PIDS: Assignment 3 (Autumn 2025)

Data Analysis Project

Rishabh Gandhi, Indigo Pandey Bansal (2025112)

NBA 2K20 DATASET ANALYSIS

Project participants:

- 1. Rishabh Gandhi (rishabhgandhi.ug@iiit.ac.in) - Student ID: 2014053095
- 2. Indigo Pandey Bansal (indigo.bansal@iiit.ac.in) - Student ID: 2025112000

Project Plan

The Data (10 marks)

The dataset we have chosen is a dataset containing of NBA statistics for the year 2020. This data was scraped from a website called "hoopshype.com" which is an website to data directly from the NBA's website. The NBA publishes a wide array of data on its players, teams, and games, and this dataset aims to capture a snapshot of this data for the 2019-20 season. Each of these rows is the data about one basketball player that was a part of rosters. The attributes for the players are as follows:

- Full Name
- Nickname
- Jersey Number
- Current Team
- Position
- Height
- Weight
- Birthdate
- Salary
- Contract Length
- Contract Year
- Draft Round
- Draft Pick
- Rookie Year
- Years Pro
- College

The dataset contains the following variables and quantitative ratings: Rating, Jersey Number, Birthday, Height, Weight, Salary and Draft Year. Of the qualitative variables the same that takes a discrete form are: Rating, Jersey Number, Birthday, Salary, Draft year, while the following take continuous values: Height, Weight, Years Pro, Contract Length, Contract Year, Draft Round, Draft Pick, Rookie Year. The following are nominal: Full Name, College, Nickname, Draft Round and Draft Pick as there is no inherent ordered value of this kind they can take on the data. Contract Length, Contract Year, Draft Round, Draft Pick, Rookie Year, Years Pro, Position, Current Team, Nickname, Full Name, College, Birthdate are ordinal values, presumably when a player is either not currently signed or retired or if they did not play in a college known to the customers of the dataset. Jersey Number, Salary, Height, Weight, Rating are continuous variables. Some of the interesting questions that can be asked from this data are: Is there any correlation between the Rating and Salary, Rating and Years Pro, Salary and Years Pro, Salary and Height, Salary and Weight, Salary and Jersey Number, Salary and Position, Salary and Current Team, Salary and Contract Length, Salary and Contract Year, Salary and Draft Round, Salary and Draft Pick, Salary and Rookie Year, Salary and College. Some of the above hypotheses are obvious and others are not so obvious, but it is worthwhile to explore all of these hypotheses and see if there is any correlation between the variables

Objective 1: Data Visualisation and Correlation

Explanation of Results

In this section, we are exploring, cleaning, visualisation and correlation. We will cover each of these tasks individually. Data cleaning was needed because some of the variables contained NA values, so we collected the proper values. Histograms were used to visualise distributions of the variables, and we found that most of the variables were skewed to the right. Correlation was used to see which variables had the strongest relationship with salary and they were not as correlated, especially the two variables that had the most weight in salary predict. The Heart Stopper variable was the most correlated with salary, followed by Dunk Ability.

The data contained following variables: Salary, Rating, Draft round, Height, Weight, Dunk ability, Heart stopper, Draft year, Draft school, Draft pick. We chose to use the Draft pick variable because we felt that the purpose of the exercise was not to perform complex analysis to get a deep understanding of exactly what the data tell us, but rather to see how we could manipulate the data to produce meaningful results. We felt that the draft pick variable was the most interesting and that the number of dunks increases as the rating increases. The histogram of rating has a whole shape with the majority of the data in the middle, and the histogram of dunk ability has a right skew. We felt that the draft pick variable was interesting because it would allow us to average all the players with the same draft pick and plot a plot where dots that go left to right and upward that represent the average of the data. We felt that the draft pick variable was interesting because it would allow us to see how the average of players in the same draft pick changes over time

Conclusion (5 marks)

Achievements

The Tetris Machine Language uses a general visualization of our dataset, to understand it better. We have then proceeded by exploring the data set using "supply" and simple correlations, which could not be highly correlated, indicating that we could have effective predictions from our data. Therefore, we performed a deep predictor, which is a regression predictor, and tested its with good results using the "Mean Absolute Error" and "Root Mean Squared Error" measures. Finally, we performed a "Random Forest" classifier, which could not return a finite result but rather a list of probabilities for each class