Ricky Lee - NBA Hall of Fame Classification Capstone Project

Your project problem statement - the underlying question you are seeking to answer or problem you are addressing i.e. what are the goals of your project?

My goal of this project was to classify current NBA players' probability of making in to hall of fame based on the model trained by historical players dataset. My personal goal was to understand different type of machine learning model and creating basic workflow of data science project from start to finish.

Background on the subject matter area of your dataset - why is this a good problem / subject area to apply data science techniques? How has it been addressed in the past?

There always has been on going argument in NBA about the fairness of NBA's hall of fame player selection. Some players were nominated to hall of fame due to their ethnical backgrounds, and their leading example. For instance, some players like Yao Ming, although statistically he may not be player of hall of fame caliber, have been enrolled in NBA as he was the first Asian NBA player to make impact in the league and bring about global presence in the league. These type of nominations brought concern to public as public viewed NBA Hall of Fame as glorified popularity vote ; other sports, likeness of MLB, have vigorous, statical structure for enrolling players in to Hall of Fame, which in turns brought much greater respect and glory to players being enrolled to.

Datasets I have used are great as it provides direct classification classes with multiple features that are not correlated to one another. Also, features used in the historic eras are current present in current NBA, allowing for easy training and fitting of the model.

Details on the source of the data and the dataset itself (including data format, structure and schema, etc.)

Dataset has been acquired through basketball-reference.com. Data formats are in CSV files, and there were no schemas present. Please see attached CSV file for the in depth detail about the dataset used.

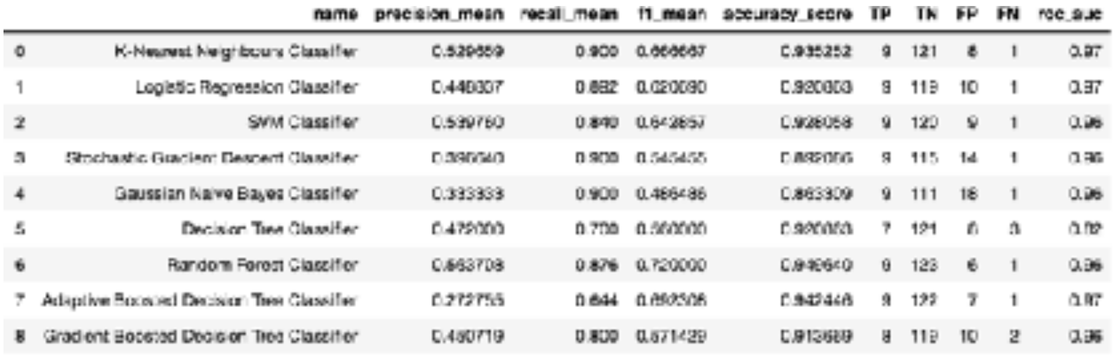
A summary of the preprocessing, feature engineering and any other data cleaning/transformation, and exploratory data analysis (EDA) performed and the motivation and reasoning behind it

- Removed player prior to 1980: Players played before 1980 were removed from dataset because there were no 3 points present in NBA. 3 points play significant role in current NBA and this should be considered for the accurate classification of current NBA players.

- Filtered players played less than 30 games: According to NBA rule, players played less than 30 games are not eligible for seasonal honours like MVP. NBA defines less than 30 games as threshold where players need to play for their data to be eligible for any awards. Therefore, players seasonal data with less than 30 games were removed.
- Replaced all null values with 0: All null values were present as these were reflective upon 0 percent of a specific feature. For instance, player A has null value of Free throw percentage in a season because the player never shot a free throw in the season.
- HOF transformation: Initial dataset did not have a feature that classifies Hall of Fame status. Instead, they had asterisk after their name to reference so. This asterisk has been replaced by actual column value.
- Custom inductee update: Some players that are recently updated are not included in the dataset. These players were manually updated to reference them as hall of fame players
- Feature selection: As the model required optimized feature amount, random forest classifier has been utilized to figure out optimum features for best fitting of the model.

A summary of all the modelling completed including the process of model evaluation, selection, and results

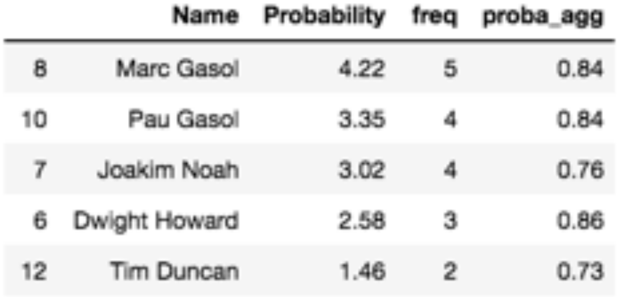
Please reference variable “position_for_model” should you want to check for other positions. Initial code will reflect NBA players in “C” (Centre) Position. Here is the summary of the model and results.
After these models give result for current NBA players, they were aggregated by each player’s respectable probability and their frequency in within a model. For example, player A may be classified as Hall of Fame player in 6 models, at average of 94% probability. Player B may be classified as Hall of Fame player in 5 models, at average of 96 % probability. Player A will be ranked higher in my aggregate resulting model.



	name	precision_mean	recall_mean	f1_mean	accuracy_score	TP	TN	FP	FN	roc_auc
0	K-Nearest Neighbours Classifier	0.529659	0.900	0.666667	0.935252	9	121	8	1	0.97
1	Logistic Regression Classifier	0.440307	0.882	0.620690	0.920863	9	119	10	1	0.97
2	SVM Classifier	0.539760	0.840	0.642857	0.928058	9	120	9	1	0.96
3	Stochastic Gradient Descent Classifier	0.380040	0.900	0.545455	0.892086	9	115	14	1	0.96
4	Gaussian Naive Bayes Classifier	0.333333	0.900	0.486486	0.863309	9	111	18	1	0.96
5	Decision Tree Classifier	0.479000	0.700	0.560000	0.920863	7	121	8	3	0.82
6	Random Forest Classifier	0.663708	0.876	0.720000	0.949640	9	123	6	1	0.96
7	Adaptive Boosted Decision Tree Classifier	0.272755	0.844	0.692308	0.942446	9	122	7	1	0.97
8	Gradient Boosted Decision Tree Classifier	0.460719	0.800	0.571429	0.913669	8	119	10	2	0.96

Findings and conclusions based on all analysis and modelling of the data - how do your results compare against your initial goals & hypotheses?

Here is example of my result:



	Name	Probability	freq	proba_agg
8	Marc Gasol	4.22	5	0.84
10	Pau Gasol	3.35	4	0.84
7	Joakim Noah	3.02	4	0.76
6	Dwight Howard	2.58	3	0.86
12	Tim Duncan	1.46	2	0.73

My initial goal was to provide model that encapsulates all players in NBA. However, as each different positions play significantly different roles, and thus have different emphasis on different features, each models were trained based on positions. My results were pretty accurate, as majority of players with expertise in field, were predicted to be in hall of fame. However, there are room for improvement as players that are noted as hall of famers are not ranked properly. (i.e: Tim Duncan is considered better player than Pau Gasol. However, my model ranked Pau Gasol higher than Tim Duncan)

A final summary of the business applications of the project as well as potential next steps and future directions

There are two major steps to improve for future directions:

- Feature engineering to improve the accuracy and F1-score
- Different clustering method to potentially allow all positions to be considered for hall of fame classification

Business Implication:

- As Hall of Fame players are often considered successful, top 1% player throughout their career, a NBA team can implement this model to figure out which player in college or high school level can reach higher potential. Thus, in draft stage, or when these players are not fully yet developed, NBA team can take advantage of the model to acquire these players whom will have higher chance to blossom as top tier level player.