Christopher Wille

Factors Contributing to High Earnings in Professional Athletes: A Supervised Learning Approach

Introduction:

The global professional sports industry has become a multi-billion dollar behemoth, with top athletes earning exorbitant salaries and lucrative endorsement deals. However, what factors contribute to their success and wealth? In this project, my goal is to analyze a dataset of the highest-paid athletes to identify any patterns or correlations that could explain why certain athletes earn more than others. Through the application of unsupervised learning techniques, I aim to explore the data and uncover hidden relationships that may shed light on why some athletes achieve greater success than others. This project has the potential to provide valuable insights into the factors that contribute to success in professional sports, and could have implications for the industry as a whole.

Goal:

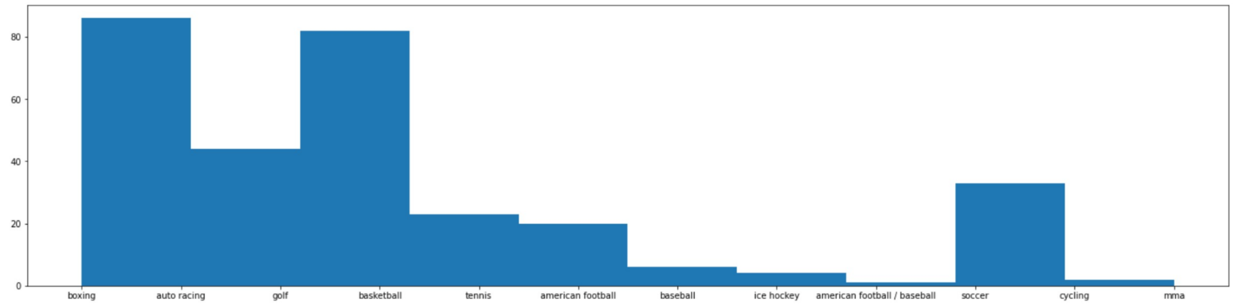
The main objective of this project is to conduct a comprehensive analysis of a dataset containing information on the highest-paid athletes and uncover key factors that contribute to their earnings. Specifically, the project aims to explore the relationships between various parameters such as sport, position, and nationality to determine if there are any patterns or correlations that can explain why certain athletes earn more than others. Furthermore, if possible, a predictive model will be developed to identify the sport of a player based on other parameters in the dataset. By achieving these goals, this project can provide valuable insights into the factors that drive the success of professional athletes and inform decision-making in the sports industry.

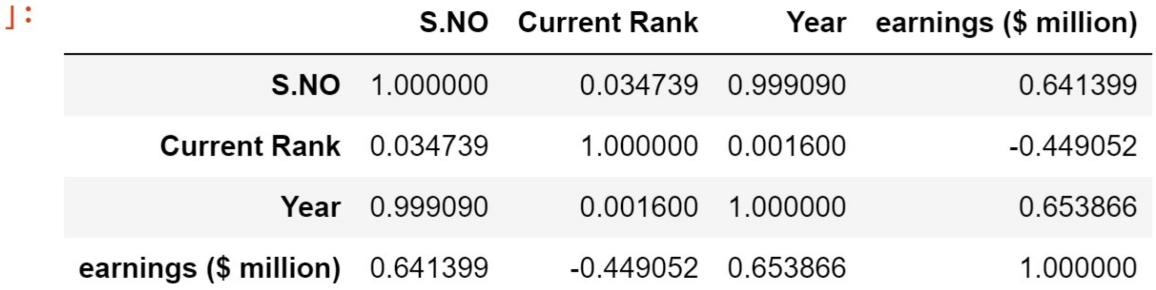
Methodology:

To analyze the data and identify hidden relationships between variables, a supervised learning approach will be utilized. The data will be imported and manipulated using Python's Pandas library, while Matplotlib and Seaborn will be used to visualize and explore the data. To ensure the accuracy of the analysis, the data will first be preprocessed by removing any missing or duplicate values, and encoding categorical variables. Following this, the data will be split into training and testing sets for further analysis. Next, I will apply a range of supervised learning algorithms to the training data, including Random Forest Regressor, K-Nearest-Neighbors, etc. To evaluate the performance of each model, I will use metrics such as accuracy score and mean squared error. Based on the results, I will select the best-performing model and fine-tune it using cross-validation to determine the optimal hyperparameters.

Results and Impact:

Our project aims to shed light on the factors that contribute to high earnings in professional sports. By utilizing supervised learning techniques, we can identify hidden relationships and correlations that can explain why certain athletes earn more than others. The insights gained from this project will be invaluable to sports analysts, managers, and athletes seeking to comprehend the drivers of success and high earnings in professional sports. Overall, this project has the potential to create a significant impact on the sports industry by providing valuable information and insights that can inform decision-making processes. To give an idea of the data, the first image shows the distribution of sport in the dataset and the second is a correlation matrix of the parameters.





	S.NO	Current Rank	Year	earnings ($ million)
S.NO	1.000000	0.034739	0.999090	0.641399
Current Rank	0.034739	1.000000	0.001600	-0.449052
Year	0.999090	0.001600	1.000000	0.653866
earnings ($ million)	0.641399	-0.449052	0.653866	1.000000

(Note: the second image only has the correlations between float values as the data is yet to be encoded)

Conclusion:

In this project, I will employ unsupervised learning techniques to analyze a dataset of the highest-paid athletes to uncover patterns and correlations that may clarify why certain athletes earn more than others. My aim is to develop a model that can predict the sport of a player based on other parameters in the dataset, and to identify factors that contribute to high earnings among athletes, including sport, position, nationality, and endorsement deals. The insights gained from this project have the potential to provide valuable information to sports analysts, managers, and athletes, contributing to a better understanding of the world of professional sports.

Annotated Bibliography

Parul Pandey. (2020). Forbes Highest-Paid Athletes (1990-2019). Retrieved March 29, 2023, fromhttps://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019