Proposal of Final Project for hands-on Data Science

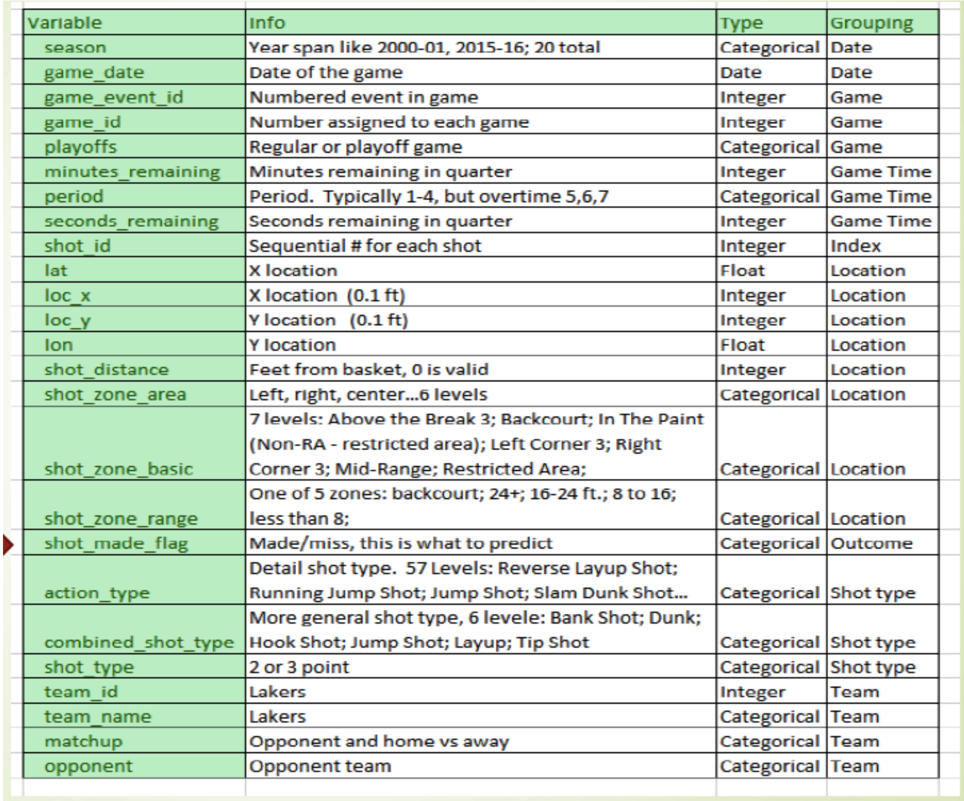
Zhi Wang
23, Sep, 2019
My Github Link: https://github.com/u0820420/1030-Final-Project.git

Overview

Kobe is a polarizing figure in sports. For one of the greatest players of all time, for others a talented but overrated ball hog who got great teams built around him. Whether either one of these assertions is true will probably remain unclear by the end of this notebook, but on the way there we will take a look into the process by which a great player chooses his shots, and how this evolves through time according to his coaches, teammates and physical capabilities.

Data Description and Preprocess

A data set from Kaggle containing the location and circumstances of every field goal attempted by Kobe Bryant took during his 20-year career. The whole dataset size is 30697 rows * 25 columns. Description follows:



Variable	Info	Type	Grouping
season	Year span like 2000-01, 2015-16; 20 total	Categorical	Date
game_date	Date of the game	Date	Date
game_event_id	Numbered event in game	Integer	Game
game_id	Number assigned to each game	Integer	Game
playoffs	Regular or playoff game	Categorical	Game
minutes_remaining	Minutes remaining in quarter	Integer	Game Time
period	Period. Typically 1-4, but overtime 5,6,7	Categorical	Game Time
seconds_remaining	Seconds remaining in quarter	Integer	Game Time
shot_id	Sequential # for each shot	Integer	Index
lat	X location	Float	Location
loc_x	X location (0.1 ft)	Integer	Location
loc_y	Y location (0.1 ft)	Integer	Location
lon	Y location	Float	Location
shot_distance	Feet from basket, 0 is valid	Integer	Location
shot_zone_area	Left, right, center...6 levels	Categorical	Location
shot_zone_basic	7 levels: Above the Break 3; Backcourt; In The Paint (Non-RA - restricted area); Left Corner 3; Right Corner 3; Mid-Range; Restricted Area;	Categorical	Location
shot_zone_range	One of 5 zones: backcourt; 24+; 16-24 ft.; 8 to 16; less than 8;	Categorical	Location
shot_made_flag	Made/miss, this is what to predict	Categorical	Outcome
action_type	Detail shot type. 57 Levels: Reverse Layup Shot; Running Jump Shot; Jump Shot; Slam Dunk Shot...	Categorical	Shot type
combined_shot_type	More general shot type, 6 levele: Bank Shot; Dunk; Hook Shot; Jump Shot; Layup; Tip Shot	Categorical	Shot type
shot_type	2 or 3 point	Categorical	Shot type
team_id	Lakers	Integer	Team
team_name	Lakers	Categorical	Team
matchup	Opponent and home vs away	Categorical	Team
opponent	Opponent team	Categorical	Team

1. I drop the rows in which shot_make_flag has null values, as they are not useful for either training or testing. 25697 remain and 5000 are dropped.
2. Since Kobe only play for Lakers in his career, so I drop the team_id and team_name and drop the shot_id and game_id, because they are not gonna help in the model. Drop lon and lat as well, because they are pairwise correlated 100% to the loc_x and loc_y. (will proof in the report later)
3. Combine the two fields to the time_remaining in seconds until the end of each period and add it to the dataframe.
4. For the purposes of creating an angle feature though we will standarise 'loc_x' and 'loc_y', to avoid zero values that may result in divisions by zero. With standarisation, the data are re-distributed around a mean of zero in one standard deviation distance. Scikit's StandardScaler provides with the facilities to transform training and testing consistently.
5. In order to use Scikit’s classifiers, we need to convert the categorical fields. This can be achieved with Scikit’s OneHotEncoder.
6. Splitting to training and testing set. I will isolate the targetble vector from the features and split the dataset to 80% training and 20% test set.

Model and Features

The problem presented is a simple one: to predict whether each shot is going into the basket or not. It is a classification problem. My data would gain from the kaggle website, named "Kobe Bryant Shot Selection". I will mainly apply decision tree model to analysis Kobe shot behavior. Other models will be tried and make a comparison, such as Random Forest, xgboost, etc. I choose to use decision tree for two reasons:

1. Easy to Understand: Its graphical representation is very intuitive and users can easily relate their hypothesis.
2. They can be used as a baseline that performs well in order to compare with more advanced models, like Random Forests, xgboost etc. They have the additional advantage that you may skip standarisation, normalistaion, feature extraction etc. and other pre-processing steps required by other algorithms.

I’ll start by determining the complexity of the tree to avoid overfitting. I will use an initial method to narrow down the area and then perform k-fold cross-validation to pick the optimal complexity. Then, I will train, test and evaluate the model across most typical metrics, i.e. accuracy, precision, recall, f1-score.

Features are: 'period', 'playoffs', 'shot_distance', 'team_id', 'month', 'year', 'time_remaining', 'home', 'angle', 'action_type', 'combined_shot_type', 'shot_type', 'shot_zone_area', 'shot_zone_basic', 'shot_zone_range', 'opponent'. Totally 16 features.

Other Projects

1. The Project from Kevin Siswandi. Features dropped:
   - shot_id (used as index, not included as a feature)
   - team_name (only one category)
   - team_id (only one category)
   - game_event_id (unique within a game, not related to shots made)
   - game_id (redundant information already contained in matchup/opponent)
   - lon (correlated with loc_x)
   - lat (correlated with loc_y)

Conclusion: A single XGBoost model without much feature engineering performed very well. XGBoost was really efficient in identifying non-linear interactions between features that not much feature engineering was required to get a good score (in fact several features that I generated ended up not used because they did not improve the final score by XGBoost).

2. Kobe Bryant’s Success Against Different Opponents and Throughout His Career. He only use 'shot_made_flag' and 'opponent'. je want this to be plotted in order of best shooting percentage to worst shooting percentage. He is also going to add in a horizontal line indicating Kobe’s overall shooting percentage, which is 43.9%. So there’s something interesting. Kobe shoots better against the Western conference than the Eastern Conference.