Evaluating Basketball Shoes: Attribute Comparison, Price Dynamics, and Player Position Prediction

Jiachen(Chris) Lu, Guanhua(Martin) Chen, Yiqiao(Jeremy) Peng

Research Question and Results Summary:

Research Question #1:

1. How can a radar graph representation be used to visually compare and evaluate the attributes of basketball shoes, thereby providing a comprehensive assessment of their relative strength and weaknesses? How does each brand prioritize what attributes and how will it affect user rating?
    a. After visualizing the correlation between overall rating and attributes of each shoe, we have found certain attributes contribute to an increase in overall rating. In addition, with a radar graph to represent each brand's attribute score and comparing them, we were able to provide a comprehensive assessment of each brand's strength and weakness.

Research Question #2:

1. How does the MSRP of basketball shoes vary based on their Overall rating and Brands?
    a. After analyzing the data visualization, it can be inferred that Nike's sneakers offer the highest cost-effectiveness based on their MSRP, overall rating, and sneaker production quantity.

Research Question #3:

1. How accurately can we predict the player position (guard, wing, big) performance score for each shoe based on the attributes of basketball shoes using machine learning algorithms? (Decision Tree)
    a. After applying the Decision Tree Classification model to our training data and fitting it to our testing data, it can be inferred that we can predict the player position (guard, wing, big) performance score for each shoe based on the attributes of sneakers with over 70% certainty.

Motivation and Background:

Research Question #1:

- Context: Understanding the relationship between the attributes of basketball shoes and their overall rating and market reception can provide valuable insights to shoe manufacturers, retailers, and consumers. Manufacturers can design shoes that meet consumer expectations and preferences, improving product satisfaction and market success. Retailers can predict market reception based on attributes, aiding in inventory management and marketing strategies. Consumers can make more informed purchasing decisions by considering their prioritized attributes and their impact on overall rating and market perception.

Research Question #2:

- Context: Understanding the relationship between MSRP, overall rating, and brands of basketball shoes provides valuable insights into the dynamics of the basketball shoe market. It involves analyzing pricing strategies and brand positioning, assessing

perceived value based on overall rating, and identifying market trends and consumer preferences.

Research Question #3:

- Context: Basketball shoes play a crucial role in a player's performance on the court, with different player positions having unique physical requirements and playing styles. Optimizing performance, minimizing injuries, and enhancing player experience require specific shoe attributes. Attributes like traction, cushioning, lockdown, lateral stability, torsion support, weight, durability, and ventilation hold varying degrees of importance for different positions.

Dataset:

https://docs.google.com/spreadsheets/d/1LmMuhapV3gaJnHSd5S6ckeK3D5BZRiHejOQNvrB6dV4/htmlview#

Our data is from the Reddit community r/BBallShoes, which is a community that publishes basketball shoe performance reviews. It has many different variables to discuss which I included below. This community helps a lot of basketball players, from amateurs to professionals to decide what shoes best fit their playing style. It also publishes PDF-like reports for individual shoes that analyze and compare them with different shoes.

Data includes Brand, Shoe name, attributes (traction, cushioning, lockdown, lateral stability, torsion support, weight, durability, ventilation), average attribute score, overall user score, guard score, wing score, big score, outdoor suitability score, number of reviews, year of releases,

MSRP, weight per shoe for US9 size, heel-to-toe offset, recent release determination, Asian brand determination, budget shoe determination, sizing recommendation, and sizing notes.

Methodology:

All Research Questions Methodology:

- Our first step is to clean the data into usable format. With our original csv file, there are several columns of data that will not be used in our research questions. For example, miscellaneous data that contains year of release, weight per shoe, heel-to-toe offset, sizing recommendations, etc. In our research question, we will be using Market Sales Retail Price (MSRP) but its original type is string, hence we will be converting them into integer type to be able to modify such data.

Research Question #1 Methodology:

- Our second step is to create scatter plots with regression lines that compares each shoes' attributes and their overall rating relationships. Since we are comparing 8 attributes to overall rating, we have defined a function that takes one attribute and overall rating as parameters. This way, we can loop over to display all attributes vs overall rating plots.
- In the radar graph, we took the average attribute score of each brand in the dataset. Since we are comparing 8 attributes for each brand, we created an octagon that each angle represents an attribute score. Since most attribute scores range from 55-95, we set ranges to make better visualization for the radar plot, so it can show which brand has their strength and weakness in each attribute.

- In the radar brand comparison graph, we decided to compare the top four brands that have the most data in the csv file. In the given dataset, we have 28 brands in total, and to compare each brand back to back, we have 378 patterns. Since some brands have less data than other brands, we decided to only compare the top four brands (has most data). In a similar manner, we created octagons for each brand to compare and overlap them to visualize which brand has a higher score than the other brand we are comparing.

Research Question #2:

- In this question, the methodology involves creating a scatter plot using Plotly Express to visualize the relationship between the average overall rating and MSRP(Manufacturer's Suggested Retail Price) by brand. The scatter plot represents each brand as a data point, with the MSRP on the y-axis and the average overall rating on the x-axis. The size of the data points is used to convey additional information, which is the sneaker production quantity here. The color of the data points can be used to differentiate between different brands.

Research Question #3:

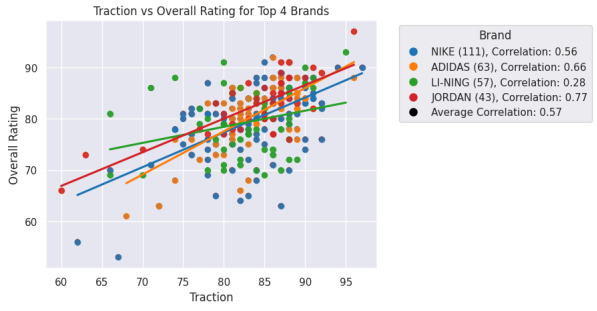
- In this question, the methodology involves training and testing a Decision Tree classifier model from scikit-learn. In this machine learning, the feature includes attributes of sneakers which are lockdown, cushioning, traction, ventilation, durability, weight, torsion support, and lateral stability. The target variable(label) is the proper position of sneakers, which are guard, wing, and big. The dataset is divided into 80% of training and 20% of testing. The Decision Tree classifier model from sklearn is then trained using the training data. Then the model's performance is evaluated by computing the accuracy of the predictions on both the training and test datasets.

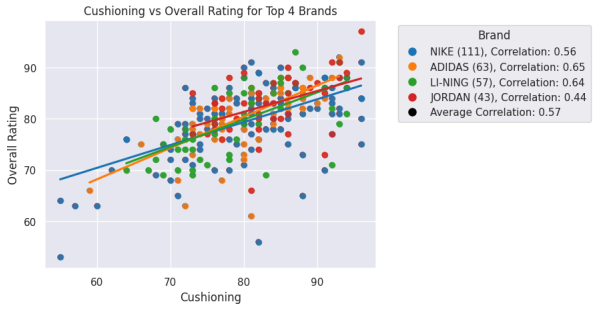
- Additionally, the decision tree visualization is created to provide a graphical representation of the model's decision-making process, which allows a better interpretation of the model's structure, rules, and how it splits. We also create a user input interface, allowing people to input the attributes they have or want for a sneaker. Therefore, our model can predict what is the proper position for that sneaker.

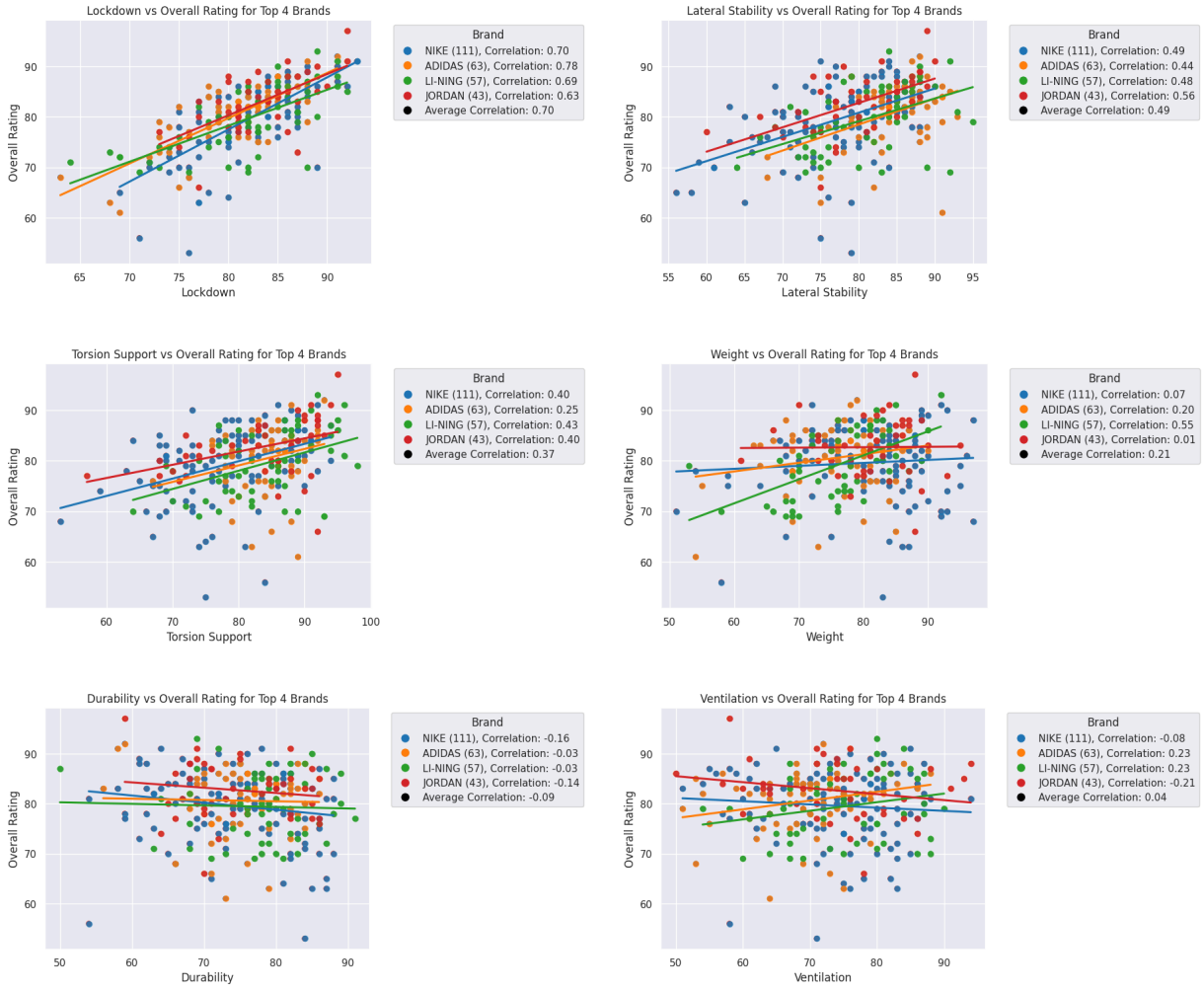
Results:

Research Question #1 Part 1: How does each brand prioritize what attributes and how will it affect user rating?

Not every attribute increases in the overall rating. For example, *Traction, Cushioning, and Lockdown* are some attributes that significantly increase in overall rating with an average correlation of 0.5+. *Lateral Stability, Tursion support, and Weight* are attributes that fairly contribute to an increase in overall rating with a correlation fall in the range of 0.2-0.5. Lastly, *Durability* and *Ventilation* are two attributes that do not contribute to an increase in overall rating with a correlation close to 0. After visualizing the relationship between each attribute and overall rating by users, we can conclude that users prioritize certain attributes such as *Traction, Cushioning, and Lockdown* over other attributes in choosing which basketball they wear. This not only benefits the shoe owners, it also benefits the companies who manufacture and design new shoes, because in the future, they can put more effort in certain attributes to increase user experience.



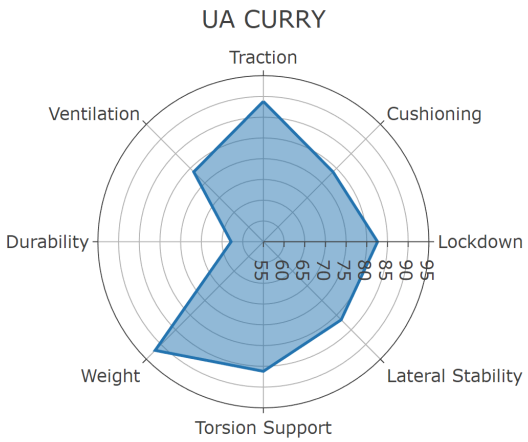


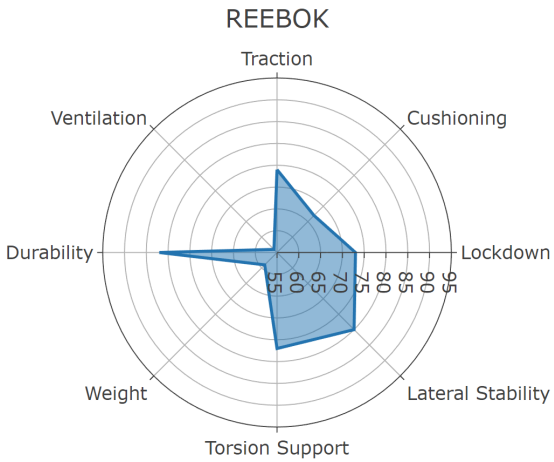


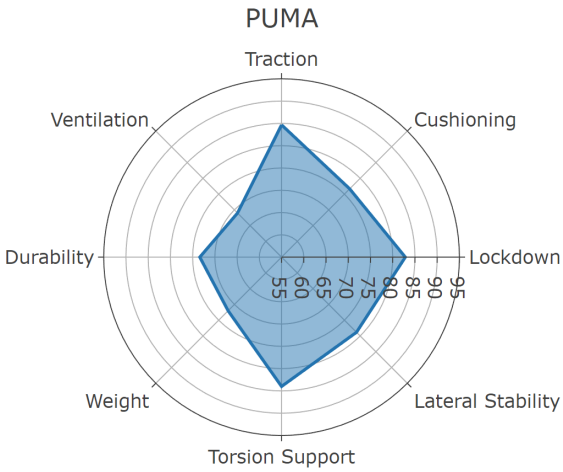
Research Question #1 Part 2: How can a radar graph representation be used to visually compare and evaluate the attributes of basketball shoes, thereby providing a comprehensive assessment of their relative strength and weaknesses?

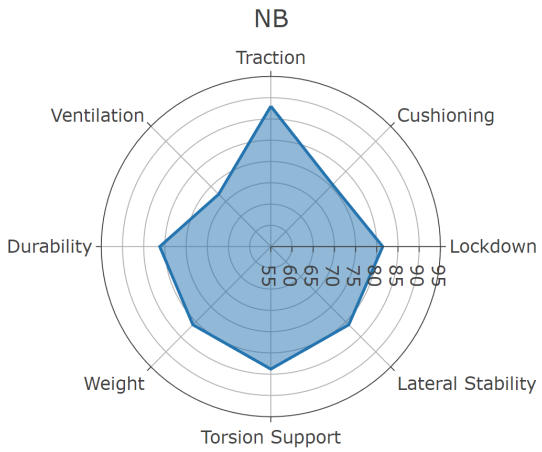
Radar graph representation successfully provides a comprehensive assessment of each brand's relative strength and weakness in their attributes. By providing a single brand radar graph, it allows users to visualize which attribute is such a brand's strength and weakness compared to other attributes. For example, *UA Curry* brand prioritizes its weight and lockdown attribute, but

lacks its durability. By providing a comparison radar graph, it allows users to visualize which brand has overall better performance across all attributes. For example, seeing *Nike vs Jordan* brand, we can see *Jordan* brand has better rating in almost every attribute than *Nike*.

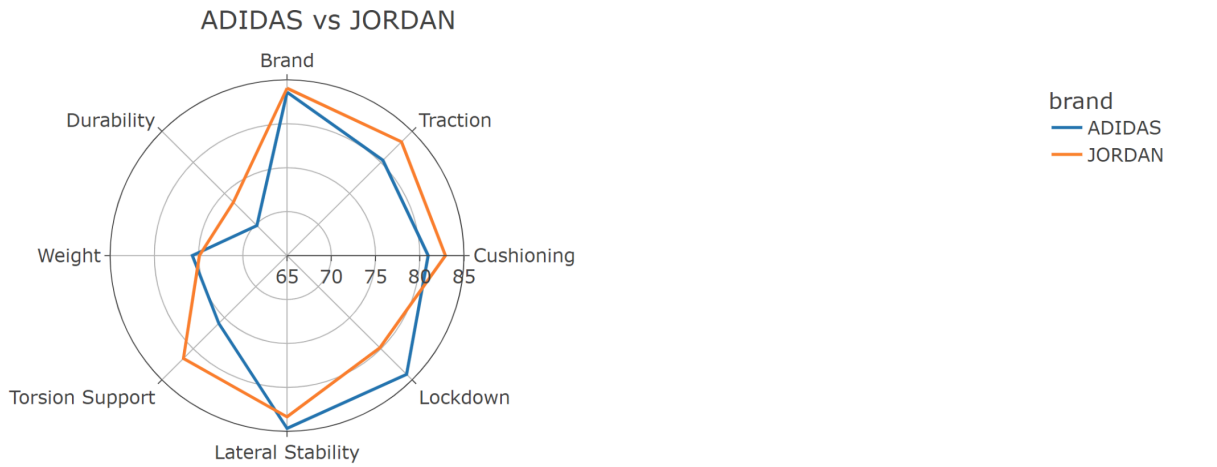


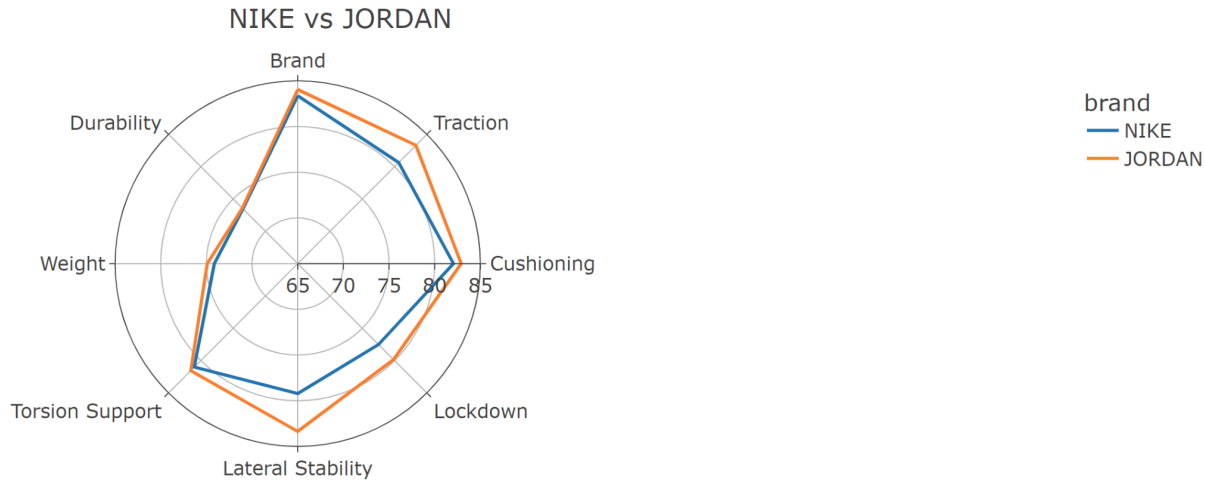






etc…





etc…

Research Question #2: How does the MSRP of basketball shoes vary based on their Overall rating and Brands?

In this question, we are trying to find out the relationship between MSRP and overall ratings of basketball shoes from different brands. From the scatter plot, we clearly see a positive correlation between these two variables. Basketball shoes ratings are often rated based on various factors, such as traction, cushioning, and so on. Shoes with higher overall ratings typically indicate better performance in every individual attribute. These shoes often have advanced materials, cushioning technology and lockdown systems, which lead to higher MSRP.

From the plot, Brands is also one of the major contributors to the MSRP. Different Brands provide shoes at different retail prices based on the marketing position and materials used for the shoes. We can see some premium brands like ASICS or Nike, whose average MSRP of basketball shoes are very high. While brands like Under Armour or ADIDAS are like mid-range brands, whose MSRP are more affordable. There are also budget-friendly brands, such as ANTA and PEAK.

By providing this graph, users can quickly find brands that can provide basketball shoes, which are in their budget, while performing well.

Research Question #3:

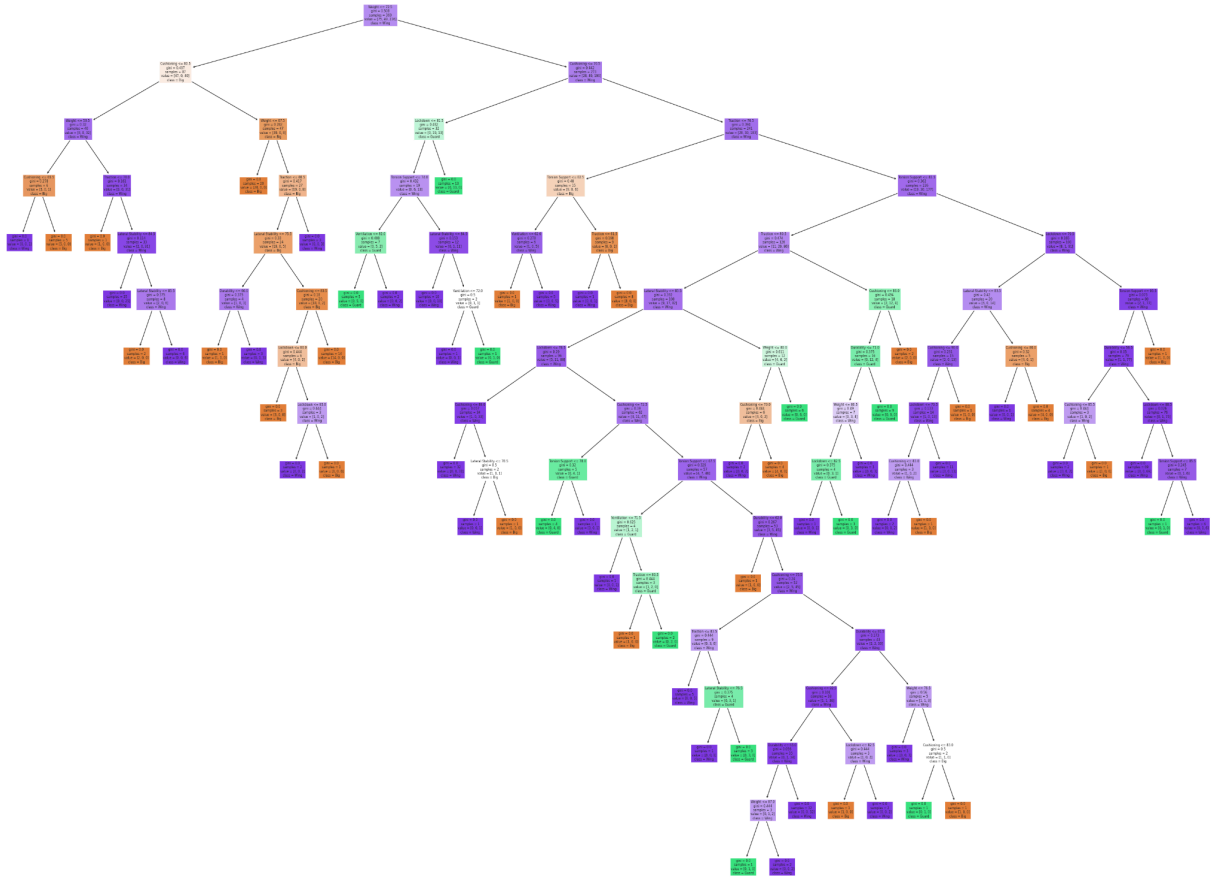
```
Train Accuracy: 1.0
Test  Accuracy: 0.7912087912087912
```

We use Decision Tree Classification as our machine learning algorithm to predict the appropriate position(guard, wing, or big) for sneakers based on different attributes of them, including traction, cushioning, lockdown, lateral stability, torsion support, weight, durability, and ventilation. Our data have been divided into 80% of the training set and 20% of the testing set.

We trained the Decision Tree Classification Model using the training data, and then tested the model using the testing data. In Decision Tree classification, the max depth is a hyperparameter that determines the maximum depth or levels of the decision tree. It controls the complexity of the tree by limiting the number of splits or branches it can have. In this question, we choose the max depth to be 20 which makes sure the accuracy of the training is 100%. On the premise of this, we found after extensive testing that we can predict the player position (guard, wing, big) performance score for each shoe based on the attributes of sneakers with over 70% certainty.

```
Enter the value for Lockdown: 78
Enter the value for Cushioning: 85
Enter the value for Traction: 79
Enter the value for Ventilation: 81
Enter the value for Durability: 80
Enter the value for Weight: 69
Enter the value for Torsion Support: 78
Enter the value for Lateral Stability: 77
Predicted Position: Big
```

We also create a user input interface, allowing people to input the attributes they have or want for a sneaker. Then our Decision Tree Model can help them know what the appropriate position is for this sneaker. Therefore, they can get more insight into that sneaker and make decisions based on this information.



Furthermore, we create a decision tree visualization that provides a clear and intuitive graphical representation of the decision-making process. It visually illustrates the model's structure and rules, allowing for easier interpretation and understanding. We can observe the Gini impurity of each decision, which qualifies the level of impurity or randomness in the target variable within each node. This visualization helps us understand how the model makes predictions based on the attribute values.

Impact and Limitations:

Research Question #1:

- Impact: The answer to this question will help develop the basketball shoe industry by reducing the risk of producing underperforming or unpopular shoes and improving

consumer satisfaction. Manufacturers can better meet consumer demands, and understanding attribute influence enhances consumer satisfaction.

Research Question #2:

- Impact: The answer to this question enables consumers to make informed decisions and select suitable options within their budget. It fosters healthy market competition, guiding manufacturers to develop better pricing strategies and promoting accessibility and affordability of high-quality shoes.

Research Question #3:

- Impact: Accurately predicting player position performance scores based on basketball shoe attributes has broad implications. It can optimize player performance, prevent injuries, advance scientific knowledge, foster industry innovation, guide evidence-based decision-making, and enhance consumer awareness. Teams and coaches can make better decisions when selecting shoes for players, contributing to improving the understanding of the relationship between shoe attributes and athletic performance.

Limitation:

- Since this dataset is completely dependent on r/BBallShoes community members' reviews, there is a possibility that this dataset is biased. In fact, for every shoe, the number of reviews barely exceeds 20 and some shoes have less reviews, which can result in a wrong score report on the dataset. In addition, basketball shoes' comfort is completely dependent on the users. It can dramatically change their opinions on one shoe based on their physical differences. Some users have narrower toes while others might have wider toes, which can affect their opinions on certain shoes. Our report only analyzes each shoe's attributes based on r/BBallShoes community's score report, so

basketball players should not make decisions of purchasing certain shoes based on the overall rating given by the dataset. Our report provides expectations for users about specific shoes and brands before users try them in store or purchase online.

Challenge Goals:

Challenge Goal 1: Improve accuracy and relevance of shoe recommendation for users

- By analyzing attributes and user preferences, our project aims to develop a model that can accurately recommend basketball shoes based on users' desired performance attributes. Through machine learning algorithms and incorporating user input scores, we can enhance the accuracy and relevance of the recommendation to make sure the model can find shoes that align with users' needs. Our project also utilizes the decision tree model to predict the proper position recommendation for different basketball shoes. This goal will have a significant impact on consumers, manufacturers, and shoe industries.

Challenge Goal 2: Make visualization for our dataset.

- By making a radar chart for each brand or each position about all the shoe attributes. Also, making a Brand Comparison chart between MSPR and Average Overall Ratings. We utilize the Plotly, Seaborn and matplotlib libraries for visualizations.The visualization can help users have a better and deeper understanding of the shoe attributes of each brand or each position and compare the MSPR and average overall ratings between different brands.

Plan Evaluation:

- Our proposed work plan worked smoothly. One of the reasons why our estimates were close to reality is because we were able to work remotely little by little every day. In addition, our coding part is separated by individual research questions. Therefore, we can work our part individually. After finishing the code, we worked together to write a report.

We can find minor bugs and errors more effectively. We can also discuss findings in our code and include them in the report.

Time Frame:	**Objective:**
By Mid May	- Complete data-clean up - Remove all unnecessary columns and NaN data. - Prepare all data for different research questions, and ready to use Expected Time: 3-4 hours
By fourth week of May	- Finish building models - Gather all data that will be used to plot Expected Time: 8-10 hours
By first week of June	- Test code and produce plots to visualize data - Write the final Report PDF Expected Time: 9-11 hours
By June 7, 2023	- Write the final report PowerPoint

Testing:

- For research questions 1 and 2, We randomly separated 20% of the original dataset into test data, and see whether the trend line (correlation for research question #1 part 1) of the test data will match the original dataset trend line. In a similar manner, we selected a random 20% of the original dataset as test data. For the radar, we visualized 20% of the dataset and determined if they match the original dataset.
- For research question 3, we initially separated our data into 80% of training data and 20% of testing data. Subsequently, the Decision Tree Classification Model was evaluated using

the reserved 20% testing set to compute its accuracy score. In addition, we also did the cross-validation test for machine learning. The purpose of cross-validation is to assess how well a model can generalize to unseen data. We utilize 5-fold cross-validation, which means the dataset is divided into 5 equally sized subsets of folds. For each fold, the model is trained on 4 folds of the data and evaluated on the remaining fold. The accuracy of the model is computed for each fold, representing how well the model can predict the label. Finally, we calculate the mean accuracy for each fold. This mean accuracy provides an overall performance estimate of the model's ability to generalize to unseen data.

```
Fold 1 Accuracy: 0.7252747252747253
Fold 2 Accuracy: 0.7222222222222222
Fold 3 Accuracy: 0.7
Fold 4 Accuracy: 0.7777777777777778
Fold 5 Accuracy: 0.7555555555555555
Mean Accuracy: 0.7361660561660561
```

Collaboration:

- We did not collaborate with anybody.

Cite:

- **https://github.com/jlu1211/CSE163-Final-Project**
- **https://plotly.com/**
- **https://stackoverflow.com/**
-