Final Project: Proposal

Michael Kuby, Sangmun Kim, Haichen Sun, Yuyang Chen

1. Research questions

1. What player metrics are correlated most strongly with an NHL player's salary?
2. Given performance data for NHL players and their salary, can we determine which players are over-valued? Under-valued? Valued appropriately? (Undervalued means they are underpaid based on their performance. Overvalued means they are overpaid based on their performance.)
3. Can we devise a recommendation system that advises trading one player for another, resulting in a positive salary/performance exchange for us/our team?

2. Dataset utilization

In our analysis, we plan to utilize two primary datasets, and potentially a third, which will primarily be obtained through web scraping techniques to ensure the most current and comprehensive data is used. Where scraping might not be used is highlighted below. Here are the details of the datasets:

Player Career Statistics:
Source: Natural Stat Trick
Description: This dataset includes detailed professional data for hockey players. It encompasses various metrics such as goals, assists, shots on goal, time on ice, and many other statistics that quantify a player's performance on the ice. Natural Stat Trick is renowned for its depth of data and analytics on hockey, making it a valuable resource for our analysis.

Player Salary information:
Source: CapFriendly
Description: CapFriendly provides comprehensive salary information for hockey players, including annual salaries, contract details, and cap hits. This dataset is crucial for our analysis as it allows us to correlate player performance metrics with their financial compensation. CapFriendly is a trusted source for salary cap and contract information in the hockey community.

Public Sentiment Data on Players [Potential]:
Source: Reddit
Description: This dataset will be compiled from various hockey-related discussion forums on Reddit to gather public evaluations and discussions about hockey players. It will include fans' and observers' opinions, ratings, and comments, offering a unique perspective on players' popularity and public image. The idea behind the Public Sentiment Data is to leverage a qualitative understanding of that player's value to their market. Public sentiment will *not* be a factor we will integrate into a player's *value*; instead, we intend to use this information to *inform trade recommendations*. For example, we envision scenarios where our model sees a player

underpaid relative to their performance but has negative public sentiment data. The recommendation system may avoid recommending a trade for that player in this situation.

By combining these datasets, we aim to thoroughly analyze how various performance metrics are associated with players' salaries. The player career statistics from Natural Stat Trick will offer insights into the on-ice contributions of players. At the same time, the salary information from CapFriendly will help us understand the economic aspect of player valuation. Our methodology will involve scripting and web scraping to collect the most up-to-date data from these sources, ensuring our analysis is accurate and relevant.

3. Methodology

- ## Data Collection:

1. Player Career Statistics will be scraped from https://www.naturalstattrick.com.
2. Player Salary information will be scraped from https://www.capfriendly.com or https://www.spotrac.com or procured through email with the appropriate sources.
3. [Potential] Public Sentiment Data on Players: This data will be scraped using the Reddit API from Reddit hockey/NHL forums. We have started to collect related subreddits and will fine-tune our scope of scraping to specific players as we progress further.

- ## Data Exploration:

Exploratory Data Analysis will have to be conducted on the data to understand the relationship between the performance metrics and salary. We are likely to find that *many* metrics we have will be highly correlated with salary. So, one of our challenges will be to identify performance metrics that are highly correlated with both salary and several other metrics and isolate them to be our features.

Important to understand is that NHL player categories fall into three general buckets: forwards, defenceman, and goalies. Modeling all players appropriately will require training three separate models since each category will have features that correlate to salary differently. So, data exploration will be necessary to understand what features are most strongly associated with the salary for each positional group.

To aid in our Data Exploration, we may also undertake PCA.

- ## Data Cleaning:

Player Career Statistics:

Player Career Statistics will have to be checked for null values, but otherwise, they have been obtained from a highly maintained table that would appropriately be assumed to be reliable.

The feature space will likely require standardization before model training since many features have significantly differing ranges.

Player Salary Information:

To use historical records and salary information, the salary values themselves will have to be adjusted for inflation.

Public Sentiment Data on Players:

Public sentiment data scraped from Reddit’s API will have to be filtered by player and text parsed into records in an analyzable fashion.

- ## Data Integration:

Once the salary data is obtained, we must undertake an entity resolution process that integrates the salary data with the player career statistics dataset.

- ## Data Analysis:

Part 1: Training ML Models and Making Value Predictions

As identified in the Data Exploration section above, the first step in our data analysis journey will be to undertake EDA and possibly PCA for all three categories of players: forwards, defencemen, and goalies.

Following the data exploration phase, we intend to develop machine learning models that output a player's value based on their *performance*. To do so, for each of the three positions – forwards, defencemen, goalies – we must identify the feature space that most influences salary. Once identified, we will train three models, one for each position, that we can then use to predict a player's value.

Due to the nature of our predictions, we will be using a regression model. Ideally, for each prediction made, we will also have a properly labeled **y** to use metrics like the Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R-squared, or one of the other standard evaluation metrics used for regression models.

Once a player’s predicted value is obtained, we can evaluate a player's performance based on their salary using **predicted value - salary = y**. If **y** is positive, our model sees them as *overperforming* based on their salary. If **y** is negative, our model sees them as *underperforming* based on their salary. Values close to zero suggest that our model considers them appropriately compensated for their performance.

Part 2: Designing a Trade Recommendation System (A)

The next step in our analysis will be implementing a trade recommendation system. The question we are trying to answer is: Given a player that we would like to trade (presumably an underperforming player), who should we trade for? To answer such a question, we will have to identify all players in the league with similar salaries within a given range and evaluate their **predicted value - salary = y**. Players with large values of **y** can be considered desirable trade targets.

Part 2: Designing a Trade Recommendation System (B) – Time permitting

To bolster our trade recommendation system, we will use NLP to train a model on player sentiment data obtained by scraping appropriate Reddit Forums. The task of the model will be to evaluate phrases and words from posts to try to understand the sentiment surrounding that player in their current local market. Players with favourable sentiment should be scored highly, while players with unfavourable sentiment will be scored poorly. The task will then be to integrate the results of this sentiment analysis into the player recommendation system: Players with highly positive sentiment have their trade recommendation bolstered. At the same time, players with poor sentiment will be negatively penalized. In essence, the player sentiment data will be used to qualitatively inform the predicted value based on a player's performance metrics.

Note: We state "Time permitting" here because, while we would like to implement this feature into our recommendation system, it may be too much work given the available time. However, we will do what we can to make it happen.

- Data Product:

We envision the final product as an interactive web app allowing users to select a player and immediately identify whether our model thinks that player is under or overvalued, based on their performance. This should be a reasonably simple interface, with a drop-down menu listing all teams in the league and then a second drop-down menu listing all players on that team. Once a player is selected, the relevant information will be shown in a clean and organized format.

A secondary feature will be the trade recommendation system. Given a player that has been selected, the trade recommendation system will offer a set of players it sees as ideal trade targets, along with their corresponding performance metrics.

4. Expected impact

Primary Impact

Consider the following scenario: You are the manager of an NHL hockey team. Your goal is to assemble the best team possible. What is the best way to go about solving this problem? Consider an additional constraint: you have limited money to spend. In the NHL, there is something called a "salary cap" that limits the summed amount a team may spend on its players. This differs from MLB or football (soccer) in that the only constraint placed on a team's spending is the amount an owner is willing to pay. The idea is to increase the parody between teams throughout the league so that all teams are competitive.

Clearly, the best team is the one that extracts the most *value* from their players, given that they can spend a finite amount of money. A great team has many *overperforming* players *relative to what they are paid.*

In its most ambitious sense, our project aims to solve this problem for the general manager. Given our product, an NHL manager should be able to select any player from their team and identify whether that player's performance aligns with their salary. If they are being overpaid based on their performance, the system can find reasonable trade targets around the league that a manager might pursue. Using this systematic approach, ideally, a team could be assembled with many overperforming players relative to their salary.

Tertiary impact(s)

Given our ML value prediction model, we should be able to extrapolate information about entire teams: Is our team's performance relative to our salary structure acceptable?

Given our ML value prediction model, we should also be able to evaluate the performance of a General Manager: Have they done a good job assembling a team?

Given our trade recommendation system and its integration with player reputation, we might understand the relationship between reputation and salary. Do well-liked players get paid more? We might further concretize other abstract abilities: leadership, off-field impact, etc. In other words, we are illuminating new ways of evaluating a player's value.

By computing the net value of a team, we are simultaneously introducing a metric that evaluates one team's performance relative to another. We can use this evaluation process to predict the winner and loser of any game. This feature could be employed, for example, in the gambling industry.

5. Potential challenges

1) Gathering sentiment data on role players

One anticipated obstacle in our analysis is the difficulty of gathering evaluations for less well-known players from Reddit. Given the nature of the platform, discussions and opinions are often focused on more prominent or controversial figures, leaving those with a lower profile without significant commentary. This gap could hinder our ability to determine these players' reputations and public perceptions accurately.

Proposed Solutions-Normalization of Data for Lesser-Known Players:
We plan to adopt a normalization technique for players with little or no evaluations available. This involves marking these players with an "average" or "normal" reputation score based on

our collected overall distribution of player evaluations. Here's how we plan to implement this method:

Establishing a Baseline: We will analyze the distribution of reputation scores for all players with sufficient data to identify a median or mean score that reflects an "average" reputation within the context of our dataset.

Assignment of Average Scores: Players needing more evaluations will be assigned this average score. This approach ensures that every player is included in our analysis without disproportionately affecting the overall insights due to the absence of data.

2) Procuring Salary Data

One foreseeable challenge is procuring enough Salary Data. We know where this data exists, but getting our hands on it may be a slight challenge. The terms of service of CapFriendly.com request that developers do not scrape their site and suggest that if the data is desired, they email them. They will determine their course of action based on the use case. Thankfully, other websites and domains are available to procure this information (*including manual labelling*), so there are ways around this problem.