摘要

英超聯賽（通常被稱為英格蘭以外的英超聯賽（EPL））是英格蘭足球聯賽體系的頂級聯賽。由 20 個俱樂部共同組成，它與英格蘭足球聯盟（EFL）一起推廣。四季從八月到五月，每支球隊打 38 場比賽（主場和客場比賽）。大多數比賽在周六和周日下午進行。自成立以來，英超聯賽共有 47 支英國俱樂部和兩支威爾士俱樂部，使其成為跨界聯賽。

介紹

本研究主要利用資料集的主隊與客隊來預測比賽結果。

資料集介紹及資料集來源

4560 場英超聯賽的比賽結果，從 2006/2007 至 2017/2018，共有 12 個賽季的 380 場比賽。本研究使用的資料集欄位含有 home_team(主隊)、away_team(客隊)、home_goals(主隊分數)、away_goals(客隊分數)、result(結果:H 表示主隊勝利,A 表示客隊勝利,D 表示平手)、season(賽季)。
資料集來源: https://www.kaggle.com/zaeemnalla/premier-league

資料預處理

2.預處理

In [224]:
```
#missing data
total = df.isnull().sum().sort_values(ascending=False)
percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data.head(20)
```

Out[224]:

	Total	Percent
season	0	0.0
result	0	0.0
away_goals	0	0.0
home_goals	0	0.0
away_team	0	0.0
home_team	0	0.0

3.欄位轉換數字

In [226]:
```
df['home_team'] = df['home_team'].replace({'Liverpool':1,'Arsenal':2,'Everton':3,'Manchester City':4 ,'Chelsea':5, 'Manchester United':6, 'Tottenham Hots
df.sample(5)
```

Out[226]:

	home_team	away_team	home_goals	away_goals	result	season
2172	12	Liverpool	1.0	0.0	H	2011-2012
3124	13	Crystal Palace	2.0	2.0	D	2014-2015
1979	32	Swansea City	2.0	2.0	D	2011-2012
3038	7	Aston Villa	3.0	0.0	H	2013-2014
1554	9	Blackpool	0.0	2.0	A	2010-2011

In [227]:
```
df['away_team'] = df['away_team'].replace({'Liverpool':1,'Arsenal':2,'Everton':3,'Manchester City':4 ,'Chelsea':5, 'Manchester United':6, 'Tottenham Hotsp
df.sample(5)
```

Out[227]:

	home_team	away_team	home_goals	away_goals	result	season
4287	4	2	3.0	1.0	H	2017-2018
433	16	14	1.0	1.0	D	2007-2008
1759	18	7	0.0	1.0	A	2010-2011
1556	8	5	1.0	3.0	A	2010-2011
1033	16	8	0.0	1.0	A	2008-2009

機器學習或深度學習方法

1. Decision tree

In [230]:
```
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics

tree = DecisionTreeClassifier(criterion='gini',max_depth=5)
tree.fit(X_train, y_train)
print(metrics.classification_report(y_test, tree.predict(X_test)))
```

```
             precision    recall  f1-score   support

          A       0.50      0.45      0.48       390
          D       0.28      0.05      0.08       349
          H       0.55      0.85      0.67       629

avg / total       0.47      0.53      0.46      1368
```

2. AdaBoost

In [231]:
```
from sklearn.ensemble import AdaBoostClassifier
adb = AdaBoostClassifier(n_estimators=100)
adb.fit(X_train, y_train)
print(metrics.classification_report(y_test, adb.predict(X_test)))
```

```
             precision    recall  f1-score   support

          A       0.55      0.41      0.47       390
          D       0.30      0.04      0.07       349
          H       0.55      0.90      0.68       629

avg / total       0.49      0.54      0.47      1368
```

3. XGBoost

In [232]:
```
import xgboost as xgb
xgbc = xgb.XGBClassifier()
xgbc.fit(X_train, y_train)
```

Out[232]:
```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0,
       max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
       n_jobs=1, nthread=None, objective='multi:softprob', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
       silent=True, subsample=1)
```

In [233]:
```
xgbc.score(X_test, y_test)
```

Out[233]:
```
0.5460526315789473
```

研究結果及討論

研究結果的準確度都偏低最高也只有 0.54，因為資料集的特徵不夠多，導致無法更準確的預測結果，也可能代表著每場球賽的主隊與客隊都有運動家精神，並無打假球的情況。

參考文獻

https://www.kaggle.com/zaeemnalla/premier-league

https://en.wikipedia.org/wiki/Premier_League

https://www.premierleague.com/results