**LARGE SCALE ANALYTICS**

**Title: Predicting Outcome for NHL Matches**

Prof: Magdalini Erinaki

**San Jose State University**

Fall 2018

**Team Members:**

Suhani Vaishnav 012489191

Mitesh Kumar 011524604

Sahana Alliyandiru Jayasheela 012416131

Canvas group : NHL Gamblers

# Introduction

Sports outcome prediction has been one of the most researched topic for Analysts. Whether its playoffs or league matches, match outcome prediction based on season’s data has been the subject of interest for audience, team coaches and players alike. Season matches of NHL (Ice Hockey) have exhaustive data which appealed to the team. Besides the availability of data, NHL matches are less goal scoring matches. Based on Nielsen analysis, NHL gambling business alone worth is more than $2 billion per year. NHL’s fame in North America has made this industry and its associated organizations a very profitable business such as Fanduel.

DATA mining is already prevalent in many sports and has shown to be extremely valuable to sports teams, fans and gambling organizations. The objective here has been to collect NHL data, research on the previous work done in Sports Analytics and come up with innovative ideas to predict winner of a NHL game. This includes contribution of different statistics like team’s overall shots, blocked shots as well as players position and their contribution . Several sports analytics groups have been working on either game statistics or player statistics to predict match outcome. Our approach has been to build data mining models by combining game, player and their position statistics in a game. Different data pre-processing techniques as well as regression and classification techniques were tried to come up with a robust model which can give accurate prediction in NHL.

A fact confirmed by Ottawa, who note that “As far as the authors are aware, there is no previous work, in the machine learning community, to predict the winner in a hockey game.” [1]. This is surprising given the fact that NHL is popular in US and Canada. Moreover, NHL has an abundance of data on previous year’s playoffs as well as league matches. The wealth of data was our motivation to work on NHL match outcome prediction.   
  
Using ML and Data Mining techniques to predict match outcome in other sports like basketball, football and Soccer has been the focus of many research groups over the past decade. And the research outcome has been fruitful. We hope that our work on NHL will try to bridge the gap of interest and research in NHL.

# System Design and Implementation

## System Design:

The objective of this project was to predict the outcome of NHL games based on various factors. In the dataset, there were multiple files with statistics for each game, team and player. It was decided to make use of these different datasets to better predict the win or loss of the game. Three approaches were followed to analyze effect of different factors on the prediction.

The approaches are as follows:

1. **Team based approach (1):** Here only the team stats table was considered for analysis which had data for each team’s goals, shots, hits, etc.
2. **Player based approach (2):** Here, the team stats table was merged with the player stats in such a way that for each game, there was information about which player played and which did not.
3. **Player and Position based approach (3**): Here, along with the player and team stats, the player’s position and plus-minus score were also added for predictions.

The flow of the project was as follows:

Choosing the right Dataset

Approach1 : Team stats

Approach2 : Team stats and player-skater stats

Approach 3: Team stats, player\_skater and player info

Data Preparation and Exploration

Joining the datasets as per requirements

Removing categorical columns

Applying Dimensionaltiy reduction

Building the Model for prediction

Random Forest

AdaBoost

XGboost

Logistic Regression

GaussianNB

KNN

Testing the model

Accuracy

F1 Score

To benchmark, various classification algorithms were used. They are discussed in the next section.

## Algorithms Used

For comparison, dimensionality reduction in the form of Principal Component Analysis (PCA) was applied to understand its impact on prediction. Following that, the below mentioned algorithms were applied for Classification:

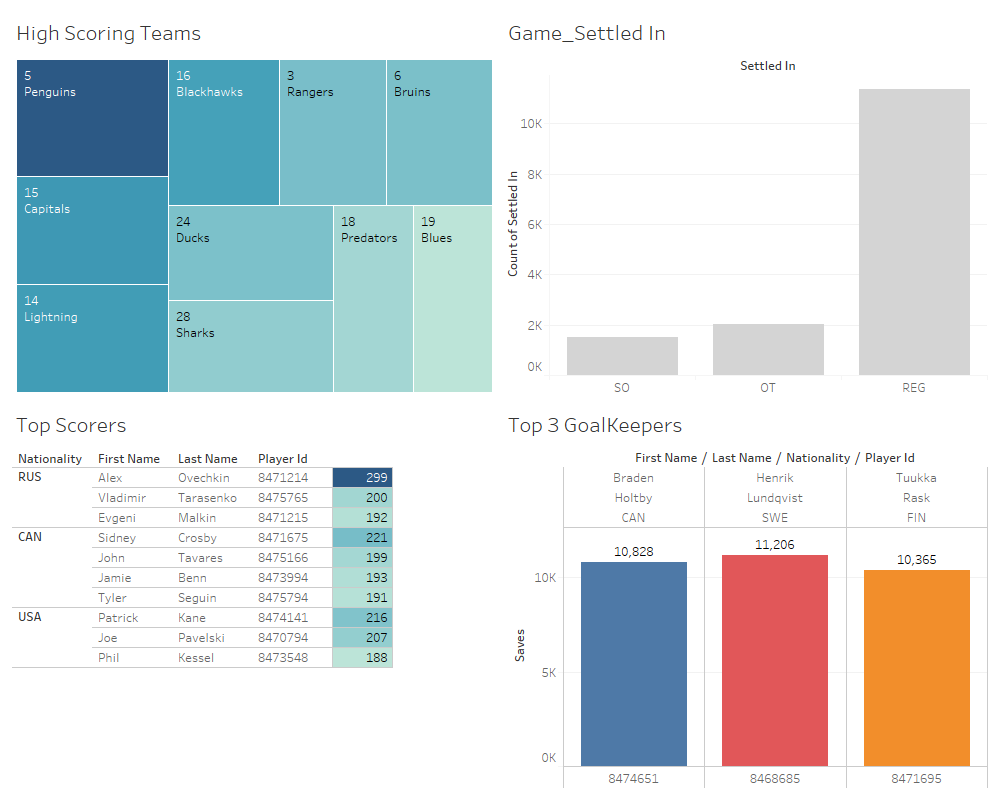
1. **Random Forest:** It is a simple and easy to use algorithm which gives good prediction results even without hyper tuning. Since its splits on the node that is the best feature amongst random set of features, it can be used to extract the feature importance’s. For this project, the outcome of feature importance helped understand which player majorly contributed towards the win.
2. **AdaBoost with Random Forest:** AdaBoost helps boost the performance by converting multiple weak learners into a single strong learner. It gives better accuracy by using multiple instances of same classifier with different parameters, hence this was chosen to work along with the random forest.
3. **Logistic Regression**: This algorithm is widely used when the target variable is binary, like in the case of this project. Since all the attributes of the train set are contributing towards the target variable, this algorithm helps build a mathematical equation that predicts the probability that the *Y*-variable takes on a value of 1. The algorithm helps in tracing the effect of variation in the measurement variable on the target variable. Since that is in line with the project idea, this was selected.
4. **XGBoost**: This is the one of the most popularly used algorithm. It is high speed- high performance model. This was selected since it is known to be efficient, accurate and feasible.

## Tools

Anaconda - Jupyter Notebook, MS Excel, Tableau

## Dashboard

To convey an idea of what information could be harnessed from the data, a Dashboard using Tableau was created.



# Experiments/ Proof of Concept Evaluation

### Dataset:

The NHL game dataset was obtained from Kaggle (<https://www.kaggle.com/martinellis/nhl-game-data>). It contains 9 CSV files. The total size of the dataset is 120MB. The table below contain details pertaining to number of instances for each attribute in respective files.

|  |  |  |
| --- | --- | --- |
| CSV FILE NAMES | #Rows | #Columns |
| game\_skater\_stats.csv | 267854 | 22 |
| game\_shifts.csv | 5564594 | 5 |
| game\_goalie\_stats.csv | 16062 | 19 |
| game\_plays.csv | 2369646 | 20 |
| game\_teams\_stats.csv | 14882 | 15 |
| player\_info.csv | 2246 | 8 |
| game\_plays\_players.csv | 3479482 | 5 |
| team\_info.csv | 33 | 6 |
| game.csv | 7441 | 15 |

The source also provided the relationship between each table. It is shown in the figure below:

A close up of a map

Description automatically generated

Tables used for analysis and visualizations:

* Game\_team\_stats
* Game\_skater\_stats
* Player\_info
* Team\_info

### Data Preprocessing

There were three separate approaches followed to achieve the goal of outcome prediction. Despite this, one dataset “ Team\_stats “was common for all. The preprocessing steps applied were as follows:

1. **One-hot Encoding**: This was implemented for the columns:
   * HoA : Home or Away – two separate columns were made to denote if the team was playing at home or away
   * Settled\_In : Three separate columns were made to share if the game ended in Regular time, Over time or ShootOut.
2. **Label Binarizer:** The target label “ Won” was converted to numerical attribute by labelling its values of True and False to 1 and 0 respectively.
3. Since the data had match statistics, there were no missing values and the zeroes were not tampered with.
4. **Merging**: Post the steps mentioned above, the final dataframe for each approach was

### Methodology followed

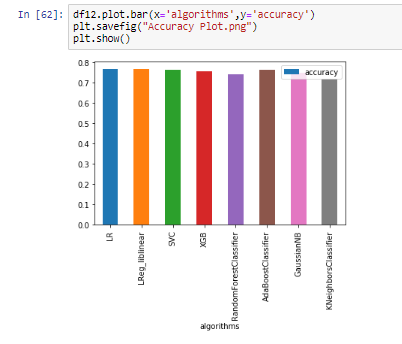
* The data, once preprocessed was split into test-train ratio of 20:80.
* Cross-fold validation was implemented with the k = 10 (as number of folds) for the output of the top 3 classifiers.

### Results & Evaluation

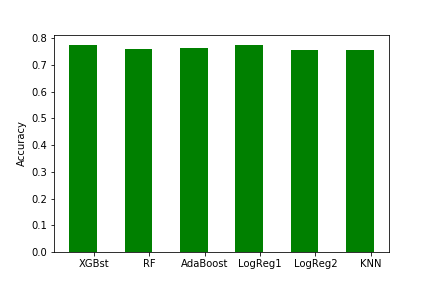
### The results are tabulated below for each approach:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Approach 1** | | **Approach 2** | | **Approach 3** | |
|  | **Without PCA** | **With PCA** | **Without PCA** | **With PCA** | **Without PCA** | **With PCA** |
|  |  |  |  |  |  |  |
| Random Forest | 0.74 | 0.59 | 0.78 | 0.68 | 0.99 | 0.99 |
| AdaBoost | 0.76 | 0.64 | 0.76 | 0.69 | 0.90 | 0.91 |
|  |  |  |  |  |  |  |
| Logistic Regression – solver ( lbfgs) | 0.77 | 0.58 | 0.76 | 0.76 | 0.90 | 0.89 |
| Logistic Regression - solver (liblinear) | 0.78 | 0.64 | 0.76 | 0.76 | 0.89 | 0.89 |
| KNNeighbours Classifier | 0.72 | .71 | 0.76 | 0.76 | 0.91 | 0.89 |
| XGBoost | 0.76 | 0.64 | 0.78 | 0.77 | 0.91 | 0.92 |

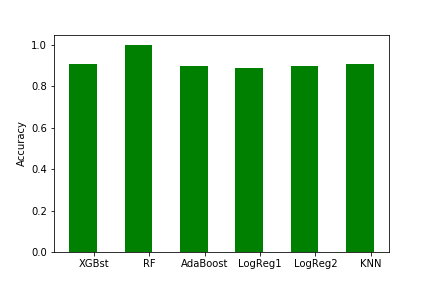
Accuracy Plots: Approach 1



Approach 2:



Approach 3:



### Analysis:

As observed from the experimental results (table no), the basic approach of just considering the team statistics yields an accuracy of 60%. Upon addition of the player statistics, the accuracy improves to be around 78%. Along with improvement in accuracy, using features importance, list of few players contributing majorly towards the win is also obtained. Further, the accuracy greatly improves once the player’s plus minus scores along with their positions are taken into consideration for the analysis.

# Discussion & Conclusion

Looking at the volume of data, it was decided to include as many attributes as possible from the available data. This led to dividing the methodology into three different approaches, which later reconfirmed our belief that more relevant data can lead to better predictions.

One idea was to first predict the goals and then develop a new metric called Adjusted Plus Minus to rank the players. This metric required data for each shift of the play and players on ice. It was observed that the data was not sufficient to do the latter part. Hence, the idea of using player statistics extended to predict the outcome of the match. This also involved using the existing plus-minus scores of the players.

The solution of using player statistics and following different approaches worked in our favor as we were able to merge the datasets and obtain good accuracy scores for each approach.

The aim of the project was to predict accurately the outcome of NHL games. The model built can be used to further predict the outcome of upcoming matches in the series. Along with it, it is also able to highlight few top -performing players. The model can be enhanced to improve the player’s plus minus metrics as well as formally obtain each player’s contribution towards the outcome. This can help the coaches finalize players for each game.

## Project Plan

The project had three different approaches towards the same goal of predicting the outcome of NHL game. The tasks were equally distributed amongst the team and the distribution is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Student name:  SJSU ID: | Suhani Vaishnav  012489191 | Mitesh Kumar  012489191 | Sahana  012466116 |
| Deciding the approach | Jointly done  Approach (2) | Jointly done  Approach (3) | Jointly done  Approach (1) |
|  |  |  |  |
| Data Preparation | Individually done – depending upon the approach | Individually done – depending upon the approach | Individually done – depending upon the approach |
|  |  |  |  |
| Dimensionality Reduction | PCA | PCA | PCA |
|  |
| Classifiers | Random Forest  XGBoost  Logistic Regression  AdaBoost | Random Forest  XGBoost  Logistic Regression  AdaBoost | Random Forest  XGBoost  Logistic Regression  AdaBoost |
|  |  |  |  |
| Report & Visualization | Jointly done  (Tableau & Matplotlib) | Jointly done  (Matplotlib) | Jointly done  (Matplotlib) |

**References:**

1. http://web.uvic.ca/~afyshe/dm\_projs/nhl\_final\_report.pdf

2. <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

3. Macdonald, Brian. "Adjusted plus-minus for nhl players using ridge regression with goals, shots, fenwick, and corsi." *Journal of Quantitative Analysis in Sports* 8, no. 3 (2012).

4. Heal, Brendan, Meara Kimball, Renee Fung, and Abdelrahman Alenazi. "Data Mining Project Final Report NHL Playoff Prediction."

5. Pischedda, Gianni. "Predicting NHL match outcomes with ML models." *International Journal of Computer Applications* 101, no. 9 (2014).