Machine Learning Engineer Nanodegree
Capstone Proposal - Winning Cartola with Machine Learning

Artur Pereira March 14th, 2017

Proposal

Domain Background

CartolaFC[1] is the name of the biggest Fantasy Football league in Brazil, where participants choose 11 players plus a coach for each round, given a limited amount of virtual cash, and score based on each player's performance for each round of the Brazilian National League, the Brasileirão. The league has been growing in popularity every year, for round 10 of the 2017 season, there were a total of 5,540,835 teams picked[2]. The usage of Machine Learning in the field of Fantasy Sports Leagues has grown in popularity over the last few years, for example, Kaggle has a yearly competition for the NCAA March Madness[3][4], which includes cash prizes and several participating teams. However, there have been few attempts at using Machine Learning for the CartolaFC League, namely, research showed only two instances, André Sakata attempted to use a Linear Regression model[5], while Arnaldo Gualberto used a Neural Network[6]. However, it seems both attempts were incomplete and did not seem to produce substantial results, nonetheless, both showed potential and that there was a lot of room for improvement. Lutz (2015)[7] makes an attempt at the Fantasy League for the NFL, using both SVR and Neural Networks, although for a different sport, and only for the Quarterback position, the approach and results show that in a Fantasy League setting, Machine Learning techniques show promising results, especially for Neural Networks. Matthews, Ramchurn, Chalkiadakis (2012)[8] used a Bayesian Reinforcement Learning Algorithm to compete on the Premier League’s Fantasy Competition, the Fantasy Premier League. Although the approach for this project is different, their success in ranking among top players of the Fantasy League shows the potential of Machine Learning techniques in these scenarios.

Given the popularity of the League, the few attempts of using Machine Learning for it, and the demonstrated relative success of its usage in other Fantasy League settings, the challenge of getting a good result with the tools available in Machine Learning seems like a worthwhile and interesting problem to tackle. This project then sets forth in the attempt to beat the best player for the 2017 season, who scored an average of 67.4 points per round, with the usage of Machine Learning techniques. The data source used for this project was obtained from the CartolaFCDados Github repository[9], which includes data for the Fantasy League from 2014 to 2017.

Problem Statement

Given the goal of this project is to get the maximum amount of points possible for each round, the model will attempt to predict the scores for each participant player in a round and make a selection of the top predicted scoring players for each position for the 2017 season, which also implies that this is a Regression Problem. The player scores for each round of the CartolaFC are a result of a sum of the points achieved for specific criteria, namely statistics such as goals scored, yellow and red cards received, missed passes, and scoring attempts. Each variable is associated with a specific amount of points awarded, positive or negative. Thus, the best fitting model will be the one that most accurately predicts the scores for every participating player in the round, as trained on the previous seasons, based on the fact that the predicted top scorers likely perform well and result in good overall final scores.

Datasets and Inputs

The datasets used for the project are available through the CartolaFCDados Github repository[9]. The data is publicly available and obtained via a scraper[10]. The repository contains 6 spreadsheets for each year from 2014 to 2017, with the following data features:

Positions: Position ID, Position Name, Shorthand Name;
Status: Status ID, Status Name;
Teams: Team ID, Team Name, Shorthand Name, Name Code;
Matches: Match Id, Round, Home Team ID, Away Team ID, Home Score, Away Score, Result ('Home', 'Away', 'Tie')
Athletes: Athlete ID, Nickname, Team ID, Position ID;
Scouts: Round, Team ID, Athlete ID, Participated (True or False), Points, Average Points, Price, Average Price, Fouls Received, Missed Passes, Assists, Shots on Post, Saved Shots, Missed Shots, Goals, Offsides, Missed Penalties,

Stolen Balls, Fouls Conceded, Own Goals, Yellow Cards, Red Card, Game without being scored, Difficult Saves (Goalie), Penalty Saves (Goalie), Goals Conceded (Goalie).

The Scouts spreadsheet has a data entry for each athlete for each round of that season. This data will be used as the basis for the project, and the other tables will be used for referencing Team, Athlete, and Position names, and whether the game played for each athlete was 'home' or 'away'. Once all players that did not participate are removed, there are a total of 43,927 entries. 10,938 are from 2017, the testing set, and 32,989, from 2014-2016, constituting the training set.

The target variable, the Score/Points are a direct results from the specified scouts for the match with the following attributed weights[11]:

Defense Scouts:

Feature	Game without being scored	Penalty Saves*	Difficult Saves*	Stolen Balls	Own Goal	Red Card	Yellow Card	Goals Conceded*	Fouls Conceded
Points	5	7	3	1.7	-6	-5	-2	-2	-0.5

** Goalie-only features*

Offense Scouts:

Feature	Goal	Assist	Shot on Post	Saved Shot	Missed Shot	Foul Received	Missed Penalty	Offside	Missed Pass
Points	8	5	3.5	1	0.7	0.5	-3.5	-0.5	-0.3

Therefore the score for a player in a round is the summation of occurrences of these features times the different weights.

An average will be taken for the previous participated rounds (maximum of 3) of each of these features that influence the scoring, for each player. This is a way of avoiding look ahead bias, as the scouts for the present round, which calculate the player score, are not used in the prediction model for that round. In addition to these averages, using the Matches dataset, a ‘home/away’ feature will be added for each instance of the Scouts dataset. Initially, the following feature set will be used for the different Machine Learning Models: All of the previously mentioned average features, a Home/Away variable, and Average Points, for a total of 20 features. However, this feature set will vary according to the position and model, and feature engineering can change this selection even further. The following table is a description of the target variable for each position:

PositionID	**Count**	**Mean**	**Std. Dev.**	**Min**	**25%**	**50%**	**75%**	**Max**
1 (Goalie)	2992	3.45	5.65	-12.00	-1.00	3.40	7.70	33.00
2 (Right/Left Defender)	6344	3.17	4.09	-9.70	0.10	2.70	5.60	27.10
3 (Defender)	6438	2.80	3.88	-11.20	0.00	2.30	5.18	23.50
4 (Midfield)	15234	2.27	3.77	-10.80	-0.10	1.40	3.90	31.90
5 (Forward)	10024	2.63	4.52	-7.30	-0.20	1.10	4.30	32.70
6 (Coach)	2895	3.39	2.11	-1.62	1.77	3.15	4.85	10.92
Total	43927	2.71	4.10	-12.00	0.00	1.80	4.90	33.00

Showing that the variation between positions can be quite significant, therefore the idea of running models for each position seems reasonable.

As for the training/validation/testing sets, the following strategy will be used: as the objective of the project is to ultimately score well on the 2017 CartolaFC Season, the 2017 dataset will be used as the testing set, as specified earlier. For the separation into training and validation sets, a cross validation strategy will be used to partition the training data into K-Fold Splits.

Solution Statement

The solution to the problem will be the final average score obtained for the 2017 CartolaFC Season. The Machine Learning models used will predict the scores for each player for each round, given a set of pre-round available statistics, and from these predictions, the team positions will be filled with the highest scoring players, and the score obtained will be the sum of the effective scores for each chosen player for that round. This process, repeated for every round of the season, allows for a final total score and average, which will then be compared to the highest scoring player for the 2017 CartolaFC Season.

Benchmark Model

Unfortunately, given the few and limited attempts at using Machine Learning for this specific scenario, there is no readily available benchmark model, therefore, I will be implementing a Linear Regression Model with some minimal feature engineering of the available data as a way of benchmarking the performance of some other techniques such as the SVR Model and a Convolutional Neural Network implementation. Given the structure of the proposed problem, a different model will be made for each of the available positions, 6 in total. A standard version of the Linear Regression Benchmark Model will be used for each of the positions with the same set of feature variables. The benchmark model will make predictions of the players' score for each round, this can then be compared with the actual score obtained for each player, thus giving an error that can be compared to the other proposed models.

Given the results obtained from the benchmark models, the players with the highest predicted scores will be chosen for each position, following the proposed formation scheme. This will yield a score for each round and thus give a final season average score, which can also be compared to the results obtained from the other models and ultimately compared to the highest scoring player for the Fantasy League. Therefore, there are two kinds of benchmarks, one based on the prediction errors for the different models, benchmarked with the Linear Regression Models, and the second one is the final average total score for the actual CartolaFC season, which will be also benchmarked with the top scoring player.

Evaluation Metrics

The evaluation metric used for this project, as mentioned earlier, will be regarding a simple difference between the predicted player score and the actual score obtained for the round. Given the difference can be either positive or negative and what matters is how far from the true scores the predicted value is, models will be evaluated given the squared difference between the predicted and actual scores for each round.

An important distinction for this specific project is that the model scores will follow a second evaluation stage, in which the highest predicted scores are then selected and the actual scores received are then counted towards a season scoring average, which is ultimately the goal of the project. This final average score is then compared to the best scoring player as a means to see whether the Machine Learning models used for this project were able to have a good final score, and beat the best player. Given there is a fairly high likelihood of variance between the players chosen for each of the models used, it could be that a model which had a higher predictive error could potentially score higher than a model with lower errors. However, choosing a winning model with higher squared errors is a “hindsight” decision, as it takes advantage of information that is not available a priori, thus it could be that the best fitting model will not necessarily be the highest scoring model.

Project Design

Given the dataset described above, the Project will be realized through the following steps:

1. With the available dataset, there will be an investigation of missing data, and incoherent variable values. As shown by the data investigation from Gualberto’s project, values are accumulated for the 2015 season, and

thus need to be modified. Also, only a fraction of the players actually play for each round, and given the scores are only calculated and meaningful when the player participates in the game, only the participating players will be considered.

2. After this initial step, some formatting will be addressed. For example, for each round, the scouts available are mostly referring to that specific round, meaning they should not be known for that round's score prediction. Hence, some data formatting will need to be made to address this issue. Not all data is numeric, therefore some one-hot encoding will also be necessary.
3. Once the data is properly formatted and arranged, some initial data exploration will be made, in order to identify correlations between dependent and independent variables and some general understanding of the importance of each feature. For example, it is somewhat common knowledge that playing "home" or "away" tends to influence the team's performance, thus such variables will be included and analyzed.
4. Once the initial feature set is determined, a Linear Regression model will be trained and cross validated on the available data for 2014-2016 for each position, and tested on the 2017 data, acting as a baseline for future models. Given these results, the predicted highest scores will be selected and the actual scores will be added which will yield an average score for the 2017 season. This value will be compared to the 2017 season winner and will also serve as a benchmark for the other models.
5. Given this baseline benchmark implementation, I will proceed to attempt some further feature engineering and the implementation of the SVR model, modifying different hyper-parameters as appropriate for each position. The associated errors will be compared to the baseline.
6. A CNN will be implemented, also with some parameter tuning for each position. The associated errors will be analyzed and compared to the previously defined models. This implementation will also follow suit and arrange a score that can be compared to the benchmarks.
    a. The idea here initially is to work with a Multi Layer Perceptron Regressor (MLPRegressor), which also somewhat works as a back-up plan.
    b. For the CNN, the idea is to try out several different Architectures, namely varying somewhat along the following parameters: epochs ∈ {10, 50, 100, 1000}, hidden layers ∈ {10, 25, 50, 100}, activators ∈ {Sigmoid, Tanh, ReLU}
7. Given the results from the different models, I will be able to identify the best scoring model, both for minimizing the errors, and overall scoring, which, as discussed previously, could turn out to be different models, and figure out whether any model was able to beat the best scoring player of the CartolaFC 2017 season.

Sources:

(1) GloboEsporte. CartolaFC. http://globoesporte.globo.com/cartola-fc/

(2) GloboEsporte. CartolaFC. *Virou rotina! Cartola bate novo recorde de times na rodada #10: 5,5 milhões*. June 24, 2017. http://globoesporte.globo.com/cartola-fc/noticia/2017/06/virou-rotina-cartola-bate-novo-recorde-de-times-na-rodada-10-55-milhoes.html

(3) Kaggle. Google Cloud & NCAA® ML Competition 2018-Men's. *Apply Machine Learning to NCAA® March Madness®*. https://www.kaggle.com/c/mens-machine-learning-competition-2018

(4) Kaggle. Google Cloud & NCAA® ML Competition 2018-Women's. *Apply Machine Learning to NCAA® March Madness®*. https://www.kaggle.com/c/womens-machine-learning-competition-2018

(5) Sakata, André. *Aplicando machine learning no CartolaFC*. September 30, 2017. https://medium.com/@andresakata/aplicando-machine-learning-no-cartolafc-4ebb5aa0a531

(6) Gualberto, Arnaldo. Github. *Análise dos Dados*. Last Update: December 10, 2017. https://github.com/henriquepgomide/caRtola/blob/master/src/python/An%C3%A1lise%20dos%20Dados.ipynb

(7) Lutz, Roman. *Fantasy Football Prediction*. University of Massachusetts Amherst. Amherst, United States. May 26, 2015. https://pdfs.semanticscholar.org/2e04/f58fa272a6eea79d23113f6a4743f8f53ac0.pdf

(8) Matthews, Tim; Ramchurn, Sarvapali; Chalkiadakis, Georgios. Association for the Advancement of Artificial Intelligence (www.aaai.org). *Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains.* University of Southampton; Technical University of Crete. Southampton, United Kingdom; Crete, Greece. 2012. http://www.intelligence.tuc.gr/~gehalk/Papers/fantasyFootball2012cr.pdf

(9) T M, Vinícius. Github. *Cartola FC Dados.* Last Update: 12 January, 2018. https://github.com/thevtm/CartolaFCDados

(10) T M, Vinícius. Github. *CartolaFCScraper.* Last Update: 3 May, 2017. https://github.com/thevtm/CartolaFCScraper

(11) GloboEsporte. CartolaFC. *Entenda as pontuações do Cartola FC e defina como escalar a sua equipe.* July 7, 2017. http://globoesporte.globo.com/cartola-fc/noticia/2017/06/virou-rotina-cartola-bate-novo-recorde-de-times-na-rodada-10-55-milhoes.html