ECE1513: Machine Learning Project

How does an NHL Playoff Team Look

Omar Ismail

University of Toronto
April 2019

Introduction

The excitement level of millions of people rises each October as the NHL season begins. 31 teams, each playing 82 games, with the top teams qualifying to the Stanley Cup playoffs. The NHL is split into 2 conferences, East and West, with each conference further split into 2 Divisions. 8 teams from each conference qualify for the playoffs: the top 3 teams in each division, plus 2 "wild card" teams from each conference. These wild card teams are the ones that finish 7^{th} and 8^{th} in their conference. Can we use previous years data to predict which teams should be in the playoffs this year? This is the question we are trying to answer. As of writing this report, the regular season has ended, so we can compare our machine-learned predictions to reality.

Making the Dataset

The first step is making the dataset. We will use historical data starting from the salary cap era (2005). Even though the NHL has been in existence since 1917, prior to 2005, there was no restriction on the amount of spending on players. With the enforcement of a salary cap, the playing field became more level. The data was collected from [1], and for each of the 391 teams, a 0 is assigned if the team made the playoffs that season, and 1 otherwise. The data consists of 33 features. As our dataset is not large and we have a lot of features, we want to avoid over-fitting. To do so, we need to cut down the number of features.

Feature Selection

We want to be able to predict a playoff team without knowing the team's ranking, or factors that influence ranking, as that would make the task trivial. So, we got rid of these features. These are:

- Ranking and Games Played
- Wins and Losses (which include overtime and shootouts)
- Strength of Schedule (this looks at strength of opponents)
- Simple Rating System (this determines how good a team is)

We also make sure we do not have repetitive features. For example, there is a Goals Scored feature, but also Even Strength Goals Scored, as well as Power Play Goals Scored, the summation of the latter two giving the former. So we removed the former. The result is that we are able to reduce our features from 33 to 19. We would like to further reduce our features. We found that the Random Forest Classifier has a tool which is able to identify the important features. We use this to identify the Top 3 features. A visualization is shown below:

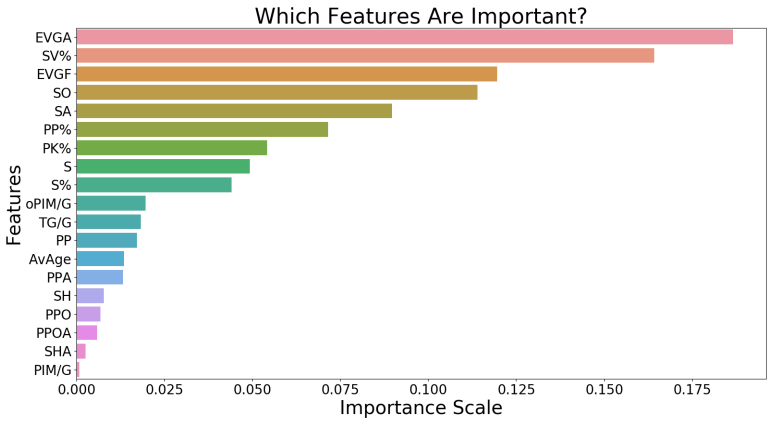

Figure 1: Feature Selection

As seen in Figure 1, the top 3 features are Even Strength Goals Against (EVGA), Goalie Save Percentage (SV%), and Even Strength Goals For(EVGF). These 3 features are the ones we will use going forward. Note, that the code has the capability of choosing more than 3 features.

Using K-Means to Classify

In this section we use K-Means to see if we can distinguish between playoff and non-playoff teams based on the 3 features selected (EVGA, EVGF, and SV%); thus, our K value is 2. To assess the accuracy of the clustered points, we will compare the assigned points to each cluster to the actual targets for both training and validation sets. We then look at how this year's teams cluster using K-Means and comment on how accurate that is.

Results and Discussion

First, we show our training and validation set plots in the figure below. We can see that after around 70 epochs, the validation set loss starts to increase, so we stop the training their:

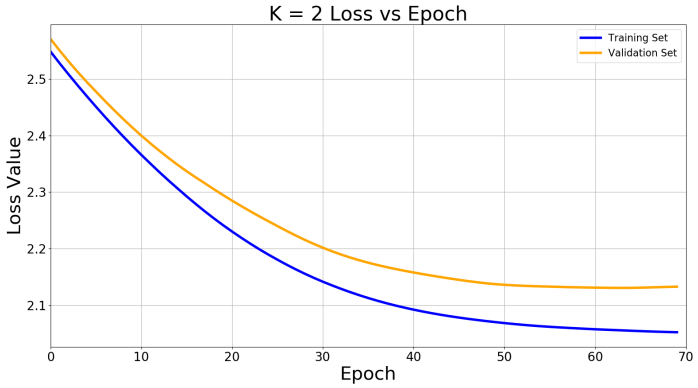

Figure 2: Training and Validation Set Loss vs Number of Epochs

When we compared the accuracy of this model to the actual data of previous years, we found the clustering algorithm was 67% accurate on both training and validation sets. This may seem low, but an argument can be made that it is reasonable. The accuracy numbers can be interpreted as: using statistics on just the goals scored, goals against, and goalie save percentage, we are able to determine if a team is a playoff team 67% of the time. In the real-world, there are many more factors that influence if a team makes the playoffs, including luck; however, these models can be very difficult to formulate.

You might ask why not use more or less features, as opposed to just 3. We tried this for different feature set sizes: 2, 4, 8, and all of them, and found the training set accuracy for each feature was 27%, 21%, 20%, 32% respectively, which is considerably less than when the feature size is 3.

Let's look at how K-Means clustered this year's teams, where red identifies teams in playoffs, and blue otherwise:

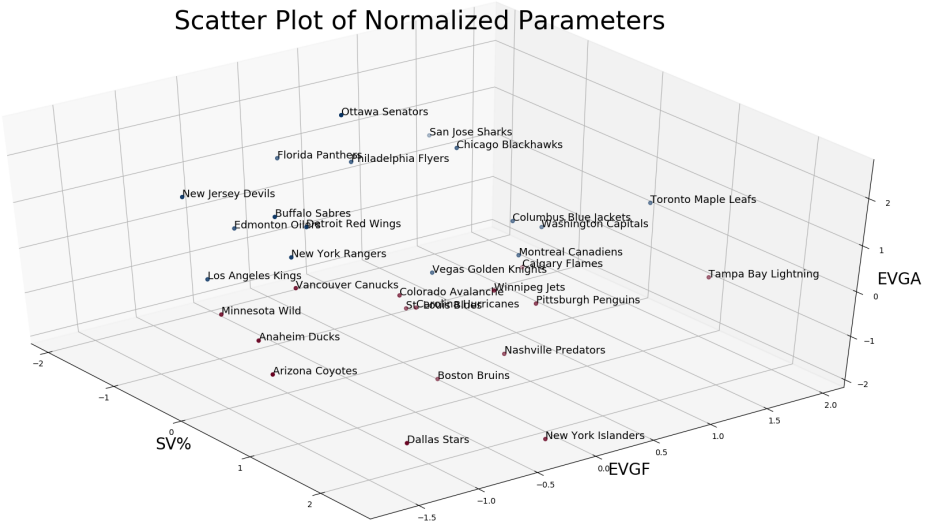

Figure 3: Who is making the playoffs according to K-Means Clustering

Some observations:

- It identified 15 teams making the playoffs, this is one short of what, where 16 teams make the playoffs
- You can easily see the difference in the plot between the best teams in the league, like the Tampa Bay Lightning, and the worst, the Ottawa Senators
- Out of the 15 predicted playoff teams, 4 of them did not and they all are in the same area: Vancouver, Minnesota, Anaheim and Phoenix. The area they are in is where they have better than league-average goaltending, but much worse scoring.
- Out of the 16 predicted non-playoff teams, 5 of them actually made it: Toronto, Columbus, Vegas, Washington and San Jose. The outlier here seems to be San Jose, who have one of the worse goaltending save percentages in the league, but make up for that by being on the higher end of scoring, and playing in the weakest division this year: the Pacific.
- With 9 incorrect predictions, the accuracy of the model on this year's data is 71%, which is close to the training and validation accuracies

Using Logistic Regression

In this section and the next, we factor into our training and validation data whether a team made the playoffs, therefore we move from unsupervised to supervised learning. The first supervised algorithm we will use is Logistic. Our model is as follows:

$$\sigma(x^T W + b)$$

In this model, the outputted vector is the probability of a team **not** making the playoffs. We use the Adam Optimizer to find the weights and biases that minimize the model above, and add a regularization term of 0.01 to avoid over-fitting the data.

Results and Discussion

The loss and accuracy plots on the training and validation data is shown below:

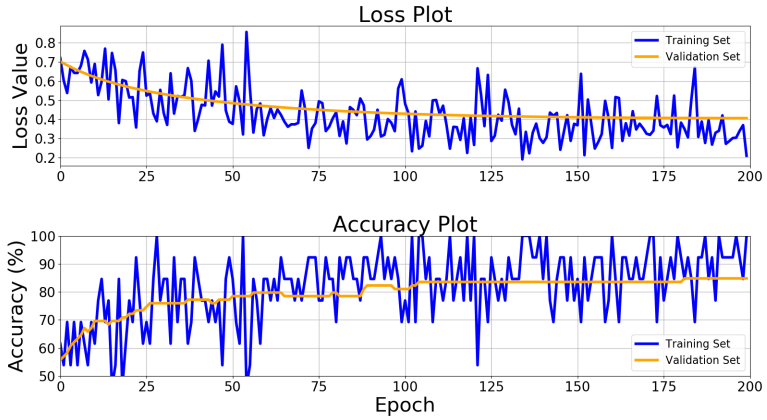

Figure 4: Logistic Regression Loss and Accuracy Plots

The noise in the training set is likely due to the small batch size 13, and the small training set size, 312 points. The final training and validation set accuracy are 100% and 84.8%. Our final weight and bias vectors are:

$$W = \begin{bmatrix} -0.66 \\ -1.67 \\ 1.75 \end{bmatrix}, b = \begin{bmatrix} -0.13 \end{bmatrix}$$

Using these values we can calculate the probability of each team making the playoffs. Unlike unsupervised learning were the algorithm only divides teams into 0 or 1, the logistic gives us a range between these two values. The model allows us to always choose the right number of playoff teams. Not only can this model predict the playoff bound teams, but it can also predict the first-round match-ups in the play-offs. For this year, the predicted playoff match-ups are as follows:

Eastern Conference Match-ups:

1. Tampa Bay Lightning vs. WC2. Carolina Hurricanes
2. Boston Bruins vs. 3. Toronto Maple Leafs
1. New York Islanders vs. WC1. Montreal Canadiens
2. Pittsburgh Penguins vs. 3. Washington Capitals

Western Conference Matchups:

1. Calgary Flames vs. WC2. Colorado Avalanche
2. Vegas Golden Knights vs. 3. San Jose Sharks
1. Nashville Predators vs. WC1. Winnipeg Jets
2. Dallas Stars vs. 3. St. Louis Blues

Apart from the Canadiens, every team predicted to make the playoffs matched reality! The Columbus Blue Jackets made it instead. So the predictor got 2 teams wrong, yeilding an accuracy of 92.3%. It might seem like Montreal got really unlucky this year; looking at the figure below, it seems the logistic predictor had them as the first Wild Card team. Of note is how the model does well in segmenting the teams that had no chance towards the end of the season in making the playoffs from teams that were in the hunt for a spot; towards the end of the season, the last 2 playoff spots was contested tightly between Montreal, Carolina, and Columbus.

Team	SV%	EVGF	EVGA	Playoff Chances
Montreal Canadiens	0.212899	0.867069	-0.153974	0.879876
Carolina Hurricanes	0.092901	0.058994	-0.878475	0.861346
Columbus Blue Jackets	-0.147094	1.037190	0.229586	0.796741
Philadelphia Flyers	-1.227071	0.314175	1.209794	0.093455
Detroit Red Wings	-0.507086	-0.493900	0.826234	0.077779
New York Rangers	-0.147094	-0.834142	0.655763	0.075583
Florida Panthers	-1.707062	-0.026067	1.167176	0.043702
Buffalo Sabres	-0.507086	-0.749082	1.209794	0.027415
Ottawa Senators	-0.987076	0.101524	2.488326	0.009014
New Jersey Devils	-1.227071	-1.089324	1.337647	0.007846

Figure 5: Wild Card East Race

In terms of predicted first round match-ups, only 3 of the match-ups were predicted correctly: Boston vs Toronto, Calgary vs Colorado, and Vegas vs San Jose. From these, just 1 had the seeding incorrect: Vegas vs San Jose.

Using Neural Network to Classify

Now we look at a more complicated supervised learning algorithm: a Neural Net. We implement a 3-layer MLP, with a drop-out layer between the 2^{nd} and the 3^{rd} layer. The hidden unit layers each have 500 nodes. The drop-out rate is set to 0.5, and just like in logistic regression, the regularization is set to 0.01.

Results and Discussion

The loss and accuracy plots are shown below:

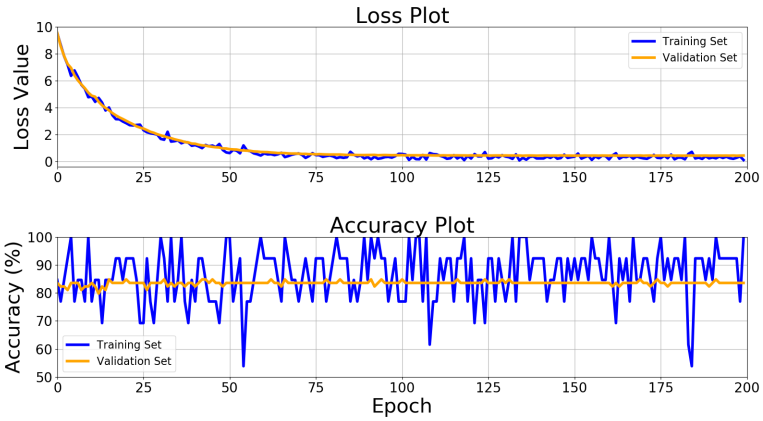

Figure 6: Neural Net Loss and Accuracy Plots

The final training and validation set accuracy are 100% and 83.5%. Note that the validation accuracy is 1% less than logistic regression, which is a minor change. It may seem like extending the classification of the dataset to a more complicated alogrithm did not make a difference. Using this neural net, the predicted playoff match-ups are as follows:

Eastern Conference Match-ups:

1. Tampa Bay Lightning vs. WC2. Carolina Hurricanes
2. Boston Bruins vs. 3. Toronto Maple Leafs
1. New York Islanders vs. WC1. Montreal Canadiens
2. Pittsburgh Penguins vs. 3. Washington Capitals

Western Conference Matchups:

1. Calgary Flames vs. WC2. Colorado Avalanche
2. Vegas Golden Knights vs. 3. San Jose Sharks
1. Nashville Predators vs. WC1. Dallas Stars
2. St. Louis Blues vs. 3. Winnipeg Jets

Again, the model misclassifies Montreal and Columbus, yeilding an accuracy of 92.3%, the same as the simpler logistic model. In terms of predicted first round match-ups, 5 of the match-ups were predicted correctly: Boston vs Toronto, and the entire Western Conference. From these, just 2 had the seeding incorrect: Vegas vs San Jose, and Winnipeg vs St. Louis. Note that Winnipeg and St. Louis finished with the same amount of points this year, but the tie-breaker for the higher seed went to Winnipeg!

Conclusion

It seems like the added complexity of the neural net model only imporves predicting the first-round match-ups compared to the logistic model. In terms of predicting the correct teams making the playoffs, both models return the same results. This makes sense: a neural nets advantage of being a non-linear classifier shows when we got a very large dataset. In th