NHL Game Predictor

Dan
Andrews

The Problem:

As the world of sports analytics grows, more and more types of statistics are amassed. Being able to use those statistics in a meaningful way can provide insight and is a good use of the rich amount of information available. Being able to predict the outcome of a game using collected statistics reliably could have big impacts on not only sports analytics, but also accessory industries such as sports betting or fantasy sports. The goal of this project is to build a reliable predictor of live NHL games, on a minute to minute basis.

The Plan:

Using the NHL's free API to extract tracked statistics from previously played games, and use that data to create predictive classifier models.

Use the same API to live-scrape data from a hockey game as it is happening, and provide real-time updated prediction on which team was going to win the game.

Compare the prediction to the collected statistics in order to consider their impact.

The Trouble with Modeling Hockey (and sports in general)

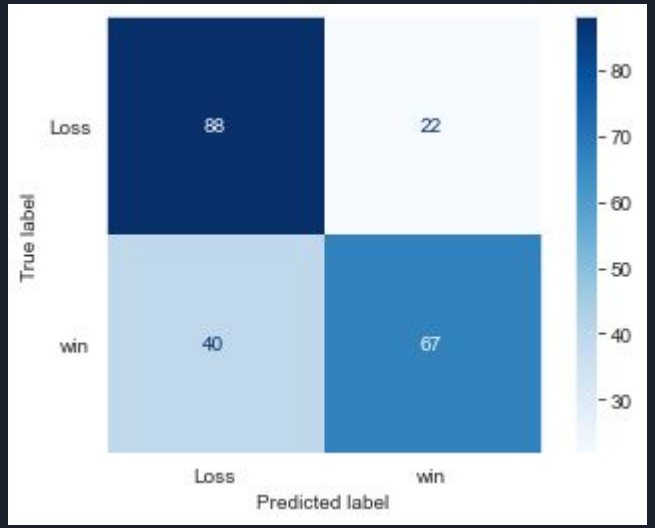
- Using in game statistics can be a strong predictor of who will win a game, sometimes. There is always a chance for a single player to make highlight plays or have a statistics defying game.
- Particularly in the NHL there can be games where a goalie will have an outstanding performance, and the team that dominated in every statistical measure will still lose.
- Small bursts of momentum can have major effects on the outcome of a game, but may not show up in the collected statistics. Examples of this are crowd noise effects or impressive plays that don't end up being recorded as a statistic.

The Approach
two pronged attack

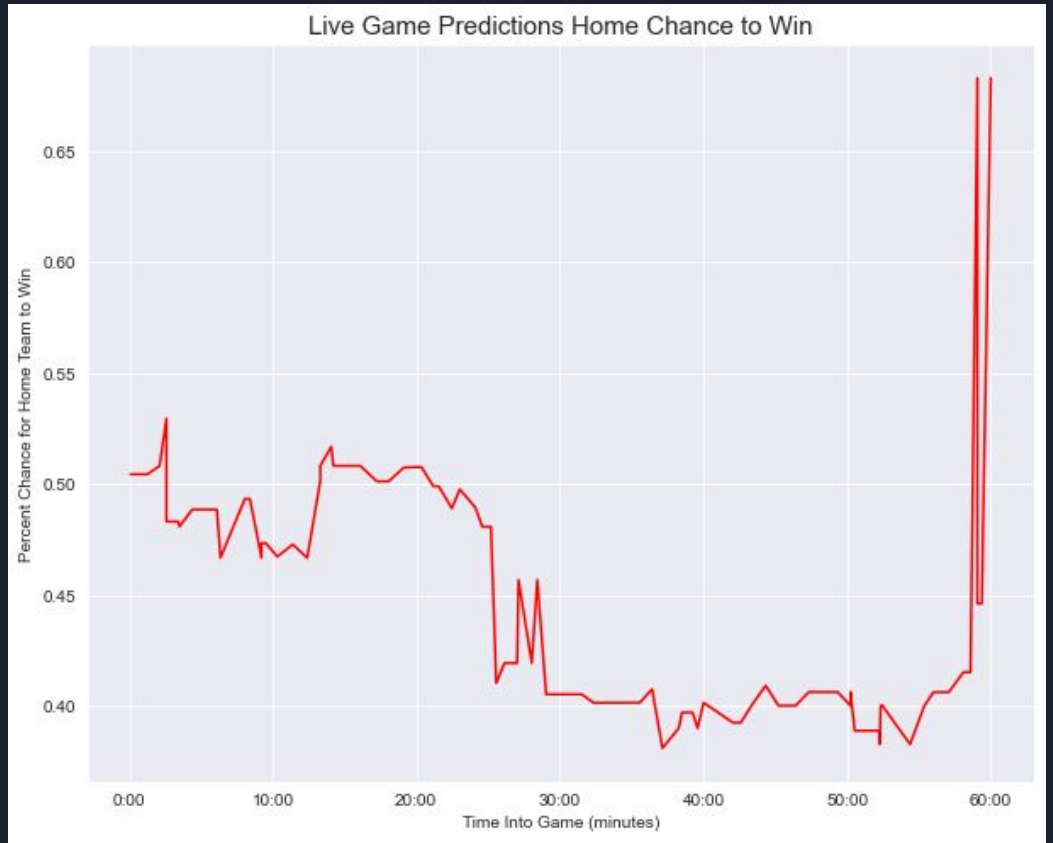
- In order to not train on full game stats, I took the difference for each games stat lines in favor of the home team, so the model could predict whether the home team will win or lose, based off of the differential in the stats.
- This model could not be trained using goal differences, because obviously if the goal difference is positive: win, otherwise: lose.
- In order to use goals in the model, I took the difference in stats for every two minute interval, and trained 30 models, one for each two minute interval in the game. That way the model would be using similar statistical amounts to the real game state at that time.

The End Game Box Score model:

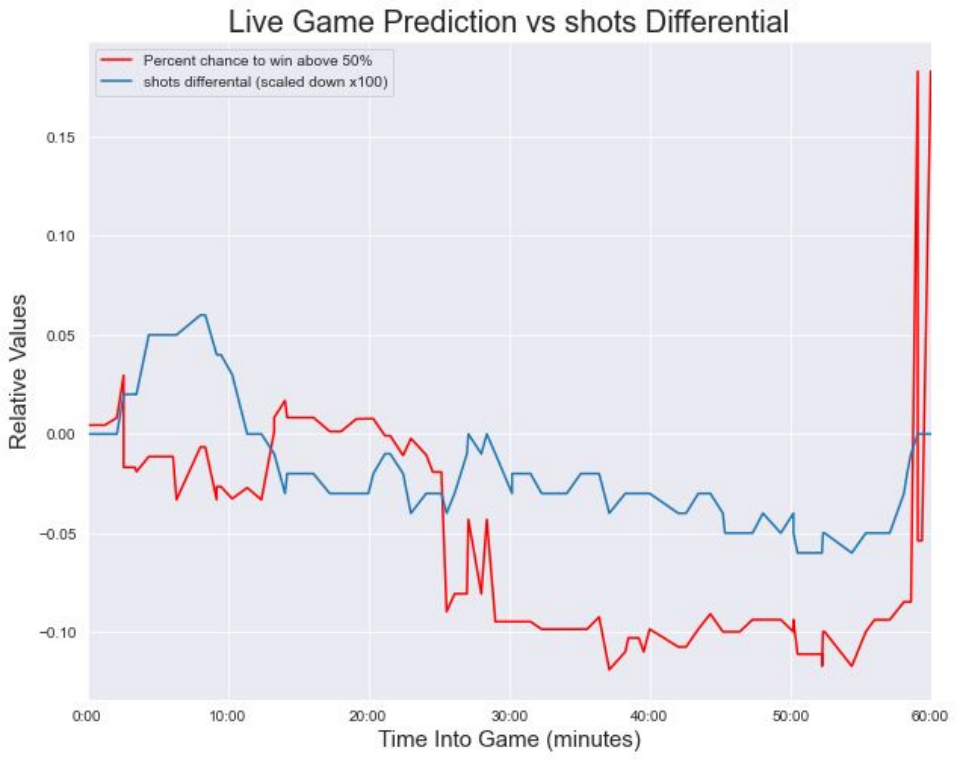
- Random Forest model scored the best out of tested models
- Train accuracy of .70 and a validation accuracy of .71
- Model is performing well on predicting game results using stats differentials
- Confusion matrix shows its more likely to predict a false negative if its going to miss a guess.

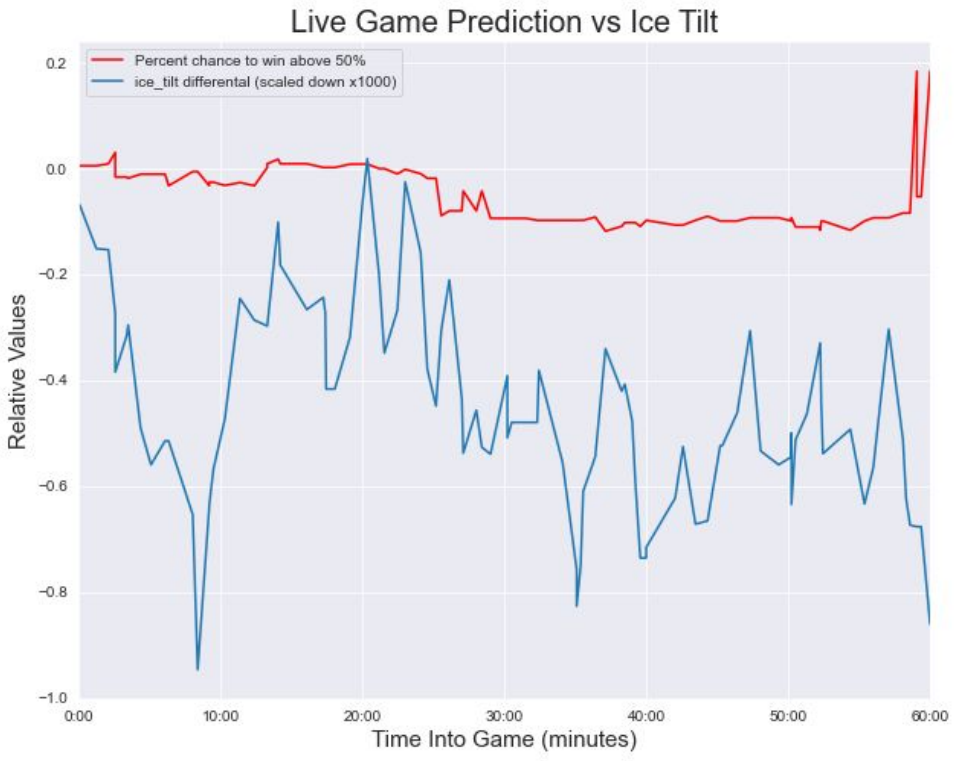


Live Game Prediction: 6/14/2021 game between host Tampa Bay Lighting and visitors New York Islanders



- This graph shows probability of the game being classified as a win for the home team Tampa Bay Lighting.
- Tampa was heavily outplayed for most of the game, but got a surprise late goal and made a strong push to try and come back.





These graphs show the percent chance to win above 50%. If the predictor line drops below the 0.00 mark then the home team are that amount below 50% to win the game. This allows for comparison with some of the model variables, too see how they trend with the game predictions.

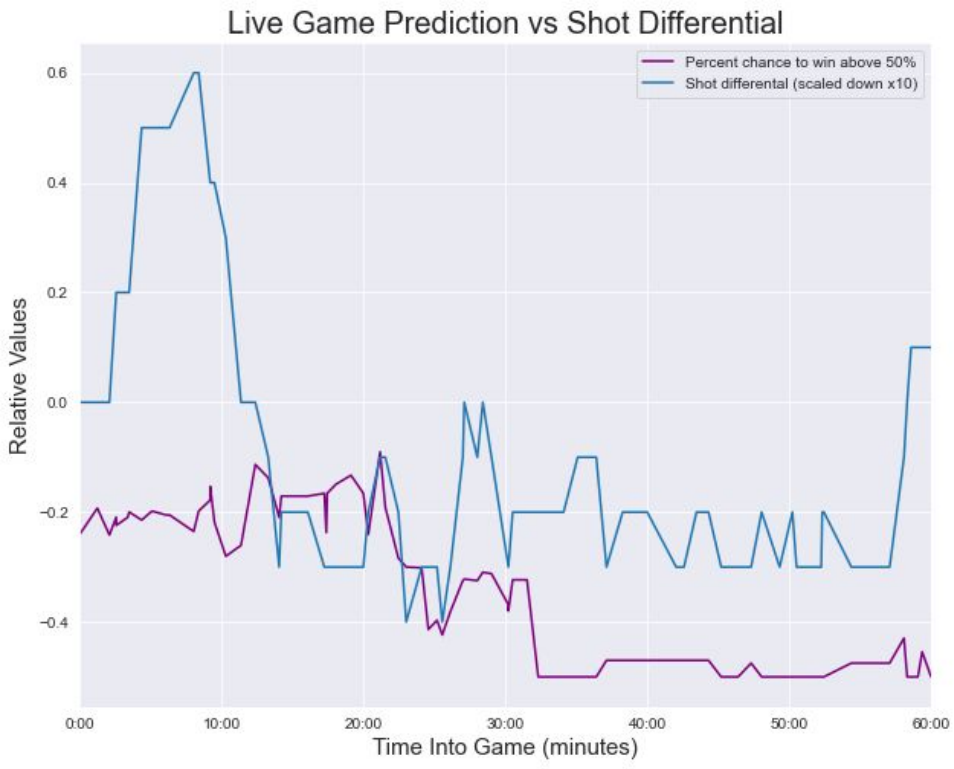
The Minute by Minute models:

- As expected, the early game models are more or less a 50/50 coin flip, as these models are trained on only the stats from the first 2 minutes of a game.
- As the amount of information the model has increases the model becomes more and more confident in its ability to predict.
- The models for the waning moments of the game are very confident in their answers.

Live Game Prediction: 6/14/2021 game between host Tampa Bay Lighting and visitors New York Islanders

- This model is much more likely to make a strong prediction.
- This model is able to train on goals, so it gets very confidant towards the end of the game.
- This model was less impressed by Tampa's late game heroics

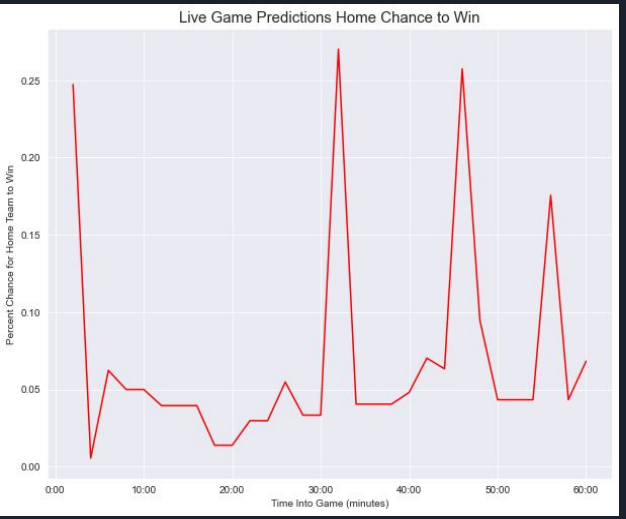
- Comparison of the two types of models, shows them following a similar trend, but the minute over minute model is much stronger in its predictions.

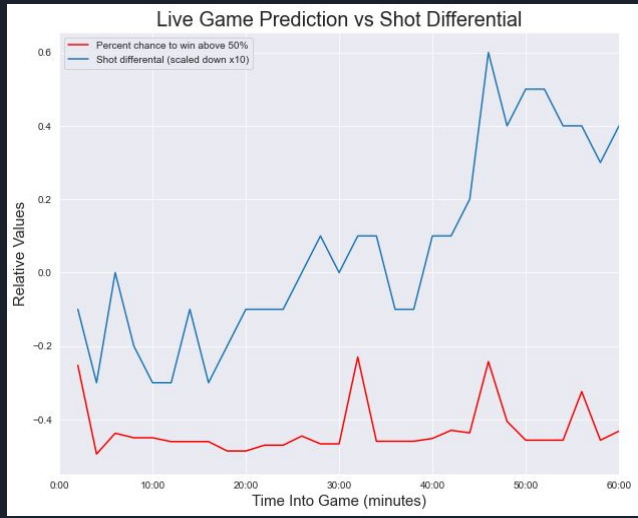


As soon as the Islanders were able to secure their first goal, their domination in all of the other statistical categories made the predictor feel very strongly that Tampa Bay was going to lose.

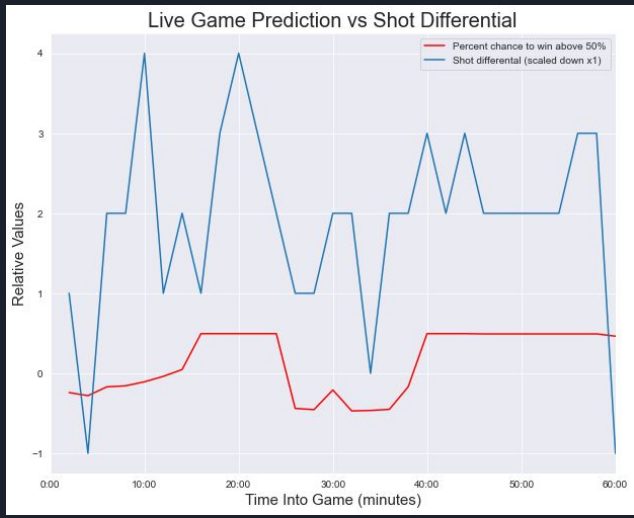
- Giveaways matter!
- It seems that in this game Tampa's high turnover rate (ending up with 6 more turnovers than NY) cost them dearly.
- Their few mistakes in the middle of the third period show up strongly, and negatively in the predictor scores.

Mock Live Game Prediction: 5/20/2021 Pittsburgh Penguins at New York Islanders:





Mock Live Game Prediction: 6/5/2021 Carolina Hurricanes at Tampa Bay Lighting:



Moving Forward with the NHL Predictor:

Next Steps:

- Automate the live scrape process more, to reduce the strain of using the system.
- Add a convenient and easy to use front end, where users can just select what game they want and see the live predictions in real time.

Improvements:

- Features, Features, Features! The NHL API provides an insane amount of data about each game. Manufacturing new features could help the model improve.
- Test more types of models! There might be other classifiers that can outperform our current model.

Conclusions and Recommendations:

Conclusion:

The minute by minute model does seem to have a good chance of predicting the winner of game, especially later on in the game (as expected). There needs to be improvement in the model being able to predict the result of a game early on, the predictor is often wrong early in the game. The full game stats predictor also shows promise but needs tuning in order to be more resistant to non-statistics related events.

Recommendations:

There seems to be a lot of promise in the live game predictor. There is a lot of room for it to grow and become even stronger at determining the game state. The model can also be improved with pre-game stats and and team based trends in order to make it even more reliable of a predictor. Moving forwards with strengthening the predictor and increasing its useability seem like the next steps in the process.

Sources and references:

API documentation: https://gitlab.com/dword4/nhlapi

Ice Tilt image: https://community.rapidminer.com/discussion/44904/using-the-nhl-api-to-analyze-pro-ice-hockey-data-part-1

Paper on the challenges of modeling sports data and hockey: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.841.8005&rep=rep1&type=pdf

All Photos: https:// nhl.com