

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 pairplot shows that data is distributed normally. and most pairs are widely scattered but some of them show clusters. Through correlogram can gain that it is no obvious linear relationship between variables. And boxplot shows the outliers are very small, which can be ignored.

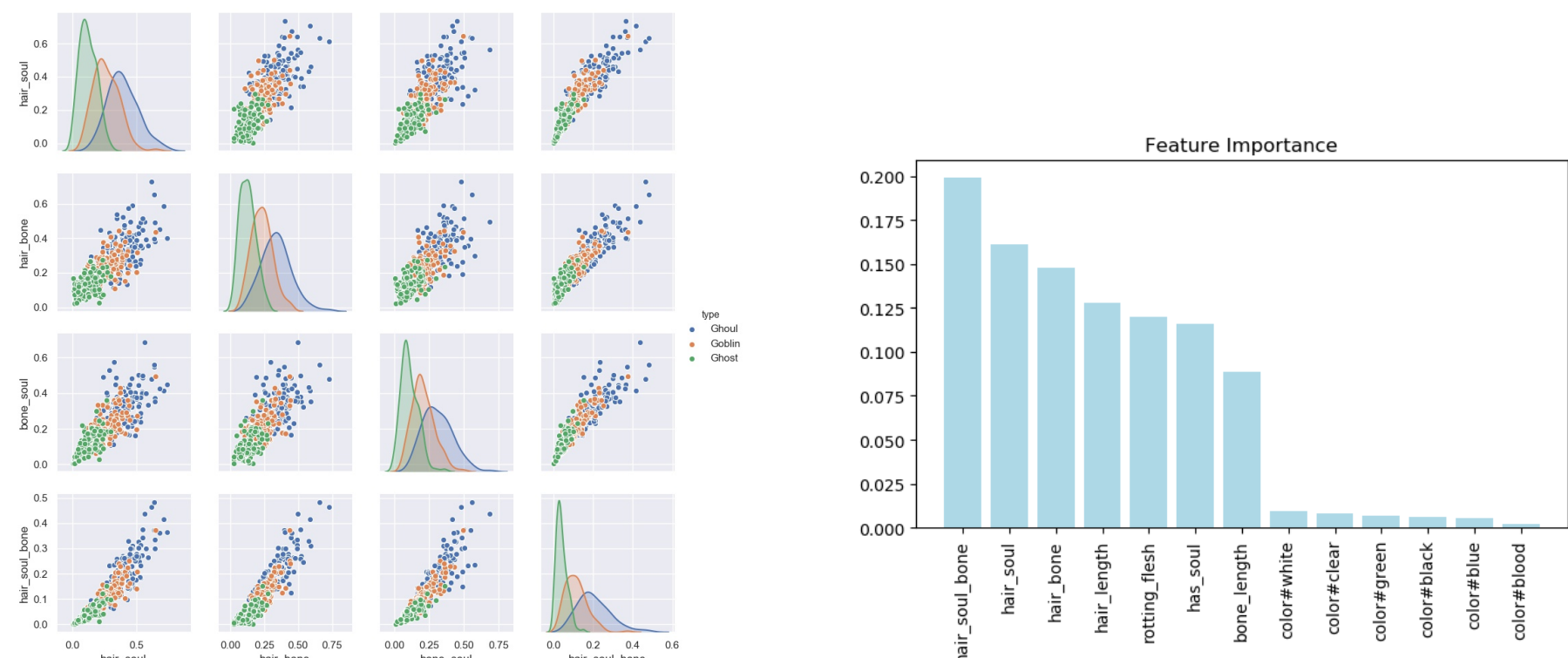


Feature Engineering

Some of attributes show clusters: hair\_length and has\_soul, hair\_length and bone\_length. So create new variables with multiplication of these columns:

- hair\_soul 'hair\_length' \* 'has\_soul'
- hair\_bone 'hair\_length' \* 'bone\_length'
- bone\_soul 'bone\_length' \* 'has\_soul'
- hair\_soul\_bone 'hair\_length' \* 'has\_soul' \* 'bone\_length'

Using the Feature Importance this function of Random Forest to select the most important features to form a new train data. The two picture, one is pairplot which is plotted by using new features data. another is the bar plot which shows the importance of features



Algorithm

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. Because the train data is relatively small, a ten-fold cross-validation is used.

- RandomForest
- LogisticRegression
- SVC
- KNeighbors
- XGBoost
- Netual Network

Take the trained models as the base models of ensemble model, and average the prediction results by using voting

Algorithm

Experiment Result

The tables below are the metrics classification report of ensemble model in original and new train data.

- Metrics Classification Report of Ensemble Model in original train data

	precision	recall	f1-score	support
Ghost	0.80	0.83	0.82	24
Ghoul	0.88	0.79	0.84	29
Goblin	0.67	0.73	0.70	22
micro avg	0.79	0.79	0.79	75
macro avg	0.78	0.78	0.78	75
weighted avg	0.79	0.79	0.79	75

- Metrics Classification Report of Ensemble Model in new train data

	precision	recall	f1-score	support
Ghost	0.84	0.88	0.86	24
Ghoul	0.93	0.97	0.95	29
Goblin	0.80	0.73	0.76	22
micro avg	0.87	0.87	0.87	75
macro avg	0.86	0.86	0.86	75
weighted avg	0.86	0.87	0.86	75

It can be observed that ensemble model performaces better in new features.

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 new features, then select the most useful features

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

Acknowledgement  
• Thanks!