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.

2. Data Exploration:

a. List of Exploration Techniques:
   i. Univariable Analysis
      - Numerical Variable
        - Min, Max, Mean, Median
        - Variance, Standard Deviation
        - R-square, Adjusted R-square, F-statistic
        - Box Plot

b. Interpretation:

i. By performing OLS Regression, we evaluated value of r-squared which is 0.203. Efficiency of the model is into question since R-squared value is significantly at a low level. Although R-squared and Adjusted R-square are having minimal differences suggest that the features in the dataset is relevant to the dependent variable. Here we are considering 2 different models, Intercept model where we consider which contains constant column and another is Specified model containing which contains all the column other than constant. As per our results p-value is closer to the 0 which shows strong evidence in favour of the alternative hypothesis. P-value is closer to 0 and f-statistics is large that means we can reject null hypothesis i.e., specified model fits the data better than intercept model. While rejecting null hypothesis, we can also eliminate features having zero value. For features selection we are using k-best Algorithm and come up with 5 most suitable features for prediction.

For prioritize features according to its importance in predicting target, We use KBest algorithms and then chose top features which play dominant role in desired forecast. Then we will experiment with number of features i.e. k in KBest to figure out appropriate k for better optimization.

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