NBA Hall of Fame Prediction Model

This project focuses on building a prediction model to determine whether an NBA player is likely to be inducted into the Hall of Fame (HoF). The model utilizes three different algorithms: logistic regression (log), random forest classifier (rfc), and k-nearest neighbors (knn) to make predictions based on various player attributes.

Datasets

The prediction model is trained on a dataset that contains information about NBA players, including their career statistics, accolades, and other relevant attributes. The dataset should be appropriately preprocessed to ensure accurate predictions.

Datasets used for this project:

1. https://www.kaggle.com/datasets/ryanschubertds/all-nba-aba-players-bio-stats-accolades
2. https://www.kaggle.com/datasets/sumitrodatta/nba-aba-baa-stats

Algorithms Used

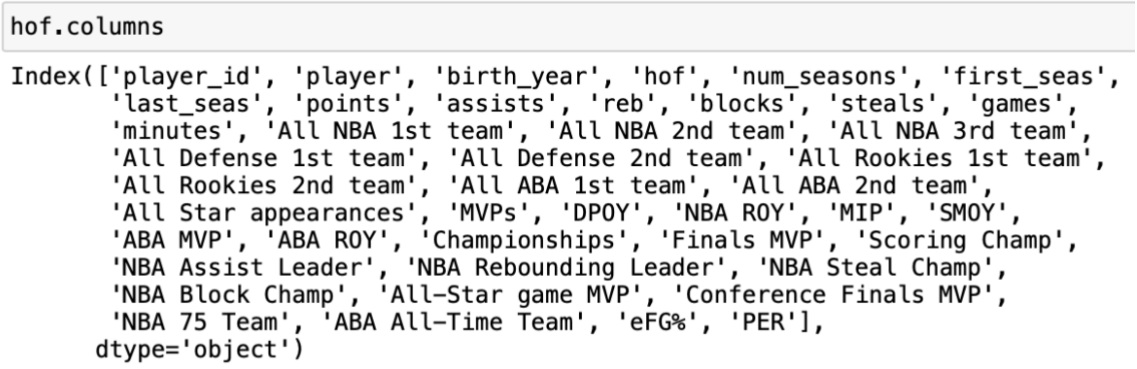
1. **Logistic Regression (log)**: This algorithm is a statistical model used for binary classification. It estimates the probability of an instance belonging to a particular class (in our case, HoF or non-HoF) based on the input features.

2. **Random Forest Classifier (rfc)**: This algorithm is an ensemble learning method that combines multiple decision trees to make predictions. It creates a "forest" of trees and aggregates their results to classify instances.

3. **K-Nearest Neighbors (knn)**: This algorithm classifies instances based on their proximity to other instances in the feature space. It determines the class of an instance by considering the majority class of its k nearest neighbors.

Features

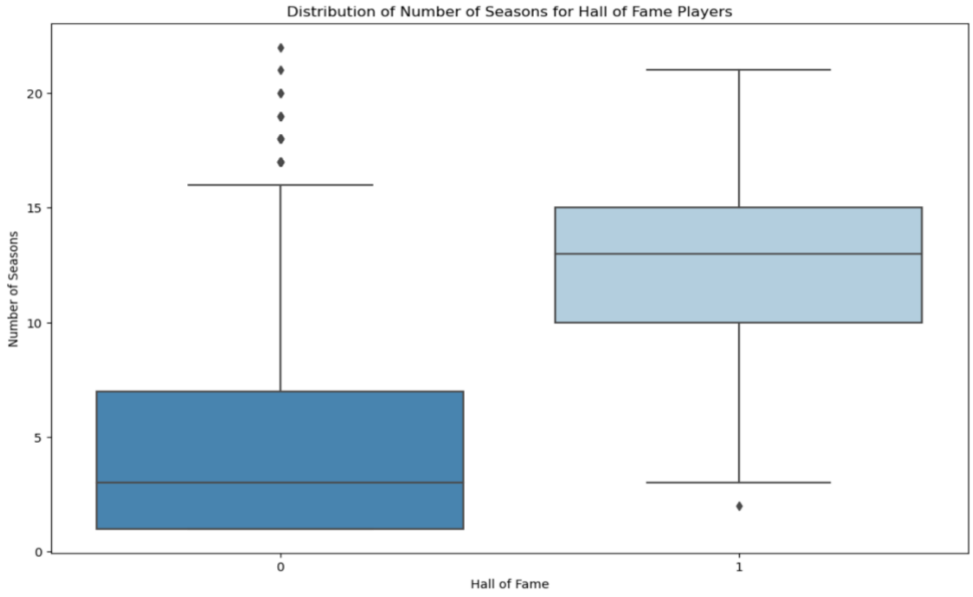
After joining and mapping different datasets, a single unified dataframe is created, incorporating relevant information from multiple sources:



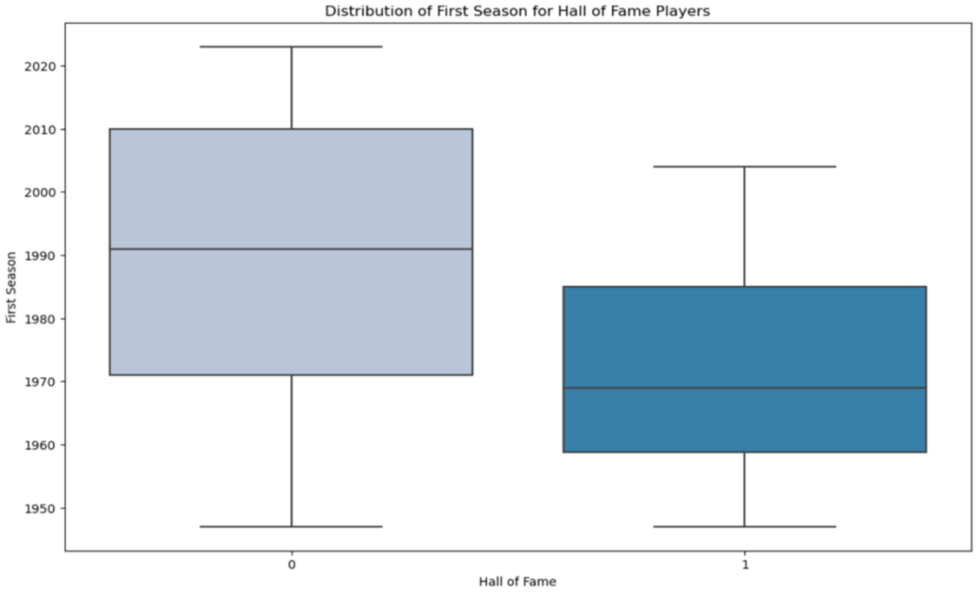
```
hof.columns
```

```
Index(['player_id', 'player', 'birth_year', 'hof', 'num_seasons', 'first_seas',
       'last_seas', 'points', 'assists', 'reb', 'blocks', 'steals', 'games',
       'minutes', 'All NBA 1st team', 'All NBA 2nd team', 'All NBA 3rd team',
       'All Defense 1st team', 'All Defense 2nd team', 'All Rookies 1st team',
       'All Rookies 2nd team', 'All ABA 1st team', 'All ABA 2nd team',
       'All Star appearances', 'MVPs', 'DPOY', 'NBA ROY', 'MIP', 'SMOY',
       'ABA MVP', 'ABA ROY', 'Championships', 'Finals MVP', 'Scoring Champ',
       'NBA Assist Leader', 'NBA Rebounding Leader', 'NBA Steal Champ',
       'NBA Block Champ', 'All-Star game MVP', 'Conference Finals MVP',
       'NBA 75 Team', 'ABA All-Time Team', 'eFG%', 'PER'],
      dtype='object')
```

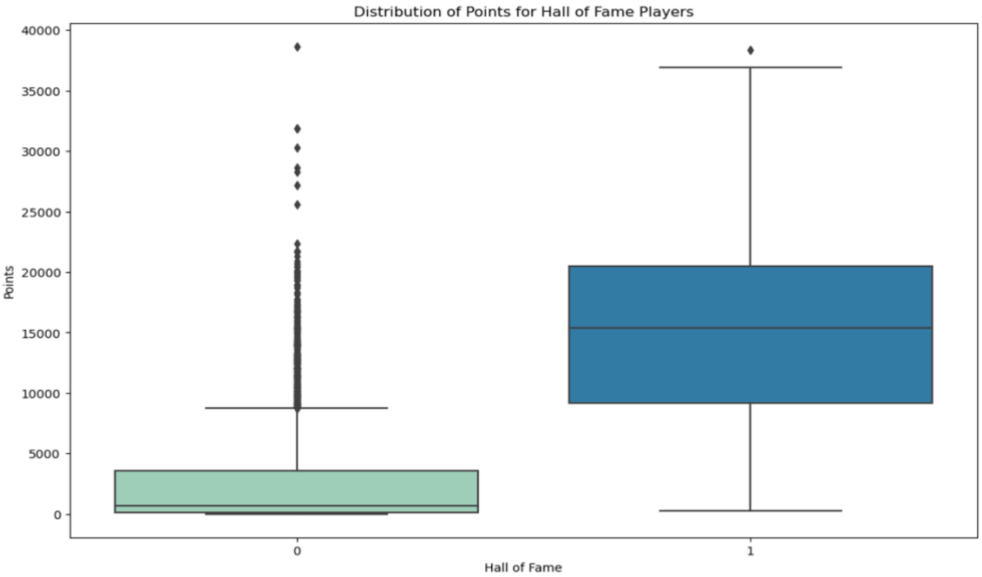
After performing necessary preprocessing steps, we proceed to visualize the data by creating various plots and visualizations.



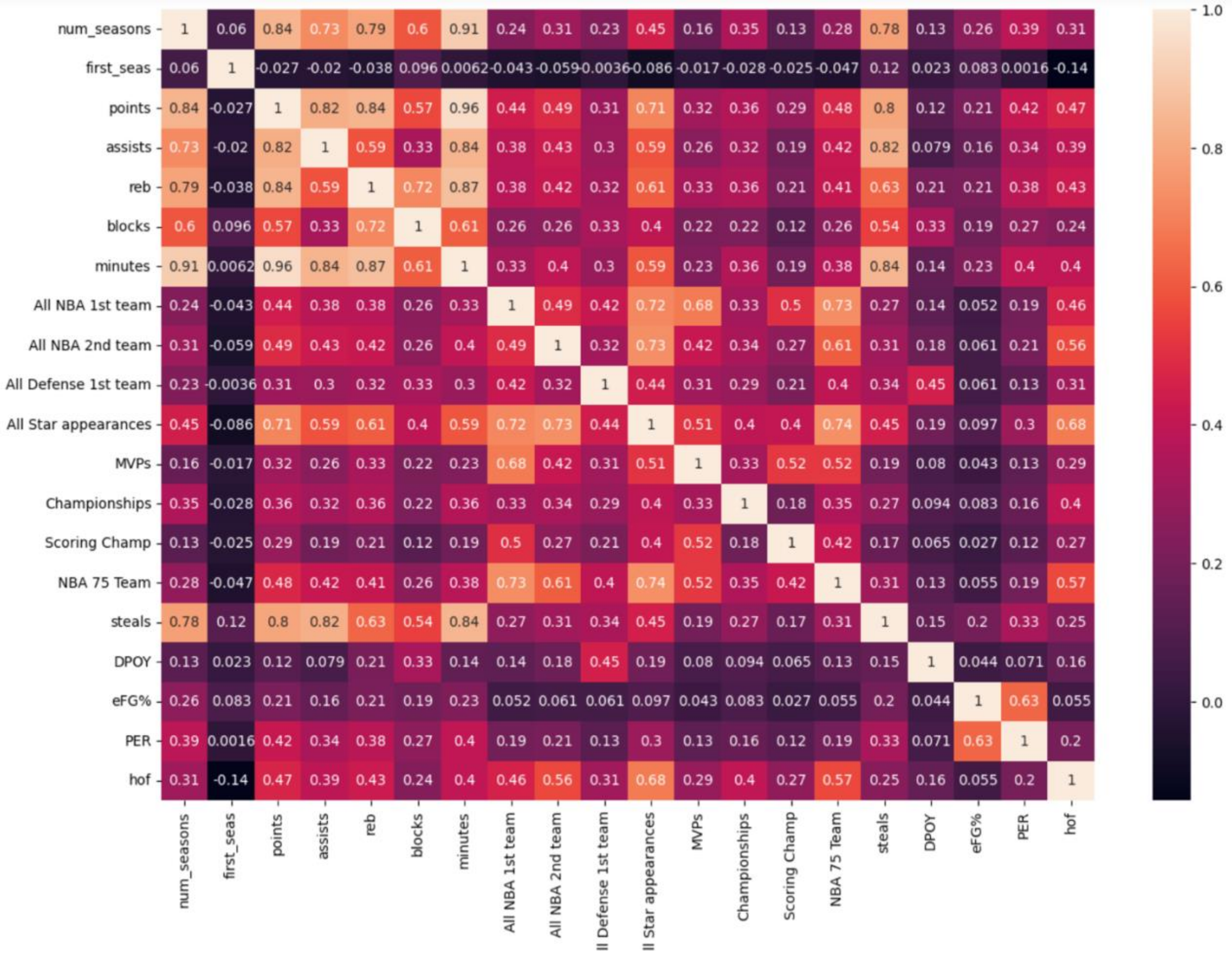
Looks like almost all Hall of Famers have at least 10 seasons played.



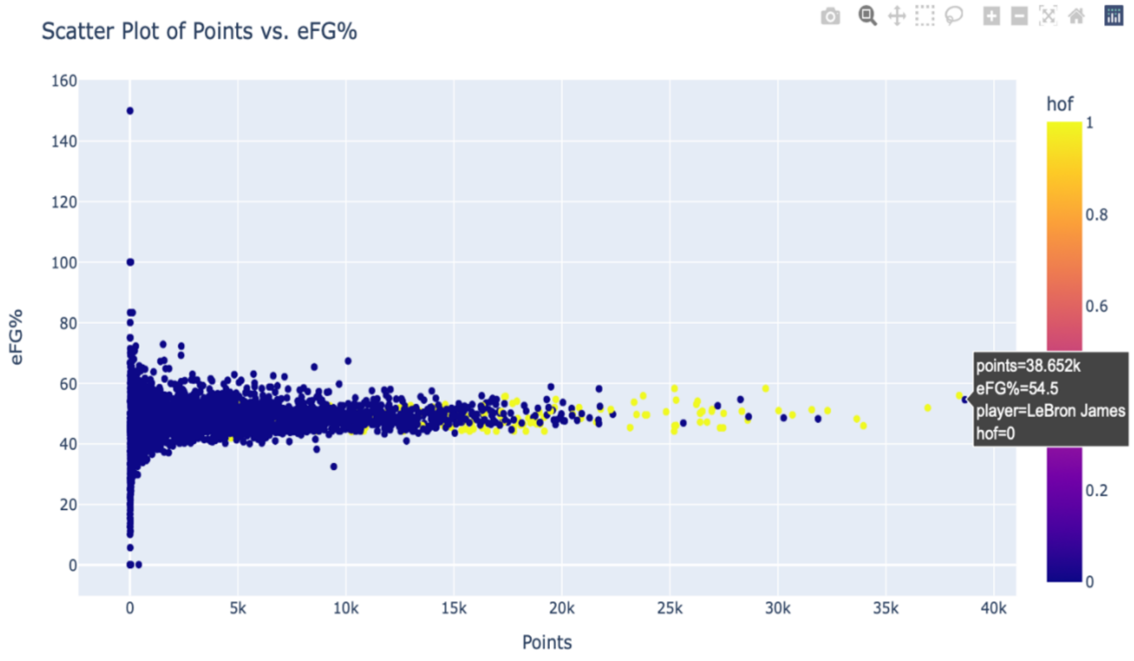
It is notable that a significant majority of Hall of Famers commenced their careers prior to 1990. This pattern can be attributed to the fact that induction into the Hall of Fame does not occur immediately following a player's retirement but rather after a certain number of years.



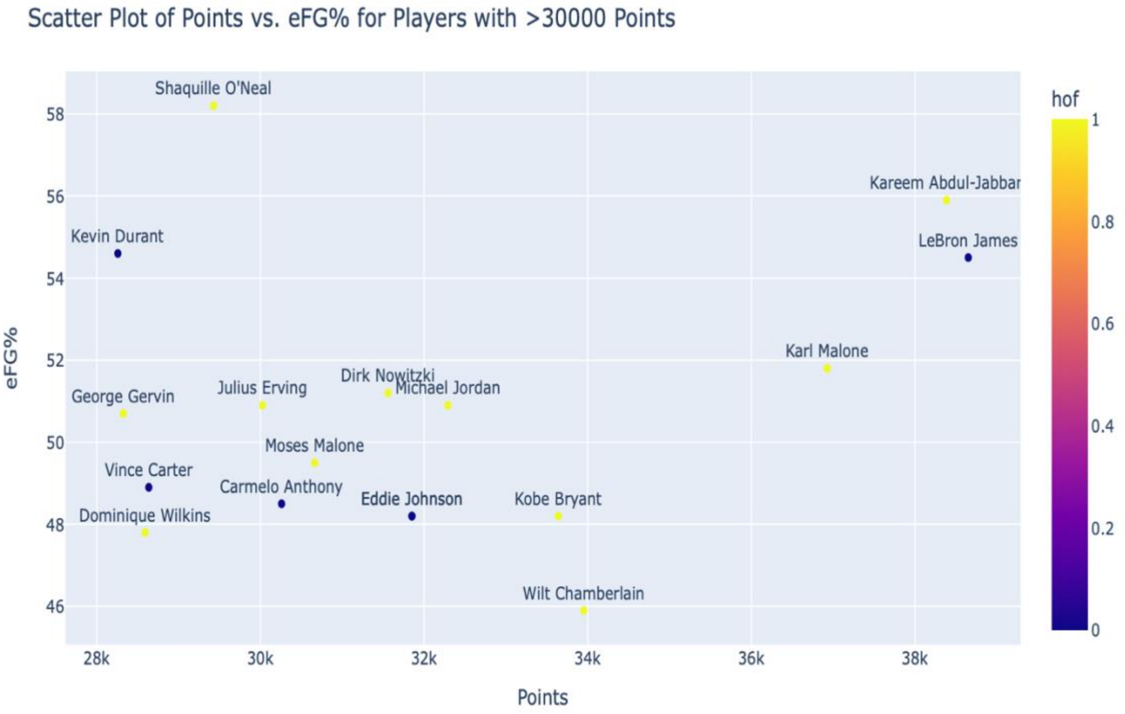
Looks like almost all Hall of Famers scored at least 10k points during their career.



In the correlation heatmap, we observe the relationships between variables such as All-Star Appearances, points, All-NBA 1st team selections, and championships. These correlations align with common expectations and provide valuable insights.

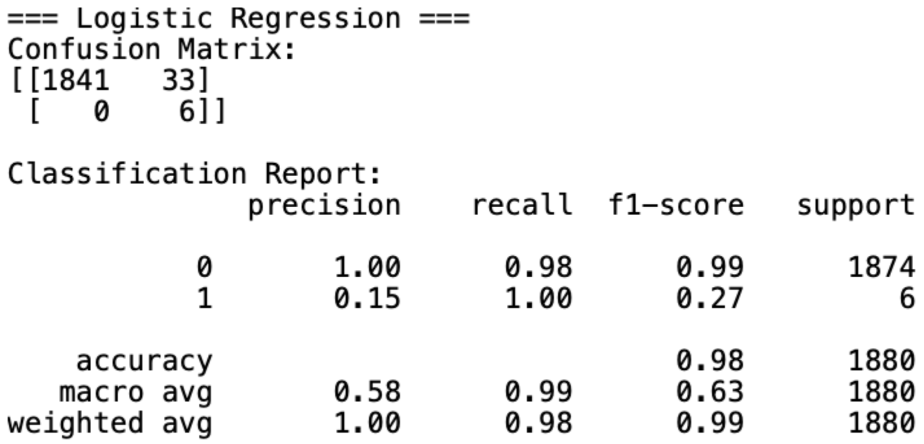


Seems incredible how Lebron James stands out in this scatter plot.

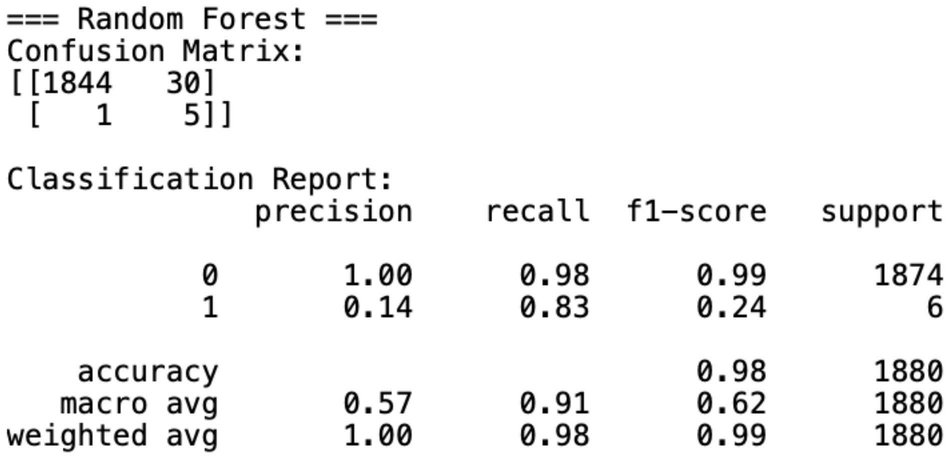


In this graph, certain inconsistencies that have been criticized over the years can be observed, such as Vince Carter not being inducted into the Hall of Fame. When comparing this graph with Dominique Wilkins, these inconsistencies become apparent.

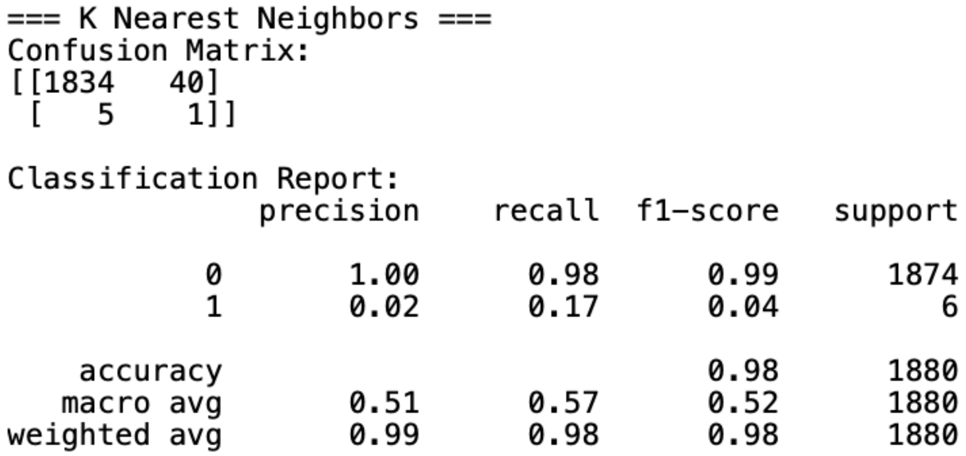
Now, let's move on to machine learning and predictions. The dataset will be split into two parts: players who started their careers before the 2000s will be included in the training data slice, while the remaining players will be used to evaluate the model's ability to predict future Hall of Famers based on their career achievements.



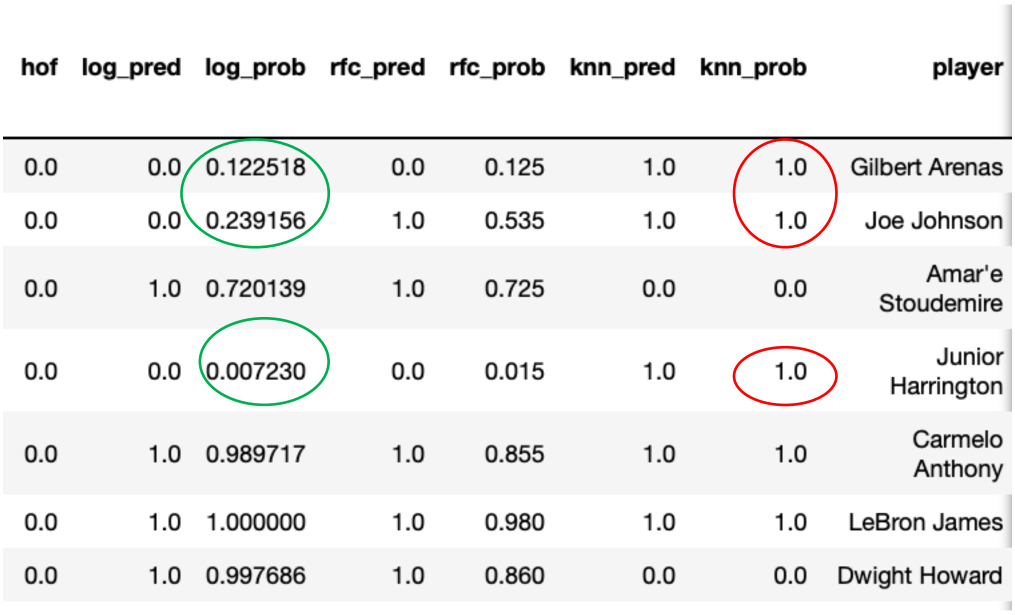
```
=== Logistic Regression ===
Confusion Matrix:
[[1841   33]
 [   0    6]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      1874
           1       0.15      1.00      0.27         6

    accuracy                           0.98      1880
   macro avg       0.58      0.99      0.63      1880
weighted avg       1.00      0.98      0.99      1880
```



```
=== Random Forest ===
Confusion Matrix:
[[1844   30]
 [   1    5]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      1874
           1       0.14      0.83      0.24         6

    accuracy                           0.98      1880
   macro avg       0.57      0.91      0.62      1880
weighted avg       1.00      0.98      0.99      1880
```

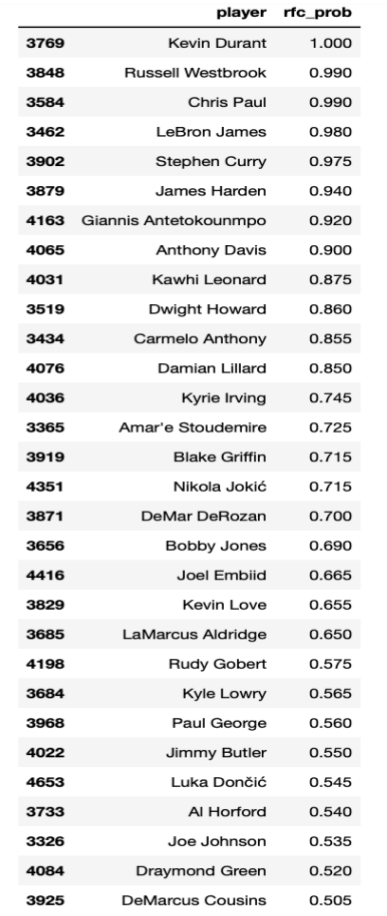


```
=== K Nearest Neighbors ===
Confusion Matrix:
[[1834   40]
 [   5    1]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99      1874
           1       0.02      0.17      0.04         6

    accuracy                           0.98      1880
   macro avg       0.51      0.57      0.52      1880
weighted avg       0.99      0.98      0.98      1880
```



hof	log_pred	log_prob	rfc_pred	rfc_prob	knn_pred	knn_prob	player
0.0	0.0	0.122518	0.0	0.125	1.0	1.0	Gilbert Arenas
0.0	0.0	0.239156	1.0	0.535	1.0	1.0	Joe Johnson
0.0	1.0	0.720139	1.0	0.725	0.0	0.0	Amar'e Stoudemire
0.0	0.0	0.007230	0.0	0.015	1.0	1.0	Junior Harrington
0.0	1.0	0.989717	1.0	0.855	1.0	1.0	Carmelo Anthony
0.0	1.0	1.000000	1.0	0.980	1.0	1.0	LeBron James
0.0	1.0	0.997686	1.0	0.860	0.0	0.0	Dwight Howard

It's evident from the results that the Logistic Regression and Random Forest models provided more realistic predictions compared to the K Nearest Neighbors (KNN) model. The KNN model misclassified players such as Gilbert Arenas or Junior Harrington as potential Hall of Famers, which seems inaccurate. Furthermore, the KNN model failed to classify any of the three players who have already been inducted into the Hall of Fame. This highlights the limitations of the KNN model in this particular classification task.



	player	rfc_prob
3769	Kevin Durant	1.000
3848	Russell Westbrook	0.990
3584	Chris Paul	0.990
3462	LeBron James	0.980
3902	Stephen Curry	0.975
3879	James Harden	0.940
4163	Giannis Antetokounmpo	0.920
4065	Anthony Davis	0.900
4031	Kawhi Leonard	0.875
3519	Dwight Howard	0.860
3434	Carmelo Anthony	0.855
4076	Damian Lillard	0.850
4036	Kyrie Irving	0.745
3365	Amar'e Stoudemire	0.725
3919	Blake Griffin	0.715
4351	Nikola Jokić	0.715
3871	DeMar DeRozan	0.700
3656	Bobby Jones	0.690
4416	Joel Embiid	0.665
3829	Kevin Love	0.655
3685	LaMarcus Aldridge	0.650
4198	Rudy Gobert	0.575
3684	Kyle Lowry	0.565
3968	Paul George	0.560
4022	Jimmy Butler	0.550
4653	Luka Dončić	0.545
3733	Al Horford	0.540
3326	Joe Johnson	0.535
4084	Draymond Green	0.520
3925	DeMarcus Cousins	0.505

In this list, we can find players ordered by their higher probability of being inducted into the Hall of Fame (HoF). The list includes prominent names of current star players and franchise icons who are likely to receive the HoF honor based on their outstanding careers. Additionally, we may come across retired players who, despite their exceptional career performances, did not attain HoF induction, but their achievements would have made them worthy candidates.

Future Improvements

There are several potential enhancements that could be explored in the future to improve the prediction model:

- Include additional player attributes or statistics that may influence HoF induction.
- Experiment with different algorithms or ensemble methods to further improve prediction accuracy.
- Consider incorporating temporal factors, such as the player's era, as it may affect HoF induction criteria.
- Explore advanced feature engineering techniques to derive more informative features.
- Continuously update the model with new data to adapt to changes in the NBA landscape.