Marie Picini and Azalea Rohr
Professor Kevin Gold
DS110: Introduction to Data Science with Python
8 December, 2023

DS110 Final Project Paper

Introduction

Sports serve as entertainment for a large population, but there are naturally also physical consequences for the players that need to be addressed, as concussions account for up to 30% of hockey-related head injuries in the NHL. Throughout this project, we decided to pose the following question to shape our project; What factors affect the severity of concussions in the NHL and how do these factors affect each other? Within our data set, there were terms used that represented positions and other terms that we needed to define to understand the real world relationship with other factors in the NHL. To define the positions of the player, the abbreviations F, D, and G are used to stand for forward, defenseman, and goalie, respectively. Another term that is used is “cap hit”. In hockey, the term cap hit refers to the annual average salary of a player and is used for the total salary cap. The salary cap for each team is set by the NHL for the total amount of money that a team can spend on each player’s salary in a season. This cap hit is then used to calculate the player's impact on the team's total salary cap. So the player's cap hit is not only the salary that they receive in a single season, but the average annual value of the player's contract over the contract's duration. The average is calculated by taking the total value of the contract and dividing it by the number of years in the contract. The term CHIP is an abrasion for “Cap Hit of Injured Players” and represents the per-game charge of a player missing a game because of an injury or illness and then divided by 82. Another way to think of CHIP is the money that a team has on reserve to hire another player while the original player is out recovering from their injury. Within our data, the column called “Games Missed” is a number between 1 and 82 representing the number of games that a player had to miss because of his concussion. When the player had to miss more games, it can be concluded that their concussion was more severe; with 82 games missed (a whole season) as the most severe concussions and 1 game missed as the least severe concussion.

Methodology

Data Preparation

Before starting on statistical analysis and machine learning, we first had to clean the data. There was a duplicate column of the number of games missed that was not needed, so we removed that in a data table manipulation. We also dropped the column of the original data that tells us the type of injury, however since each injury was a concussion, that column was not needed. In the season column, there was some variation within how the seasons were named, so instead of simply dropping the column or renaming it, we had to create a function that sorted out the unique column names and matched them to the standard way of naming the seasons. To further manipulate the season column, it initially contained a year in the format "2013/14" (one example), and in order for it to be a numerical value so we could use it within our visualizations and random forest, we had to change it from a string and into a float. Also in 2013, the Phoenix Coyotes renamed their team to be the Arizona Coyotes. To account for this, we had to replace the any row with the name Phoenix for team for Arizona to merge these two sets of data.

Statistics

We performed a chi-squared test as our statistical analysis to see the relationships between various features and the number of games missed (the severity of the concussion). In total we ran 4 chi-squared tests; team and games missed, season and games missed, cap hit and chip, and position and games missed. (For the chi-squared test for cap hit and chip it had a value of almost zero, however, we already knew that cap hit and chip were directly correlated, so we omitted this value when deciding what two factors had a significant correlation.)

Machine learning

We ran a Random Forest Classifier because we wanted to see how these different features interacted with each other. We wanted to be able to perform a thorough analysis of the different factors in our data set, not just the concussions. We varied the parameters of our random forest three times and kept the number of trees to be 400. We randomly selected 90% of our data to train on and then 10% of our data to test on.

```
df_2 = pd.read_csv('NHL Concussions Database_data.csv')

df_data = df.copy()
df_data.drop("Team", inplace=True, axis=1)
#df_data.drop['Games Missed']    ##originally, we had Games Missed as our target,
                                 ##then we realized we needed a categorical feature as our target
df_data.drop("Position", inplace=True, axis=1)
df_data.drop("Player", inplace=True, axis=1)


from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

##First Model
features_train, features_test, target_train, target_test = \
  train_test_split(df_data, df_2['Player'], test_size=0.1)

forest_pla = RandomForestClassifier(n_estimators=400)
forest_pla.fit(features_train, target_train)

forest_pla.score(features_test, target_test)

print(df_data.columns)
print(forest_pla.feature_importances_)
print(forest_pla.feature_importances_.max())

##Second Model
features_train, features_test, target_train, target_test = \
  train_test_split(df_data, df_2['Position'], test_size=0.1)

forest_pos = RandomForestClassifier(n_estimators=400)
forest_pos.fit(features_train, target_train)

forest_pos.score(features_test, target_test)

print(df_data.columns)
print(forest_pos.feature_importances_)
print(forest_pos.feature_importances_.max())

##Third Model
features_train, features_test, target_train, target_test = \
  train_test_split(df_data, df_2['Team'], test_size=0.1)

forest_tea = RandomForestClassifier(n_estimators=400)
forest_tea.fit(features_train, target_train)

forest_tea.score(features_test, target_test)

print(df_data.columns)
print(forest_tea.feature_importances_)
print(forest_tea.feature_importances_.max())
```

Results

Statistics

For the other values of the chi-squared tests; team and games missed value was 0.6309, season and games missed value was 0.1037, cap hit and chip value was 3.169e-138 (we already knew these values were directly correlated), and position and games missed value was 0.6496. Based on these p values, even though none of them are statistically significant since no value is under 0.05, the chi squared value for season and games missed is the smallest value, so we are considering it the more significant with the chi-squared test.

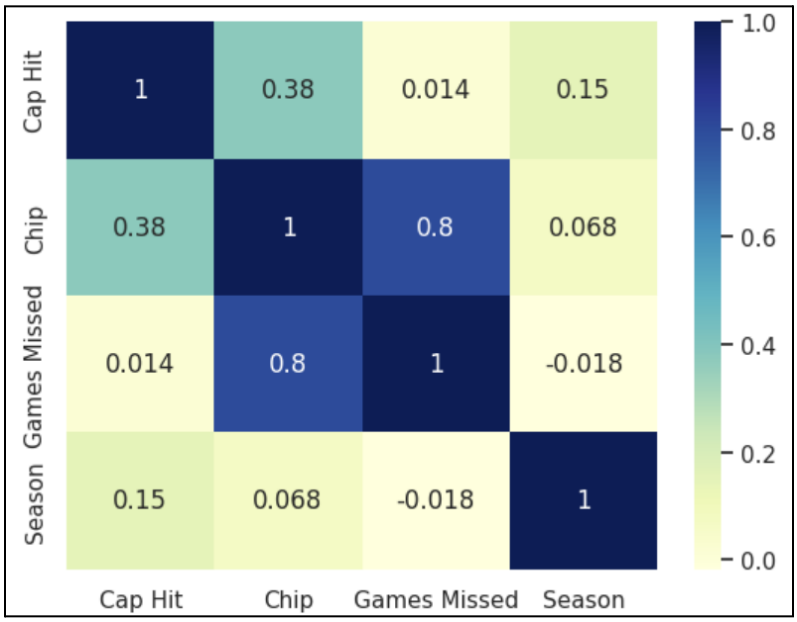
Machine learning

We found that the most influential feature with all three different targets was Cap Hit, so the amount someone gets paid is a good identifier of other features about them (Position, Player, Team). In our final model of the random forest, the highest classification value of 0.319.

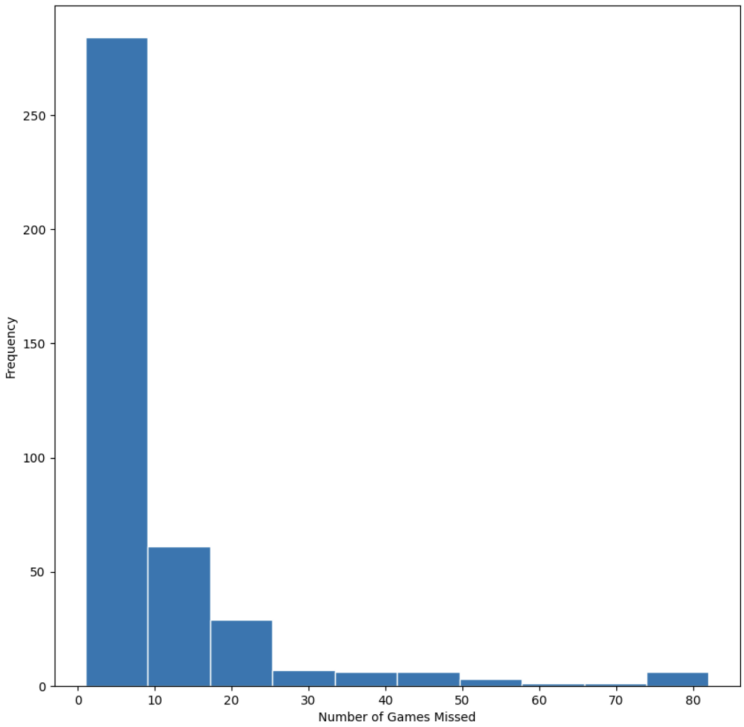
```
Index(['Cap Hit', 'Chip', 'Games Missed', 'Season'], dtype='object')
[0.33149524 0.29507199 0.19308525 0.18034753]
0.33149523765973943
Index(['Cap Hit', 'Chip', 'Games Missed', 'Season'], dtype='object')
[0.32945441 0.31211201 0.18008286 0.17835072]
0.32945441385953217
Index(['Cap Hit', 'Chip', 'Games Missed', 'Season'], dtype='object')
[0.31860906 0.29975389 0.19594199 0.18569507]
0.31860905703669373
```

Visualizations

For our first visualization, we used a heatmap to plot our correlation data since we wanted to see what 2 factors had the strongest correlation. As we can see from the heatmap, the 2 factors with the strongest correlation were Chip (the amount of money that a team has on reserve to hire another player while the original player is out recovering from their concussion) and

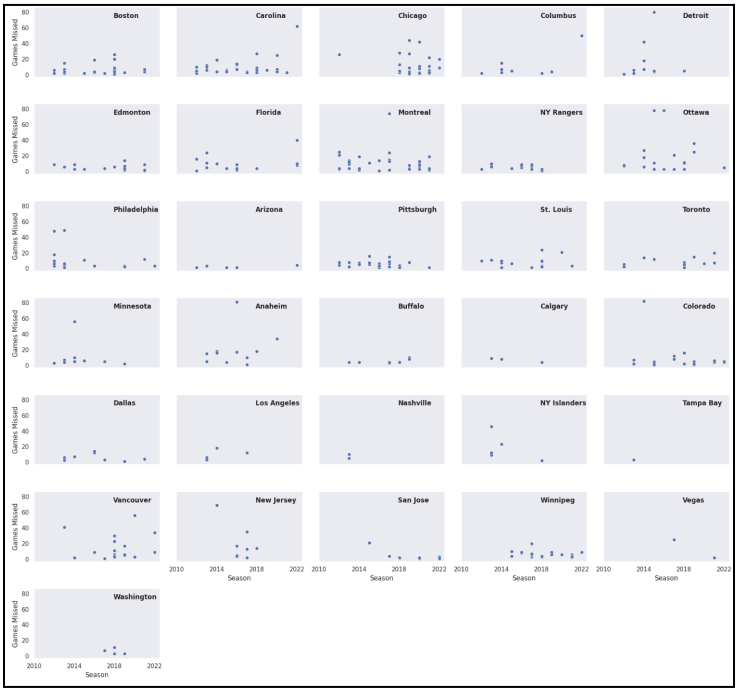


the amount of games missed (how severe the concussion was). We wanted to see how the factors from our database affected each other, and this heatmap helped us visualize that.



For our second visualization, we created a histogram that visualizes the number of games missed (severity of the concussion) and the frequency of the severity of the concussions. As we can see from the histogram, it is more common for players to miss 2-25 games (a mild to moderate concussion) than a severe concussion, however as you can see from the graph on the right, severe concussions still happen.

For our third visualization, we used a small multiple time series from seaborn to show the amount of games missed (severity of concussion) by season for each individual team. If one team had significantly more concussions (or higher severity of concussions) than the other, this visualization would show this.



Conclusion

As we saw, the only strong correlation between factors was between Games Missed and Chip and this can be explained by a number of reasons. One being that the more money a team has to spend on a player to replace the injured player, the better that player is and, thus, the less likely a team is to rush their injured player into recovery. Another being that, as we mentioned, CHIP is directly related to Cap Hit/Salary so the higher a player is paid the higher the CHIP, and the more valuable a player is, the more, or better, care he is going to receive from the team during his recovery, causing him to return later when he is 100% healed, because they don't want to risk reinjury in a very valuable player. In conclusion, we have learned about how different factors in the NHL affect each other, and based on our random forest and chi-squared test, we can conclude that the only significant relationship is the number of games missed and the chip for each team since they had the highest correlation value.