In [1]:
```python
import pandas as pd
import pickle
import numpy as np
```

In [2]:
```python
df = pickle.load(open('dataset_level2.pkl','rb'))
```

In [3]:
```python
df

# batting_team
# bowling team
# city
# current score
# ball left
# wickets left
# current rr
# last five
```

Out[3]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	venue
0	2	Australia	Sri Lanka	0.1	0	0	NaN	Melbourne Cricket Ground
1	2	Australia	Sri Lanka	0.2	0	0	NaN	Melbourne Cricket Ground
2	2	Australia	Sri Lanka	0.3	1	0	NaN	Melbourne Cricket Ground
3	2	Australia	Sri Lanka	0.4	2	0	NaN	Melbourne Cricket Ground
4	2	Australia	Sri Lanka	0.5	0	0	NaN	Melbourne Cricket Ground
...	...	...	...	...	...	...	...	...
121	964	Sri Lanka	Australia	19.3	1	0	Colombo	R Premadasa Stadium
122	964	Sri Lanka	Australia	19.4	0	0	Colombo	R Premadasa Stadium
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	R Premadasa Stadium
124	964	Sri Lanka	Australia	19.6	2	0	Colombo	R Premadasa Stadium
125	964	Sri Lanka	Australia	19.7	1	0	Colombo	R Premadasa Stadium

63888 rows × 8 columns

In [4]:
```python
df.isnull().sum()
```

Out[4]:
```
match_id               0
batting_team           0
bowling_team           0
ball                   0
runs                   0
player_dismissed       0
city                8548
venue                  0
dtype: int64
```

In [5]:
```python
df[df['city'].isnull()]['venue'].value_counts()
```

Out[5]:
```
Dubai International Cricket Stadium        2969
Pallekele International Cricket Stadium    2066
Melbourne Cricket Ground                   1453
Sydney Cricket Ground                       749
Adelaide Oval                               498
Harare Sports Club                          372
Sharjah Cricket Stadium                     249
Sylhet International Cricket Stadium        128
Carrara Oval                                 64
Name: venue, dtype: int64
```

In [6]:
```python
cities = np.where(df['city'].isnull(),df['venue'].str.split().apply(lambda x:x[0]),df
```

In [7]:
```python
df['city'] = cities
```

In [8]:
```python
df.isnull().sum()
```

Out[8]:
```
match_id            0
batting_team        0
bowling_team        0
ball                0
runs                0
player_dismissed    0
city                0
venue               0
dtype: int64
```

In [9]:
```python
df.drop(columns=['venue'],inplace=True)
```

In [10]:
```python
df
```

Out[10]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne
...	...	...	...	...	...	...	...
121	964	Sri Lanka	Australia	19.3	1	0	Colombo
122	964	Sri Lanka	Australia	19.4	0	0	Colombo
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo
124	964	Sri Lanka	Australia	19.6	2	0	Colombo
125	964	Sri Lanka	Australia	19.7	1	0	Colombo

63888 rows × 7 columns

In [11]:
```python
eligible_cities = df['city'].value_counts()[df['city'].value_counts() > 600].index.tol
```

In [12]:
```python
df = df[df['city'].isin(eligible_cities)]
```

In [13]:
```python
df
```

Out[13]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne
...	...	...	...	...	...	...	...
121	964	Sri Lanka	Australia	19.3	1	0	Colombo
122	964	Sri Lanka	Australia	19.4	0	0	Colombo
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo
124	964	Sri Lanka	Australia	19.6	2	0	Colombo
125	964	Sri Lanka	Australia	19.7	1	0	Colombo

50501 rows × 7 columns

In [14]:
```python
df['current_score'] = df.groupby('match_id').cumsum()['runs']
```

```
<ipython-input-14-ac174c139314>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['current_score'] = df.groupby('match_id').cumsum()['runs']
```

In [15]:
```python
df
```

Out[15]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	current_score
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	0
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	0
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	1
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	3
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	3
...	...	...	...	...	...	...	...	...
121	964	Sri Lanka	Australia	19.3	1	0	Colombo	125
122	964	Sri Lanka	Australia	19.4	0	0	Colombo	125
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	125
124	964	Sri Lanka	Australia	19.6	2	0	Colombo	127
125	964	Sri Lanka	Australia	19.7	1	0	Colombo	128

50501 rows × 8 columns

In [16]:
```python
df['over'] = df['ball'].apply(lambda x:str(x).split(".")[0])
df['ball_no'] = df['ball'].apply(lambda x:str(x).split(".")[1])
df
```

```
<ipython-input-16-4fd5a57c1fcd>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['over'] = df['ball'].apply(lambda x:str(x).split(".")[0])
<ipython-input-16-4fd5a57c1fcd>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['ball_no'] = df['ball'].apply(lambda x:str(x).split(".")[1])
```

Out[16]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	current_score	over	ball
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	0	0	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	0	0	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	1	0	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	3	0	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	3	0	
...	...	...	...	...	...	...	...	...	...	
121	964	Sri Lanka	Australia	19.3	1	0	Colombo	125	19	
122	964	Sri Lanka	Australia	19.4	0	0	Colombo	125	19	
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	125	19	
124	964	Sri Lanka	Australia	19.6	2	0	Colombo	127	19	
125	964	Sri Lanka	Australia	19.7	1	0	Colombo	128	19	

50501 rows × 10 columns

In [17]:
```python
df['balls_bowled'] = (df['over'].astype('int')*6) + df['ball_no'].astype('int')
df
```

```
<ipython-input-17-e7d17f656852>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['balls_bowled'] = (df['over'].astype('int')*6) + df['ball_no'].astype('int')
```

Out[17]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	current_score	over	ball
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	0	0	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	0	0	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	1	0	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	3	0	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	3	0	
...	...	...	...	...	...	...	...	...	...	
121	964	Sri Lanka	Australia	19.3	1	0	Colombo	125	19	
122	964	Sri Lanka	Australia	19.4	0	0	Colombo	125	19	
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	125	19	
124	964	Sri Lanka	Australia	19.6	2	0	Colombo	127	19	
125	964	Sri Lanka	Australia	19.7	1	0	Colombo	128	19	

50501 rows × 11 columns

In [18]:
```python
df['balls_left'] = 120 - df['balls_bowled']
df['balls_left'] = df['balls_left'].apply(lambda x:0 if x<0 else x)
df
```

```
<ipython-input-18-7ed065e8960c>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['balls_left'] = 120 - df['balls_bowled']
<ipython-input-18-7ed065e8960c>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['balls_left'] = df['balls_left'].apply(lambda x:0 if x<0 else x)
```

Out[18]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	current_score	over	ball
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	0	0	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	0	0	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	1	0	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	3	0	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	3	0	
...	...	...	...	...	...	...	...	...	...	
121	964	Sri Lanka	Australia	19.3	1	0	Colombo	125	19	
122	964	Sri Lanka	Australia	19.4	0	0	Colombo	125	19	
123	964	Sri Lanka	Australia	19.5	0	DM de Silva	Colombo	125	19	
124	964	Sri Lanka	Australia	19.6	2	0	Colombo	127	19	
125	964	Sri Lanka	Australia	19.7	1	0	Colombo	128	19	

50501 rows × 12 columns

In [19]:
```python
df['player_dismissed'] = df['player_dismissed'].apply(lambda x:0 if x=='0' else 1)
df['player_dismissed'] = df['player_dismissed'].astype('int')
df['player_dismissed'] = df.groupby('match_id').cumsum()['player_dismissed']
df['wickets_left'] = 10 - df['player_dismissed']
```

```
<ipython-input-19-e0234112d42f>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['player_dismissed'] = df['player_dismissed'].apply(lambda x:0 if x=='0' else 1)
<ipython-input-19-e0234112d42f>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['player_dismissed'] = df['player_dismissed'].astype('int')
<ipython-input-19-e0234112d42f>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['player_dismissed'] = df.groupby('match_id').cumsum()['player_dismissed']
<ipython-input-19-e0234112d42f>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['wickets_left'] = 10 - df['player_dismissed']
```

In [20]:
```python
df
```

Out[20]:

	match_id	batting_team	bowling_team	ball	runs	player_dismissed	city	current_score	over	ball
0	2	Australia	Sri Lanka	0.1	0	0	Melbourne	0	0	
1	2	Australia	Sri Lanka	0.2	0	0	Melbourne	0	0	
2	2	Australia	Sri Lanka	0.3	1	0	Melbourne	1	0	
3	2	Australia	Sri Lanka	0.4	2	0	Melbourne	3	0	
4	2	Australia	Sri Lanka	0.5	0	0	Melbourne	3	0	
...	...	...	...	...	...	...	...	...	...	
121	964	Sri Lanka	Australia	19.3	1	8	Colombo	125	19	
122	964	Sri Lanka	Australia	19.4	0	8	Colombo	125	19	
123	964	Sri Lanka	Australia	19.5	0	9	Colombo	125	19	
124	964	Sri Lanka	Australia	19.6	2	9	Colombo	127	19	
125	964	Sri Lanka	Australia	19.7	1	9	Colombo	128	19	

50501 rows × 13 columns

In [21]:
```python
df['crr'] = (df['current_score']*6)/df['balls_bowled']
```

```
<ipython-input-21-fad47fc85776>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['crr'] = (df['current_score']*6)/df['balls_bowled']
```

In [22]:
```python
groups = df.groupby('match_id')

match_ids = df['match_id'].unique()
last_five = []
for id in match_ids:
    last_five.extend(groups.get_group(id).rolling(window=30).sum()['runs'].values.tol
```

In [23]:
```python
df['last_five'] = last_five
```

```
<ipython-input-23-83d24d575aa4>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/use
r_guide/indexing.html#returning-a-view-versus-a-copy
  df['last_five'] = last_five
```

In [24]:
```python
final_df = df.groupby('match_id').sum()['runs'].reset_index().merge(df,on='match_id')
```

In [25]:
```python
final_df=final_df[['batting_team','bowling_team','city','current_score','balls_left',
```

In [26]:
```python
final_df.dropna(inplace=True)
```

In [27]:
```python
final_df.isnull().sum()
```

Out[27]:
```
batting_team     0
bowling_team     0
city             0
current_score    0
balls_left       0
wickets_left     0
crr              0
last_five        0
runs_x           0
dtype: int64
```

In [28]:
```python
final_df = final_df.sample(final_df.shape[0])
```

In [29]:
```python
final_df.sample(2)
```

Out[29]:

	batting_team	bowling_team	city	current_score	balls_left	wickets_left	crr	last_five	runs_x
31884	India	England	Colombo	40	92	9	8.571429	39.0	170
43622	South Africa	Bangladesh	Mirpur	65	78	10	9.285714	43.0	169

In [30]:
```python
X = final_df.drop(columns=['runs_x'])
y = final_df['runs_x']
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=1)
```

In [31]:
```python
X_train
```

Out[31]:

	batting_team	bowling_team	city	current_score	balls_left	wickets_left	crr	last_five
13017	Bangladesh	Pakistan	Lahore	86	31	6	5.797753	29.0
47570	South Africa	Afghanistan	Mumbai	173	17	7	10.077670	69.0
41191	South Africa	India	Mirpur	170	1	6	8.571429	45.0
5330	India	South Africa	Cape Town	135	19	6	8.019802	32.0
19865	Australia	West Indies	London	115	25	5	7.263158	41.0
...	...	...	...	...	...	...	...	...
47171	Afghanistan	Sri Lanka	Kolkata	47	61	7	4.779661	17.0
33184	West Indies	New Zealand	Pallekele	107	32	4	7.295455	22.0
25171	England	South Africa	Barbados	125	31	6	8.426966	44.0
23284	New Zealand	Sri Lanka	Colombo	49	70	8	5.880000	30.0
17856	India	Australia	Durban	90	44	8	7.105263	50.0

30781 rows × 8 columns

In [32]:
```python
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.metrics import r2_score,mean_absolute_error
```

In [33]:
```python
trf = ColumnTransformer([
    ('trf',OneHotEncoder(sparse=False,drop='first'),['batting_team','bowling_team','ci
]
,remainder='passthrough')
```

In [34]:
```python
pipe = Pipeline(steps=[
    ('step1',trf),
    ('step2',StandardScaler()),
    ('step3',XGBRegressor(n_estimators=1000,learning_rate=0.2,max_depth=12,random_stat
])
```

In [ ]:
```python
pipe.fit(X_train,y_train)
y_pred = pipe.predict(X_test)
print(r2_score(y_test,y_pred))
print(mean_absolute_error(y_test,y_pred))
```

In [ ]:
```python
pickle.dump(pipe,open('pipe1.pkl','wb'))
```

In [ ]:
```python
import xgboost
xgboost.__version__
```

In [ ]: