Predicting NHL Team Success With Machine Learning

Alessandro DeChellis
December 6, 2021
BrainStation

Background and Business Problem

Since the 2004-2005 NHL season, the league has implemented a league-wide salary cap to ensure parity between teams and to ensure that small market teams do not fall behind competitively. Since the salary cap came about, it has been widely debated on strategies to be adopted by General Managers when it comes to distributing this salary cap across a team's roster. On one hand, some executives and fans believe that top players should be paid very high in order to lure them to teams in free agency or to keep them happy on their current team and 'lock them in' for a longer term. Only recently has it been suggested that spending more on your 'depth' players (2nd and 3rd line players) has led to teams having a higher rate of success in the playoffs and brought Stanley Cup Championships to cities. Not only has a high paid player never won a Stanley Cup, there has only been one player to make over $10.5m to win a single playoff round since a salary cap was put in place

It is the goal of this report to determine if one of these strategies is more effective than the other and which players or player 'ranges' a General Manager should be spending more of their salary cap on. Through various machine learning techniques, the hope is to be able to give a clear cut answer on whether or not spending top dollar for superstar players actually makes a difference in the playoffs or not. Hockey analytics have been solely based around finding advanced statistics for actual player performance recently, with no focus on how salary distribution affects a team's success.

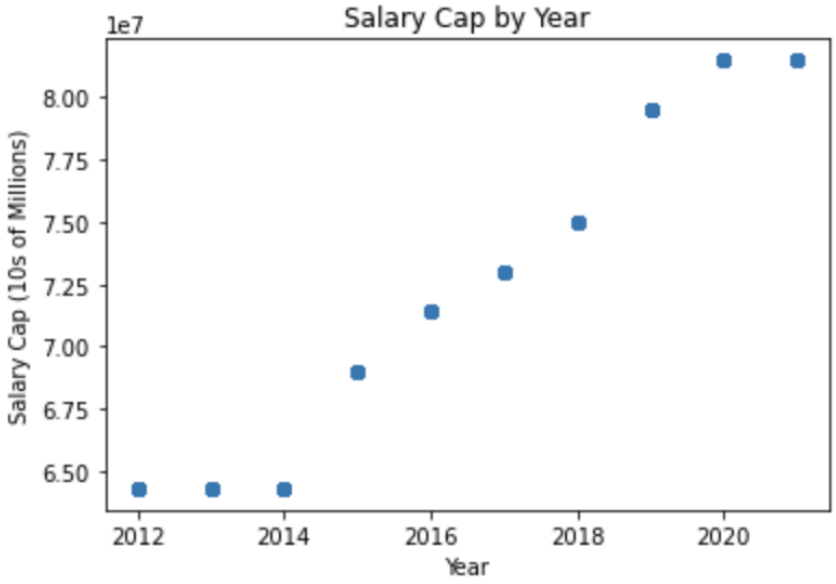
This report will be focused on finding regular season and first round playoff success.

Data Source and Cleaning

All data was retrieved from two websites via web scraping: www.capfriendly.com and www.hockey-reference.com. Both sites are run by passionate hockey fans and include all data needed. Unfortunately, CapFriendly only has players listed for their current team, even when looking at historic salaries, so the rosters were needed. Once a list of rosters and a dataframe of salaries were retrieved, there was much manual work to be done, including changing names to ensure they matched. Once the tables were joined and pivoted, all that was left was manually entering the results of each season (binary values for each round of the playoffs) and the data collection was complete.

Exploratory Data Analysis

The first step in the data analysis was to look at the change in salary cap over the years. It became clear that the changing salary cap would become an issue with the way our columns were currently set up. The individual player salaries in the columns were changed to reflect the percentage of the salary cap that they take up. The change in salary cap over the years can be seen in the figure below.



Once the salary values were changed to be a percentage of salary cap, all 20 salaries were plotted to find out how teams overall were spending on each spot in their roster. It was found that overall, teams spent roughly the same as one another in the middle of the roster, where player salaries were normally distributed. At the tail end of the roster (players 15-20), the percentage of cap spent was right skewed. At a high level, it looks as though player salaries have the same distribution in quartiles, so 4 new features were created. Four quartiles were created for players 1-5, 6-10, 11-15 and 16-20.

To see if these quartiles were successful in telling us how teams who made the playoffs spend their salary cap, the distribution of top salaries of teams by season success and the first quartile of teams by season success were compared. It was seen that the quartiles were not a very telling statistic until after the second round of playoffs. Given that this analysis is focused on the regular season and first round of playoffs, it was determined by these plots that the modelling would be focused on how individual player salaries affect team success. The quartiles were kept in the dataframe to allow the analysis to be expanded upon later.

Modelling

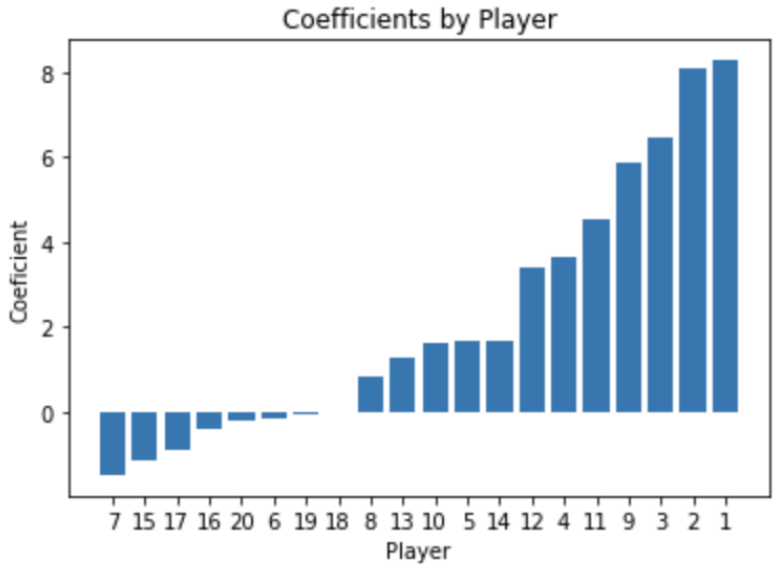
In order to allow this report to have business applications, it was decided that coefficients were needed to allow the reader to see how each individual player salary changes a team's chances at a playoff berth or a first round playoff win. For this reason, Logistic Regression was chosen as the first model for the report. Two regressions were run, one for team's making the

playoffs and one for teams making it past the first round. In order to optimize these models, the data was split into train, test and validation sets and run through a for loop in order to find the optimal value for C. Once these were found, both models were run.

The Decision Tree Classifiers were optimized with the same technique to find the optimal value for max_depth. When the train and validation scores for the loops were plotted, the max_depth value where the train and validation scores were closest were chosen to be the optimal values.

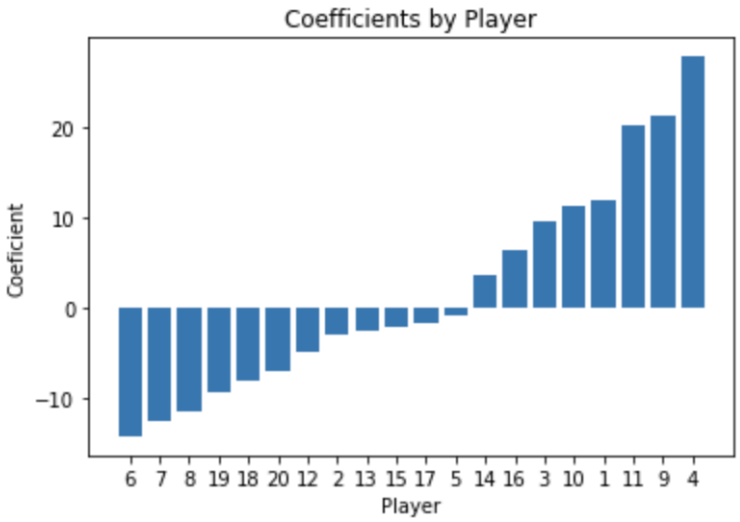
Findings

The regular season success logistic regression was found to have an accuracy score of 59.8% and a recall score of 55%. This was a low score and doesn't give much confidence in the predictive capabilities of the model as it pertains to finding regular season success. The coefficients did confirm some of the original notions that are held about higher paid players and team success in the regular season and can be seen below. It can be seen that the top three paid players on a team have the largest positive effect on a team's chances at making the playoffs. These should be taken with a grain of salt, given that the model is only 60% accurate, but can still be used for creating a spending strategy.



The second regression, run to find how playoff success can be determined by salary distribution, was more accurate than the first. It was found that this model had an accuracy score of 71.3% and a recall score of 61%, allowing us to take the coefficients of this model more seriously and use them for team strategy. The coefficients in this model were quite interesting, as it can be seen that the middle range of players has the highest effect on how the team fares in the first playoff round. These coefficients can be seen in the chart below. It is seen that the

fourth, ninth and eleventh highest paid players on a team have the highest effect on winning the first round.



The decision tree classifier for regular season success had an accuracy of 66% and a recall of 62%, while the decision tree classifier for playoff success had an accuracy of 59.7% and a recall of 62%.

Business Applications

It is quite clear that these findings can be applied by a General Manager in implementing a salary cap spend strategy when building a team or re-signing current players to new contracts. It can also be shown to players in order to coerce them to take more ‘team-friendly’ deals and allow the team to be more successful. It will have to be partnered with another analysis that will find out the true monetary worth of a player based on performance.

Next Steps

Next steps for this report are to go deeper into the playoffs with the analysis. As mentioned before, the further analysis will be done using the quartiles, as it was clear during the EDA that the quartiles start to differ more than individual salaries when playoff rounds get higher.

Along with further playoff analysis, it will be necessary to compare player salaries to performance to determine which players are worth signing for the ranges in the new strategies implemented by this report. The two analyses paired together will give a solid strategy to any General Manager looking to build out a team with data in mind.