A Naive Bayes Model for Points Prediction

Thomas Deng

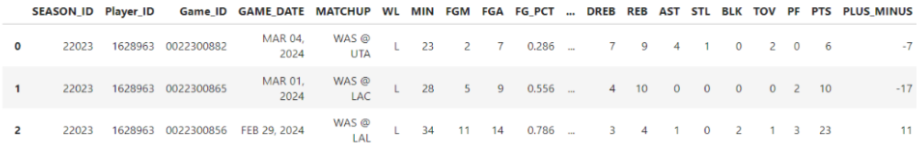
March 2024

1 Introduction

NBA players are extremely volatile by nature. As a fan, front office, coach, or sports bettor, as much as you would like your favorite player to always excel, variance, inconsistency, and many factors affect a players' performance. I have always wondered whether or not a players' past games affect their current performance, referring to the renowned hot streak or cold stretch phenomenons. If a player is performing extremely well, can we expect their next game to be better than their season average? If we were to trust word of mouth alone, we would expect players to perform much better and similarly if they were performing poorly, we would expect to perform worse than their season average the next game. However, some argue that the hot or cold streaks are simply myths, and that they are simply just byproducts of variance. I sought to see if we could create a Naive Bayes classifier to test this phenomenon with player points, and perhaps even create a model that could assist teams with statistical analysis or casual fans with sports betting.

2 Methods

. The first step in this process was collecting the data. This entailed using NBA API to scrape player data directly from NBA.com to gather CSV files of the game logs for each of the NBA players. I used GPT to help with understanding how to use NBA API. Game logs look something like this:



	SEASON_ID	Player_ID	Game_ID	GAME_DATE	MATCHUP	WL	MIN	FGM	FGA	FG_PCT	...	DREB	REB	AST	STL	BLK	TOV	PF	PTS	PLUS_MINUS
0	22023	1628963	0022300882	MAR 04, 2024	WAS @ UTA	L	23	2	7	0.286	...	7	9	4	1	0	2	0	6	-7
1	22023	1628963	0022300865	MAR 01, 2024	WAS @ LAC	L	28	5	9	0.556	...	4	10	0	0	0	0	2	10	-17
2	22023	1628963	0022300856	FEB 29, 2024	WAS @ LAL	L	34	11	14	0.786	...	3	4	1	0	2	1	3	23	11

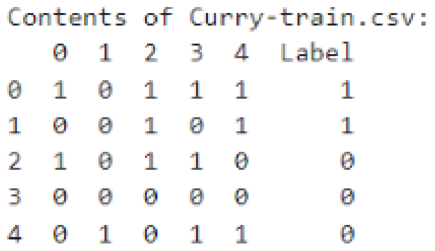
Figure 1: 3 games for a player

Each player has its own Player ID and each row represents a game that player played. Next, I created a dictionary that associated each player with his game logs. Now that we have our datasets, we have to prepare the data to fit cleanly into a Naive Bayes Classifier.

3 Naive Bayes

In this context, the difficulty was deciding how I would organize the data to fit into a Naive Bayes Classifier. Yes, player points are discrete random variables, however they do not fit neatly into binary categories. Furthermore, representing hot or cold streaks with binary variables is not obvious. Additionally, I had to decide whether I would make a model for each individual player or if I would use a one-size-fits-all model. I decided on the following methodology.

For each player, we would create training and test data by using a rolling interval of n games. Each game feature received a value of 1 if it was over the player's season points average and a value of 0 if it was under the player's season points average. The label for those features is determined by the game right after the rolling interval. I used GPT to assist in the preparation of this data. If that game surpassed the player's season average, it would receive a 1 and a 0 if it was under. The training data would look something like this:



```
Contents of Curry-train.csv:
   0  1  2  3  4  Label
0  1  0  1  1  1      1
1  0  0  1  0  1      1
2  1  0  1  1  0      0
3  0  0  0  0  0      0
4  0  1  0  1  1      0
```

Figure 2: Steph Curry Training Data

The Naive Bayes classifier assumes the conditional independence of the features given the class label. Therefore, the likelihood of the features can be expressed as the product of individual probabilities:

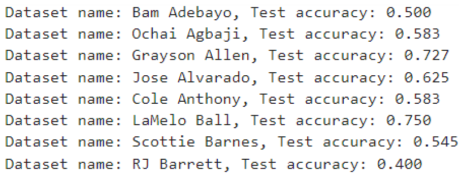
$$P(X|Y) = \prod_{j=1}^{m} P(X_j|Y) \tag{1}$$

In our NBA project, X_j for $j = 1, \ldots, 5$ represents whether the player scored above their average in each of the last five games. The class label Y indicates if the player scores above their average in the next game.

The classifier predicts the class Y that maximizes the posterior probability:

$$\hat{Y} = \arg\max_{y=\{0,1\}} P(Y=y) \prod_{j=1}^{5} P(X_j|Y=y) \tag{2}$$

We use Laplace smoothing to account for any data that contains only zero's, adding 1 to the numerator of each likelihood's probability and 2 to the denominator. We accumulate our accuracies and then represent them like this:

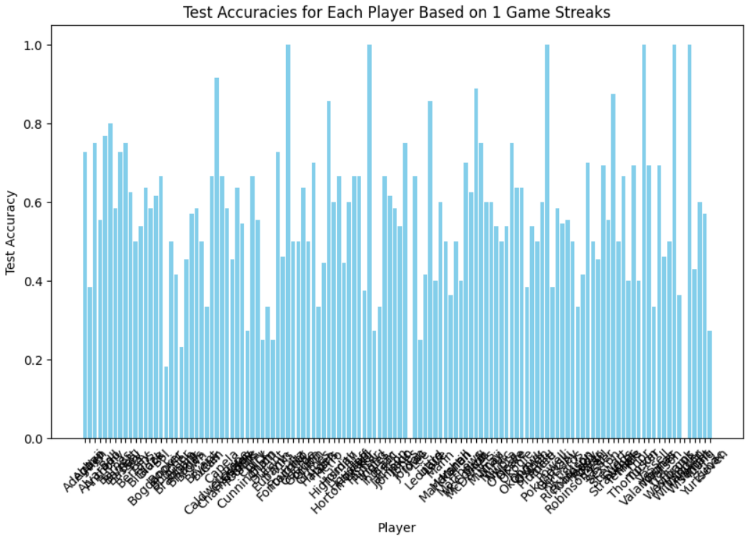


```
Dataset name: Bam Adebayo, Test accuracy: 0.500
Dataset name: Ochai Agbaji, Test accuracy: 0.583
Dataset name: Grayson Allen, Test accuracy: 0.727
Dataset name: Jose Alvarado, Test accuracy: 0.625
Dataset name: Cole Anthony, Test accuracy: 0.583
Dataset name: LaMelo Ball, Test accuracy: 0.750
Dataset name: Scottie Barnes, Test accuracy: 0.545
Dataset name: RJ Barrett, Test accuracy: 0.400
```

Figure 3: Accuracies

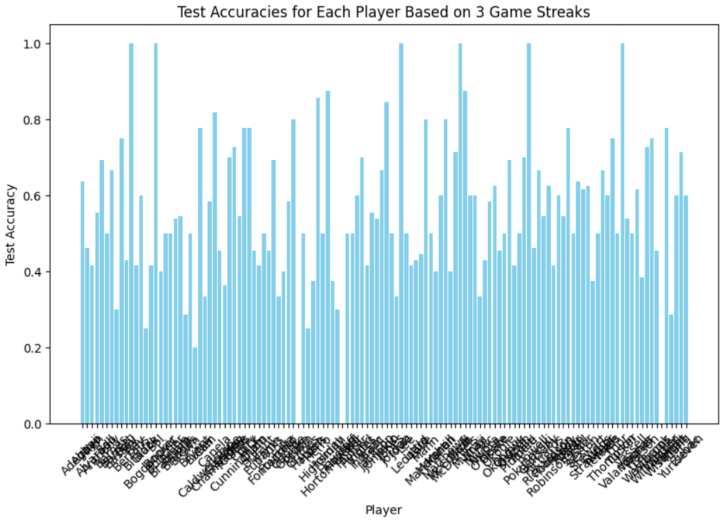
4 Results

We test on 1, 2, 3, 5, and 10 game streaks. We see for 1 game, that we can obtain a total accuracy of 0.56. For 2 game streaks, that we can obtain a total accuracy of 0.54. For 3 game streaks, we get 0.56. For five games, we get 0.54. For 10 games, we get 0.51. I find it very interesting that for the 10 game stretch, the accuracy is the lowest. It makes sense as the number of games increases, you would expect the information that a game 10 games ago would provide would be extremely minimal. It is also interesting that 1 game is tied for 3 games as the highest classification accuracy. Perhaps the only previous game that seems to indicate future performance is the game just prior. I am extremely interested in the psychology of this topic. Athletes who have been performing at the highest level would seem to be less affected by volatility mentally, however, maybe the effect can be noticed if a player has been playing a certain way for the past 1 or 3 games. In the future I would love to test on more intervals to see what number would have the highest accuracy, and I am excited to see how I can improve the model!



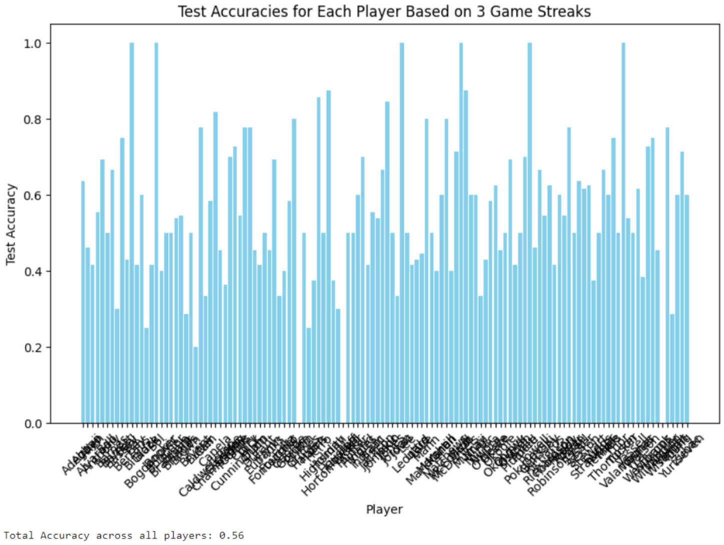
```
Total Accuracy across all players: 0.56
```

Figure 4: 1 game streaks



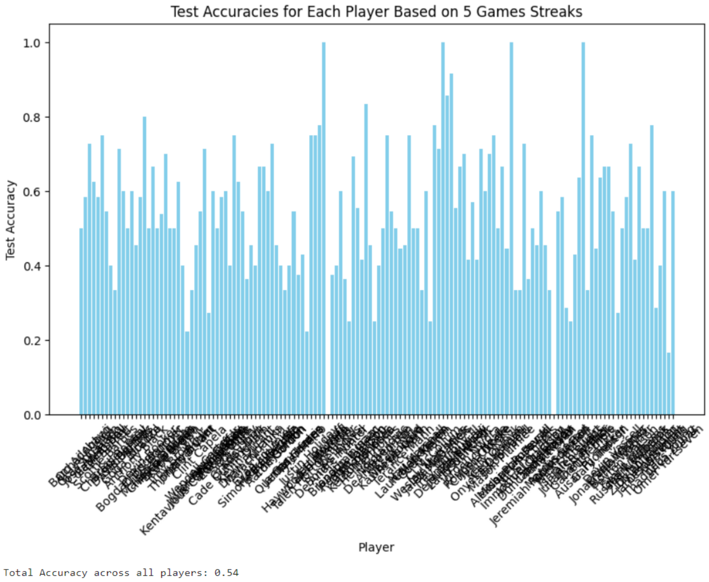
```
Total Accuracy across all players: 0.56
```

Figure 5: 2 game streaks



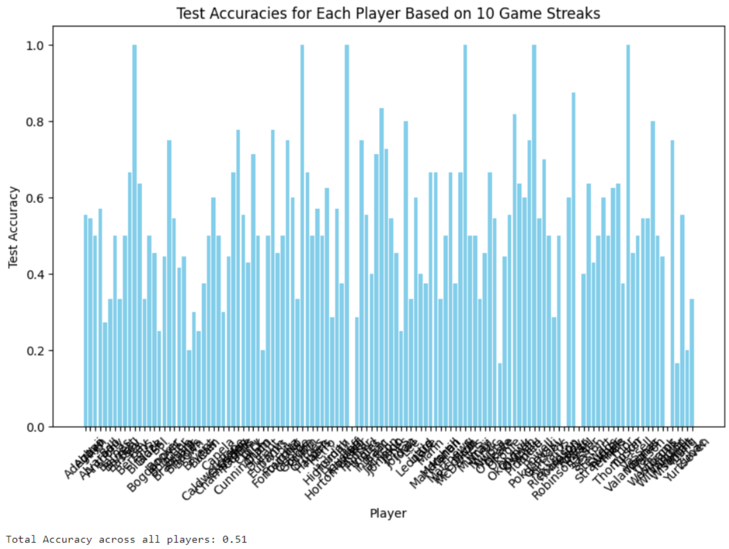
```
Total Accuracy across all players: 0.56
```

Figure 6: 3 Game streaks



```
Total Accuracy across all players: 0.54
```

Figure 7: 5 Game streaks



```
Total Accuracy across all players: 0.51
```

Figure 8: 10 Game streaks