Project Report on

Analysis of back pain using machine learning classification algorithms

Submitted by

Karthik Krishna Kamath

April 2019

ABSTRACT

Lower back is interconnected with spine , joints , nerves, ligaments or tendons and muscles. Lower back pain occurs due to various reason related to these body parts. The first step in the treatment of lower back pain is knowing the exact reasons and origin of the pain caused for the better diagnostic treatment. This project deals with the study of symptoms recorded by patients in causing the backpain and classifying them into 2 categories of backpain. The data set contains the description of symptoms and conditions of 380 patients with their records summarized in 32 variables. In the data analysis, pain is categorized as Nociceptive and Neuropathic according to the symptoms/pains in the lower back recorded by patients. The aim the project is to study the data set and classify the recording in the basis of the backpain. Few classification techniques such as Classification tree, Random Forest, Logistic regression and Support Vector Machine algorithms are used for producing these results. Finally, the best prediction model is discussed and used for the prediction of the backpain of the patients.

Table of Contents

1. INTRODUCTION

A data set consists of details of few patients suffering from lower back pain are used in this project and different classification algorithms are used to classify them according to the type of back pain into 2 categories as “Nociceptive” and “Neuropathic”. The aim of the project is to find the best classifier model that classifies the patients according to the type of the lower back pain they experience.

1.1. Data set

The data set contains 380 observations which consists the details of patients who are suffering from back pain. Their details are divided into numerical and categorical clinical indicators and are recorded in 32 variables including the response variable as “PainDiagnosis”. The classification techniques in machine learning used in this project are as below.

- Classification Tree
- Logistical Regression
- Random Forest
- Support Vector Machine
- Bagging

From the above listed techniques, the best performing machine learning model is chosen and used for the classification of the lower back pain into 2 mentioned categories.

Table 1: Details of variable in the given dataset

• *PainDiagnosis: Expert pain diagnosis (as reference standard)* • *Age : Age of the patient.* • *Gender : Gender of the patient* • *DurationCurrent : Duration of current pain.* • *PainLocation: Pain location.* • *SurityRating : Surity rating of expert clinical diagnosis.* • *RMDQ : Roland Morris Disability Questionnaire score.* • *vNRS : Verbal NRS for pain intensity.* • *SF36PCS : SF36 Physical Component Summary score.* • *SF36MCS : SF36 Mental Component Summary score.* • *PF : Physical Functioning* • *BP : Bodily Pain.* • *GH : General Health.* • *VT : Vitality.*	• *MH : Mental Health.* • *HADSAnx : HADS Anxiety score.* • *HADSDep : HADS Depression score.* • *Criterion2 : Pain assoc'd trauma, pathology, movt.* • *Criterion4 : Pain disproportionate to injury, pathology.* • *Criterion6 : More constant, unremitting.* • *Criterion7 : Burning, shooting, sharp, electric shock like.* • *Criterion8 : Localised to area of injury, dysfunction.* • *Criterion9 : Referred in dermatomal, cutaneous distribution.* • *Criterion10 : Widespread, non-anatomical distribution.* • *Criterion13 : Disproportionate, non-mechanical pattern to aggs + eases.* • *Criterion19 : Night pain, disturbed sleep.* • *Criterion20 : Responsive to simple analgesia, NSAIDS.* • *Criterion26 : Pain with high levels of functional disability.* • *Criterion28 : Consistent, proportionate pain reproduction on mechanical testing.* • *Criterion32 : Localised pain on palpation.*

In the above mentioned variables, there are 19 categorical variables and rest are numerical variables.

2. METHODS

The performance analysis of different machine learning models in classifying the type of back pain is done using R software. Libraries containing respective machine learning algorithms were also added to the workspace.

2.1. Data Preprocessing

Data preprocessing is done to make the data clean from a messy and unreadable format by removing the missing values and transforms the data more organized and readable for the downstream analysis. This preprocessing step is crucial before the model fitting process for obtaining a good fit.

The back pain data set doesn't have any missing values and half of the preprocessing steps can be avoided because of this. Out of the 13 continuous variables 1 variable "SurityRating" is a numerical categorical variable having 10 factors valued from 1 to 10 and can be converted to a factor.

2.2. Machine Learning Models

After the data is processed to a readable form, the models for performing classification of lower back pain is created using the train data set. From the total data , we split the data into train data, validation data and test data, respectively. Bootstrapping is performed on train and validation data set to increase the accuracy of the model.

The different classifier models are explained in the below section.

2.2.1 Classification Trees

Classification Trees are mostly used to predict the categories or objects in an ordinal/ categorical variable. These techniques analyses the data and a set of rules are constructed with a threshold value and these rules are then used to predict the class of the response variable. The model builds the tree like structure and it summarizes results in a simple manner so that the new observations are classified efficiently with its simpler model.

The library used in R is "rpart" and the function used for model fitting is ***rpart.***

2.2.2 Logistical Regression

Logistical regression is another technique to analyze the relationship between the predictor variables with response variable which has an outcome of only 2 possibilities. The logistic function is used to build the model for classifying the 2 outcomes(either binary , 2 numerical values or 2 category class such as "Absent" or "Present") of dependent variable.The aim of this technique is to find the best fitting model that describes the relationship of the categories of response variable. The general equation is given below,

In the above equation (Eq(1)) beta (β) values are the coefficients of regression and are estimated by the model by using the training data and X_1 *to* X_n are the predictor variables. The model finds the log odds of the probability of the default class. Odds ratio is the ratio of the probability of the event

occurring to the probability of the same event not occurring. The best model is the one which finds the coefficients that gives a value closer to 1 for default category/class and a value very close to 0 for the other category/class.
The R library used in this project for performing logistic regression is "nnet". The ***multinorm*** function is used to fit the logistic model in R.

2.2.3 Random Forest

Random Forest is an upgraded decision tree technique which uses ensemble learning method in machine learning technique for improving the performance of the model such as accuracy and stability. This algorithm generates random subsets from the dataset for finding combination of multiple Classification and Regression Tree(CART) models. This model technique follows the working of bagging method with an additional method added to it. Several CART model are randomly formed with subsamples from the train dataset. This method decides where to split the decision tree based on the random selection of features.

This level of differentiation provides the model to produce results of high accuracy and stability. The library used in R for running the Random Forest model is "***randomForest***" and the same is used to build the random forest model.

2.2.4 Support Vector Machine

Support Vector Machine (SVM) technique is a popular method for classification problems, however it can also be used in regression analysis.

SVM plots each data as points in N- dimensional space (N – number of features) such that these points are distinctly classified. The value of each feature becomes the coordinate of these data points. The figure(Fig 2.2.4 a) shows the concept of SVM. In the figure data points are grouped as 2 classes and separated by line/hyper-plane (multidimensional space that separates 2 classes). Support vectors are data points lying close hyper-plane/line. A few transformations called kernels are used to produce hyperplane which separates the data points with a maximum margin.

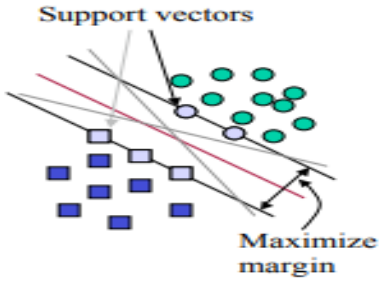


Figure 1: SVM Classifier logic diagram

When new data points are introduced to the model it tries to predict the class with its features. The R library used for SVM model is "kernlab". The R function used for fitting the model for SVM classifier was ***ksvm.***

2.2.5 Bagging Method

Bagging is another ensemble method in machine learning algorithms which is also called as bootstrap aggregation Bagging uses different bootstrap samples to train N different decision tree models. Finally, after forming the tree models an algorithm is used to compute the aggregate of the values from these decision tree models.

The final method uses voting method for classification technique and averaging technique for regression models. The R library used for bagging is "adabag" and the function for fitting the model is ***bagging.***

2.2.6 Boosting Method

Boosting is also another ensemble method which is used for improving the accuracy of the model prediction. The model analyses the misclassified data and tries to improve the accuracy by refitting the model by giving importance to the weak classified data. This process is carried out sequentially to attain maximum performance.

The function used in R to perform model fitting is ***boosting*** from the library adaboost.

2.3. Process performed in R

The below figure shows the steps followed in R with the backpain data set to analyze which classification model was successful in categorizing the type of back pain in patients.

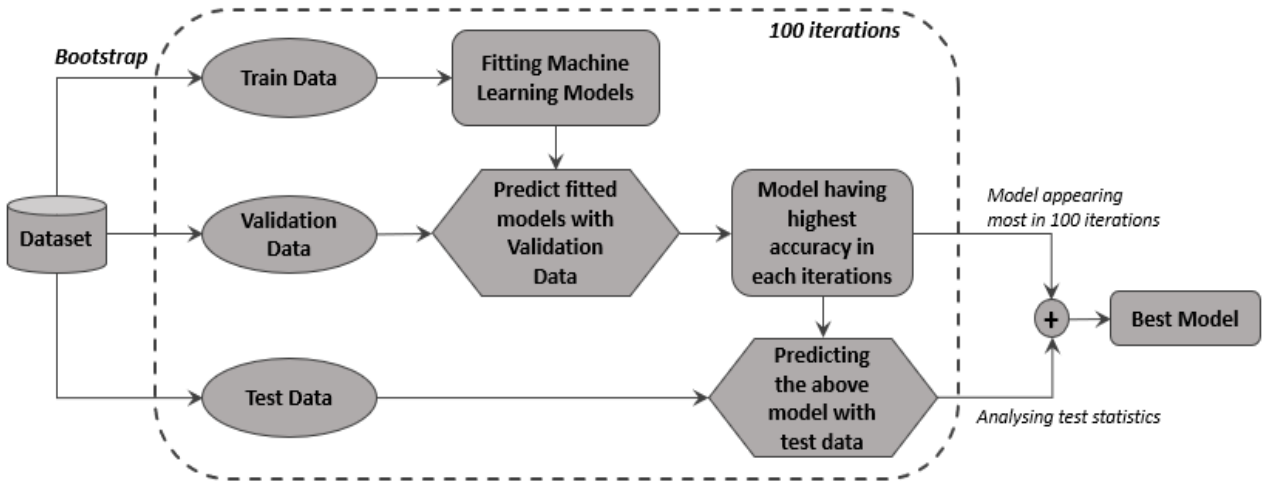


Figure 2: Steps followed in R to find the best classification model

Explanation on the above Flow chart process (Figure 2):

a) After loading the data into R, and data preprocessing is carried out.

b) Bootstrapping is done to the preprocessed data to get the train dataset. The left out samples in bootstrap process are split into validation data(50%) and test data(50%) for the validation process after the model fit step.

c) All the classifiers explained in section 2.2 is modelled using the respective libraries and functions discussed in the same section.

d) After model fitting of each classifier, the validation data is used to predict the type of back pain and the accuracy of these models are analyzed. The model with highest accuracy is considered and again predicted with the test data. Accuracy of this is stored to analyze the test statistics.

e) Steps (b) to (d) is performed 100 times to get the high performing model out of 6 classifier models

f) The frequency of models appearing as best model in step (d) is found and considered as the best model of all classifiers. The best model's coefficients and variable importance plot is drawn to analyze the contribution of these variables to the model.

3. RESULTS

Table 2: Performance analyses of all models in 100iterations

	Classification Methods						
	Classification Tree	L.R using multinorm	Random Forest	SVM	Bagging	Boosting	Test Accuracy
Min	0.7368	0.7879	0.8182	0.7576	0.7714	0.7727	0.8235
1st Qu	0.8567	0.8406	0.913	0.9038	0.8801	0.9011	0.8889
Median	0.8815	0.8723	0.9291	0.9275	0.9104	0.9198	0.9204
Mean	0.8807	0.8718	0.929	0.9206	0.9024	0.9157	0.9182
3rd Qu	0.9111	0.9014	0.9459	0.9412	0.9296	0.9317	0.9431
Max	0.9571	0.9688	0.9844	0.9726	0.9697	0.9718	0.9865

Table 2: Accuracy all models in validation dataset and best model in test dataset for first 6 iterations only.

Classification Tree	L.R using multinorm	Random Forest	SVM	Bagging	Boosting	Best Method	Test Accuracy
0.861538462	0.8	0.923076923	0.938461538	0.892307692	0.938462	SVM	0.953846
0.849315068	0.849315068	0.945205479	0.95890411	0.917808219	0.958904	SVM	0.863014
0.85915493	0.85915493	0.929577465	0.929577465	0.873239437	0.929577	Random Forest	0.944444
0.931506849	0.904109589	0.931506849	0.97260274	0.931506849	0.945205	SVM	0.932432
0.880597015	0.835820896	0.880597015	0.910447761	0.910447761	0.910448	SVM	0.895522
0.942857143	0.914285714	0.914285714	0.942857143	0.9	1	Boosting	0.928571

Table 3: Number of times each model selected as best model

Method	**Frequency**
Classification(Rpart)	4
Logistic regression	5
Random Forest	54
SVM	20
Bagging	6
Boosting	11

Table 4: Performance analysis of Random Forset in test dataset

```
Confusion Matrix and Statistics

             Reference
Prediction    Nociceptive Neuropathic
  Nociceptive          42           0
  Neuropathic           1          30
```

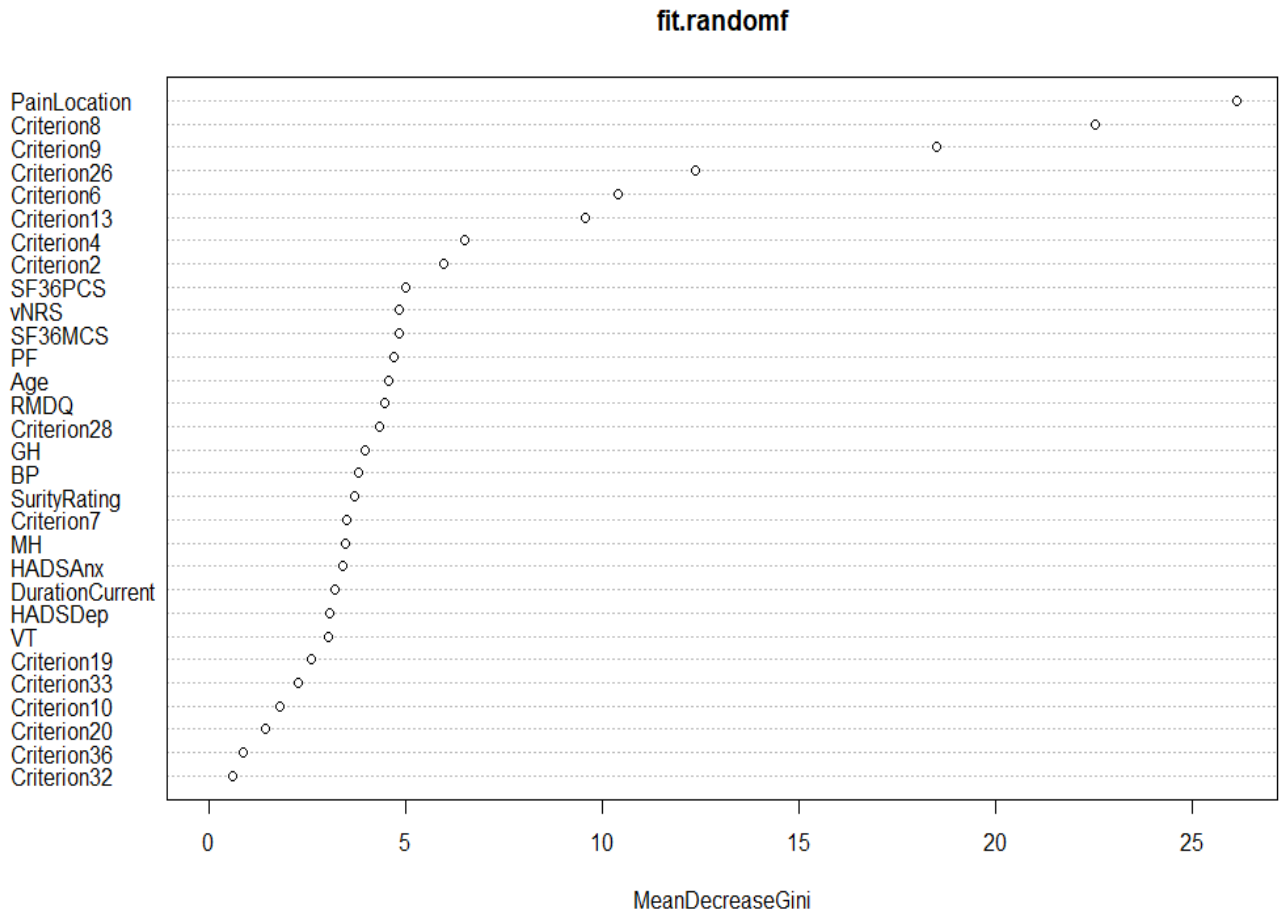
```
               Accuracy : 0.9863
                 95% CI : (0.926, 0.9997)
    No Information Rate : 0.589
    P-Value [Acc > NIR] : 8.631e-16

                  Kappa : 0.9718
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9767
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9677
             Prevalence : 0.5890
         Detection Rate : 0.5753
   Detection Prevalence : 0.5753
      Balanced Accuracy : 0.9884

       'Positive' Class : Nociceptive
```

Figure 3: Variable importance Plot for Random Forest.



4. DISCUSSION

From Table 3, it is evident that modeling with Random forest is efficient than other ensemble methods such as bagging and boosting. We have performed validation analysis of all models in 100 iterations, and Random forest is chosen as the best method most number of times . The dataset used here is validation and random forest came out as the best model among all 6 models. The best model is applied on test dataset to predict the response variable – PainDiagnosis and it is found that the mean accuracy of the random forest technique is 92.29% and highest of this being 98.44 as shown in Table 1.

Summary statistics of all models in validation dataset, and its best model in test dataset, in all iterations, is shown in Table 1. As explained before, we are supposed to check the performance of all models over different datasets obtained in sampling process. From the table 2 shows the results of first 6 iterations. Table 3 shows the frequency of each model that is selected as the best model (out of 100 iterations). It is found that the random Forest is still the best model with a frequency of 54. The second best model can be considered as the model with SVM machine learning technique which accounted 20 times. Classification tree and logistic regression cannot be considered as good models as they come out as best models in only 4 and 5 cases. The highest accuracy obtained by a model in test dataset is 98.65%. Both Random forest and SVM achieved this accuracy. However, when we compare models over 100 cases, Obviously Random forest is the winner.

Once we fit the best model with the dataset, we can obtain the important variables from varImpPlot command available for random forest models. The Figure 3 shows the graph between Variables and Mean Decrease Gini. As the value of x increases, the importance of variable also increases. So, the variables which are mentioned on the top of the y axis can be considered as important variables. From the figure 3 it is obvious that the most important variables are Pain location , Criterion8 , Criterion9 and Criterion 26.

5. CONCLUSION

From the overall model analysis, we found out that Random Forest model was the best model and the second best predicted model was SVM. These 2 methods were efficient in predicting the PainDiangonis of lower back pain in the dataset to a good limit. The mean prediction accuracy of the test data obtained was 92.29 for both the models.

6. REFERENCE

1. Confusion matric function in library **caret** in R
   https://www.rdocumentation.org/packages/caret/versions/6.0-83/topics/confusionMatrix

2. R. Berick, - An Idiot's guide to Support vector machines (SVMs)
   http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf