Analyzing the Dataset of Kobe Bryant Shot Selection

Mingyi Xue*
School of Chemistry and Chemical Engineering, Nanjing University
Wangqian Miao
Kuang Yaming Hornors School, Biophysics, Nanjing University
Rui Wang
School of Business, Nanjing University
Instructor: Prof. Cho-Jui Hsieh
Department of Computer Science & Statistics, University of California, Davis

Abstract

In the real world application of machine learning, it is always difficult to apply the algorithms from books definitely. Most of the time, we find that the result of our machine learning method does not work well without data preparation and feature engineering.

Using 20 years of data on Kobe Bryant's swishes and misses in NBA, we will predict which shots will find the bottom of the net. By this dataset from Kaggle, we practice feature engineering and classification basics with different kinds of machine learning methods. At last, we give the conclusion of our job concretely.

1 Introduction

When applying machine learning method in the real world application, we always do data preprocessing and feature engineering before we use the algorithms, especially, the dataset contains a lot of catogorical variables. However, nowadays, some algorithms can help us a lot from feature engineering (for example, xgboost method).

In this paper, firstly, we transform the dataset into data matrix and then mine the most important information through the plots. As a first try, we apply the machine learing methods including logistic regression, SVM, neural networks. Then we do more feature engineering with PCA and xgboost to make our data matrix more accurate. At last, we compared different algorithms on the new data matrix.

*Three authors are all exchange students from Nanjing University.

2 Data Describing

Name	Variable Kind	Name	Variable Kind
action_type	category	seconds_remaining	int64
combined_shot_type	category	period	int64
game_event_id	category	shot_made_flag	category
game_id	category	shot_type	category
lat	float64	shot_zone_area	category
loc_x	int64	shot_zone_basic	category
loc_y	int64	shot_zone_range	category
lon	float64	team_name	category
minutes_remaining	int64	matchup	category
period	int64	opponent	category
playoffs	category	season	category

Table 1: A summary table for the variables

3 Feature Selection and Engineering

4 Basic Machine Learning Methods

4.1 Logistic Regression

4.2 Support Vector Machine

4.3 Neural Networks

4.4 Algorithm Comparison

5 Dimension Reduction with PCA or XGboost

6 Apply Supervised Learning Method