PREDICTING GOALS SCORED FOR FOOTBALL MATCHES

Arish Dutt (ad428642@Stanford.edu)

Introduction

Is it possible to accurately predict the number of goals each team will score in a soccer match?
This project attempts to do so using objective statistics collected regarding that team's performance in previous matches. The soccer betting market is estimated to be worth over $\$700bn$[3], hence being able to accurately predict a team's performance is a lucrative challenge. The label used is the number of goals a team scores in a match

Data

The dataset off which the labels and the final set of features were built are from ***football-data.co.uk***.[2] The data contains all results from the English Premier League (EPL) along with other match statistics. 5 season (years) of data was used

Features

All features used were derived from the data.
Features were based on the team for whom we are predicting for (attacking) and the team who they are playing (defending), 25 were used in the final set of models
Many other features were tested beyond those used in the final set of models, including player specific features.
Features that added minimal predictive power were removed, largely through from a version of backward search
Features were all scaled (and the label): for feature x_j

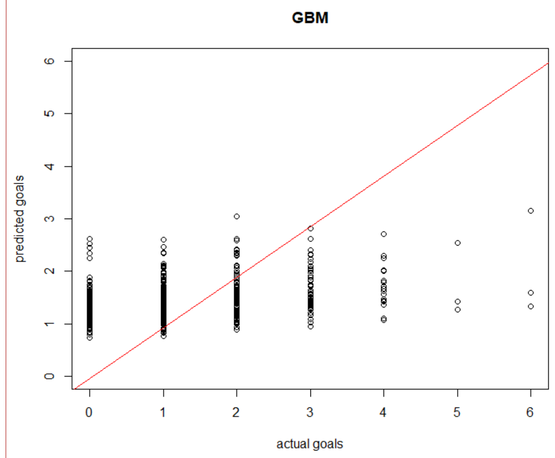
$$x_j^i = \frac{x_j^i - \min(x_j)}{\max(x_j) - \min(x_j)}$$

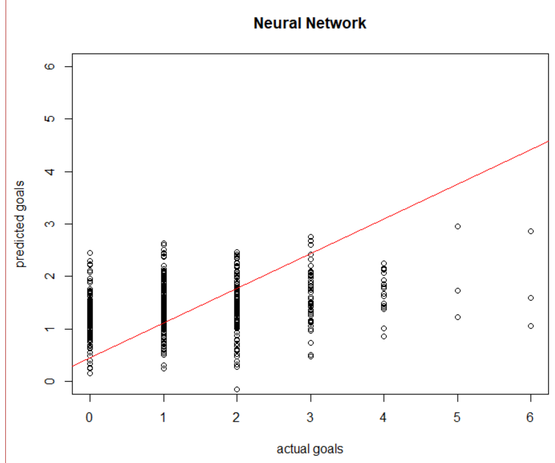
Results

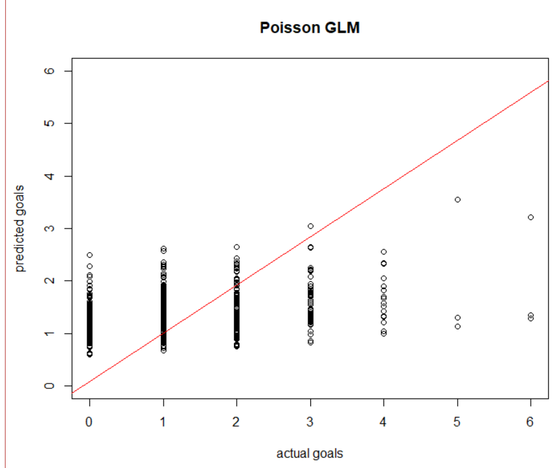
The split between training / validation / test set is 60 / 20 / 20, randomly chosen. The total number of observations is 3,800.
The error metric used is Root Mean Squared Error

Model	Train error	Test error
GBM	0.157	0.159
Neural Network	0.162	0.166
Poisson GLM	0.158	0.161

Predicted vs observed goals by model







Models

Gradient Boosted Machine
Algorithm that uses decision trees that sequentially learn from one another. Mathematically:[1]
$\hat{y} = \sum_{m=1}^{M} \gamma_m h_m(x)$ where $h_m(x)$ is the m-th decision tree. The model is built additively $\hat{y}_m = \hat{y}_{m-1} + \gamma_m h_m(x)$. The newly added tree $h_m(x)$ tries to minimize the loss L:

$$h_m = \arg\min_h \sum_{i=1n} L(y_i, F_{m-1}(x_i) + h_m(x_i))$$

Loss function used is Root Mean Squared Error

Neural Network
A single hidden layer with 2 neurons and with a $softplus$[4] activation function:

$$f(x) = \log(1 + e^x)$$

Poisson GLM with regularization
Regularization term: $\lambda(\frac{1-\alpha}{2}\left||\beta|\right|_2^2 + \alpha\left||\beta|\right|_1)$. Alpha was set to 0.5 and lambda was chosen automatically by 10 fold cross validation

Discussion

GBM performed the best, GLM close 2nd , neural network performed by far the worst.
Starting off with a wide variety of features, these were whittled down significantly to only the most significant ones.
The value of a neural network is to discover hidden, complex relationships between the features and labels. The severe feature reductions means it has far less opportunity to do so and this may explain its relatively poor performance.

Further we see that all models underperform when actual goals is >= 3. Features used cannot accurately predict this tail end of the distribution, suggests additional features are needed
The models also proved that home advantage was a real phenomenon, with it coming out as one of the strongest predictors in both the GLM and GBM

Future work

The approach used was to find features that were most predictive and remove those that did not add much value. This may have led to poor performance at the tail ends of the label distribution, if all the features had been kept the predictions could have improved. Doing so with a more complex neural network structure would be my first move if there was additional time
Also adding in more player specific features that could capture the nuances of the starting line ups of both the attacking and defending teams, e.g. difference in total Fantasy Soccer price between the 2 teams. This would hopefully do a better job at quantifying the difference in quality between the teams

Citations

1. Brownlee, J. (2019). *A Gentle Introduction to XGBoost for Applied Machine Learning*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
2. Historical football results and betting odds data. *http: //www.football-data.co.uk/data.php.* Accessed: 2019-05-10.
3. Sports Betting Dime. (2019). *The Size and Increase of the Global Sports Betting Market*. [online] Available at: https://www.sportsbettingdime.com/guides/finance/global-sports-betting-market/
4. Sefik Ilkin Serengil. (2019). *Softplus as a Neural Networks Activation Function - Sefik Ilkin Serengil*. [online] Available at: https://sefiks.com/2017/08/11/softplus-as-a-neural-networks-activation-function/ [Accessed 11 Jun. 2019].