PREDICTIVE MODEL

The probabilities of our customers

SPORT DOJO
SUMMER 2023

Lesly Zelaya
l.m.zelaya@wustl.edu

Table of Contents

Abstract

In an era when data-driven decision-making has become mainstream, this report presents a predictive model using data analytics and machine learning techniques to identify potential participants in the new Sport Dojo sports program. This program's goals and individuals in sporting activities are about engaging in and promoting an active and healthy lifestyle. It helps guide young people who are inspired to level up their coaching experience and or start their coaching career by providing them the resources to practice and engage in the game of the sport they are passionate about. Using historical data and advanced sampling techniques, the predictive model helps target potential participants most likely to participate in the intervention.

This process begins with extensive contextual data collection, including a few demographics, interests, and historical associations with sport-related activities. It conducts intensive preprocessing and feature engineering to ensure the information collection is excellent and suitable for the specimen. Exploratory data analysis (EDA) reveals valuable insights, enabling a better understanding of patterns and trends in a data system.

The methodology section outlines a selection of machine learning algorithms designed to predict participant engagement. Two specific algorithms utilized, including logic regression and simple data prediction using Python, were explored, and optimized in the best way possible to obtain optimal performance. To decide which method is best, both methods were compared to one another to interpret and break down even further. The training and test data were labeled to ensure it was appropriate for the type of data it was handled, with the addition of observational learning required to increase the accuracy of the prediction.

The results show the model's effectiveness in identifying potential participants in Sport Dojo's initiatives. Performance metrics, such as precision, recall, and F1-score, were used to underscore the model's ability to distinguish likely participants from the larger pool of individuals. The interpretability of the model is explored through feature importance analysis, shedding light on the factors driving participant engagement predictions.

The discussion delves into the broader implications of the predictive model within the context of Sport Dojo's initiative. Ethical considerations, potential biases, and challenges faced in data and model predictions are addressed, emphasizing the need for fairness and inclusivity. The report concludes by highlighting the impact of the predictive model on optimizing resource allocation, enhancing program engagement, and fostering a healthier community through increased sports and coach participation.

Looking ahead, the report suggests avenues for future work, such as incorporating real–time data feeds, refining the model interpretability, and extending the approach to other domains; in a world increasingly reliant on data-driven insights, this predictive model stands as a testament to the power of combining data analytics and machine learning to shape more effective and targeted community engagement strategies.

Introduction

The developed predictive model encompasses a dataset containing information from over 300 participants who were surveyed regarding their past sports engagement, age, gender, education, participation frequency, indices, and their past preference for working alone or in small or large groups as seen in figure 1. Based on these responses, the dataset includes a calculated percentage of independence and interdependence, reflecting the individual's inclinations.

These calculated metrics were the foundation for predicting whether participants possess the predisposition and characteristics to embark on coaching careers or volunteer opportunities to inspire children to embrace and excel in their beloved sports. Addressing the core objectives of this project, the model aids in the identification of potential candidates for Sport Dojo's sports initiatives.

The key focus lies on the interdependence percentage within the dataset, as higher values indicate a propensity for coaching or volunteering due to a background and disposition conducive to guiding and encouraging young athletes. By evaluating individuals' backgrounds, sport engagement, and cooperative tendencies, the model helps anticipate their suitability for participating in Sport Dojo's initiatives.

[7]:

	index	importance	sport	frequency	how_alone	how_small_groups	how_big_groups	independence	interdependence	sex	age	education
0	134	7	Martial Arts Tricking	5	False	True	False	4.000000	5.833333	m	26	7
1	135	7	Martial Arts Tricking	5	True	True	True	5.416667	6.583333	m	25	5
2	137	7	Basketball	4	True	True	True	4.166667	5.916667	m	27	8
3	139	7	Parkour	7	True	True	False	4.666667	5.583333	m	18	4
4	140	7	Parkour	6	False	True	False	4.833333	5.833333	m	19	7

Figure 1: The first 5 columns of the data called Sport_e2.

Column	Interpretation
Index	Number of completed questionaries
Importance	How important is your sport for you? Scale 1 (low) – 7 (high)
Sport	Which sport do you most identify with?
Frequency	How often do you play or practice the sport?
How alone	Do you prefer practicing alone
How small groups?	Do you like practicing in small groups?
How big group?	Do you like to practice in large crowds?
Independence	Mean value of independence from 1 (low) – 7 (highest)
Interdependence	Mean value of interdependence from 1 (low) – 7 (highest)
sex	Gender of participant
Age	Age of participant
Education	Highest degree of education. Scale: 1 (low) – 9 (highest)

Figure 2: A key to understand the metrics and data set and variables used.

Data Collection and Preprocessing

The dataset in Figure 1 originates from Kaggle, a free data science platform. Kaggle is renowned for its open-source datasets contributed by community members, fostering machine learning and data science. When collecting this data, the primary key points were focused on sports experience, psychology, and the interplay between sports and personality. Individuals who actively volunteer and enhance their skills by aiding others often exhibit a caring personality that inspires improvement. Analyzing backgrounds and personalities aims to uncover patterns shaping this altruistic trait.

```
[12]: data.isnull().sum()

[12]: index                0
      importance           0
      sport                0
      frequency            0
      how_alone            0
      how_small_groups     0
      how_big_groups       0
      independence         0
      interdependence      0
      sex                  0
      age                  0
      education            0
      dtype: int64
```

Figure 3: Data checked for any null values in python Jupyter notebook.

Before testing and training the data, a meticulous examination was conducted to detect potential issues such as replication, missing values, or null entries that might disrupt the predictive model. Fortunately, the data was found to be both flawless and coherent. In addition, the data was analyzed thoroughly to fully understand the values and some statistics on what this data was calculating. Specific columns featured true and false responses, which were transformed into binary values (1 for true and 0 for false). For the "Sport" column, strings were converted into integer representations, assigning numerical values to each sport.

The dataset contained essential pinpoint variables, yet it posed a challenge due to the prevalence of true/false responses instead of continuous numerical values. A solution was devised to overcome this challenge by encoding the responses as binary numbers (1 and 0). This approach accommodated the two possible responses and facilitated their integration into the training and test sets, ensuring minimal impact on the predictive model.

```
> sport_e2$sport <- as.factor(sport_e2$sport)
> sport_e2$sex <- as.factor(sport_e2$sex)
>
> sport_e2$how_alone <- ifelse(sport_e2$how_alone=="True",1,0)
> sport_e2$how_alone<-as.factor(sport_e2$how_alone)
> sport_e2$how_small_groups<-ifelse(sport_e2$how_small_groups== "True",1,0)
> sport_e2$how_small_groups<-as.factor(sport_e2$how_small_groups)
> sport_e2$how_big_groups<-ifelse(sport_e2$how_big_groups=="True",1,0)
> sport_e2$how_big_groups<-as.factor(sport_e2$how_big_groups)
> str(sport_e2)
Classes 'tbl_df', 'tbl' and 'data.frame':       399 obs. of  12 variables:
 $ index           : num  134 135 137 139 140 142 143 144 150 151 ...
 $ importance      : num  7 7 7 7 7 7 7 7 7 7 ...
 $ sport           : Factor w/ 59 levels "Acrobatics","Aerobics",..: 35 35 5 40 40 22 40 49 16 40 ...
 $ frequency       : num  5 5 4 7 6 3 6 5 8 5 ...
 $ how_alone       : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ how_small_groups: Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ how_big_groups  : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
 $ independence    : num  4 5.42 4.17 4.67 4.83 ...
 $ interdependence : num  5.83 6.58 5.92 5.58 5.83 ...
 $ sex             : Factor w/ 2 levels "f","m": 2 2 2 2 2 2 2 2 2 2 ...
 $ age             : num  26 25 27 18 19 18 16 25 17 17 ...
 $ education       : num  7 5 8 4 7 9 9 6 4 9 ...
```

Figure 4: Snip of the code that converted columns with string into binary factors (True =1, False = 0). Also, with the gender

Feature Engineering

During the variable selection process for constructing the predictive model, initial emphasis was placed on identifying the variables that exhibited the strongest correlations with the target variable. The chosen target variable, in this case, was the "interdependence" factor. As the logistic regression analysis was undertaken, the interdependence variable initially existed as a solitary numerical value with a lone factor. It needed to be transformed into a multifaceted factor with multiple attributes that the model could predict to make it a target variable.

Consequently, the target variable underwent a conversion, resulting in the incorporation of multiple factors. The variables that demonstrated the most robust correlations with the predictive model encompassed importance, age, education, frequency, and index within the logistic regression framework. The dataset was partitioned into two sections: one designated for training purposes and the other for testing. The training dataset represented 80 percent of the entire dataset, while the testing data comprised 20 percent.

The rationale behind selecting these variables was to achieve more accurate predictions by considering factors that could significantly influence, leading to higher outcomes within the interdependence sector. It was deduced that individuals exhibiting a greater frequency of engagement and a heightened degree of investment in their respective sports activities provided the closest estimations capable of impacting the interdependence sector. In the logistic regression model context, the dataset, bifurcated into training and testing subsets, was transformed into binary representations. This binary transformation facilitated the allocation of instances into one of two specific data subsets. This process delineated the procedure by which the training and testing datasets were associated with columns from the original dataset.

```
In [19]: from sklearn.linear_model import LinearRegression
         from sklearn.model_selection import train_test_split

In [223]: train = data.drop(['index','importance','sport','education','sex'], axis=1)
          test = data['interdependence']

In [224]: train_encode = pd.get_dummies(train)

In [225]: X_train, X_test, y_train, y_test = train_test_split(train_encode, test, test_size=0.4, random_state=3)
```

Figure 5: A snip of a code that displays what factors were included when testing and training the data in the python predictive model.

```
#Partition Data - Train (80%) & Test(20%)
#80% percent will be taken into my rows and the rest will be th test data.
#Take a sample first
set.seed(1234)
#help(sample)
indexSet <-sample(2,nrow(sport_e2),replace = T, prob = c(0.8,0.2))
train <- sport_e2[indexSet==1,]
test <- sport_e2[indexSet==2,]
indexSet
#Logistic  regression model
help(glm)
mymodel <- glm(as.factor(interdependence) ~ index+importance+sport+age+education+frequency, data = train, family = 'binomial')
summary(mymodel)
```

Figure 6: Snip of code that represents train and set model being prepared for

Exploratory Data Analysis (EDA):

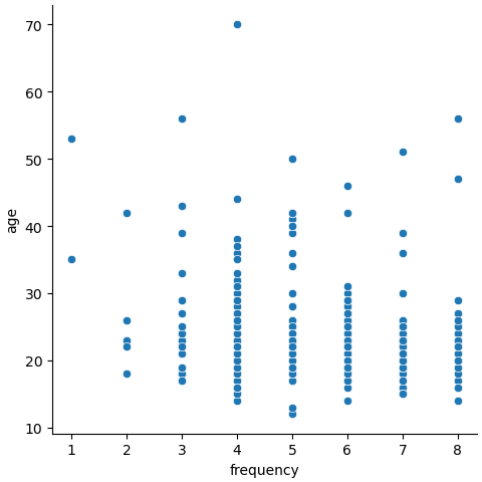

Figure 7: 2-dimensional graph showing relationship between age and frequency.

The visualization has effectively illustrated discernible trends and correlations among the selected variables. Notably, it showcases a distinct relationship between age and the frequency of engagement. Specifically, individuals aged 20 to 40 exhibit heightened levels of activity and involvement in sports-related activities. This age bracket is when individuals are more active, participating in or instructing physical endeavors. Furthermore, the visualization reveals a compelling linkage between the age range of 20 to 40 and a heightened degree of interdependence. This visualization suggests that during this phase of life, there is a greater inclination among individuals to engage in activities that involve helping others learn or practice skills they are passionate about. This phenomenon seems driven by a voluntary desire to contribute to others' growth.

To comprehensively explore the correlations between frequency, age, and interdependence, we employed visualization techniques and graphed the data. This process enabled us to discern a consistent pattern where individuals with higher engagement frequency in sports also exhibited a greater inclination towards interdependence. This inclination signifies a propensity for voluntarily aiding others. Regarding age, it is evident that the peak frequency of both high frequency and interdependence is concentrated within the 20 to 35 age range, with the broader range of 20 to 40 still showing considerable clustering. These interrelated patterns and correlations collectively suggest that these specific connections could serve as optimal variables for incorporation into the predictive model.

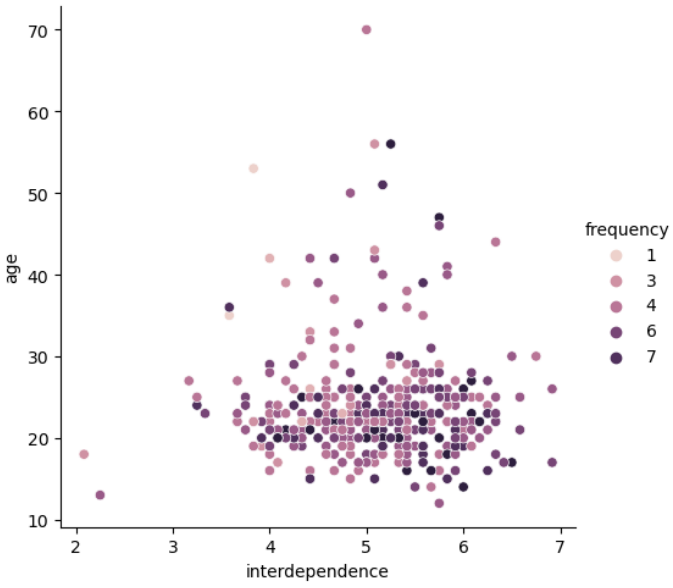

Figure 8: 2-dimensional graph representing the relationship between interdependence, age, and frequency.

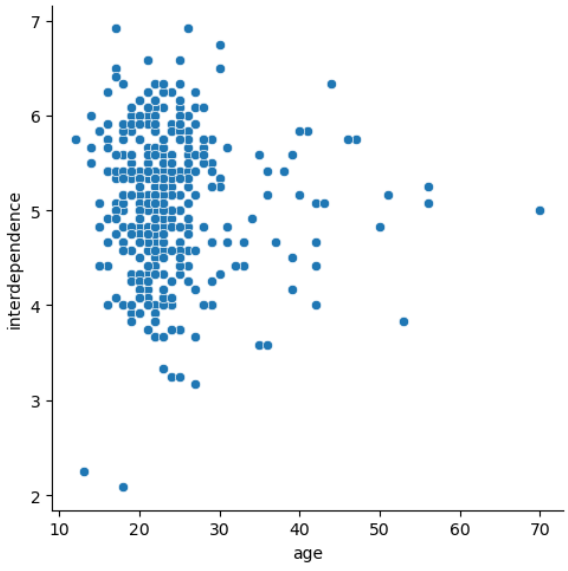

Figure 9: 2-demensional graph representing the relationship between interdependence and age.

Model Training

Two distinct approaches were employed during the model training phase: linear regression and logistic regression. Logistic regression was initially assessed to identify significant variables with the utmost impact. Among these variables, frequency and age emerged as the most influential factors. Consequently, the remaining variables were excluded to enhance the model's precision. Notably, the significance assessment was executed using a specific algorithm named "MASS," integrated into the platform, in contrast to the manual approach adopted for linear regression using Python. Once the model completed training and the relevant columns were determined following the significance assessment, efforts were directed toward achieving the optimal line fit for the final model. Additionally, the dataset's initial entry or survey data was manually processed. The formula employed for predicting the first entry's likelihood was as follows:

$$\text{Log (Odds)} = \text{Log (1 - p)} = \text{"Logit"} \rightarrow -438.854 - 2.138 * (\text{index}) + 40.120 * (\text{frequency})$$

Subsequently, another equation was utilized to calculate the probability of an individual becoming a participant or not:

$$\text{Probability} = 1 / (1 + \exp(-y))$$

This sequence of steps culminated in deriving the probability concerning an individual's potential participation and the model's efficacy in prediction, drawing upon the chosen variables. Despite a significant standard deviation error within the model, these variables remained pivotal. Despite potential gaps throughout the model-building process, they were instrumental in generating a pragmatic estimate based on the provided variables. While a definitive prediction was not achieved, the model yielded estimations aligned with the utilized equations. The comprehensive calculation applied within the system was as follows:

$$y = -438.854 - 2.138 * 134 + 40.120 * 5$$

The p-value falling below 1 was crucial, affirming that predictions remained within reasonable bounds and did not venture into overconfidence.

```
#Improving model..
library("MASS")
mymodel2<-stepAIC(mymodel)
summary(mymodel2)
#Log(Odds)=Log(p/1-p))="Logit"
            #-438.854-2.138*(index)+ 40.120*(frequency)
head(train)
#Logit
y=-438.854-2.138* 134 + 40.120* 5
#Probability
1/(1+exp(-y))

train$interdependence
mymodel2$fitted.values
```

Figure 10: Snip of code representing mathematical equations t0 solve prediction and probability in logistic regression.

Results

The outcomes obtained from this model have effectively established correlations among the variables under examination. These trends collectively indicate that individuals who frequently engage in their chosen sport and possess experience in a team-oriented environment, where mutual dependence fosters success and skill acquisition, exhibit a heightened likelihood of possessing an interdependent personality trait. This predisposition positions them as potential candidates to assume roles as coaches or voluntary members. In such capacities, they are inclined to inspire young individuals to participate in or try the sport.

The model's analysis reaffirms the centrality of variables like importance, age, and frequency in predicting an individual's potential as a participant in Sport Dojo initiatives. Notably, individuals aged between 20 and 35, who exhibit higher frequencies of sport engagement, tend to showcase heightened levels of activity, and have likely contributed to team dynamics or held leadership roles in sports settings. These individuals are more inclined to influence and assist others in joining the sport, driven by their interdependent outlook from prior experiences.

Consequently, our findings underscore the significance of passionate engagement in sports and a commitment to community assistance as critical indicators for identifying potential participants in Sport Dojo. These conclusions are rooted in past experiences, individual backgrounds, and interdependent personality traits.

Discussion

The results of our analysis highlight that the model successfully identified patterns and trends within the dataset. These patterns revealed a positive correlation between variables like frequency of engagement in sports, age, and the inclination towards interdependent behavior. This alignment supports the notion that individuals actively participating in sports and exhibiting interdependent tendencies are more likely to become coaches or contribute to sports initiatives. However, while the model captured these relationships, there might be room for enhancing the predictive power by considering additional variables. For instance, incorporating socio-economic backgrounds, prior coaching experience, or psychological factors could strengthen the model's predictive capabilities. These variables could provide a more comprehensive understanding of factors contributing to an individual's potential participation in Sport Dojo initiatives. In summary, while the current model effectively demonstrated existing patterns and trends, a more robust prediction could be achieved by introducing a broader set of variables that better capture the multifaceted nature of individuals' motivations and backgrounds.

Conclusion

The core findings of this predictive model highlight significant relationships among multiple variables. These connections reveal both independent and dependent variables that mutually influence each other. Notably, the model underscores a robust correlation between age, prior sports experience, and personality inclined towards solid interdependence. These attributes collectively identify individuals who could play pivotal roles in sport dojo's initiatives. These individuals hold critical roles due to their heightened potential for pursuing coaching careers or rekindling their coaching experiences. Additionally, they encompass individuals who possess the aspiration but need more roadmap to enter this domain. Their profound commitment to community success, fueled by a solid interdependent mindset, drives them to inspire and guide others toward achievements. The predictive model marks the inception of a vital approach for comprehending and projecting future participants in sport dojo programs. This strategic foundation empowers sport dojo to discern participants' demands and requirements, thereby enhancing its offerings and organizational trajectory. It represents the initial stride towards elevating our brand and achieving broader horizons than before.

Future work

While the predictive model did yield specific results, a significant gap was evident in its outcomes. These findings gave us preliminary insights and a general direction, allowing us to form expectations. However, these outputs still need to be improved to rely upon heavily. The model is essentially an experimental prototype, necessitating enhancement by integrating more robust data and variables.

This model marked a nascent stage, guiding us toward fresh perspectives and avenues for data exploration. Leveraging various machine learning algorithms is crucial to achieving optimal predictions. The experimental prototype employed logistic and linear regression, yielding promising outcomes. Nevertheless, the pronounced gaps in results can be attributed to challenges in sourcing alternative data implementations and navigating the constraints of existing variables.

To address this, it is strongly advised to iterate and develop new models every six months. This approach will yield a more profound comprehension of client demands and illuminate potential trajectories for Sport Dojo's future initiatives. As we refine our models, we can expect a more refined understanding of our data and a more straightforward path for the growth of Sport dojo.

Methodology:

In order to assess the predictive model's outcomes and determine potential disparities or similarities, a combination of logistic and linear regression methods was employed. This approach was selected due to the categorical nature of the predictive model's outputs, which necessitated using regression techniques suitable for categorical dependent variables.

Logistic Regression:

Strengths:

- **categorical Predictions:** Logistic regression is particularly suited for predicting categorical outcomes. It provides probabilities for the occurrence of specific categories, making it well-aligned with the nature of the predictive model.
- **Interpretability:** The coefficients obtained from logistic regression allow for unambiguous interpretation. This helps in understanding the direction and magnitude of the effects of independent variables on the categorical outcome.

Weaknesses:

- **Assumption of Linearity:** Logistic regression assumes a linear relationship between independent variables and the log odds of the categorical outcome. This assumption might only sometimes hold, potentially affecting the model's accuracy.
- **Limited to Binary Outcomes**: While extensions exist for multinomial and ordinal outcomes, logistic regression is primarily designed for binary outcomes. Handling multi-class categorical outcomes can be complex.

Linear Regression:

Strengths:

- **Wide Applicability:** Linear regression is a versatile method for predicting continuous outcomes. By extension, it can be adapted for categorical outcomes by employing techniques such as one-hot encoding.
- **Simple Interpretation**: Like logistic regression, linear regression provides interpretable coefficients that elucidate the relationship between independent variables and the outcome. This can aid in drawing meaningful insights.

Weaknesses:

- **Inaccurate for Categorical Outcomes**: Linear regression is not inherently suited for categorical outcomes, as it may lead to predictions outside the valid range for probabilities (0 to 1) when applied to such scenarios.
- **Sensitive to Outliers:** Linear regression is sensitive to outliers, which can disproportionately influence the model's coefficients and predictions. This can lead to inaccurate results, especially when dealing with categorical outcomes.

In conclusion, the combination of logistic and linear regression techniques was chosen to gain insights into the predictive model's outcomes, considering the categorical nature of the predictions. While logistic regression excels in handling categorical outcomes, linear regression was adapted with appropriate encoding. It is important to acknowledge the strengths and weaknesses of both methods to make informed decisions about their applicability in this context.

Appendices

Results from the logistic regression in R:

```
> pred1<-ifelse(p1>0.5,1,0)
> tab1 <- table(Predicted = pred1, Actual = train$interdependence)
> tab1
         Actual
Predicted 2.08333333333333 2.25 3.16666666666667 3.25 3.58333333333333 3.66666666666667 3.75 3.83333333333333 3.91666666666667  4 4.08333333333333 4.16666666666667
        0                1    0                0    0                0                0    0                0                0  0                0                0
        1                0    1                1    1                1                1    4                4                3 13               10                3
         Actual
Predicted 4.25 4.33333333333333 4.41666666666667 4.5 4.58333333333333 4.66666666666667 4.75 4.83333333333333 4.91666666666667  5 5.08333333333333 5.16666666666667
        0    0                0                0   0                0                0    0                0                0  0                0                0
        1    3                6               10   3                8               24    5               18               10 12               11                9
         Actual
Predicted 5.25 5.33333333333333 5.41666666666667 5.5 5.58333333333333 5.66666666666667 5.75 5.83333333333333 5.91666666666667  6 6.08333333333333 6.16666666666667
        0    0                0                0   0                0                0    0                0                0  0                0                0
        1   11               17               18  10               15                9   15               11                5  7                6                3
         Actual
Predicted 6.25 6.33333333333333 6.5 6.58333333333333 6.75 6.91666666666667
        0    0                0   0                0    0                0
        1    6                5   2                2    1                1
>
```

Figure 11: This is a confusion matrix where it calculates how much the machine was able to predict it accurately in which it is separated into Actual and predicted.

Results From linear regression in Python:

```
In [229]: pred = regr.predict(X_test)

In [230]: pred
Out[230]: array([5.58333333, 4.83333333, 5.33333333, 6.        , 4.5       ,
                 4.33333333, 4.33333333, 5.33333333, 5.08333333, 5.33333333,
                 5.5       , 4.5       , 6.25      , 4.91666667, 4.66666667,
                 4.91666667, 4.41666667, 4.33333333, 5.33333333, 4.66666667,
                 5.83333333, 5.25      , 5.75      , 6.16666667, 5.41666667,
                 4.        , 5.91666667, 3.83333333, 4.58333333, 5.        ,
                 5.16666667, 5.41666667, 5.75      , 4.66666667, 5.25      ,
                 4.41666667, 5.08333333, 5.        , 5.08333333, 6.16666667,
                 5.        , 5.66666667, 4.91666667, 5.75      , 5.58333333,
                 3.75      , 3.33333333, 5.66666667, 5.41666667, 5.66666667,
                 4.66666667, 5.25      , 5.08333333, 5.        , 5.25      ,
                 5.        , 4.83333333, 4.75      , 4.58333333, 6.33333333,
                 5.75      , 5.25      , 5.25      , 4.41666667, 5.16666667,
                 4.25      , 5.58333333, 4.5       , 4.08333333, 5.33333333,
                 5.83333333, 5.08333333, 4.41666667, 4.25      , 6.16666667,
                 6.25      , 5.58333333, 5.83333333, 5.33333333, 3.83333333,
                 5.25      , 5.16666667, 5.5       , 4.83333333, 4.83333333,
                 5.16666667, 5.75      , 4.91666667, 4.25      , 5.75      ,
```

Figure 12: Prediction made by linear regression as a result.

```
In [231]: regr.score(X_test, y_test)
Out[231]: 1.0
```

Figure 13: A snip of the accuracy score of the prediction model being done with linear regression.

References:

Data source:

The data set used for this project was obtained from Kaggle, a platform for data science and machine learning. The specific data set used is titled as “Independence and interdependence in Sports” and be accessed at the following link:

https://www.kaggle.com/datasets/harti28/independence-and-interdependence-in-sports

Citation:

Author: HARTI28. “Independence and interdependence in Sports” Kaggle, 2023. https://www.kaggle.com/datasets/harti28/independence-and-interdependence-in-sports

Logistic Regression Model:

Citation:

The logistic Regression Model was developed Environment for R. RStudio,

PBC, Boston, MA, 2021. https://posit.co/resources/videos/what-is-rstudio-connect/

Linear Regression Model

The linear regression model was created using Jupyter Notebook, an open-source web application that allows the creation and sharing of documents that contain live code, equations, visualizations, and narrative text.

Citation:

Kluyver, Thomas, et al. “Jupyter Notebooks – a publishing format for reproducible computational workflows”. In positioning and Power in Academic Publishing: Players, Agents, and Agendas, IOS Press, 2016. https://jupyter.org/