PREDICTING A PULSAR STAR

Ledneva Daria

May 2020

0.1 Abstract

This document is an article about project, the theme of which is "prediction of pulsar star". Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter.

0.2 IndexTerms

Natural and physical sciences, statistics, classification, astronomy, random forests, logistic regression

1 Introduction

This theme was chosen among others themes as the theme that had the most beautiful pictures. In this project, In this project, a data set is processed and 2 models are built that predict whether a star is a pulsar star or not.

2 Description of the data set

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators.

Each row lists the variables first, and the class label is the final entry. The class labels used are 0 (negative) and 1 (positive).

Attribute Information: Each candidate is described by 8 continuous variables, and a single class variable. The first four are simple

statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency . The remaining four variables are similarly obtained from the DM-SNR curve . These are summarised below:

- Mean of the integrated profile.
- Standard deviation of the integrated profile.
- Excess kurtosis of the integrated profile.
- Skewness of the integrated profile.
- Mean of the DM-SNR curve.
- Standard deviation of the DM-SNR curve.
- Excess kurtosis of the DM-SNR curve.
- Skewness of the DM-SNR curve.

Charts were constructed that can be used to estimate the spread of data for each function. The data set has been normalized to simplify working with it using this formula

$$(X - minX) * (maxd - mind)/(maxX - minX) + mind$$

where $maxd$ and $mind$ are the borders of the range in which normalized data will be located and $minX$ and $maxX$ are the borders of the range where the source data is located.

3 Description of methods

3.1 LogisticRegression

Logistic regression is a statistical model used to predict the probability of an event occurring by comparing it with a logistic curve. This regression returns the response as the probability of a binary event (1 or 0). The problem of predicting a pulsar star is a binary classification problem (to determine whether a star is a pulsar or not), so logistic regression is used here.

3.2 F-score

The samples is unbalanced (feature target class, 1,639 positive examples and 16,259 negative examples.), therefore it is best to use this metric to evaluate the quality of the model.

3.3 GridSearchCv

GridSearchCv is a method for finding the optimal hyper-parameters of the model by constructing a grid from the values of hyper-parameters and sequential training of models with all possible combinations of hyper-parameters from the grid. When searching for hyper-parameters at which the quality of the model would be highest, it is necessary to sort through all combinations of hyper-parameters, therefore this method is used here

3.4 RandomForestClassifier

A Random Forest is a learning algorithm that is mainly used for classification tasks. The forest is made up of decision trees, what removes the risk of overfitting, because the forest reduces retraining by averaging the result.

3.5 RandomizedSearchCV

This method is used to reduce GridSearchCv's run-time when iterating through combinations of hyper-parameters.

4 The experiments

4.1 Logistic Regression

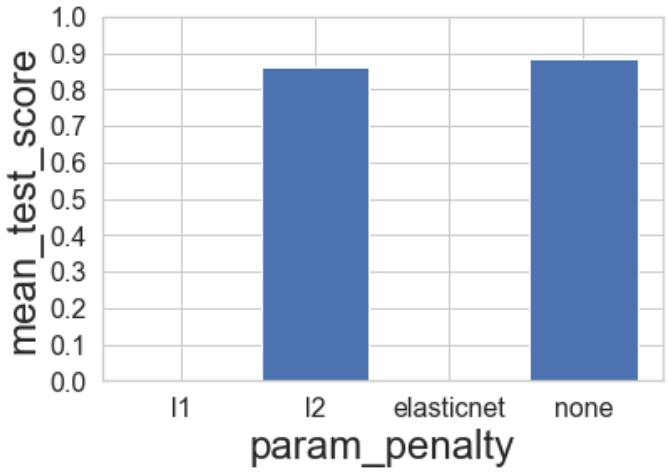
When using GridSearchCv with 4 different metrics, the following results will be obtained:

metrics	value	C	penalty
MAE	-0.0443133878675444	170	l2
MSE	-0.0886267757350888	170	l2
F1	0.8849803141410646	170	l2
R2	0.7597766170284789	170	l2

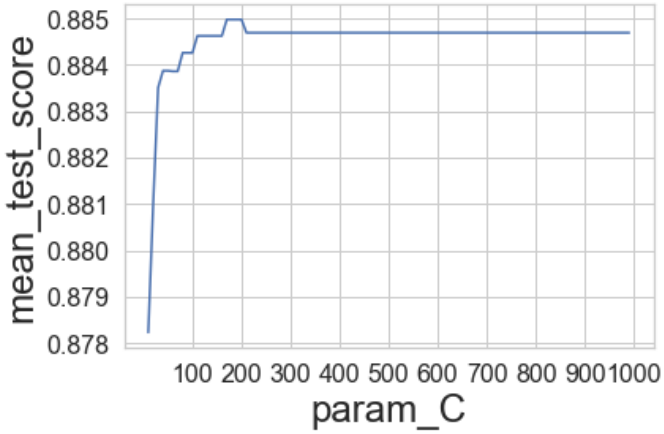
Graphs showing how the quality of the function depends on each

hyper-parameter:

Hyper-parameter *penalty*:



Hyper-parameter C:

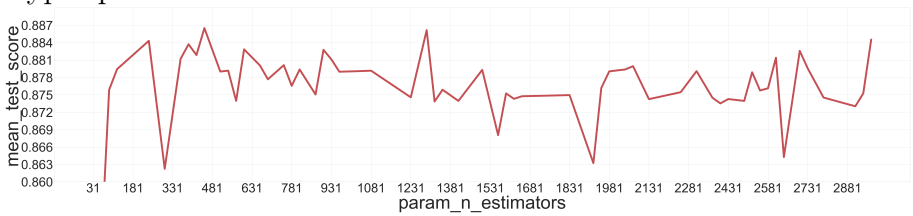


These graphs show that the model has the highest quality with the values of the hyper-parameters penalty $l2$ and *none* and C in the range from about 160 to 300.

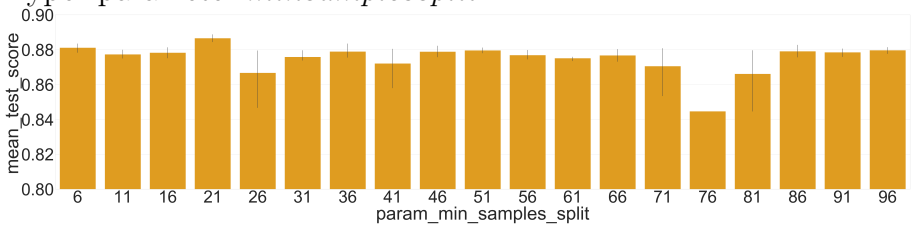
4.2 Random Forest

RandomizedSearchCV was used to reduce the ranges of hyper-parameters, thereby reducing the operating time of GridSearchCv. The dependence of the quality of the model on the values of hyper-parameters:

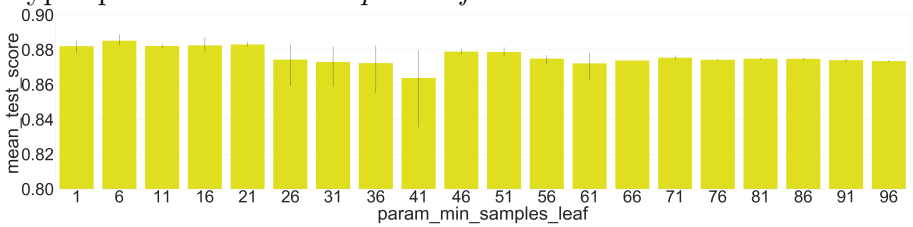
Hyper-parameter $nestimators$:



Hyper-parameter $minsamplessplit$:



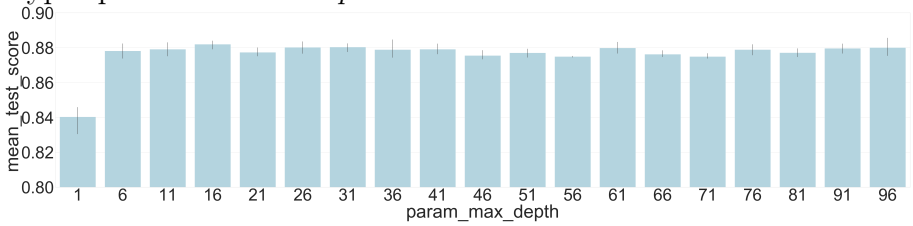
Hyper-parameter $minsamplesleaf$:



Hyper-parameter $maxfeatures$:

$maxfeatures$	$meantestscore$
sqrt	0.8764061952577415
log2	0.876362017705588

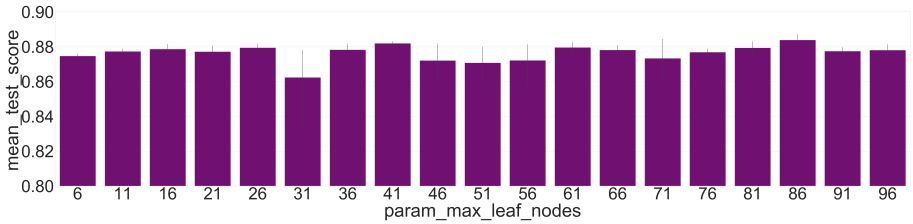
Hyper-parameter $maxdepth$:



Hyper-parameter $bootstrap$:

$bootstrap$	$meantestscore$
True	0.8759830714583473
False	0.8767481910209456

Hyper-parameter $maxleafnodes$:



Results of using GridSearchCv, in which hyper-parameters were used, at which the quality of the model is the highest (optimal hyper-parameters are taken from the graphs and tables above):

metrics	value
MAE	0.02592758158247653
MSE	0.05185516316495306
F1	0.9782902436429406

5 Conclusion

Table that shows the quality of two models using different metrics:

metrics	Random Tree	Logistic Regression
MAE	0.02592758158247653	-0.0443133878675444
MSE	0.05185516316495306	-0.0886267757350888
F1	0.9782902436429406	0.8849803141410646

Note that in all three metrics, the quality of RandomForest is higher than that of LogisticRegression. This is not surprising, since the linear classifier does not have the necessary flexibility for working with non-linear data set. In general, both models trained well enough, which was the target of my project (to build a well-functioning model for predicting pulsar stars).