PREDICTING ATTENDANCE IN THE NHL

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

Northwestern University, MSDS Capstone Project

August 24, 2024

AGENDA

1 BUSINESS PROBLEM

2 BUSINESS OBJECTIVES

3 LITERATURE REVIEW

4 MODELING APPROACH

5 OUTCOMES

6 INSIGHTS AND RECOMMENDATIONS

7 LIMITATIONS AND FUTURE WORK

8 MEET THE TEAM

BUSINESS PROBLEM

INCONSISTENT ATTENDANCE PREDICTIONS:

NHL teams lack a standardized model for predicting game attendance, leading to inconsistent and often inaccurate forecasts

FAN SATISFACTION

- Long-term loyalty may diminish if fans frequently encounter poor experiences, such as long waits or inadequate services
- Repeated negative experiences can harm a team's brand, making it harder to attract and retain fans

REVENUE LOSSES

- Inaccurate forecasts impact marketing efforts → missed opportunities for optimizing ticket sales, concessions, and merchandise

OPERATIONAL INEFFICIENCIES

- Poor attendance predictions result in overstaffing or understaffing of stadium personnel
- Can unnecessarily increase operational costs and negatively impact fan experience

BUSINESS OBJECTIVES

DEVELOP A STANDARD PREDICTIVE MODEL: Create a universally applicable, data-driven model to accurately predict game attendance across all NHL teams

OPTIMIZE OPERATIONS: Use the predictive model to align staffing levels, concession supplies, and other resources with expected attendance

ENHANCE MARKETING EFFICIENCY: Improve marketing strategies and promotions by leveraging accurate attendance forecasts, maximizing ROI on marketing spend

INCREASE FAN ENGAGEMENT AND LOYALTY: Ensure positive fan experiences by avoiding understaffing, thereby boosting long-term fan loyalty and engagement

LITERATURE REVIEW

KEY PREDICTIVE VARIABLES IN ATTENDANCE MODELING

TEAM PERFORMANCE

- Teams that perform well, particularly those on winning streaks, attract more fans
- A one-standard deviation increase in a team's winning percentage can increase attendance by ~10% (Schreyer and Ansari 2020, Coates and Humphreys 2012)

STAR PLAYERS

- The presence of star players significantly drives higher attendance
- Games featuring star players saw an increase in attendance by ~5-10%, particularly when those players were visiting (Paul, Winbach, and Robbins 2014)

RIVALRIES

- Games featuring traditional rivalries can see attendance numbers as much as 15-20% higher than non-rivalry games (Paul, Weinbach, and Robbins 2004)
- Rivalry games have strong emotional and cultural importance for fans

LITERATURE REVIEW

BROADER CONTEXT AND BUSINESS IMPLICATIONS

Similar Studies

- Mueller predicted pre and in-season MLB attendance with a random forest model (2020). Bowley and Berger predicted NFL attendance with a regression analysis (2017).
- Some researchers have studied the impact attendance has on corporate sponsors and the hospitality industry in the city (McDonald 2010). Others have examined the impact on businesses in the immediate surrounding area of the stadium (Schmidt 2012).

Other Leagues

- Since the onset of COVID, 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 have not seen fans return in numbers similar to before (Broughton 2023).

The Bottom Line

- Higher attendance boosts sales of items sold in the stadium, such as food, beverages, and merchandise. Marketing revenue generated at stadiums is also affected, as sponsored events and advertisements become more valuable with larger audiences. Moreover, attendance influences the businesses surrounding stadiums, often due to agreements teams have with nearby establishments (Schreyer and Ansari 2020).

MODELING APPROACH

BUSINESS UNDERSTANDING

CRISP-DM

DATA UNDERSTANDING

- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality

DATA PREPARATION

- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

MODELING

- Select Modeling Technique
- Generate Test Design
- Build Model Parameter Settings
- Assess Model

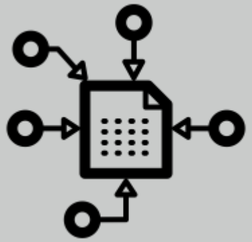
EVALUATION

- Evaluate Model
- Approved Models
- Determine Next Steps

DEPLOYMENT

- Plan Deployment
- Plan Monitoring and Maintenance
- Produce Final Report
- Review Project

DATA UNDERSTANDING



Collect Data

We leveraged data from a variety of different sources, which includes NHL attendance data from a hockey reference service, temperature data for each county in the US by a UC Davis program, star player data from All-Star team selection from the official NHL source. We use a variety of other sources to build out our remaining predictive variables, as well. In addition, we use other professional league reference sites to help us understand the impacts of COVID, too.

Describe/Verify Data

Our data, given recency, is fairly complete and requires little cleaning for missing values or conversations. We did complete some manual data creation to bring data sources together for the model, such as county assignment and COVID start/days since pandemic shut-down.

Explore Data

We use Python to explore relationships and the structure of our current dataset. We use typical EDA approaches to analyze items such as average attendance by team, attendance by day of week and month, and including feature importance of variables such as win%, winstreak, time of games, opponent, etc.

DATA PREPARATION

SELECT DATA

- NHL Game and Related Data from Seasons 2018-2024

CLEAN DATA

- Corrected errors in capacity data (i.e. games that were held in other arenas)
- Removed duplicates

CONSTRUCT DATA

- Created Order of Month and Order of Day to determine if later months in season or if later days in the week have an effect
- Created ave. shots per game
- Created lagged attendance and average moving attendance fields

INTEGRATE DATA

- Merged all datasets prior to Colab upload and model configuration
- Created unique identifiers on each data set and utilized vlookup function in excel.
- Master Data Set includes 5 unique IDs, 7141 rows of data, and 52 data points.

FORMAT DATA

- Data type conversion for correlation matrix and other aspects of model
- Time conversion for a min. after midnight data field
- Dropped or used median for missing and N/A data depending on need
- Feature selection and scaling
- Normalized data

DATA PREPARATION

TEAM PERFORMANCE

- Wins
- Losses
- Overtime Losses
- Winning %
- Previous Season Rank
- Made Playoffs Previous Season Indicator
- Ave. Goals Per Game
- Ave. Goals Against Per Game
- Ave. Shots Per Game
- Save Percentage
- Ave. Penalties per Game
- Simple Rating System Rank (SRS)
- Number of Star Players

SELECTED DATA FIELDS

GAME CONDITIONS

- Time of Day
- Day of Week
- Month
- Order of Month in Season
- Order of Day in Week
- Arena Capacity
- Lagged Attendance
- Ave. Moving Attendance
- Bobblehead Giveaway
- Games Played

OUTSIDE FACTORS

- Temperature
- Days since COVID
- NFL Game Nearby
- Rivalries
- Home Game Streak
- Ave. Age of Team

MODELING

SELECT MODELING TECHNIQUE

- Regression model to prediction attendance percentage. Needed to use percentage instead of actual attendance because every arena has a different capacity.
- Compared two models:
  - Long-Short Term Memory (LSTM) Model (Deep Neural Net Model)
  - Gradient Boosting Model

GENERATE TEST DESIGN

- LSTM model designed with a training set (80% of the data) and a testing set (20% of data). The training set is further divided into training set (80% of the training data) and validation (20% of the training data).
- Gradient models varied with no validation set, 3-fold cross validation, and 5-fold validation.
- Both modeling techniques used Mean Square Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 to analyze the performance of the model.

MODELING

BUILDING MODEL PARAMETERS

LONG SHORT-TERM MODEL

Model: LSTM Sequential
Layers: 2 LSTM, 1 Dense
Neurons: 50 in Each LSTM Layer, 1 in Dense Layer
Sequence Length = 10
Feature Scaling: MinMax
Optimizer: Adam with learning rate .001
Batch Size: 32
Number of epochs: 50
Loss Function: Mean Squared Error (MSE)
Regularization: None applied

GRADIENT BOOSTING MODEL

XGBoost Hyperparameters
Feature Scaling: MinMax
n_estimators: [100, 200, 300]
learning_rate: [0.01, 0.1, 0.2]
max_depth: [3, 5, 7]
min_child_weight: [1, 3, 5]
subsample: [0.8, 1.0]
colsample_bytree: [0.8, 1.0]
Regularization: L1 [0.01, 0.1] and L2 [1.0, 2.0]

RandomizedSearchCV parameters
estimator: model
param_distributions: param_dist
n_iter: 20 (# of random combinations to test)
scoring: MSE
Cross validation: 5 (number of folds)
verbose: 2
n_jobs: -1
random_state: 42

MODELING

ASSESSING MODELS

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

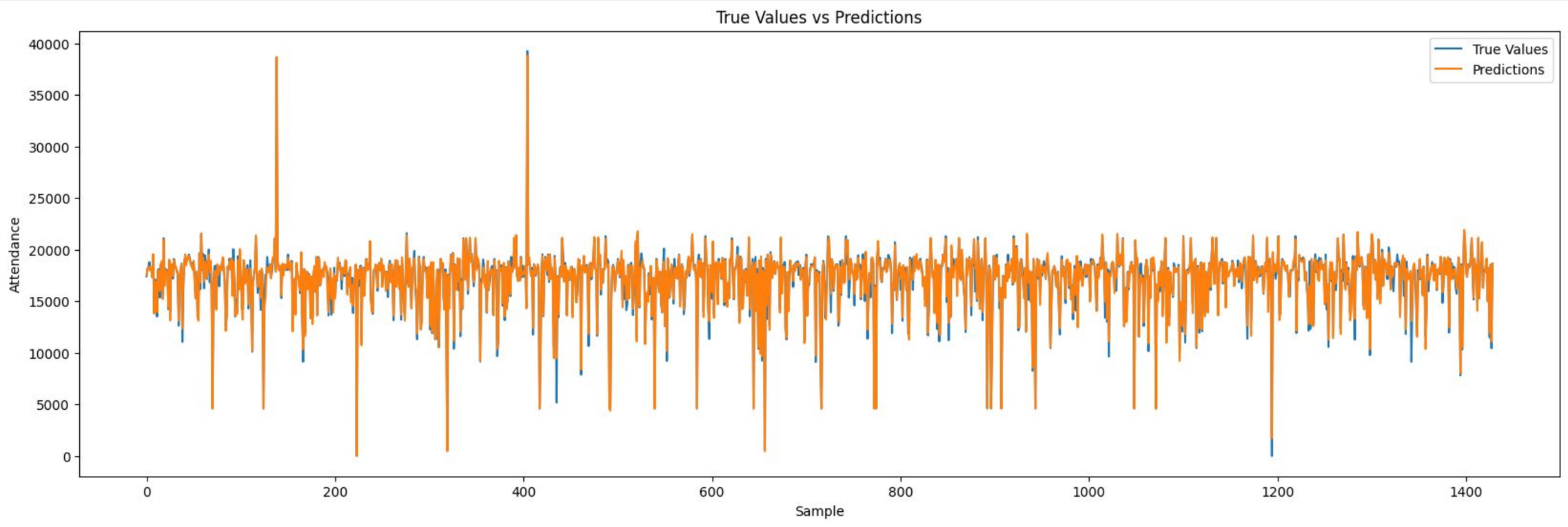
EVALUATION

Best Performing Model Key Parameters

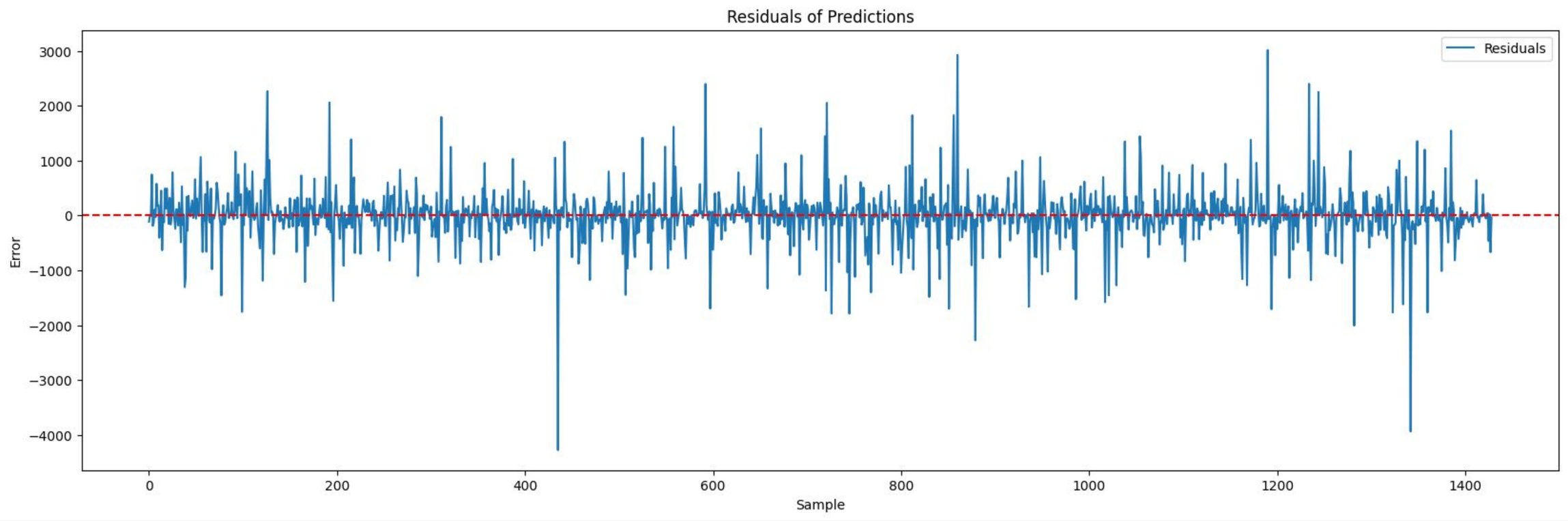
Gradient Boosting Model
RandomizedSearchCV
5-fold Cross Validation
L1&L2 Regularization
XGBoost built-in automatic feature selection

MAE	MSE	RMSE	R^2
283.22	240036.80	489.94	0.972
On average, the model's predictions are off by 283 seats when predicting attendance percentage	While most predictions might be close to the actual attendance values, a few larger errors cause the overall MSE to be high	The standard deviation of error is ~ 490 seats. A higher RMSE compared to MAE suggests large errors in the predictions	Our model explains 97.2% of the data's variance which suggests highly accurate predictions on attendance percentage

EVALUATION

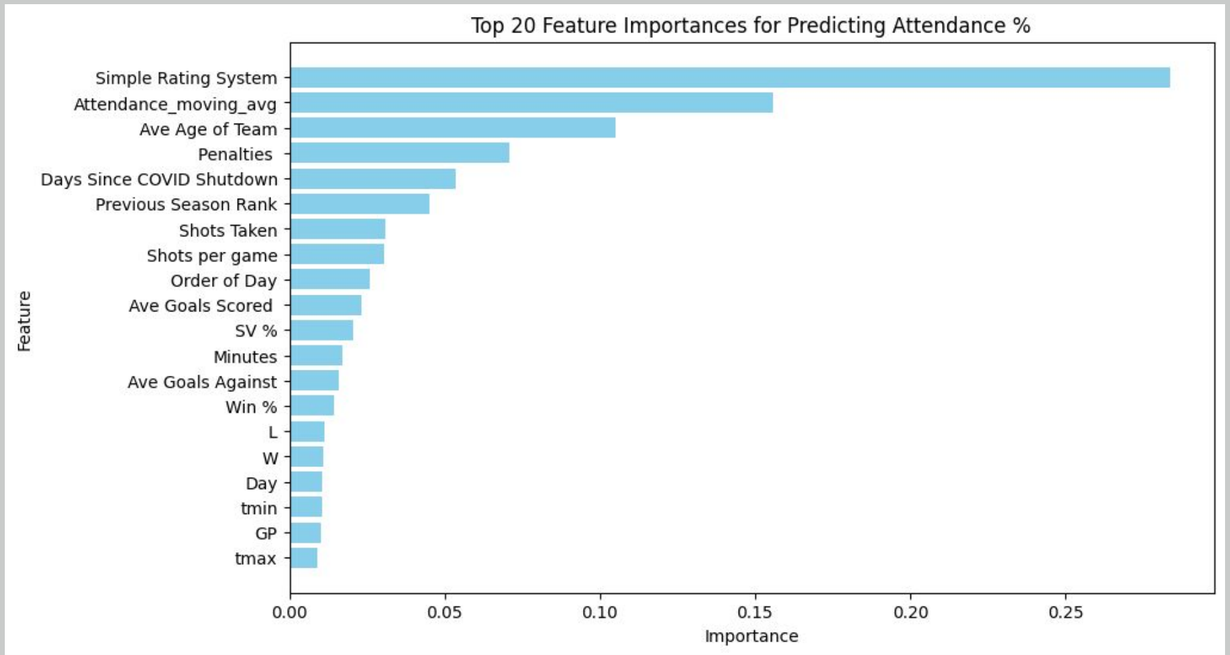

True Values vs Predictions
True Values
Predictions
Attendance
Sample
40000
35000
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200
1400

EVALUATION


Residuals of Predictions
Residuals
Error
Sample
3000
2000
1000
0
−1000
−2000
−3000
−4000
0
200
400
600
800
1000
1200
1400

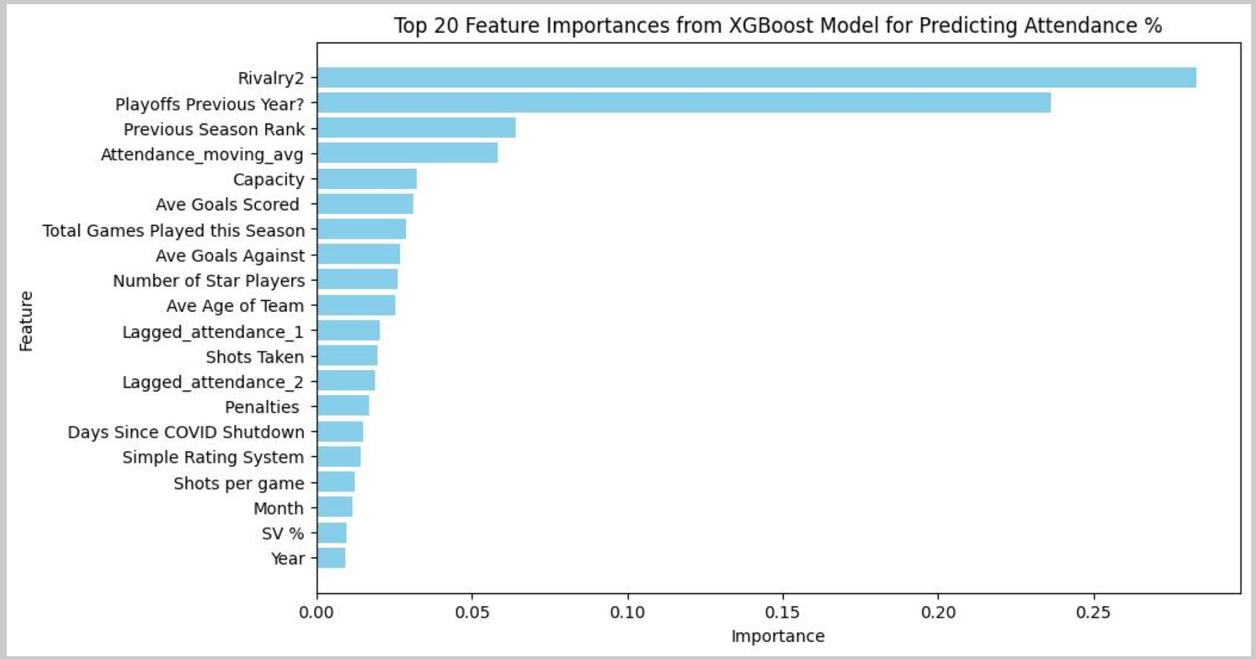
EVALUATION

Random Forest Regressor Feature Importance for Attendance %



Removed unsold seats, attendance, capacity, and lagged attendance variables.

XGBoost Model Built-in Feature Selection Feature Importance for Attendance %



Removed unsold seats and attendance variables.

DEPLOYMENT

Deployment Plan

- Our deployment plan includes leveraging Google Cloud Console(GCP).
- Upload our Machine Learning Model and integrate further data or incorporate realtime data stream.
- Develop CI/CD workflow and create pipelines via GitHub for future updates and testing.

Plan Monitoring and Maintenance

- The GCP is a outstanding cloud native deployment tool and help the clients to incorporate a wide range of applications
- Enables wide range of API Like BigQuery and cloud Build, cloud Monitor and Dashboard for their products, especially for monitoring and Billing issues and creating a dashboard.
- The GCP has an built-in API for cloud monitoring to monitor and maintain the product according to the client needs.

Final Report

- After deploying the model into the GCP and monitoring it for future improvement and error handling, a report containing all the data with a summary containing the project objectives, models used, key findings, evaluations, and the final results of the model development and deployment is written.

Review

- Stakeholder Feedback on the performance of the initial product and making further updates based on there feedback.
- Evaluate whether it meets the established KPIs and delivers the expected value.
- Detail the steps for formally closing the project, including handing over documentation, training

OUTCOMES AND FINDINGS

Best Prediction

Gradient Boosting model to be the most accurate in predicting NHL game attendance.

Run Time

The model performance was very fast and was completed in just 1.2 minutes.

Increased Performance

The model performed better after incorporating 5 fold cross validation and L1/L2 regularization methodologies into it.

Best for Deployment

Model with XGBoost automatic feature selection provided the best balance between accuracy and speed, making it a practical choice for deployment.

INSIGHTS AND RECOMMENDATIONS

OPTIMIZING MODEL SELECTION

Focus on deploying the Gradient Boosting Model with 5-Fold CV and L1/L2 regularization, which utilized XGBoost for feature selection. This model demonstrated the best predictive performance with a favorable runtime, making it the top candidate for implementation and will be used in our current deployment.

BALANCING ACCURACY AND RUNTIME

While hyperparameter optimization can dramatically improve model accuracy, it's crucial to balance these gains with potential increases in runtime. Over-optimization risks overfitting, especially with large feature sets.

LEVERAGING CROSS-VALIDATION AND REGULARIZATION

Applying cross-validation and regularization techniques significantly enhanced model performance, particularly in GB models, demonstrating the importance of these methods in refining predictive accuracy.

FUTURE FOCUS: HYPERPARAMETER TUNING

Further hyperparameter tuning is recommended for fully optimized model performance, keeping in mind the risks of overfitting. In doing this, we are ensuring that enhancements contribute meaningfully toward achieving the project's goals of accurately predicting NHL attendance.

LIMITATIONS

Relatively Small Dataset

Moving forward, we would look to expand both in features and in observations. Despite this, the accuracy of the models did not suffer.

Gaps in the Dataset

Some seasons varied in consistency and data quality. For example, 2020-21 did not include every arena and 2019-20 was cut short due to COVID. This was accounted for in the EDA and cleaning process, but still a hurdle nonetheless. Additionally, we had to make some assumptions on capacity due to the differences in arena organization/policy with regards to standing room tickets, box office seats, etc.

No Playoff Data

Since not every team makes the playoffs, we chose not to include playoff data, which could potentially alter the results of the model. However, a better solution would be to separate playoffs from regular season since they experience different attendance patterns.

Unavailability of Some Desired Variables

It was difficult to fully flesh out our dataset with all features that we wanted to include. These variables, while not limited to, include precipitation/weather patterns and unique arena theme nights.

FUTURE WORK

Expanding the Dataset

How can we refine and improve our predictive variables such that the strength of the model improves?

Potential Additional Variables

- Social Media Impact Via Google Trends
- Season Ticket/Corporate Ticket Holder Patterns
- Playoff Data
- Team Performance Metrics Beyond W/L

Understanding Revenue Drivers

Our work could be considered very high-level. With further experience within the NHL landscape and gaining a better understanding of the business side of operations, we can perhaps better apply our knowledge to building a more refined model.

Expanding on the Current Code Base

Attempt to employ more advanced modeling techniques, which was a limiting factor within the time frame due to run time constraints

OUR TEAM

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.

THANK YOU!

WORKS CITED

Bowley, Jennifer, and Paul Berger. “Predicting National Football League (NFL) Stadium Attendance.” *International Journal of Social Science and Business* 2, no. 3 (June 2017).

Mueller, Steffen Q. “Pre- and within-Season Attendance Forecasting in Major League Baseball: A Random Forest Approach.” *Applied Economics* 52, no. 41 (2020): 4512–28. https://doi.org/10.1080/00036846.2020.1736502.

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