

# Predicting Hockey Contracts

Eric Wu, Luke Kerwin, Brian Ellis, Griffin Jordan

2023-12-03

## Introduction

In the dynamic and competitive world of professional sports, where every move can make or break a team's fortunes, the business of player contracts stands as a pivotal aspect that demands precision and foresight. Our group took on the challenge of developing a machine-learning model capable of predicting the financial value of a one-year hockey player contract.

The professional sports industry, particularly the National Hockey League (NHL), has witnessed an unprecedented evolution in recent years. As teams increasingly rely on data-driven decision-making processes, the need for predictive models to estimate player contract values has become more apparent. The ability to anticipate the financial implications of signing or retaining a player not only facilitates effective roster management but also plays a crucial role in maintaining the financial health and sustainability of a team.

With the use of a multitude of player performance metrics and historical data, the goal was to create a model that could predict the size of a player's contract for a desired year. This report delves into the methodologies employed in curating and preparing the data, selecting the appropriate features, and fine-tuning the machine learning model. We navigated through a few challenges, from data availability and quality to data integration.

## Datasets

To create a model that could accomplish this task, the first step is to retrieve all the data that impacts a player's future contract. This includes all types of player informational data such as age, height, and weight. We would also need a player's historical data which is comprised of all the in-game performance statistics throughout their entire career. The first dataset we used was from a website named CapFriendly.

CapFriendly is a website that provides information and tools related to the salary cap in professional hockey leagues, particularly the National Hockey League (NHL). It offers users the ability to explore and analyze team rosters, player contracts, salary cap details, and other financial aspects of hockey teams. The site is a valuable resource for fans, analysts, and anyone interested in understanding the financial aspects of NHL teams and player contracts.

	PLAYER	PLAYER.1	AGE	POS	TEAM	DATE	TYPE	EXTENSION	STRUCTURE	LENGTH	VALUE	CAP HIT
0	Jordan Gustafson	Jordan Gustafson	19	C	VGK	Nov. 29, 2023	ELC	NaN	2-way	3	\$2,850,000	\$950,000
1	Patrick Kane	Patrick Kane	34	RW	DET	Nov. 28, 2023	Std (UFA)	NaN	1-way	1	\$2,750,000	\$2,750,000
2	Justin Bailey	Justin Bailey	27	RW	SJS	Nov. 27, 2023	Std (UFA)	NaN	2-way	1	\$775,000	\$775,000
3	Ben Hemmerling	Ben Hemmerling	19	RW	VGK	Nov. 26, 2023	ELC	NaN	2-way	3	\$2,532,500	\$844,167
4	Samuel Laberge	Samuel Laberge	26	C	NJD	Nov. 25, 2023	Std (UFA)	NaN	2-way	1	\$775,000	\$775,000

Figure 1: CapFriendly Dataset Sample

The dataset used for this project is a representation of hockey player contracts, providing information such as player names, ages, positions, teams, contract dates, contract types, contract structures (e.g., Entry-Level Contract or Standard), contract lengths, total contract values, and annual salary cap hits. Each row corresponds to a different player’s contract, with details about their contractual terms and financial aspects. For example, it includes information on Entry-Level Contracts (ELC), Standard (UFA) contracts, two-way contracts, and the financial figures associated with each player’s deal.

For the player statistical information, we acquired data from a website called Hockey Reference. Hockey Reference serves as a statistical and historical resource for ice hockey. It provides a wide range of information on players, teams, and leagues, particularly focused on the NHL. On this site, users can find detailed statistics, player profiles, team records, and other data related to the sport.

	Rk	Player	Age	Tm	Pos	GP	G	A	PTS	+/-	...	S	S%	TOI	ATOI	BLK	HIT	FOW	FOL	FO%	SEASON
0	1	Justin Abdelkader	21	DET	LW	2	0	0	0	0	...	2	0.0	19	9:18	0	3	4	3	57.1	2009
1	2	Craig Adams	31	TOT	RW	45	2	5	7	-3	...	47	4.3	391	8:41	20	67	8	13	38.1	2009
2	2	Craig Adams	31	CHI	RW	36	2	4	6	-3	...	38	5.3	314	8:43	16	53	6	10	37.5	2009
3	2	Craig Adams	31	PIT	RW	9	0	1	1	0	...	9	0.0	77	8:34	4	14	2	3	40.0	2009
4	3	Maxim Afinogenov	29	BUF	RW	48	6	14	20	-7	...	93	6.5	605	12:36	11	20	0	3	0.0	2009

Figure 2: Hockey Reference Data set Sample

This dataset summarizes the performance of individual hockey players in different seasons from 2009 to 2023. It includes various statistics such as player details (name, age, position), team information, game-related data (games played, goals, assists, points), and additional performance metrics like plus-minus, shooting percentage, time on ice, blocks, hits, and faceoff statistics, etc. Each row corresponds to a specific player's performance in a particular season.

## Data Analysis / Method

For our data analysis section, we decided to implement two different approaches for safety purposes.

### Regular Approach

In the regular procedure, we integrated the source data using programming languages like Python and R and cleared the joint data based on our needs. Afterward, we implemented a MySQL database and designed a local API to easily access the database in the local environment. The API can be accessed globally if the API and the database get hosted.

Above is a sample code that we used to generate the contract table in MySQL using SQLAlchemy, and we did the same thing with the player statistics data. Next, we ran a series of machine-learning models in a local Jupyter notebook to predict the contract length and player salary. The machine learning models included are linear regression, random forest, support vector machine (SVM), gradient boosting, and ensemble method using a voting regressor. Among the various models, SVM performs best with a mean square error of around a million, which means we were able to predict the salary in the million range. Lastly, we created a dashboard using the model and tableau. The dashboard allows the user to examine the prediction and the actual contract value.

### Snowflake

Besides the regular procedure, we also decided to try a cloud-based storage approach by repeating the procedures in Snowflake. This time, we integrated the source by performing SQL queries in the Snowflake environment and created a dashboard based on the joint table.

The dashboard examines some relationships between the position variable and other important variables. In addition to the dashboard, we also created a Jupyter Notebook that loads data from the Snowflake warehouse and performs the machine methods on the same data.

```

class Contract(db.Model):
    __tablename__ = 'contracts'
    id = Column(String, primary_key=True)
    PLAYER = Column(String)
    AGE = Column(Integer)
    POS = Column(String)
    TEAM = Column(String)
    DATE = Column(Integer)
    TYPE = Column(String)
    EXTENSION = Column(Integer)
    STRUCTURE = Column(String)
    LENGTH = Column(Integer)
    VALUE = Column(Integer)
    CAP_HIT = Column(Integer)

    def __repr__(self):
        return '<Contract %r>' % self.PLAYER

```

Figure 3: Creating a contract table using Python

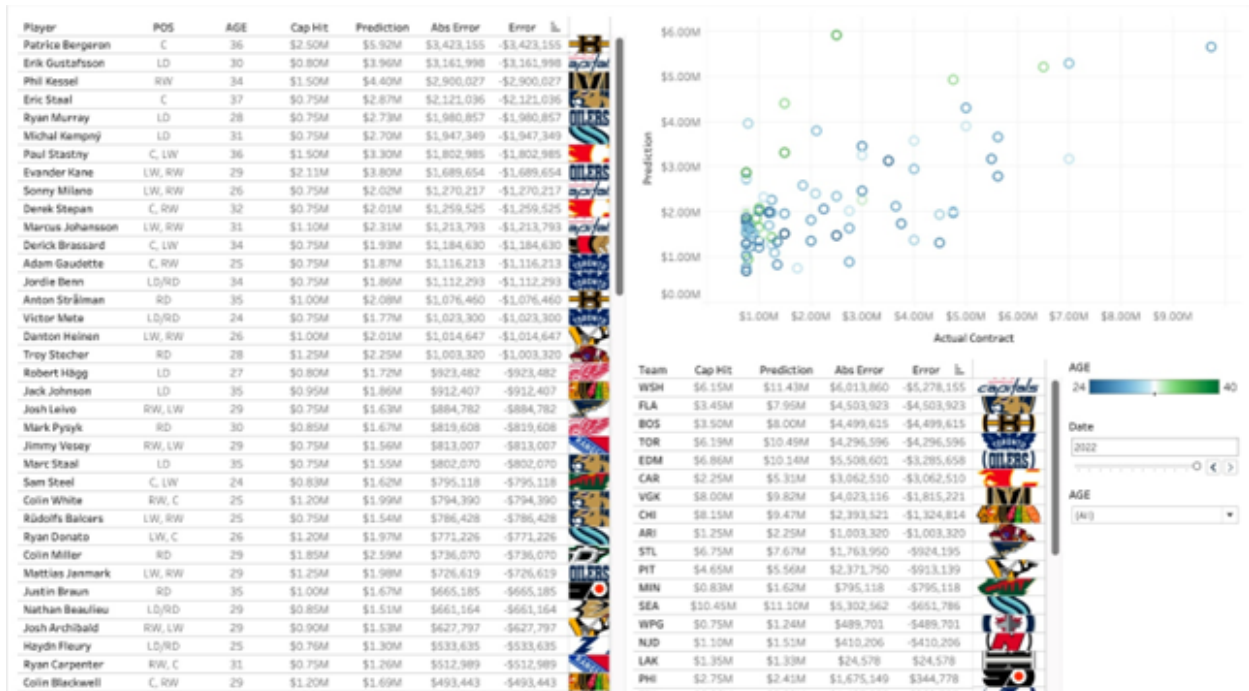


Figure 4: Tableau dashboards



## Conclusion

## Contribution

Eric Wu – Report writing, integration of the data using Snowflake and R, created Snowflake dashboard and data visualization.

Luke Kerwin – Scrape web data, build SQL database and API, and implement the model in Snowflake environments.

Brian Ellis – Integration of the data using Python, implementation of various machine learning methods, and created data visualizations on the result.

Griffin Jordan – Report writing

## Appendix

GitHub Repository - [https://github.com/lukekerwin/DS-320\\_Project](https://github.com/lukekerwin/DS-320_Project)