DSO110 - Final Project - Lottery Prediction

In [1]:
```
import pandas as pd
import numpy as np
from matplotlib import pyplot
import warnings
import re
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifi
from sklearn.model_selection import cross_val_predict, cross_val_score, train
# from xgboost import XGBClassifier
import seaborn as sns
```

In [2]:
```
df = pd.read_csv('Lottery_Mega_Millions_Winning_Numbers__Beginning_2002.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	Draw Date	Month	Day	Year	Weekday	Weekday.1	Quarter	Winning Numbers	Mega Ball	Multiplier	Fi
0	9/25/2020	9	25	2020	Fri	1	3	20 36 37 48 67	16	2.0	
1	9/29/2020	9	29	2020	Tue	0	3	14 39 43 44 67	19	3.0	
2	10/2/2020	10	2	2020	Fri	1	4	09 38 47 49 68	25	2.0	
3	10/6/2020	10	6	2020	Tue	0	4	15 16 18 39 59	17	3.0	
4	10/9/2020	10	9	2020	Fri	1	4	05 11 25 27 64	13	2.0	

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

Out[4]:
```
Draw Date            0
Month                0
Day                  0
Year                 0
Weekday              0
Weekday.1            0
Quarter              0
Winning Numbers      0
Mega Ball            0
Multiplier         903
First                0
Second               0
Third                0
Fourth               0
Fifth                0
dtype: int64
```

In [5]:
```
df.describe()
```

Out[5]:

	Month	Day	Year	Weekday.1	Quarter	Mega Ball	Mult
count	2036.000000	2036.000000	2036.000000	2036.000000	2036.000000	2036.000000	1133.00
mean	6.559921	15.726916	2011.638507	0.500000	2.521611	18.612475	3.36
std	3.424417	8.802003	5.651637	0.500123	1.111488	13.195995	1.01
min	1.000000	1.000000	2002.000000	0.000000	1.000000	1.000000	2.00
25%	4.000000	8.000000	2007.000000	0.000000	2.000000	8.000000	3.00
50%	7.000000	16.000000	2012.000000	0.500000	3.000000	15.000000	3.00
75%	10.000000	23.000000	2017.000000	1.000000	4.000000	28.000000	4.00
max	12.000000	31.000000	2021.000000	1.000000	4.000000	52.000000	5.00

In [6]:
```
df.set_index('Draw Date',inplace=True)
```

In [7]:
```
df
```

Out[7]:

Draw Date	Month	Day	Year	Weekday	Weekday.1	Quarter	Winning Numbers	Mega Ball	Multiplier	Firs
9/25/2020	9	25	2020	Fri	1	3	20 36 37 48 67	16	2.0	2(
9/29/2020	9	29	2020	Tue	0	3	14 39 43 44 67	19	3.0	14
10/2/2020	10	2	2020	Fri	1	4	09 38 47 49 68	25	2.0	9
10/6/2020	10	6	2020	Tue	0	4	15 16 18 39 59	17	3.0	15
10/9/2020	10	9	2020	Fri	1	4	05 11 25 27 64	13	2.0	5
...	...	...	...	...	...	...	...	...	...	..
11/12/2021	11	12	2021	Fri	1	4	30 32 42 46 48	15	2.0	3(
11/16/2021	11	16	2021	Tue	0	4	06 22 44 53 65	3	3.0	6
11/19/2021	11	19	2021	Fri	1	4	05 23 52 53 59	18	5.0	5
11/23/2021	11	23	2021	Tue	0	4	07 24 54 57 58	6	3.0	7
11/26/2021	11	26	2021	Fri	1	4	07 27 37 42 59	2	2.0	7

2036 rows × 14 columns

In [8]:
```
df2 = df.drop(["Month","Day","Year","Weekday","Weekday.1","Quarter",'First','
```

In [9]:
```
df2
```

Out[9]:

Draw Date	Winning Numbers	Mega Ball	Multiplier
9/25/2020	20 36 37 48 67	16	2.0
9/29/2020	14 39 43 44 67	19	3.0
10/2/2020	09 38 47 49 68	25	2.0
10/6/2020	15 16 18 39 59	17	3.0
10/9/2020	05 11 25 27 64	13	2.0
...	...	...	...
11/12/2021	30 32 42 46 48	15	2.0
11/16/2021	06 22 44 53 65	3	3.0
11/19/2021	05 23 52 53 59	18	5.0
11/23/2021	07 24 54 57 58	6	3.0
11/26/2021	07 27 37 42 59	2	2.0

2036 rows × 3 columns

In [10]:
```
d2 = df2['Winning Numbers']
df2[['Ball1','Ball2',"Ball3","Ball4","Ball5"]] =  df2["Winning Numbers"].str.
df2
```

Out[10]:

Draw Date	Winning Numbers	Mega Ball	Multiplier	Ball1	Ball2	Ball3	Ball4	Ball5
9/25/2020	20 36 37 48 67	16	2.0	20	36	37	48	67
9/29/2020	14 39 43 44 67	19	3.0	14	39	43	44	67
10/2/2020	09 38 47 49 68	25	2.0	09	38	47	49	68
10/6/2020	15 16 18 39 59	17	3.0	15	16	18	39	59
10/9/2020	05 11 25 27 64	13	2.0	05	11	25	27	64
...	...	...	...	...	...	...	...	...
11/12/2021	30 32 42 46 48	15	2.0	30	32	42	46	48
11/16/2021	06 22 44 53 65	3	3.0	06	22	44	53	65
11/19/2021	05 23 52 53 59	18	5.0	05	23	52	53	59
11/23/2021	07 24 54 57 58	6	3.0	07	24	54	57	58
11/26/2021	07 27 37 42 59	2	2.0	07	27	37	42	59

2036 rows × 8 columns

In [11]:
```
winning_numbers = df2.drop(['Winning Numbers'],axis=1)
```

In [12]:
```
winning_numbers
```

Out[12]:

Draw Date	Mega Ball	Multiplier	Ball1	Ball2	Ball3	Ball4	Ball5
9/25/2020	16	2.0	20	36	37	48	67
9/29/2020	19	3.0	14	39	43	44	67
10/2/2020	25	2.0	09	38	47	49	68
10/6/2020	17	3.0	15	16	18	39	59
10/9/2020	13	2.0	05	11	25	27	64
...	...	...	...	...	...	...	...
11/12/2021	15	2.0	30	32	42	46	48
11/16/2021	3	3.0	06	22	44	53	65
11/19/2021	18	5.0	05	23	52	53	59
11/23/2021	6	3.0	07	24	54	57	58
11/26/2021	2	2.0	07	27	37	42	59

2036 rows × 7 columns

In [13]:
```
winning_numbers['Ball1'] =winning_numbers['Ball1'].astype('category')
winning_numbers['Ball2'] =winning_numbers['Ball2'].astype('category')
winning_numbers['Ball3'] =winning_numbers['Ball3'].astype('category')
winning_numbers['Ball4'] =winning_numbers['Ball4'].astype('category')
winning_numbers['Ball5'] =winning_numbers['Ball5'].astype('category')

cat_feat = ['Mega Ball','Multiplier']
for feat in cat_feat:
    winning_numbers[feat] = winning_numbers[feat].astype('category')
```

In [14]:
```
winning_numbers.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2036 entries, 9/25/2020 to 11/26/2021
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Mega Ball   2036 non-null   category
 1   Multiplier  1133 non-null   category
 2   Ball1       2036 non-null   category
 3   Ball2       2036 non-null   category
 4   Ball3       2036 non-null   category
 5   Ball4       2036 non-null   category
 6   Ball5       2036 non-null   category
dtypes: category(7)
memory usage: 44.0+ KB
```

In [15]:
```
winning_numbers = winning_numbers.rename(columns={'Mega Ball':'MegaBall',"Mul
```

In [16]:
```
X = winning_numbers.drop(["MegaBall"],axis=1)
y = winning_numbers['MegaBall']
```

In [17]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
models = []
models.append(("LR",LogisticRegression(solver='liblinear')))
print(models)
```

```
[('LR', LogisticRegression(solver='liblinear'))]
```

In [18]:
```
# import statsmodels.formula.api as smf
# model_fit3 = smf.logit(formula='MegaBall ~C(Ball1)+C(Ball2)', data=winning_
```

In [ ]:
```
```