.165    ||   under_expectations  |    0.743    ||          debts           |     0.6     |
+--------+-------------++-----------------------+-------------++--------------------------+-------------+
```

Figure 2.1: CPDs of the variables age, last_year_performance and club_economic_situation

Regarding the ***player_overall*** variable (block 3 of *get_value_from_dataset)*, we labelled as a "*top*" each player whose transfermarkt value was above 50 million euros, "*normal*" whose value was between 10-50 million euros and "*bad*" if his value was under 10 million euros (*figure 2.2*):



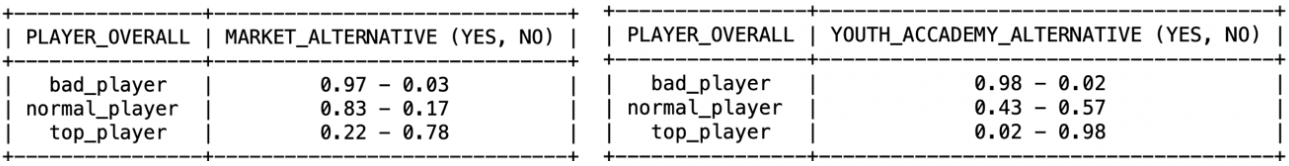
```
+-----------------------+--------+--------------------------------------------------------+
| LAST_YEAR_PERFORMANCE |  AGE   | PLAYER_OVERALL (BAD_PLAYER, NORMAL_PLAYER, TOP_PLAYER) |
+-----------------------+--------+--------------------------------------------------------+
|   above_expectations  | young  |                 0.549 - 0.333 - 0.118                  |
|   above_expectations  | mature |                 0.410 - 0.372 - 0.218                  |
|   above_expectations  |  old   |                 0.400 - 0.320 - 0.280                  |
|   under_expectations  | young  |                 0.410 - 0.372 - 0.218                  |
|   under_expectations  | mature |                 0.590 - 0.257 - 0.153                  |
|   under_expectations  |  old   |                 0.500 - 0.230 - 0.270                  |
+-----------------------+--------+--------------------------------------------------------+
```

Figure 2.2: CPD of the variable player_overall

Then, we have looked also at the ***market_alternative*** of each player, considering "*yes*" when there exists at least a player with a value in the same range and the same role as the one considered; we have applied the same procedure looking also to the youth academy of each player's club to fill the CPD of the variable ***youth_academy_alternative***. The CPDs of these three variables are shown in *figure 2.3:*

[1] An example of economic balance, AC Milan, https://www.acmilan.com/en/club/financial-report

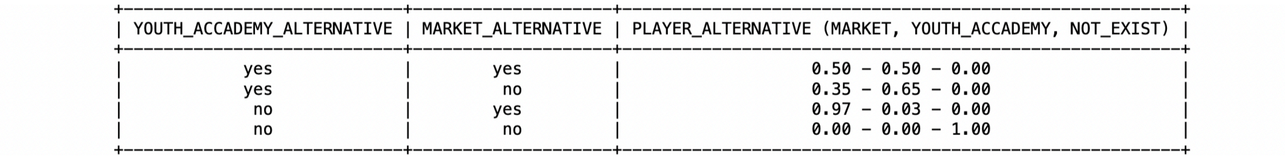
[2] fbref, dasat with players statistics for season 2019/20, https://fbref.com/en/comps/11/3260/stats/2019-2020-Serie-A-Stats#all_stats_standard



```
+----------------+------------------------------+   +----------------+-----------------------------------------+
| PLAYER_OVERALL | MARKET_ALTERNATIVE (YES, NO) |   | PLAYER_OVERALL | YOUTH_ACCADEMY_ALTERNATIVE (YES, NO)    |
+----------------+------------------------------+   +----------------+-----------------------------------------+
|   bad_player   |         0.97 - 0.03          |   |   bad_player   |              0.98 - 0.02                |
| normal_player  |         0.83 - 0.17          |   | normal_player  |              0.43 - 0.57                |
|   top_player   |         0.22 - 0.78          |   |   top_player   |              0.02 - 0.98                |
+----------------+------------------------------+   +----------------+-----------------------------------------+
```

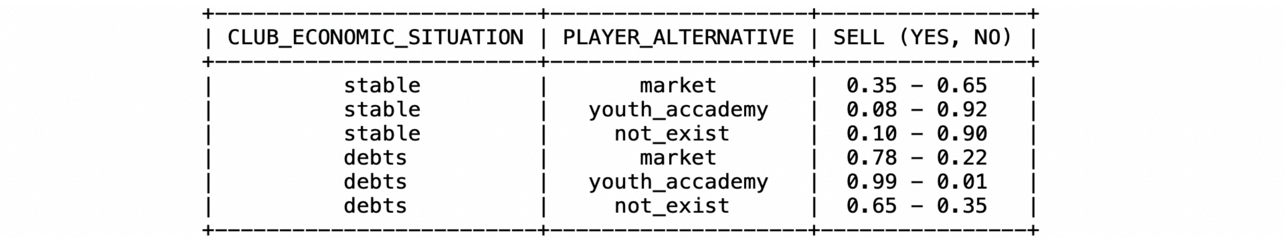
Figure 2.3: CPDs of the variables market_alternative and youth_academy_alternative

Regarding ***player_alternative***, for each player, we have merged the results of the variables market_alternative and youth_academy_alternative (*figure 2.4*). Then for the variable ***Sell***, we have intuitively fixed the probability values (*figure 2.5*). On one hand, we have considered that a club with a difficult economic situation will be prone to sell a player if there is an alternative from the youth academy (which is free), as in the case where a market alternative is available (but with a less probability, because in this case the club has also to spend the money of the sell). On the other hand, when the economic situation of a club is stable, it's not a good idea to sell a player and replace it with one from the youth academy (due to his age); the same reasoning explains also the probability values in the case of a market alternative (the alternative has always the same overall, but the age or the performance in the last year can be worse than the actual player).



```
+----------------------------+--------------------+------------------------------------------------------+
| YOUTH_ACCADEMY_ALTERNATIVE | MARKET_ALTERNATIVE | PLAYER_ALTERNATIVE (MARKET, YOUTH_ACCADEMY, NOT_EXIST) |
+----------------------------+--------------------+------------------------------------------------------+
|            yes             |        yes         |                 0.50 - 0.50 - 0.00                   |
|            yes             |         no         |                 0.35 - 0.65 - 0.00                   |
|             no             |        yes         |                 0.97 - 0.03 - 0.00                   |
|             no             |         no         |                 0.00 - 0.00 - 1.00                   |
+----------------------------+--------------------+------------------------------------------------------+
```

Figure 2.4: CPD of the variable player_alterative.



```
+-------------------------+--------------------+-----------------+
| CLUB_ECONOMIC_SITUATION | PLAYER_ALTERNATIVE | SELL (YES, NO)  |
+-------------------------+--------------------+-----------------+
|         stable          |       market       |  0.35 - 0.65    |
|         stable          |   youth_accademy   |  0.08 - 0.92    |
|         stable          |     not_exist      |  0.10 - 0.90    |
|         debts           |       market       |  0.78 - 0.22    |
|         debts           |   youth_accademy   |  0.99 - 0.01    |
|         debts           |     not_exist      |  0.65 - 0.35    |
+-------------------------+--------------------+-----------------+
```

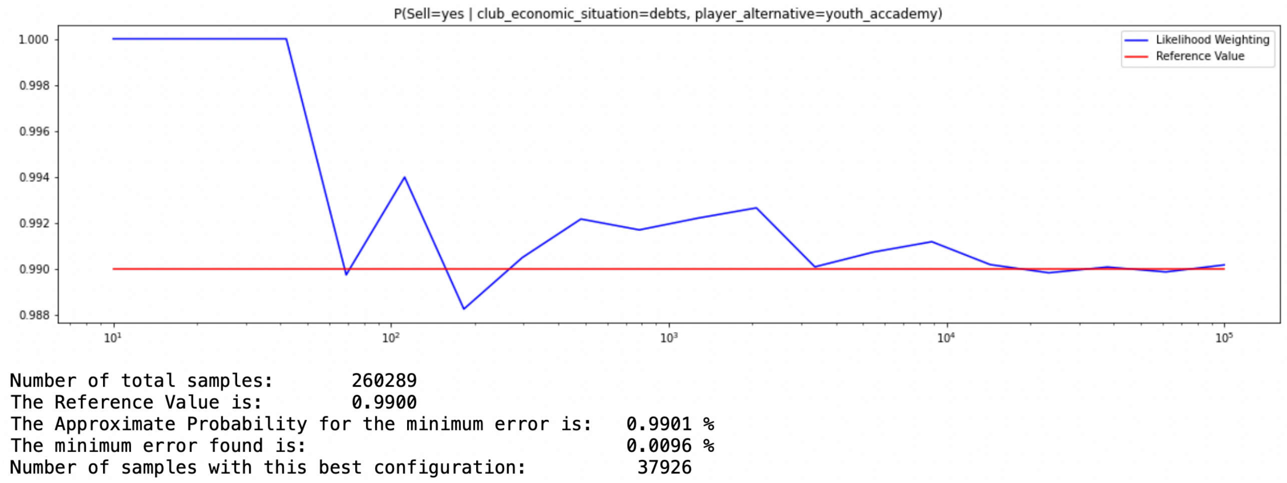
Figure 2.5: CPD for the final variable, Sell.

3. Variable elimination and Likelihood Weighting

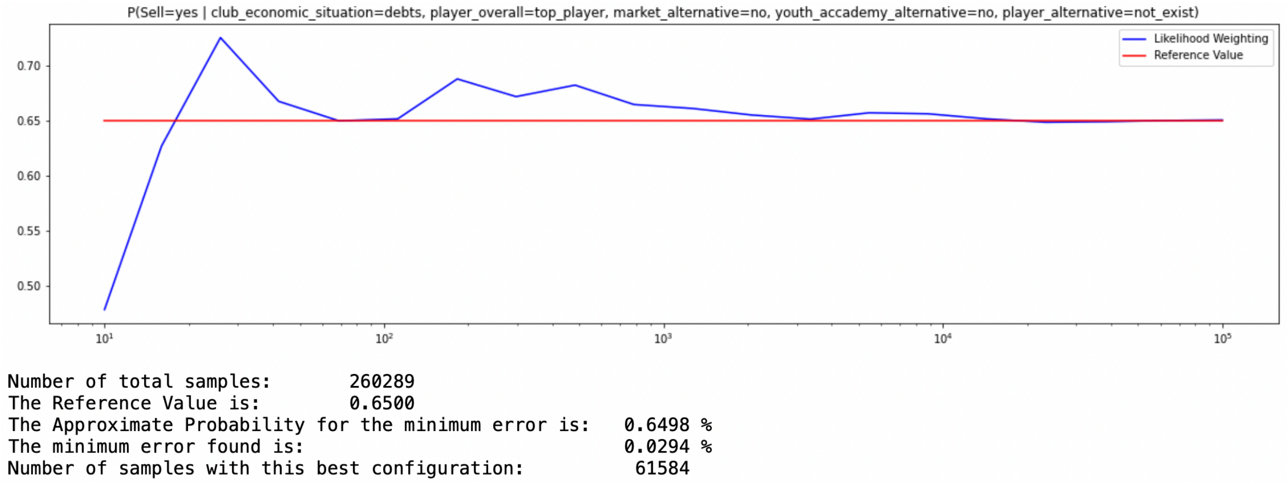
Now we have tested the network with some queries selected by the user. For each query is first calculated the probability with the technique of the variable elimination and then the program starts a loop with 20 iterations. For each iteration we generate samples with the function "logspace", which returns numbers spaced evenly on a log scale. So the function starts by 10 samples and with 20 iterations generates 100000 samples; at each iteration the probability is computed with the likelihood weighting technique, by using also the samples of the previous iterations. At the end of the loop are returned both the probabilities (likelihood and variable elimination) and the total number of samples generated. In block 6 of the code (available at https://github.com/DitucSpa/BayesianNetworkProject.git) is reported a more detailed explanation of the tools (dictionaries, loops and functions) used for this part.

4. Conclusion

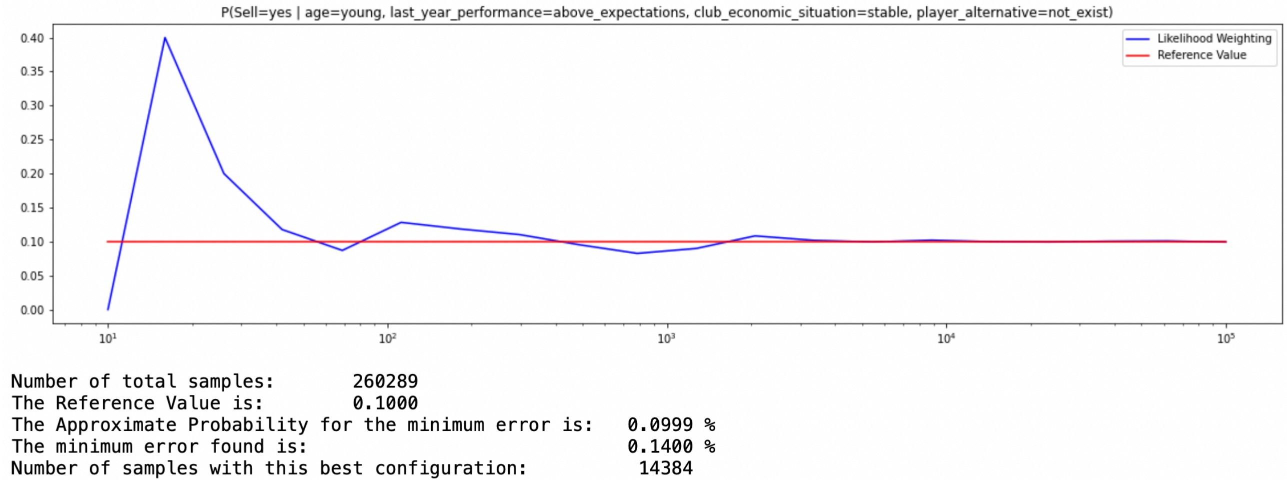
We have tested the program for several queries, and the results are shown in the following images:



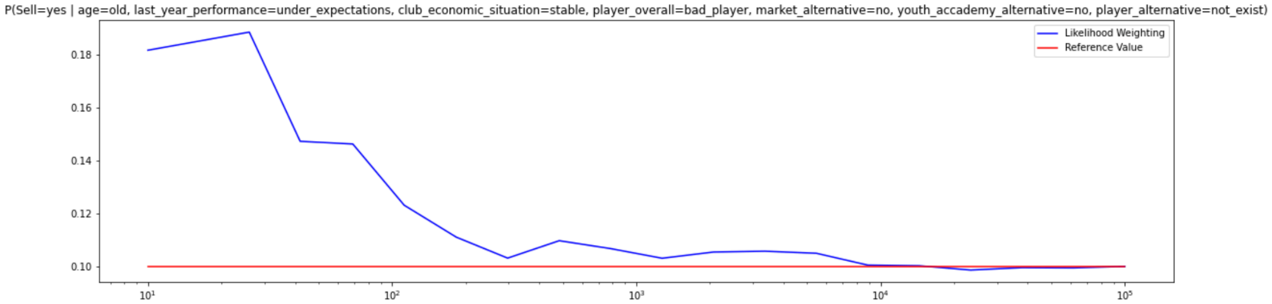
```
Number of total samples:       260289
The Reference Value is:        0.9900
The Approximate Probability for the minimum error is:   0.9901 %
The minimum error found is:                             0.0096 %
Number of samples with this best configuration:          37926
```



```
Number of total samples:       260289
The Reference Value is:        0.6500
The Approximate Probability for the minimum error is:   0.6498 %
The minimum error found is:                             0.0294 %
Number of samples with this best configuration:          61584
```



```
Number of total samples:       260289
The Reference Value is:        0.1000
The Approximate Probability for the minimum error is:   0.0999 %
The minimum error found is:                             0.1400 %
Number of samples with this best configuration:          14384
```



```
Number of total samples:       260289
The Reference Value is:        0.1000
The Approximate Probability for the minimum error is:   0.1000 %
The minimum error found is:                             0.0388 %
Number of samples with this best configuration:         100000
```