Dodgers Game Attendance

Author: Alissa Trujillo

As one of the United States' leading baseball franchises, the Dodgers draw immense crowds to their stadium yearly. Are there key factors that are linked to higher attendance? Do certain opponents encourage higher ticket sales? Are there other incentives such as gifts or themed nights that create higher demand for tickets?

We will take a look at data that has recorded attendance over the 2022 season and see if we can determine any trends. We will then plan a proposal for the team to maximize attendance for the 2023 season.

Importing Packages

In [ ]:
```
import pandas as pd
import numpy as np
```

Importing the Dataset

In [2]:
```
dodgers_df = pd.read_csv("dodgers-2022.csv")
```

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

Out[3]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	NO	NO	NO
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	NO	NO	NO
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	NO	NO	NO
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	NO	NO	YES
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	NO	NO	NO

Detecting Missing Values

In [4]:
```
dodgers_df.isna().sum()
```

Out[4]:
```
month          0
day            0
attend         0
day_of_week    0
opponent       0
temp           0
skies          0
day_night      0
cap            0
shirt          0
fireworks      0
bobblehead     0
dtype: int64
```

This dataset has no missing values that we need to fill in.

Detecting Outliers

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

Out[5]:

	day	attend	temp
count	81.000000	81.000000	81.000000
mean	16.135802	41040.074074	73.148148
std	9.605666	8297.539460	8.317318
min	1.000000	24312.000000	54.000000
25%	8.000000	34493.000000	67.000000
50%	15.000000	40284.000000	73.000000
75%	25.000000	46588.000000	79.000000
max	31.000000	56000.000000	95.000000

None of our numeric variables have any outstanding values. The range for day of the month is 1-31, which is correct. The range for attendance is between 24,312 and 56,000 which are reasonable expectations for attendance. The temperature has a range of 54 degrees to 95 degrees, which is an accurate representation of the Spring and Summer seasons in LA.

Converting Yes/No Values to Binary Features

In [6]:
```
dodgers_df['cap'] = dodgers_df['cap'].map({'YES': 1, 'NO': 0})
dodgers_df['shirt'] = dodgers_df['shirt'].map({'YES': 1, 'NO': 0})
dodgers_df['fireworks'] = dodgers_df['fireworks'].map({'YES': 1, 'NO': 0})
dodgers_df['bobblehead'] = dodgers_df['bobblehead'].map({'YES': 1, 'NO': 0})
```

In [7]:
```
dodgers_df.head()
```

Out[7]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	0	0	0
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	0	0	0
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	0	0	0
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	0	0	1
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	0	0	0

Feature Engineering

Our data has multiple features relating to different gifts that may be given out during a game. Since they each have mostly "no" values with a few "yes" values, I think it may be valuable to create a new feature that tells us whether there was any type of gift given at the game.

In [8]:
```
dodgers_df['gift'] = dodgers_df['cap'] + dodgers_df['shirt'] + dodgers_df['bobb

for i in dodgers_df['gift']:
    if i > 0:
        i = 1

dodgers_df['gift'].unique()
```

Out[8]:
```
array([0, 1])
```

In [9]:
```
dodgers_df.head()
```

Out[9]:

	month	day	attend	day_of_week	opponent	temp	skies	day_night	cap	shirt	fireworks
0	APR	10	56000	Tuesday	Pirates	67	Clear	Day	0	0	0
1	APR	11	29729	Wednesday	Pirates	58	Cloudy	Night	0	0	0
2	APR	12	28328	Thursday	Pirates	57	Cloudy	Night	0	0	0
3	APR	13	31601	Friday	Padres	54	Cloudy	Night	0	0	1
4	APR	14	46549	Saturday	Padres	57	Cloudy	Night	0	0	0

Defining Features and Target Variable

In [10]:
```
features = dodgers_df.drop('attend', axis=1)
target = dodgers_df['attend']
```

Converting Categorical Categories to Dummy Variables

In [11]:
```
features = pd.get_dummies(features)

features.head()
```

Out[11]:

	day	temp	cap	shirt	fireworks	bobblehead	gift	month_APR	month_AUG	month_JUL	..
0	10	67	0	0	0	0	0	1	0	0	..
1	11	58	0	0	0	0	0	1	0	0	..
2	12	57	0	0	0	0	0	1	0	0	..
3	13	54	0	0	1	0	0	1	0	0	..
4	14	57	0	0	0	0	0	1	0	0	..

5 rows × 42 columns

Split into Train/Test Sets

In [12]:
```
from sklearn.model_selection import train_test_split
```

In [13]:
```
X_train, X_test, y_train, y_test = train_test_split(features, target,
                                                    random_state=23,
                                                    test_size=0.2)
```

In [14]:
```
X_train.head()
```

Out[14]:

	day	temp	cap	shirt	fireworks	bobblehead	gift	month_APR	month_AUG	month_JUL
7	24	63	0	0	0	0	0	1	0	0
44	14	75	0	0	0	1	1	0	0	1
11	29	74	0	1	0	0	1	1	0	0
20	18	64	0	0	1	0	0	0	0	0
34	16	68	0	0	0	0	0	0	0	0

5 rows × 42 columns

In [15]:
```
y_train.head()
```

Out[15]:
```
7     44014
44    54014
11    48753
20    40906
34    45210
Name: attend, dtype: int64
```

Creating a Baseline Model

In [16]:
```
from sklearn.linear_model import LinearRegression
```

In [17]:
```
model = LinearRegression()
model.fit(X_train, y_train)
```

Out[17]:
```
▾ LinearRegression
LinearRegression()
```

In [18]:
```
model.score(X_test, y_test)
```

Out[18]:
```
0.21268116858039
```

By taking a look at our model's R2 score, we can see that roughly 21% of the variance in attendance can be explained by the features included in this model.

In [19]:
```
pd.DataFrame(zip(features.columns, model.coef_))
```

Out[19]:

	0	1
0	day	151.105050
1	temp	136.254647
2	cap	-3996.833769
3	shirt	-2981.103295
4	fireworks	20988.539481
5	bobblehead	7740.263068
6	gift	762.326003
7	month_APR	266.343728
8	month_AUG	5724.346344
9	month_JUL	-1091.481860
10	month_JUN	-10481.862600
11	month_MAY	2234.729191
12	month_OCT	1755.509202
13	month_SEP	1592.415996
14	day_of_week_Friday	-17770.856568
15	day_of_week_Monday	2149.938247
16	day_of_week_Saturday	3363.683743
17	day_of_week_Sunday	549.405545
18	day_of_week_Thursday	2518.608284
19	day_of_week_Tuesday	8625.386734
20	day_of_week_Wednesday	563.834014
21	opponent_Angels	18966.961021
22	opponent_Astros	-10427.375306
23	opponent_Braves	-5034.721671
24	opponent_Brewers	-11276.082767
25	opponent_Cardinals	-2345.683532
26	opponent_Cubs	-568.129354
27	opponent_Giants	-6225.688001
28	opponent_Marlins	-7510.322687
29	opponent_Mets	11637.442606
30	opponent_Nationals	5586.539911
31	opponent_Padres	4594.297696
32	opponent_Phillies	5371.602127
33	opponent_Pirates	-285.474512
34	opponent_Reds	-2423.978189

	0	1
35	opponent_Rockies	-7529.659764
36	opponent_Snakes	-9068.824549
37	opponent_White Sox	16539.096973
38	skies_Clear	1372.877538
39	skies_Cloudy	-1372.877538
40	day_night_Day	970.368936
41	day_night_Night	-970.368936

Taking a look at our baseline model, we can see the coefficients associated with each of our features. The inclusion of fireworks, all other variables held constant, is associated with an increase in attendance of 20,999 patrons. This is likely not a causal relationship, but may be due to the fact that fireworks are done on special evenings (season opener, fourth of july, etc.) which might be an outside influence on attendance.

There are a number of opponent teams that have high positive coefficients as well. All other variables held constant, a game featuring the Los Angeles Angels is associated with an average increase in attendance of 18,966 patrons and a game featuring the Chicago White Sox is associated with an average increase in attendance of 16,539. The Angels may be explained by their fans being local, however the White Sox attendance spike is not necessarily intuitive.

Fine Tuning A Model

The base model had a fairly low R2 score, so we will see if by limiting the dimensionality we can improve the model. We will select the 10 most important features in order to create a more streamlined model.

In [20]:
```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_regression
```

In [21]:
```
fs = SelectKBest(score_func=f_regression, k=10)
fs.fit(X_train, y_train)

X_train_fs = fs.transform(X_train)
X_test_fs = fs.transform(X_test)
```

Feature Scores

In [22]:
```
pd.DataFrame(zip(features.columns, fs.scores_)).head(10)
```

Out[22]:

	0	1
0	day	0.063523
1	temp	0.146998
2	cap	0.322082
3	shirt	1.206310
4	fireworks	0.098013
5	bobblehead	31.907081
6	gift	24.288936
7	month_APR	0.020963
8	month_AUG	0.329819
9	month_JUL	1.210964

This table shows us the scores for each of our feature variables. The higher the number, the more relevant the feature is in predicting our target variable: attendance.

Our New Model

In [23]:
```
model2 = LinearRegression()
model2.fit(X_train_fs, y_train)
```

Out[23]:
```
▾ LinearRegression
LinearRegression()
```

In [24]:
```
model2.score(X_test_fs, y_test)
```

Out[24]:
```
0.4157255064864367
```

Our new and improved model has an R2 score of 0.42. This is an improvement over our original baseline model. By selecting the best 10 features, we can eliminate unnecessary dimensionality and streamline our predictions. This shows that 42% of variation in attendance can be explained by these 10 variables.

In [27]:
```
mask = fs.get_support()
new_features = features.columns[mask]
pd.DataFrame(zip(new_features, fs.scores_[fs.get_support()], model2.coef_))
```

Out[27]:

	0	1	2
0	bobblehead	31.907081	10027.482787
1	gift	24.288936	250.468730
2	month_JUN	6.853000	5115.002080
3	month_MAY	6.851322	-3486.406453
4	day_of_week_Monday	7.000466	-3770.841137
5	day_of_week_Tuesday	6.212078	3443.659615
6	opponent_Angels	3.221889	1484.929026
7	opponent_Brewers	2.906064	-2642.662298
8	opponent_Mets	4.232818	834.033362
9	opponent_Nationals	3.075549	6456.300557

This table shows our most important features and their associated scores and coefficients. The model determined that receiving a gift, specifically a bobblehead, was critical in determining a game's attendance. The months of May and June were also important indicators, with May having a negative correlation and June having a positive correlation. The day the game took place was also an important indicator, with Monday having a negative correlation and Tuesday having a positive correlation. A small selection of opponent options (Angels, Brewers, Mets and Nationals) were also relevant when assessing the size of the crowd.

Conclusion

There are many factors that go into a fan's decision to attend a baseball game at Dodger Stadium. It is difficult to determine causation for many of the features described, as there are potential other variables (special events, holidays, ticket prices) that may influence both attendance and our feature variables (gifts given, fireworks displayed).

The model discarded many features I expected to be important, including weather and temperature. Though in hindsight, this makes sense as most patron's purchase tickets before that information is determined. Details that are secured well in advance, such as the day of the week or the opponent, had larger effects on ticket sales.

This model gives us a bit of insight into factors that influence crowd size, however only 42% of variance in attendance is explained by the model. Due to contractual constraints, the team does not have control over most of the features that were determined important. The only factor the stadium can control is whether they give a gift or not. Everything else, in terms of when the game is played and who it is played against, is determined pre-season.

My recommendation to management is to determine the causal element in the relationship between bobblehead gifts and attendance. If they can determine that bobbleheads are the driver for attendance, they can introduce more nights with bobblehead gifts for fans. If there

is in fact a third casual element (special nights or player tributes that are accompanied by bobblehead gifts), then adding more special themed nights would help attendance increase.

I do not believe that the team has control over the schedule, but if they are able to influence it, manipulating the game days would help the Dodgers fill their stands. If they were able to concentrate more games into the month of June rather than May, that would allow more guests to attend. Also, if they are able to host more games on Tuesdays rather than Mondays they may draw larger crowds.

Another suggestion that is more feasible: if the team is able to increase incentives, such as gifts, on days where they are playing less popular teams (such as the Brewers) they may be able to gather larger crowds for less enticing games.

References

Emily Inestroza. (2019). Dodgers Game Day Information, Version 1. Retrieved March 31, 2022 from https://www.kaggle.com/datasets/meluchatrojan123/dodgers-game-day-information