NHL Draft Forecasting Analysis

By: Grant Culbertson and Andrew Mayer

Background

Hockey is a pretty unique sport as most of its players do not play NCAA hockey before getting drafted in the NHL. While the number of college athletes in the NHL is increasing, a majority of players still make their way to the league through junior hockey leagues. These leagues give a lot of prospects opportunities to play a large amount of games and thus a large amount of data. Using this data for draftees from 1998-2008, we were looking to see if this data would provide insight into these players' performance and longevity in the NHL.

We approached this task by looking at a variety of plots, PERMANOVA analysis, and CatBoost models to help identify patterns in the data and predictive input variables in determining career length and performance.

Methods

- **Permutational Multivariate Analysis of Variance (PERMANOVA)**
  PERMANOVA is a non-parametric multivariate variant of standard ANOVA. This method was ideal for our data as most values of our response variables of interest were 0 making for very non-normal data. PERMANOVA operates using distance matrices of permutations of the original dataset to identify significant differences between categorical variables.
- **CatBoost Gradient Boosting**
  CatBoost is a type of gradient boosting algorithm just like XGboost, which we have seen in class before. The main difference between CatBoost and XGboost is that CatBoost has native support for categorical variables and generally has better performance.

Results and Takeaways

After 999 permutations of the data using Euclidean distance measures, PERMANOVA found significant differences in the multivariate pattern of response variables (made up of CareerLength, sum_7yr_GP, and sum_7yr_TOI) between the country of origin, position, and junior hockey league. Unfortunately, PERMANOVA does not lend itself to identifying specific differences in groups so we moved on to CatBoost models to get some insight on this and more.

Our CatBoost models had respectable RMSE values after tuning each through 1642 different model variations. The model for career length had an rmse of 3 years and our model for seven years time on ice had an rmse of 1430 minutes. Overall, we were happy with their performance. In interpreting our CatBoost models we were able to identify some valuable insights into the drafting of NHL players: Looking at SHAP values we found some interesting breaks in the data that could be used as potential rules of thumb for drafting, such as refraining from drafting players with over 50 penalty minutes in their junior league season high in the draft. Overall, we did not find anything too surprising with the CatBoost models, but further research into SHAP values for all factors could provide more impactful insights.