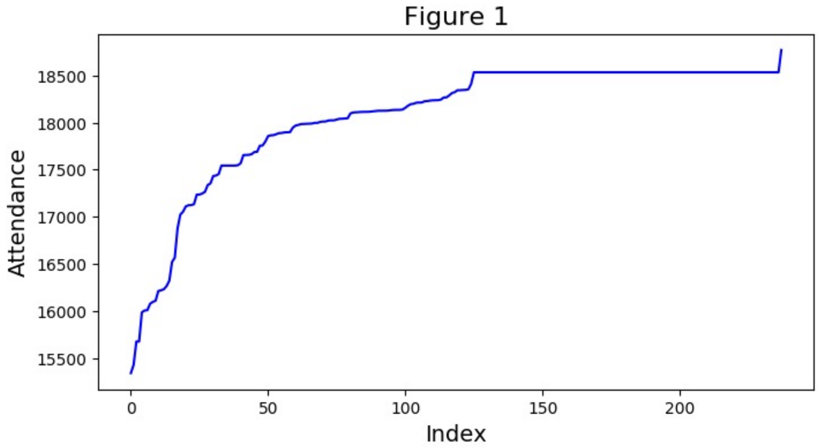
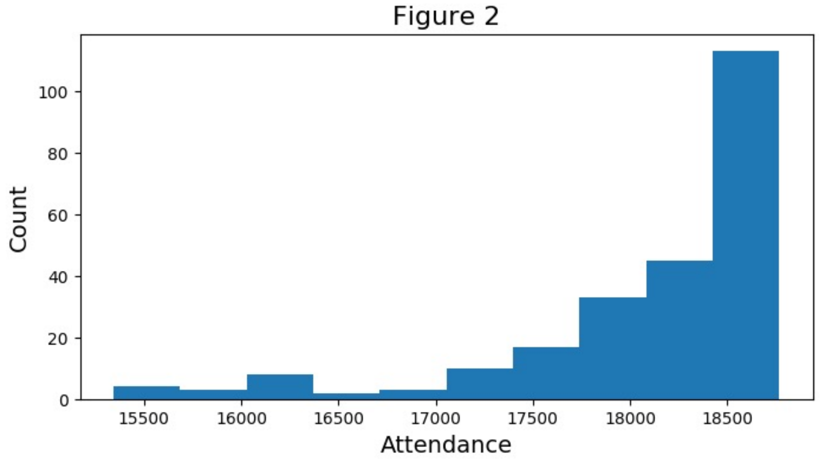
Overview

- The goal of this project is to develop a reliable attendance model which is capable of predicting attendance for the upcoming season's home games.
- We only select a subset of the given data which doesn't contain any preseason games and any home games in the 20-21 season.
- The distributions of attendance from the 14-15 season to the 19-20 season show that the number of games at full capacity is about half of the total number of games, which indicates there is a potential imbalanced problem for regression (Figure 1 and Figure 2).
- We select "Day of the Week", "Puck Drop", "Opponent", "Division", and "Month" as model features, which are all available information before the game starts.
- We regard our attendance model as a tool for preliminary planning of ticket sales and marketing before the season starts. Therefore, last game's result and points percentage before the game starts are not considered as model features since they are only available after the season starts.
- However, during the season, we believe that last game's result and points percentage before the game starts are important to be considered in a new attendance model, which can be used to adjusted the planning of ticket sales and marketing.





One-Layer Model

- The one-layer model represents that we directly use a regression model to predict attendance without any data transformation or over-sampling techniques, given that there is a potential imbalanced problem for regression.
- We build three regression models, including k-nearest neighbors, random forest, and LightGBM, and tune their hyperparameters using Bayesian Optimization with 5-fold cross validation.
- LightGBM is a gradient boosting framework that uses tree-based learning algorithms. Bayesian Optimization is far more efficient in saving time and has better overall performance than grid search and random search.
- The LightGBM regressor outperforms the k-nearest neighbors regressor and the random forest regressor on both test MSE (Mean Squared Error) and test MAPE (Mean Absolute Percentage Error).
- However, the distributions of the true attendance and the predicted attendance of the three regressors show that they are all largely affected by the high number of games at full capacity, which confirms there is an imbalanced problem for regression (Figure 3, Figure 4, and Figure 5).

	MSE	MAPE (%)
K-Nearest Neighbors	526732.59	2.86
Random Forest	481752.00	2.72
LightGBM	481315.11	2.58

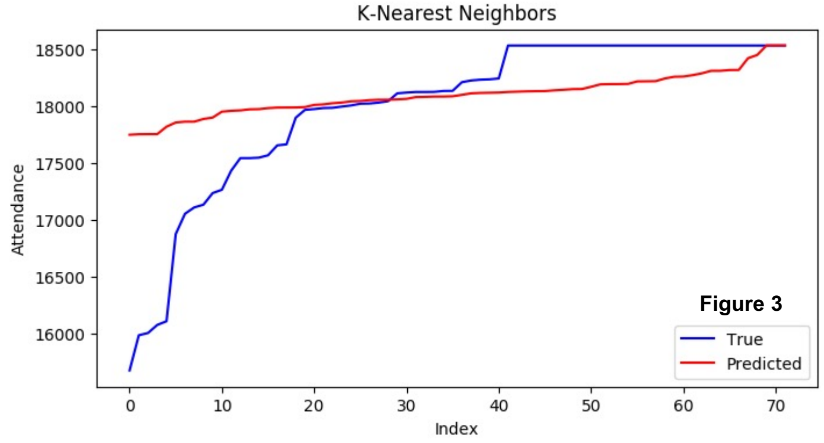


Figure 3

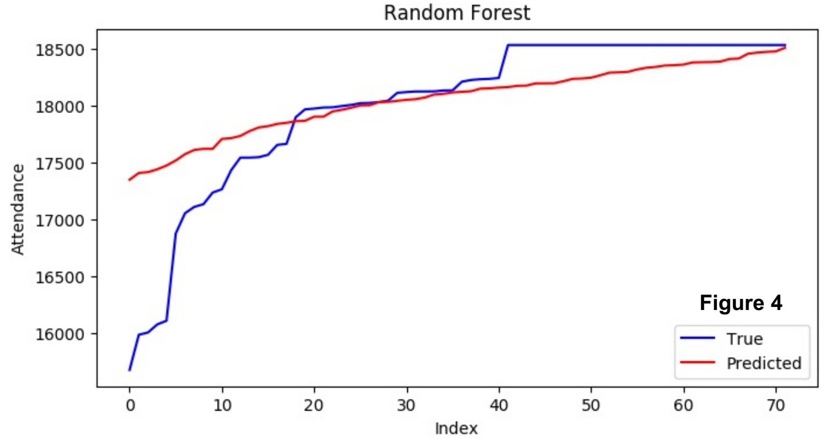


Figure 4

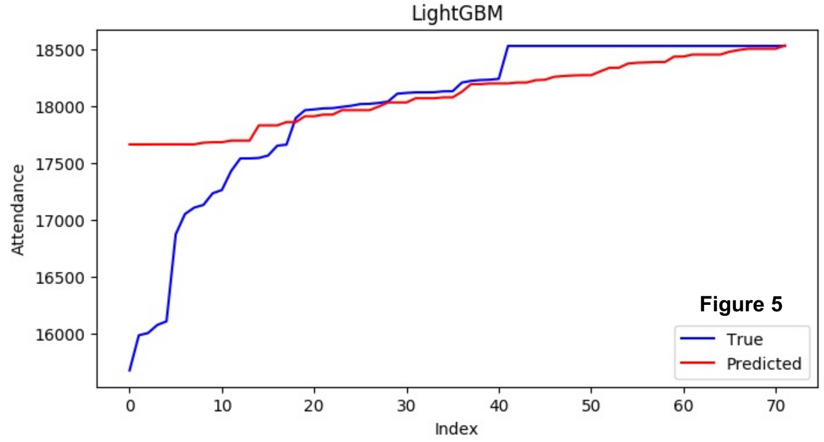
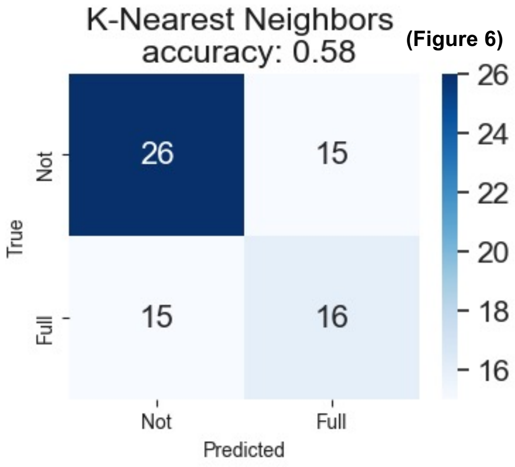


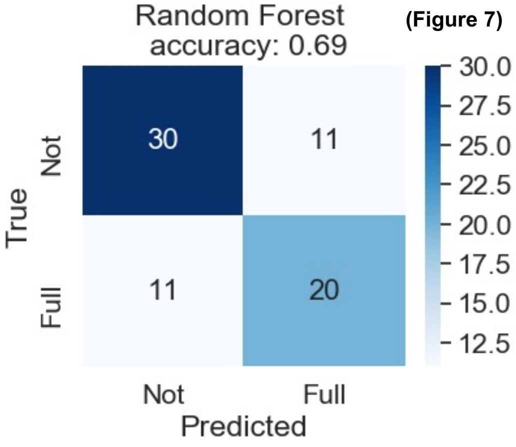
Figure 5

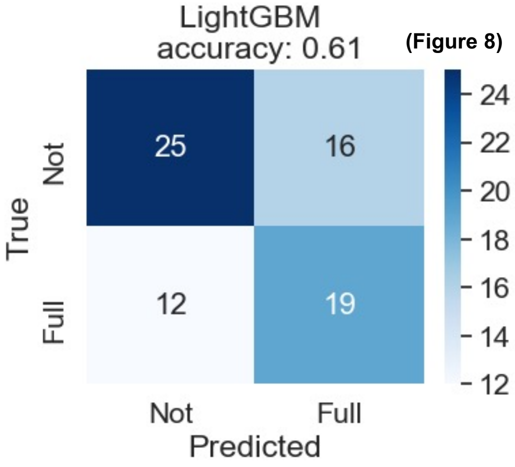
Two-Layer Model (I)

- If there is an imbalanced problem for regression, it's usual to perform a data transformation or a over-sampling technique such as SMOGN. However, since about half of games are at full capacity, we propose a two-layer model to solve the imbalanced problem.
- In short, we first use a classification model to classify whether or not a given game is at full capacity. If it is not at full capacity, we then use a regression model to predict its attendance. If it is at full capacity, its predicted attendance would be directly equal to 18,532.
- In the first layer, we build three classification models, including k-nearest neighbors, random forest, and LightGBM, and tune their hyperparameters using Bayesian Optimization with 5-fold cross validation. The random forest classifier outperforms the k-nearest neighbors classifier and the LightGBM classifier on the test accuracy (Figure 6, Figure 7, and Figure 8). Therefore, we decide to use the random forest classifier in the first layer.
- In the second layer, we build three regression models, including k-nearest neighbors, random forest, and LightGBM, and tune their hyperparameters using Bayesian Optimization with 5-fold cross validation. The k-nearest neighbors regressor outperforms the LightGBM regressor and the random forest regressor on both test MSE and test MAPE. Therefore, we decide to use the k-nearest neighbors regressor in the second layer.

	MSE	MAPE (%)
K-Nearest Neighbors	871190.79	4.30
Random Forest	937740.27	4.49
LightGBM	935007.08	4.53


(Figure 6)


(Figure 7)


(Figure 8)

Two-Layer Model (II)

- We then combine the random forest classifier and the k-nearest neighbors regressor into our two-layer model. Obviously, the proposed two-layer model outperforms the one-layer models on both test MSE and test MAPE.
- Most importantly, unlike the one-layer models, the distribution of the predicted attendance of the two-layer model is pretty similar to the distribution of the true attendance, which indicates the two-layer model is pretty decent and more appropriate than the one-layer models (Figure 11).
- No matter for the random forest classifier or the k-nearest neighbors regressor, "Opponent", "Day of the Week", and "Month" are the top three important features (Figure 9 and Figure 10).

	MSE	MAPE (%)
K-Nearest Neighbors	526732.59	2.86
Random Forest	481752.00	2.72
LightGBM	481315.11	2.58
Two-Layer Model	395365.49	1.95

Note: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$; $MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$

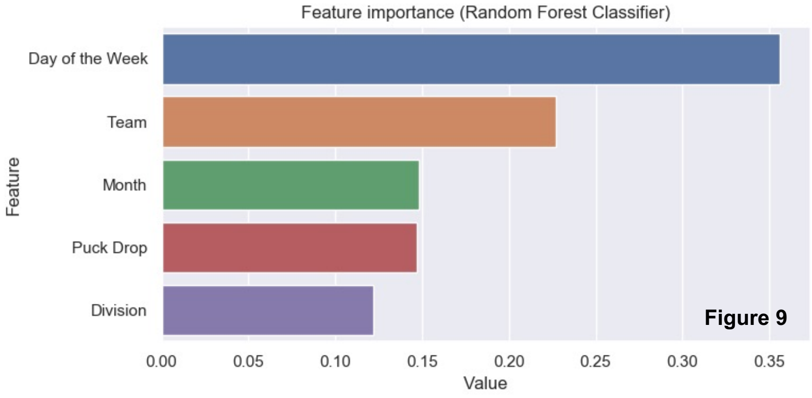


Figure 9

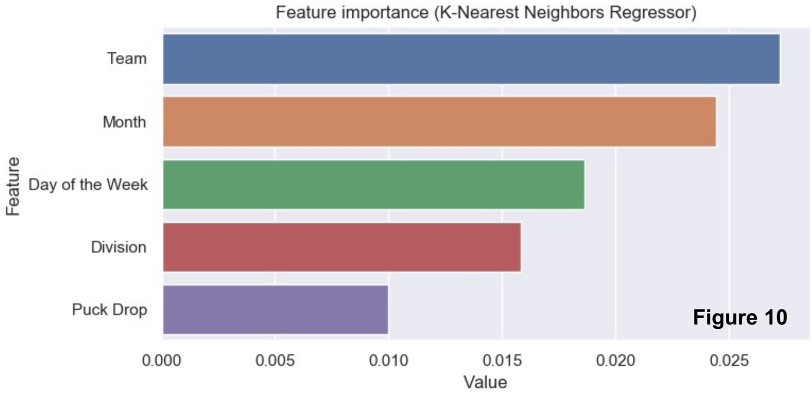


Figure 10

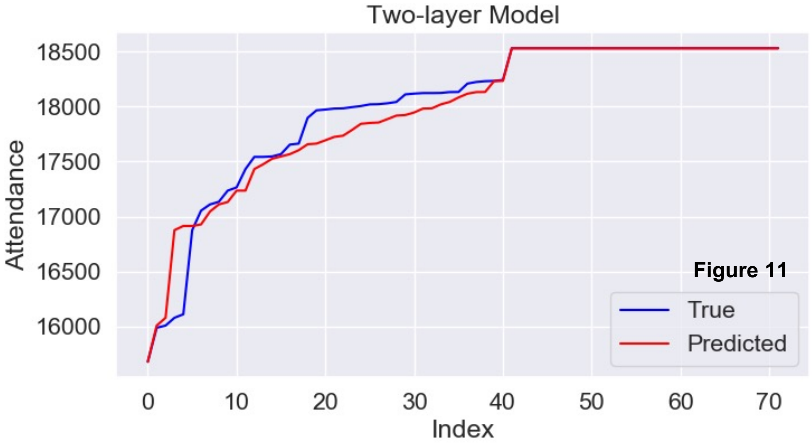
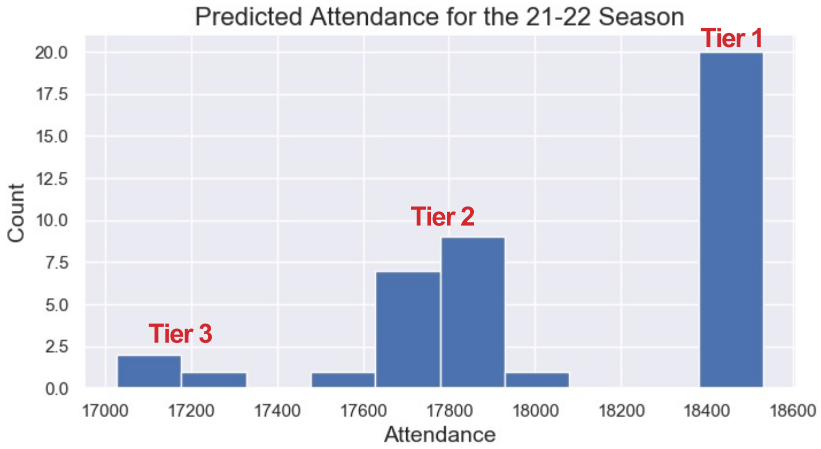


Figure 11

Attendance Prediction

- Now, we are able to use the proposed two-layer model to predict attendance for the 21-22 season's home games. The distribution of the predicted attendance shows there are three groups of the predicted attendance. Therefore, we decide to place the 21-22 season's home games into three tiers.
- The first tier represents the games which are projected to be at full capacity. The second tier represents the games whose attendance is projected to be between 17500 and 18000. The third tier represents the games whose attendance is projected to be lower than 17500.
- In addition to our last game's result and points percentage before the game starts, opponent's last game's result and opponent's points percentage before the game starts can also be considered in a future attendance model during the season. Moreover, it will also be helpful to consider several external factors such as ticket price, temperature, probability of precipitation, and so on.

	Count
Tier 1	20
Tier 2	18
Tier 3	3



Season	Day of the Week	Date	Puck Drop	Home / Away	Opponent	Predicted Attendance	Tier
21-22	Friday	2022/2/11	19:00:00	Home	Calgary Flames	18532	1
21-22	Friday	2021/12/3	19:00:00	Home	Columbus Blue Jackets	18532	1
21-22	Friday	2022/3/25	19:00:00	Home	Nashville Predators	18532	1
21-22	Friday	2021/11/26	19:00:00	Home	New York Islanders	18532	1
21-22	Friday	2021/12/17	19:00:00	Home	Seattle Kraken	18532	1
21-22	Friday	2022/1/14	19:00:00	Home	Winnipeg Jets	18532	1
21-22	Monday	2022/1/10	19:30:00	Home	Montreal Canadiens	18532	1
21-22	Saturday	2022/3/19	16:00:00	Home	New Jersey Devils	18532	1
21-22	Saturday	2022/4/2	19:00:00	Home	Arizona Coyotes	18532	1
21-22	Saturday	2022/1/8	19:00:00	Home	Nashville Predators	18532	1
21-22	Saturday	2022/4/16	19:00:00	Home	San Jose Sharks	18532	1
21-22	Saturday	2022/1/29	20:00:00	Home	Detroit Red Wings	18532	1
21-22	Saturday	2022/1/22	20:00:00	Home	Vancouver Canucks	18532	1
21-22	Sunday	2022/2/13	18:00:00	Home	Carolina Hurricanes	18532	1
21-22	Sunday	2022/3/27	18:00:00	Home	Colorado Avalanche	18532	1
21-22	Sunday	2022/1/2	19:00:00	Home	Buffalo Sabres	18532	1
21-22	Thursday	2021/10/14	19:30:00	Home	New York Rangers	18532	1
21-22	Thursday	2022/3/31	19:30:00	Home	Philadelphia Flyers	18532	1
21-22	Thursday	2022/1/27	19:30:00	Home	Tampa Bay Lightning	18532	1
21-22	Tuesday	2021/11/16	19:30:00	Home	Chicago Blackhawks	18532	1
21-22	Friday	2021/10/29	19:00:00	Home	Minnesota Wild	17839	2
21-22	Monday	2021/11/22	19:30:00	Home	San Jose Sharks	17816	2
21-22	Saturday	2021/11/20	13:00:00	Home	Los Angeles Kings	17855	2
21-22	Saturday	2022/3/5	16:00:00	Home	Winnipeg Jets	17838	2
21-22	Saturday	2021/10/16	18:00:00	Home	St. Louis Blues	17893	2
21-22	Saturday	2021/10/23	18:00:00	Home	Toronto Maple Leafs	18005	2
21-22	Thursday	2022/3/3	19:30:00	Home	Anaheim Ducks	17706	2
21-22	Thursday	2021/12/30	19:30:00	Home	St. Louis Blues	17795	2
21-22	Thursday	2022/3/17	19:30:00	Home	Washington Capitals	17744	2
21-22	Tuesday	2022/3/15	19:30:00	Home	Chicago Blackhawks	17542	2
21-22	Tuesday	2022/1/25	19:30:00	Home	Colorado Avalanche	17665	2
21-22	Tuesday	2022/3/1	19:30:00	Home	Florida Panthers	17748	2
21-22	Tuesday	2022/4/12	19:30:00	Home	Minnesota Wild	17927	2
21-22	Tuesday	2021/11/2	19:30:00	Home	Pittsburgh Penguins	17806	2
21-22	Tuesday	2021/10/19	19:30:00	Home	Seattle Kraken	17777	2
21-22	Tuesday	2022/3/29	19:30:00	Home	Vegas Golden Knights	17727	2
21-22	Wednesday	2022/2/9	19:30:00	Home	Edmonton Oilers	17725	2
21-22	Wednesday	2022/1/12	19:30:00	Home	Ottawa Senators	17865	2
21-22	Monday	2021/12/13	19:30:00	Home	Arizona Coyotes	17151	3
21-22	Thursday	2021/11/4	19:30:00	Home	Anaheim Ducks	17026	3
21-22	Tuesday	2021/12/28	19:30:00	Home	Boston Bruins	17223	3