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

By: Christopher Wille

Introduction

In the ever-growing professional sports industry, top athletes earn exorbitant salaries and enjoy lucrative endorsement deals. However, the underlying factors that contribute to their success and wealth remain somewhat obscure. With this project, my primary objective is to conduct a thorough analysis of a dataset containing information on the highest-paid athletes and identify any patterns or correlations that may explain why certain athletes earn more than others. By employing supervised learning techniques, the intention is to delve deeper into the data and uncover any hidden relationships that could potentially shed light on why some athletes achieve greater success than others. The potential insights gleaned from this project could prove to be invaluable for the sports industry and could be leveraged by teams, organizations, and fans alike to make more informed decisions when it comes to hiring and paying athletes. At present, there exists a dearth of comprehensive solutions or workarounds for this specific problem, which further underscores the importance and urgency of this project. Ultimately, the findings of this project could have far-reaching implications and serve as a crucial resource for shaping future decision-making in the professional sports industry.

Goals

The main goal of this multifaceted project was to delve into the intricacies of the professional sports industry and identify the crucial factors that contribute to the high earnings of top athletes. To accomplish this, a comprehensive analysis of the dataset was conducted, and various parameters such as sport, year of payment, and nationality were scrutinized. By meticulously examining these factors and unearthing any correlations or patterns that may exist, the project attempted to unveil the reasons why some athletes earn more than others. Additionally, the development of a predictive model that can identify a player's sport based on other parameters in the dataset will be a crucial element of this project. The ultimate goal is to generate valuable insights into the underlying factors that drive the success of professional athletes and assist in making informed decisions within the sports industry. Since there are currently no comprehensive solutions or workarounds for this specific problem, the outcomes of this project can make a significant contribution to the field and potentially shape the future directions for the industry and the fervent fans who are keenly interested in what factors contribute to the salary of their favorite athletes.

Methodology

To facilitate the analysis of the data and uncover any hidden relationships between variables, an unsupervised learning approach was employed. The data was be imported and manipulated using Python's Pandas library, and to gain better insights and explore the data more thoroughly, Matplotlib and Seaborn will also be utilized. Prior to conducting the analysis, it is crucial to ensure the accuracy of the results. To achieve this, the data was preprocessed by removing any missing or duplicate values and encoding categorical variables. The process began

with use of data analyzing tactics to see what potential features correlate to their very high ranking in the dataset. Then, the data was divided into training and testing sets for further analysis. A variety of supervised learning algorithms, such as using various Machine Learning estimators. Additionally, further analysis was conducted to answer specific questions regarding the top-20 highest paid athletes in the dataset. To evaluate the performance of each model, several metrics, including accuracy score and mean squared error, were utilized. Finally, based on the results, the best-performing model was be selected, and cross-validation was used to determine the optimal hyperparameters.

Get the Data

	S.NO	Name	Nationality	Current Rank	Previous Year Rank	Sport	Year	earnings ($ million)
0	1	Mike Tyson	USA	1	NaN	boxing	1990	28.6
1	2	Buster Douglas	USA	2	NaN	boxing	1990	26.0
2	3	Sugar Ray Leonard	USA	3	NaN	boxing	1990	13.0
3	4	Ayrton Senna	Brazil	4	NaN	auto racing	1990	10.0
4	5	Alain Prost	France	5	NaN	auto racing	1990	9.0
...	...	...	...	...	...	...	...	...
296	297	Stephen Curry	USA	6	9	Basketball	2020	74.4
297	298	Kevin Durant	USA	7	10	Basketball	2020	63.9
298	299	Tiger Woods	USA	8	11	Golf	2020	62.3
299	300	Kirk Cousins	USA	9	>100	American Football	2020	60.5
300	301	Carson Wentz	USA	10	>100	American Football	2020	59.1

(Figure 1)

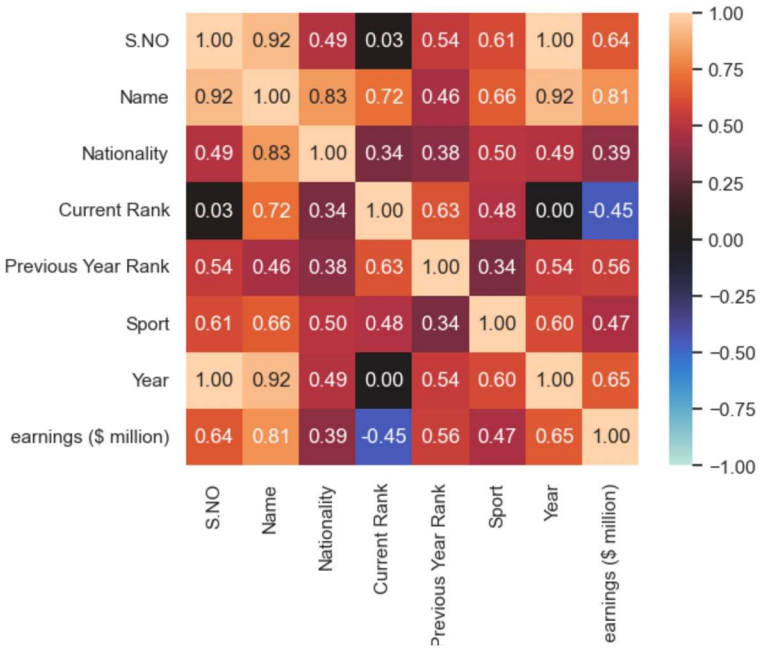
Figure 1 is a dataset from Forbes Top 300 Highest Paid Athletes From 1990-2020 with 301 rows of data and the following column names and data type of each column :

- Current Rank - Where they rank in Forbes' most recent list: int64
- Nationality - Nationality of the athlete: object
- S.NO - Rank in the entire dataset: int64
- Name - Name of the athlete: object
- earnings ($ million) - Earnings the year the athlete was ranked: float64
- Sport - Sport the athlete made the earnings from: object
- Year - Year the athlete earned the money: int64

Data Exploration

To start the exploration off in the barest bones ways possible, it was started off with a simple correlation matrix which will compare all the features to one another. This gave an initial idea of how the features in the dataset are connected or lack thereof. In data analysis, a correlation matrix is a table that shows the correlation coefficients between a set of variables. The correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive

correlation), with 0 indicating no correlation. By using a correlation matrix, one can quickly see which features in a dataset are strongly correlated with each other and which are not. For example, if two features have a high positive correlation coefficient, it means that they tend to increase or decrease together. On the other hand, if two features have a high negative correlation coefficient, it means that they tend to have an inverse relationship, whereas one feature increases, the other decreases. By examining the correlation matrix, analysts can get an initial idea of how the features in a dataset are connected or not, which can be useful for further analysis and modeling. It can also help identify potential problems such as multicollinearity, which occurs when two or more features are highly correlated with each other, making it difficult to distinguish the effects of each feature on the target variable.



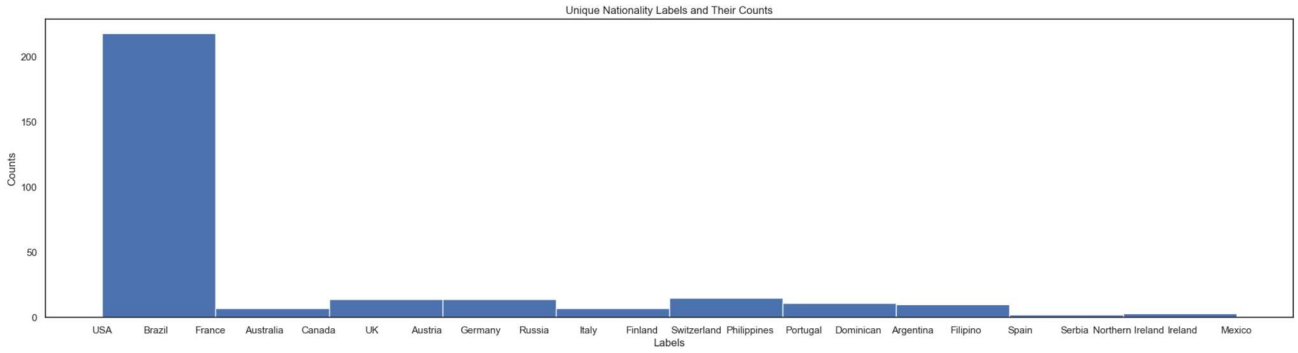
(Figure 2)

As the color bar indicates, the features in the correlation matrix (Figure 2) seem to be relatively strongly correlated with each other, meaning that there are potentially meaningful relationships among the variables that could be explored further. This initial observation is promising and suggests that there may be valuable insights to uncover from the data. To ensure the accuracy of the analysis, each column of the data was analyzed for potential anomalies, such as missing data, incorrect labels, or repeated information. This is an important step in data cleaning and preparation, as any anomalies could impact the accuracy of the analysis. Additionally, it was decided to exclude three of the variables - S.NO, Current Rank, and Previous Year Rank - from the analysis. These variables may be related to Forbes' ranking methodology rather than the earnings themselves, so including them in the analysis may not have provided useful insights. This decision demonstrates an understanding of the specific research question and the importance of carefully selecting variables that are relevant to that question. Overall, the preliminary analysis of the correlation matrix suggested that the data is in a good state to move forward with the analysis, and that there may be useful insights to uncover. By carefully

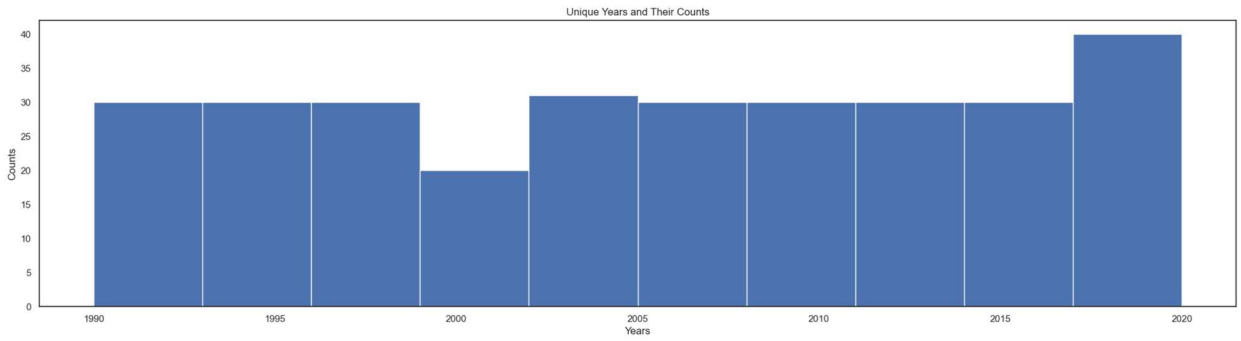
examining the data and selecting relevant variables, this led the way to start taking steps to ensure that the analysis was accurate and meaningful.

```
: Sport
  American Football              17
  American Football / Baseball    1
  Auto Racing                    10
  Auto Racing (Nascar)            2
  Auto racing                     1
  Baseball                        3
  Basketball                     54
  Boxing                         29
  F1 Motorsports                  5
  F1 racing                       8
  Golf                           24
  Hockey                          1
  Ice Hockey                      2
  MMA                             1
  NASCAR                          3
  NBA                             1
  NFL                             3
  Soccer                         22
  Tennis                         18
  auto racing                     7
  baseball                        3
  basketball                     27
  boxing                         17
  cycling                         1
  golf                           20
  ice hockey                      1
  motorcycle gp                   4
  soccer                         11
  tennis                          5
  dtype: int64
```

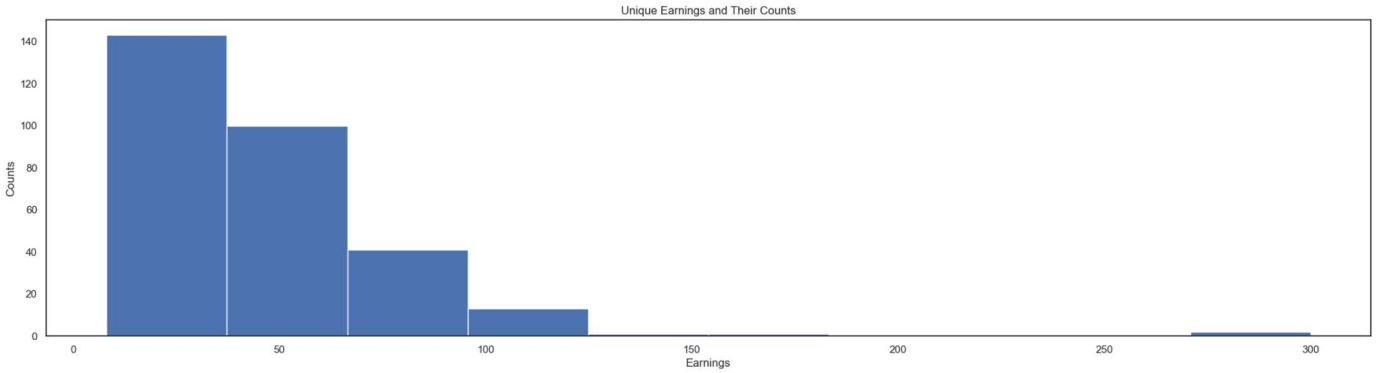
(Figure 3)



(Figure 4)



(Figure 5)



(Figure 6)

```
Athlete Name Non-NA's:

301
```

(Figure 7)

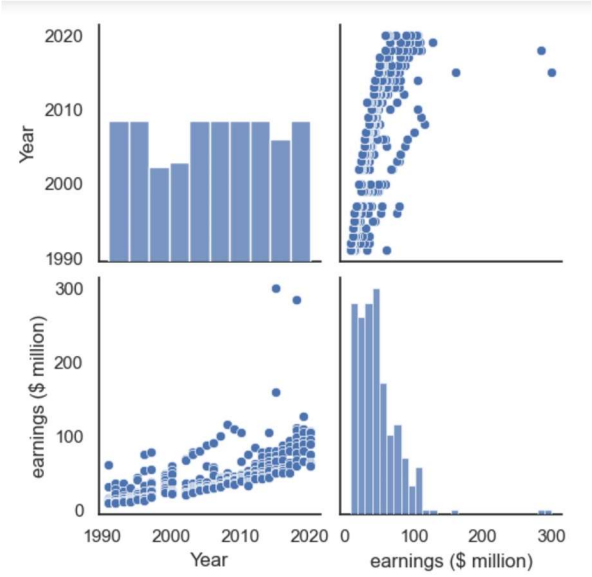
The 'Name' (Figure 7) and 'Sports' (Figure 3) columns were analyzed differently than the other columns due to the nature of the data they contain. The 'Name' column contains categorical data, specifically the names of over two hundred athletes, which requires different analysis techniques than numerical data. The 'Sports' column is also a categorical variable, which had some inconsistencies in the labeling of the sports. Upon closer inspection, there were some sports that are labeled as unique when in fact they are the same sport with different labels. For example, 'Hockey' and 'Ice Hockey' are both listed as unique sports, even though they are the same sport. This type of inconsistency could have led to inaccurate analysis and misleading results if not properly addressed. To address this issue, it was decided to relabel some of the sports in the 'Sports' column. By consolidating redundant categories and ensuring consistent labeling across the dataset, it can ensured that the analysis is accurate and meaningful. This process is called data cleaning, and it is an important step in preparing data for analysis. As for the other features, Figure 4 suggested a bias in the dataset to USA born athletes while Figure 5 implied there is a

pretty even distribution of years of the athlete's earnings. Figure 6 showed the lower earnings make up a majority of the 'earnings ($ million)' feature.

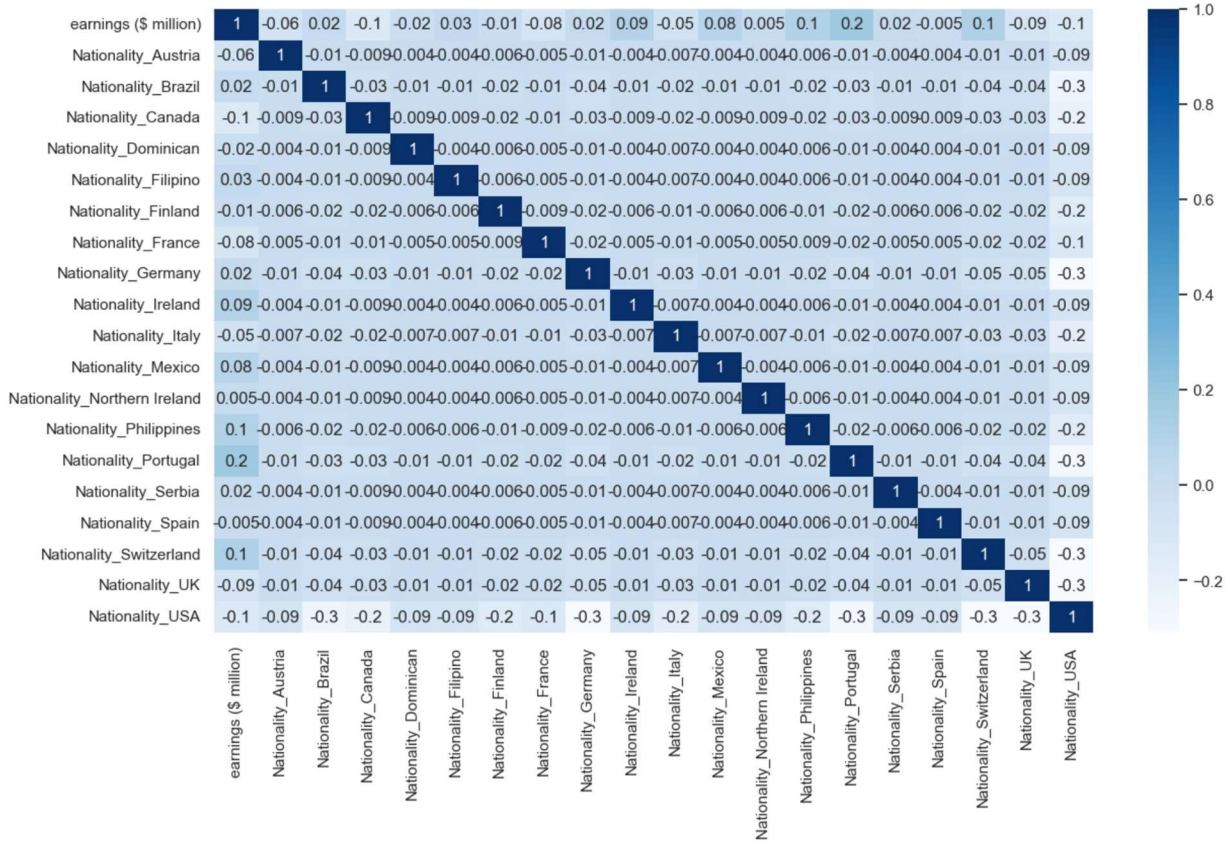
```
Sport
american football               20
american football / baseball     1
auto racing                     40
baseball                         6
basketball                      82
boxing                          46
cycling                          1
golf                            44
ice hockey                       4
mma                              1
soccer                          33
tennis                          23
dtype: int64
```

(Figure 8)

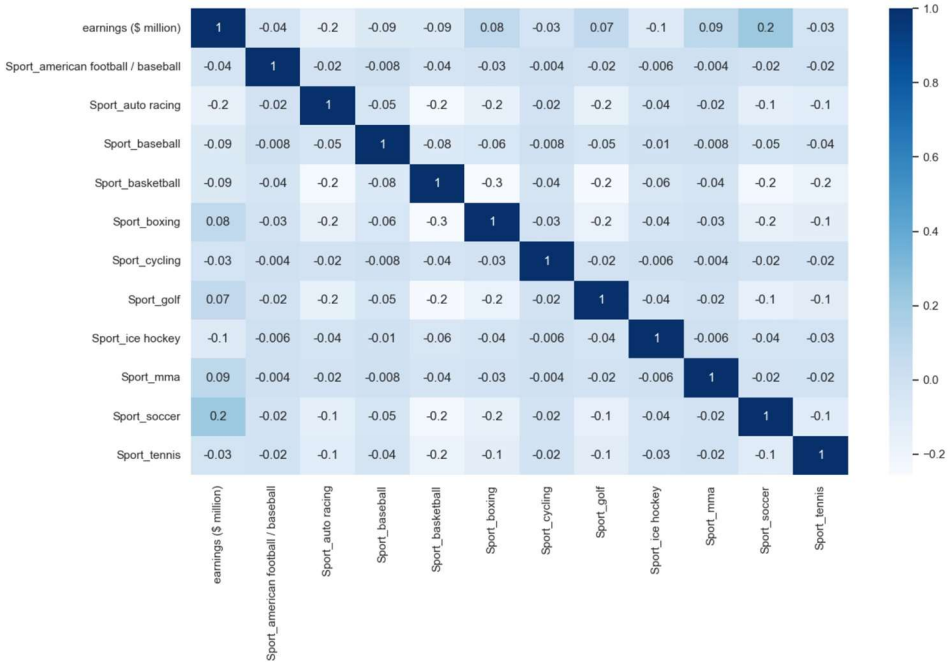
Figure 8 showed that there are potential biases if the model chosen included 'Sport' as a feature since there were more data entries for some sports than others. There were potential biases in the 'Name' feature to keep an eye out as well. The other features seemed relatively even in terms label variance and there also did not seem to be any NA values. Dummy variables were then used on the object data type features so correlation plots and maps comparing all the features to one another using the Seaborne library package can be developed. 'Name' was also not be included as a feature here on out since for the repeat athletes that were in there may create a bias in a model building process as well it would make sense for datapoints to be independent of who you are and more dependent on the statistics behind who you are if that makes sense. A target parameter of 'earnings ($ million)' made sense as the goal is to see what correlates to the money the athlete makes. These correlation plots gave an early look at what features may be good for building a model later.



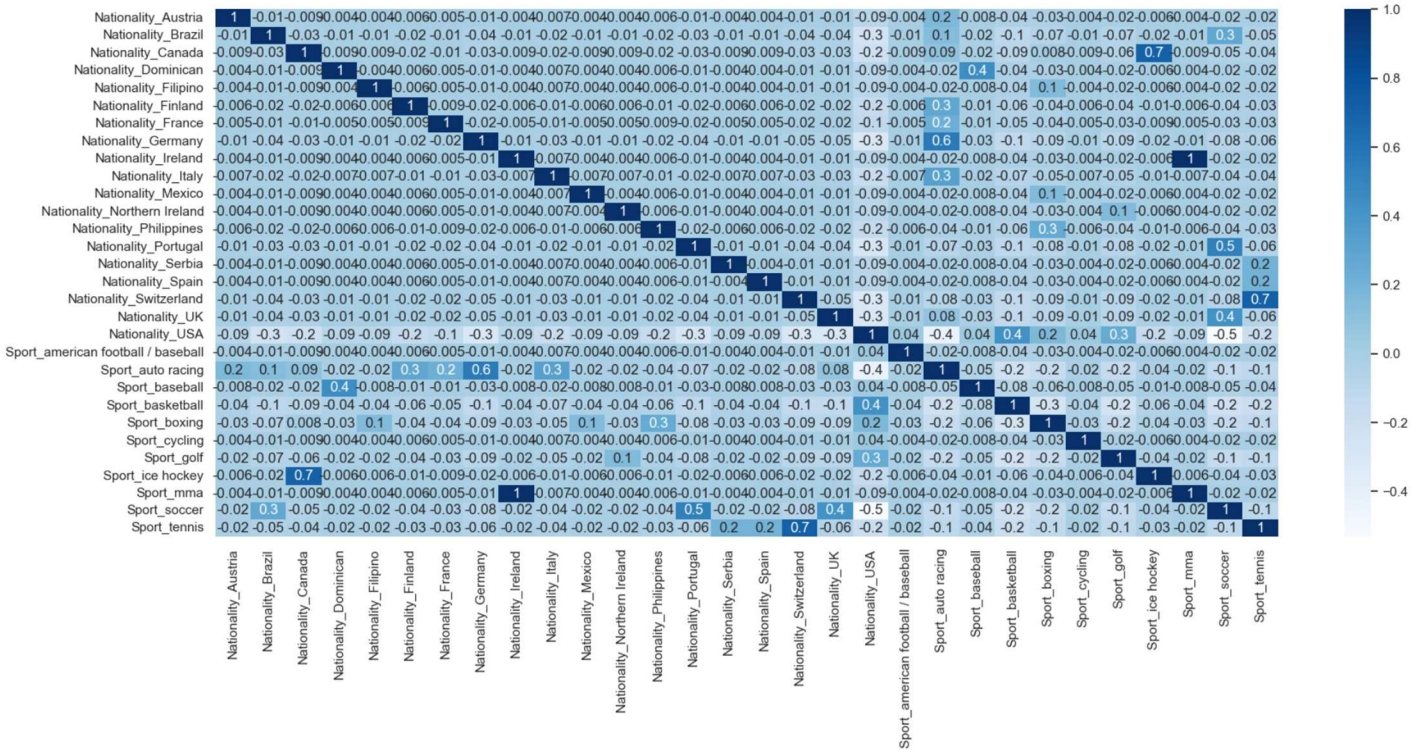
(Figure 9)



(Figure 10)

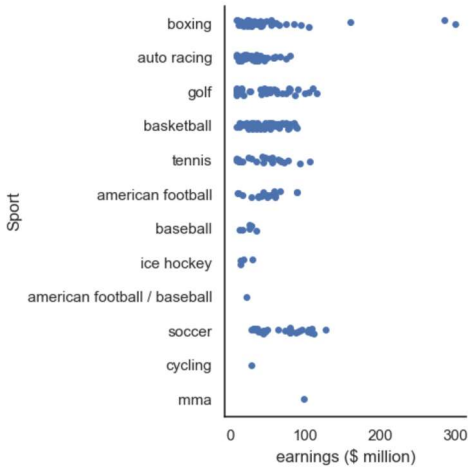


(Figure 11)

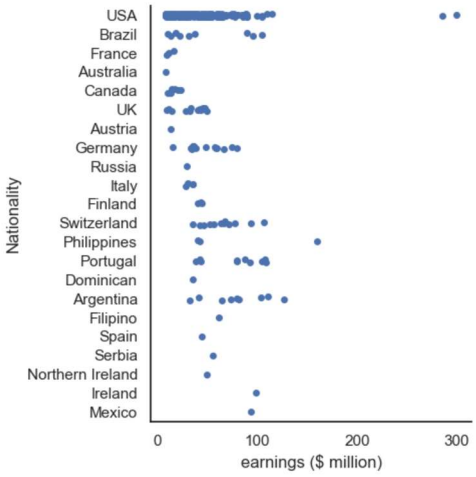


(Figure 12)

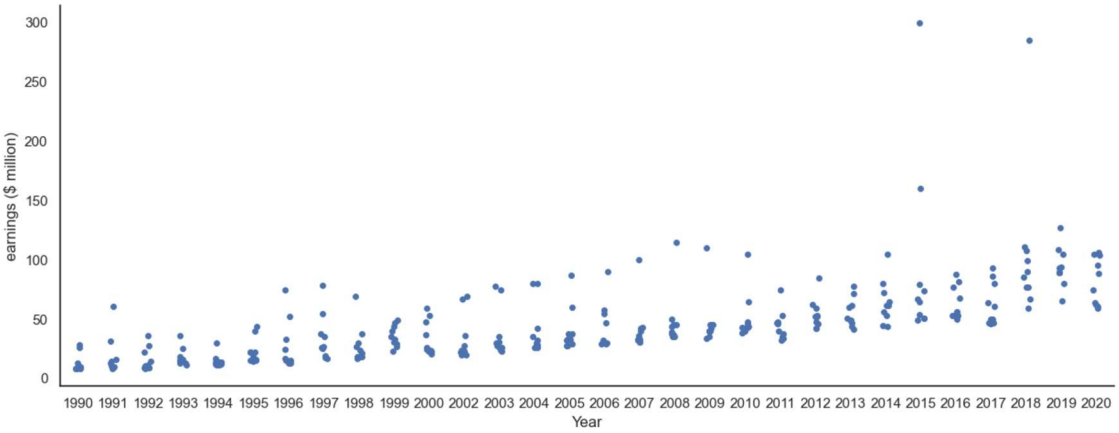
Regarding Figure 9, it appeared that there is a clear positive correlation between the year and earnings, which is the strongest correlation that was analyzed. This could potentially be attributed to inflation or perhaps athletes being more highly valued over time, or a combination of both factors. Moving on to Figure 10 and Figure 11, it appeared that they were both in a similar position, with the former examining the correlation between nationality and earnings and the latter looking at the correlation between sport and earnings. However, neither of them showed any particularly strong correlations. Figure 12, on the other hand, which examined the correlation between nationality and sport, yielded the second-strongest correlations. It was interesting to see if the trend held up not only in model building process but also when analyzing the top-20 highest-paid athletes. The distribution of earnings across the three other features in question were observed next.



(Figure 13)



(Figure 14)



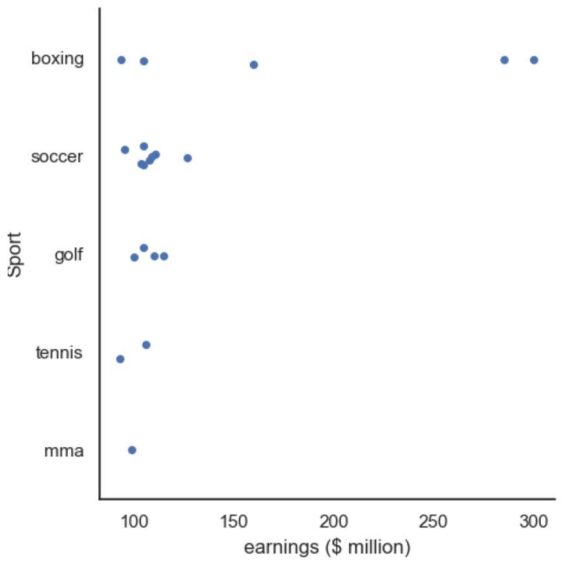
(Figure 15)

For Figure 13, the sport of boxing contained the highest paid athletes, with three clear outliers. An outlier is a data point that is significantly different from the others in the dataset. In this case, there were three athletes in boxing who earned significantly more than other athletes in the same sport. The earnings of athletes in other sports appeared to be clustered on the lower end, which could of made it difficult to use this feature in later analysis. This suggested that the sport of boxing may be more lucrative for athletes than other sports. However, it is important to note that outliers can sometimes skew the results of an analysis, and it may be necessary to investigate these outliers further to determine if they were legitimate data points or errors in the data.

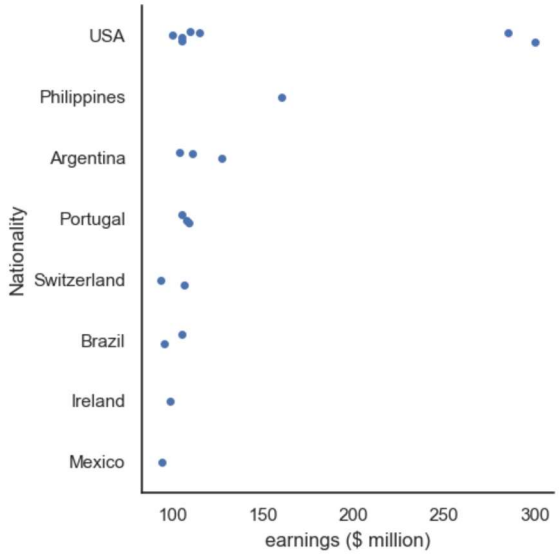
Figure 14 showed the distribution of earnings by the nationality of the athletes. The figure revealed that there is a large number of US-born athletes present in the data, but the rest of the nationalities appear to be relatively evenly distributed in terms of earnings. There was less clustering of earnings by nationality in that figure than in Figure 9. It is important to note that the distribution of nationalities in the dataset may not have been representative of the overall population of athletes, and this could have potentially impacted the generalizability of the results.

Figure 15 showed the relationship between earnings and the year in which the athlete earned those earnings. The analysis revealed another positive correlation between the two variables, supporting the earlier analysis of this relationship. A positive correlation means that as one variable (in this case, the year) increases, the other variable (earnings) also tends to increase. These finding suggested that athletes were earning more over time, which could be due to a variety of factors, such as changes in the sport, increases in viewership, or changes in sponsorship opportunities. However, it is important to note that correlation does not imply causation, and further analysis would be necessary to determine the underlying factors driving this relationship.

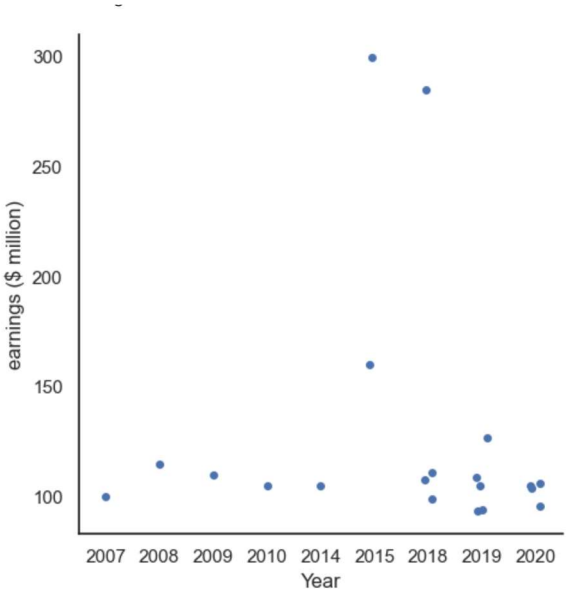
Results of Analyzing the Top 20 Athletes in the Dataset



(Figure 16)



(Figure 17)



(Figure 18)

Figure 16 had a lot of similarities to Figure 13. The top-3 highest paid athletes were still in the same sport obviously, but it is interesting to note that the distribution still looked similar as

when the whole dataset was being analyzed this way. One possible explanation could be that boxing has a relatively small number of high-profile athletes who command the majority of the earnings in the sport. Additionally, it is worth pointing out that the only MMA datapoint is in the top-20 as well, which could suggest that MMA is a growing sport that is starting to attract higher earnings. Moving on to Figure 17, the most interesting detail is that US born athletes lead all the nationalities in the top-20. This could be due to the fact that the US has a larger sports industry compared to other countries and is therefore able to pay athletes more. Lastly, for Figure 18, there was an expectation that the earnings in the top-20 would increase with time but this was not actually the case. The top-paid athlete received their earnings in 2015 and there has not been any athlete nearly paid that high up to 2020. This could mean that the top-3 salaries may truly be outliers, or it could indicate that there has been a shift in the sports industry where earnings are not increasing at the same rate as they once were.

Results of Shortlisting Promising Models

Starting in the right direction of testing Machine Learning (ML) models can be difficult at times. To begin, strong predicting ML models were used such as Random Forest Regressor, Gradient Boosting Regressor, and K-Nearest Neighbors. Random Forest Regressor builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It is a powerful model that can handle a large number of input features, and is often used in both regression and classification problems. The Gradient Boosting Regressor model builds an ensemble of decision trees sequentially, where each subsequent tree tries to correct the errors of the previous tree. It is also a powerful model that can handle a large number of input features, and can be used in both regression and classification problems. K-Nearest Neighbors classifies new data points based on the k-nearest neighbors in the training set. It is a simple but effective algorithm that can handle non-linear relationships between features and can be used for both regression and classification problems.

All these models initially used 'Year', 'Sport', and Nationality as X features with 'earnings ($ million)' as the target feature. This is a good starting point as these variables had been identified as important in the analysis done previously. It's important to note that the performance of the models may vary depending on which variables are included in the model, as well as how those variables are preprocessed. To measure the success of the models, the data was split into a 75% training set and a 25% testing set. Each model was trained using the training set X features which then the model tried to predict the testing set target features. The performance of the respective model was measured using R-Squared(R2) and mean-squared-error (MSE). R-Squared (R2) is a statistical measure that represents the proportion of the variance in the target variable that is explained by the independent variables in the model. It is a value between 0 and 1, where 0 indicates that the model explains none of the variability in the target variable and 1 indicates that the model explains all of the variability in the target variable. Mean Squared Error (MSE) is a measure of how well the model fits the data. It is the average of the squared differences between the predicted values and the actual values. A lower MSE indicates that the model is better at predicting the target variable. In the context of the ML models mentioned in the paragraph, R2 and MSE are used as evaluation metrics to measure the performance of each model. The R2 score indicates how well the model fits the data and how much of the variance in the target variable is explained by the features. The MSE score indicates how close the predicted values are to the actual values, with a lower MSE indicating better performance. This gave an

idea of how well each model is able to predict the earnings of the athletes based on these selected features. Performance of each model was then compared to determine which one was the most accurate and useful for further analysis. The next step was to see which of the three features of that went into the model had the greatest effect on predicting by running the best performing estimator except with one feature per model rather than all of them combined. The same performance measures as before were used to observe the importance of the feature when predicting.

```
R2: 0.76180210312553
MSE: 177.87332475495037
```

(Figure 19)

```
R2: 0.7668894063191765
MSE: 174.07440148583663
```

(Figure 20)

```
R2: 0.629270212517364
MSE: 276.8409828571429
```

(Figure 21)

The results obtained after utilizing the three ML models were remarkably promising. All three of the models yielded an R2 value above 0.60, which was quite impressive. After rounding off, it was observed that the GBR model (Figure 20) obtained the highest R2 value of 0.77, followed closely by the RFR model (Figure 19), which obtained an R2 value of 0.76. On the other hand, the KNNR model (Figure 21) performed comparatively worse, obtaining a value of 0.63. The MSE values were ranked in the same order as well, which indicates that the average error was also the smallest with the most accurate model out of the three. Advancement to the next step can begin, which involved fine-tuning the GBR model to achieve even better results. Therefore, the GBR estimator was used to determine which feature matters the most when predicting by using the same performance metrics for each respective model.

```
Nationality GBR R2: -0.00016451439980036398
Nationality GBR MSE: 746.8688423053851
```

(Figure 22)

```
Sport GBR R2: 0.12481775370592896
Sport GBR MSE: 653.5388345467675
```

(Figure 23)

```
Year GBR R2: 0.4795057013726358
Year GBR MSE: 388.67703127385664
```

(Figure 24)

When analyzing the features individually, it appeared that Year (Figure 24) was by far the best predictor out of the three, with an R2 value of 0.48 when rounded, indicating that it explains nearly half of the variance in athlete earnings. Sport (Figure 23) was the next closest with an accuracy of 0.12, suggesting that it is a weaker predictor compared to Year. Interestingly, Nationality (Figure 22) had a negative R2 value, indicating that when used on its own, it may be more likely to incorrectly predict an athlete's salary. However, it is important to note that Nationality still played a role in predicting athlete earnings when used in combination with other features such as Year and Sport. Overall, it appears that Year was the most important feature when it comes to predicting athlete earnings, followed by Sport and Nationality.

Results of Fine-Tuning the Gradient Boosting Regressor (GBR) Model

To further improve the predictive power of the GBR model, it's hyperparameters can be tweaked to find the optimal combination. Specifically, the learning rate, number of estimators, minimum samples split, and max depth were experimented with. Learning rate controls the contribution of each tree in the ensemble, the number of estimators is the number of sequential trees to be modeled, represents the minimum number of samples required to split an internal node, and max depth is the maximum depth of each decision tree. To perform this optimization, the GridSearchCV object from the Scikit-Learn Python Package was utilized, which allowed search over the specified hyperparameter grid and evaluated each combination using cross-validation. After fitting the GridSearchCV object to the training data, the best estimator was obtained, which was the GBR model with the optimal hyperparameters. Next, the tuned model was used to make predictions on the test set and calculate its performance metrics. Finally, the best hyperparameters were printed out and the same performance metrics used previously to evaluate the initial models. By doing so, best possible model for predictive purposes can be assured.

```
Best parameters: {'learning_rate': 0.05, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100}
R2: 0.7666826230187335
MSE: 174.2288160866212
```

(Figure 25)

According to Figure 25, fine-tuning the GBR model did not decrease the performance of the model, which is great news. However, the process of fine-tuning the model did not result in any significant improvement of R2 or MSE either. Even though the fine-tuning process did not provide much improvement, it is still important to note that the model was capable of making relatively accurate predictions. This means the GBR model can be used with confidence to make predictions on future earnings of athletes, as it can still provide a reasonable estimation of their potential earnings.

Conclusion

In this project, the primary objective was to conduct a thorough analysis of a dataset containing information on the highest-paid athletes and identify any patterns or correlations that may explain why certain athletes earn more than others. The analysis aimed to delve into the intricacies of the professional sports industry and identify the crucial factors that contribute to the high earnings of top athletes. By scrutinizing various parameters such as sport, year of payment, and nationality, the project attempted to unveil the reasons why some athletes earn more than others. The exploration of the dataset yielded valuable information to help achieve these goals. The correlation plots identified that the features in the dataset had high correlations to one another, and revealed the features that would be most relevant for helping major sports organizations and fans observe and predict the salaries of some of the highest paid athletes in the world. Once the salary was determined to be the main target for the model being built, the preprocessing analysis revealed possible biases and anomalies in the Sport, Name, and Nationality features of the dataset. This would be addressed by moving forward with just the year the athlete was paid the salary, sport the athlete played, and the nationality of the athlete in question. The analysis of the figures and the performance evaluation of the ML models provides valuable insights into the sports industry and its potential for prediction using machine learning techniques. The Gradient Boosting Regressor (GBR) model was found to be the most accurate, with an R2 value of 0.77. The most important feature for predicting athletes' earnings was found to be Year, followed by Sport and Nationality. The potential insights gleaned from this project could prove to be invaluable for the sports industry and could be leveraged by teams, organizations, and fans alike to make more informed decisions when it comes to hiring and paying athletes. In conclusion, the findings of this project could have far-reaching implications and serve as a crucial resource for shaping future decision-making in the professional sports industry. The development of a predictive model that can identify a player's sport based on other parameters in the dataset will be a crucial element of this project. The outcomes of this project can make a significant contribution to the field and potentially shape the future directions for the industry and the fervent fans who are keenly interested in what factors contribute to the salary of their favorite athletes.

Reference Links

- Forbes Dataset
  - https://www.kaggle.com/code/suvojithaldar/analysis-of-world-s-richest-athletes
- Random Forest Regressor Info
  - https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
- Gradient Boosting Regressor Info
  - https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
- K-Nearest-Neighbors Regressor Info
  - https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

Appendix

- Github Link
  - https://github.com/willechr/willechr_492