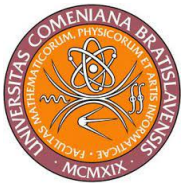
Project MAD

NBA statistics

Lukáš Jankola



Fakulta matematiky, fyziky a informatiky
Univerzita Komenského
June 2023

Contents

Chapter A

Introduction

In this project, we created a simple website for NBA statistics. It contains basic graphs for every team in NBA since 1990. Then we observed how 3 point shot has evolved over the years and created a model with which we can predict what position a player most likely plays based on some stats. We used a public dataset https://www.kaggle.com/datasets/wyattowalsh/basketball and only one file game.csv from there to get data from all the games. For model training we used this dataset of players from 2021-2022 season https://www.kaggle.com/datasets/vivovinco/nba-player-stats.

Chapter B

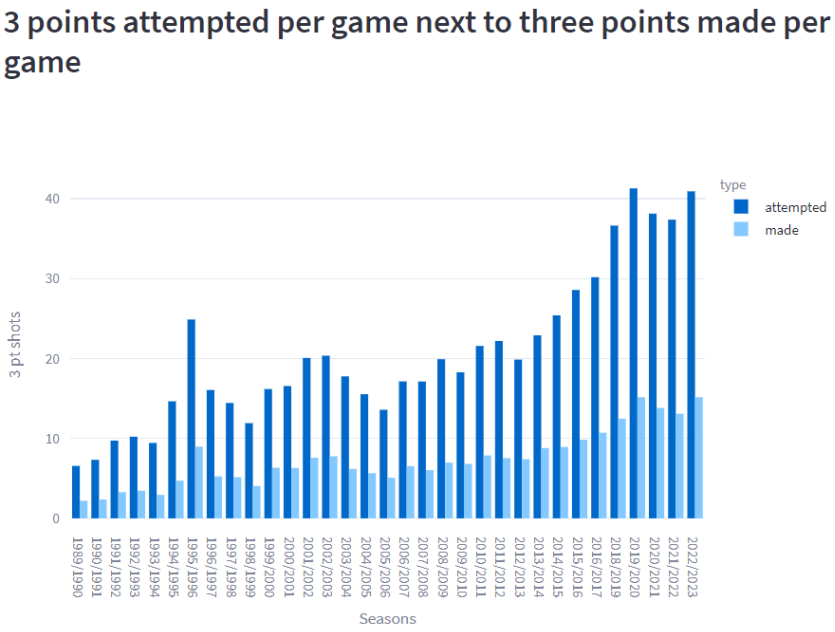
Description of methods

We decided to use streamlit to create the website for this project because it is straightforward to use and one can make visualizations with it and cache data. For some visualizations, we used plotly as it is good for interactive visualizations and compatible with streamlit. To work with data we used pandas as it is the best option. For model training, we used xgboost which implements boosted random forests. We chose it because it performs extraordinarily on similar problems. To evaluate the model we used some functions from scikit-learn. So we could get a latitude and longitude of cities we used geopy.

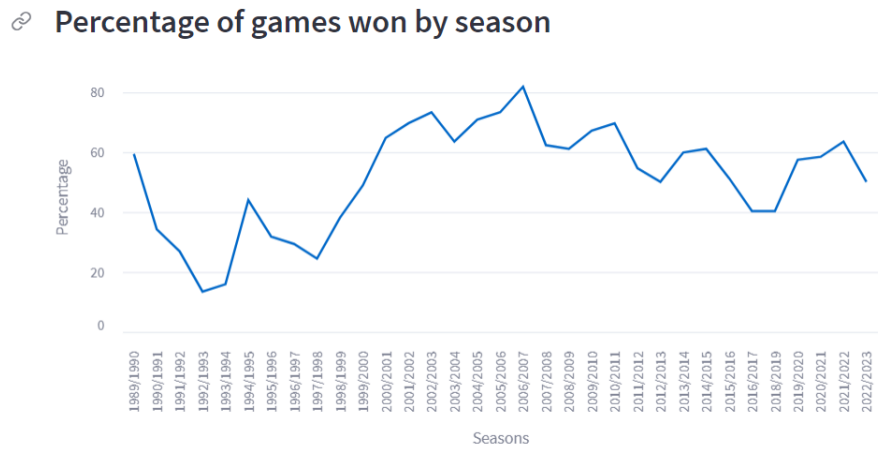
Chapter C

Visualizations for every team

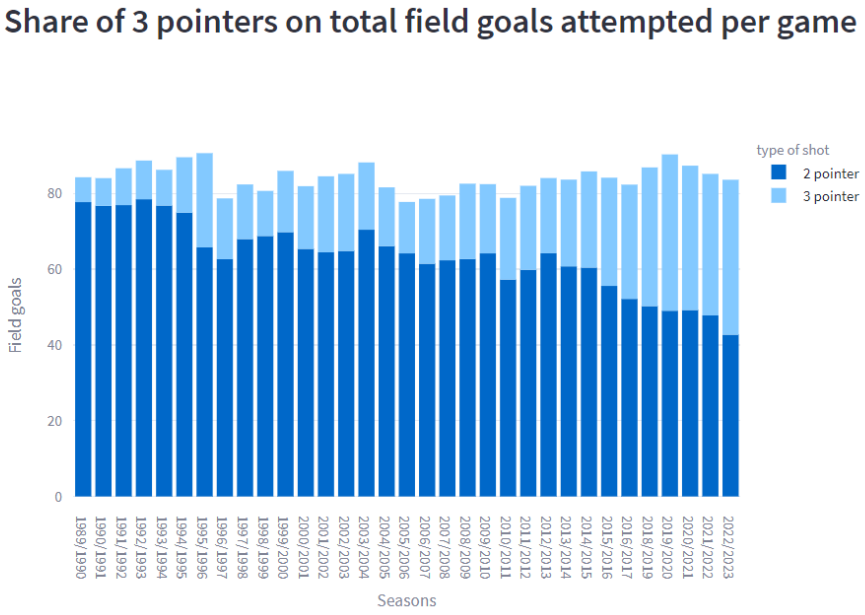
In this chapter, all charts are for Dallas Mavericks. On the first chart, we can see that over time the number of three-pointers attempted has grown, but the ratio has stayed roughly the same.



Here, we can observe how well a team was doing in a certain season and how they were improving in some years or getting worse in others.

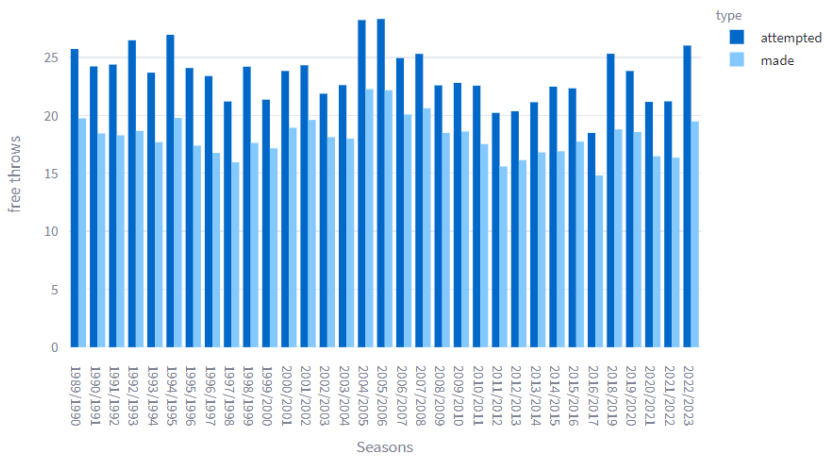


We can see that the number of field goals attempted has stayed roughly the same over time, but as in the previous graph, we can see that the number of 3-pointers attempted has been increasing which means, that the number of 2-pointers attempted has been decreasing.



This visualization was the most shocking for me because I assumed that the number of free throws over time would be increasing because I thought that the referees nowadays are more strict and that the players have figured out how to get more free throws. I also assumed that the ratio would change over time as I thought the players nowadays are better shooters, but none of my assumptions were correct.

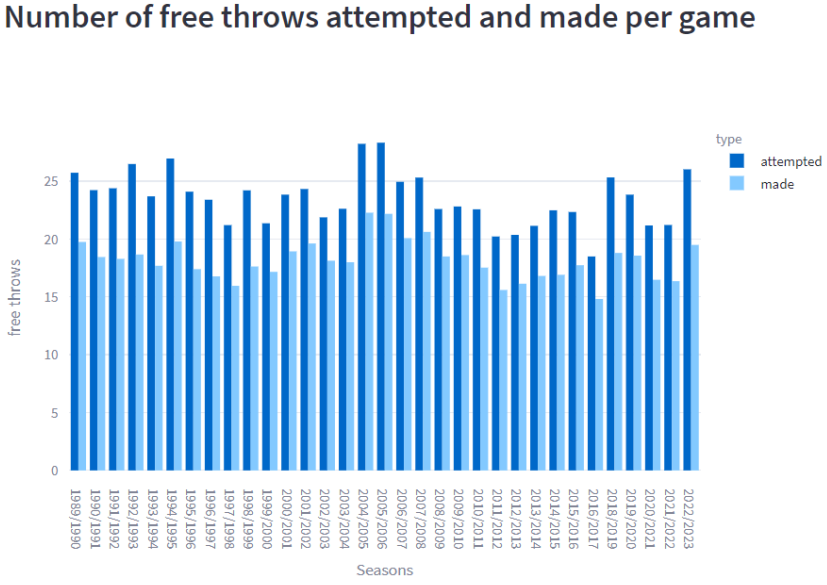
Number of free throws attempted and made per game



Chapter D

Evolution of 3 point shots

We can see: The number of 3-pointers attempted has really started growing around the year 2011 which might be because of the GSW team. Field goals attempted have stayed basically the same over the years because the number of 3-pointers has grown almost the same way as their share of total field goals. The accuracy from three has stayed the same over the years, only their number has grown. The growth of threes has slowed over the past years, so maybe it will stay the same for the years to come.



Chapter E

Training

Based on some variables we want to tell what position a player most likely plays. We used xgboost library for training. Variables used are points per game, assists per game, free throws attempted per game, and so on. As a training metric we used log-loss and an activation function was softmax, because we have more than two classes and softmax tells us a probability distribution for these classes. We didn't finetune the model as we were running out of time. Also, the dataset was very small. As a test set, we took 15 % of the data.

Chapter F

Conclusion

The most challenging part was learning how to work with the streamlit library as I was new to it. Also, when I wanted to plot some things I got stuck, but then found a solution in documentation. The part where I could do better was the training of the model. On the internet, I only found data for individual seasons for players, so if I had more time I would aggregate them and have a bigger dataset for training. Also, next thing to do would be to fine-tune the model using a cross-validation set and techniques such as GridSearch. The part that was easy for me was data manipulation with pandas as I work with it in my job.