Flip00 Project Final Presentation

Jin Chen
HuNan University

November 29, 2019

Overview

Problem Statement

Problem Definition

Defn 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).

Train Data and Test Data

There are 30697 lines of data in the training set.I will split the dataset as training sets and testing sets. They have removed 5000 of the shot_made_flags (represented as missing values in the csv file). These are the test set shots for which we need submit a prediction. We are provided a sample submission file with the correct shot_ids needed for a valid prediction.

Data Set

Defn The action_type,shot_made_flag, shot_type and shot_zone_area are part of the attributes of each sample, the fllowings are the meaning of some attributes.

Data List

Attribute	Note
action_type	Jumpshot,Layup,Dunk,Tipshot,Hookshot,Bankshot
loc_x ,loc_y	shots point
shot_made_flag	1=Yes,0=No
shot_type	2PT Field Goal,3PT Field Goal
shot_zone_area	shots area by area
shot_zone_basic	shots area by NBA rules
shot_zone_range	shots area by radius

Exploratory Data Analysis

Data Visualization

Exp

Use EDA to plot the distribution of the data, can observate the data intuitively and find the relation between the attribute values.

- Figures
  - Histogrm
  - Scatter Plot
  - Line Chart

Data Visualization

Exp

It can be seen that dunk is the highest hit rate, followed by bank shot is about 80%, while jump shot and Tip shot are relatively difficult.

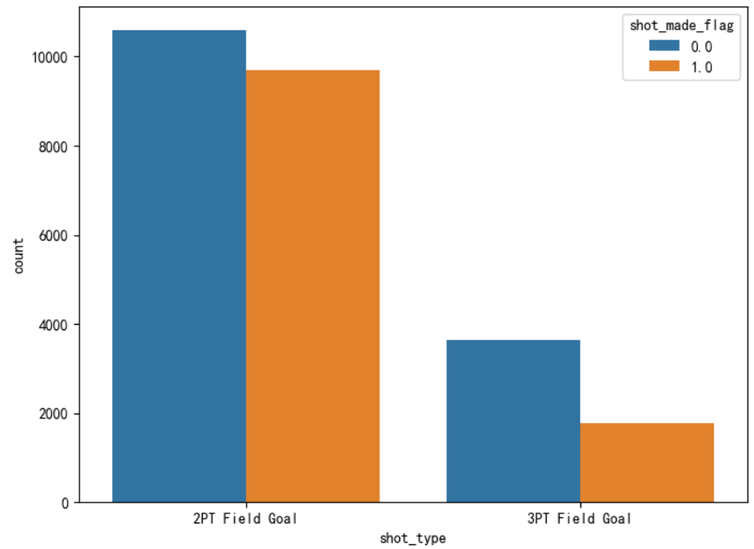

Figure 1: The hit distribution histogram of two shot types

Data Visualization

Exp

It can be seen that dunk is the highest hit rate, followed by bank shot is about 80%, while jump shot and Tip shot are relatively difficult.

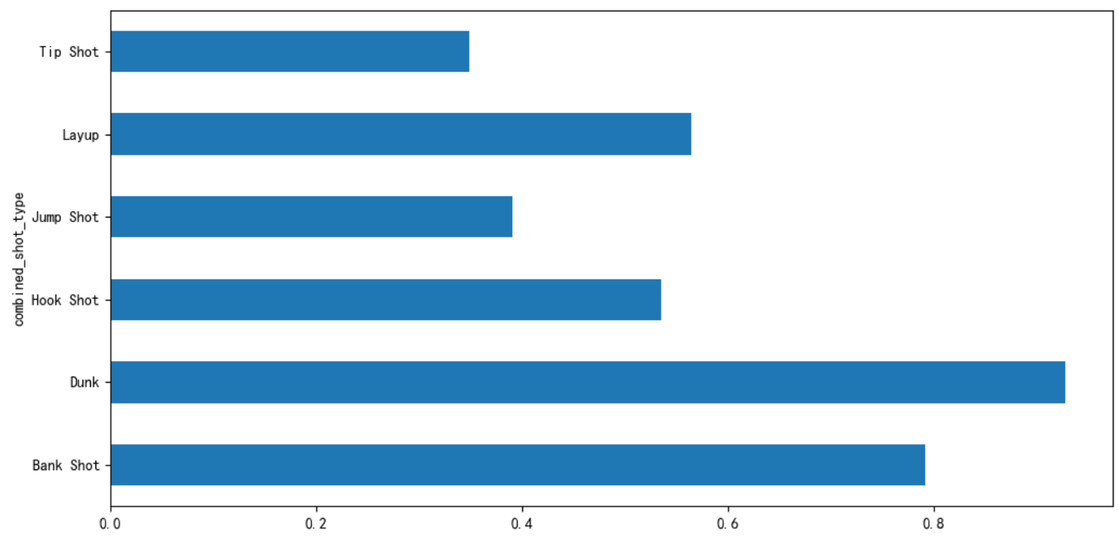

Figure 2: the shot accuracy of various action type

Data Visualization

Exp

Lets get some understanding about the different zones and the shots made from zones.

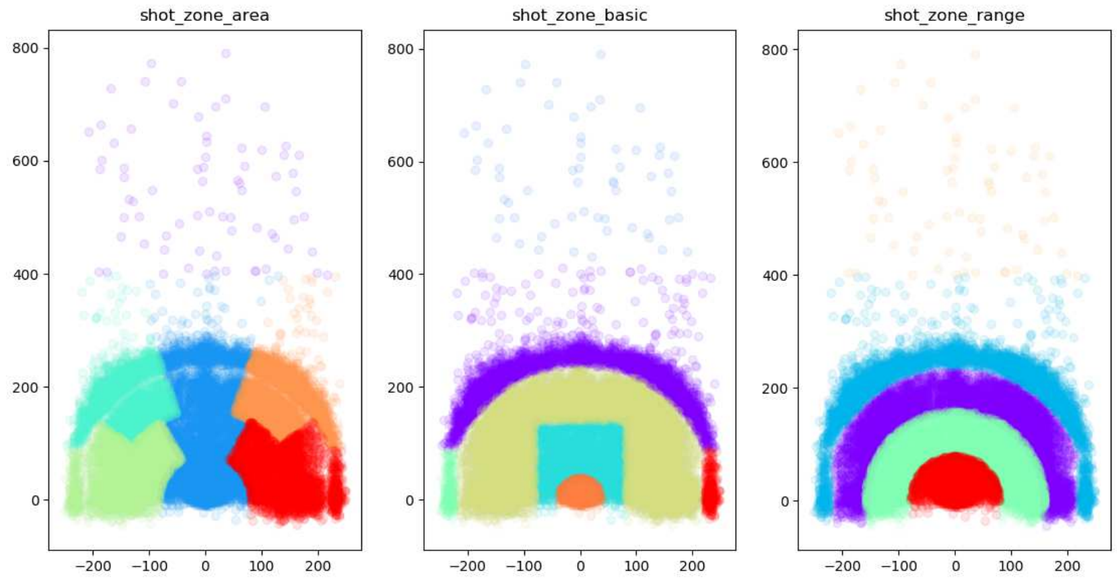

Figure 3: Division of shooting area

Data Visualization

Exp The line chart can not only show the quantity, but also clearly see the increase and decrease of data. Lets now see the Kobe's shots positioning with the time and distance.

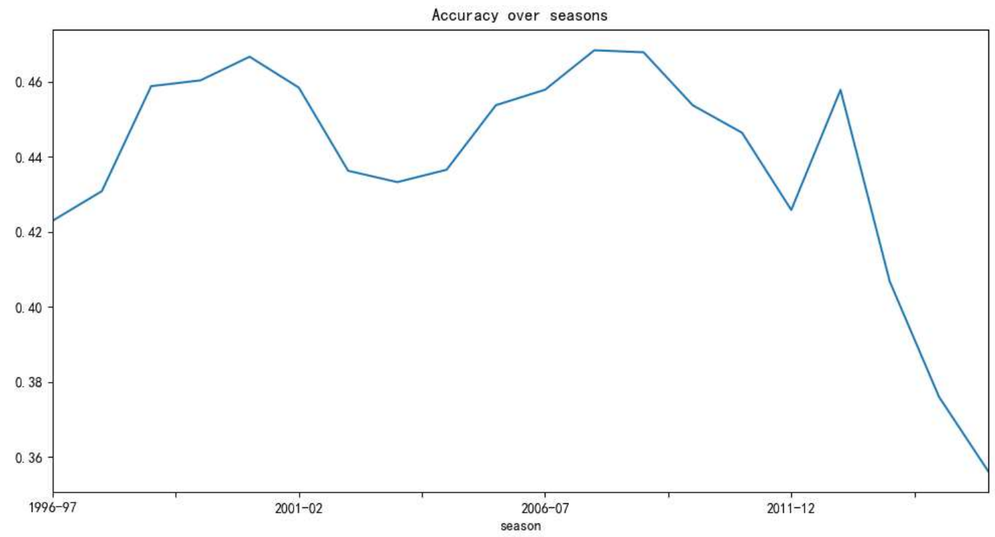

Figure 4: shot accuracy of each seasons

Data Visualization

Exp The line chart can not only show the quantity, but also clearly see the increase and decrease of data. Lets now see the Kobe's shots positioning with the time and distance.

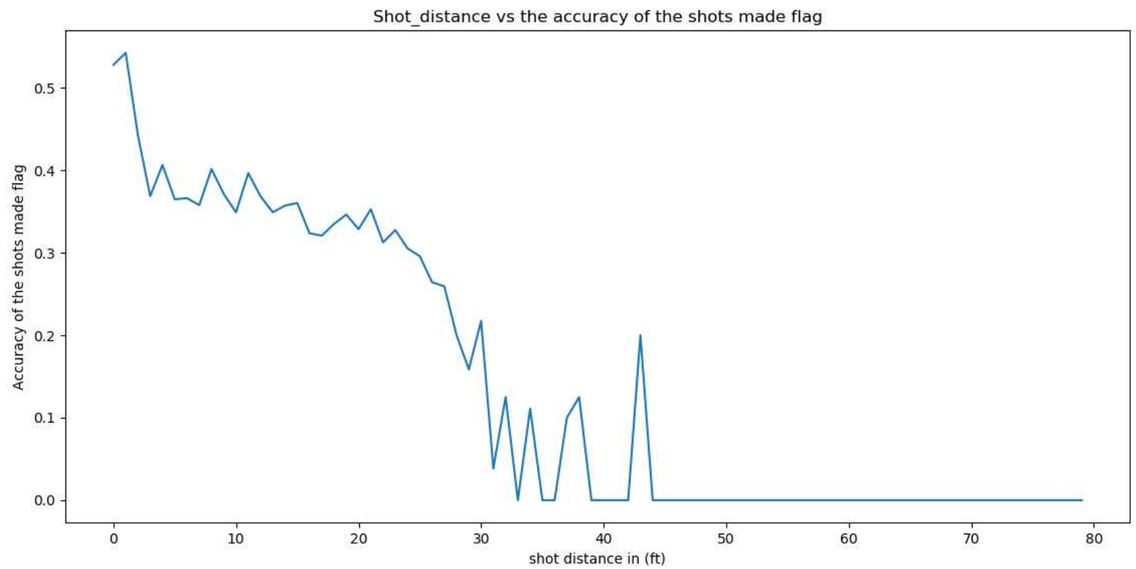

Figure 5: Shot_distance vs the accuracy of the shots made flag

Data Preparation

Data Cleaning

As it can be seen from the picture that, (loc_x ,loc_y) and (lat , lon) represent the same. So, drop one of those.

Meanwhile,some attributes have no attribution for our model, Therefore some columns might be dropped.

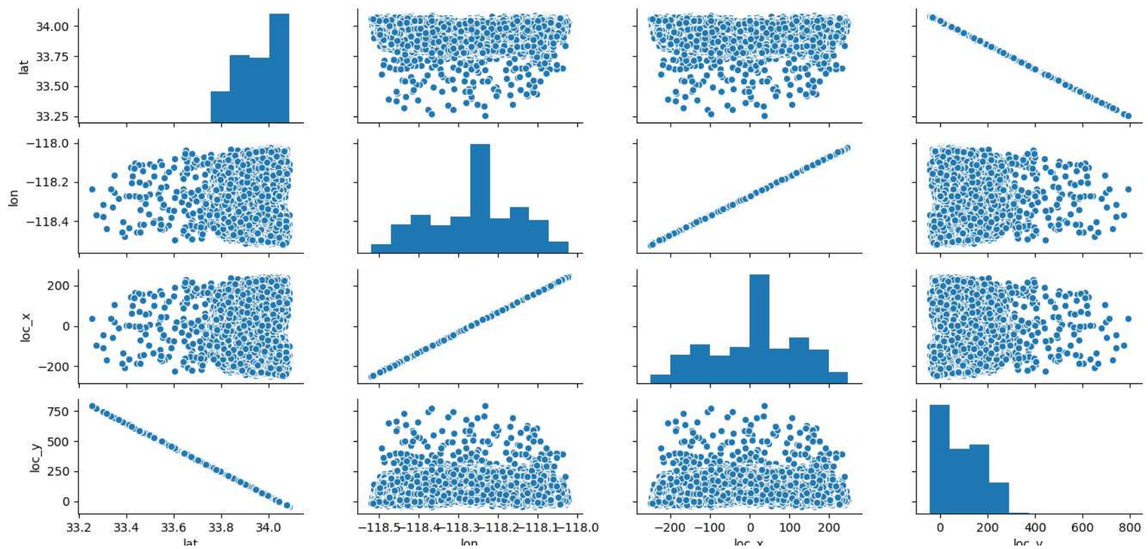

Figure 6: Pairplot of (loc_x ,loc_y) and (lat , lon)

Data Transformation

After deleted all the useless columns,we need to **merge some features**,and **create dummy variables**. First,Let's convert the minutes and seconds to single column.

total_seconds = row[seconds_remaining]+60*row[minutes_remaining]

After that,we can remove the minutes and the seconds columns.

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.

action_type_Cutting Layup Shot	action_type_Driving Bank shot	action_type_Driving Dunk Shot	action_type_Driving Finger Roll Layup Shot	action_type_Driving Finger Roll Shot	action_type_Driving Floating Bank Jump Shot
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	1	0	0	0

Figure 7: part of the converted dataset

Methods

Methods

There are many machine learning algorithms, use the machine learning algorithms below as Ensemble Model's base models. Through Grid Search and ten-fold cross-validation to find the optimal parameters. Then use the ensemble model on test data.

- Base Models
  - RandomForeset
  - LogisticRegression
  - KNeighbors
- Voting ensemble

Parameter Adjustment

The following are optimal parameters of three models.

- 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'

Make final predictions

Forecast Results

The tables below are the accuracy of each model with the adjusted optimal parameter.

- Metrics Classification Report of Ensemble Model in original and new train data

Accuracy of three Models			
	RF	LR	KNN
Best Score	0.637509727626	0.68186770428	0.568404669261

It has shown that logistic regression runs better than others.In the final model, the weights of LR is larger.

Then,use the ensemble model make the final prediction. The final ensemble model's prediction accuracy is **0.738894059077**.
Finally, generate the forecast result and save them in the csv file.

Conclusion

Conclusion

Exploratory Data Analysis It is an exploratory analysis of the data to 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 as more as poosible features, then select the most useful features.

Model Training The models have many parameters, and can use Grid Search to find the optimal paratemers.

Thanks for attending and welcome for questions