Capstone report
Expected goals model

Sari Farah
BrainStation
Data science bootcamp
April 4, 2022

Problem statement

The goal of this project is to bring more data usage into the beautiful game of football across all levels of the sport and the organizations that operate within it. My project tackles just one aspect of what could be numerous analytical questions in football. This question is "What makes a shot successful in football?" By using mainly numerical and some categorical data I utilize machine learning techniques to build a classification model that predicts if a shot ends up being a goal or not. This will give coaches and players alike some extra information on how to approach a football match and the areas they should try to exploit to increase their chance of scoring more goals and in turn winning the match.

Background

Football is the most followed sport in the world, and its following is only getting larger. More than 1.1 billion people tuned in to watch the world cup final in 2018 [1]. In England 40% of the population watched the English premier league in 2020/2021 season [2]. For football clubs football isn't just a sport, it's a business and they need to win matches to be successful. In the champions league, every victory lands a team €2.8m [3]. This is where my project comes in to try to aid football teams in adjusting their tactics to try to increase their chances of scoring a goal.

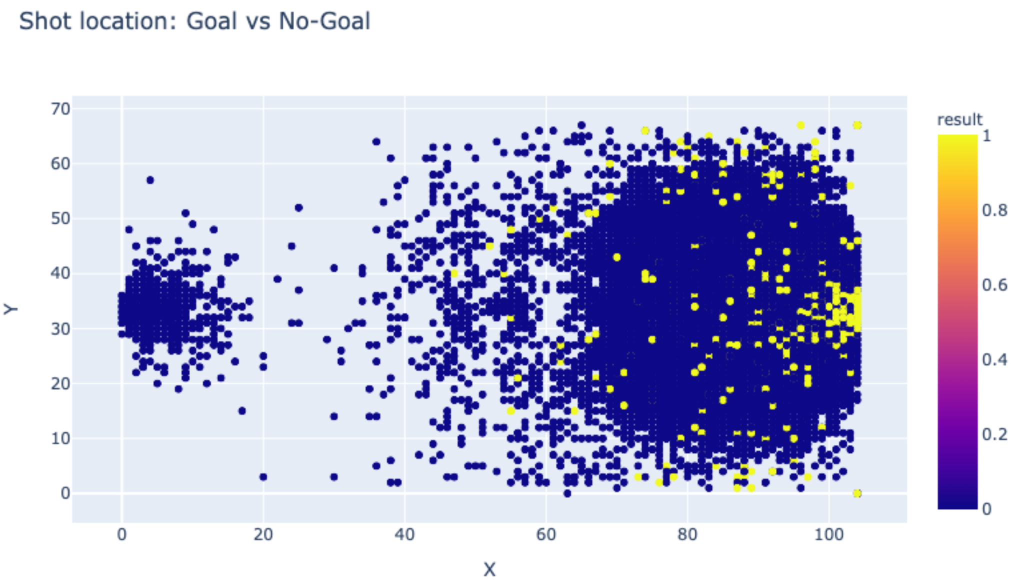
Data Source

The data for this project was acquired from PyPI. The Python Package Index (PyPI) is a repository of software for the Python programming language. PyPI helps you find and install software developed and shared by the Python community. In this case I found a web scraper for a football website that tracks the top five football leagues in Europe (English premier league, Spanish La Liga, German Bundesliga, French Ligue 1, and Italian Serie A). I scraped the events data (shots) for seasons 2015/2016-2019/2020, for the leagues mentioned above. The Data includes 227,614 rows and 19 columns including the minute, result, X coordinates, Y coordinates, player, home or away, situation, season, shot type, home team, away team, home goals, away goals, player assisted, last action, player id, match id, and date.

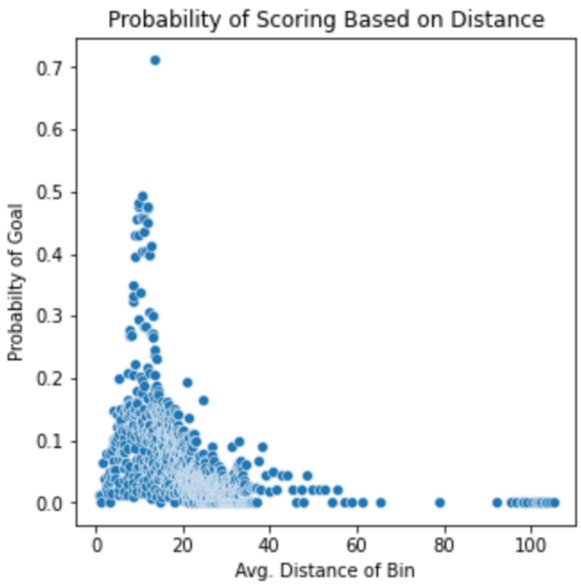
Data processing

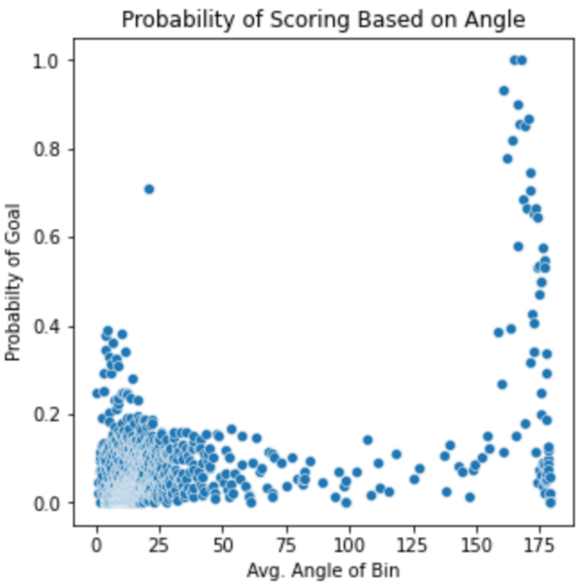
An exhaustive preprocessing of the dataset was performed. Since I wanted to use machine learning models, I had to turn all the categorical data into numeric data, this mainly done by binarizing certain columns, dropping columns that did not describe the shot data, and eventually one hot encoding the team names. I had very few null values, these were encoded with the applicable data.

Exploratory Data Analysis and Feature Engineering
The figure below shows the location of all successful vs non successful shots in my data set. There are some seemingly obvious characteristics of successful vs non successfully shots however they weren’t entirely clear. For this reason, I did some feature engineering using the X and Y columns (the coordinates) to extract the distance and angle of the shot.



We were able to deduce using the new features that we created how distance and angle affect the probability of scoring.





As you can see in the figure above, we are able to see that as the distance of the shot increases the probability of the shot resulting in a goal decrease. For the angle of the shot, we can see that the opposite is true, the probability of scoring based on angle increases as the angle of the shot increases.

Modelling

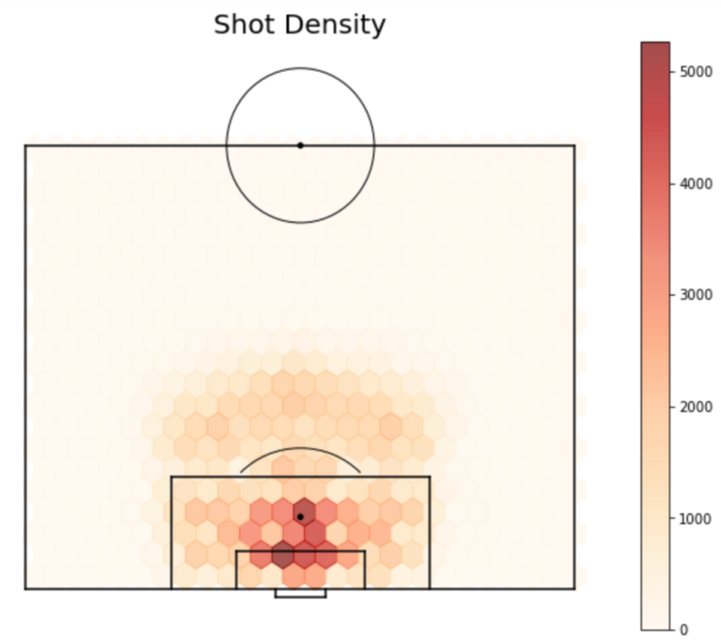
We modelled using classification models, as we had labeled data with result as our dependent variable. The three supervised machine learning models that we used for modelling were logistic regression, KNN model and Random Forest. For the purposes of this project, the most important metric to evaluate our models was the F1 score. The reason for this is that we had an extremely large imbalance in our dataset. To solve class imbalance problems is to use better accuracy metrics like the F1 score, the F1 score considers not only the number of prediction errors that your model makes, but that also look at the type of errors that are made. Therefore, I chose to evaluate my model based on the F1 score.

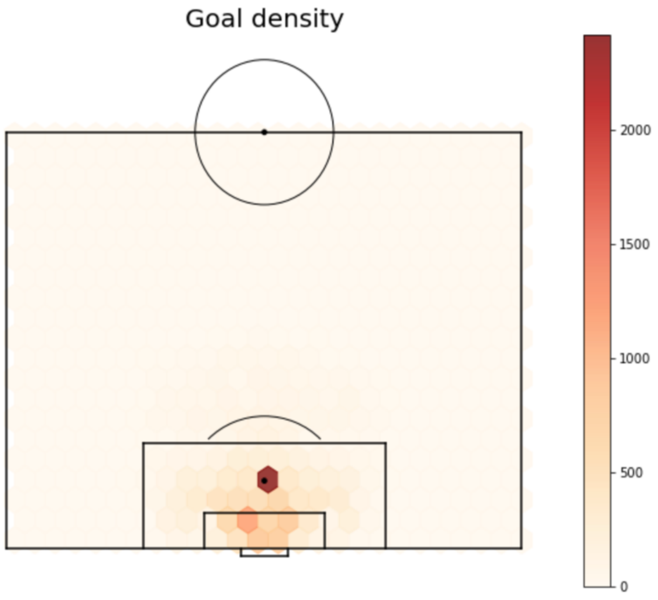
Rank	Model	F1 score
1	Random Forest	0.35
2	Logistic Regression	0.33
3	KNN	0.22

Findings

Our main findings is that the shots that result in the highest probability of scoring are the ones taken from the shortest distance and widest angle. However, the amount of shots taken from favorable locations and angles also do result in misses quite often. For this reason, when modelling, we couldn’t get the results that we'd hoped for since shots that end up being scored or missed often do have very similar characteristics. This is displayed in our two figures below. We were able to see this in our modelling by getting a lot of false positives and false negatives.

Furthermore, as we saw in our heatmap plots in EDA notebook, the most densely populated areas for shots on the pitch are the areas Infront of goal. When we filtered for goals only, we saw a massive decrease in the heatmaps intensity, however the observations were still in the same location, which leads us to further believe the fact that shots that were missed and scored have very similar characteristics even though the large majority end up being a miss.





Conclusion

In this study, we have trained a model that can classify a shot as a goal or not with a F1 score 0.35. Compared to simply predicting the if the shot resulted in a goal. We were able to do this with a ROC curve of 0.8 as well. The overall test accuracy for this model was 70%.

Next Steps

For next steps I would Find statistical data for each player, such as current goals ratio, conversion ratio, and so forth to incorporate into my analysis. The reason for this would be that the model would be able to evaluate which players are better finishers and that would increase the likelihood of their shot. I would add more features because different shots from the same distance and angle could be entirely different if there is a defender in front of the shooter versus no defender Infront of the shooter. If I had defenders' positions at the time of the shot, that could better classify if the shot was successful or not. I would incorporate statistical measures for the teams such as current average goals per game, or current average clean sheets per game. I would be able to incorporate that into my analysis to assess whether a team is likely to score and be scored on, which would give the machine learning models more information on the outcome of the shot. Also, I would run more models to with more hyperparameter tuning. I would train-test-split my data using techniques other than up-sampling to see what yields the best results. This would include down-sampling and SMOTE.

Works cited

Richter, F. (2022, February 11). *Infographic: Super Bowl pales in comparison to the biggest game in soccer*. Statista Infographics. Retrieved April 3, 2022, from https://www.statista.com/chart/16875/super-bowl-viewership-vs-world-cup-final/

A record-breaking season. Premier League Football News, Fixtures, Scores & Results. (n.d.). Retrieved April 3, 2022, from https://www.premierleague.com/season-review/the-fans/2164581?articleId=2164581#:~:text=40%20per%20cent%20of%20the,of%2043%20in%202019%2F20.

Desk, F. T. (n.d.). *Champions League prize money: How much do teams make?* Champions League prize money: How much do teams make? | FootballTransfers.com. Retrieved April 3, 2022, from https://www.footballtransfers.com/en/transfer-news/2021/05/champions-league-prize-money-how-much-teams-make