Project proposal: Survival Analysis on Football Players and the impact on their valuation.

Introduction

High-profile sports transcend mere professional disciplines; they are thriving multi-billion-dollar enterprises heavily dependent on diverse factors, with player value standing out as a crucial element. In the realm of football, the FIFA International Transfer Snapshot recently unveiled a staggering USD 7.36 billion expenditure on transfer fees from June to September 2023. This marks a remarkable 47.2% surge compared to the mid-year period in 2022 and a substantial 26.8% increase over the previous mid-year record set in 2019.

Unraveling the intricacies of how football clubs, both as sellers and buyers, determine a player's value proves to be a challenging endeavor. Numerous variables come into play, making it clear that there is no precise algorithm for accurately establishing a player's cost. Nevertheless, each season, clubs find themselves compelled to estimate the worth of their players, as well as those from other clubs to the best of their ability. Specialized platforms like Transfermarkt leverage the "wisdom of the crowd" principle, akin to the law of large numbers. On Transfermarkt, users engage in discussions on market values through forums, presenting evidence-backed arguments. The platform's staff then evaluate this collective information, combining it with their own research and knowledge to formulate a robust estimate.

Amid the myriad variables influencing a player's market value, one crucial factor is age. The rationale is straightforward: as a player ages, the remaining years of their career diminish, impacting their perceived value in the market.

This project aims to use survival analysis as a technique to assess the weight of players' age (via their risk of retirement) in their market value.

Data

The dataset for this project is gathered from Transfermarkt, a renown website that gathers football information, such as scores, results, statistics, transfer news, and fixtures and player estimated values. Historical data on players' contracts can contribute to build the "survival" probabilities needed for the analysis.

The dataset includes 30,000+ players from international clubs, 400,000+ player market valuations historical records and 1,200,000+ player appearance records from all games.

Techniques

Survival Analysis, traditionally applied to evaluate life expectancy, insurance rates, and annuities, involves assigning a probability model to assess risk and life expectancy. In this project, we extend these statistical principles to a less somber context: predicting the expected remaining years of a player's professional career.

At the core of survival analysis is the "survival function," quantifying the probability that a specific event has not transpired by a given time. In our case, the focal event is the retirement of a professional player, irrespective of the underlying cause. Another significant measure is the "hazard function," depicting the instantaneous failure rate at any given time, considering survival up to that point. By attributing a hazard risk for retirement, our aim is to correlate this risk with the current market valuation of players, with the aspiration of developing a model that can effectively evaluate and forecast future valuations.

Survival analysis employs various statistical techniques, including Kaplan-Meier curves, utilized for estimating the survival function; log-rank tests, employed for comparing survival curves among different groups; and the Cox proportional hazards model, which assesses the impact of multiple covariates on survival. These methodologies collectively form the analytical framework for our exploration into the interplay between players' age, retirement risk, and market valuation.

Survival Analysis, traditionally applied to evaluate life expectancy, insurance rates, and annuities, involves assigning a probability model to assess risk and life expectancy. In this project, we extend these statistical principles to a less somber context: predicting the expected remaining years of a player's professional career.

At the core of survival analysis is the "survival function," quantifying the probability that a specific event has not transpired by a given time. In our case, the focal event is the retirement of a professional player, irrespective of the underlying cause. Another significant measure is the "hazard function," depicting the instantaneous failure rate at any given time, considering survival up to that point. By attributing a hazard risk for retirement, our aim is to correlate this risk with the current market valuation of players, with the aspiration of developing a model that can effectively evaluate and forecast future valuations.

Survival analysis employs various statistical techniques, including Kaplan-Meier curves, utilized for estimating the survival function; log-rank tests, employed for comparing survival curves among different groups; and the Cox proportional hazards model, which assesses the impact of multiple covariates on survival. These methodologies collectively form the analytical framework for our exploration into the interplay between players' age, retirement risk, and market valuation.

Notes:

Pros: extensive dataset publicly available. Easy to manipulate and clean. Many possible extensions: clubs, games, appearances, etc. Possible to join to other datasets if needed.

Cons: retirement is not recorded for players, and sometimes not easy to define (some players might be inactive, but not formally retired, or come back from retirement in other cases). Some assumptions need to be made (e.g., a player with no contract, no club and no appearances in the last X months is deemed retired).