ProjectDescription

DATA 1030

Kaiwen Yang

Git Repository: https://github.com/Melo96/data1030-project.git

Introduction

This project aims to use machine learning methods to predict the offensive win share of a NBA player base on his shooting tendency. Win share is an advanced player stat that indicates how much a player contributes to success of his team. It is an important attribute to evaluate a player. In this project, two datasets will be used. "Seasons_Stats.csv" contains the basic stats of all NBA players from 1950 to 2017. "nba_shot_types.csv" contains data about players' shooting tendency (e.g. the ratio of a player's shots inside three point line to shots outside the three point line). The target variable would be "ows", which is the offensive win share of a player. It is a continuous variable, so regression model would be implemented.

The topic of this project is highly related to the ongoing revolution of NBA. A lot of teams intorduced data science to game play analysis in order to improve offense and defense efficiency. They use more complex and detailed dataset to find out the most efficient offense choices, and most of the results suggest that NBA teams should encourage their players to shot more three-point and shot less mid-range jump shots. In this project, we could use simpler dataset to reproduce this analysis.

Dataset

In "nba_shot_types.csv", there are 3007 data points, and each has 23 features. According to Kaggle, this dataset has not been used in other projects. Here are column descriptions for columns that will be using:

"YEAR": NBA season. Category. Will be changed to discrete numerical data for the convinience of merging.

"PLAYER": Player name. Category.

The following data type are all float:

"PCT_FGA_2PT": Percentage of Field Goal Attempts That Were 2 PT Shots

"PCT_FGA_3PT": Percentage of Field Goal Attempts That Were 2 PT Shots

"PCT_PTS_2PT": Percentage of Points That Came From 2 PT Field Goals Made

"PCT_PTS_MR": Percentage of Points That Came From Midrange

"PCT_PTS_3PT": Percentage of Points That Came From 3 PT Field Goals Made

"PCT_PTS_FSTBRK": Percentage of Points That Came on Fast Breaks

"PCT_PTS_FT": Percentage of Points That Came From Free Throws

"PCT_PTS_OFF_TOS": Percentage of Points That Came Off Turnovers

"PCT_PTS_INTHEPT": Percentage of Points That Came In the Paint

In "Seasons_stats_complete.csv", there are 26.1k of data points and each has 50 features. This dataset was used to make some data visualization about players' season stat. In this project, only data between 2013-2019 will be used. Here are column descriptions for columns that will be using:

"Year": NBA season. Category. Will be changed to discrete numerical data for the convinience of merging.

"Player": Player name. Category.

"Pos": Position. Category.

"OWS": offensive win share. Continuous numerical.

"2PA": 2-point attempt. Discrete numerical.

"3PA": 3-point attempt. Discrete numerical.

"FTA": free throw attempt. Discrete numerical.

Preprocessing of Datasets

Two tables will be merged base on the "Year" and "Player" columns. After merging, these two columns will be droped.

one-hot encoder will be applied to: "Pos". The 5 different position are not ordinal. They just stand for the position of the player plays in that season.

MinMaxEncoder will be applied to columns "2PA", "3PA", "FTA" and columns start with "PCT*". "PCT*" columns are bounded by 0 and 100. "2PA", "3PA", and "FTA" are bounded because players can only attempt shots in a limited amount of time, so the shot attempts are bounded.