Flip 00 project final report

Jin Chen
HuNan University, China

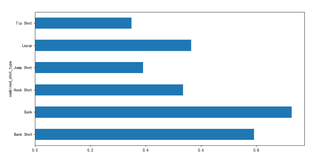
Introduction

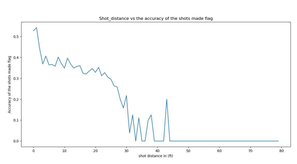
The data contains the location and circumstances of every field goal attempted by Kobe Bryant took during his 20-year career. The task is to predict whether the basket went in (shot_made_flag). The following is the attributes list of data:

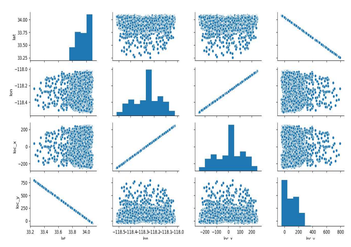
shot_made_flag Yes=1No=0
action_type *Jumpshot,Layup,Dunk, Tipshot,Hookshot,Bankshot*
loc_x ,loc_y shots position
shot_type 2PT Field Goal,2PT Field Goal
shot_zone_area shots area by area
shot_zone_range shots area by radius
shot_zone_basic shots area by NBA rules
shot_made_flag *Yes=1, No=0*

Data Visualization

The following figures show the distribution of the data. The histogram shows the shot accuracy of various action type. The line chart shows the Kobe's shots positioning with distances. And the scatter plot shows that data is distributed normally. and most pairs are widely scattered but some of them show clusters.







Feature Engineering

Data Cleaning:
Some of attributes show clusters: hair_length and seconds_remaining, hair_length and bone_length. So create new variable with multiplication of these columns:
row[total_seconds] = row[seconds_remaining]*row[minutes_remaining]
After that,we can remove the minutes and the seconds columns.

Data Transformation:
Categorical variables such as action_type,combined_shot_type,season,shot_type, shot_zone_range and opponent,we can create the dummy variables for further analysis.

Algorithm

There are many machine learning algorithms for classification problem. Choose the following algorithms, use original train data and new train data to train the models, and determine a set of optimal parameters through Grid Search.

- RandomForest
- LogisticRegression
- KNeighbors

Forcast Result

The following are the best parameters and the Best Score in training of the base models in original train data.
Best Parameters of Models

RandomForest 'criterion': 'entropy', 'max_depth': 5, 'max_features': None, 'n_estimators': 100

LogisticRegression 'C': 1, 'penalty': 'l1'

KNeighbors 'algorithm': 'auto', 'leaf_size': 10, 'n_neighbors': 20, 'p': 5, 'weights': 'uniform'

- Accuracy of these Models

	RF	LR	KNN
Best Score	0.638	0.682	0.568

From the table, it shows that the accuracy of each model is not much different. and it has shown that logistic regression runs better than others.

Conclusion

Exploratory Data Analysis It is an exploratory analysis of the data can provide the necessary conclusions for data processing and modeling.

Data Preprocessing This step contains dealing with missing data and outliers, changing categorical variable into one-hot code and so on.

Feature Engineering It's the most important thing. Create features, then select the most useful features.

Model Training The models have man… Grid Search to find the optimal paratemers.

Acknowledgement
- Thanks!