Data Mining

CSCI-6674-01

DAT (Data Analysis Team)

UNIVERSITY OF NEW HAVEN

TAGLIATELA COLLEGE OF ENGINEERING, WEST HAVEN, CT.

SUBMITTED TO:

Prof. Reza Sadeghi

DAT (Data Analysis Team)

Head of Team:

Anuj Parikh	apari6@unh.newhaven.edu

Team Members:

Anuj Parikh	apari6@unh.newhaven.edu
Rahul Akkineni	rakki1@unh.newhaven.edu

TABLE OF CONTENTS

1. Research Question:

Predicting the 5-year career longevity of the NBA Rookies.

That is, we try to predict whether if a rookie stays for 5-years after coming to the NBA based on his performance during his time.

1.1 Objective:

By predicting longevity of NBA Rookies, our objective is to predict which player is going to last long in NBA and how much score he can make in a single game. At the same time, Coach can choose healthy players for greater winning strategy and energetic game.

1.2 Dataset:

Dataset link: https://data.world/exercises/logistic-regression-exercise-1

- **Description:** The dataset is taken from https://www.basketball-reference.com/ where all the data about the NBA players and their careers is recorded it is an official website for all the basketball players, seasons, and all. So, the data from here is very accurate.

2. Data Optimization:

2.1List of Modeling Techniques:

- Data Modeling can be done in different ways. As we have implemented 15 different modeling techniques in our modeling phase. Now, for data optimization we have done data modeling on selected techniques which have high accuracy in modeling phase. So that we can predict the future with high accuracy. Following list are the modeling techniques which we have used for before standardization and After standardization of the selected features.

- Classification Methods
  - Logistic Regression
  - Naïve Bayes Gaussian
  - SGD (Stochastic Gradient Descent)

2.2 Parameter and hyperparameter:

- **Following hyperparameters are used in above mentioned classification methods:**

- **Logistic Regression:** We used different algorithm for nonlinear optimization as a solver and other parameters such as ‘C’, ‘max_iter’ and ‘Penalty’. For penalty ‘l2’ is the default one which is used by most of the solver with C parameter to control the penalty strength.

- **Naïve Bayes Gaussian:** We set hyperparameter of variance smoothing to 1e-9 for achieving stability.
- **SGD classifier:** In SGD (Stochastic Gradient Descent), we defined parameters such as ‘alpha’, ‘n_iter_no_change’, ‘loss’ and ‘Penalty’. We performed SGD with boundary of 1000 maximum iterations.

Modeling Technique	Parameters/Hyper Parameters
Logistic Regression	‘C’, Penalty, max_iter, Solver
Naïve Bayes Gaussian	Variance smoothing
Stochastic Gradient Descent	Alpha, n_iter_no_change, loss, Penalty

2.3 Optimization Techniques:

- **GridSearchCV**
  - GridSearch is used to modify supervised learning parameters and improve the efficiency. It tries all possible parameters of interest and find the best possible fit. In GridSearchCV implements we used a ‘fit’ method. For Logistic regression we used ‘C’ as a parameter. In Naïve Bayes we used ‘Variance smoothing’ to have a high accuracy. We used scikit learn to implement GridSearchCV.

- **RandomSearch CV**
  - It is like GridSearchCV but in contrast to GridSearchCV, not all parameter values are checked, but a fixed number of parameter settings are sampled from the distributions you specify. n_iter provides the number of parameter settings that are attempted.

- **Bayesian Optimization**
  - In Bayesian Optimization, the algorithm makes an informed guess for each run of hyperparameters on the objective function, which set of hyperparameters is most likely to increase the score and should be attempted in the next run. It is achieved by having regressor forecasts on several points

and selecting the point based on the so-called acquisition function that is the best guess.

3. Performance Metric

- **Accuracy:** We got an accuracy of 72.56% for our model After Standardization by using All features of the dataset which is highest in all modeling techniques we used. By taking 5 best features, we got an accuracy of 72.25% using Naïve bayes gaussian modeling technique and with 10 features we got an accuracy of 71.95% using Logistic Regression.

  - **All Features:**

    - Before Standardization – Logistic Regression
      - GridSearchCV: 71.95%
      - RandomSearch CV: 72.56%
      - Bayesian Optimization: 70.33%

    - After Standardization – Logistic Regression
      - GridSearchCV: 71.64%
      - RandomSearch CV: 72.56%
      - Bayesian Optimization: 70.94%

  - **5 Features:**

    - Before Standardization – Naïve Bayes Gaussian
      - GridSearchCV: 68.90%
      - RandomSearch CV: 68.59%
      - Bayesian Optimization: 69.72%

    - After Standardization – Naïve Bayes Gaussian
      - GridSearchCV: 70.12%
      - RandomSearch CV: 72.25%
      - Bayesian Optimization: 68.80%

    - After Standardization – SGD
      - GridSearchCV: 69.51%
      - RandomSearch CV: 68.90%
      - Bayesian Optimization: 70.32%

- **10 Features:**
  - Before Standardization – Logistic Regression
    - GridSearchCV: 71.34%
    - RandomSearch CV: 71.03%
    - Bayesian Optimization: 71.04%

  - After Standardization – Logistic Regression
    - GridSearchCV: 71.64%
    - RandomSearch CV: 71.34%
    - Bayesian Optimization: 70.83%

- **Precision and Recall:** Precision-Recall is a measure of success of prediction when the class is imbalance. Precision is a measure of result relevancy, while recall is a measure of truly relevant results. When a system has high recall but low precision, then most of the predicted labels are incorrect compared to training labels. Whereas system with high precision and low recall returns few results but all the resulted labels are correct.

- **Features selection graph with respect to importance:**
  - The Table and graph below show the importance of different features used in the model. We find the importance of different feature using Extra Tree Classifier. As we can see that GP is the most important feature which is Game Played.

Features	Importance
GP	0.09583504
Min	0.05336147
PTS	0.05921336
FGM	0.05435793
FGA	0.04830798
FG%	0.06062864
3P Made	0.03063386
3PA	0.04315281
3P%	0.04378697
FTM	0.05529345
FTA	0.05039769
FT%	0.04992243
OREB	0.05490673
DREB	0.05167443
AST	0.05884954
STL	0.04842301
BLK	0.04497474
TOV	0.04562186
Const	0.05065805

Figure 1 Feature Importance table

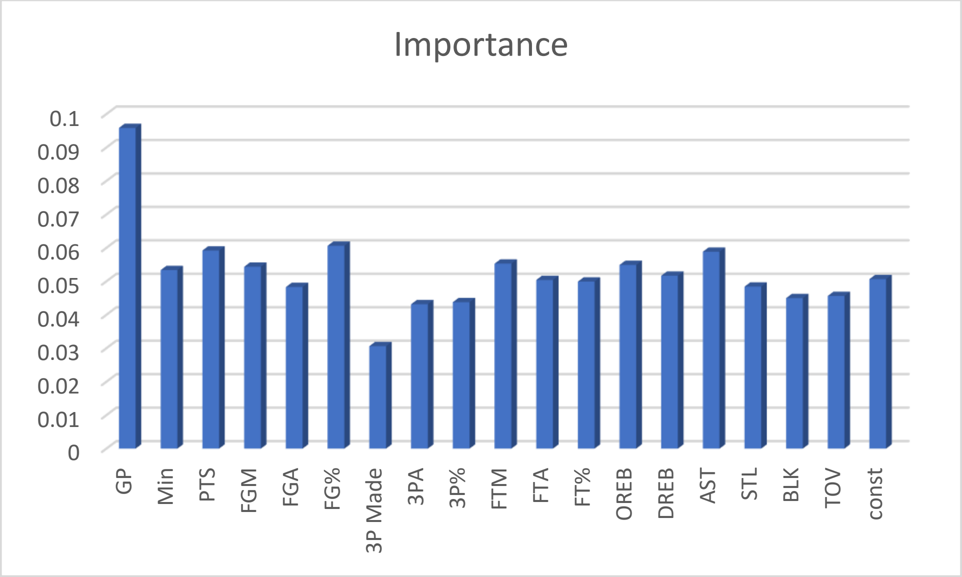

Figure 2 Feature Importance

4. Conclusion:

Initially we performed various tasks such as Data exploration, data modeling and data optimization. After performing various modeling techniques, we get 72.56% accuracy on learning methods, which strongly support the argument of relating player historical performance to longevity of player. Since we observed accuracy of 70+%, we also take few more performance metrics such as precision, recall and f-score under consideration which help us to ensure the result of our outcomes. However, 72.56% is pretty good to deduce that pattern exist in asked research question. We strongly believe that Data Model with larger data can more accurately pinpoint relations between performance and longevity.

GitHub link: https://github.com/Anuj-parikh/DAT