Thesis Question:

Can the Lake Placid data be used to predict future outcomes?

Spencer Fascetta

Methodology:

Using the shot data from Lake Placid, build a logistical model to determine how predictive the dataset is.

Why did I choose this methodology?

This methodology is nearly identical to that used in my previous thesis paper, which will also be released this week, and therefore provides readers an introduction or example of the work done in that paper, as well as utilizes a proven method.

The Experiment

There are two event variables: Shots and Goals. Shots encapsulate any shot attempt that was not a goal scored. Shot attempts have four attributes: Shot Type, Shot Destination, Shot Traffic, and One-Timer. Shot Traffic and One-Timer are Boolean attributes, either defined as true or false. Shot type has 6 possible variations: wristshots, slapshots, snapshots, wrap arounds, fans, and deflections. Shot destination has three possibilities: on net, blocked, or missed. By grouping the data first by event, then by shot type, shot destination, traffic, and one-timers in that order, there are 144 unique possible combinations. Of those combinations, 67 of them occurred within the Lake Placid dataset. This gives us our initial dataset to use for the model.

So, that means we need to build our model. First, split the dataset into two random sets, one to train the model, and one to test it. I defined the categorical variables for the model using the previously noted attributes for shots. Then, I built a function to generate a list of dummy variables to help artificially expand the size of the dataset. I then used Synthetic Minority Oversampling Technique, or SMOTE, in order to balance the model into a more even training dataset. This allows us to maintain the integrity of the testing dataset while smoothing the training one.

I then tested each feature of the model independently to see how well it performs using recursive feature elimination. I tested it with every possible variable and pared it down to only the features that returned TRUE, which increases the accuracy of the model.

A picture containing text

Description automatically generated

Cool, let’s RUN the model then!

Analysis

Text

Description automatically generated with medium confidence

Graphical user interface, application, table

Description automatically generated

So, what does this tell me? Well, right off the bat I don’t love the R2 value sitting at a 0.45. It’s not great. I also find it interesting that, of the variables tested, only three provided a non-negative coefficient. This tells me that shot type was not a significant predictor of whether or not a shot attempt was actually a goal. Obviously, that attempt being on net is a significant indicator of conversion. What seems more interesting to me is that shot type seems to be almost a superfluous indicator. While it’s clear that deflections and wristshots are the most effective shot types given they have the highest coefficient values, there isn’t a lot of difference between each of the 6 shot types. Interestingly, slapshots are less effective than fans, meaning players who whiff on a shot attempt have a greater probability of scoring a goal on that shot than if they took a full slapshot. Ideally, slapshots should be used to generate deflections, not to score.

It’s also rather intuitive that the presence of traffic is a significant indicator of conversion. It is much easier to score when the goaltender cannot see the shot. However, that actually isn’t what the data shows. It seems that, by a slight amount, players who got a clean look at the goaltender were much better at converting their attempts than those shooting through a screen. Now, without looking at the individual players who this was true for, it would stand to reason that the league’s elite shooters were probably a large driver of this factor, and in a larger sample size, the importance may be reversed.

The accuracy here of our logistic regression classifier using the set of testing data is 0.83. I ran a confusion matrix to see how the model performs, and strangely, it produces zero false negatives.

A picture containing text

Description automatically generated

Table

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

Finally, let’s take a look at what this data actually looks like. I mapped shot attempts, goals, and shots onto a rink to visualize the data. Here is where the shot attempts originated:

Chart

Description automatically generated

And here are the goals:

Chart

Description automatically generated

What I find interesting here is how much goal scoring was actually skewed to one side of the zone. There is a dearth of goals from the low and mid danger range in the goalies’ right, but goalies struggled with long-distance shots from their left. That feels bizarre but is worth additional exploration in the future.

Conclusion

So, what does this tell us? Overall, I’d say this data is reasonably accurate to the true talent of the players in Lake Placid, to the point where I’d probably be comfortable extrapolating additional information from it moving forward. However, it’s still a relatively small sample size, and these tests did not factor any external variables (game time, ice conditions, etc), nor did it consider strength, score, or time left on the clock. This would indicate that it’s probably not a good enough dataset to make definitive conclusions about individual players just yet but is worthwhile to analyze for league-wide trends.

I felt this dataset was less useful than that used in my previous paper, but that dataset was also significantly larger. Moving forward, there are aspects of this project that can be related to that one, however. This project allowed me to define my shot danger locations on an actual coordinate plane rather than as categorical data, and while I have not yet determined the best way to sort the attempts into those categories, it does give me an excellent starting point. So, while the results of this particular analysis were rather underwhelming, they should prove fruitful for future endeavors.

I have attached my thesis paper as well as a reference point, as it serves as my bibliography. The code involved in both projects was identical, save for a few naming differences. Consider the thesis my appendix to this project.