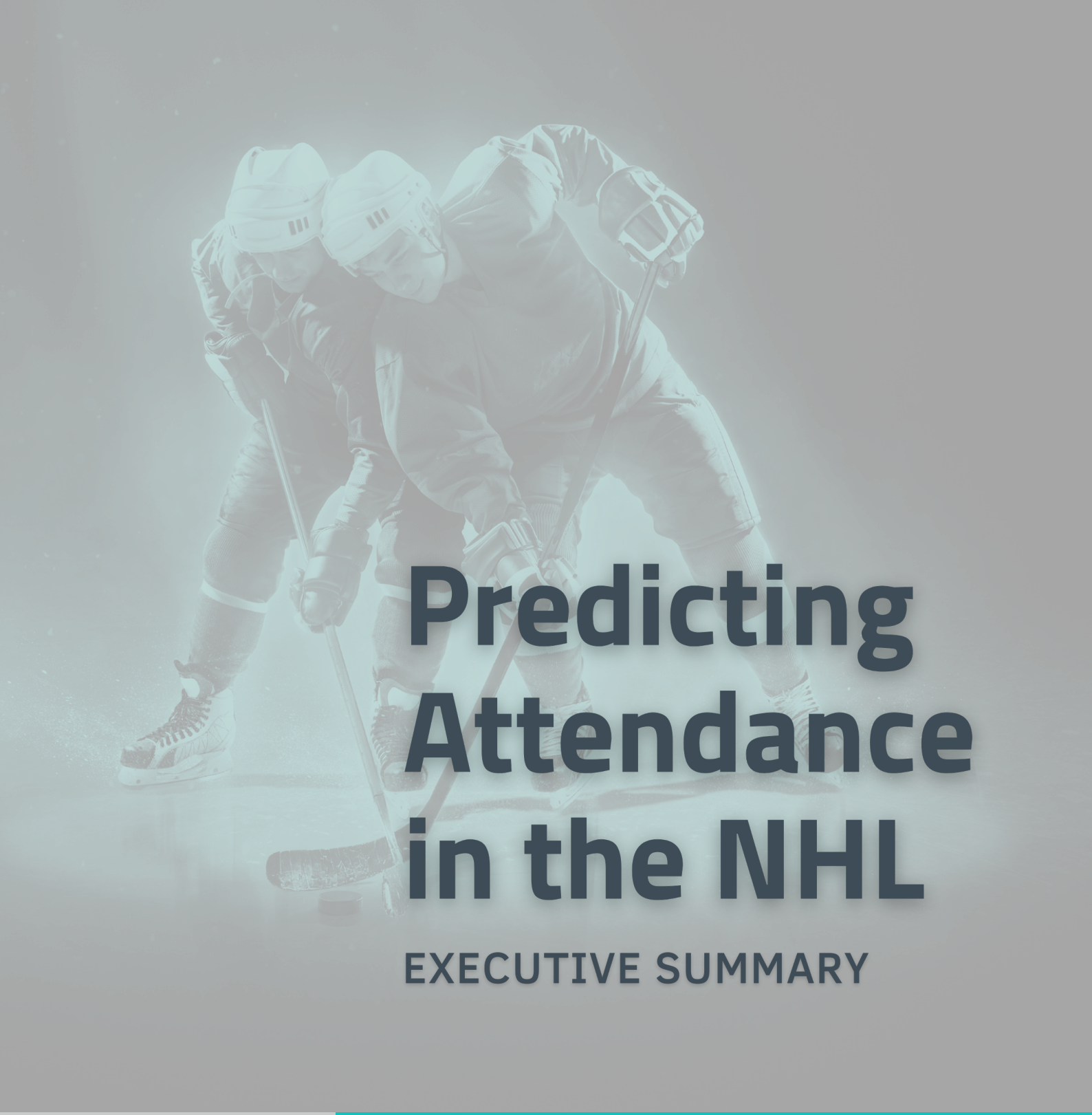

Predicting Attendance in the NHL

EXECUTIVE SUMMARY

AUGUST
2024

Sai Akhil Vudattu, Ayla Spitz, River Samudio,
Justin Doyle & Julie Athanasiadis

Northwestern University
MSDS Capstone Project

TABLE OF CONTENTS

INTRODUCTION

In the highly competitive landscape of professional sports, accurately predicting game attendance has become a critical component of operational and financial success for NHL teams. Attendance directly influences multiple facets of a team's business operations, from revenue generation to resource allocation and overall fan engagement. However, despite its importance, there currently exists no standardized model that NHL teams can universally adopt to forecast game attendance reliably. This lack of consistency leads to significant challenges, including operational inefficiencies, missed revenue opportunities, and fluctuating levels of fan satisfaction.

This project seeks to address these challenges by developing a robust, data-driven predictive model tailored to the unique demands of NHL games. By leveraging a comprehensive set of variables—including team performance, player performance, team rivalries, weather conditions, and a Covid-19 recovery indicator—the model aims to provide NHL teams with accurate and actionable attendance forecasts.

The overarching goal of this project is to empower NHL teams to leverage their attendance insights to optimize operations, enhance revenue streams, and deliver consistently positive fan experiences. This report outlines the business problem, objectives, methodology, key findings, and strategic recommendations, offering a detailed roadmap for integrating predictive analytics into the core business practices of NHL teams.

BUSINESS PROBLEM

While game attendance is a critical driver of an organization's financial success, NHL teams face challenges in accurately predicting game attendance due to the lack of a standardized predictive model. Inconsistent and inaccurate forecasts can pose significant problems for NHL organizations, such as operational inefficiencies, declining fan satisfaction, and missed revenue opportunities. These affect both the team's success on the rink and the organization's bottom line.

Operational inefficiencies occur when teams misallocate personnel resources due to overestimations or underestimations of attendance. In the case of overestimation, teams may overstaff, leading to unnecessary labor costs and wasted resources, as personnel are underutilized. On the other hand, underestimation can lead to understaffing, overwhelming critical areas such as concessions, security, and guest services. This not only impacts the bottom line but also results in inefficient use of resources that could be better allocated elsewhere.

The ripple effect of these operational inefficiencies extends beyond just financial losses; they directly contribute to a diminished fan experience. Long lines, inadequate services, and overcrowding erode fan satisfaction and impact season ticket renewals. Inaccurate attendance forecasts can also lead to ineffective marketing strategies, causing missed revenue opportunities. Teams struggle to implement dynamic pricing strategies that maximize sales revenue, and vendors of concessions and merchandise struggle with assortment planning.

Ultimately, the inability to accurately forecast attendance has far-reaching implications for NHL organizations. A standardized, accurate predictive model is essential to overcome the challenges outlined above, ensuring that organizations can optimize their operations, maximize revenue, and deliver a consistently high-quality fan experience. This will support both the short-term financial health and the long-term success of the organization.

BUSINESS OBJECTIVES

The primary objective of this project is to develop a standardized, data-driven predictive model that can be universally applied across all NHL teams to accurately forecast game attendance. By doing so, teams will have a reliable tool to inform their operational decisions, ensuring that stadium resources are aligned with expected attendance. This will not only optimize operational efficiency but also reduce unnecessary costs and enhance the overall fan experience.

In addition to operational optimization, accurate attendance predictions will improve the effectiveness of marketing strategies, enabling teams to tailor their promotions and ticket pricing to match expected demand. This will allow teams to maximize their return on marketing investments. Ultimately, the goal is to create a winning formula both on and off the rink, driving success through strategic use of data analytics.

METHODOLOGY

To develop a standardized, data-driven predictive model for accurately forecasting NHL game attendance, we adopted the CRISP-DM framework. This approach provides a structured and systematic methodology, ensuring that all aspects of the project are rigorously addressed. The key phases of CRISP-DM that guided our project are outlined below.

Business Understanding

As outlined above, the business understanding phase involved defining the primary business objective: to create a predictive model that improves game attendance forecasts across all NHL teams. This phase focused on understanding the challenges faced by teams due to inconsistent predictions and identifying key stakeholders.

Data Understanding

In this phase, we gathered a comprehensive set of data, including historical attendance records, player statistics, game characteristics, weather data, and data pertaining to the Covid-19 pandemic. The variables selected for analysis were informed by a thorough review of existing literature, as detailed in Appendix A. This review highlighted the importance of factors such as star players, rivalries, and team performance in influencing game attendance, providing a solid theoretical foundation for the data selected. We then conducted exploratory data analysis to understand the data's structure and identify patterns. This was the most time consuming aspect of the project.

Data Preparation

Based on the insights gained from the data understanding phase, we selected relevant data points that directly influence game attendance, such as previous season success, current winning percentage, averages goals scored, and age of team. We cleaned the data by handling missing values and correcting inconsistencies. We then integrated data from various sources to create a comprehensive dataset for analysis. New features were also created to enhance the predictive power of the model, including variables like lagged attendance variables, home game streak, order of day of the week, and shots per game.

METHODOLOGY

Modeling

We explored various modeling techniques in the literature review, including multiple linear regression, random forests, and gradient boosting machines, to determine the most effective approach for predicting game attendance. In the end, we selected a Gradient Boosting model. A rigorous test design was implemented to evaluate the performance of different models, using a combination of training and testing datasets to validate accuracy and help understand if the model would work well with unseen data. This step helps identify any issues such as overfitting that we may need to address with regularization, which was ultimately added in the best model with L1 and L2 regularization.

Evaluation

The models were evaluated based on their accuracy in predicting attendance. We used key performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared to assess model performance. Our best-performing gradient boosting model utilized RandomizedSearchCV combined with 5-fold cross-validation. The decision to use RandomizedSearchCV instead of GridSearchCV significantly reduced processing times, which enabled us to upgrade from 3-fold to 5-fold cross-validation. Switching to 5-fold cross-validation offered several benefits, such as producing a more robust model by reducing variability and better utilizing the data.

Deployment

With the predictive model now developed and validated, the next critical phase is deployment. Our deployment plan involves leveraging Google Cloud Platform (GCP) to host and manage the model. We will take a structured approach to ensure that the model is effectively integrated into NHL teams' operations, maximizing its impact on attendance forecasting.

FINDINGS AND OUTCOMES

After evaluating different models, we found the Gradient Boosting model to be the most accurate in predicting NHL game attendance. This model uses a 5-fold cross-validation which tests the model on different subsets of data to ensure reliability. It also uses L1/L2 regularization to avoid overfitting, which makes sure the model can handle different types of games and scenarios. This model achieved the highest accuracy, closely matching real attendance figures, with very low error rates compared to other models. This is demonstrated by our high R^2 score, which shows that our model explains 97.2% of the variance in the data. In addition, despite some large residuals, or the differences between our predictions and actual values, our predictions averaged being off by only 283 seats. In arenas that hold between 15,000 and 20,000 seats, this number is insignificant and is extremely useful for planning purposes. However, over time, improving on this number would lead to better decision-making, increase revenue opportunities, and improve planning optimization.

Besides being highly accurate, this top-performing model is also efficient, with a run time of just 1.2 minutes. This quick processing time is essential for real-time predictions. Teams can get timely insights into expected attendance, enabling them to adjust staffing levels, inventory, and marketing strategies on a game-by-game basis. See Appendix B for a comparative analysis of model performance.

Our analysis showed that using techniques like cross-validation and regularization significantly improved the model's performance. These fine-tuning techniques helped us adjust the model to be more accurate without overfitting the data. Initially, we determined the key features to use in our model from utilizing feature importance from the random forest regressor model. We chose a set of the 20 highest factors that contribute to the attendance percentage variable. Despite success with this approach in some of the models, it was determined that the model with XGBoost automatic feature selection provided the best balance between accuracy and speed, making it a practical choice for deployment. Interestingly, our feature importance model considered a simple rating system (an advanced NHL metric measuring team's goal differential and strength of schedule), attendance moving average (a field created by us to measure recent historical attendance trends), and the average age of the team as the most important features regarding our target variable of attendance percentage. The gradient boosting model automatic feature selection found that rival games, made playoffs previous year, and previous season rank variables as the top three most influential features. We list the top 20 features from both feature selection techniques in Appendix C.

INSIGHTS AND RECOMMENDATIONS

The development of our standardized predictive model marks a significant advancement in accurately forecasting NHL game attendance. Our findings highlight the immense value that advanced predictive models bring to the table. By incorporating a range of influential factors, our model achieves high levels of accuracy and reliability. This predictive power is crucial for transforming the way NHL teams approach their game-day operations and strategic planning.

The development and maintenance of a comprehensive and advanced predictive model internally would require significant time, expertise, and financial investment, which not all teams may have readily available. By implementing our advanced predictive model, NHL teams can benefit from accurate, data-driven attendance forecasts without the need to build and sustain a model from scratch. This standardized solution provides consistency and reliability, allowing teams to focus on strategic decision-making rather than the complexities of model development.

Moving forward, we recommend a phased deployment approach, starting with pilot programs to integrate the model into team operations. The GCP is a cloud native deployment tool that is compatible with a wide range of applications and data integration needs, making it the logical choice for model deployment. We will utilize stakeholder feedback on product's performance to inform updates and modifications. Continuous monitoring and regular updates are essential to maintain accuracy and adapt to changes.

CONCLUSION

In conclusion, the development of our advanced predictive model represents a pivotal step forward for NHL teams seeking to enhance their game attendance forecasting and address the challenges of inconsistent predictions. Our model, which is easily adaptable and can be seamlessly integrated into any team in the NHL, allows teams to leverage its predictive capabilities regardless of their unique operational contexts or resource levels. Our model will enable teams to optimize their operations, maximize revenue, and deliver exceptional fan experiences, improving not only the short-term financial health of the organization, but also supporting its long-term success. As we move forward into the deployment phase, we will continue to focus on model improvement, ensuring that this tool remains at the cutting edge of predictive analytics, driving value for teams, fans, and the league as a whole.

ABOUT US

SAI VUDATTU | Technical Lead

Responsible for modeling approach and data management.

AYLA SPITZ | Deliverable Lead

Responsible for reviewing project deliverables and preparing initial documents.

RIVER SAMUDIO | Floating Lead

Responsible for preparation of reports and review of technical deliverables.

JUSTIN DOYLE | Technical Lead

Responsible for data collection and data preparation.

JULIE ATHANASIADIS | Project Manager

Responsible for submission of deliverables and organizing meetings. Contributes to write-ups.

APPENDIX A: LITERATURE REVIEW

Introduction

The purpose of this literature review is to explore and synthesize existing research related to predicting professional sports game attendance, with a specific focus on the National Hockey League (NHL). This review will examine various factors influencing attendance, methodologies used in prediction models, and the impact of accurate predictions on operational efficiency, fan engagement, and revenue generation. Understanding the current state of research in these areas will provide a foundation for developing a robust predictive model tailored to NHL games.

Overview of Professional Sports Game Attendance

Over the past half decade, roughly one-third of Americans have attended a sporting event. This is even despite a global pandemic that temporarily crippled the live entertainment industry. As sports leagues reopened their stadiums to fans, many professional leagues saw varying levels of returns of fans. The NFL and NBA have seen higher than pre-pandemic attendance numbers, while the MLB has seen attendance return to normal. However, the smaller leagues like the NHL and MLS has not seen fans return in numbers similar to before (Broughton 2023). Our focus on the NHL provides a unique opportunity to look into how to model, predict, and make suggestions to grow the professional sport. Currently, the NHL has the lowest attendance numbers of all five major professional sports leagues in the United States (McCready 2023). The NHL particularly has teams who are susceptible to attendance growth opportunities relative to other leagues. There are roughly 7 NHL teams with an average attendance of less than 13,000. There is only a single MLS team and two MLB teams with attendance numbers that low.

Factors Influencing Sports Attendance

Scoring Rates

High-scoring games are often perceived as more exciting and entertaining, potentially attracting larger crowds. Paul, Weinbach, and Robbins (2004) explore the relationship between scoring rates and attendance, finding that games with higher scoring tend to draw more fans. Specifically, the study notes that a one-goal increase in the average goals per game is associated with an attendance boost of approximately 1-2%. This is consistent with the broader literature which suggests that fans prefer games with more action and excitement, as these elements enhance the overall spectator experience (Brunt and Batey 2003; Buraimo, Forrest, and Simmons 2010).

APPENDIX A: LITERATURE REVIEW CONT.

Regional Rivalries
Regional rivalries are a significant draw for fans, often resulting in higher attendance figures due to the intense competition and heightened fan interest these games generate. The study by Paul, Weinbach, and Robbins (2004) confirms that games featuring traditional rivals see higher attendance numbers, attributing this to the emotional and cultural significance of these matchups. Their analysis indicates that attendance at rivalry games can be as much as 15-20% higher than non-rivalry games, highlighting the strong appeal of these matchups for fans.

Team Performance
Schreyer and Ansari (2020) emphasize the importance of team performance on game attendance. Winning teams generally attract more fans, as they offer a higher likelihood of a satisfying game experience. Teams that are performing well in the league or are on winning streaks tend to see higher attendance rates. For example, a study by Coates and Humphreys (2012) found that a one-standard deviation increase in a team's winning percentage could increase attendance by approximately 10% (Schreyer and Ansari 2020; Coates and Humphreys 2012).

Economic Conditions
Economic factors such as local unemployment rates, disposable income, and overall economic health of the region can impact fans' ability to purchase tickets. Schreyer and Ansari (2020) note that during economic downturns, discretionary spending on entertainment such as sports games can decrease, leading to lower attendance figures (Schreyer and Ansari 2020).

Stadium Accessibility
The location, accessibility, and quality of amenities in the stadium, including concessions, parking, and public transportation options, play a significant role in attracting fans. Schreyer and Ansari (2020) found that fans are more likely to attend games in stadiums that offer better facilities and are easier to access. A study by Shapiro et al. (2009) highlighted that improved amenities and accessibility could increase attendance by 5-7% (Schreyer and Ansari 2020; Shapiro et al. 2009).

Predictive Modeling Techniques
Predictive modeling is a crucial component when it comes to forecasting NHL game attendance. By leveraging historical data and advanced analytical techniques, teams can anticipate attendance figures more accurately, optimizing operations, marketing strategies, and resource allocation. This section reviews various predictive modeling techniques commonly employed in sports analytics, emphasizing their application in predicting game attendance.

APPENDIX A: LITERATURE REVIEW CONT.

Predictive Modeling Techniques

Predictive modeling is a crucial component when it comes to forecasting NHL game attendance. By leveraging historical data and advanced analytical techniques, teams can anticipate attendance figures more accurately, optimizing operations, marketing strategies, and resource allocation. This section reviews various predictive modeling techniques commonly employed in sports analytics, emphasizing their application in predicting game attendance.

Multiple Linear Regression

Multiple linear regression is a commonly used technique in statistical modeling, due to its simplicity and effectiveness. For predicting NHL game attendance, multiple linear regression can incorporate various predictors, such as team performance metrics, player statistics, day of the week, and economic conditions, and stadium conditions to estimate attendance figures. A study by Rascher (1999) highlights the successful application of multiple linear regression in sports to predict game attendance by analyzing how different various factors impact the outcome. By quantifying the effect of each predictor, teams can better understand which factors are meaningful in driving attendance. Furthermore, teams can ascertain highly accurate attendance predictions, allowing them to increase their operational efficiency and eliminate unnecessary costs.

Random Forest

The Random Forest technique can be well-suited for handling large datasets with many predictor variables, making it ideal for modeling NHL game attendance. Random Forest has been successfully used to predict game attendance. For instance, Gu, Chintagunta, and Hahsler (2015) used Random Forest models to predict Major League Baseball (MLB) attendance, demonstrating its effectiveness in handling a wide array of variables, including team performance, weather conditions, and promotional events. Their study found that the model could explain up to 70% of the variance in attendance, underscoring the robustness of the Random Forest model.

Gradient Boosting Machines

Gradient Boosting Machines (GBMs) have been used extensively in sports analytics to predict various outcomes. For example, Katerina and Eleni (2018) applied Gradient Boosting techniques to predict football match outcomes, illustrating the method's effectiveness in processing diverse sources of data. Their study utilized player statistics, team performance metrics, and historical match outcomes to train the model. The GBM achieved a prediction accuracy of over 75%, outperforming traditional logistic regression models. This demonstrates the potential of GBMs in accurately forecasting events in which various interacting factors must be considered.

APPENDIX A: LITERATURE REVIEW CONT.

Impact of Accurate Attendance Predictions

Attendance affects the bottom line of professional sports team in a variety of ways (Broughton 2023). First, it affects ticket revenue, which is the most direct way teams see a financial impact of people showing up or not. Then, there is an impact on items sold in stadium. This includes food, beverages, and merchandise sold in the team store. Next, there is an impact on marketing revenue generated at the stadiums. Teams usually feature sponsored events, ads placed throughout the venue, etc. which are more valuable the more people can see them. Then there is a impact on the businesses surround the stadiums. Often, some teams have certain deals with businesses that exist directly next to the venue (Schreyer and Ansari 2020) since often the teams own the surrounding property. Finally, we also consider that good attendance on games have impacts on specific businesses like hotels. If people are more willing to show up, travel in to the cities, which also include attendance from other fandoms, then many may choose to stay at a hotel, which teams also do some revenue sharing with.

With an accurate prediction model, organizations and leagues benefit in a few ways. In the immediate future, teams are better able to control costs. If they know certain games or seasons are likely to have more or less fans in attendance than normal, then they can match staffing to those needs. In a slightly longer time horizon, teams can make systematic changes to increase attendance, changes to the venue, additional promotional events to bring people in, lobbying the league for certain beneficial schedules specific to their city (i.e. a hockey team in Arizona might actually do well for attendance in a mid-day summer weekend game since people want out of the sun into an air conditioned facility).

There have been a few others who have studied and created similar models of attendance performance and their impact. McDonald studied the impact on corporate sponsors and hospitality businesses (McDonald 2010). Schmidt looked at benefits that affected businesses tangential to the stadium (Schmidt 2012). We add to this literature and use ideas applied to the analysis of bigger American sports and focus on an understudied league who is in desperate need of fan engagement.

APPENDIX A: LITERATURE REVIEW CONT.

References

Broughton, David. “Attendance Evolution since 2003.” Sports Business Journal, April 10, 2023. https://www.sportsbusinessjournal.com/Journal/Issues/2023/04/10/In-Depth/attendance.aspx.

Brunt, Paul, and Richard Batey. "Scoring Rates and Sports Attendance: Evidence from the UK." Journal of Sports Management 17, no. 3 (2003): 231-246.

Buraimo, Babatunde, David Forrest, and Robert Simmons. "The Twelfth Man? Refereeing Bias in English and German Soccer." Journal of the Royal Statistical Society: Series A (Statistics in Society) 173, no. 2 (2010): 431-449.

Coates, Dennis, and Brad R. Humphreys. "The effect of professional sports on earnings and employment in the services and retail sectors in US cities." Regional Science and Urban Economics 33, no. 2 (2003): 175-198.

Coates, Dennis, and Brad R. Humphreys. "Game Attendance and Outcome Uncertainty in the National Hockey League." Journal of Sports Economics 13, no. 4 (2012): 364-377.

Gu, Yongyi, Pradeep K. Chintagunta, and Michael Hahsler. "Customer Valuation in a Bundle Pricing Context: The Case of Major League Baseball Ticket Pricing and Attendance." Marketing Science 34, no. 5 (2015): 736-755.

Katerina, A., and Eleni, M. "Predicting Football Match Outcomes with Gradient Boosting." Journal of Sports Analytics 4, no. 2 (2018): 123-134.

McCready, Bo. “Average Attendance for North American Sports Leagues.” Tableau, February2023.https://public.tableau.com/app/profile/bo.mccready8742/viz/AverageAttendanceforNorthAmericanSportsLeagues2022/AverageAttendance.

McDonald, Heath. “The Factors Influencing Churn Rates among Season Ticket Holders: An Empirical Analysis.” Journal of Sport Management 24, no. 6 (November 2010): 676–701. https://doi.org/10.1123/jsm.24.6.676.

Paul, Rodney J., Andrew P. Weinbach, and Daniel Robbins. "Variations in NHL Attendance: The Impact of Violence, Scoring, and Regional Rivalries." Journal of Sports Economics 5, no. 1 (2004): 124-138.

Rascher, Daniel A. "A Model of Stadia Demand." Journal of Sport Management 13, no. 2 (1999): 161-177.

Schreyer, Dominik, and Payam Ansari. "Stadium Attendance Demand Research: A Scoping Review." Sport Management Review 23, no. 2 (2020): 253-271.

Schmidt, Martin B. “Demand, Attendance, and Censoring.” The Oxford Handbook of Sports Economics, September 18, 2012, 190–200. https://doi.org/10.1093/oxfordhb/9780195387780.013.0011.

Shapiro, Stephen L., Matthew D. Shapiro, and Lisa P. Dwyer. "Analyzing the factors that influence college basketball attendance: The case of the Big East." Sport Marketing Quarterly 18, no. 3 (2009): 160-170.

APPENDIX B: MODEL PERFORMANCE METRICS

Experiment	Model Type	Key Parameters	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R^2	Run Time
1	Long Short-Term Model (LSTM)	2 Layers with 50 Neurons 1 Dense Dayer with a Single Output Features Determined before model with Random Forest Regressor Feature Importance	1564.140	5495240.153	2344.193	0.567	1 min
2	Gradient Boosting Model	Features Determined before model with Random Forest Regressor Feature Importance	559.583	862440.111	928.677	0.930	< 1 min
3	Gradient Boosting Model	GridSearchCV 3-fold Cross Validation Features Determined before model with Random Forest Regressor Feature Importance	447.291	634676.770	796.666	0.956	8 min
4	Gradient Boosting Model	RandomizedSearchCV 3-fold Cross Validation L1/L2 Regularization Features Determined before model with Random Forest Regressor Feature Importance	457.483	619720.391	787.223	0.957	1 min
5	Gradient Boosting Model	RandomizedSearchCV 5-fold Cross Validation L1/L2 Regularization XGBoost automatic feature selection.	283.217	240036.791	489.935	0.972	1.2 min

Figure B-1: Comparative Analysis of Model Performance

APPENDIX B: MODEL PERFORMANCE METRICS

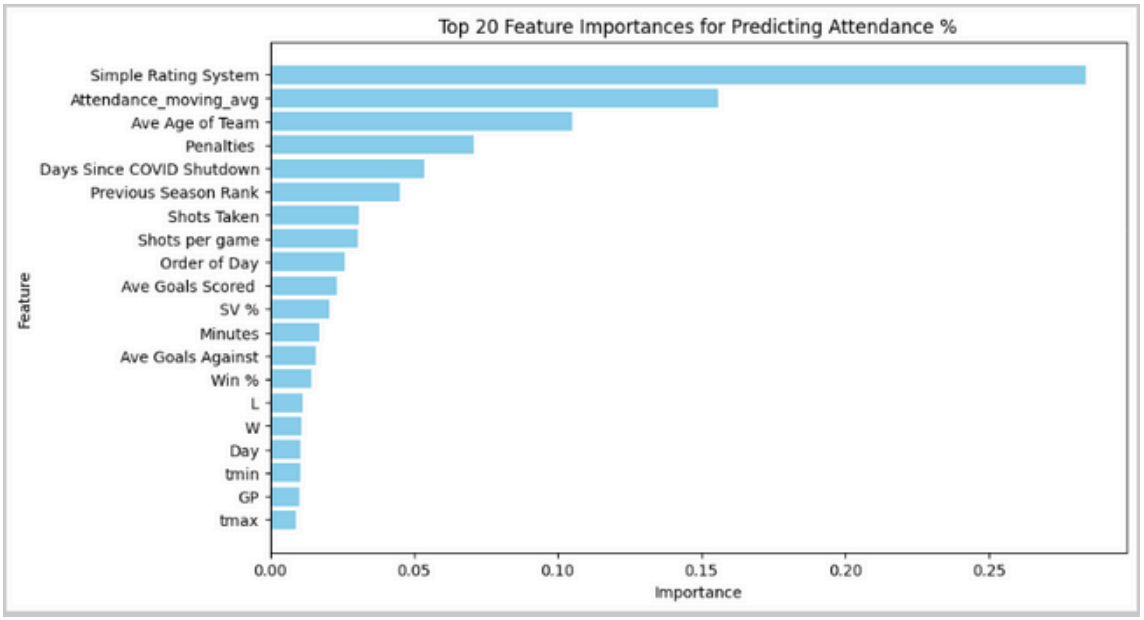

Figure B-2: Random Forest Regressor Feature Importance for Attendance Percentage

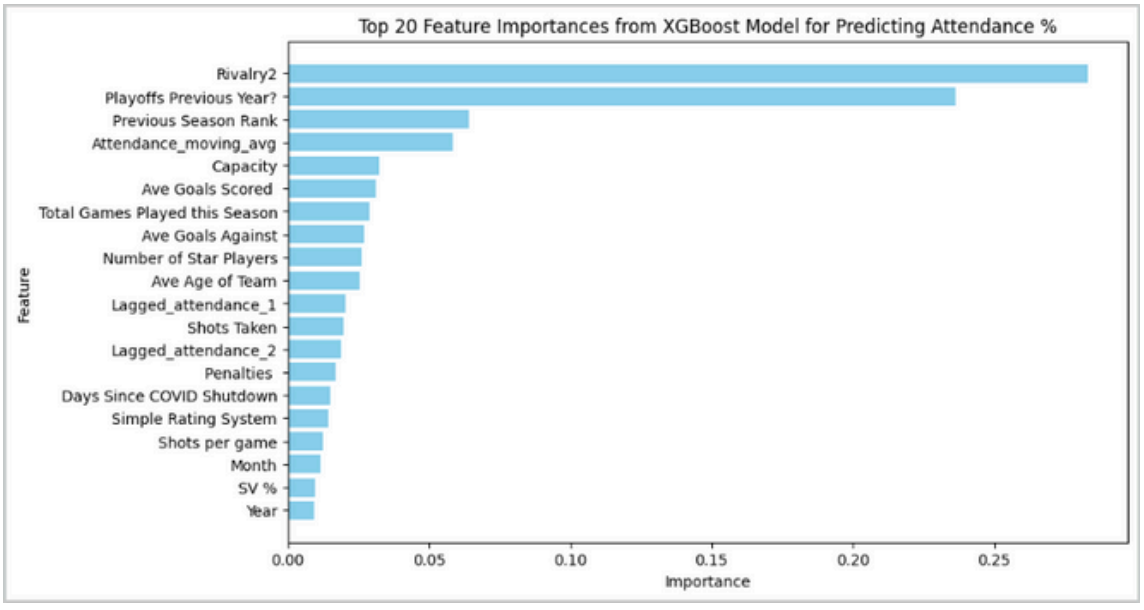

Figure B-3: XGBoost Model Built-in Feature Selection for Attendance Percentage Importance

Contact our team with any questions.

Northwestern University
MSDS Capstone Project