Expected Turnovers in Soccer and Basketball

Team: Kevin Bernat, Alton Wiggers, Ruifan Wang **Project Mentor TA:** Alex

1) Abstract

For this project, we focused on “turnovers” in soccer and explored expected goals in basketball[1]. More specifically, we looked at individual moments in these sports and tried to use machine learning to determine whether the individual play would result in one team giving over possession of the ball to the other. This is important in sports for seeing how risky a team is being on offense if they are giving up higher than expected turnovers and seeing how accurately that risk can be assessed using machine learning. Our primary target contribution is to featurize existing data to accommodate for the notion of turnovers, as it was not an inherent feature of any dataset we were able to find. Then, using a collection of methods, we found how well turnovers can be expected using machine learning. We found that XGBoost was able to perform the best on this problem, although the feature set may still be too small for practical application on real world games.

2) Introduction

The problem we are trying to solve is looking at an individual play in sports and trying to say what the chance is that it will result in the player with the ball turning it over to the other team. Here we looked at plays for soccer and basketball, but the methodology could be applied to any sport with a notion of possession (hockey, rugby, American football, etcetera). These plays could be an individual shot, pass, or run down the field/court for example. A data point is one of these individual plays and will have a large number of features including play type, player positions, ball position, and time in the match. The goal of the algorithm is to output either what the result of the play will be or the chance that the play will be a turnover (depending on the algorithm).

This is important for understanding the intermediate play of a lot of sports. A stat that is largely tracked in soccer is “possession,” which is the percentage of the match an individual team held the ball. However, this stat does not do a very good job of representing the events of the match. It does not give a sense of whether a low possession was from one team simply holding the ball without attempting goals or from poor playmaking of the low possession team when they did have the ball. The hope of this work is to see how accurately turnovers can be expected in a match. If we can accurately predict turnovers, then we would be able to say how risky a team played the ball and whether or not it worked out for them. If a team ended up having a lot higher turnovers than expected, then this is due to poor playmaking and may be able to highlight how and why a team failed to create scoring opportunities. Likewise, if the opposing team had a lot fewer turnovers than expected, you may be able to better assess the performance of your defense outside of scoring opportunities.

[1] Expected goals in basketball, in this case, is defined as the probability of a shot being successful

3) Background

We are building upon the prior work detailed by Jayanth Nair [A]. His article explains the benefits of looking at expected goals for soccer analytics. The author created his own expected goals metric using various machine learning models, ultimately creating an ensemble model with his best performing three. The article ends with other metrics that can be explored such as Expected Goals Against (xGA) which gives an indication of the defensive strength of a particular team. This, however, was not explored at all in the article. We addressed this shortcoming by not looking specifically at Expected Goals Against, which would still limit the events of interest to shot events only. Instead we examined a different metric, Expected Turnovers, so that we can look at the complicated passing dynamics.

Another limitation of the article is that it only looked at the women's Statsbomb data and it worked with the Expected Goals metric which has now become commonplace in any serious soccer analytic work. This shortcoming was addressed by looking at the men's Statsbomb data and by looking at two different metrics: Expected Turnovers in soccer and Expected Goals in Basketball. Looking at passing events, particular turnovers, have not been explored nearly as much as Expected Goals. Since each soccer game only has about a few dozen shots, looking at passing and dribbling events which number in the hundreds explores a broader and more common interaction in soccer. Predicting baskets in basketball further expands the work of the article to a completely different sport where this metric is not quite as common as other basketball metrics.

A. "Expected Goals - How I combined learning data science with my soccer obsession"
    a. URL: https://medium.com/analytics-vidhya/expected-goals-how-i-combined-learning-data-science-with-my-soccer-obsession-f81d721432c7
    b. Code Link: https://github.com/j-v-n/Springboard/tree/master/capstone_projects/capstone1
    c. Relevance: We built upon this article and expanded the project changing from expected goals to expected turnovers which have an emphasis on the defense rather than offense which is often neglected.

4) Summary of Our Contributions

The purpose of this project is to explore how advanced metrics can be used to help further improve our understanding of the complex dynamics involved in both soccer and basketball. While Expected Goals have become a mainstay in any serious soccer analysis, more defensive metrics are often overlooked. This project is two-pronged: it will first analyze a new metric, Expected Turnovers (xT). This metric will give an indication of the defensive strength of a team. This allows us to explore the complex passing interactions that occur in any soccer game and gain some insight into how influential they can be in determining the outcomes of matches. The second prong will explore scoring baskets in basketball. Expected Goals are most commonly seen in soccer, so applying this metric to a completely different sport will also shed light on the complicated shooting mechanics in basketball.

5) Detailed Description of Contributions

Note: several figures are referenced throughout this and following sections. These figures can all be found at the end of the document.

For the soccer data, we looked at all matches given for La Liga, the highest level of Spanish soccer, in the StatsBomb open data [4]. Our goal was to featurize the dataset to accommodate analysis on expected turnovers, and then do that analysis with several different Machine Learning methods. In this dataset, there were 485 matches with a total of 528,421 events relevant to our analysis (those being passes and dribbles) with 118 columns. We chose to focus on these types of events as they would likely have the most meaningful analysis. Event types like “fouls” that always result in a turnover or “kickoff” that never do would not add meaningful information to our results. Of these events, 30,611 resulted in turnovers. Turnovers were not an inherent feature of the data, so we encoded it by ordering the events for each match and observing where the possession changed from one team to another. We had to remove certain types of events first to make sure passes and dribbles were considered the catalyst event for the turnover and not the failed “ball reception” event type, for example.

To featurize our data, we had to make sure all features were numerical, so we one-hot encoded all numerical features. There were many features only relevant to certain types of events such as “goalkeeper_outcome,” which tells us what the result of a goalkeeper’s block was. We dropped all these features that were not relevant to our chosen event types. There were also features such as “pass_outcome” that directly supplied what the result of the play was, so we dropped these features as well.

Originally we dropped many events so we had an equal number of plays that did and did not result in turnovers, however, this resulted in an overestimation of turnovers when looking at individual games.

One possible reason for this is that turnovers, in relation to the total number of pass events, are relatively rare events. By splitting the dataset into an equal number of events with and without turnovers would naturally make it predict for turnovers more often, even if in comparison to the total dataset it is a rare event. From these findings, we revised our preprocessing step so that the machine learning models were trained on the full dataset. One issue that emerged from this was that the logistic regression model had difficulty converging on the dataset with many zero values and many features. Many of the features involved looking at shot events in addition to the passing and dribbling events. Removing the shot events and reducing our overall pool of features (as detailed above) allowed for the logistic regression model to converge in a reasonable time.

We decided to train our preprocessed soccer data on four different machine learning models: KNN, Random Forest, Logistic Regression, and XGBoost. The main goal was to see which machine learning model would provide the best predictor of Expected Turnovers. KNN really struggled with this problem, most likely due to the complexity and varying relevance of the features, as well as the disparity in the positive and negative labels. This resulted in a very poor F1 score, as the model tended to predict an overabundance of negative labels. Random Forest was one of our better-performing models, having improved metrics from both KNN and logistic regression. The F1-score was still not as high as would likely be required to use this model for real work decision making but does predict an expected number of turnovers more consistent with actual games.

For our logistic regression model, we found that it was a fairly poor predictor of the Expected Turnovers even after proper convergence and hyperparameter finetuning. Expected Turnovers were often trending above the actual number of turnovers in most cases and the precision, recall, and F1 scores were still fairly poor. This led us to try using XGBoost which works well with weak learners. This resulted in a classification report where the precision, recall, and F1 scores all improved. Similar to Random Forest, the F1-score was still not as high as we would have hoped. However, it also resulted in Expected Turnovers that were closer in line with the actual turnover numbers (See fig: Comparing Turnovers For Logistic Regression and XGBoost).

Since XGBoost was the machine learning model that resulted in the best Expected Turnover metric, we explored how it performed on the World Cup 2018 dataset which was also part of the Statsbomb dataset. We looked at the World Cup Final match between France and Croatia and calculated the Expected Turnovers over the course of the match. Using the predicted probabilities for each event, we were able to get a clear progression of how the number of Expected Turnovers evolved over the course of the game. It is clear that France was able to keep the total number of turnovers lower throughout the game when compared to Croatia. While this could also be due to the fact that France attempted fewer passes than Croatia, it is also evident that Croatia had a lot more dangerous passes that could have led to turnovers in comparison to France (see fig: Expected Turnovers (xT) over time for the World Cup 2018 Final). Using XGBoost as the most successful model, we also took a look at the feature importance metric (see Fig: Feature Importance). The most important included pass end location, the duration and length of each pass, and the location of the player on the field, which follows from intuition: longer passes that are slower are more likely to be intercepted, and regions of the field where there are a lot of opposing players, like near the penalty area, would also lead to more turnover events.

We initially wanted to compare basketball turnovers vs soccer turnovers. However, since NBA's official data is not open to the public anymore, we only had access to the limited movement data provided in the past [1]. The movement data was very low level, it only had the coordinates of the players and the ball. Therefore, we were not able to create the turnover column due to the lack of information, so we switched to looking at expected goals. With such raw data, we had to manually decide which event is a shot attempt by filtering the events to the ones where the ball reaches higher than 10 feet (the height of the rim) while the ball coordinates are also next to the rim. Finally, we also had to create a y-column that indicates whether it was a successful shot or not. We utilized the shot clock for this part since the shot clock will reset to 24 seconds when a basket is made, however, there are some rare scenarios that the shot clock will also reset to 24 seconds such as for a flagrant foul, but in such cases, the ball wouldn't reach the rim. Our main focus was on the soccer data since it was very time-consuming to clean up basketball data since it was very low level. There were no real means of encoding turnovers without manually labeling each set of events. The data was also nested where each event contains moments ranging from 14 to 2000. So we only applied it to a few models such as KNN and random forest.

Finally, to address the feedback from the mid-report, we actually want the events to be independent from each other. Instead of making them into a time series where the previous events could correlate to the next event, we actually wanted to predict the shot solely based on the current event and where every player and the ball is during the event of shooting.

6.1 Methods

For our machine learning analysis, we used several different methods to predict expected turnovers.

KNN: For K-nearest neighbors we used Euclidean distance and trained using k's from one to fifteen, validating on 10% of the training set. We see a steady decline in the F1-score for the validation set after k=3, so we tested with this k-value (see fig: KNN Validation).

Random Forest: For the random forest classifier we trained with max depths ranging from 10 to 42 (our total number of features) and between 50 and 200 estimators. We optimized using 5-fold cross-validation F1-scores and used the model and parameters with the highest validation accuracy. The best setup was using 300 estimators with no max depth (see fig: F1-Scores for Random Forest).

Logistic Regression: For the Logistic Regression classifier we used L2 Regularization as the penalty and a tolerance of 0.1. We used the default solver Limited Memory BFGS since the Saga solver provided poor results. Due to the difficulties in converging, the max_iter was increased from 100 to 1000 to ensure proper convergence. Despite this hyperparameter tuning, Logistic Regression still resulted in a poor F1-Score and a low amount of predicted turnovers per game (see fig: F1-Scores for Logistic Regression)

XGBoost: For the XGBoost classifier we used the binary logistic classifier with 10 boosting roundings. Other parameters included a learning rate of 0.5, a max depth of 5, gamma of 1, and the "exact" tree method which considers all candidates for splitting. This proved to be the most successful model by F1-Score and the Area Under the Curve (AUC) for the ROC curve.

6.2 Experiments and Results

Our Results using each of the methods above are presented in the table below. While all of them had relatively high accuracy, they varied quite a bit in F1-score and the area under the curve for ROC graphs. XGBoost was the superior method in all metrics, although the F1-score is still quite low, likely due to the imbalance of data and limited feature set, as mentioned above.

Method	F1-Score	AUC for ROC	Accuracy
KNN	0.21	0.68	0.93
Random Forest	0.36	0.88	0.95
Logistic Regression	0.09	0.84	0.94
XGBoost	0.37	0.89	0.95

(See fig: ROC curves)

7) Compute/Other Resources Used

We didn’t use any resources other than Google Colaboratory.

8) Conclusions

Projects of this nature always carry with them some ethical considerations. While the intention of this project is to foster an increased understanding of soccer and basketball, with the sports analytics and sports betting both becoming more mainstream we understand that risks of a project of this nature can result in using similar machine learning models for personal monetary gain. Furthermore, there has been no noticeable environmental impact from the free sample data that we used for this project. Perhaps access to the full Statsbomb dataset would result in the need for more computational power, but at this scale environmental impact is negligible. Finally, while we are given individual player positions and information, we are not using any individual player information at any point in our training process. This prevents any biases for or against certain players that could have otherwise emerged. For the basketball dataset, we were mostly limited by the data. Since we only usually have the position of the ball carrier and not all the other players, it’s hard for any ML algorithm to make use of the features at hand. In the future, if players can be tracked throughout the match in detail, we could see a great improvement in these results.

One of the main takeaways from this project is the complexity and meticulous attention to detail needed to preprocess sports data. While we expected an extensive preprocessing process, the level of detail and complexity of the datasets used presented a great challenge. Especially when working with the NBA data which was composed of raw player and ball positions with numerous nested events, it became clear why most of sports analytics are left to an expert group of data scientists. Despite this, a lot was learned in the machine learning process even with all the roadblocks. The main takeaway from the project is that there is still a lot to explore in the world of sports analytics and much can be learned even with the limited free data that is available. The Expected Turnovers metric hopefully turns fans passionate about soccer and the media’s preference for forwards and goals toward other equally important factors such as passing and defensive strategies.