Written Tutorial for Running PAT: Introduction

Video Link

Table of Contents

1. What is PAT?

The Predictive Analytics Toolkit (or PAT) is a cloud-based web platform used to facilitate automated development and testing of predictive models. For most users, the main purpose of PAT is to give simplified access to the analytics power of the vast array of the R programming language's packages and commands that are useful for detecting, analyzing, quantifying, and visualizing associations and other relationships in datasets using standardized, well-documented, and well-supported algorithms. The intended workflow of PAT is for users to be able to upload a dataset, select specific data analysis parameters, and receive predictive analytics and associated diagnostic results at the push of a button.

PAT uses the framework of the CARET (Classification And REgression Training) package in the R programming language (see Kuhn, 2008). It is a module of the Causal Analytics Toolkit (or CAT) and both CAT and PAT were developed by Cox Associates Consulting.

Please note that this introduction and all associated documentation found here on the Cox Associates PAT website correspond only to the Data and Predict tabs within CAT. There are several additional modules within CAT that are not covered by this documentation. For more information on CAT, contact Cox Associates.

This introductory tutorial discusses the intended users of PAT, the type of input data PAT requires, general concepts of predictive analytics and machine learning and how they are used in PAT, the types of outputs PAT generates, and additional training materials that are available for PAT.

1.1. Who should and should not use PAT?

PAT was developed to be used by individuals with an understanding of predictive analytics concepts, as a tool for those with little to no familiarity with the R programming language. Additionally, PAT can aid those who are familiar with R but want a quick, user-friendly way of performing predictive analytics.

PAT is designed to be very easy to use. Nevertheless, those unfamiliar with machine learning and predictive analytics are encouraged to study these topics outside of PAT. Doing so will facilitate correct use of the features and appropriate interpretation of the outputs of PAT. A very brief overview of machine learning in the context of PAT is provided in Section 1.3; however, for users to be able to fully interpret and utilize the outputs of PAT properly, Section 1.3.1 lists a few resources that provide more in-depth information on machine learning.

1.2. PAT Input Data

There are seven example datasets available to users on the PAT webpage for utilizing the functions of PAT, with a variety of features between them. These include sample datasets with data on chemical mutagenicity, asthma risk factors and outcomes, and a banking institution's telemarketing campaign. The asthma dataset comes from Cox 2017, and the banking dataset comes from Moro et al., 2014. Users can also upload their own datasets to the website and make simple modifications before applying the machine learning algorithms to the selected dataset, such as rearranging columns and selecting a subset of columns to be included in the analysis.

The final dataset that is processed through PAT, however, must have the dependent variable (the outcome that will be predicted) defined as the first column of the dataset. ***This column must contain only ones and zeros.*** "1" signifies the presence of the outcome (i.e., 'True', or a positive outcome) while "0" signifies the absence of the outcome (i.e., 'False', or a negative outcome). A subset of the asthma example dataset provided on the PAT website is presented in Table 1 as an illustration of the structure of the required input. The AsthmaEver column in this dataset will be taken as the dependent variable when developing prediction models.

Table 1. Subset of the asthma example dataset within PAT.

Dependent Variable	Potential Predictors			
AsthmaEver	Education	Smoking	Pm25Average	Age
0	5	0	4.0	51
0	6	1	6.3	69
0	5	1	6.3	68
0	4	1	9.7	72
0	6	0	9.7	83
0	5	0	9.7	54
1	4	1	4.8	65
0	6	0	5.0	56
0	6	1	6.5	56
0	6	0	7.3	96

All columns to the right of the dependent variable are taken as potential predictors of the outcome, and these are used in training the prediction models. As can be seen in Table 1, these potential predictor columns can be binary or non-binary.

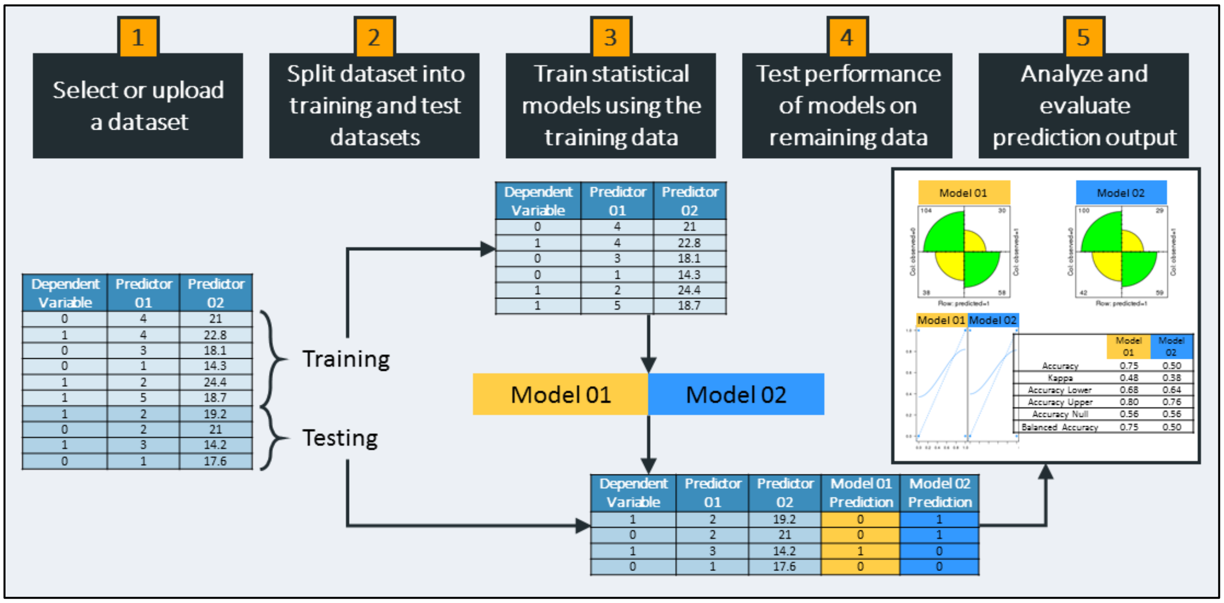
1.3. Overview of Machine Learning and PAT

1.3.1. What is machine learning?

Machine learning is a method of data analysis that can be used to automatically construct computational prediction models. Use of machine learning in the context of PAT can be described as a five-step process (see

Figure 1 below). Specifically:

1. The user selects or uploads a dataset of interest.
2. The user then selects how to split the input dataset into training and test subsets.
3. PAT then trains statistical models to predict the observed outcome of interest.
4. PAT tests the performance of these models on the remaining portion of the dataset (i.e., the portion not used in training the models).
5. PAT then analyzes the predictions of the various models, and evaluates the ir performance.


1
Select or upload a dataset
2
Split dataset into training and test datasets
3
Train statistical models using the training data
4
Test performance of models on remaining data
5
Analyze and evaluate prediction output
Training
Testing
Model 01
Model 02
Model 01
Model 02
104
30
38
58
Col: observed=0
Col: observed=1
Row: predicted=1
100
29
42
59
Col: observed=0
Col: observed=1
Row: predicted=1
Model 01
Model 02

Dependent Variable	Predictor 01	Predictor 02
0	4	21
1	4	22.8
0	3	18.1
0	1	14.3
1	2	24.4
1	5	18.7
1	2	19.2
0	2	21
1	3	14.2
0	1	17.6

Dependent Variable	Predictor 01	Predictor 02
0	4	21
1	4	22.8
0	3	18.1
0	1	14.3
1	2	24.4
1	5	18.7

Dependent Variable	Predictor 01	Predictor 02	Model 01 Prediction	Model 02 Prediction
1	2	19.2	0	1
0	2	21	0	1
1	3	14.2	1	0
0	1	17.6	0	0

	Model 01	Model 02
Accuracy	0.75	0.50
Kappa	0.48	0.38
Accuracy Lower	0.68	0.64
Accuracy Upper	0.80	0.76
Accuracy Null	0.56	0.56
Balanced Accuracy	0.75	0.50

Figure 1. Flow diagram of how machine learning is used in PAT.

For a more detailed discussion of the R package that is at the core of PAT, see the R documentation page or Kuhn, 2008. For additional information on machine learning in general, see Mitchell, 1997 and Cox et al., 2018.

1.3.2. What does PAT do?

PAT is intended to produce predictive analytics results with very little user input. For the simplest use of PAT, the following are the only items required to be defined by the user:

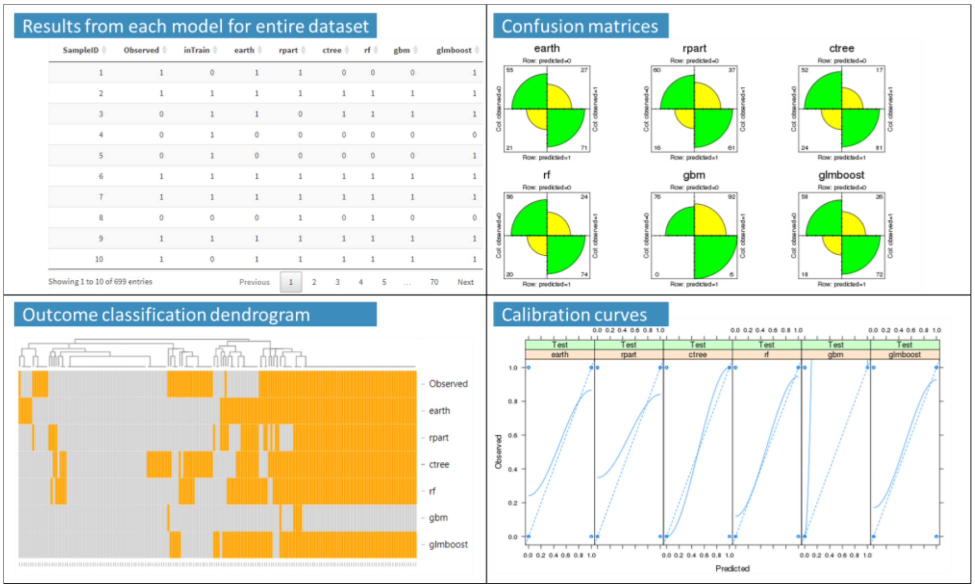
- An input dataset with one column being composed of ones and zeroes (i.e., the dependent variable, where 1 = presence of the outcome and 0 = absence of the outcome).
- The percentage of the data, and which section of the rows of the data, to use as the training dataset (i.e., only the rows from the first percentage of the dataset, or a percentage of the rows selected randomly from throughout the dataset).

With these two items defined, PAT will develop six prediction models based on the input data. These models are:

- earth: The R programming language name for Multivariate Adaptive Regression Splines (which is trademarked and licensed to Salford Systems).
- rpart: Recursive partitioning.
- ctree: Classification tree.
- rf: Random Forest.
- gbm: Gradient boosting machines.
- glmboost: Boosted generalized linear model.

1.3.3. What does PAT produce?

Predicted outcomes are generated for the input dataset's dependent variable using the suite of six prediction models. Additionally, several model diagnostic outputs are generated (see Figure 2). These include confusion matrices, accuracy and balanced accuracy calculations, calibration curves, and an outcome classification dendrogram (i.e., heatmap).


Results from each model for entire dataset

SampleID	Observed	inTrain	earth	rpart	ctree	rf	gbm	glmboost
1	1	0	1	1	0	0	0	1
2	1	1	1	1	1	1	1	1
3	0	1	1	0	1	1	1	1
4	0	1	0	0	0	0	0	0
5	0	1	0	0	0	0	0	1
6	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1
8	0	0	0	1	0	1	0	0
9	1	1	1	1	1	1	1	1
10	1	0	1	1	1	1	1	1

Showing 1 to 10 of 699 entries

Figure 2. Example of PAT outputs using the mutagenicity input dataset.

1.3.4. How much control does the user have?

While much of what goes on 'under the hood' in PAT is automated and doesn't require direct user input, several functionalities are modifiable by the user for their specific needs. These include:

- The ability to upload a dataset, make modifications to the columns (potential predictors) that are to be included in the analysis, and save this to the cloud for future use.
- Optional pre-processing filters that narrow down the number of potential predictors analyzed (to reduce runtime).
- The option to sample all training data from the top rows of the input dataset (e.g., for analysis of a time series or a dataset with distinct sections) or to sample training rows randomly from the entire dataset.
- The option to impute data for or remove all rows with missing data in the dependent variable. Please note that missing values in the dependent variable are not allowed within PAT.

1.4. Overview of Other Documentation

In addition to this introductory tutorial to PAT, several written and video tutorials have been developed to help guide new users through the process of using PAT to explore development of prediction models by applying machine learning algorithms to their datasets and to use the PAT features to characterize the performance of such models.

These additional documentation materials include the following (links are provided to both the written and video versions of the tutorials):

- **Existing Dataset Tutorial:** Guide for how to run PAT on one of the existing datasets available on the website (mainly employing the tool's default settings to obtain predictive results and diagnostics) and how to interpret some of the basic output figures of PAT [Written Tutorial] [Video Tutorial].
- **User-provided Dataset Tutorial:** Guide for how to run PAT on a dataset uploaded by the user. This includes details on some of the more in-depth functionalities of PAT, such as how to format and save the data to the cloud, how to export the model-generated data, how to interpret some of the more advanced figures of PAT, and other details not covered in the Existing Dataset Tutorial [Written Tutorial] [Video Tutorial].

2. References

Cox, L.A., Jr., Popken, D.A., and Sun, R.X. (2018). Causal Analytics for Applied Risk Analysis. Springer. doi: 10.1007/978-3-319-78242-3

Cox, L.A. (2017). Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010-2013. Environmental Research, 155: 92-107. Doi: https://doi.org/10.1016/j.envres.2017.01.003

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5). Available online at: https://www.jstatsoft.org/article/view/v028i05

Mitchell, T.M. (1997). Machine Learning. *McGraw-Hill*. Available online at: https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill%20-%20Machine%20Learning%20-Tom%20Mitchell.pdf

Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, Elsevier, 62:22-31. Available online at: https://archive.ics.uci.edu/ml/datasets/bank+marketing