

Prediction of Major Factors affecting Fans Attendance for the Teams of Major League Baseball

MSc Research Project
Data Analytics

Rahul Gupta
Student ID: x18131115

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland
Project Submission Sheet – 2019/2020
School of Computing



Student Name:	Rahul Gupta
Student ID:	18131115
Programme:	MSc Data Analytics
Year:	2019/20
Module:	Research Project
Lecturer:	Dr. Cristina Muntean
Submission Due Date:	12/12/2019
Project Title:	Prediction of Major Factors affecting Fans Attendance for the Teams of Major League Baseball
Word Count	8294
Page Count: 26	

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Rahul Gupta
Date:	December 12, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Major Factors affecting Fans Attendance for the Teams of Major League Baseball

Abstract

Fans attendance has evolved to be a major factor in Major League Baseball. The early knowledge of how many fans will watch the match in stadium could aide team owners and managers to make promotions effectively. Fans engagement has evolved over the past decades. This development calls for an approach which capitalizes the engagement of sports fans deliberately and economically. An extensive literature review was conducted regarding the area of fans attendance in Major League Baseball, and several valid conclusions were drawn from them. This research was carried out to investigate the contribution of game statistics on the fans attendance figures in their Major League Baseball games, using and statistical techniques and machine learning algorithms. The study analysed the results from Multiple Linear Regression (MLR), Random Forest Regression, Support Vector Regression (SVR), and Artificial Neural Networks (ANN). All the results were evaluated using 4 performance metrics. Random Forest Regression model produced the best results with RMSE of 0.02, MAD of 0.09 and MAPE of 2%. The model explained that different game factors affect the teams in a different way. It was also concluded that the fans attendance depends highly on in-game statistics, for the old franchises than for the new ones.

Keywords: Major League Baseball, machine learning, attendance, team performance

1. Introduction

Major League Baseball (MLB) is the professional league played for Baseball in United States of America (USA). MLB comprises of 30 teams or franchises. Every franchise is scheduled to play a game in either their own stadium or its opposition's stadium. Sports fans are the "consumers" which are very emotionally connected to their team. Fan engagement has evolved over the course of years, and now the franchises are adapting more deliberate and economic approach to monetize this engagement. This can be achieved by filling the stadiums with as many people as possible. Since attendance figures has become the driving factor in franchise's economics, the forehand knowledge of the figures could greatly help franchises in devising effective marketing and sales strategies for monetary purposes. The initial analysis have shown that the Attendance figures have continuously dropped in the past decade. This research tried to study the influence of various game statistics fans' attendance figures in Major League Baseball (MLB). The initial analysis have shown that the Attendance figures have continuously dropped in the past decade.

Ormiston (2014) studied the effect of number of star pitchers in the team, on the fans attendance figures in a season. The effect of star power of a pitcher continuously declined from 1994 to 2010, and therefore the presence of a star pitcher did not attract fans for the games. But there have been a lot of new players since 2010, and so have their style of play. So, the presence of star players may now have a considerable effect on the fans. This factor was included in the research. Waltering (2018) studied the effect of win percentage of teams, on the fans attendance figures. The study concluded that the win percentage has positive influence on the fans, and the teams who won the most matches has higher number of fans attendance. The study utilised

ANOVA, MLR and Logistics Regression to compute the results. The statistical techniques does not give the best results when the relationship is non-linear. This project will employ machine learning models to interpret any linear or non-linear relationship. Şahin and Erol (2017) used Neural Networks and ANFIS models to forecast the attendance of soccer games. Both the models were able to forecast attendance figures with error less than 10%. This research will employ the same models, but with the increased number of inputs. Jiang, Huang and Zhang (2017) reviewed the scope of Support Vector Regression (SVR) model, for prediction purpose. The study concluded that SVR has the ability to map both the linear and non-linear relation between input and output. These findings and observations were the motivation for conducting the research.

The motivation for this research lead to the formulation of following research question:

“To what extent can the in-game factors affecting the attendance figures for the teams of Major League Baseball be predicted based on in-game statistics and historical attendance figures, using Machine Learning algorithms, to enable team owners and managers to select the best players for the team?”

The study had the following objectives:

- To analyse the contribution of various factors which significantly affect the attendance figures in a game of Major League Baseball. The datasets contain a variety of in game statistics. The study aims to identify the factors which will be helpful in predicting the attendance with the maximum accuracy. By eliminating the redundant factors, and keeping the important factors, the predictions will be very accurate.
- To study and select the best machine learning model for the prediction of fans attendance. Various models will be applicable for prediction purposes. The model with the least erred results will be adopted for predicting the results for all the 30 teams of the league. These underlying trends will then be studied to explain the extent of effect of the right factor.
- To build the best machine learning model. The model will take into account the most suitable parameters for predicting the attendance figures. The gathered data will be tailored to the input requirements of every model. The results of the model will also be checked for overfitting or underfitting. The model will be trained with major portion of gathered data, which in turn will be helpful for a good prediction.
- To identify and propose evaluation criteria for the study. The evaluation criteria will evaluate the following – Is the model making good prediction? Is the value of the error within the acceptance range?

Although the study was done with all the due considerations, there were still few limitations. The study utilised the in-game statistics, tickets price and twitter data to investigate their effects on the fans attendance figures in the tournament. The game statistics is the aggregated figures, and the statistics of a team is independent of its opposition in a particular game. So, any effect of the opposition team is not considered. Average price of the tickets price is taken in the study. And hence the deviation from the actual prices might translate into results. And hence, biased results could be produced.

Section 2 will give the critical analysis of the previous work done on predicting the attendance figures in various sports league. It will also provide a critical analysis of the machine learning

algorithms used in this area. Section 3 will detail the step taken to complete in the methodology to complete the research. Section 4 will include the design decision undertaken and the step-by-step details of data processing. Section 5 will provide the details of the implementation of various machine learning models. Section 6 will present the evaluation of the results of the study. Section 7 will present the conclusion of the study, and the future work.

2. Literature Review

In the past, several researches or study have been done in this area. These studies ranges from calculating the attendance figures of only one season based on success rate of all the teams to studying the attendance figures due to effects of post-season promotion, to doing a case study on Major League Baseball to examine the relationship between attendance and competitive balance of the league. Following section will provides a critical analysis of the existing studies.

2.1 Critical review of the studies predicting several factors affecting the attendance figures in various sports

Davis (2009) studied the contribution of percentage wins for a team, on the fans attendance in the league games. Generalized Auto Regressive Conditional Heteroscedastic (GARCH) effect was used to analyse the underlying trends, and then the Multiple Regression Analysis was done to quantify the effect of winning percentage on the fans attendance figures. This study successfully established for every team which has winning percentage of more than .500, that there is almost an increase of 15,200 fans per game in attendance, i.e. the team has won more than half of its played matches. There were some exceptions in the results, even after this concrete findings. For the small dataset or the newer teams, the model was unable to explain the effect of coefficients of various factors.

Chen and Lin (2010) took opposition team, rival leagues, weekend matches and location of the game (whether the match is played home or away), and studied its effect on the fans attendance figures in the Chinese Baseball league. This study used Multiple Regression analysis to predict the values of attendance figures. This study established that the weekend games attracted most audience and there was an increase in the fans attendance, whenever the home team played a higher ranked team than them in the league. The model successfully explained 54.2% variance in attendance. This model could be expanded to include other variables to a new model, explaining more amount of variance in the output.

Ormiston (2014) studied the effect of number of star pitchers¹ on the total fans attendance figures of the team, for a season. This study used Linear and Binary Logistics Regressions model to calculate the attendance figures based on the basis of number of star pitchers. The results showed that there is a sharp decline in a player's career star power to attract audience in the stadiums from 1994 to 2010. Similarly, the Age Adjusted star power also failed to attract the audience in the stadiums. Only 3 of the legendary players showed a statistically significant impact in bringing the fans to the stadiums. This study also established that the star power of the visiting team's pitcher is equally important as the star power of home team's star pitcher. Finally, the study also established that there has been a decline in fan responsive ness to star pitchers. A significant issue

¹ Pitcher is a player in baseball, who throws the ball from pitcher's mound to the catcher to start every play.

with this study was that the process that was used to calculate the star power of a particular pitcher. It was contentious, and thus could have resulted in biased findings.

Waltering (2018) studied the effect of Winning percentage of a team on its attendance figures. This study followed the innovative approach of calculating the average attendance as a portion of stadium capacity, instead of calculating the aggregated figures. The data was collected for a period of 16 years, i.e. 1998 to 2013. This study used four types of models – Cross tab analysis, ANOVA, Multiple Regression and Binary Logistics Regression. Regression analysis provided the least erred results. The results successfully established that the winning percentage is highly responsible for the attendance figures. This study also established that, over the course of the seasons, the stadium were only 67% occupied. These results will be very helpful in this research. The same analysis could also work for all the teams.

Soebbing and Watanabe, (2014) examined the effect of variance in ticket prices on the attendance figures. The ticket-price data was collected for the period 1975-2008. This study used the two-step generalized method of moments (GMM) model. This study utilized two measures for price dispersion: Number of price levels and Gini co-efficient of the dollar value of all the price levels. A regression model was build using the price dispersion with other variables and then the GMM model was applied on the resultant dataset. The results showed that the increase in dispersion in price has a negative impact on the attendance figures. These results could be helpful for a franchise to evaluate opportunities to maximise revenue.

Meehan JR., Nelson and Richardson, (2007) calculated the difference between winning percentage home and away or “competitive balance”, and studied its effect on the attendance figures in Major League Baseball. This study used Multiple Regression model to calculate the attendance figures with competitive balance, capacity of stadium, weather conditions, day on which the game was played. This study proposed 3 models depending on the type of independent variables used. The degree of competitive balance was common in all the models. Of all the models, 2 models showed a decrease in attendance, when the home team has poor winning percentage than the visiting team. There was a very small impact on the attendance figures when the home team has a better winning percentage than the visiting team. Although the study was successful in giving valuable insights, it did not consider some factors such as games remaining for the team and rank of the team on the effect of change in competitive balance on the attendance figures. This could make the results, biased. And hence, a more broad and extensive study could result in better results.

Gitter and Rhoads (2010) studied the effect of quality of team on the attendance figures of Minor League Baseball (MiLB). Regression model was built to predict the attendance figures, and the predictors were MiLB team quality, MiLB team characteristics and same characteristics of their corresponding MLB teams. This study established that those MiLB teams, for whom the corresponding MLB teams saw increase in attendance, saw an increase in attendance figures too. This might help some teams to adopt the marketing strategies from the senior team, for the junior team. So, the effect of proximity could also be studied in the future research

Tainsky and Winfree, (2010) examined the effect of presence number of international players on the fans attendance figure for a single Major league Baseball season. This study utilized multiple regression analysis to predict the attendance figures. The data was taken for the period 1985-2000. The results showed that the demography of population is very importance in the attendance figures. Initially with the increase of an international player in the league, the attendance figures

decreased steadily from 1985 to 1992. These figures then increased from 1992 to 2000, and started falling from 2000 to 2005. So, the effect remained unstable over the time. There is a need to further investigate the underlying trends, including other variables to explain the instability.

Wakefield (2016) used Fans' Passion, Media Consumption and Social Media behaviour to predict Attendance figures. This study showed that intimacy, commitment and team identification impacted the attendance figures, severely. Just because of the passion for the team, almost 60% of the surveyed people attend games in stadiums. These findings would help franchises significantly to formulate relationship with ticket sales and sales possibilities.

Mohan (2010) studied the effect of image of away team on the home fans attendance figures in professional hockey. The study considered fans at different levels: fans having seasonal tickets, mini-plan ticket and game day purchasers. This study showed that the image of away team has a positive effect on fans. Cost, weather, hospitality factors and safety, of the away team, were the most important attributes that contributed to the image of the away team. This study also established that the seasonal ticket holders and game-day ticket holders, seriously considered the above factors, when travelling to different cities to watch the game. This data can be further used to market and monetize the road trips more effectively, by home as well as away teams.

2.2 Critical Assessment of models predicting Attendance in Sports (2009-2019)

Ahn and Lee (2014) used Panel Factors model to predict attendance figures of Major League Baseball. This study utilised the historical attendance figures from 1904-2012. The study was broken down into two periods 1904-1957 and 1958-2012. The model established that winning percentage of home team has the major influence in the attendance figures during the period 1904-1957, whereas quality of stadium and it's size, playing styles and uncertainty in outcome became the major determinants during the period 1958-2012. 3 Regression models were built by taking several variables into account, and then a panel data model was built with the common factors of these three regression models. The major advantage of this model is that the factors loading accounts for the time-invariant and time varying nature of the omitted variables.

Mills and Fort (2018) carried out the team level time-series analysis on the Attendance figures and outcome certainty for the Major League Baseball (MLB). The study failed to reject that there is any impact of the breaks on the attendance in MLB. This study also concluded that there was a significant impact of World War 1, World War 2 and the Great Depression Era on the aggregate attendance figures of each team. The reasons of these variations was not explained in the research. And hence a more extensive research would be very helpful.

Groothuis, Rotthoff and Strazicich (2015) also performed time series analysis to determine the trends or structural breaks in the game over the past years. This study showed that there were 2 structural breaks in 1921 and 1992 for mean slugging percentage. 1921 was the era of Babe Ruth, who was famous for his free swinging style. So, more players adopted the free swinging style in 1921 which resulted in first structural break in 1921. MLB banned the use of steroids in 1992. So, time-series analysis proved to be significant in determining the different eras in sports and provide meaningful insights.

Şahin and Erol (2017) performed a comparative analysis between Adaptive Neuro Fuzzy Inference System (ANFIS) and Neural Networks (NN) to predict attendance figures for the soccer games. Training data was used to train the model and testing data was used to evaluate the performance of model. The results showed that Neural Network (NN) model produced better results than ANFIS model. Mean Absolute Deviation (MAD) and Mean Absolute Percent Error (MAPE) statistics were

used to evaluate the performance of the models. The main advantage of Neural Networks is that it reduces the error between training data and predicted results by adjusting the weightage. NN model is also an alternative to the regression models and time-series analysis for larger data sets. Whereas ANFIS, which combines the concepts of Neural Networks and fuzzy inference systems, learns faster than Neural Networks, and captures the non-linear nature of the data rapidly and adapts accordingly. The effective implementation of ANFIS can produce better results.

Wang et al. (2011) used the structured time series analysis model to forecast the fans attendance in Chinese Professional Baseball League (CPBL) of Taiwan. This study established that the impact of fixing scandal was disastrous and attendance figures in not affected by GDP. This study was able to present concrete results because it considered several trends, cycle effect to forecast short-term results and seasonal factors too. The accuracy of the results was evaluated using testing data of attendance figures of 2009 and 2010.

Jiang, Huang and Zhang (2017) reviewed recent developments, the existing theory, scope and methods of Support Vector Regression (SVR). This study was done on the gray use of SVM on gray images. SVM compressed the gray images comprehensively. This study also concluded that SVR is capable of performing prediction with even a million points. SVR can also deliver excellent results in time series application and regression purpose. SVR model will be utilized in this study for prediction purposes.

2.3 Conclusion

This section provided detailed critique of the existing work in this area of research. This section detailed the performance of various machine learning models that were applied to find the fans attendance figures. It also detailed the factors which effect the fans attendance figures in Major League Baseball (MLB). This research will apply the Random Forest regression model to the baseball data, to calculate the major factors affecting the fans attendance. And will also report the factors for each team, and not on the championship as a whole. These factors will help team owners and coaches to trade players in the league.

3. Research Methodology

Every research in the field of Data Analytics follows one of the 3 methodologies: Knowledge Discovery in Data (KDD), Cross-Industry Standard Process for Data Mining (CRISP-DM), or SEMMA. Figure 1 shows the various stages of methodology employed in the research.

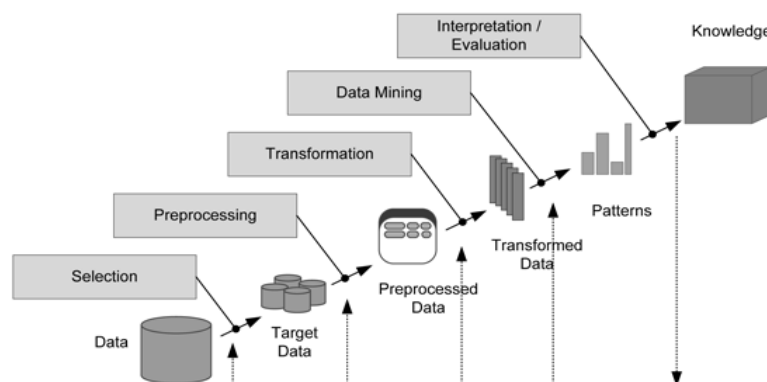


Figure 1 - Stages of KDD methodology²

²https://www.google.com/search?tbm=isch&sxsrf=ACYBGNTJiRKEoiVA1lwjShV7P4lgsb5y0A:1575111571033&q=kdd+methodology&chips=q:kdd+methodology,g_1:data+science+project&usg=AI4-

This research project employed KDD methodology. There are 5 stages of KDD methodology:

3.1 Data Extraction

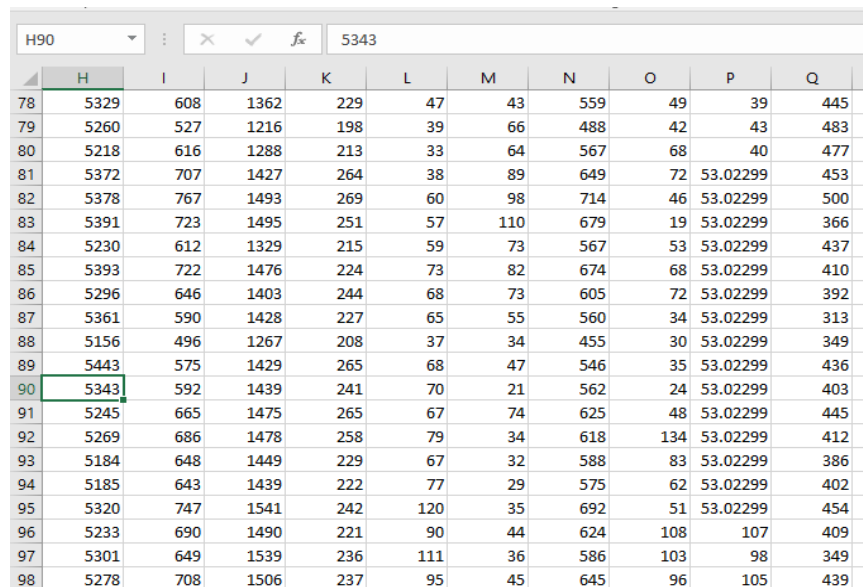
At this stage, the required data was collected. The data was collected from multiple sources. The data was collected from following sources:

- (i) <https://www.baseball-reference.com>
- (ii) https://en.wikipedia.org/wiki/List_of_current_Major_League_Baseball_stadiums

The details of past attendance figures and in-game statistics was collected from the first data source. The data was downloaded in the structured format. In-game statistics includes historical batting, pitching and fielding details. The CSV file of the batting contained 24 features. The CSV file of the pitching contained 16 features. The CSV file of the historical stats contain 10 features.

3.2 Data Preprocessing

3.2.1 Data Cleaning – The collected data had N/A values in few records. Those cells were imputed with the mean value of that column. Figure 2 shows mean values in a column.



	H	I	J	K	L	M	N	O	P	Q
78	5329	608	1362	229	47	43	559	49	39	445
79	5260	527	1216	198	39	66	488	42	43	483
80	5218	616	1288	213	33	64	567	68	40	477
81	5372	707	1427	264	38	89	649	72	53.02299	453
82	5378	767	1493	269	60	98	714	46	53.02299	500
83	5391	723	1495	251	57	110	679	19	53.02299	366
84	5230	612	1329	215	59	73	567	53	53.02299	437
85	5393	722	1476	224	73	82	674	68	53.02299	410
86	5296	646	1403	244	68	73	605	72	53.02299	392
87	5361	590	1428	227	65	55	560	34	53.02299	313
88	5156	496	1267	208	37	34	455	30	53.02299	349
89	5443	575	1429	265	68	47	546	35	53.02299	436
90	5343	592	1439	241	70	21	562	24	53.02299	403
91	5245	665	1475	265	67	74	625	48	53.02299	445
92	5269	686	1478	258	79	34	618	134	53.02299	412
93	5184	648	1449	229	67	32	588	83	53.02299	386
94	5185	643	1439	222	77	29	575	62	53.02299	402
95	5320	747	1541	242	120	35	692	51	53.02299	454
96	5233	690	1490	221	90	44	624	108	107	409
97	5301	649	1539	236	111	36	586	103	98	349
98	5278	708	1506	237	95	45	645	96	105	439

Figure 2 N/A values replaced by mean values

The data was otherwise consistent and it had no garbage or inconsistency. After the cleaning of data, exploratory data analysis was done to get insight about the data.

3.2.2 Exploratory Analysis

After the cleansing of data, gaining insight about the data is one of the most important step in the research of data analytics. Exploratory Data Analysis is essential to build a robust model. Statistical techniques are widely used to detect any underlying information from data. Multivariate Data Analysis is used to identify underlying patterns and analyse the dominance of any variable in a dataset. Principal Component Analysis (PCA) and Factor Reduction (FA) will be used to obtain a set of variables. For regression purposes the data must be linearly

distributed, or the relationship between dependent and independent variables must not follow a curvilinear relationship. Scatter plots were drawn to assess the relationship. These plots plotted various independent variables against the dependent variable, i.e. “Attendance.G” It can be seen from the Figure 3 shows that most of the data is linearly separable, and does not follow any curvilinear relation. The figure only shows the scatter plot of some of the dependent variables, but the linear relationship was checked for every independent variable.



Figure 3 Normal Distribution check for Attendance Figures

Correlation plot was used to check the multicollinearity among the predictors. It was found out several variables were highly correlated. As seen in Figure 4, the predictors are highly correlated. So, most of the predictors are telling the same thing. And, hence the variation in the output could be explained by less number of such predictors. So, such highly correlated variables were dropped to obtain a set of variables which are not correlated. The same process was undertaken to drop the highly correlated variables from all the 3 CSVs. And then a consolidated dataset was prepared in which the predictors are not highly correlated.

Figure 6 shows the cumulative variance explained by all the principal components. It is the graph plotted with Number of Principal Components against the Cumulative Proportion of Variance with the help of those many variables. The first 2 Principal Components (PC1 & PC2) together were able to explain 55% of the variation in output. Similarly first 3 (PC1, PC2 & PC3) together were able to explain 68% variance in the output, and so on.

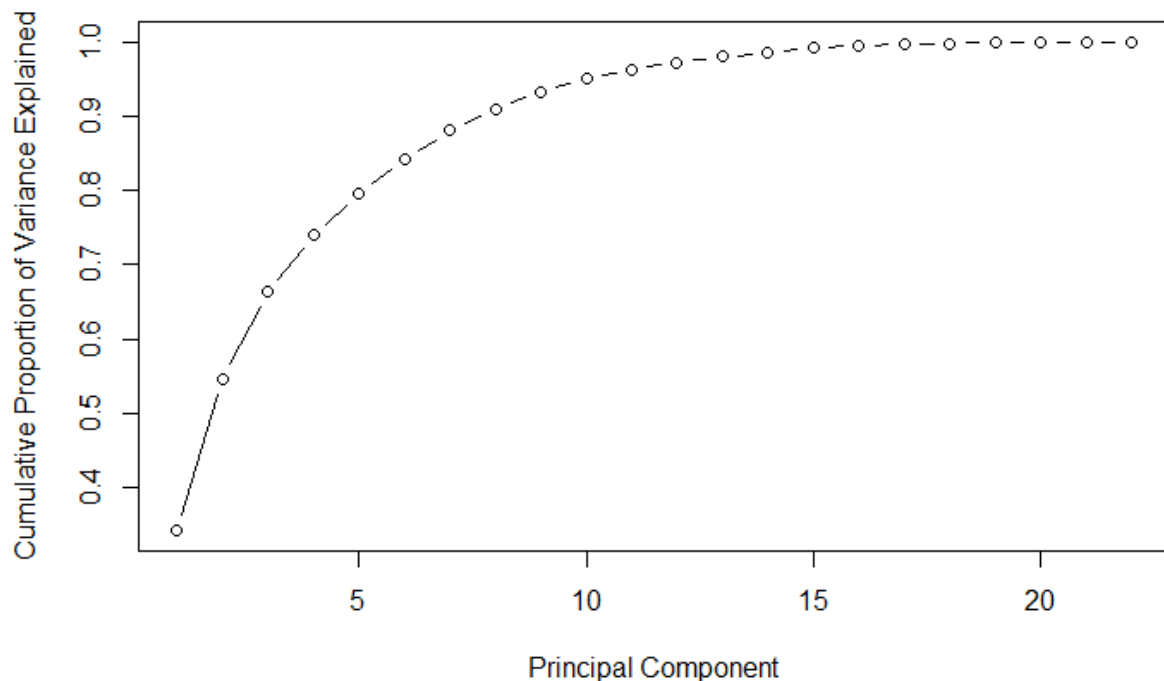


Figure 6 Cumulative contribution of the Principal Components (PCs)

With the help of PCA, 37 predictors in the original dataset were reduced to 22. These variables could explain the maximum variance in the output. These PCs were then consolidated to form the final dataset. Figure 7 shows the correlation plot between the variables in the dataset with the Principal Components (PCs). It can be seen from the figure that these new variables are not correlated. These new variables or PCs were then used to form the training and testing data for Artificial Neural Networks model, and the predictions were made on this data.

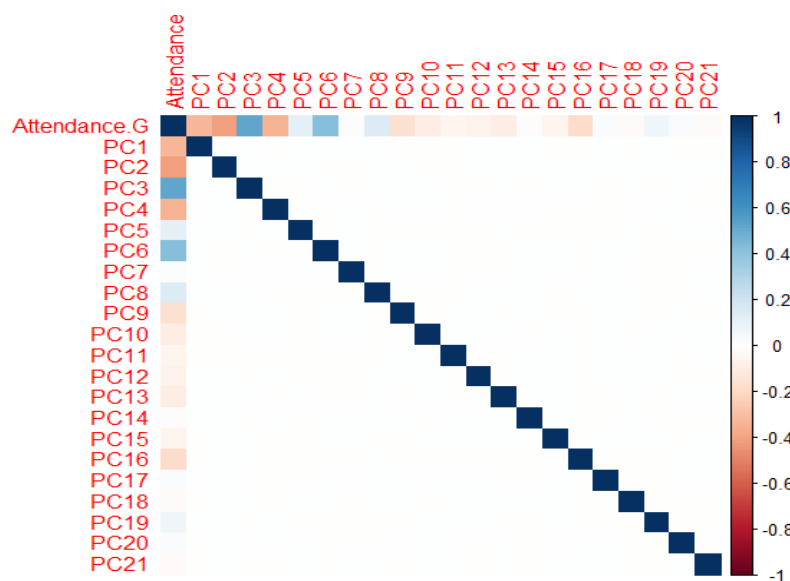


Figure 7 Correlation Plot after Principal Component Analysis (PCA)

3.4 Data Mining

At this stage, various machine learning models were applied to the data. Following model were applied in this study:

- Multiple Linear Regression (MLR)
- Random Forest (RF)
- Artificial Neural Network (ANN)
- Support Vector Regression (SVR)

The models were trained and tested with the appropriate training and testing datasets respectively. The predictions and the description of the models are given in the next section. All the predictions were consolidated, and were then analysed to produce the best results.

3.5 Evaluation and Interpretation

After all the results, from all the models, were compiled, they were evaluated. The evaluation was done to find out the model that predicted best or least erred results. The results were evaluated using following evaluation criteria:

- Root Mean Square Error (RMSE)
- Mean Absolute Deviation (MAD)
- Mean Absolute Percentage Error (MAPE)

After the evaluations, the study was able to suggest one model which produced best or the least erred results. These predictions were then analysed to study the major factors which affect the fans attendance figures in Major League Baseball (MLB). These results were visualized using Tableau to present the graphical view of the extension of effect of each factor on the attendance. These results could be utilised, by the team coach and/or team owners to either trade players based on good game stats or improve their existing performance.

4. Design Specifications

Figure 8 outlines the proposed design approach undertaken for the research project. The above figure gives the visual description about what steps were taken, during the execution of the project.

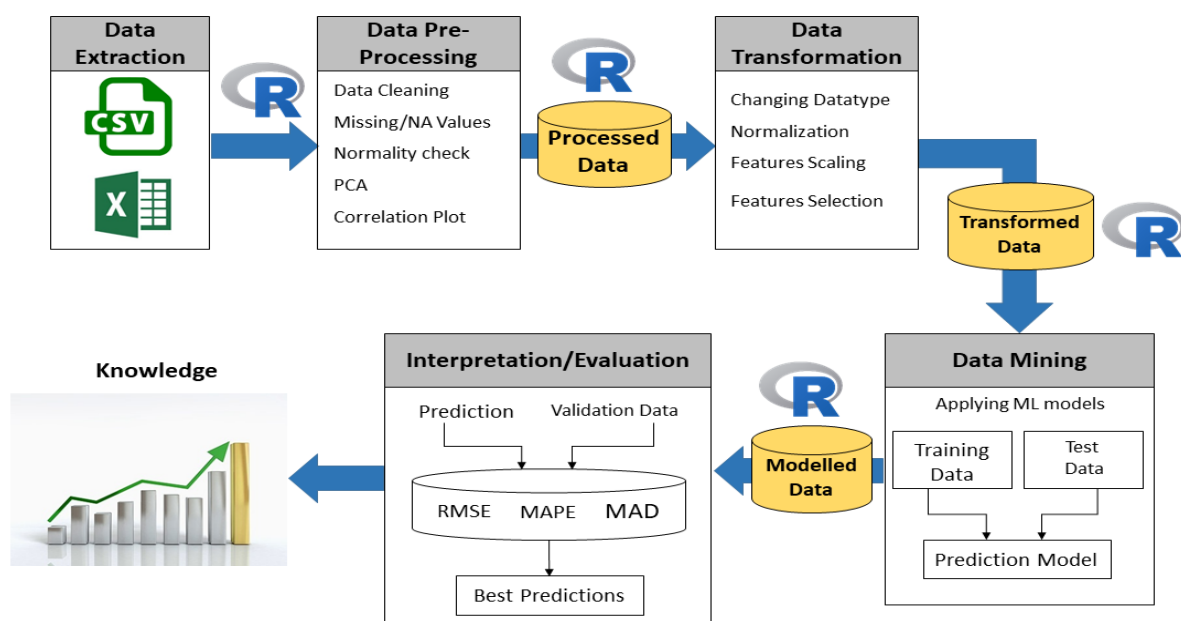


Figure 8 Proposed Design Approach

The research focused on determining the in-game factors which are affecting the fans attendance in the games of Major League Baseball (MLB). The project included following steps:

- 1) Firstly, the data was extracted both in structured and unstructured format.
- 2) The extracted data was cleaned to remove all or any inconsistencies, garbage values and N/A values. Exploratory Data Analysis (EDA) was then done to get an overview of the data. Few important inferences were made during EDA.
- 3) The processed data was then transferred into the format, so that the machine learning models could be applied on them. The transformations include: normalization of data, changing the datatype of data and replacing the missing data with the mean values.
- 4) The transformed data was then applied to the 4 machine learning models for making predictions. The models were adjusted to get the best predictions.
- 5) The results were consolidated, and analysed. These results were evaluated using 3 performance metrics. The result of the model, which gave least erred results was taken. The extent of effect of various factors affecting the fans attendance figures was analysed and reported.

5. Implementation

The implementation or the machine learning models were made using R programming language. “caret” package was installed to run the machine learning models. The dataset was collected for all the 30 teams of the league. The number of records varied for every team. The data contains figures from the inception of the franchise to the figures for 2019 season. So, the machine learning models were tested with the data of 2018 and 2019, i.e. 2 rows. And, rest of the rows were used as training dataset.

5.1 Implementation of Multiple Regression model

Multiple Regression analysis is a statistical method which is used to identify the relationship between dependent and independent variables variable. This regression model is very helpful in understanding the extent of effect of various factors on the attendance figures. “mlrbench” library was loaded in R studio to run the model. The model was tested using the data of the last 2 years, i.e. 2018 and 2019. Rest of the data was used for training the model. The model was trained and tested 30 times, to get the predictions for all the 30 teams of the league. Figure 9 gives the description of the model and the values of descriptive of the model.

```
Call:
lm(formula = Attendance.G ~ ., data = training_set_indians)

Residuals:
    Min       1Q   Median       3Q      Max
-3783.6 -1149.3   69.1  1225.4  5219.7
```

Figure 9 Model description

Figure 10 gives the summary of the implemented model. It can be seen that the model is giving the Adjusted R squared value of 0.817, i.e. the model is able to explain 81.7% of variance in the output. So the model is giving the desired values, and hence the model was adopted to implement for all the teams.

```
Residual standard error: 2166 on 85 degrees of freedom
Multiple R-squared:  0.8671,    Adjusted R-squared:  0.817
F-statistic: 17.33 on 32 and 85 DF,  p-value: < 2.2e-16
```

Figure 10 Summary of the model

Figure 11 shows the value and significance of all the coefficients. “*” sign indicate that the variable is significant at 95% confidence interval. So, with this model we can say that, we are 95% sure that the fans attendance figures, for Cleveland Indians team depends upon PA (Number of plate appearances), 2B (Number of double-base hits) and Page (Average Age of Pitchers in the team). So, there is a chance of getting more fans in the stadium if the team can manage more plate appearance or a batter completes his turn batting, or the audience sees more double plays or even if the Average Age of pitchers is more or the pitchers are very experienced. The same model was applied for all the teams, and the results are reported and explained in the Results section.

Coefficients: (1 not defined because of singularities)				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.247e+05	4.014e+05	0.311	0.7597
Finish	-5.272e+02	7.599e+02	-0.694	0.4966
R.G	-1.511e+03	1.893e+04	-0.080	0.9373
PA	5.786e+01	2.688e+01	2.153	0.0452 *
AB	-7.347e+00	2.843e+01	-0.258	0.7990
R	-9.816e+01	1.444e+02	-0.680	0.5052
H	4.073e+00	3.336e+01	0.122	0.9042
X2B	-7.458e+01	3.454e+01	-2.160	0.0445 *
X3B	-6.241e-01	1.101e+02	-0.006	0.9955
HR	4.066e+00	3.918e+01	0.104	0.9185
RBI	6.002e+01	9.158e+01	0.655	0.5205
SB	-3.206e+01	1.962e+01	-1.634	0.1197
CS	2.195e+01	5.975e+01	0.367	0.7176
BB	-5.131e+01	2.825e+01	-1.817	0.0860 .
Fld.	-2.294e+05	3.525e+05	-0.651	0.5234
BatAge	3.524e+02	6.977e+02	0.505	0.6196
RA.G	1.458e+04	2.816e+04	0.518	0.6109
ERA	-1.018e+04	5.003e+04	-0.203	0.8410
CG	2.643e+01	8.610e+01	0.307	0.7624
tSho	3.360e+02	2.238e+02	1.501	0.1506
SV	5.772e+01	1.179e+02	0.489	0.6304
IP	-1.604e+02	1.455e+02	-1.103	0.2847
H.1	-9.164e+00	1.859e+01	-0.493	0.6279
R.1	-1.023e+02	1.884e+02	-0.543	0.5937
ER	1.196e+02	3.329e+02	0.359	0.7235
HR.1	1.228e+01	3.701e+01	0.332	0.7439
BB.1	-2.580e+01	1.586e+01	-1.627	0.1211
SO	5.625e-01	7.245e+00	0.078	0.9390
PAGE	1.824e+03	7.859e+02	2.321	0.0322 *
PAGE.1	NA	NA	NA	NA
X.Bat	3.422e+02	1.674e+02	2.044	0.0559 .
X.P	-1.709e+02	3.189e+02	-0.536	0.5986
Win pctg	5.331e+04	3.728e+04	1.430	0.1698

Figure 10 Output of Multiple Linear Regression model

5.2 Implementation of Random Forest model

The model was built using R programming language, in R studio. “randomForest” and “ie2misc” libraries were installed to build the model. Random Forest is a combination of decision trees. Multiple trees are combined to give more accurate result. Main advantage of Random Forest is that the multiple uncorrelated trees work in parallel, and so the accuracy of each decision tree is combined to give more accurate result. In the project, the model was tested using the data of the last 2 years, i.e. 2018 and 2019. Rest of the data was used for training the model. The model was

trained and tested 30 times, to get the predictions for all the 30 teams of the league. An output of the model for one of the teams is given below:

The forest was constructed using 400 trees. And, the model plotted the importance of significant variables.

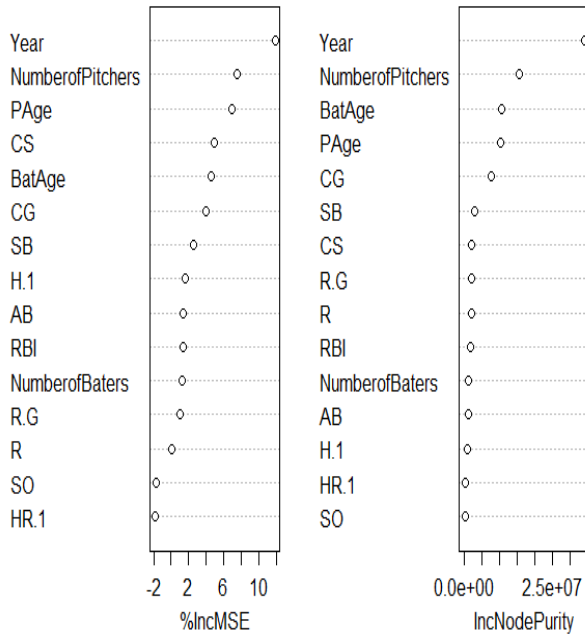


Figure 11 Variable Importance plot

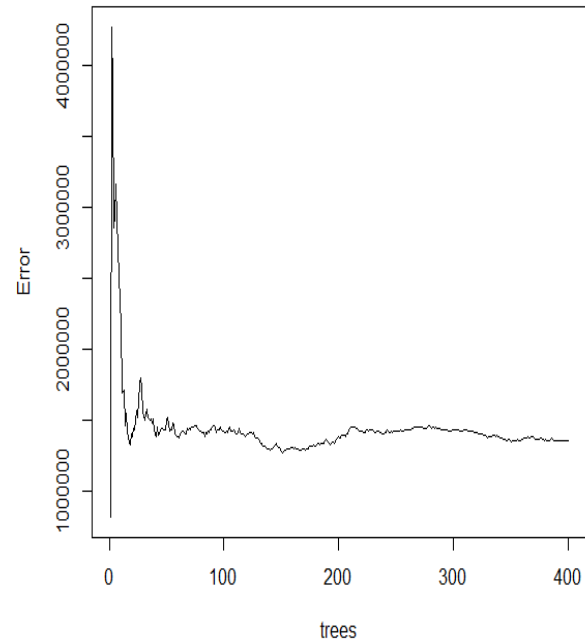


Figure 12 Error vs Number of Trees graph

%IncMSE or Mean Decrease Accuracy graph indicate the value by which the accuracy of the model will decrease if a particular variable is left out. IncNodePurity or Mean Decrease Gini Imp is a measure of variable importance. Gini Impurity Index is defined as the probability of a variable being randomly classified incorrectly, when chosen randomly. So, higher the values of %IncMSE and IncNodePurity, the more important is the variable. Figure 11 gives the Variable Importance plot for the implemented model. From the figure it can be noted that Year, Number of Pitchers, Batter Age, Pitcher Age and Completed games have high values for both %IncMSE and IncNodePurity. So, these variables have greater effect on the fans attendance. There is a chance of more fans attending the game if there are experienced players in the team or there are higher number of pitchers in the team.

Figure 12 plots the error rate with the increase in the number of trees. It can be easily seen that error decreases with the increase in the number of trees.

5.3 Implementation of Support Vector Regression model

The model was built using R programming language, in R studio. “penalizedSVM” & “e1071” libraries were loaded in the R studio for the execution of the same. Support Vector Machine algorithm can be used for the purpose of both classification and regression. For classification, SVM creates a hyperplane to separate different classes. But for regression purposes, SVM sets a margin of tolerance (ϵ), as the prediction is based on real numbers. The main advantage of SVM over traditional regression models, is that it minimizes the general error to achieve high performance. In the project, the model was tested using the data of the last 2 years, i.e. 2018 and 2019. Rest of the data was used for training the model. The model was trained and tested 30 times, to get the predictions for all the 30 teams of the league. An output of the model for

one of the teams is given below: Linear function was used as the kernel function in the model, as the data was normally distributed.

```
Call:
svm(formula = Attendance.G ~ ., data = train_ori_svm, kernel = "linear", cost = 10,
    scale = FALSE)

Parameters:
  SVM-Type:  eps-regression
 SVM-Kernel:  linear
      cost:  10
    gamma:  0.03125
  epsilon:  0.1

Number of Support Vectors:  19
```

Figure 13 Output of SVM model

Figure 14 gives the summary of applied model. “Cost” is defined as the cost of misclassification of data. The ultimate goal of SVM model is to identify hyperplane that would best separate the points of two classes. So, cost parameter is given to handle the cost of misclassification. “gamma” is the parameter of Gaussian kernel, to handle the non-linear classification. Small value of gamma indicates low bias and high variance in output. So, model is able to explain large variance in the output.

Parameter tuning was then done to choose a set of optimal parameters. 10-fold cross validation method was used for sampling the data. Figure 15 shows the output of parameter tuning. As seen from the figure, the margin for error tolerance (ϵ) was obtained as 0. Also, the cost of classification or C value is 2, i.e. maximum of 2 points were only misclassified by the model.

```
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
  epsilon cost
    0      2
- best performance: 1855561
```

Figure 14 Parameter tuning in SVM

After the parameter tuning and creation of various models, the best model was selected. Figure 16 shows the summary of the best SVM model. Now we have the optimized SVM which has best possible parameters for the model. The prediction was then made based on this model.

```
Call:
best.tune(method = svm, train.x = Attendance.G ~ ., data = train_dbk_svm, ranges = list(elsilon = seq(0,
1, 0.1), cost = 1:100))

Parameters:
  SVM-Type:  eps-regression
 SVM-Kernel:  radial
      cost:  3
    gamma:  0.03030303
  epsilon:  0.1

Number of Support Vectors:  19
```

Figure 15 Best SVM model

Figure 17 shows the values of the support vectors and constants. These values are used to tell the extent of effect of each variable on the output. The values are combined with the constant to form an equation which gives us the best hyperplane. The values of different variables are then placed to come up with an attendance figure.

```
> W = t(dbk_svm_model$coefs) %*% dbk_svm_model$SV
> W
      Finish      R.G      PA      AB      R      H      X2B      X3B      HR      RBI      SB
[1,] -7.979342  0.1621962 -12.21044 -18.26035  30.61933  36.48124 -51.22713  80.04268 -13.17063 -23.42577 -32.91217
      CS      BB      Fld. BatAge      RA.G      ERA      CG      tSho      SV      IP      H.1
[1,]  59.34942  20.48761 -0.00934287  7.925  0.03458366 -0.02345188  40.87924 -23.35123  11.11596  48.17905 -13.59948
      R.1      ER      HR.1      BB.1      SO      PAge      Year NumberofBaters NumberofPitchers Win_pctg
[1,]  5.535453  9.840278 -33.16913  14.7036  1.712268  7.56038 -49.0251      -28.17383      -42.06374  0.1238121
      Loss_pctg
[1,] -0.1238121
> b = dbk_svm_model$rho
> b
[1] -172630.2
```

Figure 16 Predictions of SVM model

5.4 Implementation of Artificial Neural Network (ANN) model

The model was built using R programming language, in R studio. “neuralnet” library was loaded in the R studio for the execution of the same. ANN is a deep learning algorithm which is capable for the purposes of both classification and regression. It consists of 3 layers: input, hidden and output layer. The input layer receives input, hidden layer comprising of neurons, normalizes input and does the processing, and output layer, also comprising of neurons, gives the output. The main advantage of ANN is that it can easily detect any non-linear relationship, and hence no prior assumptions are required. Also, there is no possible way defining the number of hidden layers. It depends upon the data. In the project, the ANN model was trained with different number of hidden layers, and the model which gave the least erred results is presented below. In the project, the model was tested using the data of the last 2 years, i.e. 2018 and 2019. Rest of the data was used for training the model. The model was trained and tested 30 times, to get the predictions for all the 30 teams of the league.

Figure 17 shows the plot of implemented ANN model. Black lines shows the connection between layers, and the blue line shows the bias added on each step. This bias is same as the intercept in the regression model. The numbers written on the lines shows the weight on each of the connections. The model contains 3 hidden layers and the output layer. Different models were built using different number of hidden layers. But this configuration gave the best execution time and best results and hence, this model was adopted to calculate the attendance for each team.

```
> summary(dbk_ann_model)
      Length Class      Mode
call           5  -none-   call
response       20  -none-  numeric
covariate     380  -none-  numeric
model.list      2  -none-   list
err.fct         1  -none-  function
act.fct         1  -none-  function
linear.output   1  -none-  logical
data           20 data.frame list
exclude         0  -none-   NULL
net.result      1  -none-   list
weights         1  -none-   list
generalized.weights 1  -none-   list
startweights    1  -none-   list
result.matrix   67  -none-  numeric
```

Figure 17 Summary of the ANN model

Figure 18 shows the obtained neural network from one of the implemented models.

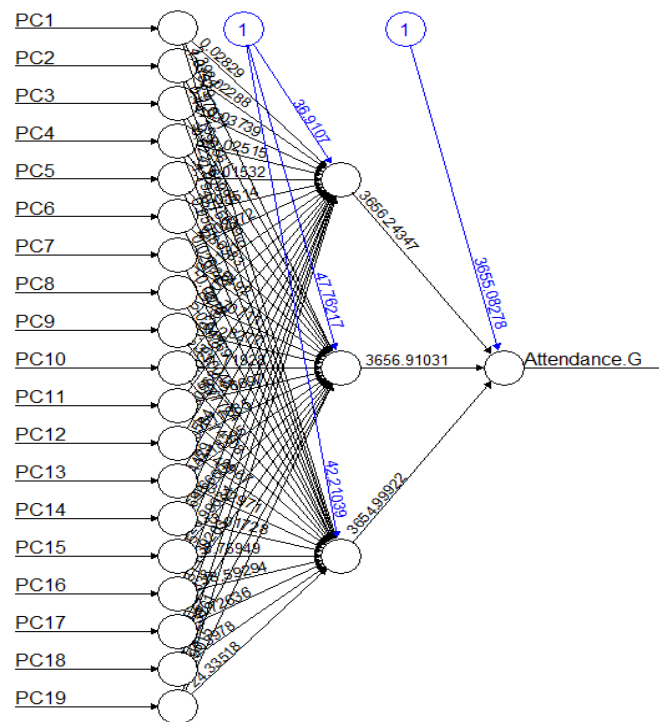


Figure 18 Neural Network plot

6. Evaluations and Results

In order to accomplish the desired results, multiple models were applied on the data to predict the fans' attendance figures and the various factors affecting it. After getting the prediction results, the models were evaluated for the precision in predictions. The closer the predicted and actual values are, the better is the model. So, the models were evaluated using some measures. Root Mean Square Error (RMSE), Mean Absolute Error or Deviation (MAD) and Mean Absolute Percentage Error (MAPE) were chosen as evaluation criteria for the models. These values were recorded for the results from all the models and are compared below. Figure 20(a) shows the comparison of RMSE values for all the models. It can be seen that the Random Forest model is giving the lowest RMSE values for all the teams.

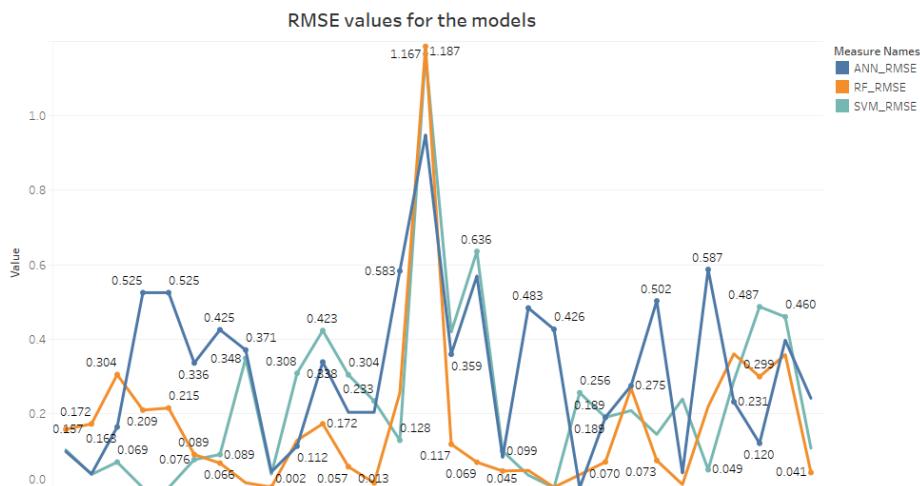


Figure 20(a) Comparison of RMSE values

Figure 20(b) shows the comparison of MAPE values for all the models. It can be again seen that the Random Forest model is giving the lowest MAPE values for all the teams.

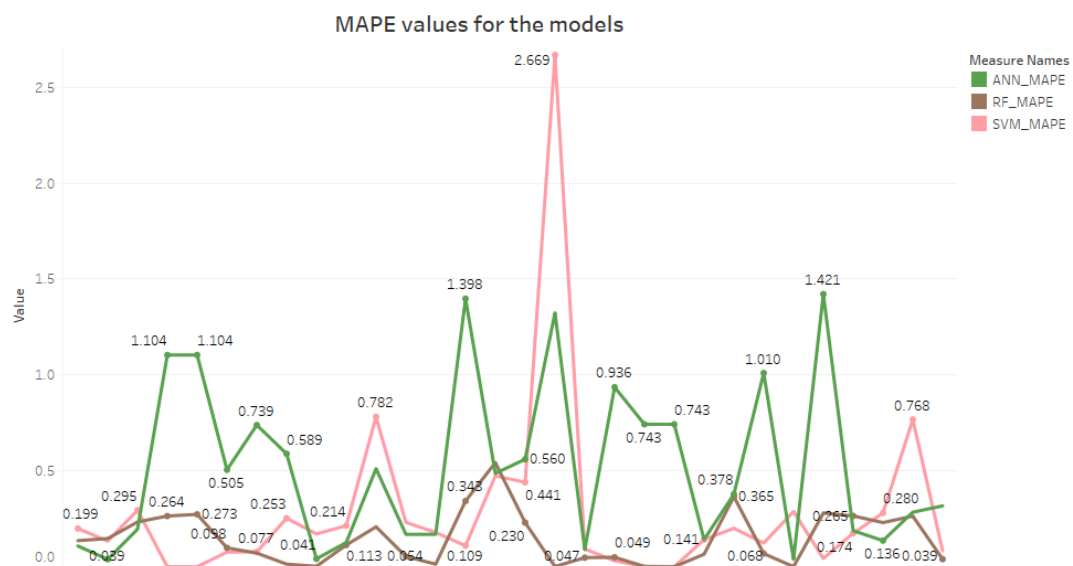


Figure 20(b) Comparison of MAPE values

Figure 20(c) shows the comparison of MAD values for all the models. It can be again seen that the Random Forest model is giving the lowest MAD values for all the teams.

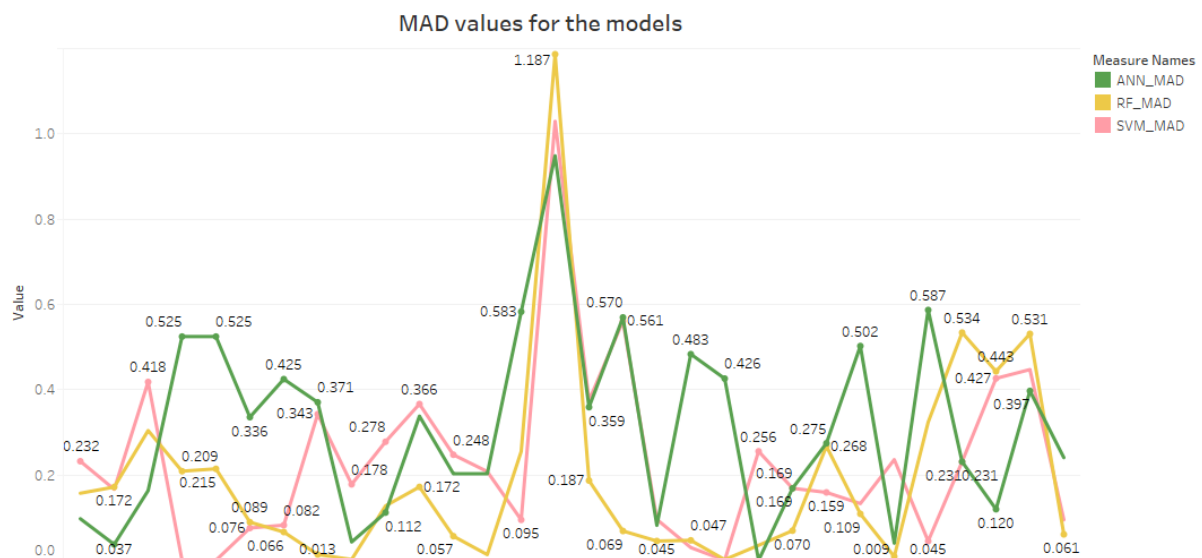


Figure 20(c) Comparison of MAD values

The average values of all the evaluation measures were taken to find out the model which works best for the teams. Figure 20(d) gives the comparison of the average values of RMSE, MAPE and MAD for the 3 models. From the figure, it is clear that the Random Forest Regression model gave the least erred results, i.e. the values of errors is least. Random Forest regression model was applied to calculate the attendance figures for the 30 teams of the league. This project will report the results for the 6 teams with lowest attendance figures.

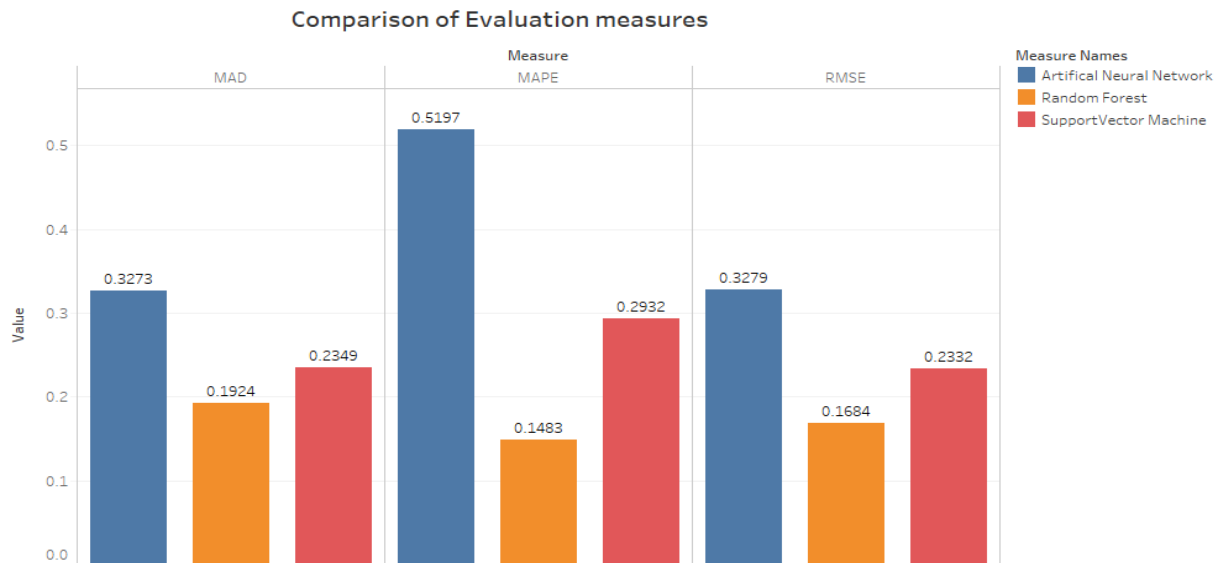


Figure 20(d) Comparison of average values evaluation measures from all the models

Table 1 show the RMSE, MAD and MAPE values for all teams, from Random Forest model.

Team	RMSE	MAD	MAPE
Arizona Diamondbacks	0.1572735	0.1572735	0.1359
Atlanta Braves	0.1718567	0.1718567	0.1466534
Baltimore Orioles	0.3043272	0.3043272	0.2333212
Boston Red Sox	0.2089785	0.2089785	0.2641882
Chicago Cubs	0.214516	0.214516	0.2731005
Chicago White Sox	0.0894844	0.0894844	0.0982788
Cincinnati Reds	0.0663411	0.0663411	0.0710549
Cleveland Indians	0.0134662	0.0134662	0.0132873
Colorado Rockies	0.0018494	0.0018494	0.0018528
Detroit Tigers	0.1276483	0.1276483	0.1131987
Houston Astros	0.1721206	0.1721206	0.2079054
Kansas City Royals	0.0565619	0.0565619	0.0535339
Los Angeles Angels	0.0132935	0.0132935	0.0131191
Los Angeles Dodgers	0.2551718	0.2551718	0.3425916
Miami Marlins	1.1867173	1.1867173	0.5426935
Milwaukee Brewers	0.1172715	0.1869908	0.2299984
Minnesota Twins	0.0687733	0.0687733	0.0007385
New York Mets	0.0452038	0.0452038	0.047344
New York Yankees	0.0468722	0.0468722	0.0491772
Oakland Athletics	0.0015525	0.0015525	0.0015501
Philadelphia Phillies	0.035181	0.035181	0.0339853
Pittsburgh Pirates	0.0695447	0.0695447	0.0650227
San Diego Padres	0.2675359	0.2675359	0.3652545
San Fransisco Giants	0.0734835	0.1089466	0.0684533
Seattle Mariners	0.0087863	0.0087863	8.71E-05
St. Louis Cardinals	0.2184946	0.3239401	0.2795817
Tampa Bay Rays	0.3601872	0.5340135	0.2648071
Texas Rangers	0.2988633	0.4430947	0.230096
Toronto Blue Jays	0.3582276	0.5311082	0.2637464
Washington Nationals	0.0410193	0.0608152	0.039403

6.1 Results for Miami Marlins team

RMSE is 1.86, MAD is 1.45 and MAPE is 0.54. Variable Importance plot of the Random Forest model gives the %IncMSE and IncModePurity plot. These are also called the Variable Importance plot. The greater the values for a particular factor in the plot is, the more important the factor is. From Figure 19, it can be clearly seen that factors like At Bats (AB), Earned Runs

or Runs that are scored by batting team's offense and not due to the error of opposition (ER), Average age of Pitchers (Page), tSho or Shutouts by a team (no runs conceded by one or more pitchers), HR or number of home runs scored by a team, 2B or double-base hits and RA/G or Runs allowed per game are the some of the important factors for the fans attendance. But these are of less importance as the prediction is highly erred. This means that the fans attendance for Miami team only marginally depends on the in-game statistics.

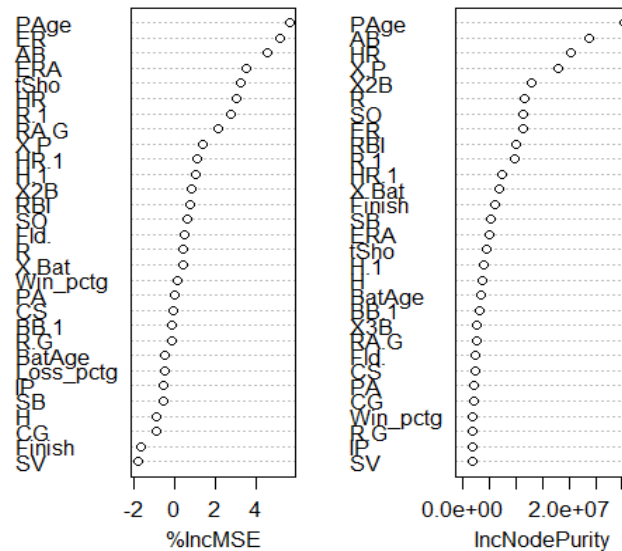


Figure 21 Variable Importance plot for Miami Marlins

6.2 Results for Tampa Bay Rays team

Figure 22 shows the variable importance plot for Tampa Bay Rays team. The is indicating that the factors like Pitcher Age, Caught Stealing (CS), Runs scored, Finish position, Earned runs and win percentage are important in fan attendance. RMSE is 0.36 and MAPE is 26%. These values are a little lower than the average RMSE value of the model. This is because there are several other factors, other than in-game statistics, that serves as a determinant for Tampa Bay Rays team's fans.

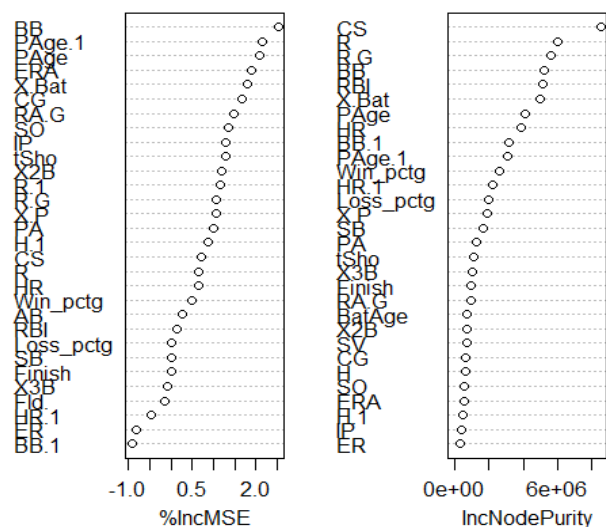


Figure 22 Variable Importance plot for Tampa Bay Rays

6.3 Results for Pittsburgh Pirates team

Figure 23 shows the variable importance plot for Pittsburgh Pirates team. The RMSE value is 0.06 and MAPE is 5%. So, the model was successfully able to predict the fans attendance accurately. And it can be safely said factors like Home Runs, Completed Games (CG), Number of Saves (S), Strikeout, Finish position and Double-base hits (2B) are the major factors in the fans attendance. The fans of Pirates team are very passionate about the game and they appreciate good players.

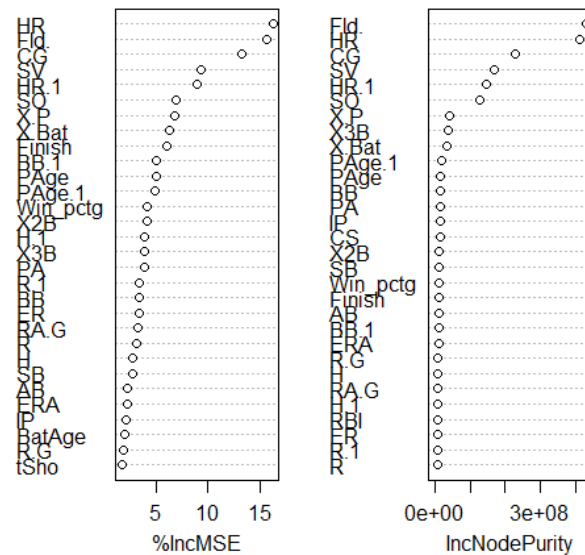


Figure 23 Variable Importance plot for Pittsburgh Pirates

6.4 Results for Baltimore Orioles team

Figure 24 shows the variable importance plot for Baltimore Orioles team. RMSE value is 0.30 and MAPE is 0.23. These are significantly higher than the average values, and so it could be said that in-game statistics are not the only factors for fans attendance figures and the fans of Orioles also keep other factors in mind while deciding to watch the match or not. But, in-game factors

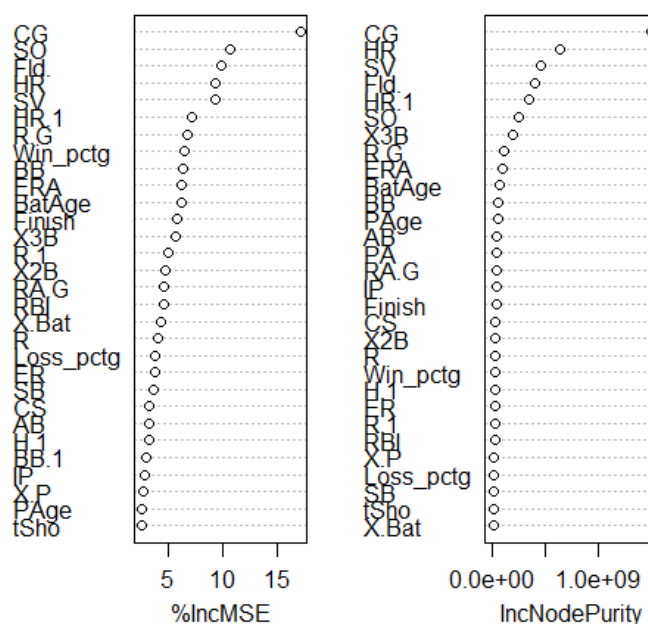


Figure 24 Variable Importance plot for Baltimore Orioles team

6.5 Results for Chicago White Sox team

Figure 1 consists of two side-by-side scatter plots. The left plot shows the relationship between %IncMSE (x-axis, 0 to 15) and various predictors (y-axis). The right plot shows the relationship between IncNodePurity (x-axis, 0e+00 to 3e+08) and the same predictors (y-axis). Both plots show a positive correlation between the predictor and the metric.

6.6 Results for Detroit Tigers team

Figure 1 consists of two scatter plots. The left plot shows the relationship between %IncMSE (x-axis, 0 to 20) and node labels (y-axis). The right plot shows the relationship between IncNodePurity (x-axis, 0e+00 to 3e+08) and node labels (y-axis). Both plots show a clear trend where higher node purity corresponds to higher %IncMSE.

Figure 26 Variable Importance plot for Detroit Tigers team

Home Runs Allowed, Saves and Strikeouts. All these factors are credited to the pitcher, and so it can be concluded that the fans of Tigers team like the pitching facet of the game. A good pitching performance will drive fans to watch the game in stadium.

7. Conclusion and Future Work

Major League Baseball (MLB) has become the most famous tournament among the baseball fans. Fans come to support their favourite team in the stadium. And the team/franchise owners look to monetize the passion of baseball fans for their team. So, fans attendance is of utmost importance for the team owners and managers. The fans attendance depends on several factors. This study was conducted to effect of in-game statistics and trends in historical attendance figures, on the fans attendance. The study analysed in-game statistics of all the 30 teams of the Major League Baseball (MLB) and applied several machine learning models to predict the fans attendance figures, based on the observed patterns. The results from the models were recorded, and Random Forest Regression model gave the best results, i.e. the least erred results. This model was then used for making predictions for all the teams. The results were then analysed to figure out major factors that affect the attendance figures. It was concluded that, for the new franchises (the teams which have joined the league recently), the model produced highly erred results than for the old ones. This maybe because the new franchises are not performing well or the sport is new to the fans and the fans are still not passionate about it. For instance, Tampa Bay Rays joined the league in 1890 and the model made the prediction with 94% accuracy. Whereas, Tampa Bay Rays team joined the 1998 and the model made prediction with accuracy of 64%. But the model still explained the extent of importance of various in-game stats on fans attendance.

In addition to studying the in-game statistics, further research could consider studying weather of the day and when is the game being played, i.e. day or night. The results would strengthen the basis of prediction. As most of the stadiums don't have a roof over them, fans might think before attending the game. In addition, more study could be done to implement the SVR model, as that model gave the second best results in this study, also supported in the study found by Jiang, Huang and Zhang (2017).

Acknowledgement

I would like to express my sincerest gratitude to my supervisor Dr. Cristina Muntean for her continuous support, enthusiasm, motivation and immense knowledge on the subject matter. Her guidance helped me immensely during the completion of research project. I could not have imagined a better mentor than her. I would also like to thank my classmates for their support and help. I also owe a sense of gratitude to my parents who encouraged and supported me for the completion of research.

References

- Ahn, S. and Lee, Y. (2014). Major League Baseball Attendance. *Journal of Sports Economics*, 15(5), pp.451-477.
- Chen, C. and Lin, Y. (2010). An Analysis of Game Attendance in the Chinese Professional Baseball League. In: 2010 International Conference on Management and Service Science. Wuhan: IEEE, pp.1-3.
- Davis (2009), Analysing the Relationship between Team Success and MLB Attendance with GARCH Effects, *Journal of Sports Economics*, 10(1), pp. 44-58.
- Gitter, S. and Rhoads, T. (2010). Determinants of Minor League Baseball Attendance. *Journal of Sports Economics*, 11(6), pp.614-628.
- Groothuis, P., Rothhoff, K. and Strazicich, M. (2015). Structural Breaks in the Game. *Journal of Sports Economics*, 18(6), pp.622-637.
- Jiang, H., Huang, K. and Zhang, R. (2017) 'Field support vector regression', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10634 LNCS(10), pp. 699–708.
- Meehan JR., j., Nelson, R. and Richardson, T. (2007). Competitive Balance and Game Attendance in Major League Baseball. *Journal of Sports Economics*, 8(6), pp.563-579.
- Mills, B. and Fort, R. (2018). Team-Level Time Series Analysis in MLB, the NBA, and the NHL: Attendance and Outcome Uncertainty. *Journal of Sports Economics*, 19(7), pp.911-933.
- Mohan, L. J. (2010) 'Effect of Destination Image on Attendance at Team Sporting Events', *Tourism and Hospitality Research*. Nature Publishing Group, 10(3), pp. 157–170.
- Ormiston, R. (2014). Attendance Effects of Star Pitchers in Major League Baseball. *Journal of Sports Economics*, 15(4), pp.338-364.
- Şahin, M. and Erol, R. (2017). A Comparative Study of Neural Networks and ANFIS for Forecasting Attendance Rate of Soccer Games. *Mathematical and Computational Applications*, 22(4), p.43.
- Tainsky, S. and Winfree, J. (2010). Discrimination and Demand: The Effect of International Players on Attendance in Major League Baseball. *Social Science Quarterly*, 91(1), pp.117-128.
- Wakefield, K. (2016). Using Fan Passion to Predict Attendance, Media Consumption, and Social Media Behaviours. *Journal of Sport Management*, 30(3), pp.229-247.
- Waltering, M. (2018). Attendance Still Matters in MLB: The Relationship with Winning Percentage. *The Sports Journal*, 20(1), pp.1-14.
- Wang, H., Wu, M., Lin, Z. and Chang-Chien, I. (2011). Forecasting the consumption of professional baseball in Chinese Taipei-new evidence from structural time series model. In: 2011
- Watanabe, N., Yan, G. and Soebbing, B. (2015). Major League Baseball and Twitter Usage: The Economics of Social Media Use. *Journal of Sport Management*, 29(6), pp.619-632.