Journal of Computer Science and Information Technology
December 2018, Vol. 6, No. 2, pp. 15-22
ISSN: 2334-2366 (Print), 2334-2374 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jcsit.v6n2a2
URL: https://doi.org/10.15640/jcsit.v6n2a2

Predicting Attendance at Major League Soccer Matches: A Comparison of Four Techniques

Barry E. King[1] & Jennifer Rice[2]

Sport team managers need to predict attendance levels at sporting events to plan staffing levels, plan inventories, and decide upon possible promotions. This paper discusses predicting attendance at Major League Soccer events using data from the 2014 and 2015 seasons. Panel data is obtained for each team, season, and weather category. A traditional least squared dummy variable linear regression technique is used along with three machine learning algorithms – random forest, M5 prime, and extreme gradient boosting. Extreme gradient boosting provides superior results with respect to out-of-sample root mean square error statistics. Well-founded technique for working with different methods is presented and the efficacy of contemporary algorithms is offered.

Keywords: Major League Soccer, machine learning, least square dummy variable linear model, random forest, M5 prime, extreme gradient boosting

1. Introduction

Attendance at sporting events has been well-researched as evidenced by the numerous references in the Literature Review section. Being relatively new, attendance at Major League Soccer (MLS) matches has not received as much attention. Prediction models for attendance mostly have been multivariate linear regression attempts. This study focuses on attendance at MLS matches and examines the efficacy of three machine learning regression methods in addition to a panel adjusted linear regression approach. The goal of the article is to demonstrate well-found practice in developing machine learning models and to examine the appropriateness of methods for constructing prediction forecasts.

2. Literature Review

2.1. Machine Learning

(Zhu and Chen, 2016) provide a thorough overview of extreme gradient boosting pointing out that the XGBoost library runs substantially more quickly and with fewer resources than other machine learning algorithms. This study's dataset is too small to take advantage of the speed of XGBoost; instead, it was chosen for its recent success at machine learning competitions. (Raut, 2016) provides good advice for selecting a machine learning algorithm. Unfortunately, it does not mention M5 prime or extreme gradient boosting, two techniques used in this analysis. (Trawinski, et al., 2012) discusses nonparametric statistical analysis for comparing machine learning regression algorithms. They show that pairwise Wilcoxon test, when employed to multiple comparisons, results in overoptimistic conclusions. For this reason, we employ Tukey Honestly Significance Difference test in this analysis.

[1]king@butler.edu, +1 317 9405464, Butler University, Lacy School of Business, 4600 Sunset Avenue, Indianapolis, IN 46208, United States of America
[2]jlrice@butler.edu, Butler University, Lacy School of Business, 4600 Sunset Avenue, Indianapolis, IN 46208, United States of America

2.2. Attendance at Sporting Events in General

Douvis (2007) and Douvis (2014) offer a thorough review of why fans attend professional sports. Although thorough, the work is now old. The author suggests that sport managers segment the customer base and then identify factors that influence the spectator decision-making process. A review of international literature on the demand for sport is provided by Borland and MacDonald (2003). The authors mention there are no simple lessons to be drawn from existing literature, but that uncertainty of outcome, quality of contest, and quality of viewing are important factors. Deshande and Jensen (2016) capture and compare highly paid National Basketball Association players who have a low impact to those players with high impact. The authors mention that existing metrics do not provide this comparison.

2.3. Attendance at Major League Soccer

An exploratory examination conducted by Karakaya, Yannopoules, and Kaflaki(2016) mentions"the results indicate that there are three major motivations – emotional excitement, socialization, and soccer atmospherics – and two identity salience factors – ardent soccer fans and rational soccer fans – for attending soccer games. The most important factor for attendance is being an ardent soccer fan closely followed by the emotional excitement factor. Among the demographic factors considered, only gender significantly affects soccer game attendance."Deshande, et al., (2016) observe that soccer-specific stadiums and proximity to the fan base were important to attendance. We too report that arena distance is negatively correlated with attendance.Uncertainty of outcome has been a factor in drawing attendance although the importance of this is debated (Paul and Weinbach (2007), Sung and Mills (2017), and Weinbach and Paul (2013)).

3. Research Question

Can machine learning models produce better predictions of Major League Soccer attendance than can traditional models such as a linear model configured for panel data?

4. Data

We acquired 572observations of Major League Soccer (MLS) matches for the 2014 and 2015 seasons. The raw data consisted of 62 box office variables including the number of full tickets sold, average ticket price, event date and time, and the attendance at the match. The score of the match was not included in the data. Most data were unusable. Only total attendance (the target variable), the number of full season tickets, event date and time, and average ticket price were used from the 2014 and 2015 data sets.Other data, such as weather, venue distance from downtown, Metropolitan Statistical Area (MSA) population, and other data were attached to the team data.

4.1. Data Cleansing

Rows with missing data were eliminated including those for Chivas USA (no 2015 data) and Sporting Kansas City (incomplete data). This resulted in 556 usable observations.

4.2. Legend for R Variables

This study used the R statistical language. R does not provide for variable labels to accompany variable names. Meaningful variable names must be constructed. Table 1 is a legend showing the R variable name, comment about the variable, and source of the variable's data.

Table 1: Initial model variables, comments, and sources

Initial Variable	Comment	Source
Total_Attendance	Target variable	Proprietary source
Arena_Distance_from_Downtown	In miles	Wikipedia by team
Average_Ticket_Price	Over all ticket categories	Proprietary source
Capacity	Published capacity not including standing room	Wikipedia by team
Full_Ticket_Quantity	Number of full season tickets	Proprietary source
Home_Team_Total_Salaries	Includes base salaries and total compensation for designated players	MLS players association (n.d.)
Lagged_Attendance_One_Match	Attendance one home match earlier	Derived variable

Lagged_Attendance_Two_Matches	Attendance two home matches earlier	Derived variable
MSA_Hispanic_Percentage	Proportion of MSA that is Hispanic	Census Information Center (n.d.)
MSA_Population	Metropolitan Statistical Area population	Census Information Center (n.d.)
MSA_White_Percentage	Proportion of MSA that is white	Census Information Center (n.d.)
Number_of_Home_Designated_Players	Count by team, by season	MLS designated players (n.d.)
Number_of_Visiting_Designated_Players	Count by team, by season	MLS designated players (n.d.)
Points_per_Season	3 points for a win; 1 for a draw	MLSsoccer (n.d.)
Visiting_Team_Popularity	Number of Google searches by team	Google searches by sport (n.d.)
Visiting_Team_Total_Salaries	Includes base salaries and total compensation for designated players	MLS players association (n.d.)
Home_Team	Categorical variable with 19 levels for 19 teams with usable data	Proprietary source
Season	Categorical variable with two levels: 2014 and 2105	Derived from date of match
Weather_Category	Categorical variable with three levels: Good, Moderate, and Bad	Weather underground (n.d.)
Early_Afternoon_Match	Binary indicator variable	Derived from date and time of match
Early_Evening_Match	Binary indicator variable	Derived from date and time of match
Late_Afternoon_Match	Binary indicator variable	Derived from date and time of match
Friday_Match	Binary indicator variable	Derived from date of match
Saturday_Match	Binary indicator variable	Derived from date of match
Sunday_Match	Binary indicator variable	Derived from date of match

4.3. Data Partition

The data were partitioned into an 80 percent training data set (n=435) and a 20 percent test data set (n=121). The sections Transformation through Final Variable Selection only used the training data set.

4.4. Transformation

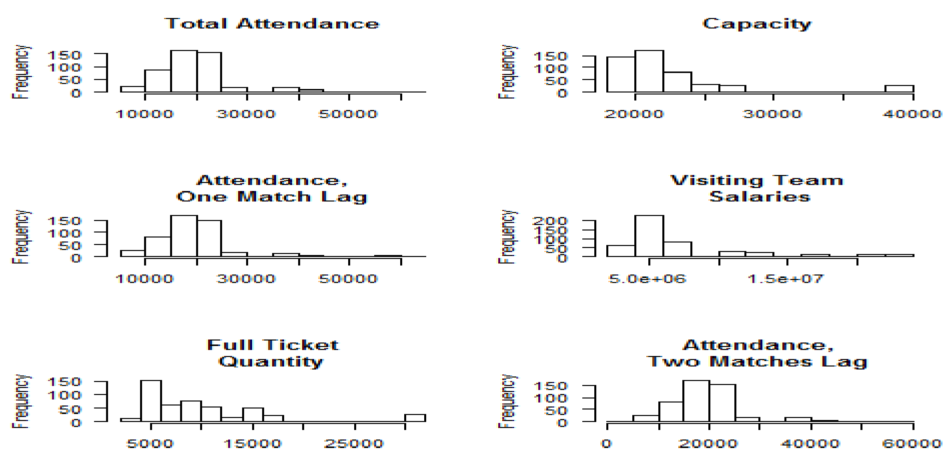
A test for skewness and kurtosis indicated six features may be right-skewed. See those variables with skewness greater than 2.00 in Table 2.

Table 2: Skewness and kurtosis for predictor variables.

Variable	Skewness	Kurtosis
Total_Attendance	2.47	8.97
Capacity	2.34	5.23
Lagged_Attendance_One_Match	2.32	7.30
Visiting_Team_Total_Salaries	2.27	4.55
Full_Ticket_Quantity	2.11	4.29
Lagged_Attendance_Two_Matches	2.07	6.23
Home_Team_Total_Salaries	1.93	3.01
Visiting_Team_Popularity	1.67	2.61
MSA_Population	1.64	1.31
Arena_Distance_from_Downtown	0.94	-0.30
Average_Ticket_Price	-0.66	-0.30
Number_of_Visiting_Designated_Players	-0.46	-1.07
Points_per_Season	-0.44	-0.59
MSA_Hispanic_Percentage	0.34	-1.27
MSA_White_Percentage	-0.32	0.93
Number_of_Home_Designated_Players	-0.29	-1.19

Histograms of the six variables with skewness greater than 2.00 are shown in Figure 1.The six variables in this figure were replaced with log transformations of the original variables.

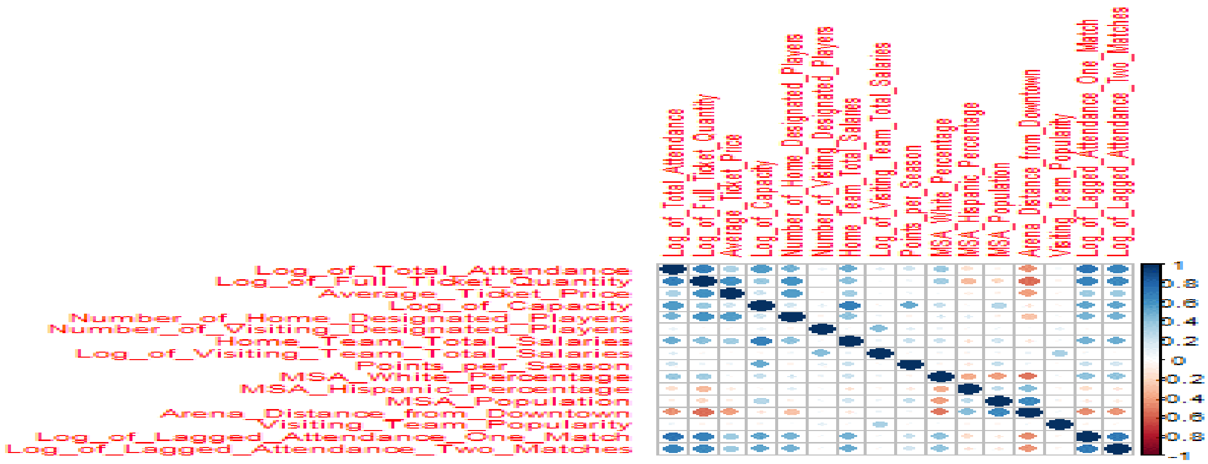
Figure 1: Six right-skewed variables.



4.5. Correlation Plot

Figure 2 plots the correlations of the numeric model variables, including the six log transformed variables. Note that the first column displays the correlation of the predictor variables to the target variable, Log_of_Total_Attendance.

Figure 2: Correlation plot



Logs of lagged attendance are highly correlated with the target,Log_of_Total_Attendance.Number_of_ Visiting_Designated_Players and Visiting_Team_Popularity are not well-correlated with Log_of_Total_Attendance.

4.6. Collinearity

Columns 2 (sqrt(VIF)) and 3 (Rejected for Collinearity) of Table 3 report the results of constructing a variance inflation factor (VIF)linear model. The sqrt(VIF) critical value is 2.00 for this analysis. Log_of_Full_Ticket_Quantity is eliminated due to collinearity.

Table 3: Collinearity and final feature selection.

Variable	sqrt(VIF)	Rejected for Collinearity?	Rejected by Boruta Selection Algorithm?	In Final Model?
Log_of_Total_Attendance	-	-	-	Yes
Arena_Distance_from_Downtown	1.66	No	No	Yes
Average_Ticket_Price	1.64	No	No	Yes
Home_Team_Total_Salaries	1.45	No	No	Yes
Log_of_Capacity	1.85	No	No	Yes
Log_of_Full_Ticket_Quantity	2.35	Yes	-	No
Log_of_Lagged_Attendance_One_Match	1.89	No	No	Yes
Log_of_Lagged_Attendance_Two_Matches	1.71	No	No	Yes
Log_of_Visiting_Team_Total_Salaries	1.32	No	No	Yes
MSA_Hispanic_Percentage	1.48	No	No	Yes
MSA_Population	1.60	No	No	Yes
MSA_White_Percentage	1.40	No	No	Yes
Number_of_Home_Designated_Players	1.66	No	No	Yes
Number_of_Visiting_Designated_Players	1.10	No	Yes	No
Points_per_Season	1.26	No	No	Yes
Visiting_Team_Popularity	1.24	No	No	Yes

4.7. Boruta Feature Selection Algorithm

Rather than use Akaike Information Criteria to finalize the predictor variable set, the Boruta feature selection algorithm was used.

"Boruta is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest. This package derives its name from a demon in Slavic mythology who dwelled in pine forests." (Analytics Vidhya, 2016.)Number_of_Visiting_Designated_Players was eliminated by the Boruta algorithm. See column 4 (Rejected by Boruta Selection Algorithm?) of Table 3.

4.8. Final Variable Selection

The final set of predictor variables is reported in column 5 (In Final Model?) of Table 3. Also included in the model are the panel data variables, Home_Town, Season, and Weather_Category, along with the six binary indicator variables of Table 1.

5. Algorithms and Tuning

Four algorithms were tuned and trained on the training data. In-sample statistics and 10-fold cross-validated root mean squared error (RMSE) out-of-sample statistics were developed.

5.1. Least Squares Dummy Variables Linear Model

The linear model used in this examinationused dummy variables for each team, season, and weather category. The variablesHome_Team, Season, and Weather_Category were presented to the lm() function of R as factors as were the six binary indicator variables for day of week and time of match.

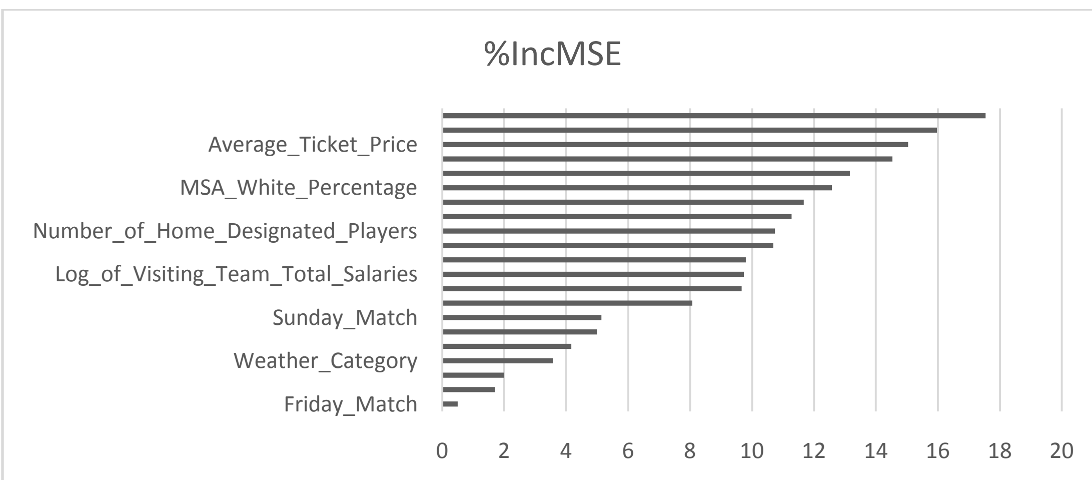
5.2. Random Forest

6. "Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set." (Wikipedia, n.d.).The mtry parameter was optimized by using the tuneRF() function of the randomForest package.

Variable importance

Figure 3 shows the relative importance of each variable to the random forest model.

Figure 3: Variable importance as the percentage increase in mean square error.



"IncMSE is the most robust and informative measure. It is the increase in [mean squared error] of predictions (estimated with out-of-bag-CV) [because of] variable j being permuted (values randomly shuffled)." (Welling, 2015)

6.1. M5 Prime

M5 prime is a tree-based piecewise linear modeling algorithm with linear models at the terminal nodes. (Quinlan, 1992.) It was tuned and trained using the train() function of the caret package.

6.2. Extreme Gradient Boosting

"XGBoost (eXtreme Gradient Boosting) is one of the most loved machine learning algorithms at Kaggle. Teams with this algorithm keep winning [machine learning] competitions. It can be used for supervised learning tasks such as Regression, Classification, and Ranking. It is built on the principles of gradient boosting framework and designed to 'push the extreme of the computation limits of machines to provide a *scalable*, *portable* and *accurate* library.'" (Nishida, 2017)Extreme gradient boosting was selected primarily because it has been performing well in machine learning competitions. It was tuned using one hot encoding for the panel data and then employing the xgb.train() function of the xgboost package for training.

Results

All four methods were tuned and trained on the training dataset. In-sample statistics were generated by running the tuned and trained methods against the training dataset. In-sample statistics are reported in Table 4.

Table 4: In-sample performance.

Statistic	Linear Model	Random Forest	M5 Prime	Extreme Gradient Boosting
Mean Absolute Error (MAE)	2532	1130	2596	2168
Mean Absolute Percent Error (MAPE)	14	6	13	12
Root Mean Square Error (RMSE)	3379	2134	4421	2940

Better statistics are obtained by running the methods against the test dataset. These 10-fold cross-validated out-of-sample statistics are reported in Table 5.

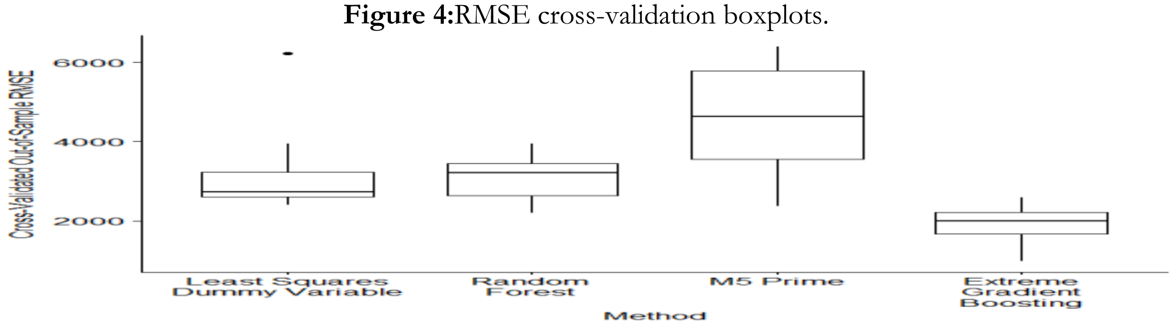
Table 5:Out-of-sample performance.

Statistic	Linear Model	Random Forest	M5 Prime	Extreme Gradient Boosting
Mean Absolute Error (MAE)	2584	2209	2701	1219
Mean Absolute Percent Error (MAPE)	13	13	15	6
Root Mean Square Error (RMSE)	3315	3094	4631	1925

7.1. Comparison Across Methods

Figure 4 contains boxplots of the RMSE cross-validation vectors.

Figure 4:RMSE cross-validation boxplots.



Note. Extreme gradient boosting appears to have the most favorable RMSE by a considerable margin. Tukey honestly significance difference test was applied to the RMSE data. The results are reported in Table 6.

Table 6: Tukey honestly significance difference test.

Method Pairs	Difference	Lower	Upper	p Adjusted
Random Forest-Least Squares Dummy Variable	-115	-1286	1055	0.99
M5 Prime-Least Squares Dummy Variable	1421	251	2592	0.01
Extreme Gradient Boosting-Least Squares Dummy Variable	-1285	-2456	-114	0.03
M5 Prime-Random Forest	1536	366	2707	0.01
Extreme Gradient Boosting-Random Forest	-1170	-2340	1	0.05
Extreme Gradient Boosting-M5 Prime	-2706	-3877	-1535	0.00

At the 0.05 p-value level of significance, extreme gradient boosting is significantly different than the other three methods.

Discussion

Extreme gradient boosting has emerged as the far superior technique for this study. Good practice warrants assessing the performance of a variety of techniques and not using just the favorable technique-of the-day for any regression problem.

We observe that a larger Hispanic population does not translate into larger attendance, that the number of home team designated players has a larger positive impact on attendance than does the number of visiting team designated players, and that weather is not as important as was initially thought.

References

Analytics Vidhya (2016), retrieved from https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/.

Borland, J. and MacDonald, R., (2003) Demand for sport, Oxford review of economic policy, 19(4), 478-502.

Census Information Center (n.d.) Retrieved from www.census.gov/cic.

Deshande, S. and Jensen, S., (2016) Estimating an NBA player's impact of his team's chances of winning, retrieved from https://arxiv.org/abs/1604.03186.

DeSchriver, T., RAscher, D., and Shapiro, S., (2016) If we build it, will they come? Examining the effect of expansion teams and soccer-specific stadiums on Major League Soccer Attendance; Sport, business and management: an international journal, 6(2) 205-227, https://doi.org/10.1108/SBM-05-2014-0025.

Douvis, J. (2007). A review of attendance and non-attendance studies at sporting events. Biology of Exercise. 3. 10.4127/jbe.2007.3.5-20.

Douvis, J. (2014). What makes fans attend professional sporting events? A review. Advances in Sport Management Research Journal, Vol. 1, p. 40-70.

Google searches by sport (n.d.) Retrieved from https://fivethirtyeight.com/features/theres-a-big-five-in-north-american-pro-sports/.

Karakaya, F., Yannopoula, P., and Kafalaki, M., (2016), Factors impacting the decision to attend soccer games: an exploratory study, Sport, business and management: an international journal, (6) 3.

MLS designated players (n.d.) Retrieved from https://www.mlssoccer.com/glossary/ mdesignated-player.

MLS players association (n.d.) Retrieved from https://mlsplayers.org/resources/salary-guide.

MLSsoccer (n.d.), Retrieved from www.mlssoccer.com/stats/season.

Nishida, K. (2017) Retrieved from https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7.

Paul, R. and Weinbach, A., (2007) The uncertainty of outcome and scoring effects on Nielsen ratings for Monday night football, Journal of economics and business, 58(3) 199-211.

Quinlan, R. (1992). Learning with continuous classes. Proceedings of the Australian Joint Conference on Artificial Intelligence, 343–348. World Scientific, Singapore.

Raut, S. (2016) Want to know how to choose machine learning algorithm? Retrieved from http://customerthink.com/want-to-know-how-to-choose-machine-learning-algorithm/.

Sung, H. and Mills, B. (2017) Estimation of game-level attendance in major league soccer: Outcome uncertainty and absolute quality considerations, Sport management review, article in press.

Trawinski, B., Smetek, M., Telec, A., and Lasota, T., (2012) Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. International Journal of Applied Mathematics and Computer Science, 22(4) 867-891.

Weather underground (n.d.) Retrieved fromhttps://www.wunderground.com.

Weinbach, A. and Paul, T. (2013) Uncertainty of outcome and television ratings for the NHL and MLS, Prediction markets, 7(1) 53-65.

Welling, S.H., (2015) Retrieved fromhttps://stats.stackexchange.com/questions/162465/in-a-random-forest-is-larger-incmse-better-or-worse?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa.

Wikipedia (n.d.) Retrieved from https://en.wikipedia.org/wiki/Random_forest.

Zhu, N. and Chen, T. (2016) XGBoost: implementing the winningest Kaggle algorithm in Spark and Flink, retrieved fromhttps://www.kdnuggets.com/2016/03/xgboost-implementing-winningest-kaggle-algorithm-spark-flink.html.