Predicting Home Attendance For The New York Mets

Using Regression & Time-Series Models

Yasir Karim

Overview

- Covid-19 has hurt the global sporting industry severely. After months of stoppage, top-level sporting leagues return to action without the fans.
- This has led to a significant loss of revenue for teams.
- *New York Mets*, for example, generated over a *$100 million* from matchday revenue in *2019* that they will lose out on.

Problems to solve

1. How accurately can we predict game by game attendance for the New York Mets home ground?

2. How much matchday revenue will the Mets lose for the 2020 season?

3. Which factors are most significant for ticket sales?

4. Can we identify time periods such as months or days of the week that are more significant to attendance?

The Data

1620 Games (10 Seasons)

1 target

Date

Opponent

Streak

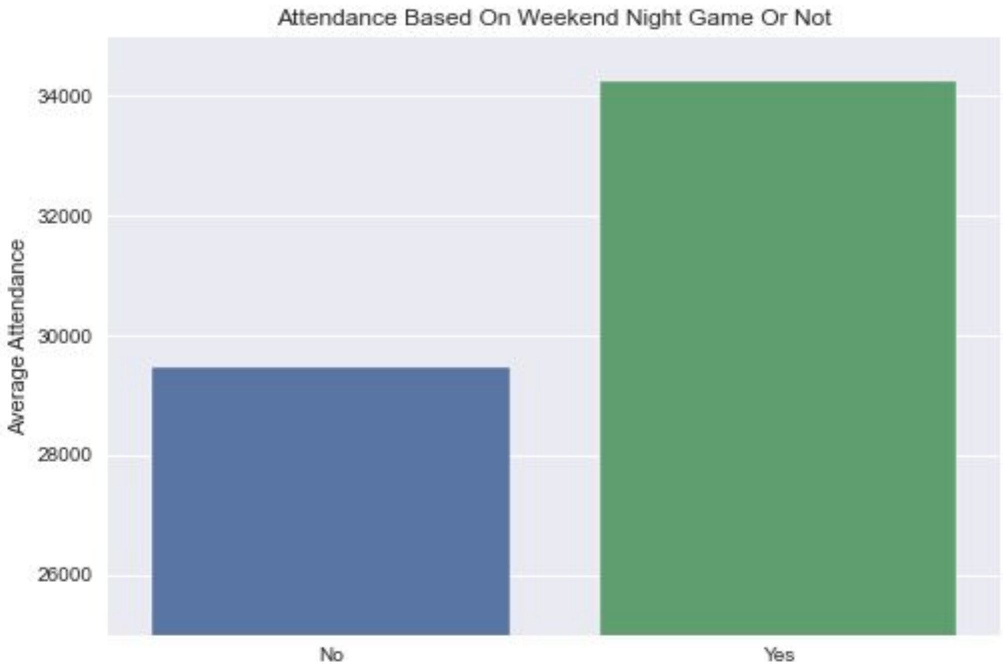
Game by game attendance

Train (9 seasons) - Test (1 season) Split

Data Cleaning

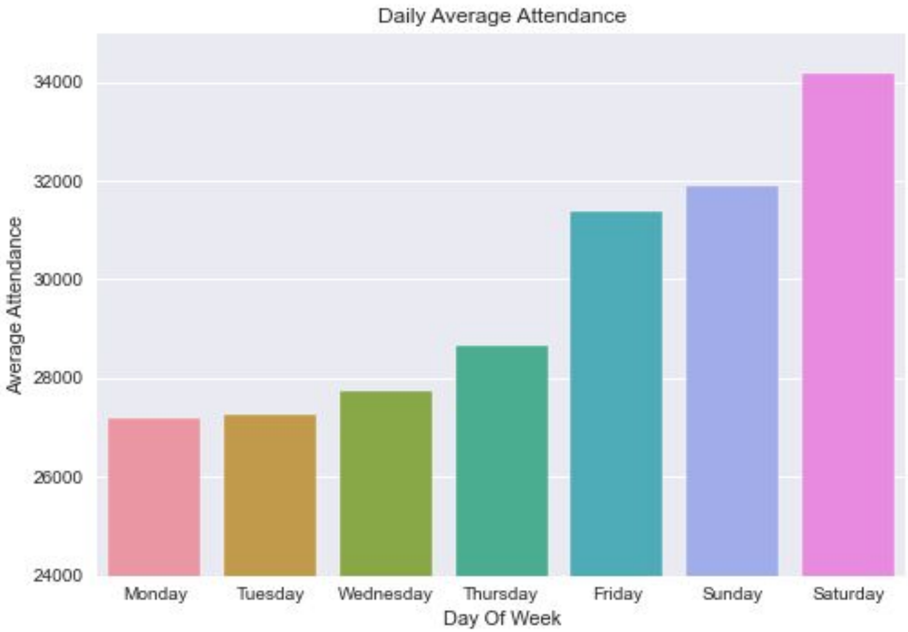
- We had to remove any games that were not played in Citi Field, which is the home ground of the Mets.
- We had to add missing year values on the date column to indicate which season the game was played.
- On days where multiple games were played, first of those entries had missing attendance value. We had to impute these values from the second game that day.
- We converted the columns *streak*, *games_behind* & *d/n* into numeric types.

EDA



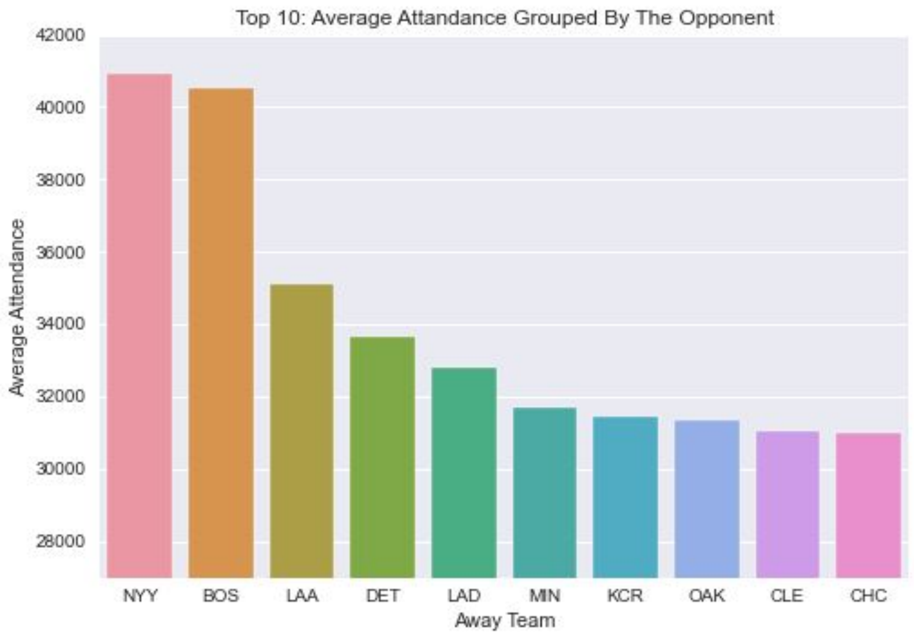
- Night games that were played during the weekend had higher attendance than Non-night/Non-weekend games

EDA (continued)



- Games played during friday till the end of the week have higher attendance as opposed to the other days.

EDA (continued)



- Some opponents are more popular than others.
- Such as the New York Yankees, who are city rivals, and Boston Red Sox.
- They draw a lot bigger crowd than the other opponents.

Modelling Process (Linear Regression)

- We **normalized** our train set before splitting it into two parts having a 9:1 train to holdout ratio.
- We used RMSE as our evaluation metric.
- Then, we used and iterative to run through different models in order to get the best scores.
- The K-Best Linear Regression model had the best RMSE scores, where K = 25.

	RMSE
k-Best	4691.09
Poly-K-Best	4719.73
Linear Regression	4763.06
Dummy Regressor	5987.98
Poly-Lasso	42548.70

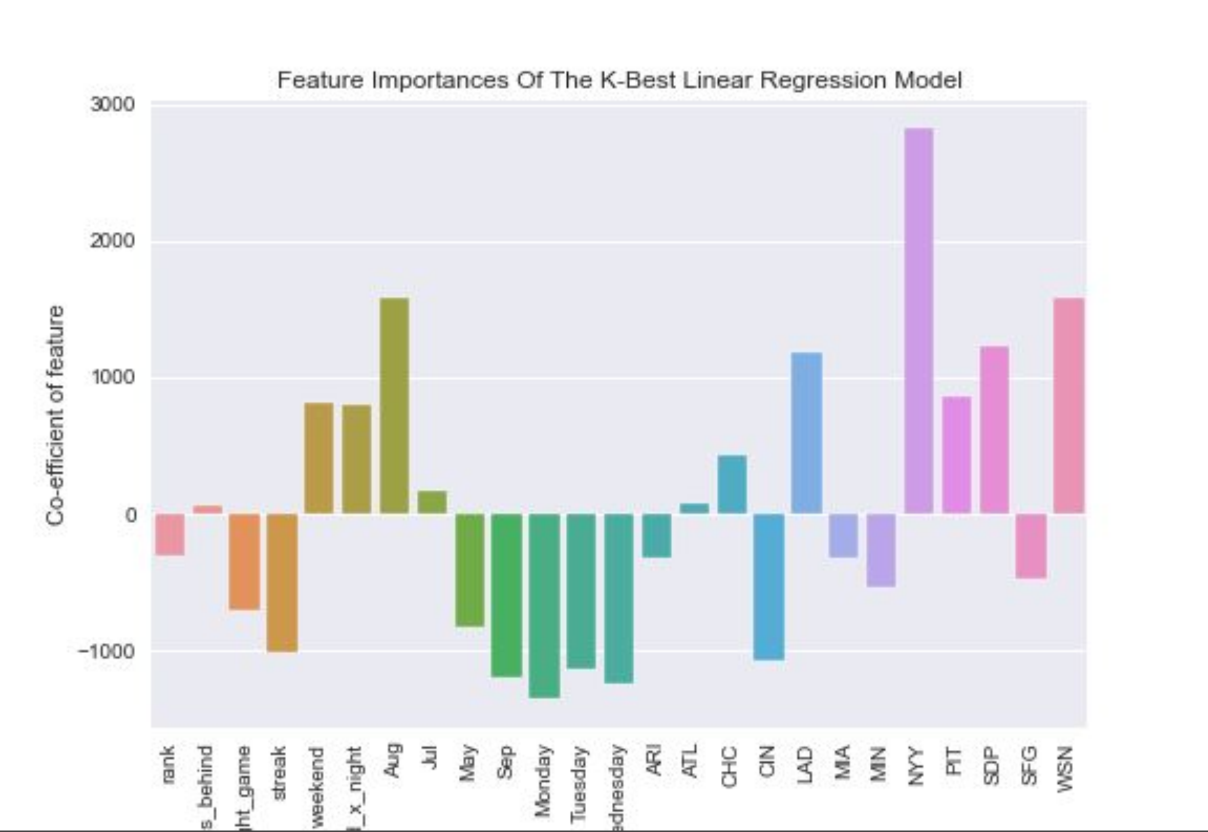
Modelling Process (Time-Series)

- After settling on our best regression model, we moved on to fitting time series models to the cleaned data.
- We started off with a simple ARMA model as our baseline and kept increasing the complexity of the models.
- Our evaluation metrics were not as good as those from the regression models.
- Our previous best model K-best, achieved mean absolute error of 2119 on the holdout data.

Model	RMSE
Baseline (ARMA)	6575.41
ARIMAX	5389.56
SARIMAX #1	5827.97
SARIMAX #2	5326.13
SARIMAX #5	**5006.82**

Best Model	Holdout RMSE	Holdout MAE
K-Best	2947	2119

Feature Importances of Linear Regression



Recommendations

Proposal 1	• Increase prices for weekend-night games and games against popular opponents.
Proposal 2	• Offer reduced prices for weekday games and less popular opponents.
Proposal 3	• Using the data from the 2020 season to calculate lost revenue from the predicted attendance.
Proposal 4	• Improve on-field performances as negative streak & games behind have adverse effect on attendance.

Future Work:

- Implement a recurrent neural network model to our data.
- Introduce more features for our data such as the weather of that day and in-game stats such number of injured players.
- Incorporate the impact of different categories of tickets sold such as premium and non-premium tickets and look at how that impacts revenue.

Sources

- Data source

  https://www.baseball-reference.com/teams/NYM/2020.shtml

- Revenue Information

  https://www.forbes.com/teams/new-york-mets/?sh=6b494cda3215

-

Thank you.