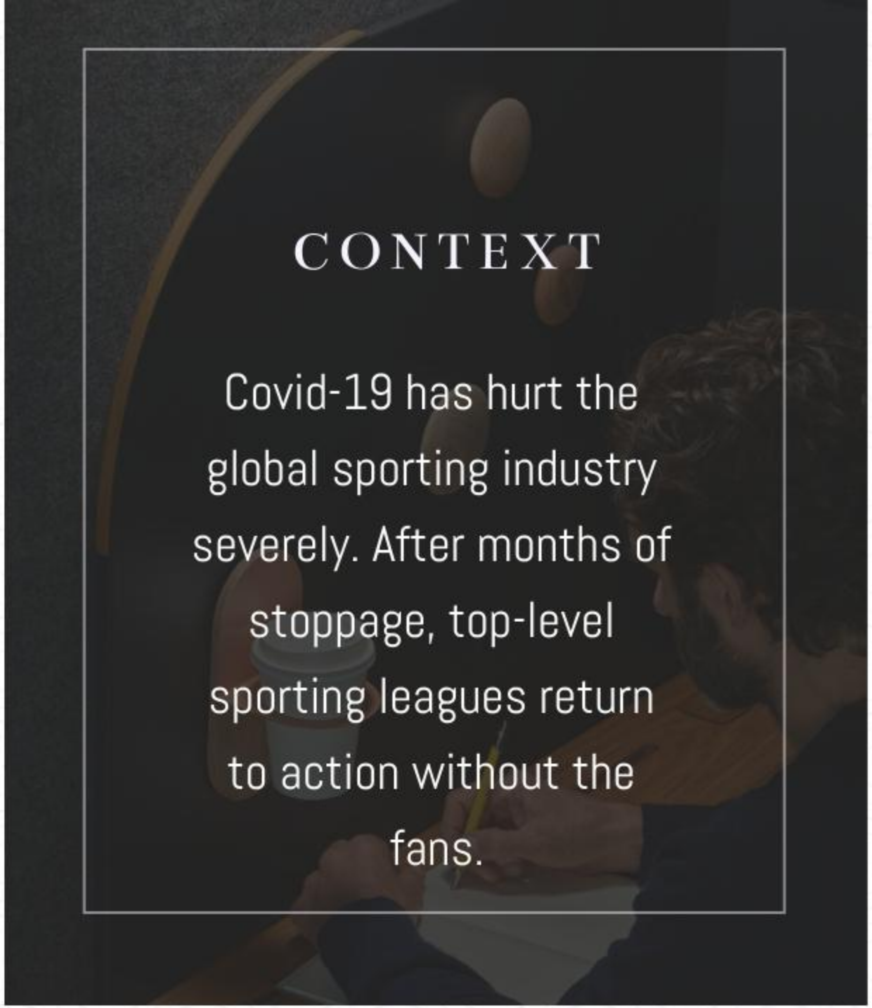
Predicting Home Attendance For The New York Mets

Yasir Karim



CONTEXT

Covid-19 has hurt the global sporting industry severely. After months of stoppage, top-level sporting leagues return to action without the fans.

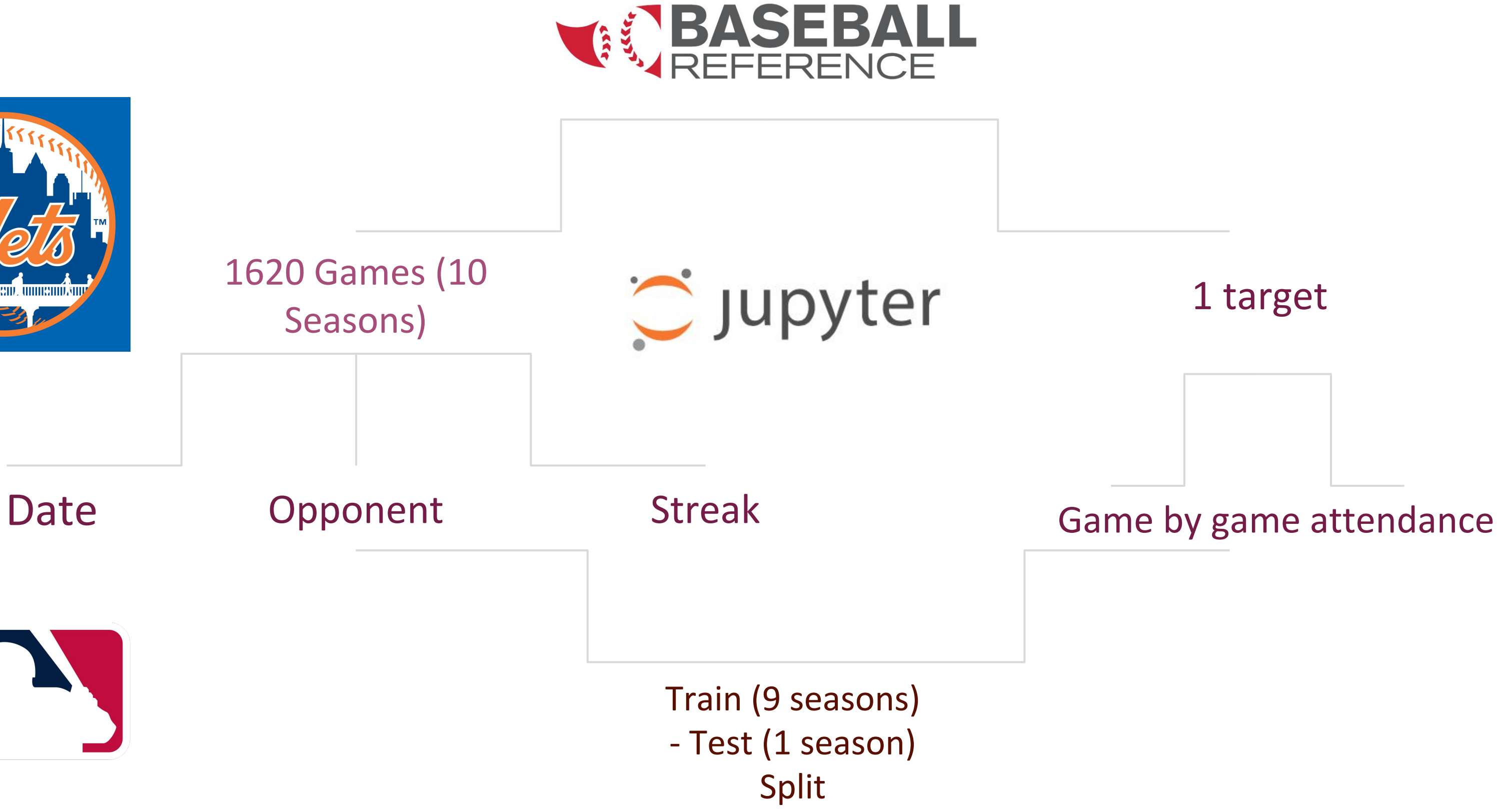
This has led to a significant loss of revenue for teams.

New York Mets, for example, generated over a $100 million from matchday revenue in 2019 that they will lose out on.

AGENDA

- How accurately can we predict game by game attendance for the New York Mets home ground?
- How much matchday revenue will the Mets lose for the 2020 season?
- Which factors are most significant for ticket sales?
- 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)

Date

Opponent

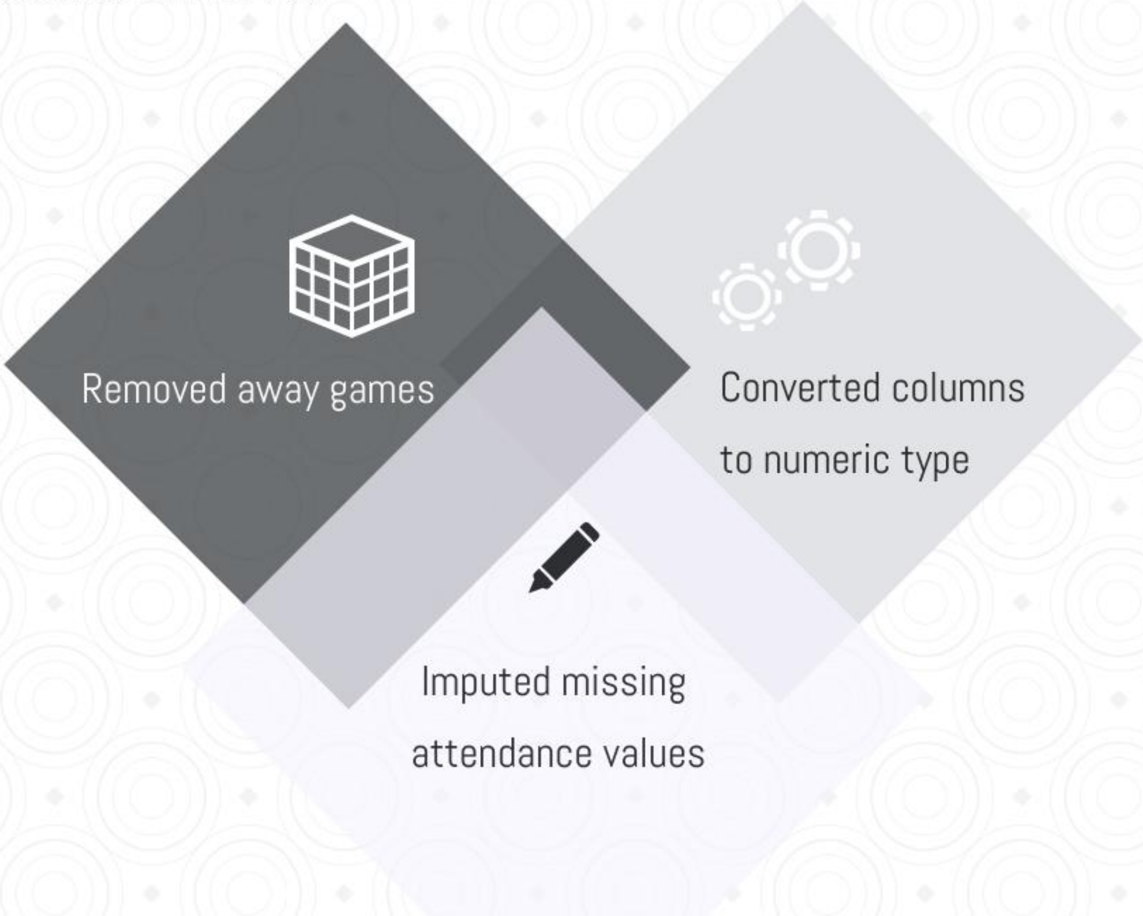
Streak

1 target

Game by game attendance

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

DATA CLEANING

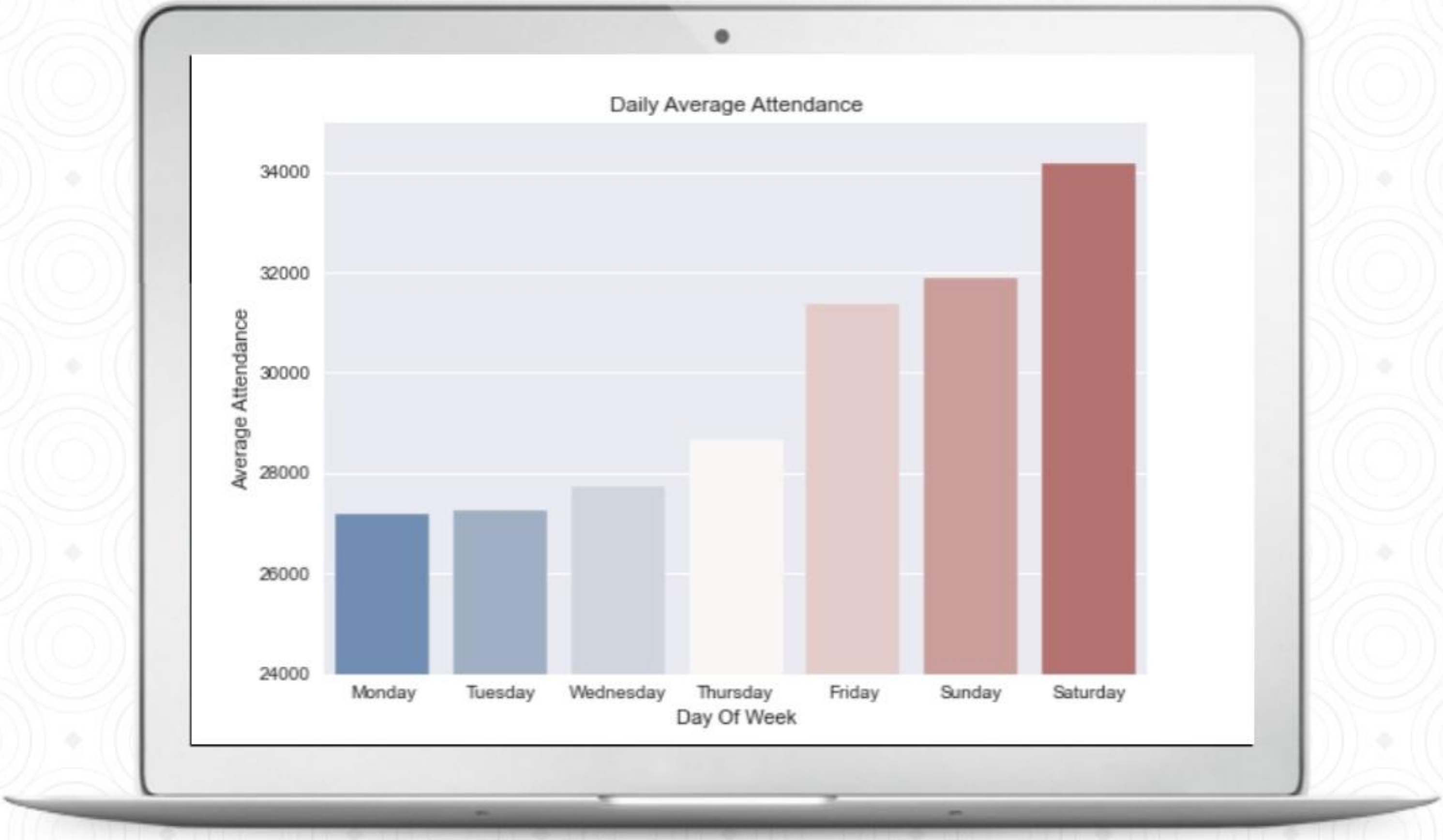


Removed away games

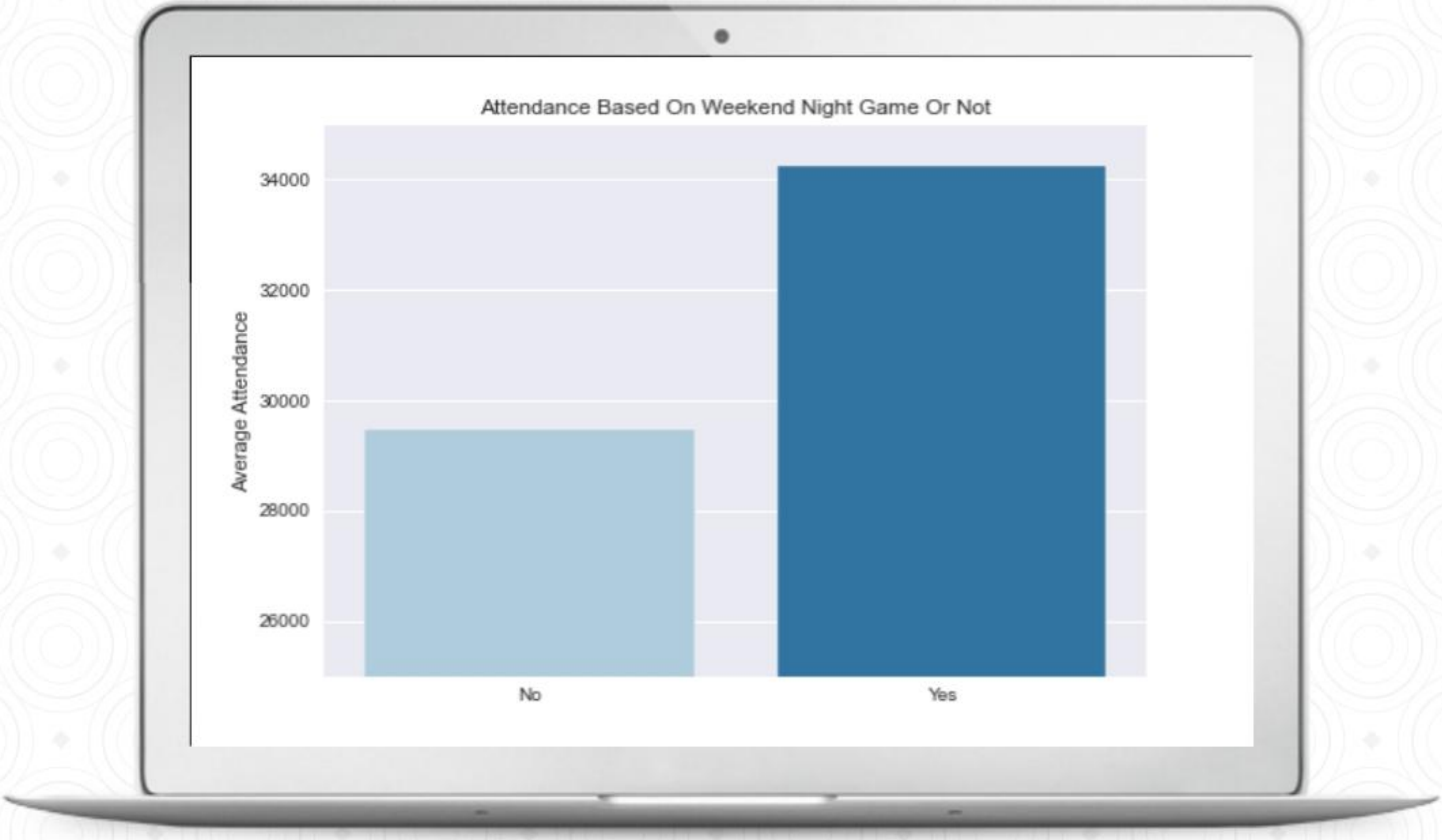
Converted columns to numeric type

Imputed missing attendance values

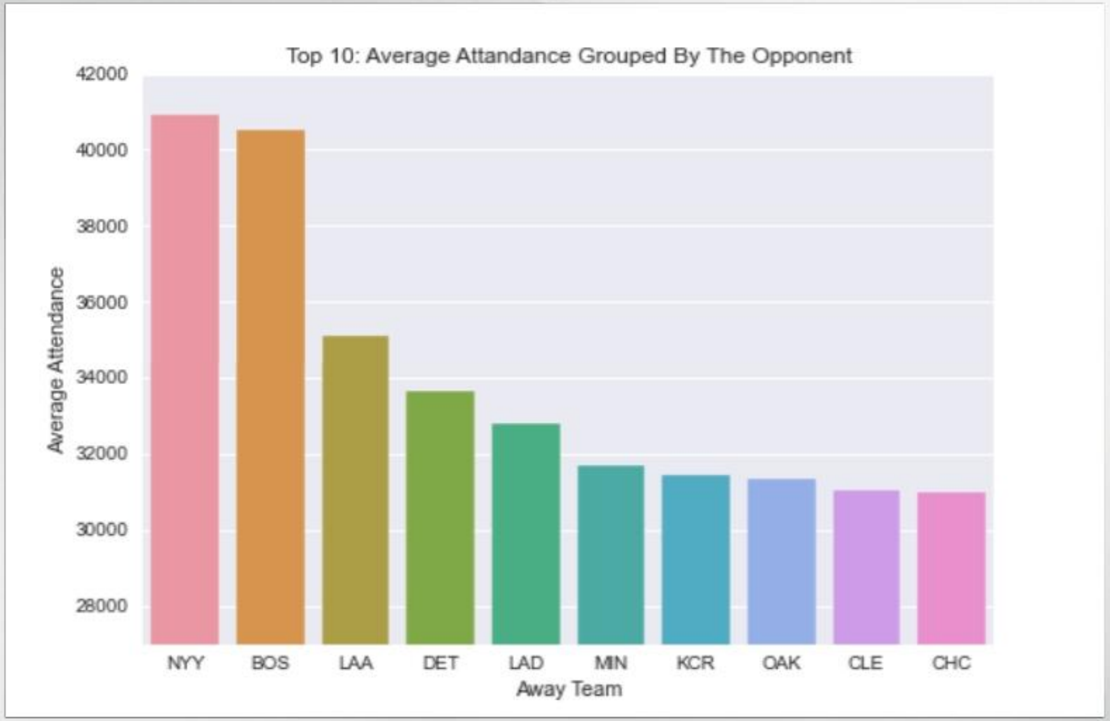
EDA


Daily Average Attendance
Average Attendance
34000
32000
30000
28000
26000
24000
Monday
Tuesday
Wednesday
Thursday
Friday
Sunday
Saturday
Day Of Week

EDA (continued)



EDA (continued)



LINEAR REGRESSION MODELLING

Model Fitting

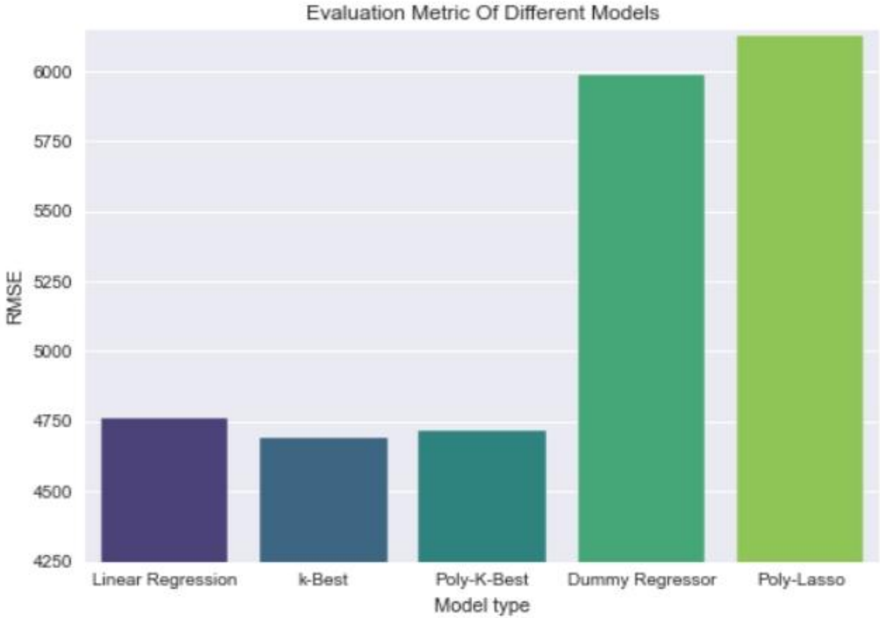
We iterated through a set of regression models in order to determine the best one

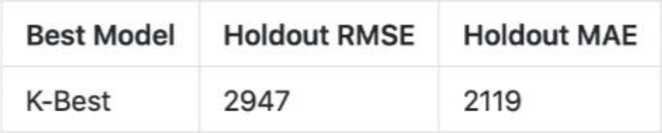
Evaluation

We evaluated our models using RMSE scores since they penalize high errors more

Feature Selection

We selected our best features using K-Best and Lasso filtering in order to simplify the models.


Evaluation Metric Of Different Models



Best Model	Holdout RMSE	Holdout MAE
K-Best	2947	2119

TIME SERIES MODELLING

Baseline Model (ARIMA)

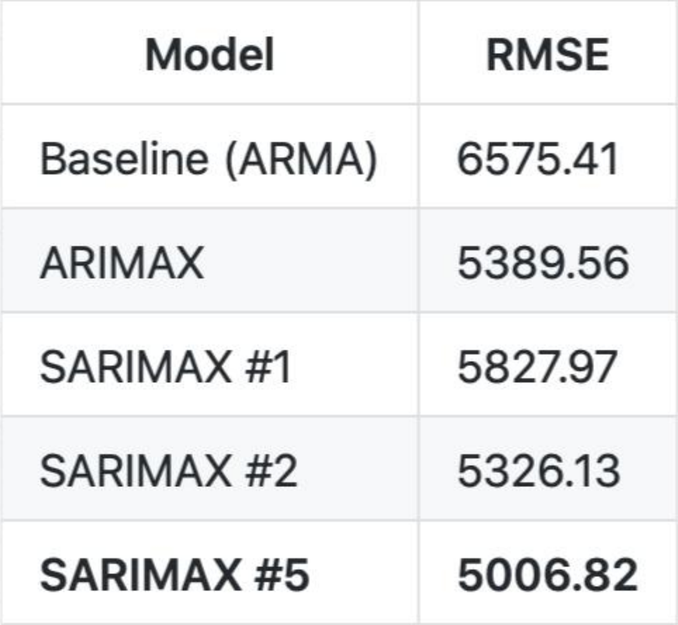
A baseline model that was basically predicting the mean

ARIMAX

Added the K-Best features from the regression model as exogenous variables

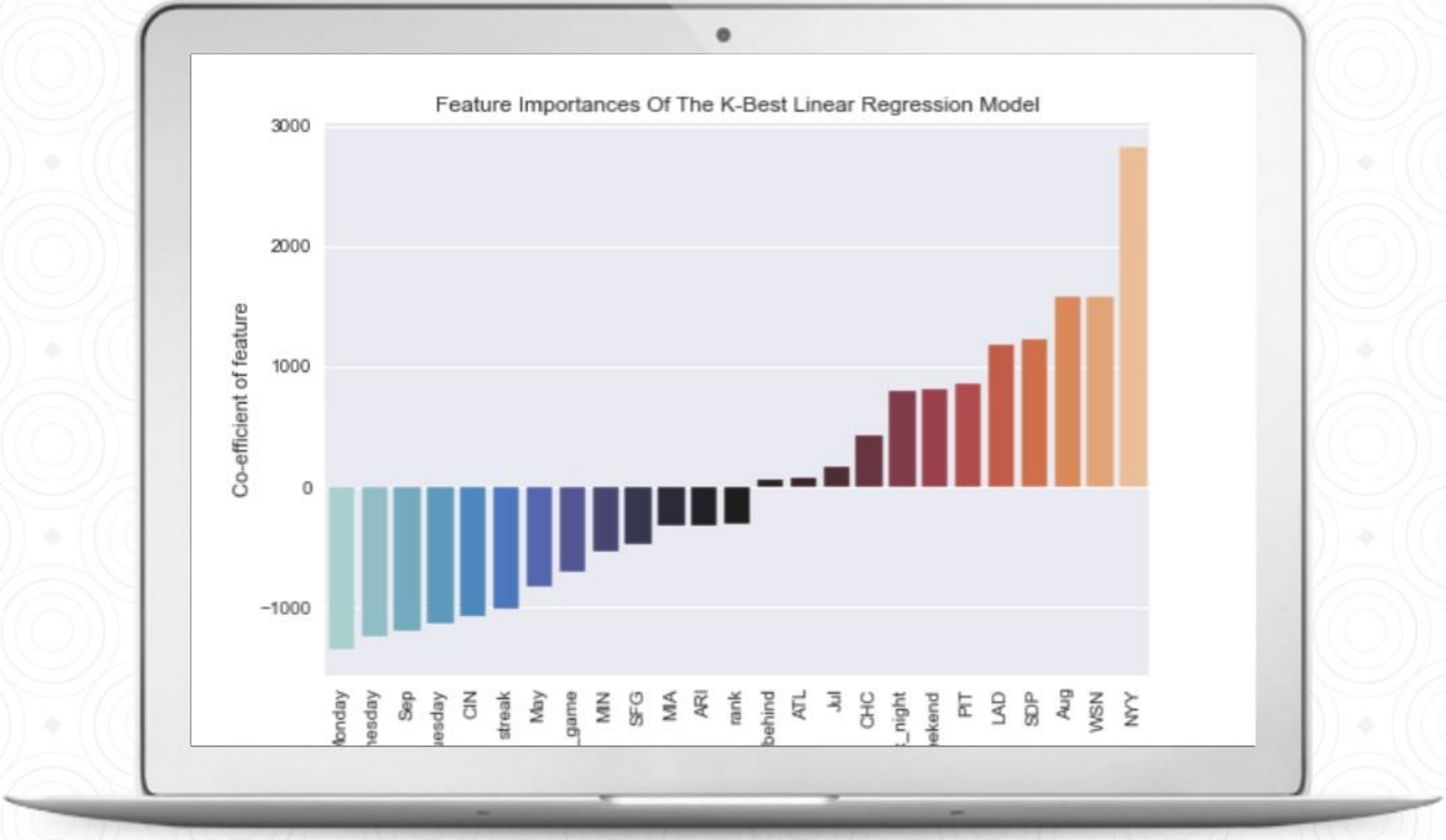
SARIMAX

Added seasonality to the model since we know baseball games are played in seasons with 81 home games



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

FEATURE IMPORTANCE



CONCLUSIONS

ADJUST PRICE

Increase/reduce ticket prices based on date and of the popularity opponent

CALCULATE LOSS

Use the stats and features from the 2020 season to calculate revenue loss.

IMPROVE PERFORMANCE

Improve on-field performances as negative streak & games behind have adverse effect on attendance.

NEXT STEPS

- 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.

THANK YOU FOR LISTENING