University of Michigan

Major Heading

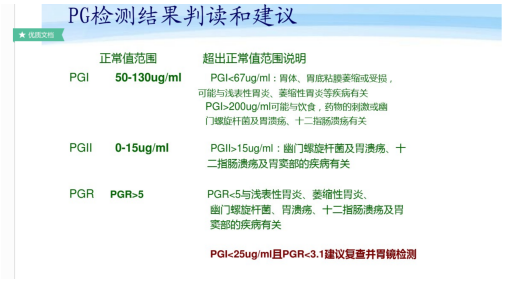
Minor Heading

Title

Author:
Zhen Qin

Supervisor:
Dr. James Smith

March 11, 2018



Abstract

Your abstract.

1 Introduction

Your introduction goes here! Some examples of commonly used commands and features are listed below, to help you get started.

If you have a question, please use the support box in the bottom right of the screen to get in touch.

2 Data Exploration

This dataset features the salaries of 874 NHL (National Hockey League) players for the 2016/2017 season. The players are randomly split into a training and test populations. There are 151 predictor columns as well as a leading column with the players 2016/2017 annual salary. The league has a salary cap, meaning that each team is limited in the amount of money it can spend on players. It is worthwhile to note that the dollar figure that counts toward the salary cap isn't the actual amount of money a player makes in a season, it is the average of their yearly compensation over the length of their contract. After removing missing values, the sample sizes of trainning set and test set are 491 and 227.

2.1 Dimension Reduction

In order to avoid the effects of the curse of dimensionality, firstly dimension reduction is performed prior to applying machine learning algorithms. Thanks to the fact that Salary is numerical, some robust regression models are available for variable selection. Both feature selection and feature extraction methods are used to select numerical variables in this report.

In the data set, there are 11 categorical variables. After removing some irrelevant variables (for example, last name and first name), Hand (L or R) and Country (USA, CAN or others) are chosen to be involved in the the analysis framework.

Variable	Description	Type
Salary	The player's salary	numerical
DftYr	Year drafted	numerical
DftRd	Round in which the player was drafted	numerical
G	Goals	numerical
A1	First assists, primary assists	numerical
A2	Second assists, secondary assists	numerical
PTS	Points. Goals plus all assists	numerical
TOIX	Time on ice in minutes	numerical
iFF	Unblocked shot attempts taken by this individual	numerical
iHDf	The difference in hits thrown by this individual minus those taken	numerical
iTKA	Takeaways by this individual	numerical
iFOW	Faceoff wins by this individual	numerical
dzFOW	Faceoffs win in the defensive zone	numerical
CA	Shot attempts allowed while this player was on the ice	numerical
FA	Unblocked shot attempts allowed while this player was on the ice	numerical
HF	The team's hits thrown while this player was on the ice	numerical
PS	Point shares, a stats that measures contributions in points	numerical
OTOI	The amount of time this player was not on the ice	numerical
Hand	Handedness	categorical
Cntry	Country of birth	categorical

Table 1: Selected Variables

2.1.1 Model Selection - Lasso

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. This method shrinks many coefficients to exactly 0, cross validation to choose the parameter t. Under this condition, 66 variables have coefficients greater than 0. This result is good but not satisfactory. Aiming to keep less variables, we may adjust the parameter to increase the penalty, or use criterion methods.

2.1.2 Model Selection - BIC

Bayes information criterion (BIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based,

in part, on the likelihood function. Comparing with AIC, BIC has better performance with respect to prediction tasks. The picked model minimizes BIC, which is 13950.33. After stepwise regression, 18 variables including salary are kept in the final model.

The result is relatively good, however, model selection methods do not always benefit interpretability. In addition to regression methods, we require the variables must seem to contain most information in terms of definition. Due to each variables' description, iHDF (The difference in hits thrown by this individual minus those taken) has more information than a selected variable, iHF (Hits thrown by this individual), so iHF is replaced by iHDF . Table 1 shows the selected variables, their description and type.

2.1.3 PCA

2.2 Comments

Comments can be added to the margins of the document using the, as shown in the example on the right. You can also add inline comments too:

2.3 Tables and Figures

Use the table and tabular commands for basic tables — see Table 1, for example. You can upload a figure (JPEG, PNG or PDF) using the files menu. To include it in your document, use the includegraphics command as in the code for Figure **??** below.

2.4 Mathematics

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

2.5 Lists

You can make lists with automatic numbering ...

1. Like this,
2. and like this.

... or bullet points ...

- Like this,
- and like this.

We hope you find writeLaTeX useful, and please let us know if you have any feedback using the help menu above.

References

[1] Fred G. Martin *Robotics Explorations: A Hands-On Introduction to Engineering*. New Jersey: Prentice Hall.

[2] Flueck, Alexander J. 2005. *ECE 100*[online]. Chicago: Illinois Institute of Technology, Electrical and Computer Engineering Department, 2005 [cited 30 August 2005]. Available from World Wide Web: (http://www.ece.iit.edu/ flueck/ece100).