# **DS 320: Predicting Hockey Contracts**

Brian Ellis Luke Kerwin Eric Wu Griffin Jordan 2023-12-05

### **Abstract**

This report encapsulates our approach to predicting National Hockey League (NHL) player contract values by utilizing a sophisticated data integration and machine learning pipeline. Our study primarily focuses on integrating and analyzing player performance data and historical contract information, spanning from 2009 to 2022. The data is meticulously sourced from CapFriendly and Hockey Reference through python web scraping techniques.

A key aspect of our methodology involves rigorous data cleaning and preprocessing, where we exclude Entry-Level contracts (ELCs), Restricted Free Agent contracts (RFAs), and Contract Extensions to refine the dataset for more accurate model training. These three types of contracts are restricted in ways that would skew our modeling. This cleaned data is then stored in a locally hosted MySQL database, with a Flask and SQLAlchemy built API facilitating remote data access, demonstrating potential for online deployment.

The cornerstone of our project is the development of a predictive model, derived from the merged dataset of contracts and performance statistics. This model, encapsulated as a Python class, is integrated into our API, enabling real-time predictions as player statistics update.

Further, we have designed an interactive dashboard using Tableau, connected directly to our MySQL database, to visually represent our findings and predictions. While currently limited to manual updates due to resource constraints, the dashboard is designed for potential future automation and easy management, should Tableau Server be utilized.

This report not only presents a viable model for predicting NHL player contracts but also exemplifies the effective use of data integration, machine learning, and visualization techniques in sports analytics. The implications of this study extend beyond contract predictions, offering insights into data-driven decision-making in professional sports.

### **Table of Contents**

- 1. Introduction
- 2. Methodology
- 3. Implementation
- 4. Results
- 5. Discussion
- 6. Conclusion
- 7. References
- 8. Appendices

#### Introduction

In an era where data reigns supreme, sports analytics has emerged as a cornerstone of strategic decision-making in professional sports. The ability to parse through vast amounts of data and extract meaningful insights is invaluable, particularly in the high-stakes world of professional hockey. This paper delves into the realm of data integration within sports analytics, with a focus on the National Hockey League (NHL).

Data integration, the process of combining data from different sources into a unified view, plays a pivotal role in modern sports analytics. It enables a comprehensive analysis of player performance, a crucial factor in determining contract values in the NHL. However, predicting player contracts is a complex task, riddled with challenges due to the dynamic and multifaceted nature of player performance metrics and contract negotiations.

Our project confronts this challenge head-on. We aim to predict NHL player contracts by leveraging a decade's worth of player performance data (2009-2022) and historical contract information, sourced from CapFriendly and Hockey Reference By integrating these diverse datasets, we seek to unveil patterns and correlations that influence contract values.

The objectives of this study are twofold: firstly, to develop a robust data integration pipeline that can efficiently process and merge diverse data types; and secondly, to create a predictive model that can accurately forecast NHL contract values based on player performance data.

The significance of this study extends beyond the immediate realm of contract prediction. It serves as a case study in the effective application of data integration techniques in sports analytics, potentially paving the way for similar analyses in other sports disciplines.

This paper is structured as follows: we begin with a comprehensive methodology section outlining our data collection, cleaning, and integration processes. This is followed by a detailed account of our model development and implementation strategies. We then present our results and discuss their implications, before concluding with key takeaways and potential avenues for future research.

### Methodology

#### 1. Data Collection

#### **Sources**

#### CapFriendly

- Used to gather all historic NHL contract signings from 2012-2022.
- Used BeautifulSoup to scrape each individual table month by month due to the websites web-scraping prevention tools.
- Specific URL here

#### • Hockey Reference

- 2009-2022 due to needing n-3 previous seasons of statistics to predict year n
- Much easier to scrape than CapFriendly due to the tables being plain HTML tables.
- Benefited from using Pandas' read\_html() function.
- Specific URL here

#### Webscraping Techniques

During the process of collecting data some insights we gained were profound in how we set up and deployed our data pipeline. For example, normally you can retrieve data by changing the URL parameters. In this case, the web URL was almost a dummy url. We later found an Ajax request call that had similar parameters in it. The next challenge was adapting for the pagination. If we ever called a page that didnt exist, the code would instantly break and we would lose our data. To combat this, we were able to use caching and try-except statements to safegaurd our data collection.

With the Hockey Reference data source since the statistics were not paginated, we didnt have to iterate from month to month. We could get a whole years data in one request call. All we had to do was change the year in the URL; https://www.hockey-reference.com/leagues/NHL\_2022\_skaters.html. One unique thing about this dataset was that the HTML column headers were built into the actual dataset. We were able to handle this in the Data Cleaning and Preprocessing stage.

Python code used to webscrape can be found here.

#### 2. Data Cleaning and Preprocessing

#### **Contract Data**



Figure 1: Contract Raw Data

Above is what the contract data looked like as soon as we scraped it from CapFriendly. Some key features that we need to clean and preprocess are:

- 1. Handling PLAYER.1 column
- 2. Standardizing the DATE to a more accessible data point
- 3. Filter TYPE to Stnd (UFA) to get the true contracts
- 4. Replacing EXTENSION values with binary (0,1)
- 5. Standardizing VALUE and CAP HIT to integers from strings
- 6. Adding an id column for easy database storage

After implemeting these changes our final contract dataset looks like this:

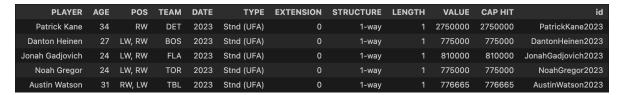


Figure 2: Contract Cleaned Data

#### **Statistics Data**



Figure 3: Statistics Raw Data

As you can see from above, there are some key changes we will need to make inorder to make our data friendly not only for machine learning but also for storing in our SQL database. Heres what we did:

- 1. Remove Rk as it has no value to us
- 2. Use TOT (Team) rows for players who played on multiple teams in one season (example: Craig Adams above)
- 3. Fix duplicate column names such as EV, PP, SH
- 4. Fix column names with % in them as they are not MySQL compliant names
- 5. Add an id column
- 6. Remove goalie statistics

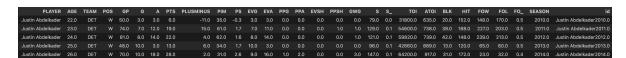


Figure 4: Statistics Cleaned Data

Python code used to clean and preprocess the data can be found here.

After formatting and filtering our data, we have 1006 contract observations and 9186 statistical observations to support.

Now we need to store the data.

- 3. Database Setup and Management
- 4. Data Integration
- 5. Model Development
- 6. API Integration
- 7. Visualization and Dashboard
- 8. Challenges and Limitations

## Implementation

## Results

### Discussion

### Conclusion

### References

## **Appendices**