Prediction of NHL Rookie Salary with ML Models

Chandler Brooks John Carmack Lillian Coar Brian Horsburg

***Abstract*— NHL player salaries are capped by a set amount per team. It is then imperative that recruiting managers and talent scouts determine the appropriate salary to offer new recruits. This can be difficult given the subjective nature of 'scoring' candidate recruits. One way to do so is to look at how the NHL at large is allocating their salary budget. In this paper several machine learning models designed to predict the salary of NHL players given some common stats among the players.**

I. Introduction

II. Data Discussion

The data acquired for this report was collected from a popular dataset sharing resource, Kaggle.com. The data consists of 874 samples of rookie NHL player salaries and their corresponding statistics. Each sample contains 153 features. These features contain a mixture of categorical and numeric data.

Initially the dataset was broken into multiple files, one purpose-split for training and another for testing. The two files were conglomerated into a single file such that it could be split again using a standard ratio of 80% training data and 20% testing data.

The author of the data set conveyed that the data set was incomplete, and some features were missing for some of the represented samples. To remedy this, mean imputation was applied to the data set such that all missing features were replaced with the mean.

After doing this the data was then visualized to gain some intuition behind the features. A histogram of the salary values was produced and is presented below. It demonstrates a non-normal distribution of salaries heavily weighted towards smaller values.

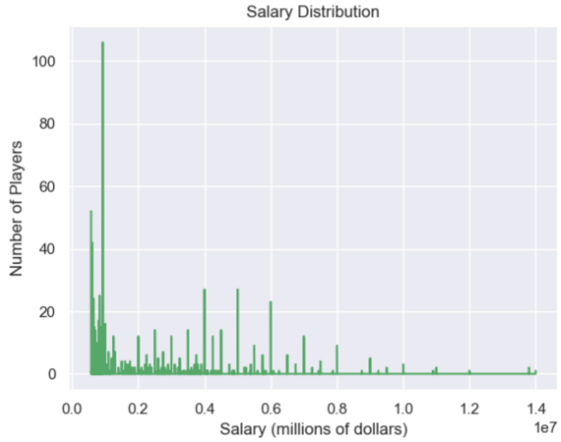

Figure 1 - Histogram of Salaries

Plots were also produced to determine the correlation between select features and salary. The features selected were ones believed to be correlated with athletic performance, such as points per game and overall draft pick.

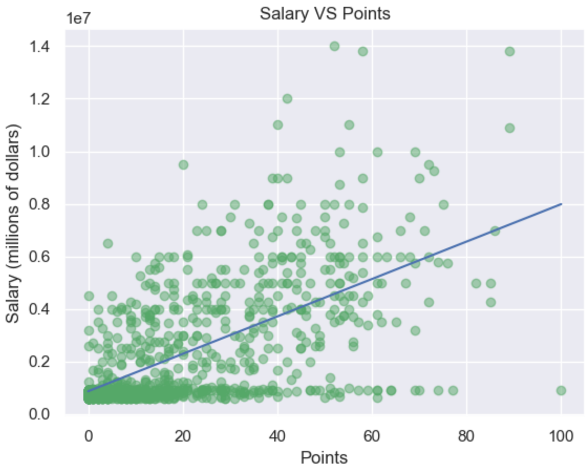

Figure 2 Salary vs. Points Per Game

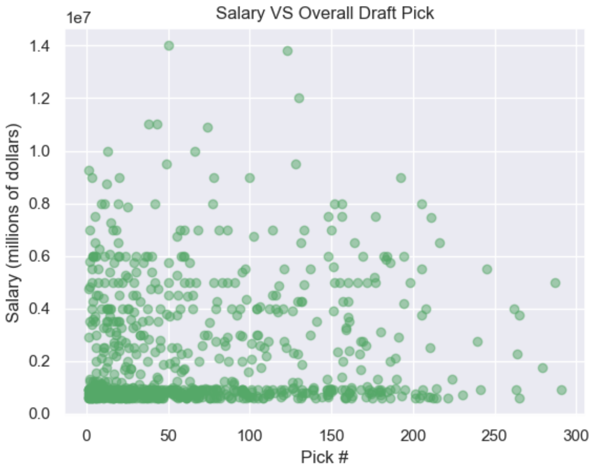

Figure 3 - Salary vs Overall Draft Pick

Of these two features, only points per game showed a marginal correlation to salary, but one is present. It was then decided that some features must be highly correlated. With this in mind, and considering the relatively high dimensionality of the data, the decision was made to perform some measure of feature selection.

Firstly, all non-numeric features were stripped from the data. This reduced the dimensionally from 153 features to 144. Highly correlated features (>90% correlated) were compared and trimmed stochastically. This further reduced the number of features to 74. From there, the number of features was reduced even further by only selecting the 30 features most closely correlated with salary.

III. Model Discussion

IV. Results