Crowd Tracker:

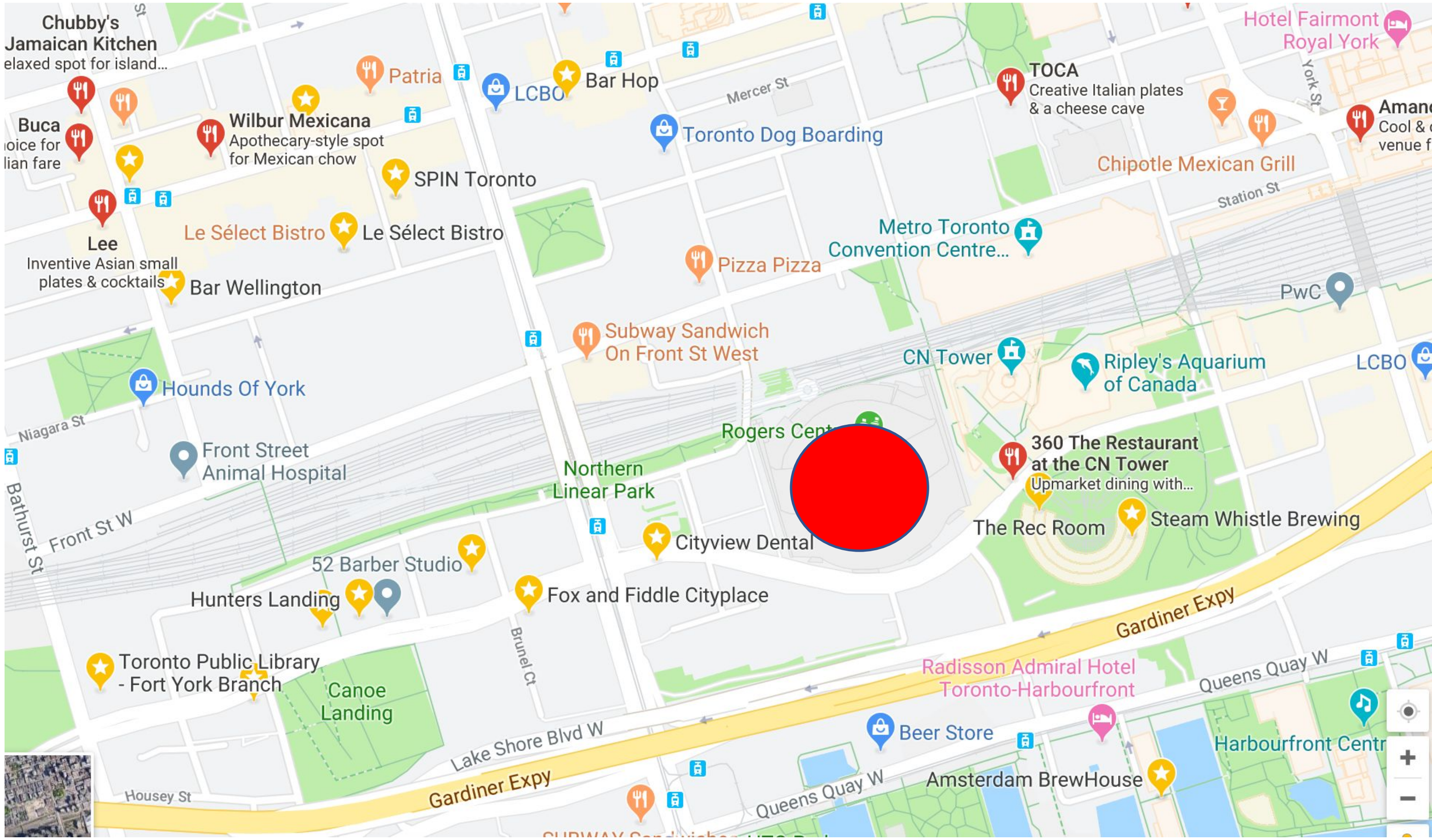
Blue Jays Attendance Predictor

David Maillet, PhD

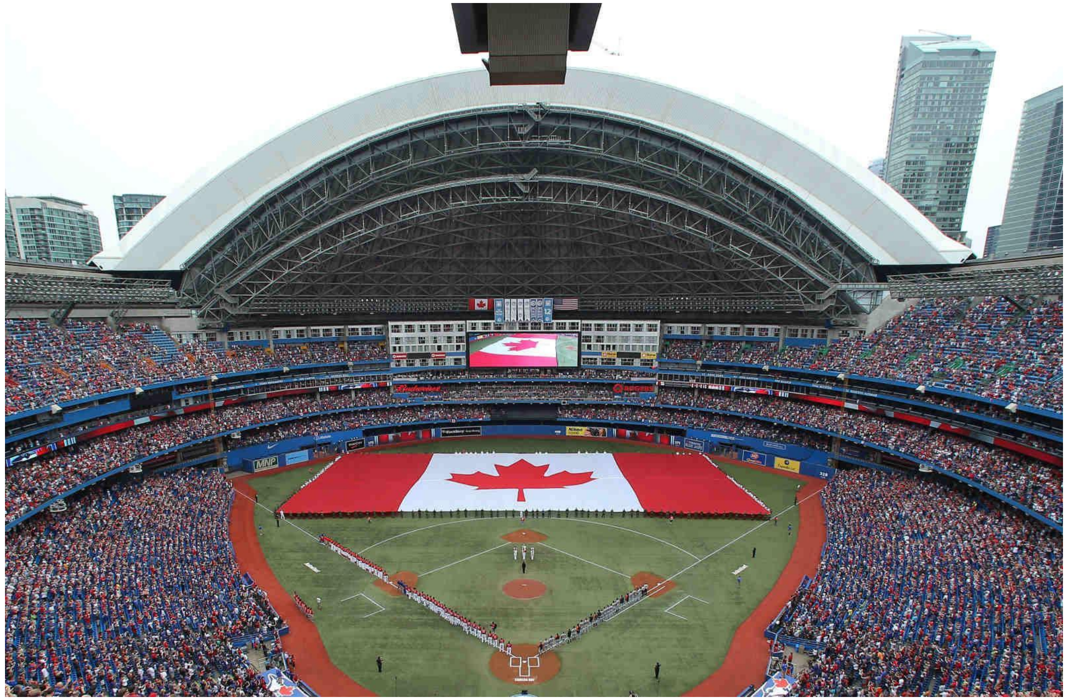
The Toronto Blue Jays Play at the Rogers Centre

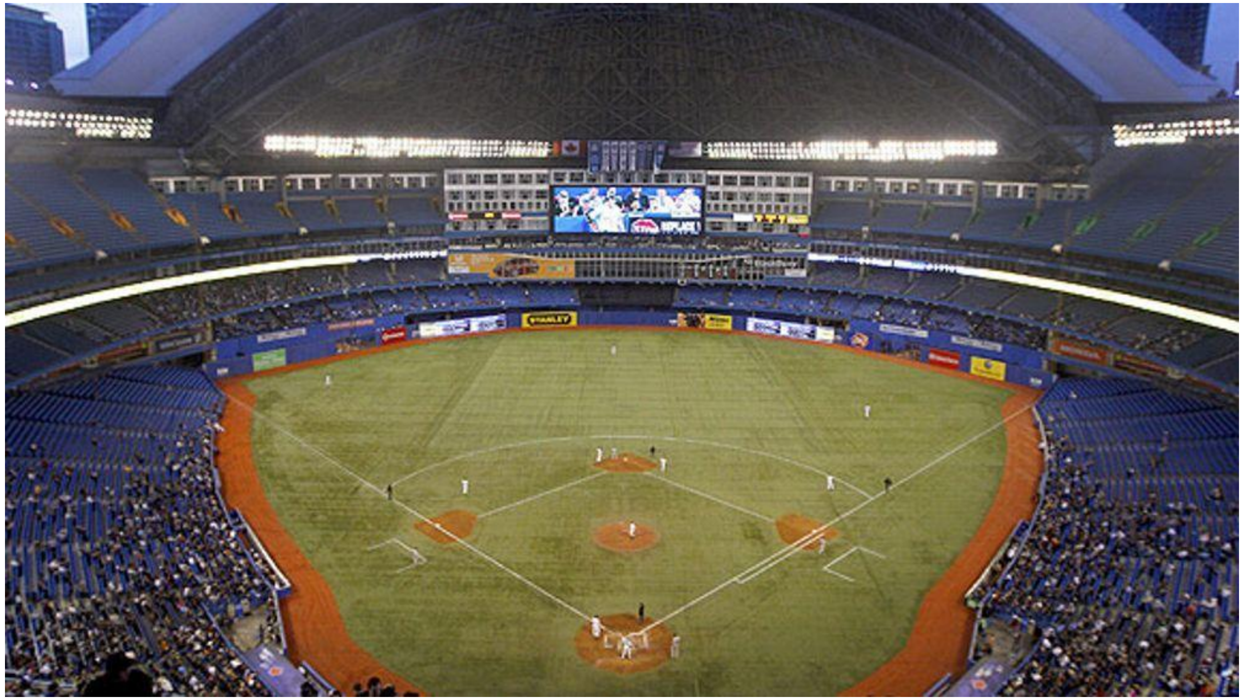
Businesses around the Rogers Centre stand to benefit from increased traffic generated by the games

- Restaurants
- Ice cream shops
- Bars
- Pharmacies
- Convenience stores
- Tourist shops
- Hotels
- Etc.



Attendance can vary by up to 40,000 people - Thus there is uncertainty about how many people will show up to any single game





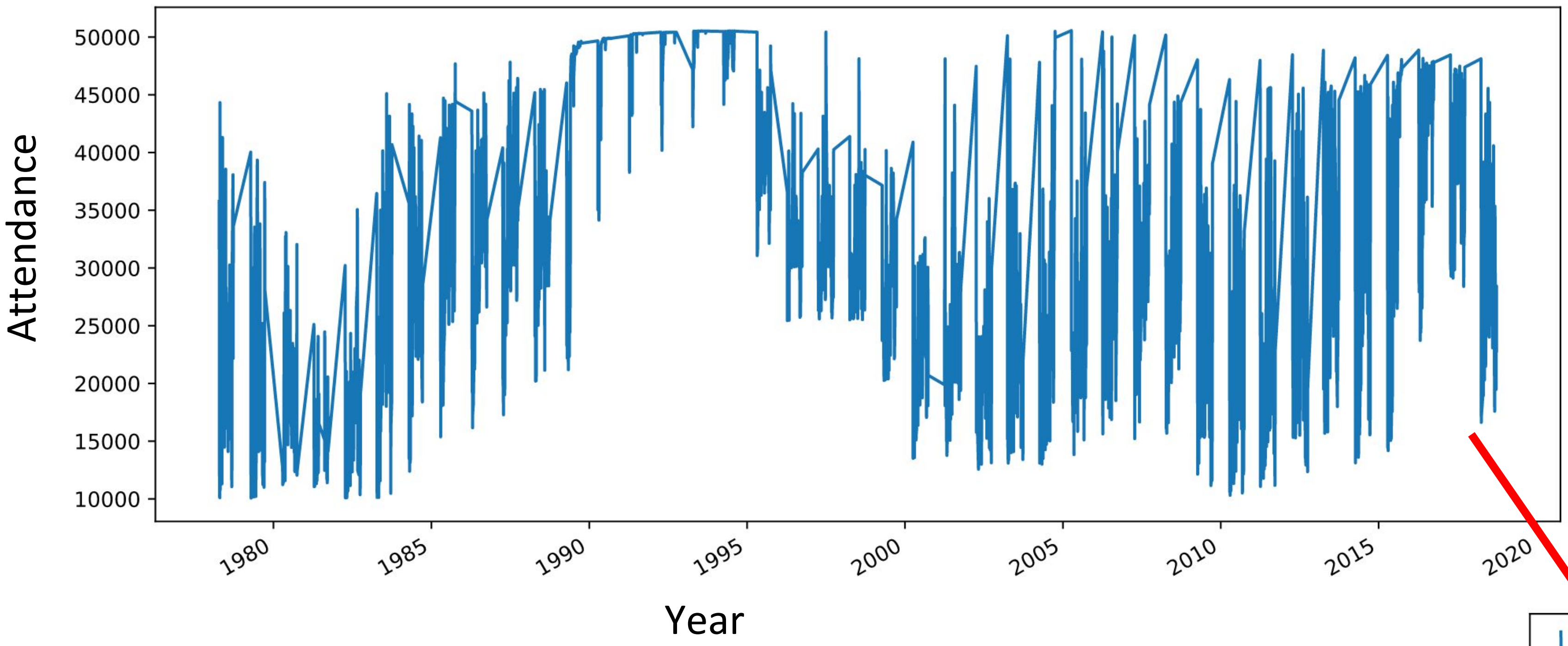
Goals

- Use historical data from 1978-2018 to develop a machine learning model that predicts how many people will go see each baseball game in 2019 so that businesses can better prepare staffing, stocking and pricing decisions on those days.

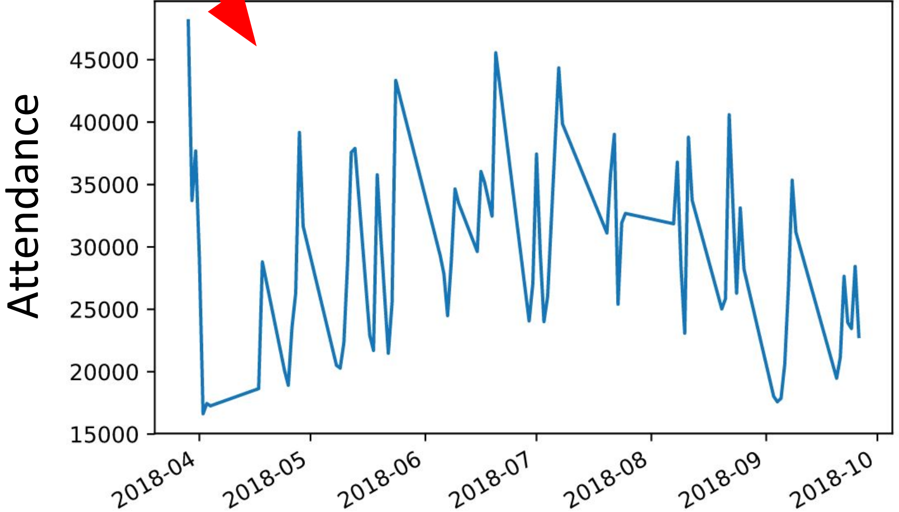
- Deploy model online as an interactive dashboard that anybody can consult

Target variable: Attendance (3207 data points)

Attendance from 1978 -2018



Attendance in 2018

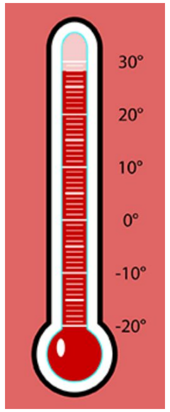


Total of 53 features included (all scraped from the internet)

Baseball-related features

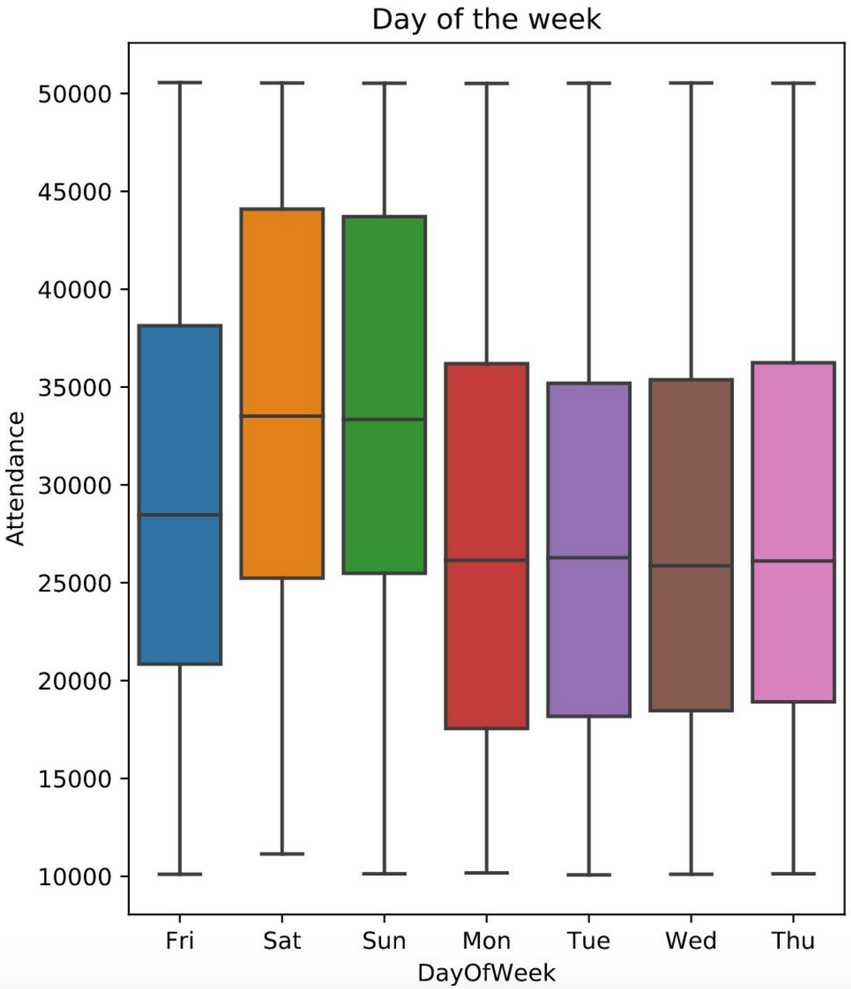
Competing events

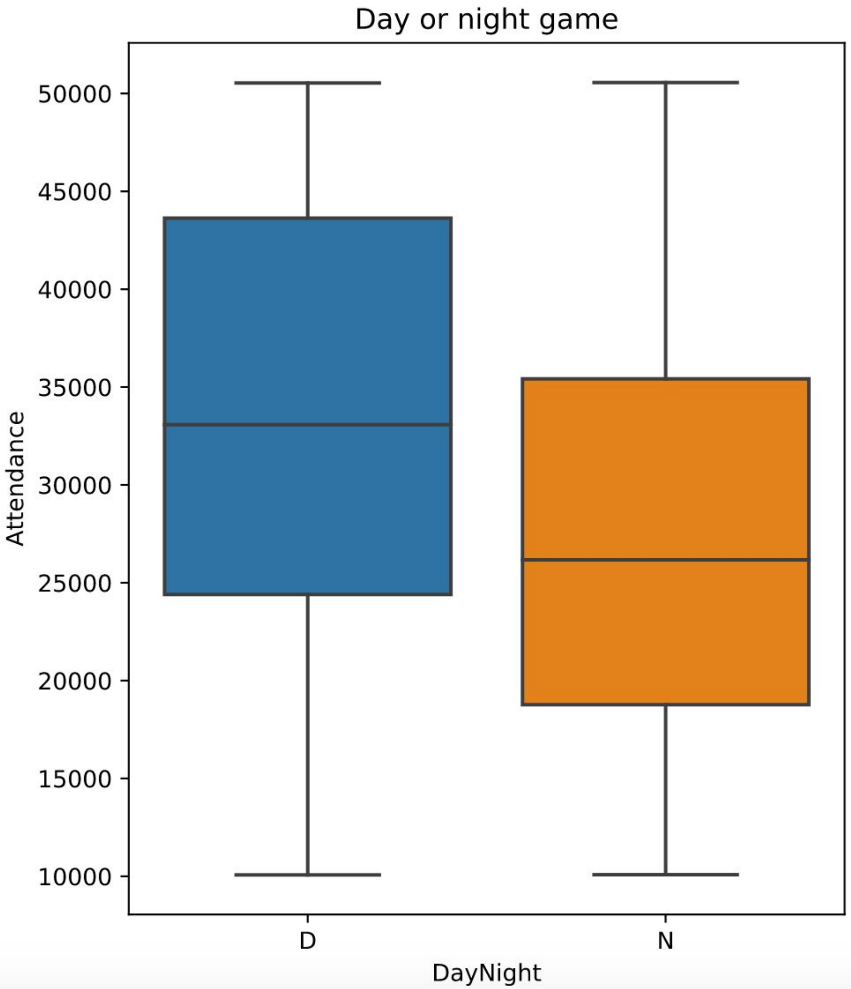
Weather

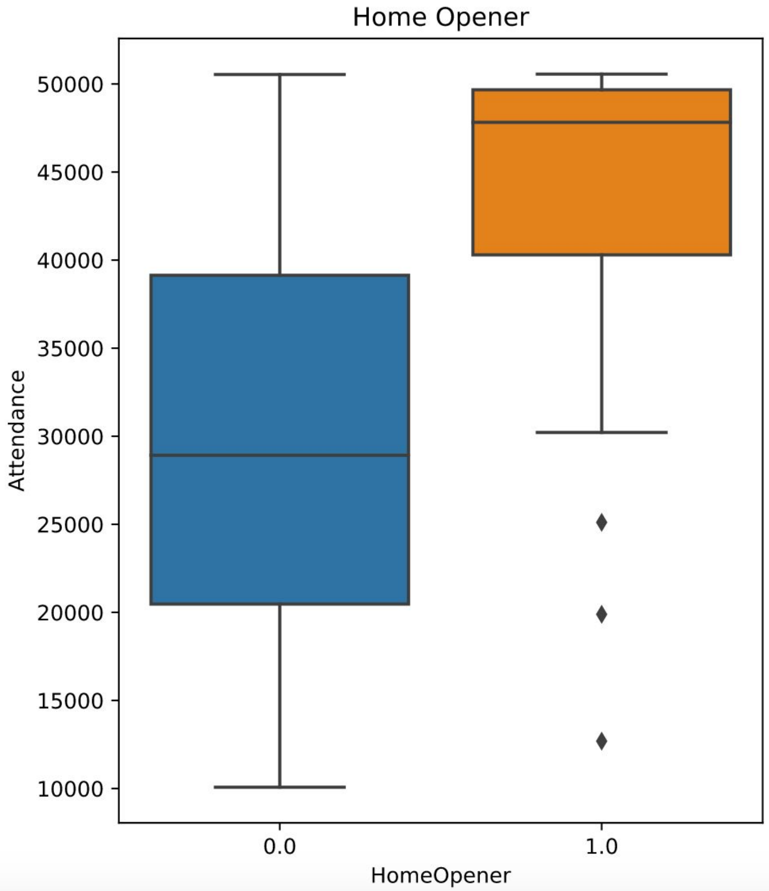


Time-related features

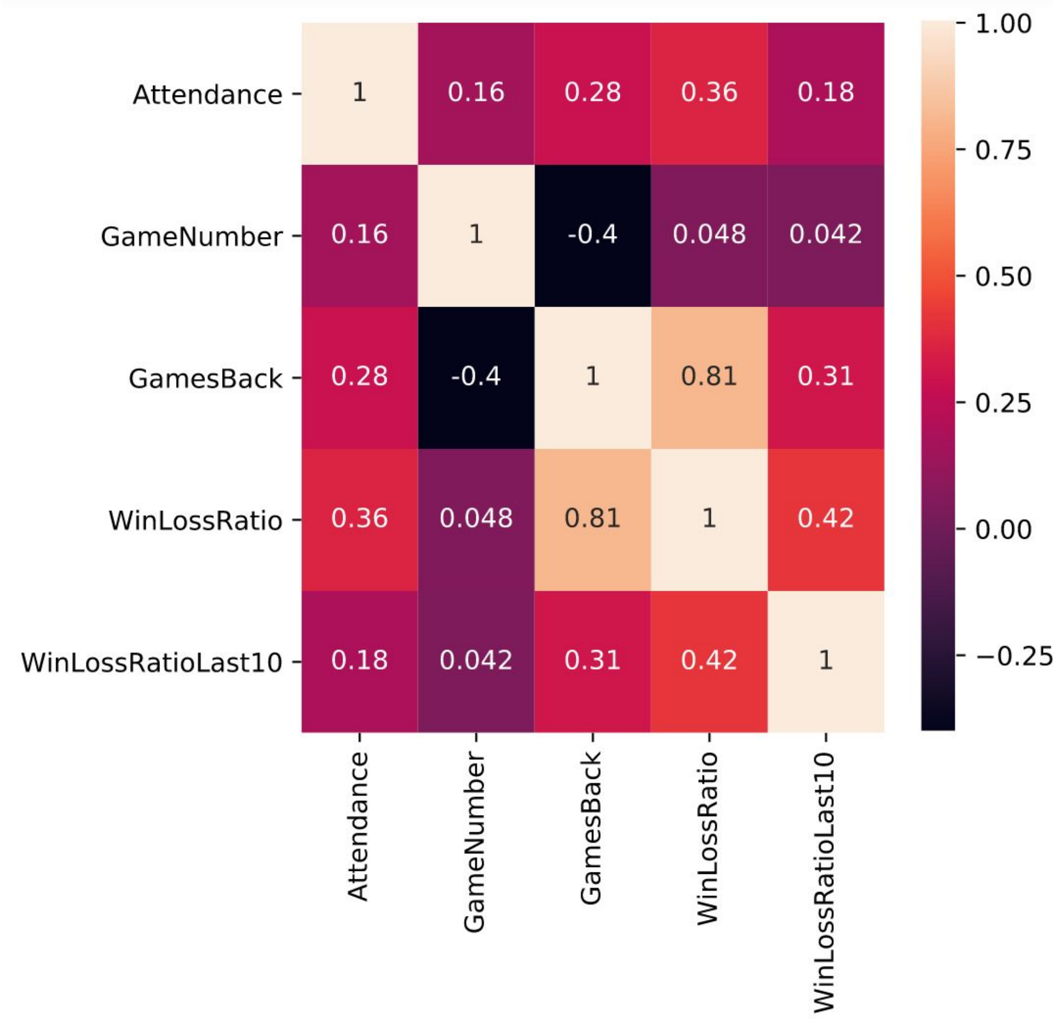
Exploratory analysis: Attendance is higher on weekends, for day games, and for the 1st game of the season


Day of the week


Day or night game


Home Opener

Exploratory analysis: Attendance is higher later in the season and when the team is doing well (when their win/loss ratio is high)



Machine learning

- Random forest model trained on data from 1978 to 2018

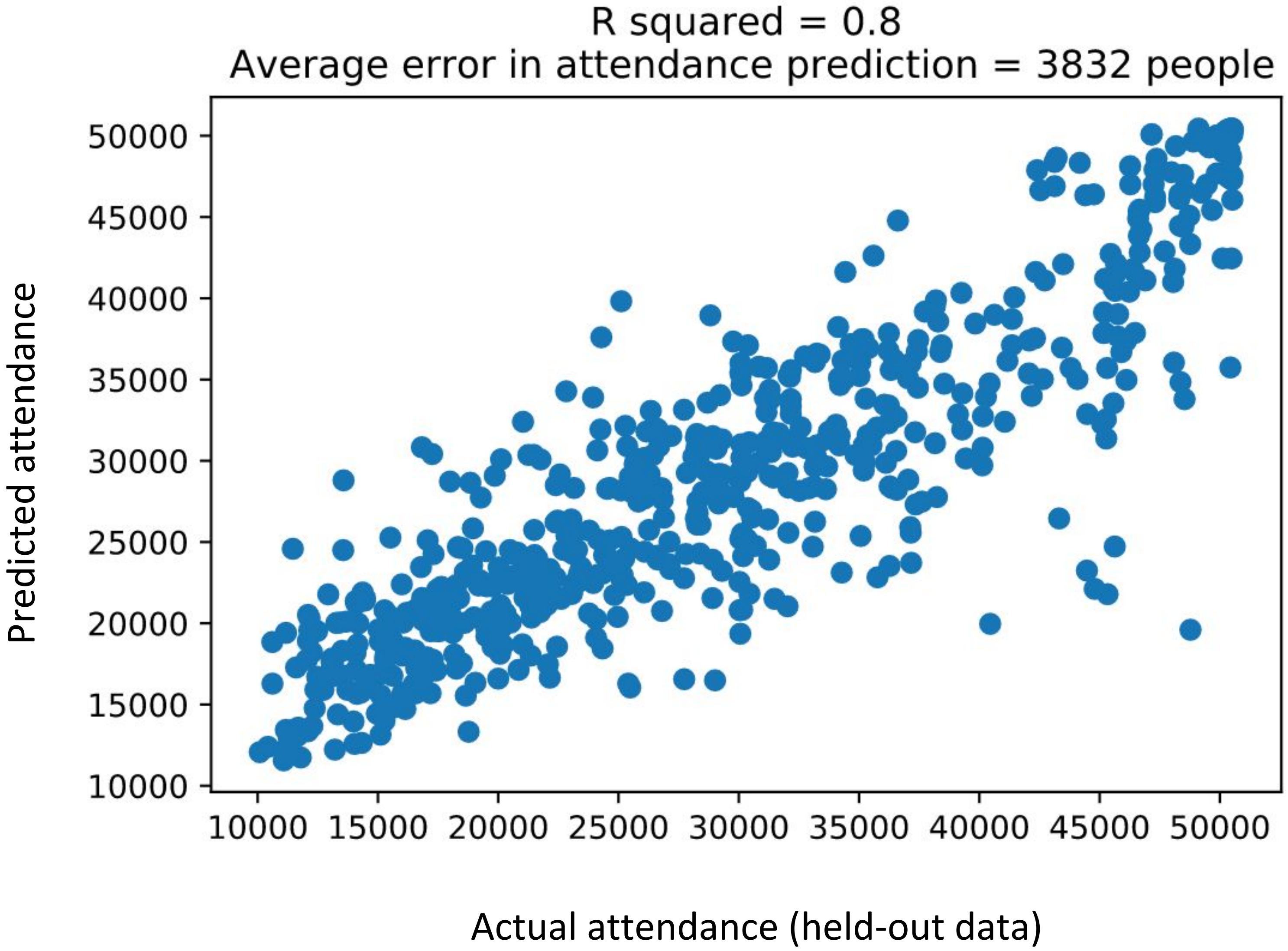
- Data split into training and test sets

- Cross validation used to find hyperparameters (tree depth, etc.)

- Model used to make predictions for games in 2019

- Feature importance was evaluated

Model performance



Best features

- Most important features have to do with how well the Blue Jays are doing
  - Attendance is higher when Blue Jays’ win/loss ratio is high
  - Attendance is higher when the Blue Jays are doing well relative to other teams in their division

- Time features are also important
  - More people go on weekends vs. weekdays

Website

- www.bluejaysattendance.com

- Developed with Dash and hosted on AWS

Future directions

- The scatterplot presented earlier indicates that the model frequently under-predicts but rarely over-predicts. It is likely missing features that explain high attendance for certain games
    - Promotion days (e.g., loonie dog day)
    - Player statistics (e.g. is star pitcher playing)

- Make different models for predictions far into the future
    - Some features, like blue jays win/loss ratio, become more uncertain as we go further into the future. Different models could be built to simulate different outcomes (best/worst case scenario)