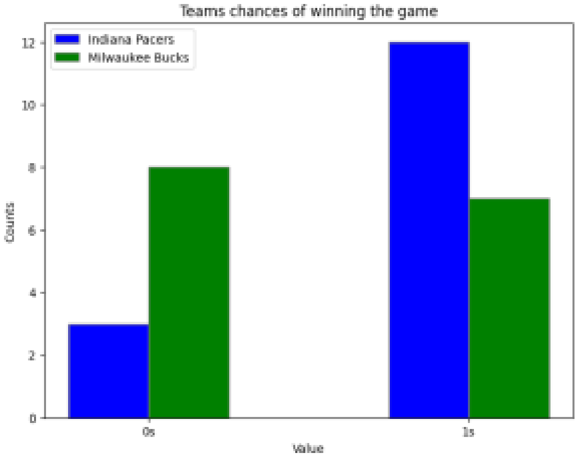
Dribble, Predict, Win: Leveraging PySpark & ML for Real-Time NBA Game Forecasts

Summary: Our project aimed to harness the power of predictive analytics in the context of NBA basketball games using PySpark, with the overarching goal of enhancing decision-making processes for stakeholders in sports analytics and the betting industry. Our journey began with the extraction of data from the official NBA website, scraping various statistics including team, player, and game data. This process involved meticulous extraction and processing to compile a comprehensive dataset conducive to training sophisticated machine learning models.

Having amassed a rich dataset comprising historical game performances, player statistics, and team dynamics, our next step was preprocessing. We meticulously cleaned and formatted the data, ensuring consistency and compatibility with our chosen machine learning algorithms. With a refined dataset in hand, we proceeded to integrate live streaming data from ongoing NBA matches into our framework. The integration of real-time data posed a new set of challenges, necessitating further preprocessing to align it with our historical data. Once both historical and real-time datasets were harmonized, we employed various encoding techniques and trained machine learning models using PySpark. Leveraging three classifiers—logistic regression, random forest classifier, and decision tree classifier—we evaluated their performance and selected the most effective model for predicting game outcomes in real time (which has 72 % accuracy).

Using the selected model, we input real-time game statistics to predict the likelihood of each team winning the match. By evaluating the predictions for both teams, we determined the eventual winner, thus fulfilling our objective of accurately predicting NBA game outcomes in real time. A pivotal aspect of our project was visualizing the predictions, providing stakeholders with clear insights into the probable results of basketball games. The plot generated from our analysis, showcasing which team had more '1's, symbolizing a higher likelihood of victory, served as a tangible representation of our predictive capabilities.

In summary, our project successfully developed a robust pipeline leveraging PySpark and machine learning algorithms to predict NBA game outcomes in real time. By integrating historical and real-time data, preprocessing, training models, and visualizing predictions, we provided actionable insights for coaches, players, analysts, and betting companies, contributing to the advancement of sports analytics and decision-making processes.



Plot of a live game (INDIANA PACERS Vs MILWAUKEE BUCKS) on 28 April 2024