CISC 451 – Course Project

Predicting Match Outcomes in Professional Ice Hockey
Midterm Report

October 22nd, 2020

Gavin McClelland – 10211444
Marshall Cunningham - 20249991

1. Introduction, Background, and Problem Definition

Data analytics has been rapidly adopted in the realm of professional sports over the past decade and has created an increasing appetite for the application of data-driven methods to develop a better understanding of different concepts. This is especially the case is professional ice hockey, where advanced statistics started being collected by the National Hockey League (NHL) in the 2007-2008 season.

The accessibility of open-source NHL datasets allows for different analytical approaches to be explored and has created a community of data scientists who have collectively brought to light the impact analytics can have on the professional hockey landscape. However, there is a common understanding that there is much more parity in the NHL relative to other sports. As such, there exists an inelastic demand to explore data collected by the NHL to improve intelligent models regarding match outcome. On this topic, the NHL established a relationship with MGM as an official betting partner in 2018 (Rosen, 2018).

The problem to be addressed is how to analytically address the parity in the NHL through the wealth of data readily available by predicting whether a team will win. From there, this would allow for the components of a match to be deconstructed, at which point the individual impact of each feature could be analyzed. This bridges the gap professional teams face in getting the most out of their players and game tactics.

The initial approach specified in the proposal specified the sole exploration of event-driven data—specifically shot locations—but similar attempts in literature are much more sophisticated and are out of the scope of this project's timeline. More importantly, these previous approaches have been solely focused on determining the likelihood of a shot becoming a goal, instead of these micro-level events contributing to the outcome of a game. The expectation is to use this event-driven dataset as a part of the analysis alongside datasets at a higher level of abstraction to simplify the methodology and analytics process.

2. Brief Dataset Description

The dataset that has been explored is the same as the one specified in the proposal, entitled "NHL Game Data", found on Kaggle (Ellis, 2019). This can also be obtained through calling the NHL API (Hynes, 2020). This dataset contains multiple files from the NHL Real-Time Scoring System (RTSS), arranged as a relational database. These files encompass game outcomes, individual events (i.e. hits, shots, face-offs, etc.), and many others dating back to the 2007-2008 NHL season. The files of interest—or those that have been explored to date—are highlighted in the Entity-Relationship Diagram (ERD) provided Figure 1. Due to the size of some of the files, all data used has been condensed to the scope of the 2017-2018 and 2018-2019 seasons to allow for more efficient analyses given the available resources.

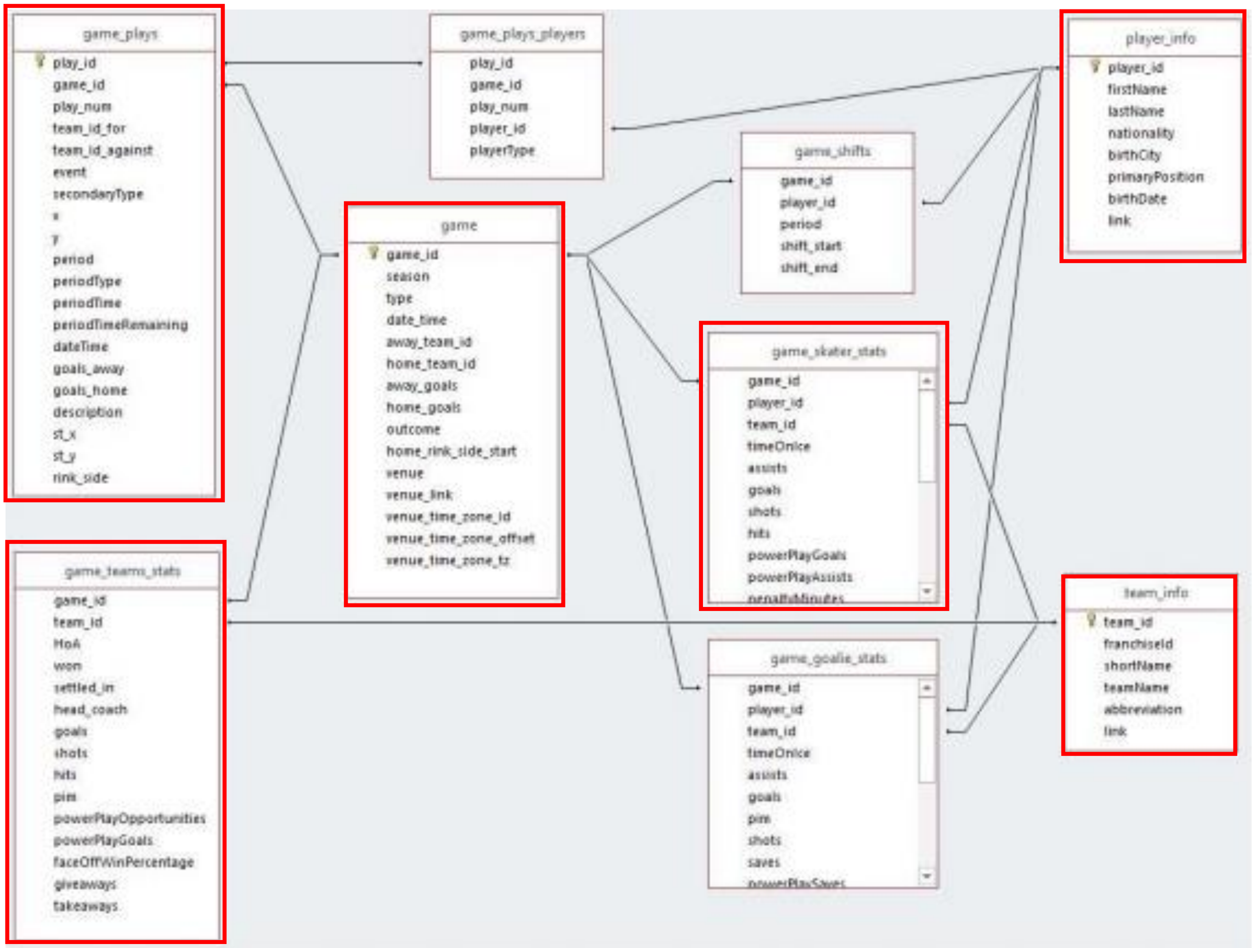


Figure 1 - ERD from provided Kaggle dataset, csvs that are of interest are highlighted in red.

In addition to the ERD, data from other online sources (such as MoneyPuck and Hockey Reference) could be included moving forward to supplement work completed to date (MoneyPuck, 2020) (NHL Advanced Stats / Analytics, 2020). Reasons for this consideration will be discussed in the future work section.

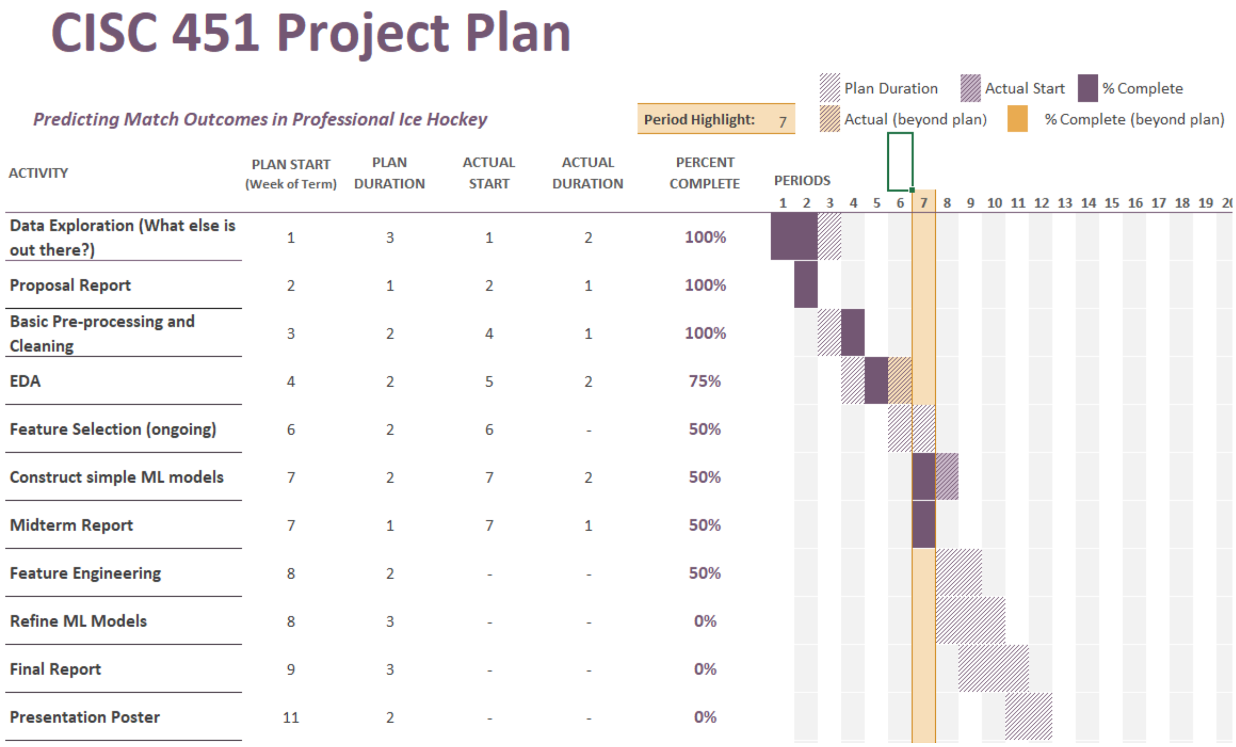
3.0 Assessment of Challenges and Obstacles

Challenges encountered to date include the following:

1. Large volume of data to start with (approximately 1 GB). This was addressed by condensing the provided dataset from 11 seasons to two.
2. Despite condensing the dataset, the feature space to explore was still very rich nonetheless and resulted in a lot of time being dedicated to exploratory data analysis (EDA).
3. Team members are in the process of getting comfortable using Scikit-learn and this has posed a small development hurdle thus far but more experience throughout the term is making this less of a challenge.

4.0 Project Management

A detailed project plan is shown in the below Gantt chart (Figure 2). This timeline has been refined since the proposal as the scope of the project became clearer during EDA. To date, EDA is not complete as there are a few more datasets that can be incorporated to supplement existing work. Moreover, this is the reason why feature selection is only partially complete. Some simple models have been created following EDA, but the expectation is to compare them against other simple models before engineering more complex features prior to the final submission. Otherwise, the project is on track with the main points in the timeline specified in the proposal.



CISC 451 Project Plan

Predicting Match Outcomes in Professional Ice Hockey

Period Highlight: 7

▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ■ % Complete (beyond plan)

ACTIVITY	PLAN START (Week of Term)	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Data Exploration (What else is out there?)	1	3	1	2	100%
Proposal Report	2	1	2	1	100%
Basic Pre-processing and Cleaning	3	2	4	1	100%
EDA	4	2	5	2	75%
Feature Selection (ongoing)	6	2	6	-	50%
Construct simple ML models	7	2	7	2	50%
Midterm Report	7	1	7	1	50%
Feature Engineering	8	2	-	-	50%
Refine ML Models	8	3	-	-	0%
Final Report	9	3	-	-	0%
Presentation Poster	11	2	-	-	0%

Figure 2 – Detailed project timeline

5.0 Methodology and Analytics Process

5.1 Software Packages and Download Instructions

All code written for this assignment was written in the Python programming language using Jupyter notebooks. The main software packages used in this assignment were, pandas, scikit-learn, and numpy. A tree structure of the submitted directory is shown below.

```
|  CISC 451 - Midterm Report.pdf
|  requirements.txt
+---code
|      1_basic_preprocessing.ipynb
|      2_event-based shot analysis.ipynb
|      3_player-level analysis.ipynb
|      4_team-level analysis.ipynb
\---data
       data.zip
       rink.png
```

Instructions to replicate all work completed as submitted are listed below:

1. Install all dependencies by running the command “pip install -r requirements.txt” in the root directory of the submitted folder.
2. Extract the contents of the data.zip folder to the same directory
3. Before running any notebook in the code folder, add your working directory at the top where indicated (i.e. %cd “<your directory here>”). Then, in increasing numerical order, run all cells in each notebook except for “1_basic_preprocessing.ipynb”, which was only used to condense the initial dataset to a manageable size.

5.2 EDA

Exploratory data analysis (EDA) was conducted in increasing levels of abstraction. In other words, the most micro-level datasets were explored, building up to the exploration of the highest-level dataset provided (teamstats_2017-2018_2018-2019.csv).

The first dataset explored was 'plays_2017-2018_2018-2019.csv', containing individual plays with an associated game_id, event description, and the location of the event on the playing surface, among other features. First, the different unique events were identified, then shots and were isolated to explore different properties. Shot events with a null shot type indicated that those shots either missed the net or were blocked by an opposing player, as shown in the code. The main takeaways from this analysis were the ability to visualize shot and goal locations, which was useful to indicate where the density of different events come from. Additionally, generic metrics such as the volume of shots taken, and goals scored by different shot types were created to identify different baseline metrics.

Next was a short analysis of player-level data contained in the file 'skaterstats_2017-2018_2018-2019.csv'. This dataset contains aggregates of individual events attributed to each player by game. This can be deemed useful for modeling moving forward since most features are numeric. Simple metrics were created at both the player and team-level to demonstrate that this dataset could be used to create insights at both the player and team-level. More importantly, this data could be joined with the previous dataset to create a better feature space.

The last dataset that was briefly explored was 'teamstats_2017-2018_2018-2019.csv', which was used to create a simple model in line with similar past approaches that have been published. Most published attempts at predicting game outcome have involved feature spaces at the team-level and have demonstrated that there is more room for improvement. This once again gives this project purpose in addressing this room for improvement. Two simple logistic regression models were created from this dataset to establish a benchmark model, and to explore the impact of individual features.

5.4 Other approaches explored

Similar models in the space of hockey analytics concern the prediction of whether a given shot will be a goal or not—called “expected goals” (xG)—based on several features. It appears this process is much more sophisticated than anticipated, and that there is much work to be done to construct a successful expected goals model before match outcome can be considered as a next step (EvolvingWild, 2018). Moreover, expected goals is a different classification problem altogether, which was not apparent at the time of proposal. So, the desired scope is to consider the binary classification of match results and not classifying the results of individual events. Nonetheless, this feature is included in datasets from the other sources previously mentioned and could be used to supplement existing work by creating a better feature space in predicting match outcome.

6.0 Evaluation of Work Completed

The only work to evaluate out of the work completed so far are the two logistic regression classifiers created at the end of the EDA phase. The first classifier contained all numeric features in the team level dataset and were used to predict the match result (i.e. whether the 'won' column would be "True" or "False"). The confusion matrix and ROC curve are shown below in Figure 3 and Figure 4.

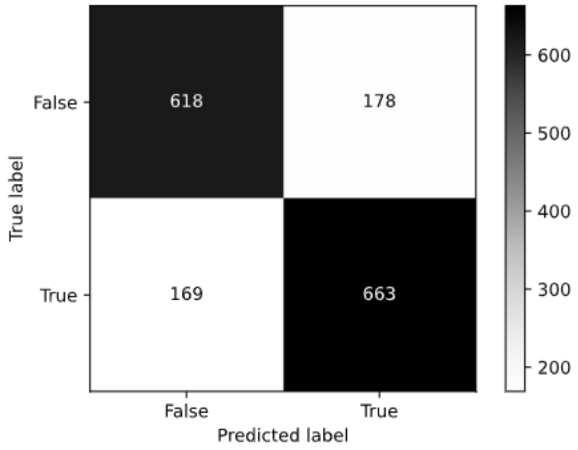


Figure 3 - Confusion matrix for the first logistic regression classifier.

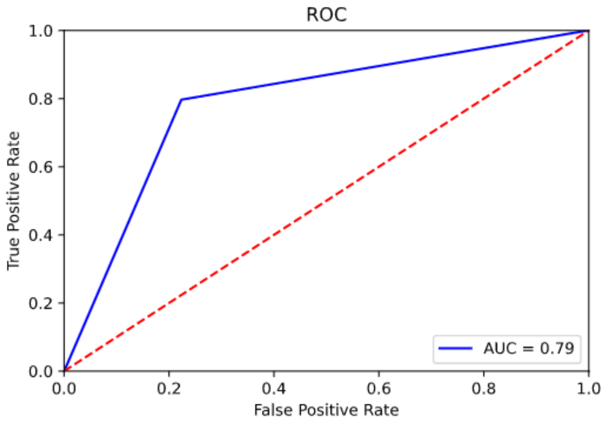


Figure 4 - ROC plot for first logistic regression classifier.

At first sight the ROC AUC of 0.79 was astonishingly high. This was because the classifier included both goals and powerplay goals as part of the feature space, which makes sense. This essentially meant "score more goals to win more games", which is obvious. These features were promptly removed to eliminate this obvious bias and another classifier was trained on the remaining features. The confusion matrix and ROC curve for the second classifier are shown below in Figure 5 and Figure 6.

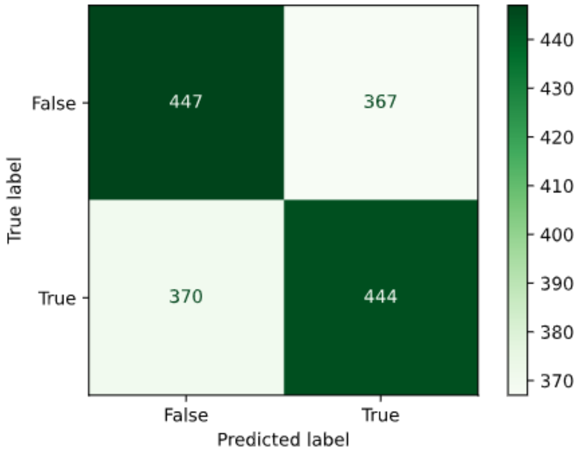


Figure 4 - Confusion matrix for the second logistic regression classifier.

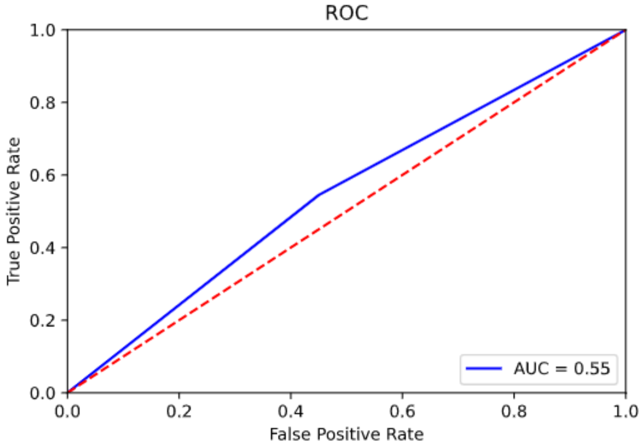


Figure 6 - ROC plot for the second classifier

The results of the second classifier make more sense now without the impact of goal-related features. This ROC AUC of 0.55 is just barely better than a coin flip (50%, shown by the dotted red line), and will serve as a benchmark moving forward.

7.0 Future Work

From evaluating a (very simple) benchmark model, features that are known to impact match result (such as goals) are going to be more important than others when retrospectively predicting match outcome. However, what can be done when these features are not available, such as in the case of predicting the outcome of a match in near-real-time, or scheduled matches that have yet to be played? Clearly the absence of goal-related features has a negative impact on model performance (ROC AUC of 0.55), therefore, a more sophisticated feature space needs to be developed.

There are two approaches that will be pursued for the remaining duration of the project to address this issue. The first is to enrich the existing simple feature space by including player-level aggregates and individual events from the datasets initially explored. This would add depth, data, and sophistication to any models to be developed moving forward. This is also made doable with the use of the schema provided in the ERD (Figure 2). The second approach will be to supplement existing datasets with additional data from other sources that have been previously identified to provide more information about individual teams and/or players (MoneyPuck, 2020) (NHL Advanced Stats / Analytics, 2020). If these approaches are unsuccessful, the existing dataset contains enough data that will support any features to be engineered in future work.

References

Ellis, M. (2019, June 21). *NHL Game Data*. Retrieved from Kaggle: https://www.kaggle.com/martinellis/nhl-game-data

EvolvingWild. (2018, June 7). *A New Expected Goal Model for Predicting Goals in the NHL.* Retrieved from rpubs: https://rpubs.com/evolvingwild/395136/

Hynes, D. (2020, August 5). *NHL Stats API Documentation*. Retrieved from Gitlab: https://gitlab.com/dword4/nhlapi/-/blob/master/stats-api.md

Leung, M. (2018, Aug 27). *$4,718 — Using Machine Learning to Bet on the NHL.* Retrieved from Medium: https://medium.com/coinmonks/4-718-using-machine-learning-to-bet-on-the-nhl-25d16649cd52

MoneyPuck. (2020, September 21). *Download Player and Team Data*. Retrieved from MoneyPuck: http://moneypuck.com/data.htm

NHL Advanced Stats / Analytics. (2020, September 20). Retrieved from Hockey Reference: https://www.hockey-reference.com/analytics/