Springboard Data Science Capstone 3 Project
Author: Michael Palazzolo
Springboard Mentor: Kenneth Gil Pasquel

MLB Game Attendance Prediction

Table Of Contents

1 Introduction

The MLB generates approximately 4 Billion USD in revenue annually. In person stadium attendance is a connection between the owners of sports teams and their fan base. It is a staggering investment to build and maintain a sports stadium, as such if we were able to more accurately predict the fan attendance rates at a stadium it would allow the owners to reduce food waste and allocate staffing more precisely.

In this report we seek to investigate factors that are key drivers for predicting fan attendance and build machine learning models to support this investigation.

2 Data Sources

The Data for this machine learning project was taken from two sources.

2.1 Baseball Game Statistics Data

Source - https://www.baseball-reference.com

This data source offers data going back to 1980 for all professional baseball games that have taken place, it includes information such as:

- Home team
- Opponent

- MLB ranking
- Win/Loss streak
- Game date
- Score
- Team data
- Innings played
- Day or night game

2.2 MLB Stadium Characteristics

Source - https://www.wikiwand.com/en/List_of_current_Major_League_Baseball_stadiums

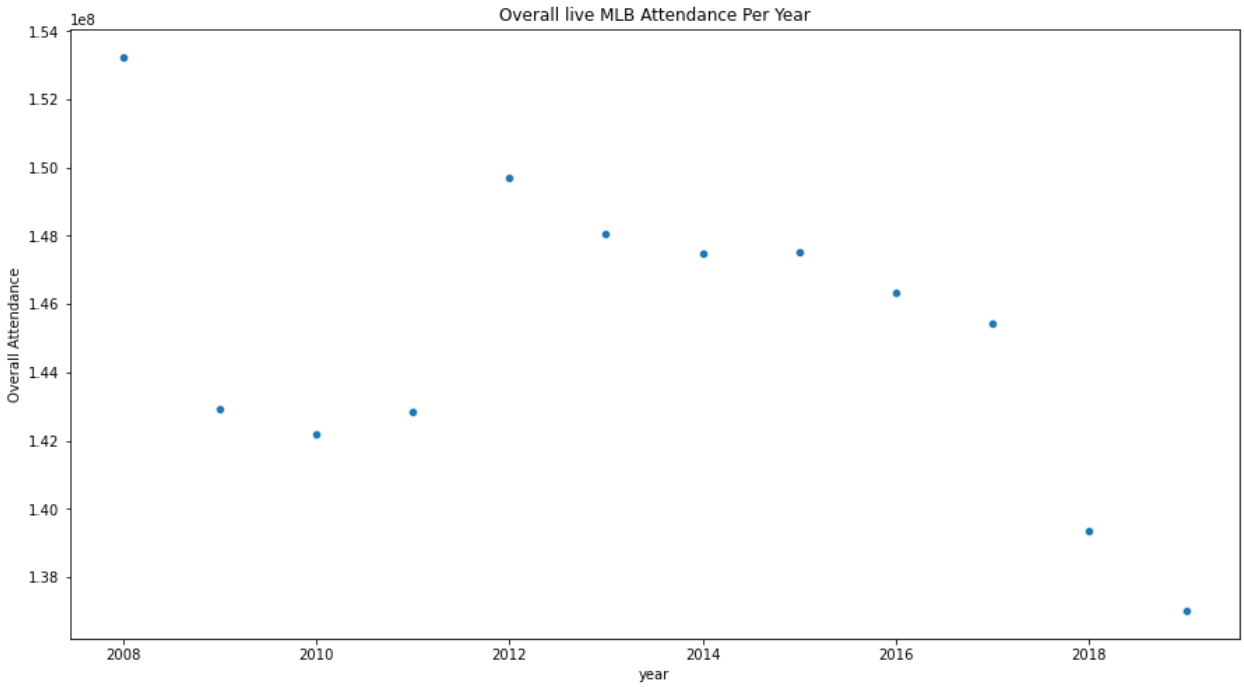
This data source provides data about the 30 MLB stadiums that includes:

- Stadium Capacity
- Distance to center field
- Whether the stadium is in open air or covered
- The turf type

3 Exploratory Data Analysis

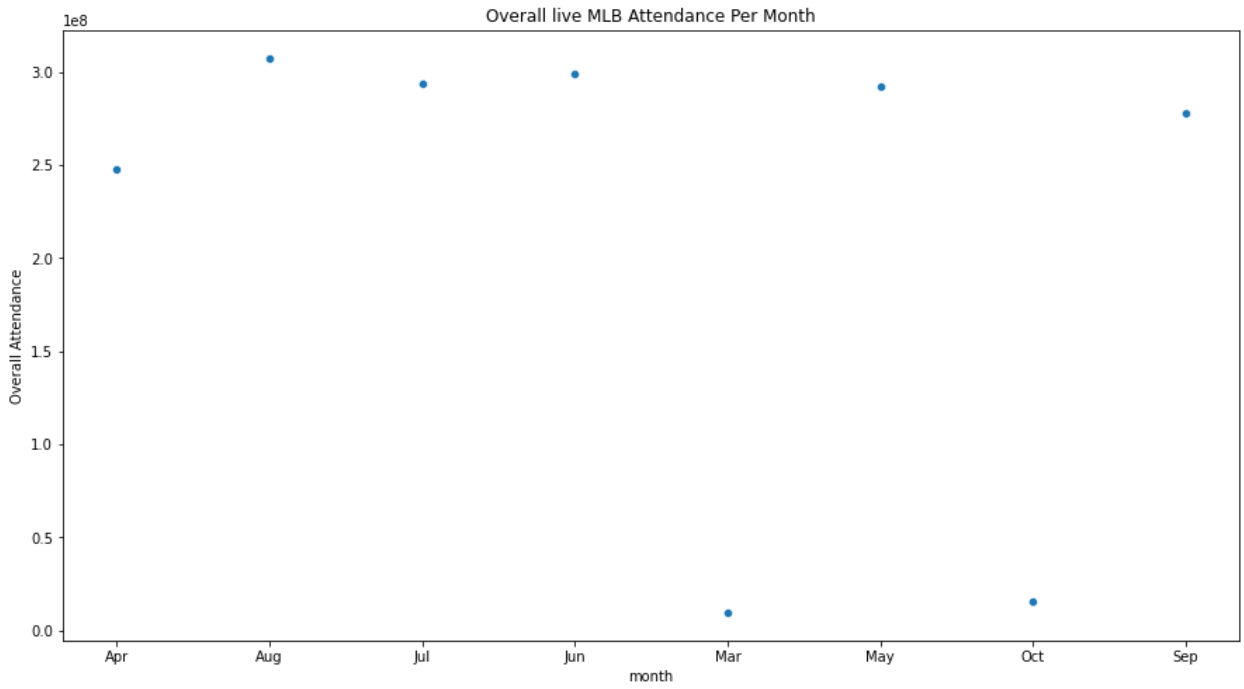
3.1 Yearly In Person MLB Attendance Trends

In person attendance has been on a decline since 2012 and leading into 2019

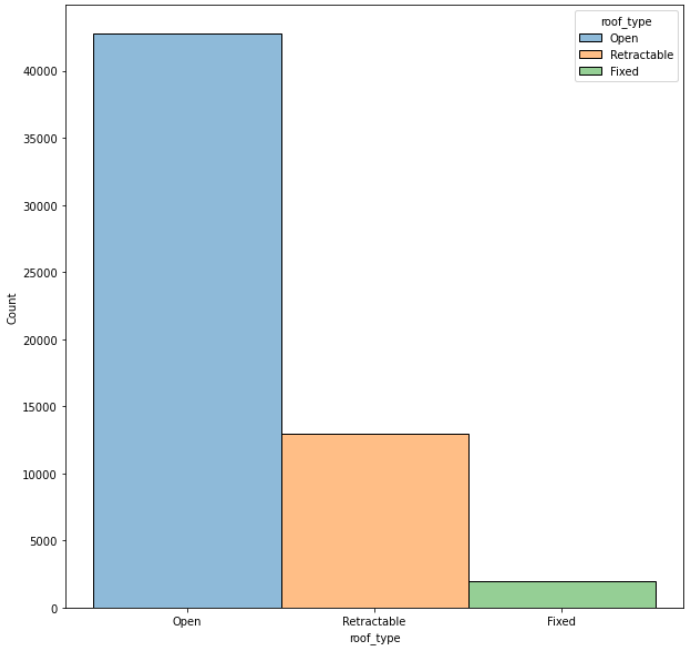


3.2 Monthly In Person MLB Attendance Trends

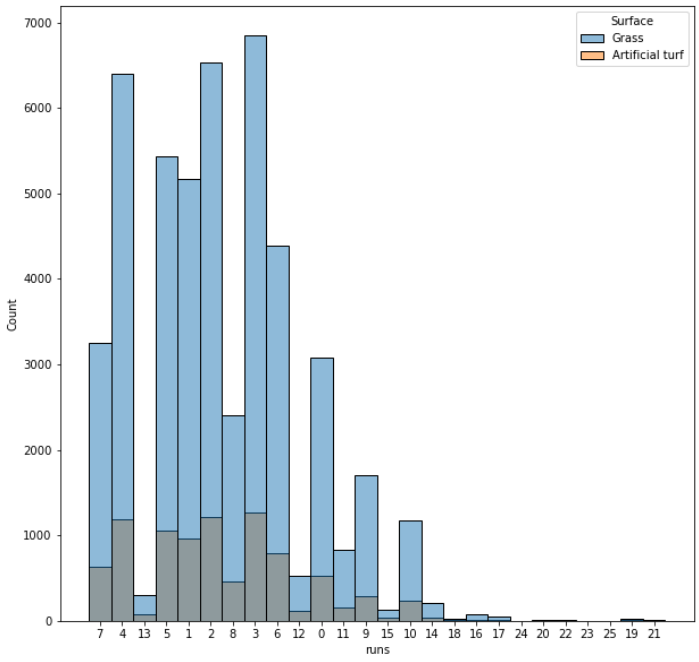
With the exception of March and October the attendance rates are consistent for in person attendance by month across the dataset from



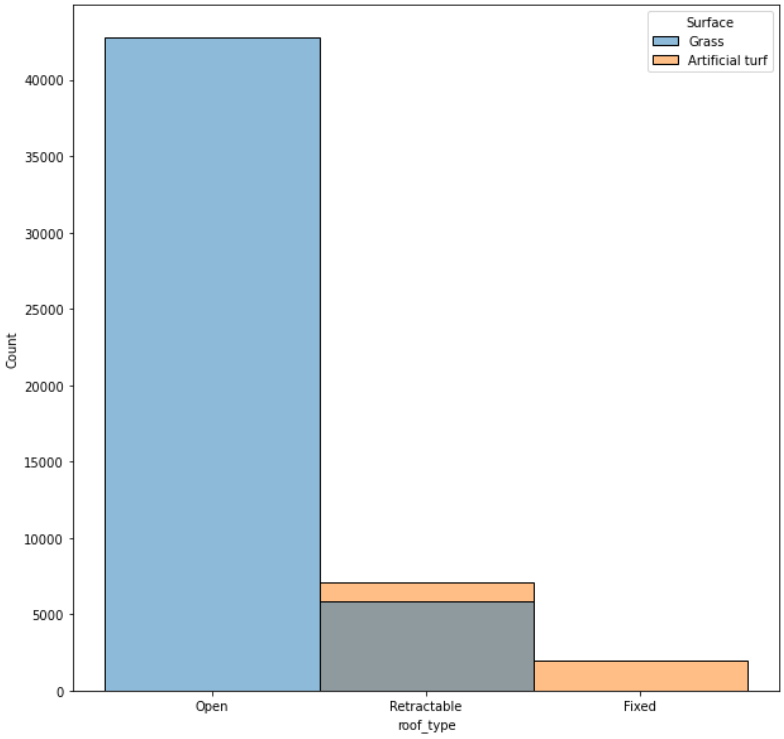
3.3 Distribution of Stadium Roof Types



3.4 Runs Vs Surface Type



3.5 Runs Vs Surface Type



4 Data Pre-Processing

In the process of defining model parameters, some variables were dropped to avoid data leakage or because they were redundant. The following table describes what was provided to the model.

4.1 Feature data prepared for model

The following table are the features that were provided to the model

Feature Name	Final State
Team Rank	Integer
cLI	Float
Streak	Int
Year	Int
Numeric Date	Int
Wins	Int
Losses	Int
Capacity	Int
Distance to Center Field	Float
Roof Type	One Hot Encoded Boolean
Stadium Type	One Hot Encoded Boolean
Surface	One Hot Encoded Boolean
Month	One Hot Encoded Boolean
Day Of Week	One Hot Encoded Boolean
Opponent	One Hot Encoded Boolean
Home Team	One Hot Encoded Boolean

4.2 Target feature prepared for model

The stadium attendance was provided to the model

Name	Final State
Attendance	Int

5 Approach Taken

Due to the nature of the data, we seek to predict the fan attendance as a numeric value for a particular game. This type of prediction requires the use of a regression model.

Three types of models will be built and tuned in order to see which one provides the best prediction capability.

5.1 Predictive Power Metric

Mean Absolute Percent error was used to measure the prediction capability of the model, the chart below shows ranges as guidance to measure the performance and they will be ranked against this.

MAPE	Forecasting power
<10%	Highly accurate forecasting
10%~20%	Good forecasting
20%~50%	Reasonable forecasting
>50%	Weak and inaccurate forecasting

Source: Lewis (1982)

5.2 Models Used

1. XGBoost Regressor
2. Random Forest Regressor
3. TBD

5.3 XGBoost Regressor

A 5 fold cross-validation randomized search approach was implemented with the XGBoost Regressor in order to improve the predictive power. The tuned parameters are listed below.

5.3.1 Model Parameters Tuned

```
params = {
        'min_child_weight': [1, 5, 10, 15, 18],
        'gamma': [0.5, 1, 1.5, 2, 5, 7],
        'subsample': [0.6, 0.8, 1.0],
        'colsample_bytree': [0.6, 0.8, 1.0],
        'max_depth': [3, 4, 5, 7, 9],
        'learning_rate': [0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5],
        'n_estimators' : [200, 500, 1000, 2000, 4500]
        }
```

5.3.2 XGBoost Model

```
xgb = xgb.XGBRegressor( n_estimators=50, random_state = 42)
XGB_random = RandomizedSearchCV(estimator = xgb, param_distributions = params, n_iter =
15, cv = 5, verbose=4, random_state=42, n_jobs = -1,
scoring =  neg_mean_absolute_percentage_error')
XGB_random.fit(X_train_scaled, y_train)
```

5.3.3 Best Estimator

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
            colsample_bynode=1, colsample_bytree=1.0, gamma=0.5, gpu_id=-1,
            importance_type='gain', interaction_constraints='',
            learning_rate=0.2, max_delta_step=0, max_depth=7,
            min_child_weight=15, missing=nan, monotone_constraints='()',
            n_estimators=4500, n_jobs=0, num_parallel_tree=1, random_state=42,
            reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=0.8,
            tree_method='exact', validate_parameters=1, verbosity=None)
```

5.3.4 XGBoost Model Result

Mean Absolute Percent Error	18% - Good Forecasting

5.4 Random Forest Regressor

A 5 fold cross-validation randomized search approach was implemented with the Random Forest Regressor in order to improve the predictive power. The tuned parameters are listed below.

5.4.1 Model Parameters Tuned

```
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
```

5.4.2 Random Forest Model

```
RandomizedSearchCV(cv=5,
                   estimator=RandomForestRegressor(n_estimators=50,
                                                   random_state=42)
```

5.4.3 Best Estimator

```
RandomForestRegressor(bootstrap=False, max_depth=60, min_samples_leaf=4,
                      min_samples_split=10, n_estimators=854, random_state=42)
```

5.4.4 Random Forest Model Result

Mean Absolute Percent Error	25.2% - Reasonable Forecasting

6 Selected Model

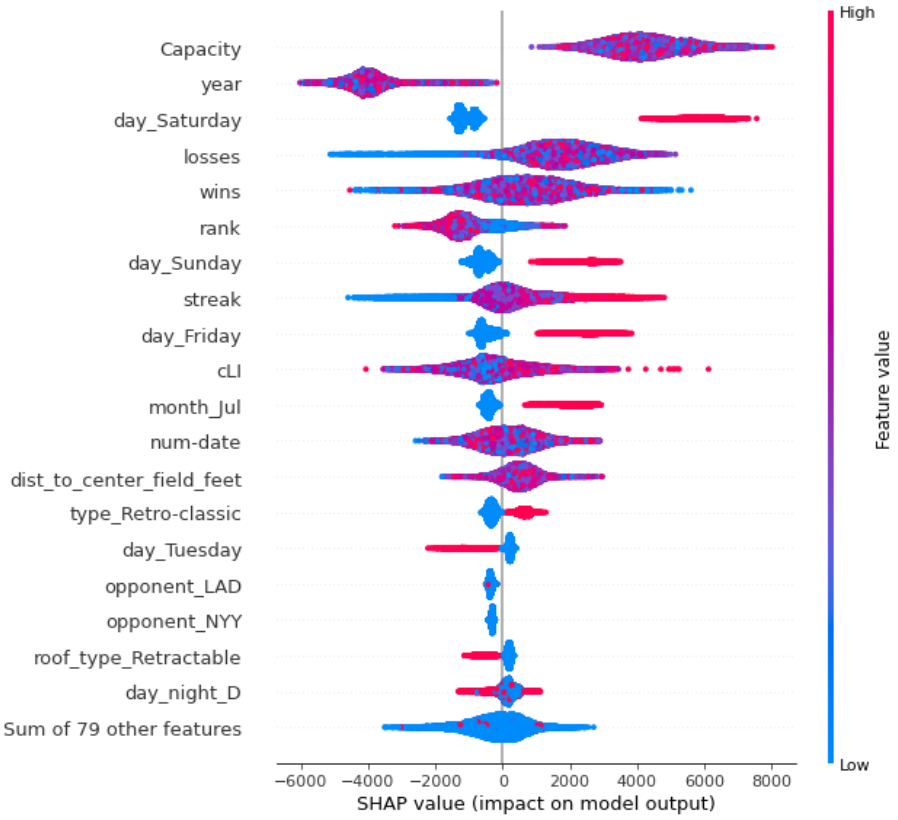
The model selected as the final candidate was the XGBoost model due to the great explainability and the prediction performance.

6.1 Summarized Model Results

MAPE	**Performance**
Random Forest Regressor	25.2% - Reasonable Forecasting
XGBoost Regressor	18.1% - Good Forecasting
OLS Regression	

6.2 XGBoost Model Output Review

6.2.1 Feature Importance



6.2.2 Model Insights

The following insights can be taken from the model output:

1. Games that fall on Friday, Saturday and Sunday have a consistent positive impact on the in person attendance rate.
2. When the team is on a winning streak, it positively impacts the in person attendance rate.
3. Stadiums with retractable roofs are not as desirable to the population of in person fans and negatively impacts attendance rate.
4. Fans are aware of the Championship Leverage Index (cLI) and how it represents the importance of a game towards the championships. It appears to be correlated to the in person attendance rates.
5. July Is an important month for In person attendance.
6. Fans respond positively to a classically styled stadium.
7. When a team is playing the LA Dodgers or the New York Yankees the in person attendance is negatively affected.