Amelia Grevin
Dylan Murphy

NBA Player Retirement Prediction

Problem:
How accurately can we predict when NBA players will retire?

Goals and Relevance:
Goals:

We chose two specific methods of analyzing retirement, by modeling:

- Predicting career length
- Predicting retirement age

Relevance/Interest:

- Basketball is among the top 3 watched sports in America
- There is a lot of recorded basketball data/stats(we were able to find a good dataframes to work with because of the records kept on player stats)
- We both enjoy watching the NBA

Steps:
- Data Cleaning
- Preparing the data
- Modeling
- Loss analysis
- Results

Dataset:
- The datasets we used are from Kaggle
  - Player statistics: https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats
  - Height and weight: https://www.kaggle.com/datasets/drgilermo/nba-players-stats/data?select=Players.csv
  - Injuries: https://www.kaggle.com/datasets/ghopkins/nba-injuries-2010-2018?select=injuries_2010-2020.csv
- Important column info:
  - player, playerID, mp, g, gs, fg, fga, fg_percent, 'x3p', 'x3pa', 'x3p_percent', 'x2p', 'x2pa', 'x2p_percent', 'e_fg_percent' ,'ft', 'fta', 'ft_percent', ‘height’, ‘weight’, ‘num_injuries’
- Important row info:

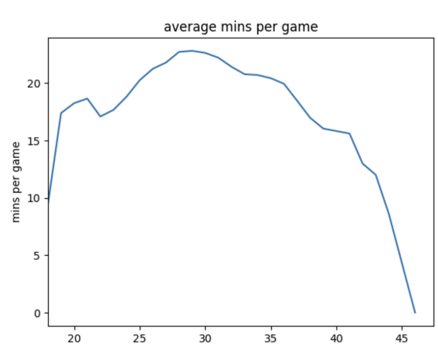
- The original dataset contains rows for each unique data piece per player per season(ex: there are two different rows, within the same year, for a player who played for two different teams in the same season)

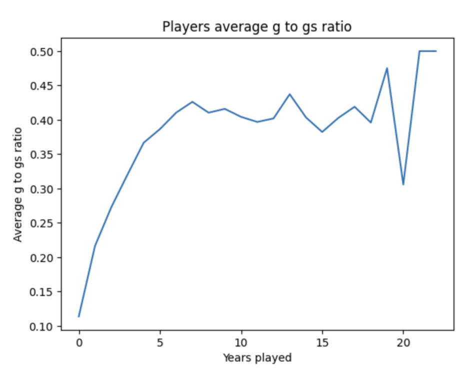
Data Cleaning:

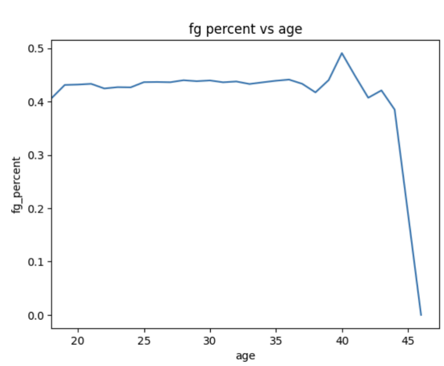
- Merging:
  - Merge multiple player stats datasets together on the player_id column(shared col across the main player stats datasets)
  - Include additional data sets that include height/weight information
  - Include a dataset with injury data(between the years of 2010-2020)
- Cleaning/deletion:
  - Create a filtered DF, only including rows between 1947-2025(We excluded the 2026 data because the season was not yet completed and we didn't want the incomplete data to affect the training.)
  - Replace missing data with Nan to start with(later on we replace with mean col values for linear regression)
- Create a column for the true values of:
  - Retirement year(found by determining last year played by a player)
  - Career length(found by counting number of unique seasons for player)

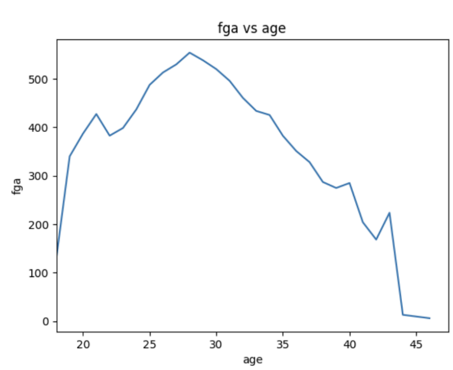
Initial Plotting:

We made a few initial plots that we thought might show interesting relationships:









Looking at these plots(along with some other initial plots)helped us to determine which columns we thought we should keep an eye out for as high impact features related to retirement.

Preparing Data for Modeling:

- Only use the data from retired players(exclude active players because we cant accurately train the data without a true retirement age(or true career length))
- Groupby playerID:
    - Each player should have only a single row
    - The numeric columns are each averaged across all seasons played
    - Include a column that shows the true retirement age
    - Include a column that shows the true career length
    - Include the player and pos columns
    - Include height, weight, and injuries columns for applicable players
- Split the data into a train and test set (test size = 0.2)

Modeling:

- Create a multiple linear regression model function to predict the retirement age and the career length for each retired player
- Training data with features(x) and true values(y) used for fitting model
- Feature columns include FG%, MP, Height, Number of injuries
- Test data used to evaluate predictions
- Fill in missing values with the column mean because not too many missing values

Predicted Retirement Ages Results:

Results of calling the predictor function and passing in the features and y arrays. We are using the feature(cols) to predict the retirement age. So we are running multiple linear regression to determine how accurately we can predict retirement age.

	player	retirementAge	predictedRetirementAge
0	0	0.0	24.625370
1	Alaa Abdelnaby	26.0	27.195733
2	Kareem Abdul-Jabbar	41.0	36.630127
3	Walt Hazzard	31.0	31.162181
4	Mahmoud Abdul-Rauf	31.0	31.112303
5	Tariq Abdul-Wahad	28.0	28.265437
6	Zaid Abdul-Aziz	31.0	27.900310
7	Shareef Abdur-Rahim	31.0	35.536842
8	Tom Abernethy	26.0	27.270291
9	Forest Able	24.0	23.996286
10	John Abramovic	28.0	26.671437
11	Álex Abrines	25.0	28.256299
12	Alex Acker	26.0	24.338062
13	Don Ackerman	23.0	24.857912
14	Mark Acres	30.0	28.604436
15	Bud Acton	26.0	24.316645
16	Quincy Acy	28.0	27.563520
17	Alvan Adams	33.0	32.958187
18	Don Adams	29.0	30.802637
19	George Adams	25.0	28.482041

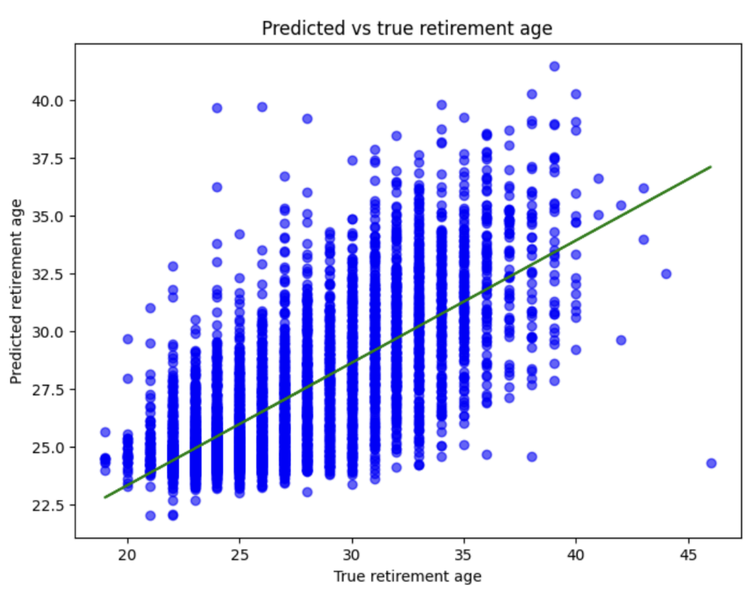
Predicted Career Length Results:

Results of calling the predictor function and passing in the x and y arrays. We are using the feature(cols) to predict the career length.

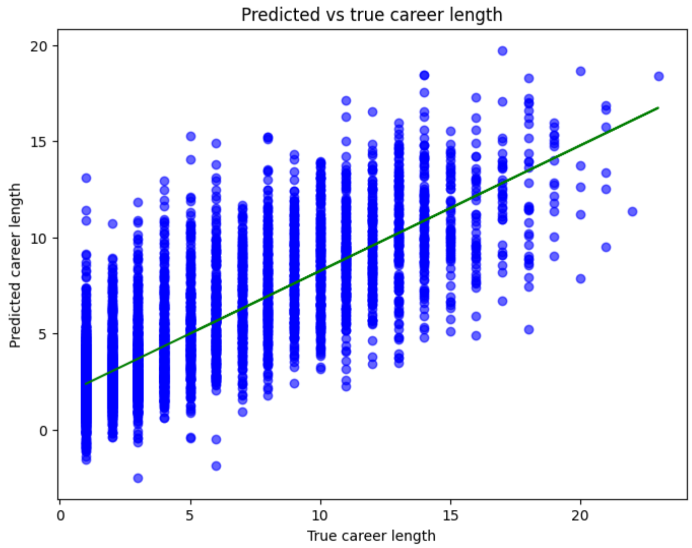
	career_length	predicted_career_length
1149	11.0	9.488108
393	10.0	6.830872
1268	4.0	4.904366
4233	2.0	3.101261
4181	10.0	9.799169
2680	14.0	10.059506
3949	16.0	9.190214
4854	2.0	0.575836
3757	1.0	3.047487
3979	6.0	8.312091
3418	4.0	4.275128
2509	2.0	4.041470
1371	1.0	0.591065
3478	2.0	1.791224
4748	7.0	4.791690
2577	16.0	14.496677
538	7.0	8.370778
718	1.0	2.636539
4502	16.0	9.897090
4473	2.0	0.558572

Plotting with line of best fit:

- The plot shows the true retirement age vs our predicted retirement age with a positive relationship!
- We included the line of best fit to show the positive relationship

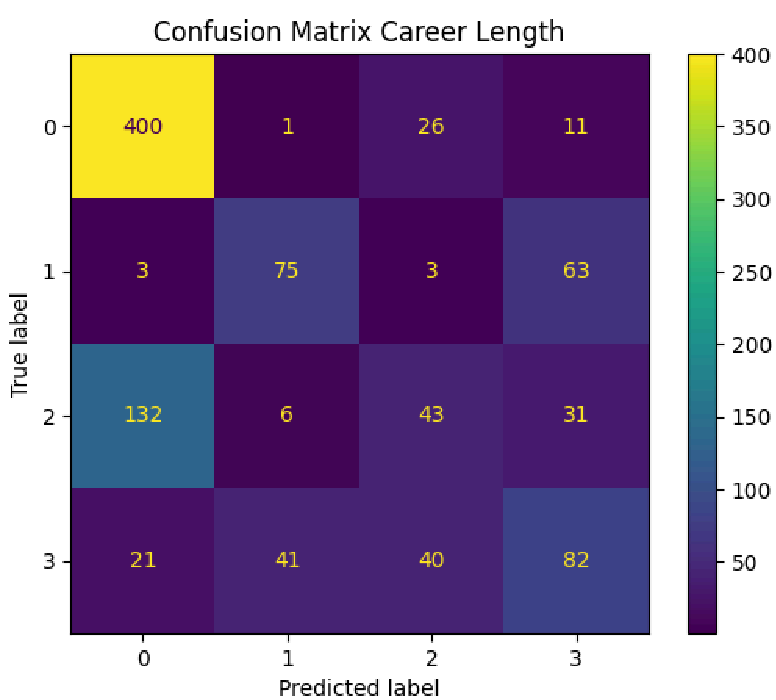


- This plot shows the true career length vs our predicted career length, which also has a positive relationship!

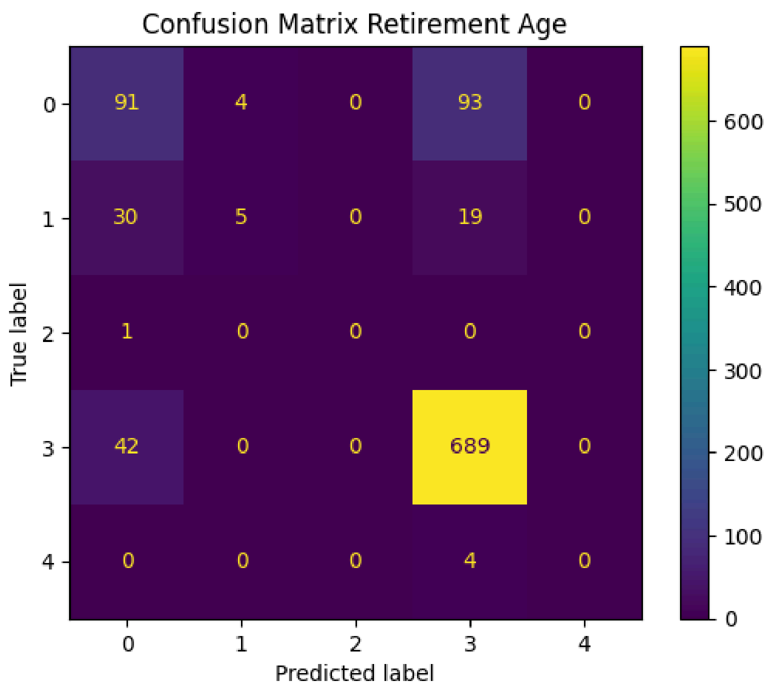


Classification:

- Career Length
  - Classes = [0-2, 2-5, 5-10, 10+]



- Retirement Age
    - Classes = [<30, 30-35, 35-40, 40+]



Both of the confusion matrices show generally effective and accurate classification, especially for Retirement Age prediction.

Results:

- Use the MSE loss function to determine the accuracy of prediction(We selected MSE instead of MAE because we didn't believe there to be large enough outliers to hugely affect the data).
- RMSE gives us the number of years we are expected to be off by in our predictions.
- For the career length linear regression:
    - RMSE=2.57
    - R^2 = 0.6699
- For the age predictor linear regression:
    - RMSE=3.10
    - R^2=0.5901
- After logistic regression, accuracy tells us the percentage of predictions our model got right, alongside other metrics such as precision and recall.

- For the career length logistic regression

```
Career Length Accuracy: 0.6134969325153374
              precision    recall  f1-score   support

           0       0.72      0.91      0.80       438
           1       0.61      0.52      0.56       144
           2       0.38      0.20      0.27       212
           3       0.44      0.45      0.44       184

    accuracy                           0.61       978
   macro avg       0.54      0.52      0.52       978
weighted avg       0.58      0.61      0.58       978
```

- For the retirement age logistic regression

```
Retirement Age Accuracy: 0.8026584867075665
              precision    recall  f1-score   support

           0       0.55      0.48      0.52       188
           1       0.56      0.09      0.16        54
           2       0.00      0.00      0.00         1
           3       0.86      0.94      0.90       731
           4       0.00      0.00      0.00         4

    accuracy                           0.80       978
   macro avg       0.39      0.30      0.31       978
weighted avg       0.78      0.80      0.78       978
```

Analysis:

- Our multiple linear regression model more accurately predicts career length over retirement age, while the logistic regression model more accurately predicts retirement age over career length.

- RMSE being 2.57 for career length means that we are approximately 2.57 years off in predicting career length per prediction on average.
- R^2 being 0.6699 means our model explains 66.99% of the differences in career length.
- Our plotting results show us that the relationship between true and predicted values is positive!
- The logistic regression model accurately predicts about 61% of career lengths, and about 79% of retirement ages

Additional Plans:
- Find more complete data for injuries
- Use injury information in a more complex way (severity of injury)
- Use data on player income
- Try different modeling methods and loss functions to reduce error
- Wait for currently active players to retire to see how accurate our predictions are