Written Tutorial for Running PAT: User-provided Dataset

Video Link

Table of Contents

1. Introduction

1.1. Purpose of this Tutorial

This Tutorial covers some of the advanced features of the Predictive Analytics Toolkit (PAT) that were not covered in the Existing Dataset Tutorial for PAT. That Tutorial, which guides users through how to run PAT on a dataset provided on the PAT website, how to configure and run PAT, and how to interpret many of the basic output figures of PAT, can be found at this web address on the PAT website. Users are encouraged to read that Tutorial, along with the Introduction to PAT Tutorial (located at this web address) before proceeding with the following Tutorial.

The following are the main topics covered in this Tutorial:

1. Uploading and saving a new dataset to PAT.
2. The use of the top rows of a dataset as the training dataset, the bottom rows as the test dataset, and how this can be useful.
3. Exporting the results from PAT.
4. How to interpret the advanced output figures of PAT.

An abbreviated version of this Tutorial can be viewed in video form at this web address.

1.2. Top Rows as Training Dataset, and Dataset to be Used

We will use an example dataset developed for the purposes of this Tutorial, which can be downloaded via this link on the Cox Associates website. The basis of this dataset is the **mutagens** dataset used within the Existing Dataset Tutorial, with several differences. The main difference is that, for demonstration purposes only, the rows of data have been sorted such that the molecular weight of the chemicals increases down the rows of the dataset. In this Tutorial, we will configure PAT such that it uses only the top 80% of the rows of data (559 rows) to train the prediction models (i.e., chemicals with a molecular weight of less than ~300 g/mol), and the bottom 20% of the rows (i.e., chemicals with a molecular weight greater than ~300 g/mol) to test the developed prediction models (see Section 2.3). An additional column of data, the **Row_Use** column, denotes whether a given chemical will be used in the training or testing dataset for the prediction models. This sorting of rows and the molecular weight cutoff in the example dataset are purely arbitrary, and are only intended to demonstrate the user's ability to organize their data such that rows with certain characteristics can be used for training versus testing of the prediction models and the possibility for exploring different questions based on this organization.

A portion of the example dataset can be seen in Table 1-1.

Table 1-1. Subset of the example dataset used in this Tutorial.

Observation	Row_Use	MW	mutagen	Mp	nAB	nH
555	Training	296.29	1	0.7	16	12
556	Training	296.35	1	0.68	10	16
557	Training	298.32	1	0.68	12	14
558	Training	298.33	1	0.75	24	12
559	Training	298.33	1	0.75	24	12
560	Testing	300.33	0	0.66	5	12
561	Testing	300.34	1	0.66	12	16
562	Testing	300.35	0	0.69	11	17
563	Testing	300.76	0	0.72	12	13
564	Testing	300.79	0	0.78	24	13

Note: Due to the fact that 80% of the 699 rows of data (i.e., 559 rows) will be used in training the prediction models, rows 1-559 have the **Row_Use** column labeled **Training**, while rows 560-699 are labeled **Testing**.

The organization of the input data in this way, such that the top and bottom portions of the rows of data are organized according to some category of the data (in this case, lower versus higher molecular weight) is just one example of how to organize a dataset to take advantage of PAT's ability to use the top rows of the dataset in training the prediction models and the bottom rows to test the models.

Using this row organization technique, users can also effectively apply models previously generated within PAT to a new set of data. This can be done by appending additional rows of data onto a dataset that had been analyzed with PAT, but only allow the rows from the previous training dataset to be used when analyzing this new 'hybrid' dataset. For example, say the user had an original dataset with 100 rows and generated prediction models using the first 75 rows as the training dataset. After generating or obtaining 100 additional observations, the (roughly) same

prediction models that had been applied to the original dataset can be developed again in PAT and applied to these new 100 rows. This would be done by appending the new 100 rows to the bottom of the old 100 rows, importing this new 200-row dataset into PAT, and setting the training data percentage to 37% so that only the first 74 rows (i.e., roughly those used in the original analysis) will be used to re-develop the models, which will then generate predictions for all 200 rows of data.

Another possible use of this feature would be to have a time series of observations, where the initial observations (those in the first rows) are used to train the prediction models on what the future outcomes (those in the bottom rows) will be.

Whatever the top versus bottom row categorizations are for a given input dataset, it is important to determine the exact percentage of the dataset that is composed of the rows the user intends to use in training the models so that the user-defined percentage will instruct PAT to extract only the correct rows for training and only the correct rows for testing of the models.

2. Importing and Using a New Dataset in PAT

2.1. Uploading and Modifying a Dataset

To upload a dataset to the PAT website, from the CloudCAT landing page (i.e., the **Data** tab) click **Upload File**. A dialog box will open. From here navigate to the dataset of interest and click **Open** (see Figure 2-1).

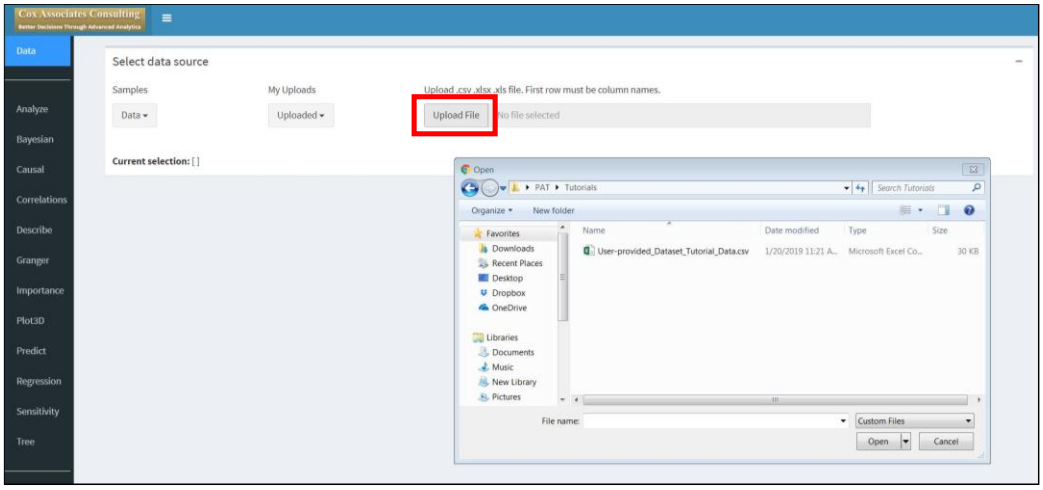

Figure 2-1. Location and use of the 'Upload File' option within the 'Data' tab of PAT.

Once the data have been uploaded, additional options and a preview of the data will appear below. The **Select columns** field allows users to select specific columns of the input dataset to be included or excluded from analysis within PAT. By clicking within this selection box, a drop-down list of the available columns of the dataset will appear, from which users can select the column names of the dataset to be included in the analysis (see Figure 2-2). Alternatively, users can click the **Select/deselect all columns** button below the drop-down box to load all columns of the dataset into the selection box (1; see Figure 2-3).

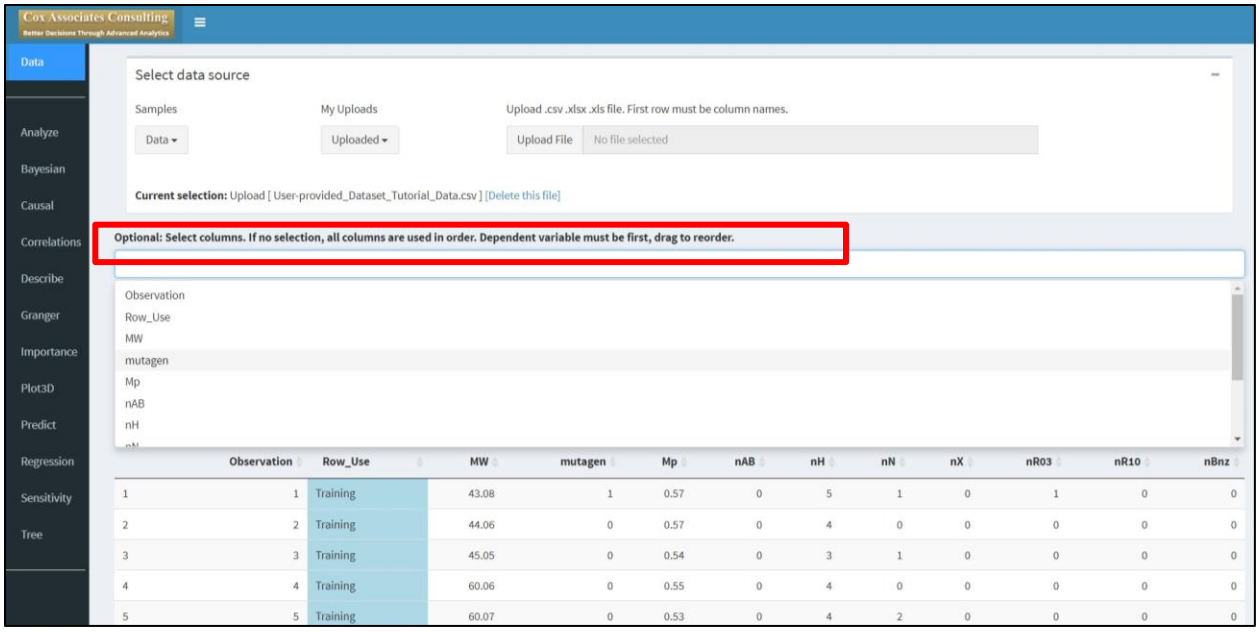

Figure 2-2. Example use of the 'Select Columns' box within the 'Data' tab of PAT.

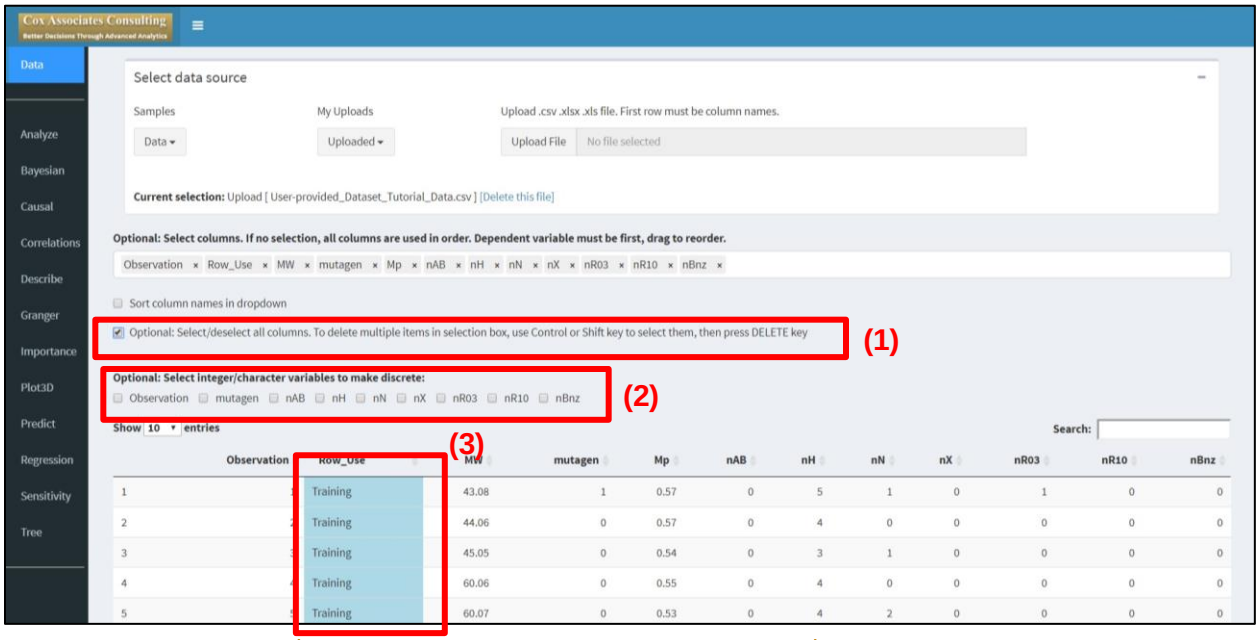

Figure 2-3. The 'Select/deselect all columns' and 'Select integer/character variables to make discrete' fields within the 'Data' tab of PAT.

Once the desired list of columns for analysis has been added into this field, users can rearrange the order of the columns by clicking and dragging the column name of interest and placing it in the desired order. The main purpose of this would be to reorder the columns such that the first column (which will be used as the dependent variable in PAT) is of the required binary format (i.e., each row contains a '0' or '1'). Users can also remove a column name from the list of those to be analyzed by clicking the 'X' next to each column name in the list. Changes made to the columns of the dataset will be reflected in the data preview table below these selection boxes. For this tutorial, we will click

the **Select/deselect all columns** button, remove the **Observation** column by clicking the 'X' button next to the column name, and rearrange the **mutagen** column to be the first column of the dataset, as this is the desired dependent variable of the dataset and is in binary format. As explained in the Existing Dataset Tutorial, this **mutagen** column denotes whether the chemical represented in each row was observed to be mutagenic ('1') or non-mutagenic ('0').

An additional option on this page is the **Select integer/character variables to make discrete** option (2; see Figure 2-3). By checking the selection box for this feature next to a given column name of the input dataset, the variable will be treated within PAT as a discrete factor (i.e., as a category variable). An example of such a discrete category variable is ethnicity. While the **Row_Use** column of the example dataset is taken by default as a discrete field by PAT (as indicated by the blue shading of this column, see box 3 in Figure 2-3), we will not set any other columns of the dataset to be discrete fields for this tutorial.

2.2. Saving a Dataset to the Cloud

The uploaded dataset, along with any changes that have been made to it within the **Data** tab of the PAT tool, can be saved onto the Cox Associates Consulting cloud for future access. Under the data preview table, in the **Table name in cloud** field, enter a name for the dataset, excluding the file extension, and then click the **Save table in cloud** button (see Figure 2-4).

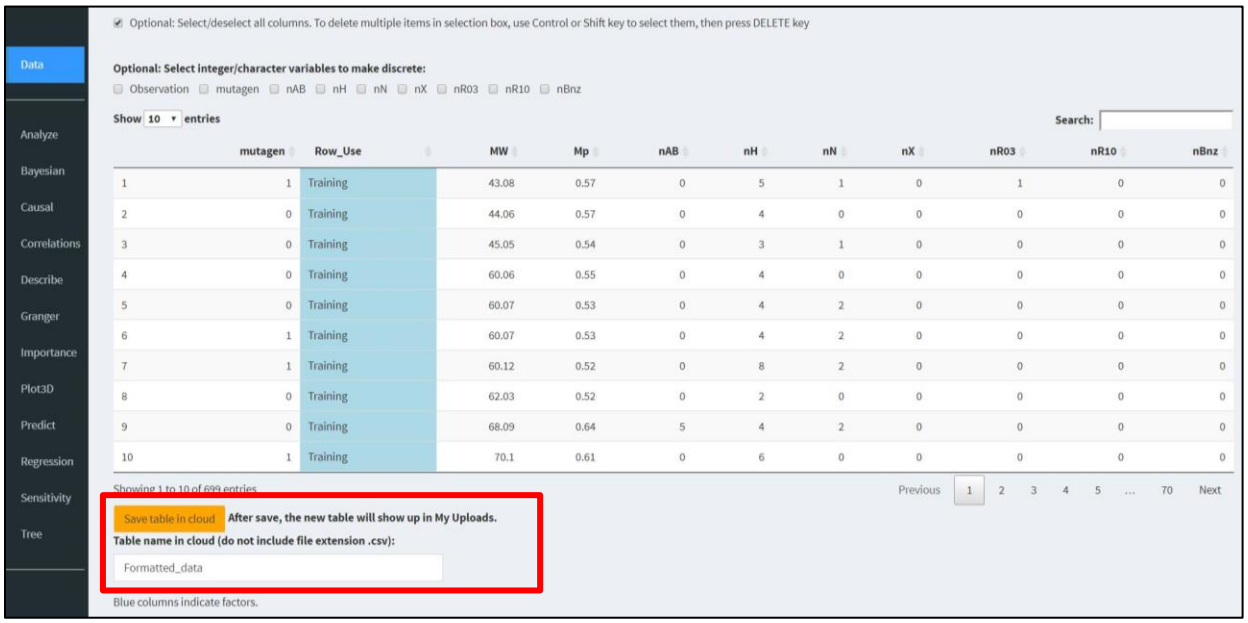

Figure 2-4. Location and use of the 'Save table in cloud' option in the 'Data' tab of PAT.

This will save the input file, along with any changes that have been made to it, to the Cox Associates Consulting cloud, and take the user back to the original PAT webpage. The saved dataset will now be available to the user via the drop-down box under the **My Uploads** heading (1; see Figure 2-5).

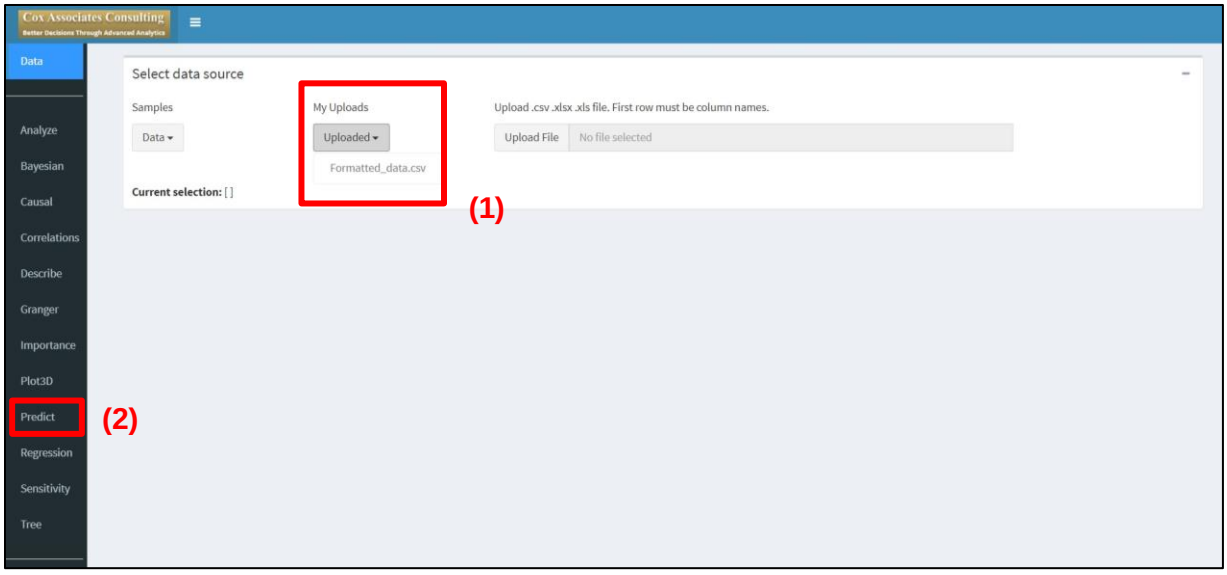

Figure 2-5. Location of the saved user-provided dataset under the 'My Uploads' heading and the 'Predict' tab.

By selecting this saved dataset from the **My Uploads** drop down menu, the dataset along with the changes that were previously implemented will be loaded back into PAT.

2.3. Using Top Rows of the Dataset as Training Data

Once the dataset of interest has been loaded into PAT, select the **Predict** tab from the ribbon on the left-hand side of the webpage (2; see Figure 2-5). This will take users to a new page of **Prediction Options**. Most of the options on this page are described at length in the Existing Dataset Tutorial. Here we focus only on the **Train data percentage** box and the use of the **Use top rows as train data only** option.

As discussed in Section 1.2, for this tutorial we will use the top 80% of the rows to train the prediction models, so we set the **Train data percentage** box to 80 (1; see Figure 2-6). For the **Use top rows as train data only** selection box, we will leave this option checked, which will instruct PAT to use the user-defined top percentage of the rows (80%) to train the prediction models, and the remaining 20% of the rows (in the bottom portion of the dataset) to test the prediction models.

To increase the processing speed of PAT, we will also select the **Tree** filter from under the **Select filters** heading as well as both of the **Pre-process data** options (see Figure 2-6). See the Existing Dataset Tutorial for more details on these pre-processing options.

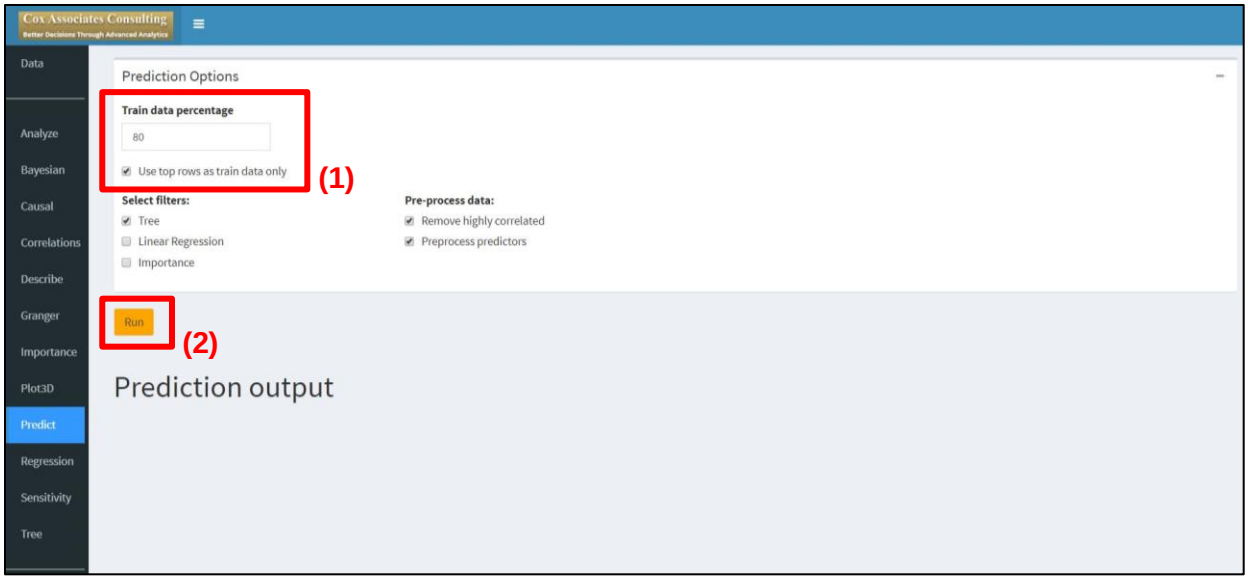

Figure 2-6. Train data percentage and use of the 'Use top rows as train data only' option in PAT.

Once all of the **Prediction Options** are set, click the **Run** button (2; see Figure 2-6). With the input dataset used here employing the configurations outlined in this Tutorial, PAT will take less than five minutes to finish processing. When processing is complete, the outputs of PAT are displayed in several sections under the **Prediction output** heading.

3. Advanced Outputs of PAT

Many of the output figures and tables of PAT are discussed at length in the Existing Dataset Tutorial. In this Tutorial, we focus on describing several of the more in-depth output figures of PAT, as well as how to export the prediction data generated by PAT.

3.1. Exporting the Predicted Results to a File

The **Predicted results from all models** section presents in tabular format the outputs of PAT (see Figure 3-1). The table in this section contains the observed outcomes for each row of the uploaded dataset (the **Observed** column), an indicator for whether or not the row was used in the training dataset (the **inTrain** column), and the predicted outcome from each of the prediction models developed within PAT (the **earth** through **glmboost** columns. Since we parameterized PAT to use the first 80% of rows of data to train the prediction models, the **inTrain** column shows that the first 559 of the 699 rows in the dataset were used in training the prediction models (i.e., the **inTrain** column contains a '1') while rows 560 through 699 (the last 20% of the rows) were not used in training (denoted by a '0' in the **inTrain** column; see Figure 3-1).

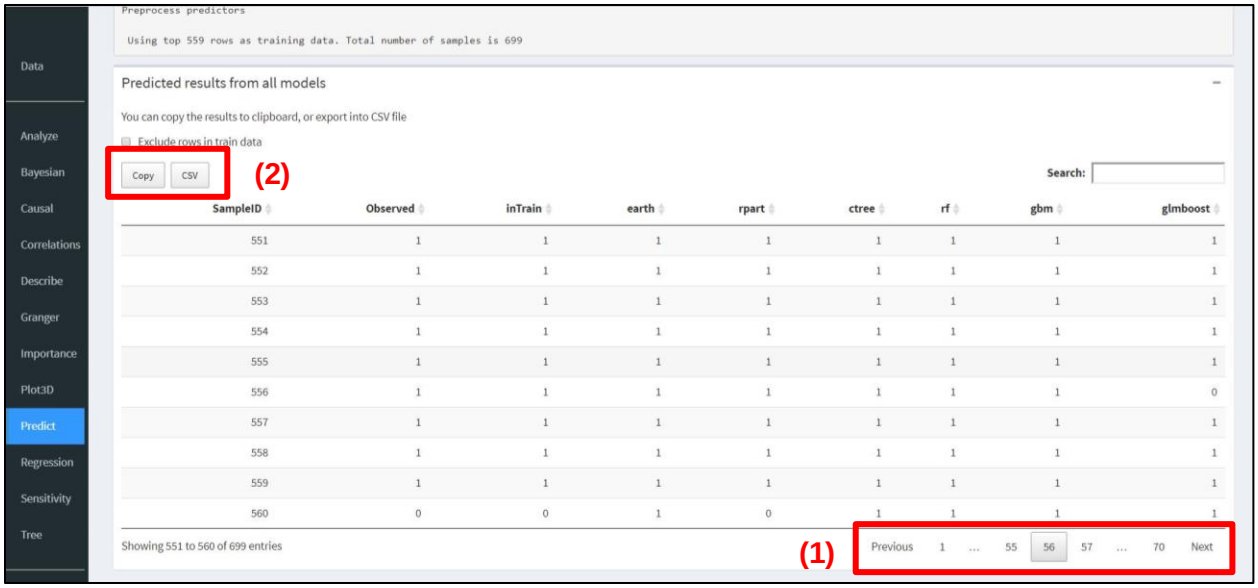

Figure 3-1. Predicted results from all models table, denoting the transition between rows used in training and those not used in training the prediction models. Buttons for navigating the output data and saving off the data to the user's computer are also indicated.

Users can peruse the information in this table by using the navigation buttons below the bottom right of the table (1; see Figure 3-1). To navigate these data more easily, however, the information in this table can be exported to an external text file or a Microsoft Excel file. To do this, click the **Copy** button (2, Figure 3-1) above the top-left portion of the table, which will copy the information in the entire table to the user's clipboard (a **Copy to Clipboard** notification will then appear briefly on the screen). This information can then be pasted into an empty text or Microsoft Excel file. Alternatively, by clicking the **CSV** button (2, Figure 3-1), PAT will export the information in this table directly to a comma-separated values (CSV) file that will automatically download to the default location that the user's web browser downloads files to (when using Google Chrome a download bar will appear along the bottom of the webpage; see Figure 3-2). This CSV file will be named "prediction-YYYY-MM-DD.csv", where "YYYY-MM-DD" denotes the year, month, and day that the PAT analysis is performed. This CSV file can then be opened using either a text file viewer (such as Notepad) or by using Microsoft Excel (see Figure 3-2).

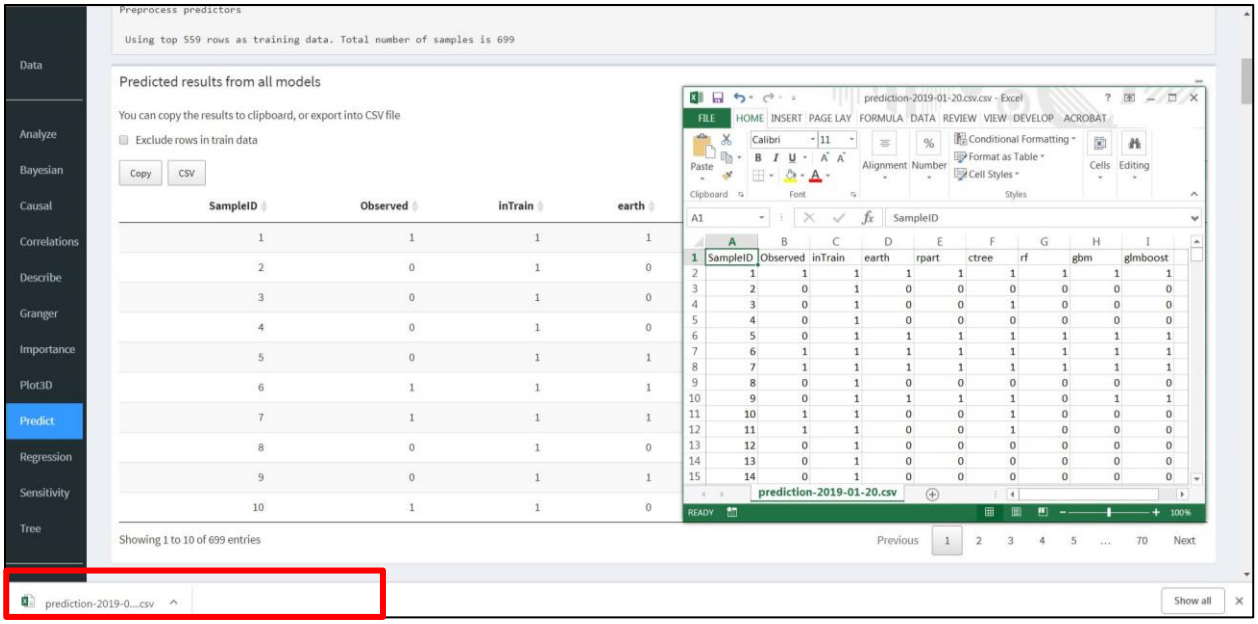

Figure 3-2. Downloaded CSV file of the 'Predicted results from all models' table.

3.2. Advanced Output Figures

The **Confusion Matrix**, Areas Under Curve (**AUC**), and **Observed vs. predicted heatmap** output sections of PAT are described at length within the Existing Dataset Tutorial. Here, we focus on the **Observed vs. predicted for each model**, **Earth Prediction**, and **Lift Charts** sections of the PAT output.

3.2.1. Observed vs. Predicted for each model (Calibration Curves)

All of the PAT output figures discussed in the Existing Dataset Tutorial (namely the confusion matrices, the AUC plots, and the heatmap) deal with binary outputs ('1' or '0', for this particular instance denoting 'mutagenic' or 'non-mutagenic', respectively). However, PAT generates these predicted *binary* outcomes based on a modeled *probability* of the outcome being positive, with separate probabilities being generated for each outcome (or row) for each model. For a given row of data that a particular model is making a prediction for, if the probability of the outcome being positive is at least 0.5, then the outcome in this row is slated as a 'positive' prediction. If the probability of the outcome being positive is less than 0.5, it is predicted by that model to be a negative outcome.

The relationship between the *predicted* probability of outcomes being positive and the frequency of these classifications being *observed* to be positive is summarized across all of the data in the **Observed vs. predicted for each model** section of the PAT outputs (see Figure 3-3), with one panel of the figure denoting results for each prediction model.

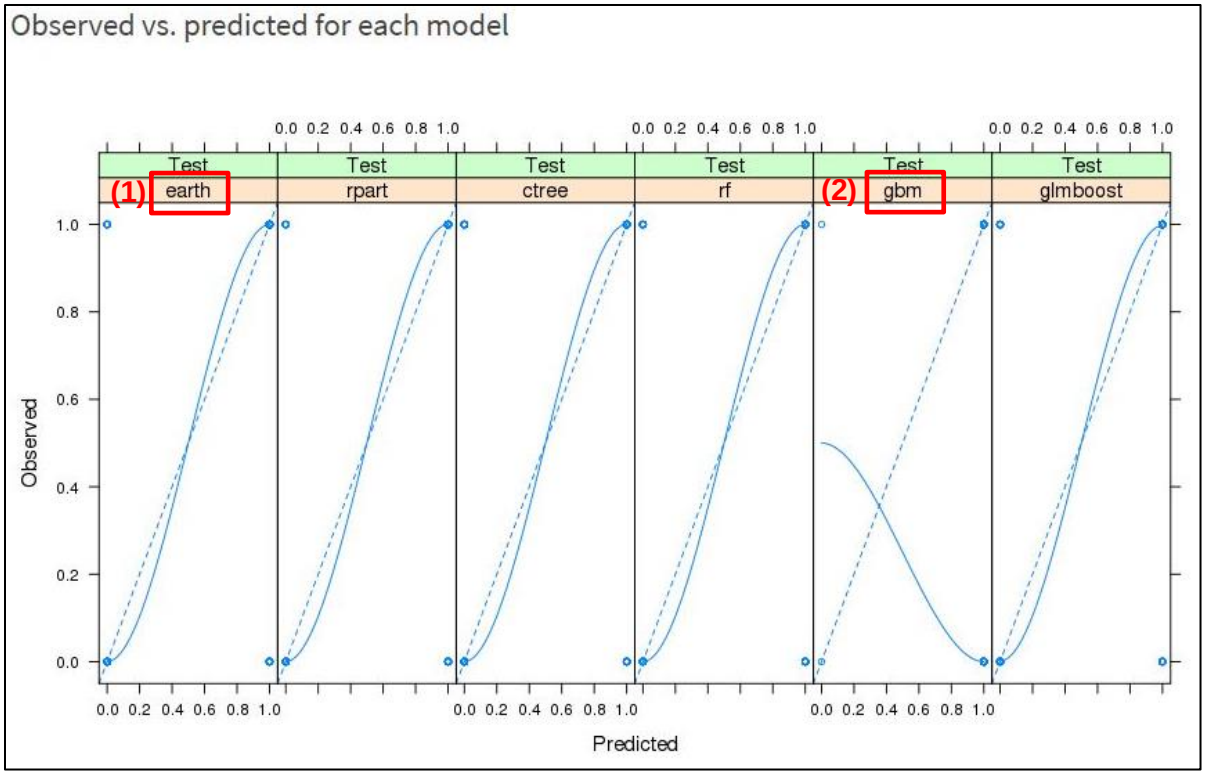

Figure 3-3. Observed vs. predicted (calibration) curves for each model.

The plots within this figure, referred to as "calibration curves", provide information on how well the probability-based prediction models performed relative to observations. The x-axis denotes the model-predicted probability of the outcome being positive (i.e., in the case of this dataset, that a chemical is mutagenic). The y-axis denotes the corresponding frequency with which chemicals were observed to be positive (i.e., mutagenic). In other words, for these data the panels for each of the models in Figure 3-3 can be interpreted as follows: when the given prediction model provides an [x-axis probability] of a chemical being a mutagen, [y-axis] of these chemicals are observed to actually be mutagens.

Perfect calibration (i.e., x-axis values equal y-axis values) is denoted by the dashed blue lines in the panels, while the probability distribution generated by PAT for each model is denoted by the solid blue lines. The shapes of these solid blue lines provide insight into how the models performed at different places in the probability distribution. For example, the **earth** model (1 in Figure 3-3) shows values below the perfect calibration line below 0.5 on the x-axis, but above the perfect calibration line above 0.5 on the x-axis. This implies that the **earth** model overestimates mutagenicity in the lower part of the predicted probability distribution (i.e., there are actually fewer observed instances of mutagenicity than the model predicts), and underestimates mutagenicity at the upper end of the probability distribution (i.e., there are actually more instances of mutagenicity than the model predicts). Conversely, the **gbm** calibration curve (2) denotes a pronounced underestimation of positive outcomes in the lower half of the probability distribution, and a pronounced over estimation of positive outcomes in the upper half of the distribution. For information about the R function used to generate these figures, see the information at this website.

It should be noted that calibration curves that are poorly calibrated do not necessarily translate to poor model performance. This is because these continuous probability distributions are converted to binary outcomes by simple rounding (i.e., values of 0.5 and above are set to 1, or 'mutagenic' in this analysis, while values below 0.5 are set to 0, or 'non-mutagenic'). For a better metric delineating how well a particular model performs at predicting outcomes for a dataset, see the discussion of Confusion Matrices and Balanced Accuracy in the Existing Dataset Tutorial.

3.2.2. 'Earth' Prediction Plot

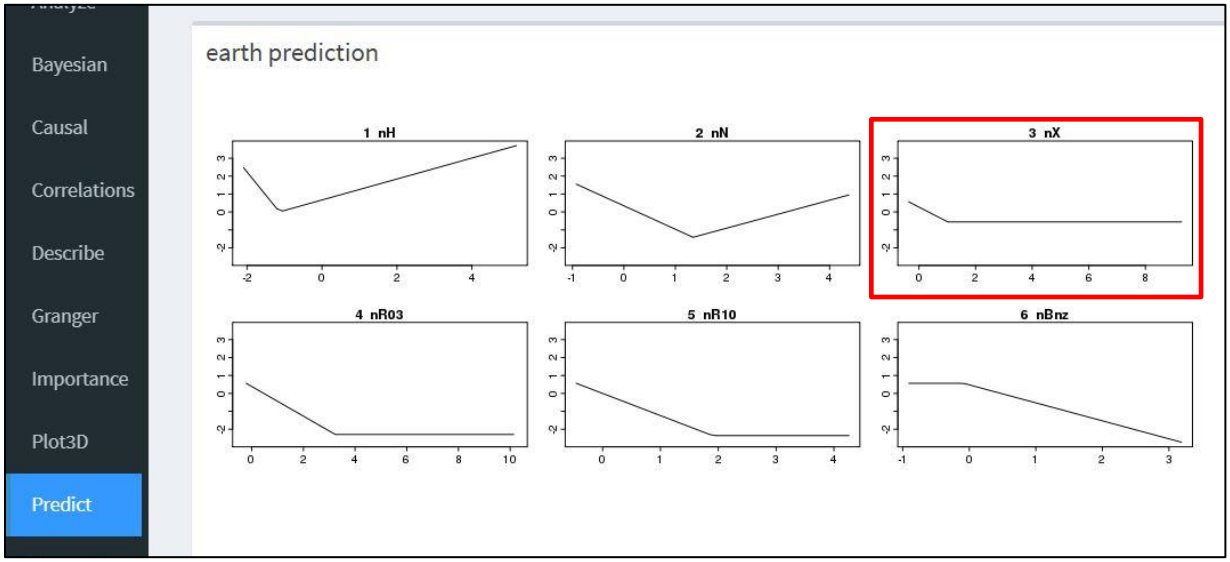

Figure 3-4. The 'earth prediction' plot generated within PAT.

The **earth prediction** section of the PAT outputs displays several figures that contain results of the **earth** prediction model when values for only one of the predictors are varied at a time, with all other predictors being held constant (see Figure 3-4). For example, the **3 nX** panel of this section displays the results of the **earth** prediction model when the **nX** predictor is varied between its lowest and highest values. The values displayed along the x-axis are the range of values of the **nX** predictor that are input to the **earth** prediction model, while all other predictors input to the model are held constant at their median value. The results of the **earth** prediction model, with these values of the **nX** predictor being the only varying inputs, are then plotted on the y-axis.

Each panel of the **earth** prediction output section contains the results of a similar analysis, run with a different input predictor being the only varied parameter. For more information on these figures, see the information about the R function used to generate these figures at this website.

3.2.3. 'Lift Charts' Section

Under the **Lift Charts** heading of the PAT output, three different model performance plots are displayed for each of the generated prediction models (see Figure 3-5).

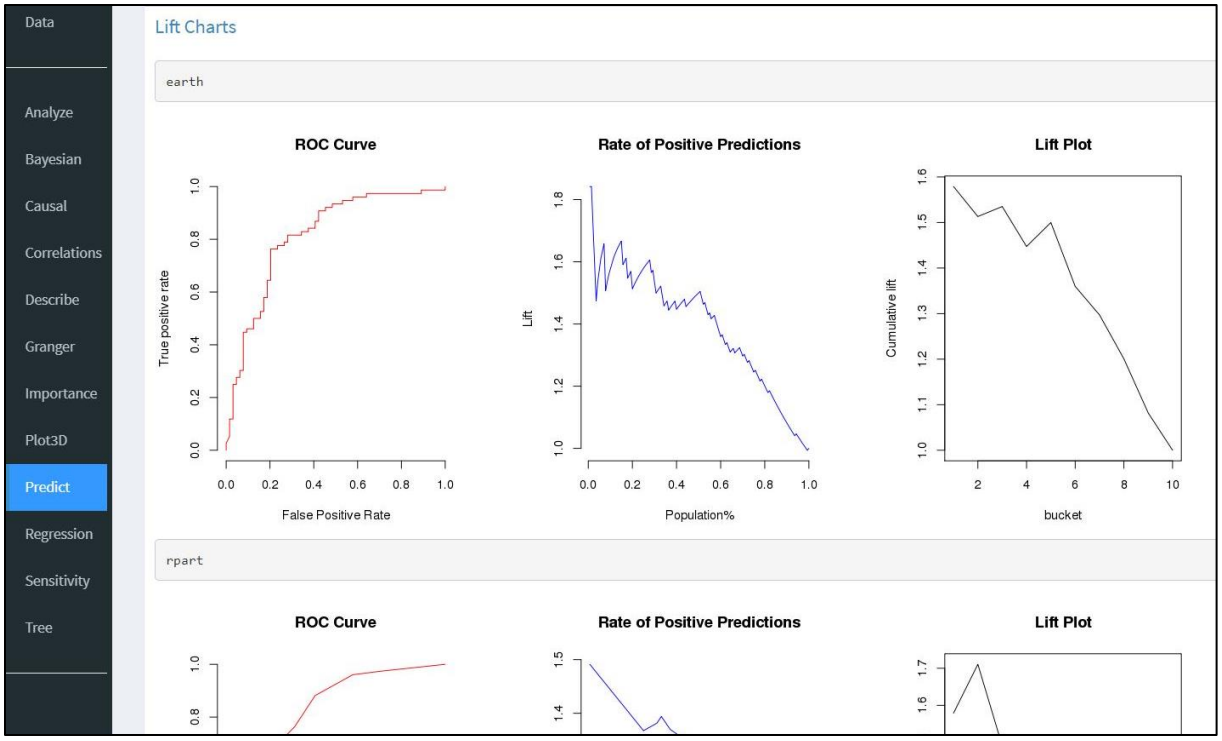

Figure 3-5. Example of the output figures for a prediction model under the 'Lift Charts' section.

These separate plots are discussed in the following sub-sections, focusing on the Lift Chart output for the **earth** prediction model.

3.2.3.1. ROC Curve Plot

Like some of the other output figures of PAT, receiver operating characteristics (ROC) curves display a graphical representation of how well a prediction model performed at predicting the outcomes of a given dataset (see Figure 3-6). As explained in Sayad, 2019, ROC curves plot the false positive rate (i.e., the rate at which a binary '1' is predicted for observations that were in reality '0') versus the true positive rate (i.e., the rate at which a binary '1' is predicted for observations that were in reality '1'; see the red line in Figure 3-6). For a prediction model that generates outcomes at random, an ROC curve would display as a diagonal straight line (i.e., values along the x-axis equal those on the y-axis).

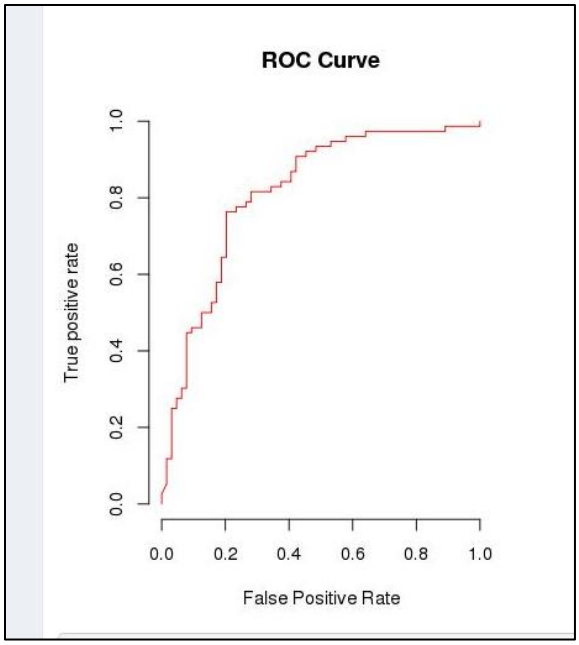

Figure 3-6. Example of an ROC Curve plot for the 'earth' prediction model.

For models that predict outcomes of a dataset relatively well, the resultant ROC curve generally has a very steep slope in the left-portion of the graph, which then flattens out towards the top and right part of the graph (i.e., with increasing false positive rate, the true positive rate rapidly becomes larger than the false positive rate). This pattern can be seen in the ROC curve for the **earth** prediction model in Figure 3-6.

For a given prediction model, the closer the ROC curve line is to the diagonal line of true positive rate=false positive rate, the closer the model is to being a random outcome predictor (i.e., it is not a good prediction model for the dataset). For more information on ROC Curves, explore the material provided at Sayad, 2019 or at the Data School website.

3.2.3.2. Rate of Positive Predictions and Lift Plots

As explained by Sayad, 2019, a Lift Plot enumerates how much more likely a user is to select positive outcomes from a dataset when using a given prediction model as compared to trying to select positive outcomes by random selection (i.e., not using the prediction model). In other words, it is a "ratio between the results obtained with and without the model" (Sayad, 2019). PAT generates two types of Lift plots, labeled as a **Rate of Positive Predictions** plot (shown in Figure 3-7) and a **Lift Plot** Figure 3-8.

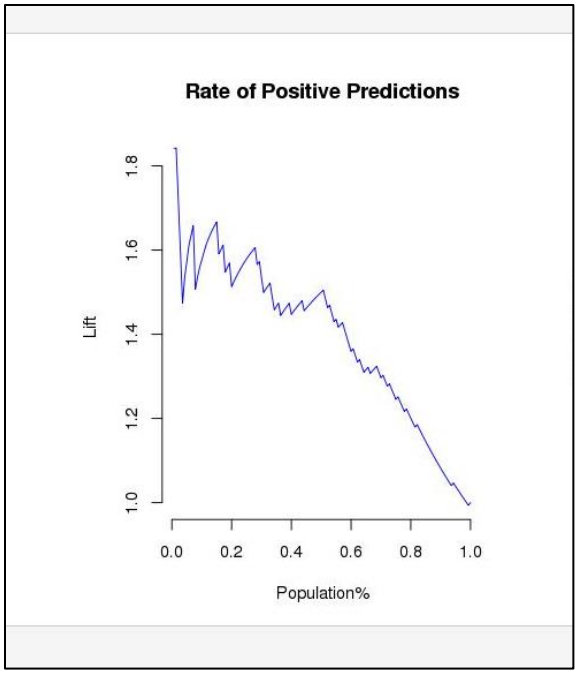

Figure 3-7. Example of a 'Rate of Positive Predictions' lift plot for the 'earth' prediction model.

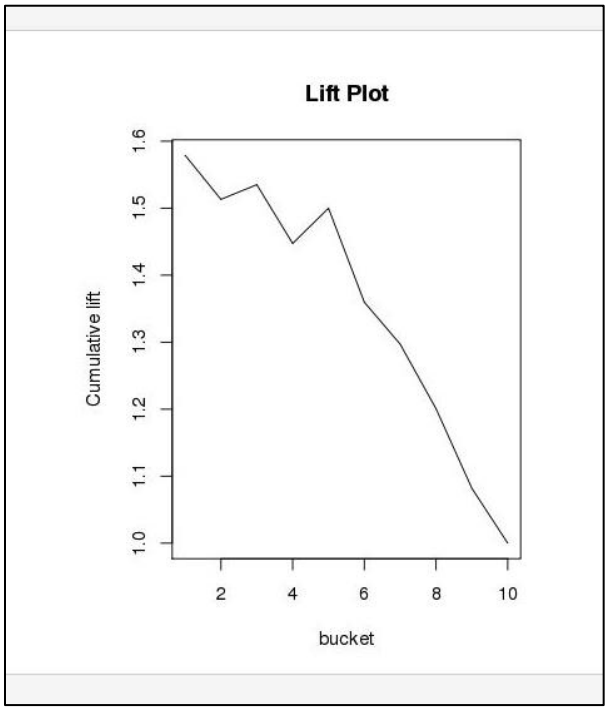

Figure 3-8. Example of a 'Lift Plot' for the 'earth' prediction model.

In the Lift Plots generated by PAT, the x-axis denotes the percentages of the data being selected (e.g., in Figure 3-7, the 0.2 denotes 20% of the data, while in Figure 3-8 'bucket' 2 = refers to the 2nd decile of data (i.e., 20% of the data)). The y-axis denotes the increased fraction of positive outcomes that will be identified using the given prediction model relative to using random guessing (i.e., not using the model).

Based on the data in Figure 3-8, for the **earth** prediction model in the current PAT analysis, if 40% of the input data is being analyzed, using the **earth** prediction model to identify positive outcomes (i.e., mutagenic chemicals) will result in 1.45 times more mutagenic chemicals being identified than by using random guessing to identify mutagenic chemicals. When 100% of the data are selected (i.e., '1.0' along they x-axis in Figure 3-7 or 'bucket 10' in Figure 3-8), the prediction model does not add any predictive power to identifying mutagenic chemicals over random guessing because all chemicals in the dataset have been selected.

For more information on Lift Plots, see Sayad (2019) or the information at this webpage.

3.2.4. Other output from each model

Under the **Other output from each model** section, PAT generates several additional figures, one for each of the models generated within PAT. For more information on these figures, please refer to this webpage.

4. References

Sayad, S. (2019). An Introduction to Data Science: Model Evaluation — Classification (webpage). Accessed February 21st, 2019. Available online: http://www.saedsayad.com/model_evaluation_c.htm