CS424 Fall 2021
Team 3 - Step 3

John Carmack
Brian Horsburgh
Lillian Coar
Chandler Brooks

Salary Predictions of NHL Players in Rookie Season by Numerical Data

Team 3's project aims to predict rookie NHL player salary by performing a linear regression on a dataset obtained through the big data sharing site Kaggle. The data set collected consists of 874 samples of rookie players and their salary. Each data point contains 153 features. The features are a mixture of discrete, categorical, and continuous values. This report will detail the initial analysis of the data, the cleaning methods that have been considered and applied to the data, and some preliminary model exploration with various linear regression techniques.

I. Data analysis

The first question asked was: what is the distribution of salaries? To answer this question the distribution of salaries was plotted and is presented below. From there an analysis of the features was conducted to determine if data could be cleaned intuitively. Features such as points scored and overall draft pick seem like sensible predictors of salary, so a plot of salary as a function of both was produced.

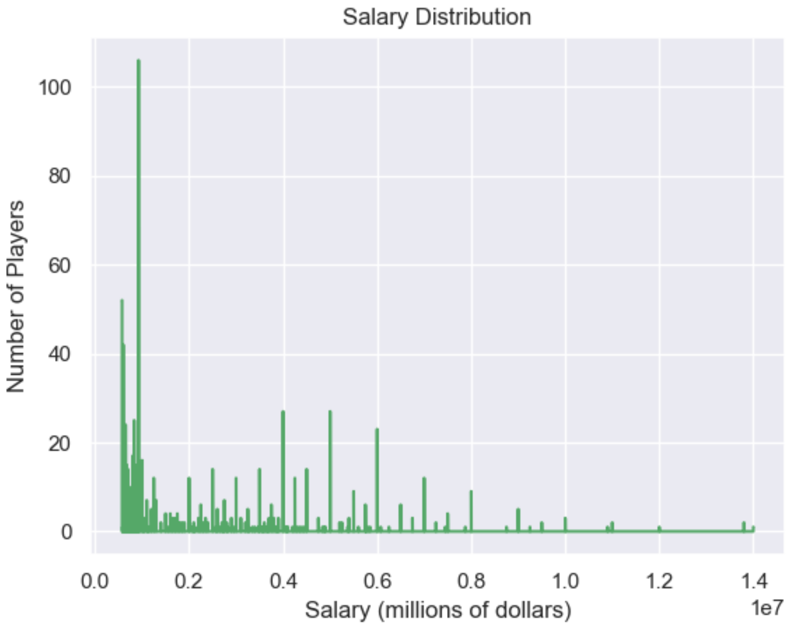

Figure 1 - Salary Distributions for NHL

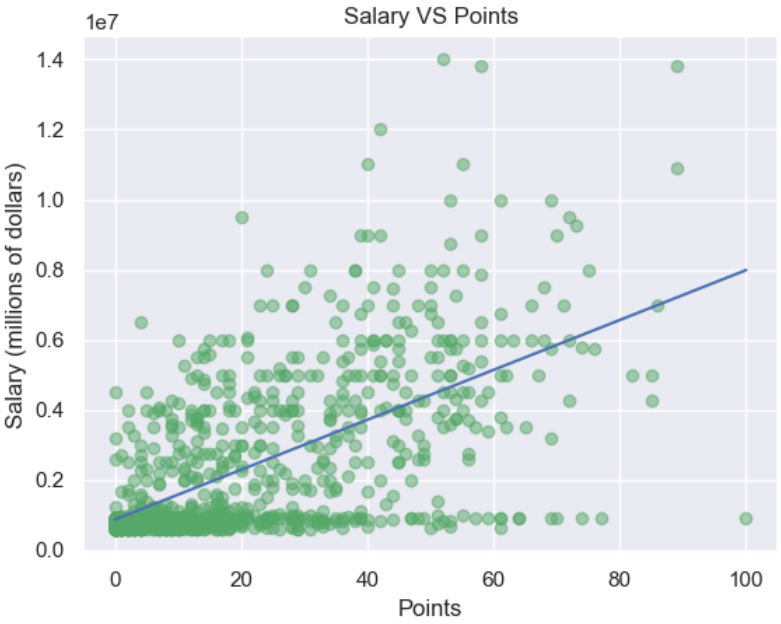

Figure 2 - Salary as a Function of Points Scored

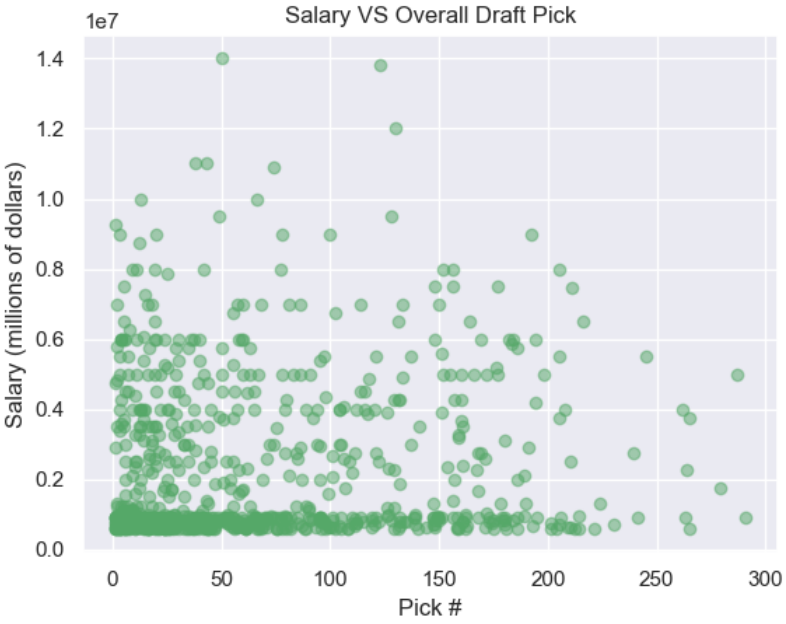

Figure 3 - Salary as a Function of Overall Draft Pick Number

As illustrated by the above plots, there is a high degree of variance in salaries, and intuitive predictors of salary like overall draft pick and points scored are correlated with salary in the expected direction. A higher number of points score corresponds to a higher salary in general, and a lower draft pick seems to correlate to a higher salary as well.

II. Data cleaning

As this data set contains many features, and many of those features are categorical, the first step in cleaning the data was to remove all non-numerical features. This reduced the number of features from 153 to 144. The second stage of cleaning was to find all pairs of features that were correlated within some threshold, and discard one. The data was highly correlated, so a threshold of .9 was selected and further reduced the number of features to 74. This threshold value will most likely be used as a hyperparameter when model selection is finalized. The final step in cleaning was to select the 20 features of the remaining 74 that are most highly correlated with salary. From there, exploratory model selection was performed.

III. Model selection

Team 3 has decided to employ a linear regression model to predict salary, but there are many to choose from in the sklearn data science library. As the features of the data are highly correlated, the models selected for preliminary testing all employ some measure of regularization to combat this. The models selected for testing were Lasso, Ridge, and ElasticNet, which are implemented in the sklearn library. A GridSearchCV was performed with alpha values of 0.001, 0.01, 0.1, 1, 10, 100, and 1000. The GridSearchCV was scored on R-Squared value. All initial regression models performed poorly to the same degree with Ridge regression performing the best with an R-Squared of 39%. Both ElasticNet and Lasso regression produced R-Squares of 38%.

These figures point to an over-fitting of the data most likely explained by the cleaning methods employed.

IV. Future Work

Clearly the data cleaning and model selection methods currently being employed require revision. Values such as the correlation threshold and number of features selected will play a large role in the success of our models to predict salaries.

Model selection will also require further research. Sklearn’s linear regression models offer little in the way of hyper-parameter adjustment so perhaps a deeper dive into what’s possible with them is required.