FLIP00 FINAL PRESETANTION REPORT

JIN CHEN

Abstract. This report contains four parts. first, introduce the problem and describe the data. Second, visualize the data to find some potential relationships between the attribute values. Afterwards,process the data like creating the dummy variables, feature engineering, feature selection etc. Third, explain the method of experiment.Then, experiment and analyze the performance of different models. The last part is conclusion .

Contents

Date: 2019-11-29.
Key words and phrases. Machine Learning, Data Mining, Binary Classification.

1. Introduction

1.1. **Problem Statement.**

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).

1.2. **Data List.**

For example,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.

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

Table 1. Data Information

1.3. **Problem Analysis.**

1.3.1. *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.

1.3.2. *Problem Possible Solutions.* After analyse the dataset by some simple visualizations. In the process of date preparation,I will engineer the feature to improve the model accuracy, then,create some dummy variables. There are many machine learning moodel can solve the two classification problem, such as The RandomForestClassifier, KNeighbors Classifier and Logistic Regression.Use CV to find the best parameters of the algorithms,and then use these base model construct our ensemble model. Finaly,make the final prediction.

2. Exploratory Data Analysis

2.1. **Data Information.**

The following table 2 is the statistical result of the columns. From this table,we can discover that some of columns have a similar meaning or may not contribute much to the model. so we may need remove and convert them.

Table 2. Data Information

	loc_x	loc_y	lon	min_remaining	sec_remaining
count	30697	30697	30697	30697	30697
mean	7.110499	91.107535	-118.262690	4.885624	28.365085
std	110.124578	87.791361	0.110125	3.449897	17.478949
min	-250.000000	-44.000000	-118.519800	0.000000	0.000000
25%	-68.000000	4.000000	-118.337800	2.000000	13.000000
50%	0.000000	74.000000	-118.269800	5.000000	28.000000
75%	95.000000	160.000000	-118.174800	8.000000	43.000000
max	248.000000	791.000000	-118.021800	11.000000	59.000000

2.2. **Data Visualization.**

Use exploratory data analysis to plot the distribution of the data, can observate the data intuitively and find the relation between the attribute values. For example histogram can visually observe the distribution of numerical variables, scatterplot can show their distribution trends and whether exists outliers. For classification problems, the data with the same label is drawn in same color, which is very helpful for the construction of the Feature.

2.2.1. *Histogram.*

The figure 1 shows the hit distribution of Kobe's shots the figure 2 is the visualization of two kinds of shots(two-point shot and three-point shot). It can be seen that the number of shots and hit rate of the 2PT ball are extremly high. While the percentage of 3PT shots is relatively low. the figure 3 shows the shot accuracy of various action type. 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.

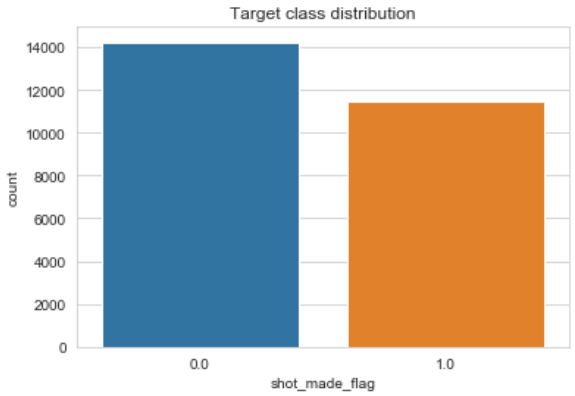

FIGURE 1. target class distribution

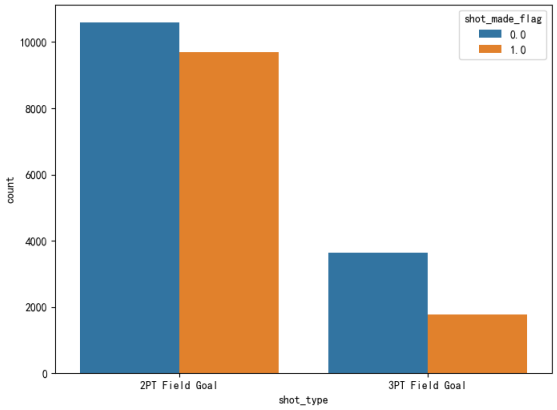

FIGURE 2. The hit distribution histogram of two shot types

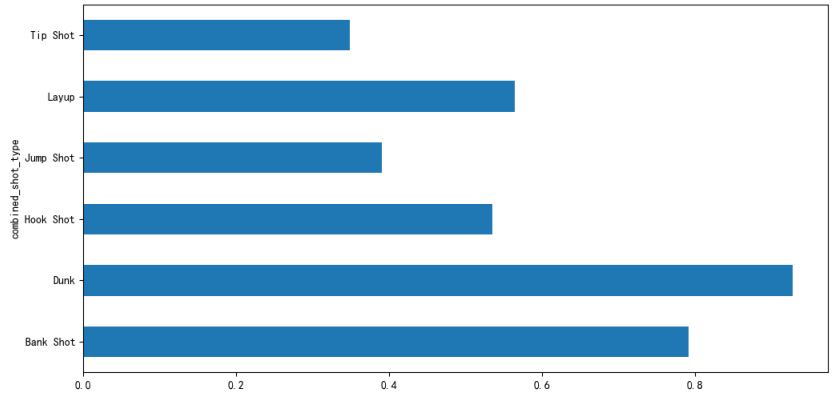

FIGURE 3. the shot accuracy of various action type

2.2.2. *Scatter Plot.*

Using the scatter plot we can combine multiple categorical value series on to the same chart distinguishing them using color or variation in symbol. Lets get some understanding about the different zones and the shots made from zones.

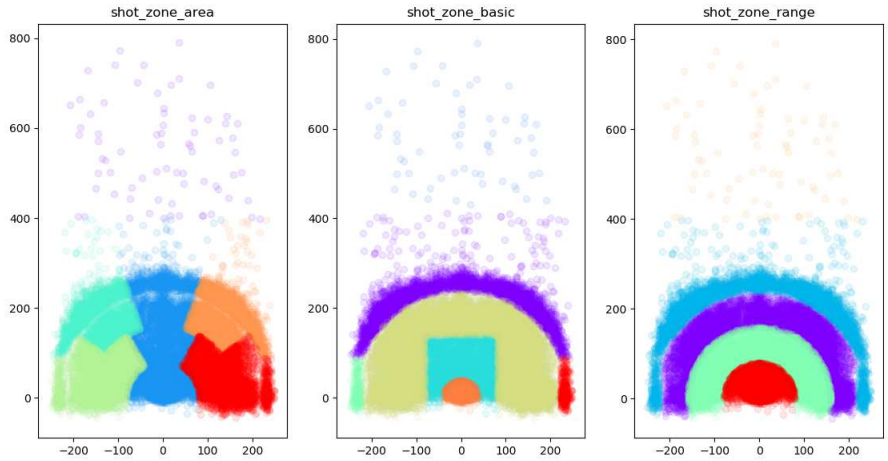

FIGURE 4. Division of shooting area

2.2.3. *Line Chart.*

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.

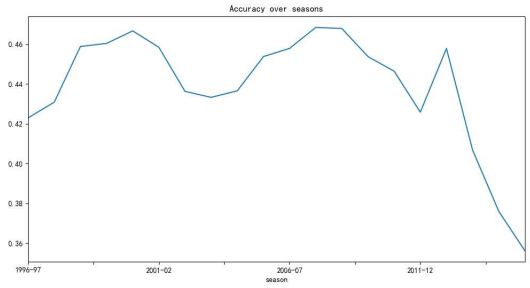

FIGURE 5. shot accuracy of each seasons

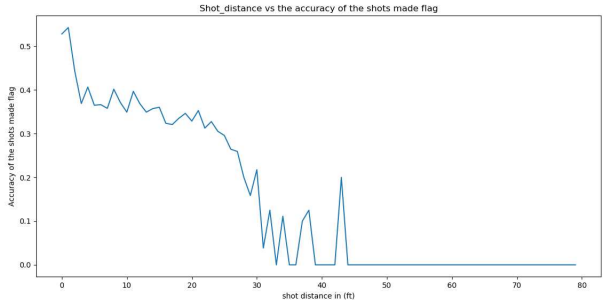

FIGURE 6. Shot˙distance vs the accuracy of the shots made flag

2.3. **Data Preparation.**

2.3.1. *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. similarly,there is no real use of the columns team˙id, team˙name, game˙event˙id, game˙id. So, removing them is a good option.The opponent and the matchup also represents the same thing, so remove the matchup column.The game˙date and shot˙id also has no use. Since, we have equal attacks from both sides we can remove the shot˙zone˙ares as it doesnt contribute much to the model, we can also remove the shot˙zone˙basic for the same reason.

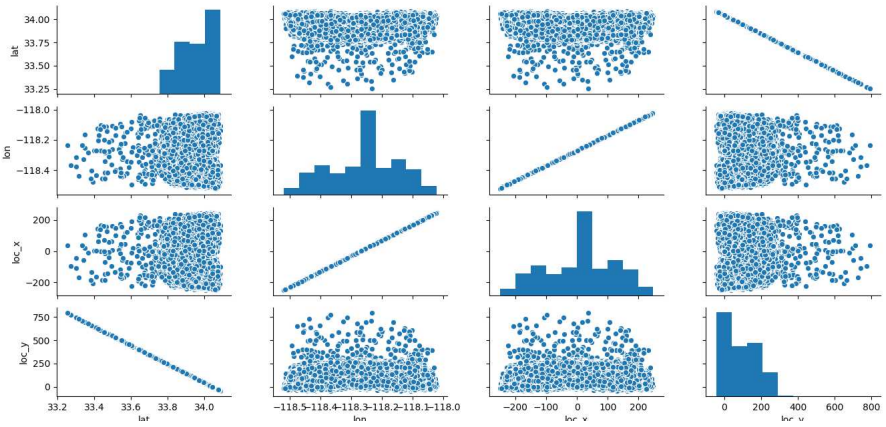

FIGURE 7. Pairplot of (loc_x ,loc_y) and (lat , lon)

2.3.2. *Data Transformation.*

After deleted all the useless columns,we need to merge some features,and create the 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 8. part of the converted dataset

3. Methods

There are many machine learning algorithms for classification problem. Choose the following algorithms as the base models of ensemble model,show the most important parameters.

- RandomForest
- LogisticRegression
- KNN

3.1. **Models.**

The base models have many parameters, select the some parameters that have a larger impact on the forecast results, the use Grid Search to find the optimal paratemers set. The following is training result.

3.1.1. *RandomForest.*

Random forest is a classifier with multiple decision trees, and the output is determined by the mode of the individual tree output.

n˙estimators: the number of decision trees
criteriom: criterion of choosing the most appropriate node
max˙depth: The maximum depth of the tree, the default is None
max˙features: Number of features in feature subset cannot exceed this value.

3.1.2. *LogisticRegression.*

Logistic regression is the algorithm that processing a large amount of observation data to obtain mathematical expressions that are in line with the internal laws of things

Penalty: Regular function
C: Regular coefficient

3.1.3. *KNN.*

The meaning of the knn algorithm is that enter new data without tags, compare each feature of the new data with each feature in the training set, and select the classification tag with the most similar feature (nearest neighbor: k)

n˙neighbors: number of neighbors to use by default for kneighbors queries
leaf˙size: leaf size passed to BallTree or KDTree
p: power parameter for choosing the distance calculation formula
weights: used in prediction
algorithm: compute the nearest neighbors

3.2. **Forcast Result.**

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'
  KNN: 'algorithm': 'auto', 'leaf_size': 10, 'n_neighbors': 20, 'p': 5, 'weights': 'uniform'
- Model Accuracy
  From the Table 3, it shows that the accuracy of each model is not much different. and it has shown that logistic regression runs better than others. In the final model, the weights of LR is larger.

Table 3. Accuracy of these Models

	RF	LR	KNN	ensemble
Best Score	0.637509727626	0.68186770428	0.568404669261	0.738894059077

It can be seen that the ensemble model performs better in testing. Then,use the ensemble model make the final prediction.Finally, generate the forecast result and save them in the csv file.

4. Conclusion

- Exploratory data analysis is very important for the competition, that is an exploratory analysis of the data to provide the necessary conclusions for data processing and modeling.
- The data that we have, needed processed in many cases. Data preprocessing includes deal with missing data and outliers, change categorical variable into one-hot code and so on.
- The most important thing is feature engineering. We can create as more as poosible features, then select the most useful features.
- Model training is also very important. There are many algorithms, in my opinoin, if the time permits, we can We can try all the algorithms.
- The last thing is adjustment, for example, the models have many parameters, can use Grid Search to find the optimal paratemers.