Tel Aviv University
Raymond & Beverly Sackler Faculty of Exact Sciences
Blavatnik School of Computer Science
Data Science Workshop

Guided by:
Professor Daniel Deutch
Mr. Amit Somech

Predicting Major League Baseball Attendance

Submitted by:
Dana Rapoport, 305750663
Eidan Wasser, 204589972
Elkana Gamliel, 026594333
Ran Erez, 200663300

Introduction

Popular sporting events have a major impact on the economy and the society residing in the venue's vicinity. Recent papers have found that an increase in 1000 spectators results in 10.74 annual hours of traffic delay [1]. Furthermore, increased attendance results in substantial degradation to surrounding communication networks[2]. In this work, we chose to focus on predicting Major League Baseball attendance. Since MLB seasons contain a substantial number of games per team (81 home games over 6 months, excluding playoffs), we determined that the multitude of data points will help us achieve higher accuracy which in turn can result in better preparations for communication bandwidth allocation and traffic congestion reduction.

Our final results show a **29.2% improvement** on current industry benchmark utilizing Gradient Boosting Machine (GBM) model and enable predictions with a **RMSE of 4.73, in bins of 1K spectators**.

Dataset Description

Our initial dataset was extracted from retrosheet.org and contained MLB game logs for 1990-2017 containing 65,322 individual games. The data contained basic information such as home & away teams, stadium, opening lineup and in-game score and our response variable - Attendance.

Characteristics & Nature of Dataset -Our initial dataset contained 98 columns with 65,322 rows. The data included 483 games with missing attendance. Such games were part of a "double-header". Our target variable ranged from a minimum of 746 to a maximum of 80227. For more details regarding the nature of our dataset, see Step 4 of the data science process.

Problem Formulation - *"Baseball is like church. Many attend but few understand" - Wes Westurm, New York Giants.*

Following our preliminary research, we have decided to focus on predicting MLB attendance by bins of 1000. This will ensure municipalities can take measures to decrease traffic congestion on game-days and eventually save costs resulting in these delays. Each game will receive a prediction which will indicate "how many thousands of people will arrive to the game".

Project Point of Focus

Baseball game attendance is not just a question of ticket price. The decision whether to attend a baseball game is dependent on many different factors from disparate domains such as; quality of opposing teams, weather, cost, importance of the game (to reach the playoffs), presence of star players and historical team fixtures. Furthermore, information regarding the players is both on a per-game basis as well as cumulative metrics within a single season and even across seasons. We also wanted to address the importance of players without having to resort to encoding distinct players which would result in increased dimensionality (thousands of players over 28 seasons and 30 teams).

Therefore, our project's point of focus is **quantifying and integrating disparate domains into one dataset while controlling the increase in dimension**.

We accomplished our point of focus with 4 key activities:

- Detect which domains influence the game attendance and their interconnections (financial, climate and player's experience)
- Quantify the domains into numeric and categorical variables
- Normalize the data across seasons and within a single season (for both player and team metrics)
- Integrate player information without explicitly using individual players in the model.

We also chose to use a dedicated python package (H2O) which assisted us in controlling the high dimensionality and handles both categorical and numerical variables at once [4].

Our Data Science Process

Step 1 - Business & ML Problem framing, setting benchmarks

We began our work by clarifying our business problem, to detect traffic delays caused by MLB attendance. In order to answer our business problem, we framed our Machine Learning problem as predicting how many thousands of people will arrive to a given MLB game. We researched for related work done in the field (see "Related Work" section) and defined them as our benchmarks (RMSE 6.113, MAE 5.665, R^2 0.83).

Step 2 - Data Collection and Integration - Our initial analysis showed some level of correlation between our data's variables and attendance and we decided to augment our data from the following external sources which were scraped from baseball-reference.com:

- Weather Data - Each game's starting temperature, wind speed and precipitation. We collaborated with Jordan Bean, another researcher in the field (see "Related Work" section) and received his dataset containing average city temperature and stadium condition (closed or open roofed).
- Financial Data - Team payroll for each season, as well as individual player salaries for each season.
- Player Metrics – Each player's offensive and defensive metrics for each game played.

Step 3 - Data Preparation and Cleaning

Our data preparation focused on addressing the high seasonality in our dataset. Seasonality in our case exists on two levels: intra-season (some team metrics build up as the season progresses) and inter-season (adjusting for trends in gameplay and inflation – for financial metrics)

- Intra-season Seasonality –Metrics that describe team/player quality in sports are cumulative in nature and tend to be reset at the start of each new season. Thus, if we are aiming to characterize which teams are considered by fans "good"/"bad" on a given day, we must measure a team's metrics compared to metrics of all teams on that given day. All team/player quality metrics for each game were normalized against their distribution on that game's day in the following manner: For metric X on day d:

  $$(X - \bar{X}_{on\ d})s_{X\ on\ d}$$

  (subtract the mean of all X's on a day d and multiply by the standard deviation of X on day d).

  In Baseball in particular, team/player performance in any one game is highly erratic. Since these metrics are reset at the start of each season, these metrics are not meaningful for the first 10 games of the season. Thus, we have chosen to set these metrics to 0 for those first 10 games.
- Inter-season Seasonality – Financial metrics such as player salaries are highly susceptible to this type of seasonality. This is mainly influenced by inflation, but also by trends in gameplay – how different skills are valued in different points in time (offense vs. defense for instance). To address this, we chose to normalize each financial feature in a given season against the distribution of that feature in that season. For metric X in season y:

$$(X - \bar{X}_{in\ y})s_{X\ in\ y}$$

(subtract the mean of all X's in that season and multiply by the standard deviation of X in y).

- Missing Values - Our data was mostly complete except for 483 games where the response variable was missing. This was due to a MLB phenomenon called double-headers where two teams play against each other twice on the same day (attending fans purchase only one ticket for both games). In these instances, MLB reports attendance for the second game only, marking the attendance for the first game as zero. We chose to drop the first games of double-headers, since they are highly influenced by the second game.

Step 4 - Data Visualization & Analysis

We started out by plotting our response variable across all our data. Figure 1 shows the distribution of our target variable. As can be seen attendance ranged from a minimum of 746 to a maximum of 80,227 (excluding missing values which were described in step 3).

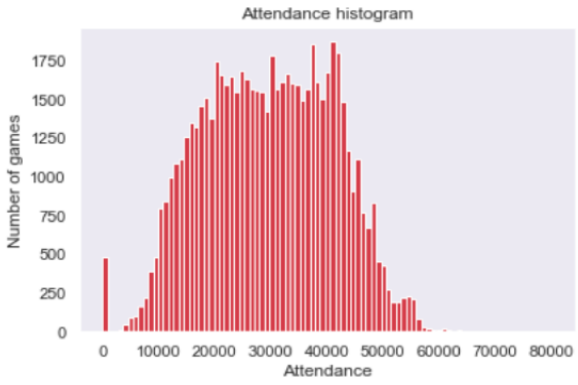

Figure 1 - Target Variable Distribution

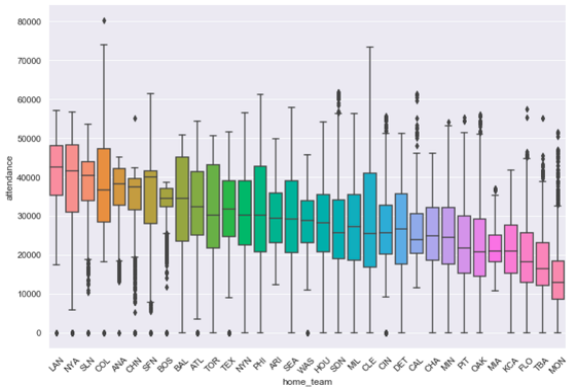

Figure 2 - Attendance By Team

We also analyzed our target variable's behavior across the different baseball teams in our dataset. As can be seen in figure 2, teams enjoy different levels of popularity and the resulting difference in attendance can be almost twice the spectators for the two extremes. In order to better understand our initial dataset and gain insights into possible directions for feature selection, we plotted the correlation matrix between dominant feature. As can be seen from figure 3, we learnt that in-game data & financial data are highly correlated with the response variable. This motivated us to seek data which could capture similar information.

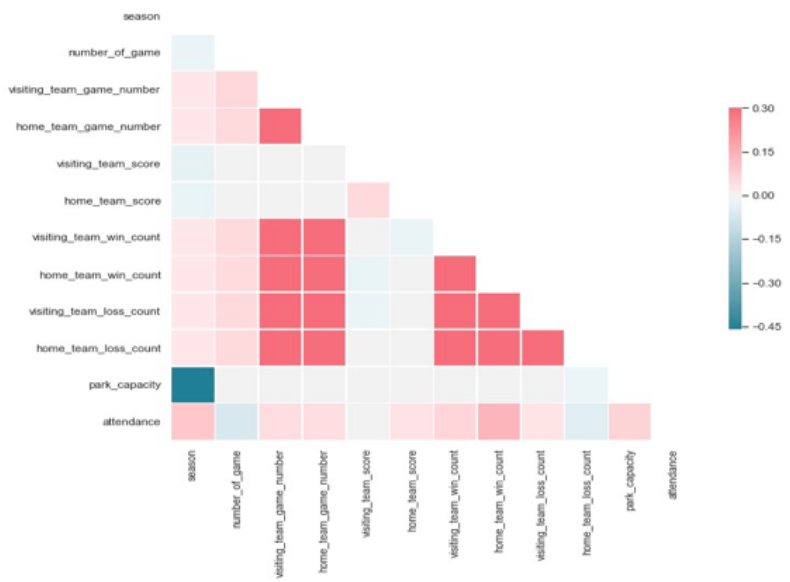

Figure 3 - Initial Correlation Matrix

Step 5 - Initial Model Run and Feature Augmentation

After gaining some intuition regarding the dataset at hand, we ran initial models to establish our baseline. Our choice was to start with a Linear Regression which yielded a RMSE of 8.834 and R^2 of 0.48. Clearly, there was room for improvement. We then focused on building our metrics for our enriched data containing the data integrations and normalizations described in Steps 2 and 3 to improve from our baseline. We chose to explore two main paths in regard to our initial models, Ridge & Lasso Regression and KNN Clustering. Our initial models performed only slightly better as can be seen in Table A.

Model	R^2	RMSE
Baseline Linear Regression	0.48	8.834
Ridge Regression	0.56	7.97
KNN Clustering	0.68	9.49

Table A - Initial Models

At this point, we understood that with our current data and models we will not be able to reach our benchmarks. We focused on the insights gained in step 4 to produce more features which relate to

quantifying demand for game viewership (using financial metrics) and game importance We added a few substantial features which boosted our model's accuracy:

- Seasonal Average Ticket Prices - We decided to add features which provide a proxy for the "demand side", that is the fans' willingness to pay for a game. We used data provided by Rodney D. Fort, a leading sports economics professor from the University of Michigan[3]. This metric was normalized to address currency inflation.
- Cumulative Player's "Baseball Age" - One of our hypotheses was that a player's popularity builds over time and we wanted to capture features which impact the team-crowd relationship. "Veteran" players have much better name recognition, and players that stay with the same team for many years become icons. One of the major features we set out to build was each of the player's "baseball age" up to the current game, that is the number of games the player appeared in an opening lineup. In order to construct this complex feature we had to use baseball game logs starting from 1970s. We used the data to create team's average "baseball age" for each game, which proved to be an important indicator for game attendance.
- Contention Score - How important a given game is to the ability of the team to contend for a spot in the playoffs. Given a team's current rank, the amount of games left in the season and its current win record, what is the probability that the team will win enough games to capture first place from its top division contender before the season is over.

  Let PCT = current win record (wins/games played). GL = games left to be played in the season. GB = "games behind" (team loss count – division rival loss count). Subscript α = metric for the team with division rank $\alpha(\alpha > 1)$

  X_α= a random variable - the number of wins that the team with division rank α will win out of its remaining games. $X_\alpha \sim Bin(PCT_{\alpha,}GL_\alpha)$. X_c = same as X_α for team α's division contender. Contention Score = $P(X_\alpha \geq X_c + GB_\alpha)$

Step 6 - Benchmarking Models

As we understood during the initial model, we were required to perform feature augmentation. The results on the new dataset were superior and beat the benchmarks utilizing out of the box Ridge & Lasso regressions. We wanted to further improve our accuracy and researched potential ensemble models which very much like baseball games, are popularity contests. Our results with Random

Forest showed a slight improvement. Since our target variable is a categorical variable, we chose to use the H2O python package which can handle such scenarios.

Model	R^2	RMSE
Ridge & Lasso Regression	0.78	5.34
Random Forest	0.77	5.13
Random Forest By Team	0.95	7.01
XGBoost	0.91	4.99
GBM	0.89	4.73

Table B - Benchmarking Models

Step 7 - Model Selection & Hyperparameter Tuning

In Step 6, Gradient Boosting Techniques achieved the most accurate results. During our analysis we discovered that what impacted the score was the Gradient Boosting regularization. XGBoost has a more regularized model formulization [5] and thus our decision was to proceed with GBM in order to further calibrate the regularizations, which showed the best results at 0.06. Our final model achieved a **RMSE of 4.73 with R^2 of 0.89**.

Our Findings & Statistical Evaluation

Our final model reached R^2 of 0.89, this can testify for overfitting, but we also got RMSE of 4.73 and even lower MAE (3.46). Thus, the final model is relatively successful on both fitting the whole dataset and for prediction. The model includes boosting of many trees which may suggest that the model is complicated, the early layers fitted simple models and the later fitted models to the errors of the early ones. On the other hand, the stopping criterion was set to 0.06 while the default value is 0.001, this indicates that the data can easily be overfitted if not controlling this parameter.

Figure 4 shows the most important variables of the model, we can see that 4 out of the top 10 variables are features engineered by us. Among them some that influence the ability and willing of a person to attend the game (ticket price, holiday) and some that indicate the importance of the game or abilities of the team (salary, contention score). The park ID was found as the most important variable, maybe because it incorporates both the home team identity and the park

capacity or because the park conditions (accessibility, visibility, convenience) themselves influence the decision to attend a game

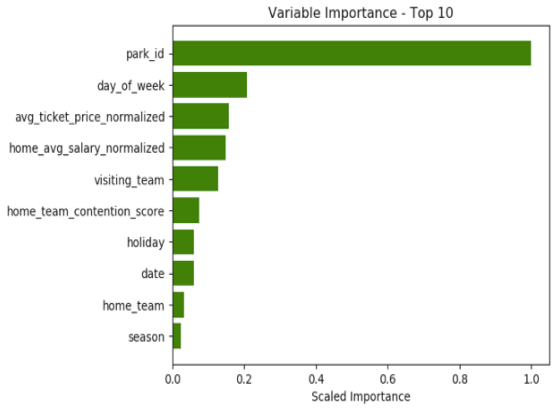

Figure 4 – Variable Importance

Insights and Applications

Our results improve the existing industry benchmarks by **29.2%**. Although the ability to predict baseball attendance has a dramatic financial impact on the MLB teams pricing strategy, our project delivers a more important social benefit. Research has shown the dramatic effect of MLB attendance on traffic congestions, where an additional 1000 spectators cause 10.74 annual hours of traffic delay. By extension, our results enable a 29.2% improvement in pre-game traffic preparations which could potentially improve urban pollution, economic waste and citizen satisfaction.

Related Work

Previous work	Description	Model	Results
Predicting MLB Game Attendance [6]	Game data from 1990-2016 and player metrics from baseball-reference.com	GBM	R^2: 0.83
Predicting Baseball Games Attendance [7]	Game data from 1990-2014 and stadium capacity from Wikipedia	Random Forest	R^2: 0.72 RMSE: 6.113
MLB Attendance [8]	Predicting % of capacity utilizing financial data and game metrics	Linear Regression	R^2: 0.60
Can Temperature Predict Attendance at MLB Games? [9]	Utilizing average monthly temperature and stadium condition (Dome/Open Roof)	Ridge Regression	MAE: 5.665

Annex I - Citations

[1] Traffic Congestion due to MLB Games- http://busecon.wvu.edu/phd_economics/pdf/17-05.pdf
[2] Communication Network Degradation - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.391.7813&rep=rep1&type=pdf
[3] Rodney Fort's homepage - https://sites.google.com/site/rodswebpages/home
[4] H2O Python library Documentation - http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html
[5] The difference between XGBoost and GBM https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting/answer/Tianqi-Chen-1
[6] Predicting MLB Game Attendance - https://towardsdatascience.com/predicting-mlb-game-attendance-c36cdc1b8de6
[7]- Predicting Baseball Games Attendance https://r-dir.com/blog/2015/02/predicting-baseball-game-attendance-with-r.html
[8] MLB Attendance -http://washusportsanalytics.weebly.com/uploads/5/0/0/9/50097143/final_project_-_group_5_aronson_finch_justus_parker_(1)_(1).pdf
[9] Can Temperature Predict Attendance at MLB Games? -https://medium.com/coinmonks/can-temperature-predict-attendance-at-mlb-games-d0b6a50217e0