JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning Model Selection Tool For Supervised Learning Algorithms

Contributing to an efficient ML Model Selection Process for Researchers and Practitioners.

[1]Rehmah Ahmed Batki

[1]Student
[1]Department of Computer Science,
[1]Royal College of Arts, Science and Commerce, Mumbai, India

Abstract : Machine Learning has simplified the task of teaching machines manually and have worked on making them think on their own. While it is widely used and have advanced to greater heights, there is room for much more efficient systems just by choosing the right model for your data. Machine Learning Model Selection Tool web application was created and put into use in this research with the goal of helping users choose the best supervised learning algorithms for tasks involving regression and classification. The tool's design considerations, implementation specifics, and underlying technique are covered in this document. It emphasizes mainly the importance of mean R-squared values and F1 scores as assessment metrics for tasks involving regression and classification, respectively. Machine Learning Model Selection Tool helps researchers and practitioners streamline their machine learning processes by providing an effective and user-friendly solution for algorithm selection.

Keywords **- Machine Learning, Model Selection, R-squared, F1 score**

I. INTRODUCTION

A crucial part of developing predictive models is choosing a machine learning (ML) model. Because there is such a wide variety of algorithms available, researchers and practitioners frequently encounter difficulties when trying to choose the best model for their datasets. This work presents a web-based tool that helps users choose models more quickly by giving them information about how various algorithms perform.

Machine Learning Model Selection Tool proposes report generating capability for Supervised Learning Algorithms. Under this umbrella it gives users 2 sections to use-

(i) Classification

It is the prediction of category for a row of attributes. This section requires the dataset to have a categorical target class. The dataset can have string as well as numeric values. String values are encoded with LabelEncoder.

(ii) Regression

Regression is the prediction of numeric values for continuous data. This section requires the dataset to have all numeric data.

II. METHODOLOGY

2.1 Evaluation Requirements:

- The dataset must be of filetype ".csv" or ".xlsx".
- The dataset must be a single file.
- Classification dataset must contain target class of categorical type.
- Regression dataset must contain all numeric data type values.
- The datasets should be preprocessed by the user according to their requirements.

2.2 Evaluation Models:

The Machine Learning Model Selection Tool is implemented using Flask, a Python web framework. Users can upload datasets to each section mentioned above, where the data is processed and analyzed using a variety of popular algorithms.

Modals used for Classification:

1. Multinomial
2. BernoulliNB
3. KNeighboursClassifier
4. RandomForestClassifier
5. AdaBoostClassifier
6. GradientBoostingClassifier
7. BaggingClassifier
8. ExtraTreesClassifier
9. SGDClassifier
10. SVC
11. LinearSVC
12. NuSVC
13. LogisticRegressionCV
14. LogisticRegression

Modals used for Regression:

1. Linear Regression
2. Lasso Regression
3. Elastic Net Regression
4. KNeighborsRegressor
5. Decision Tree Regressor
6. Gradient Boosting Regressor
7. Random Forest Regressor
8. Support Vector Regressor

2.3 Evaluation Metrics:

The tool evaluates classification models based on F1 score and regression models based on mean R-squared values.

Metric used for Classification: F1 Score

F1 Score combines 2 metrics Precision and Recall of a model and gives a single metric to score that balances the Precision-Recall tradeoff.

Precision?

The percentage of positive class predictions that came true is known as precision. For example, if a model labels 100 samples as positive and among those 80 are actually positive class of the dataset (the other 20 were negative but mistakenly marked as "positive" by the model), then the precision is said to be 80%.

Precision = True Positives Samples / (True Positives Samples + False Positives Samples)

Recall?

The percentage of actual positive class samples that the model was able to identify is known as recall. In other words, how many of the 100 samples in the dataset's positive class that make up the test set were identified? The recall is 60% if 60 of the positive samples were accurately identified.

Recall = True Positives / (True Positives Samples + False Negatives Samples)

Precision is the ratio of true positive samples to the total predicted positives samples , while recall is the ratio of true positive samples to the total actual positives samples .

F1 Score?

The harmonic mean of a classifier's precision and recall returns the F1-score as a single statistic. Its main purpose is to compare two classifiers' performances. Assume classifier B has greater precision and classifier A has a larger recall. When deciding which classifier performs better in this situation, one can look at the F1-scores of both of them.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

For this project I have taken average='micro' which means metrics are calculated globally by counting the total true positive samples, false negative samples and false positive samples.

Metric used for Regression: Mean R-Squared Value

The statistical measure "R-squared", or "coefficient of determination", is a metric to know how much of the variance in the dependent variable can be predicted from the independent variables. It gives an idea of how effective are the independent factors accountable for the dependent variable's variability.

R-squared is calculated using the formula:

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Sum of squared regression (SSR) is the difference between the actual and predicted values of the dependent variable.

Total Sum of Squares (SST) is the difference between the actual values and the mean of the dependent variable.

R-squared ranges from 0 to 1, where:
0 indicates that the model doesn't explain any variability of the dependent variable around its mean.
1 indicates that the model explains all the variability of the dependent variable around its mean.

III. RESULTS

The tool generates a report table listing the models along with their respective F1 scores or mean R-squared values, depending on the section. The models are ordered in descending order, with the highest score indicating the best-performing model.

3.1 Classification Report on Obesty_classification dataset

Classification Report

Model	f1 Score
NuSVC	0.6696208764155588
AdaBoostClassifier	0.6696208764155588
BernoulliNB	0.6696208764155588
Multinomial	0.6696208764155588
LogisticRegressionCV	0.6696208764155588
BaggingClassifier	0.6696208764155588
ExtraTreesClassifier	0.6696208764155588
GradientBoostingClassfier	0.6696208764155588
KNeighboursClassifier	0.6696208764155588
LinearSVC	0.6687592319054653
LogisticRegression	0.6446331856228459
RandomForestClassifier	0.6395864106351551
SGDClassifier	0.5420974889217134

3.2 Classification Report on gender classification dataset

Classification Report

Model	f1 Score
KNeighboursClassifier	0.9984003199360127
BernoulliNB	0.9984003199360127
NuSVC	0.993001399720056
AdaBoostClassifier	0.9832033593281344
Multinomial	0.9802039592081584
LogisticRegression	0.979004199160168
SGDClassifier	0.978604279144171
LogisticRegressionCV	0.9784043191361728
LinearSVC	0.9766046790641871
GradientBoostingClassfier	0.9764047190561888
SVC	0.9762047590481904
BaggingClassifier	0.9756048790241951
ExtraTreesClassifier	0.9750049990002

3.3 Regression report on Real Estate dataset

Regression Report

Model	R-Squared
Random Forest Regressor	0.6879876558097554
Gradient Boosting Regressor	0.6692863021013042
KNeighbors Regressor	0.6027222860406153
Support Vector Regressor	0.5827116181336207
Linear Regression	0.5823304449734418
Lasso Regressor	0.5662855509783788
Elastic Net Regressor	0.5460989708638087
Decision Tree Regressor	0.43095106361361396

3.4 Regression report on Candy dataset

Regression Report

Model	R-Squared
Elastic Net Regressor	0.09688594755066474
Lasso Regressor	0.05390317230722965
Random Forest Regressor	0.03835229947566003
Gradient Boosting Regressor	-0.023073394020058313
KNeighbors Regressor	-0.029343362110107428
Support Vector Regressor	-0.07517688158423655
Linear Regression	-0.11580418434069495
Decision Tree Regressor	-0.30179690034869344

The tool is available online for users to access and test with their datasets. Users can also provide feedback and suggestions through the contact page. Additionally, the tool allows users to download the report table in CSV format, enabling further analysis and integration with other tools. Its results thus makes it a powerful analysis tool for all researchers and practitioners in the field.

It is important to note that other factors such as interpretability, computational complexity and domain-specific considerations can influence choice of a model despite the fact that the report table's top-ranked model represents the best performing learner according to the evaluation measure used. The tool is invaluable for researchers and practitioners interested in automating and speeding up the process of selecting models.

IV. CONCLUSION

By offering a practical and effective model selection method, the Machine Learning Model Selection Tool for Supervised Learning Algorithms makes a significant addition to the field of machine learning. The tool assists academics and practitioners in selecting the best algorithms for their datasets by utilizing the F1 score for classification tasks and the mean R-squared value for regression activities.

When choosing a model, it's crucial to take into account additional aspects such model interpretability, computational complexity, and domain-specific needs, even though the tool offers insightful information on algorithm performance. The tool could also be improved and refined further by adding more assessment metrics, supporting more algorithms, and integrating with other machine learning platforms and tools.

All things considered, the Machine Learning Model Selection Tool is evidence of continuous attempts to make machine learning easier to understand and more efficient, enabling practitioners and researchers to make better decisions and produce better outcomes for their machine learning initiatives.

Tool URL:- http://rehmahahmed.pythonanywhere.com/

REFERENCES

[1] A Review of Automatic Selection Methods for Machine Learning Algorithms and Hyperparameter Values Gang Luo (corresponding author)

[2] Approach towards Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning Mrs .Swati Dhabarde

[3] A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction Nicholas Pudjihartono1, Tayaza Fadason1,2, Andreas W. Kempa-Liehr 3 * and Justin M. O'Sullivan

[4] Model Selection for Machine Learning Algorithm on Decision Making in Oil and Gas Upstream Project Malaysia Mohd Shahrizan Abd Rahman[1] and Nor Azliana Akmal Jamaludin[2]

[5] Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning Sebastian Raschka