AMLR Auto Machine Learning Report

Exploratory Data Analisys

In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Wikipedia

Preliminar Results

Configuration | Regression Analisys | Unbalance Classes | Correlation | Multicollinearity | Residual Analisys

Data Set

Shape	891 / 12
Classes	2
Classes Found	['Yes' 'No']
Duplicated	none
Excluded Features:	

Feature	Freq
PassengerId	1.0
Name	1.0
Ticket	0.7643097643097643

column	dtype	not_null	percent
Survived	object	891	0.0
Pclass	int64	891	0.0
Sex	object	891	0.0
Age	float64	714	0.1987
SibSp	int64	891	0.0
Parch	int64	891	0.0
Fare	float64	891	0.0
Cabin	object	204	0.771
Embarked	object	889	0.0022

Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them, such as standard deviation and kurtosis. The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values are normally all of the same kind. However, there may also be missing values, which must be indicated in some way.

Wikipedia

Grid - Hyperparameter optimization

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

AMLR Auto Machine Learning Report

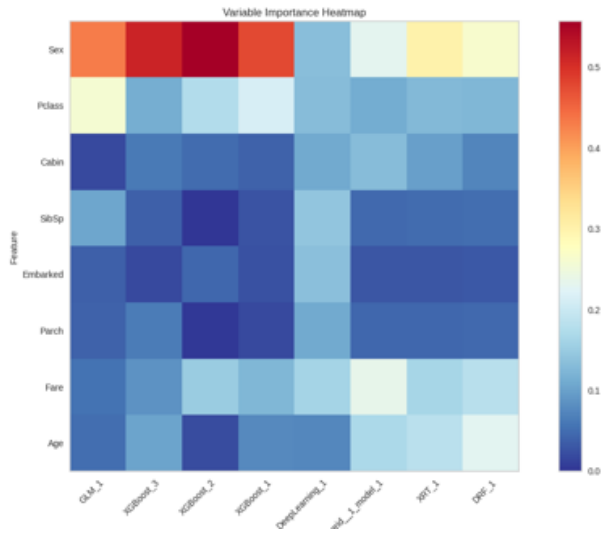
AutoML - Results The models are classified by a standard metric based on the type of problem (the second column of the scoreboard). In binary classification problems, this metric is AUC, and in classification problems in several classes, the metric is the average error per class. In regression problems, the standard classification metric is deviation.

model_id	auc	logloss	aucpr	mean_per_class_error	training_time_ms
StackedEnsemble_AllModels_AutoML_20210312_183930	0.844421101774043	0.455287378458646	0.8700386454164557	0.2433367286308463	237
StackedEnsemble_BestOfFamily_AutoML_20210312_183930	0.8435510567863509	0.4572602699928686	0.8718251674919805	0.2458513708513708	222
XGBoost_1_AutoML_20210312_183930	0.8342564298446652	0.4841015306738469	0.8602276825788912	0.2731516849163908	485
GLM_1_AutoML_20210312_183930	0.8323837110601817	0.4655844584338352	0.8655179105928286	0.2504562431033019	203
XGBoost_grid__1_AutoML_20210312_183930_model_1	0.8306330107800696	0.5047491125223659	0.8781263973468549	0.2261798658857482	1368
XGBoost_3_AutoML_20210312_183930	0.8292695866225277	0.4986974502753629	0.8625016423150934	0.2446311858076564	83
DRF_1_AutoML_20210312_183930	0.825492318139377	1.7892353902771394	0.849143304474938	0.220291146761735	44
XGBoost_2_AutoML_20210312_183930	0.8107015533486122	0.5456942878920749	0.8565910762350545	0.2895339954163483	121
XRT_1_AutoML_20210312_183930	0.8060807656395892	1.274138758410555	0.8346404897827888	0.2899796282149223	34
DeepLearning_1_AutoML_20210312_183930	0.7596394618453441	0.6106566471787331	0.7998649566824941	0.3484105763517528	83

Partial dependence plot (PDP) gives a graphical depiction of the marginal effect of a variable on the response. The effect of a variable is measured in change in the mean response. PDP assumes independence between the feature for which is the PDP computed and the rest.

An Individual Conditional Expectation (ICE) plot gives a graphical depiction of the marginal effect of a variable on the response. ICE plots are similar to partial dependence plots (PDP); PDP shows the average effect of a feature while ICE plot shows the effect for a single instance. This function will plot the effect for each decile. In contrast to the PDP, ICE plots can provide more insight, especially when there is stronger feature interaction.

Variable Importance by Model



AML - Partial Dependence

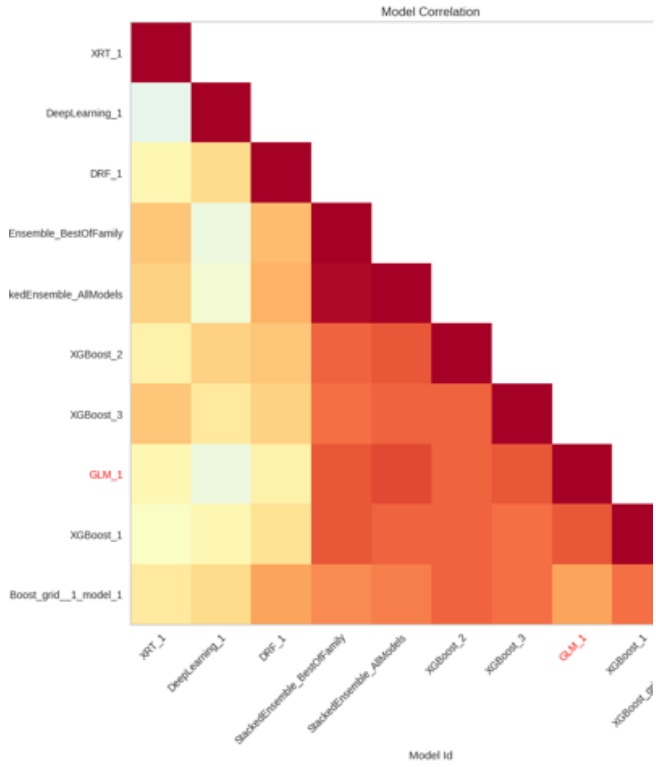
Select a Feature

Ensemble - (ICE) Individual Condition Expectation

Select a Feature

Correlation Heatmap by Model

AMLR Auto Machine Learning Report



Model Performance

Analytical Performance Modeling

Analytical Performance Modeling is a method to model the behaviour of a system in a spreadsheet. It is used in Software performance testing. It allows evaluation of design options and system sizing based on actual or anticipated business usage. It is therefore much faster and cheaper than performance testing, though it requires thorough understanding of the hardware platforms

Wikipedia

Comparative Metrics Table

Algo	Overall ACC	Kappa	Overall RACC	SOA1(Landis & Koch)	SOA2(Fleiss)	SOA3(Altman)	SOA4(Cicchetti)	SOA5(Cramer)	SOA6(Matthews)	TNR Macro	TPR Macro	FPR Macro
GLM	0.8125	0.5741	0.5598	Moderate	Intermediate to Good	Moderate	Fair	Strong	Moderate	0.7647	0.7647	0.2353
Random Forest	0.7875	0.5499	0.5279	Moderate	Intermediate to Good	Moderate	Fair	Relatively Strong	Moderate	0.7719	0.7719	0.2281
GBM	0.8438	0.6662	0.5319	Substantial	Intermediate to Good	Good	Good	Strong	Moderate	0.8266	0.8266	0.1734
xGBoost	0.8625	0.6986	0.5438	Substantial	Intermediate to Good	Good	Good	Strong	Strong	0.8337	0.8337	0.1663
Deep Learning	0.6938	0.3708	0.5133	Fair	Poor	Fair	Poor	Moderate	Weak	0.689	0.689	0.311

AMLR Auto Machine Learning Report

Description	RF	GLM	GBM	XGB	DL
overall	3.7	3.86667	4.48333	4.68333	2.38333
class	6.9	8.4	7.8	9.2	5.6

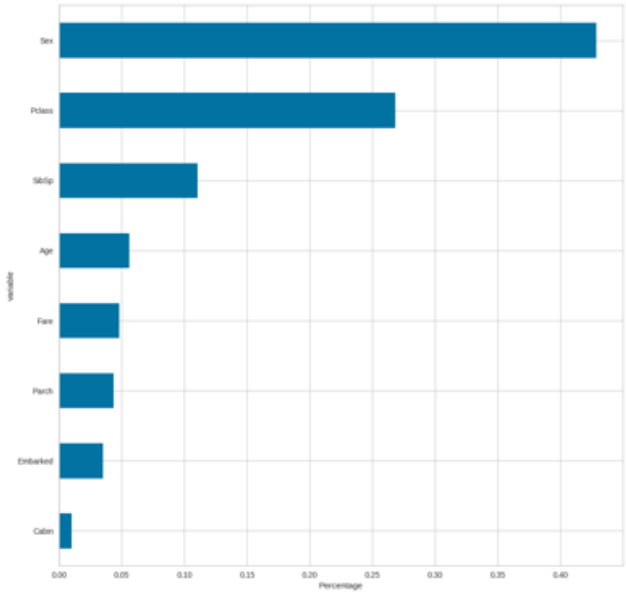
The best name: XGB

Gradient Linear Estimator

Confusion Matrix

Description	precision	recall	f1-score	support
Yes	0.9714	0.5397	0.6939	63.0
No	0.768	0.9897	0.8649	97.0
accuracy	0.8125	0.8125	0.8125	0.8125
macro avg	0.8697	0.7647	0.7794	160.0
weighted avg	0.8481	0.8125	0.7975	160.0

Feature Importance



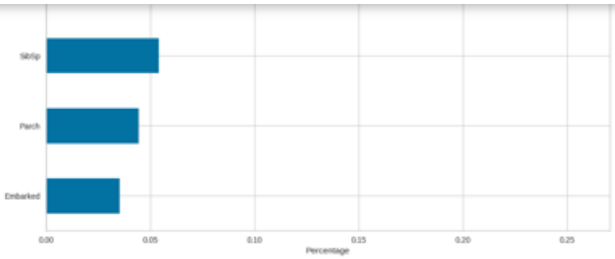
Dynamic Random Forest

Confusion Matrix

Description	precision	recall	f1-score	support
Yes	0.7458	0.6984	0.7213	63.0
No	0.8119	0.8454	0.8283	97.0
accuracy	0.7875	0.7875	0.7875	0.7875
macro avg	0.7788	0.7719	0.7748	160.0
weighted avg	0.7858	0.7875	0.7862	160.0

Feature Importance

AMLR Auto Machine Learning Report

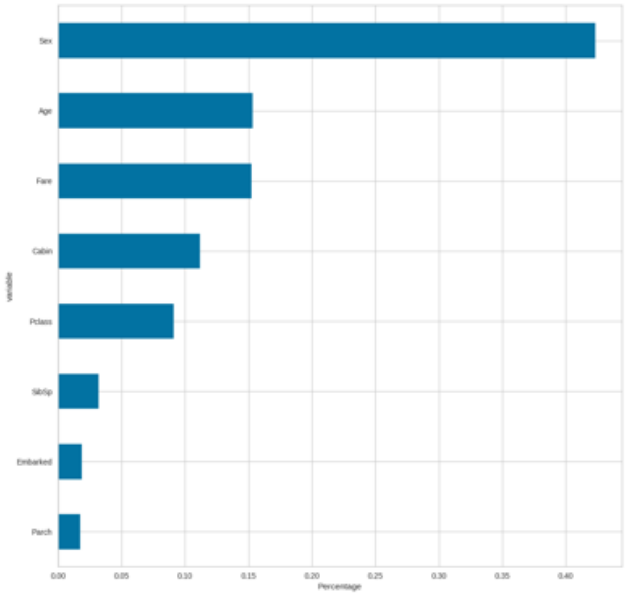


Gradient Boost Machine

Confusion Matrix

Description	precision	recall	f1-score	support
Yes	0.8393	0.746	0.7899	63.0
No	0.8462	0.9072	0.8756	97.0
accuracy	0.8438	0.8438	0.8438	0.8438
macro avg	0.8427	0.8266	0.8328	160.0
weighted avg	0.8434	0.8438	0.8419	160.0

Feature Importance

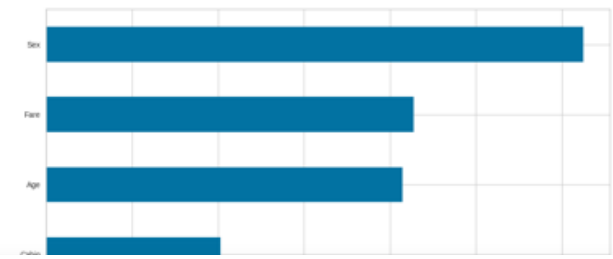


XGBoost

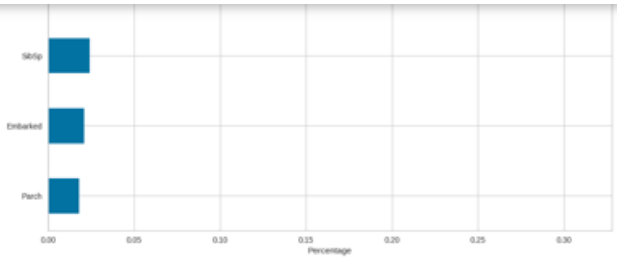
Confusion Matrix

Description	precision	recall	f1-score	support
Yes	0.9362	0.6984	0.8	63.0
No	0.8319	0.9691	0.8952	97.0
accuracy	0.8625	0.8625	0.8625	0.8625
macro avg	0.884	0.8337	0.8476	160.0
weighted avg	0.8729	0.8625	0.8577	160.0

Feature Importance



AMLR Auto Machine Learning Report



Deep Learning

Confusion Matrix

Description	precision	recall	f1-score	support
Yes	0.6	0.6667	0.6316	63.0
No	0.7667	0.7113	0.738	97.0
accuracy	0.6938	0.6938	0.6938	0.6938
macro avg	0.6833	0.689	0.6848	160.0
weighted avg	0.701	0.6938	0.6961	160.0

Feature Importance