A User Guide to TwoRavens: An overview of features and capabilities

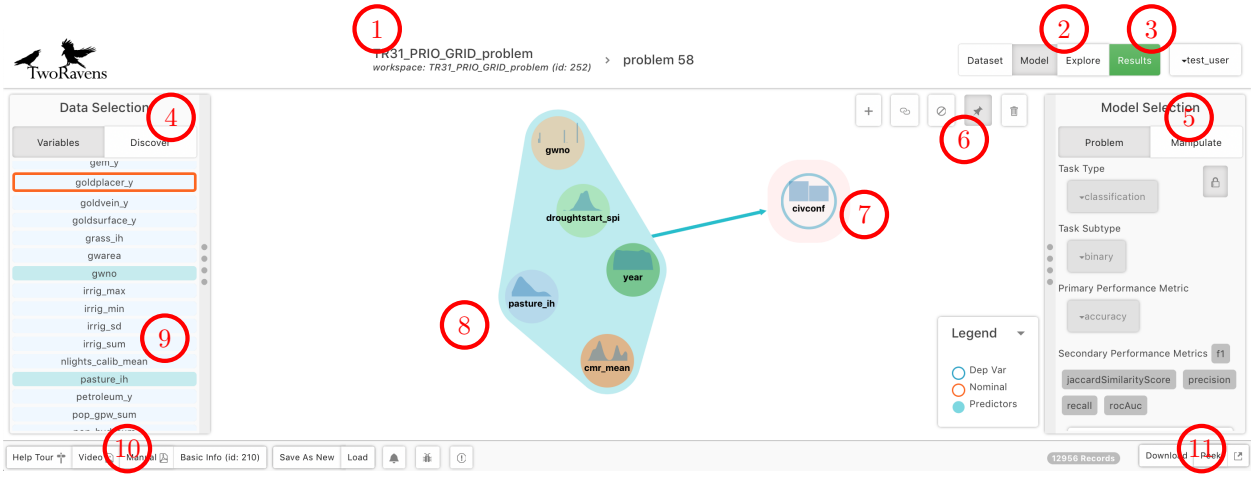
January 16, 2020*

1 Overview

The DARPA Data Driven Discovery of Models (D^3M) program automates the methods in data science to create empirical models of real, complex processes. D3M enables non-expert users to make predictions from data without the need for data scientists. The program accelerates scientific discovery and intelligence analysis by automatically searching the complex model space and discovering and explaining models to users.

TwoRavens (a D^3M project) is a data-driven web application designed to bring researchers to insights fast. It facilitates intuitive machine learning, model discovery, and data exploration for researchers. As our intelligent back-end automatically seeks interesting relationships in the data and builds models to predict outcomes, researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis in an interactive paradigm we call *human-guided machine learning*.

Below is a quickstart guide to features available when the interface opens:



Core Features

1	Dataset Description	Provide high-level description of dataset.
2	Explore Mode	Click this link to change the display to move to data exploration mode.
3	Results Mode	Press this button to start finding solutions to this problem.
4	Problem Discovery	Press this button to go to a table of suggested problems found in this dataset.
5	Manipulations Tab	Create new features, subset, or augment dataset.
6	Control Buttons	These icons help control the model in the center panel.
7	Target Variable	In this problem view, the target variable to be predicted.
8	Predictor Group	The set of variables currently being considered as predictors.
9	Variable List	Remove or add variables from the problem view by clicking their name.
10	Help Buttons	Press these buttons to bring up a help manual or a help video.
11	Peek Button	Press this button to see selected observations of the raw data.

*Current version of this document available at http://2ra.vn/guide

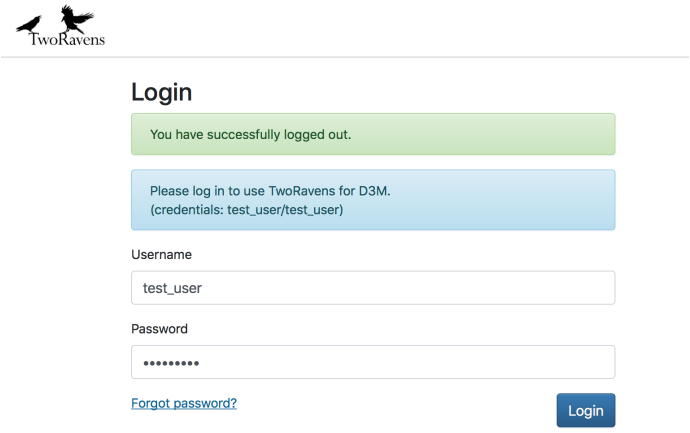
2 Getting Started

2.1 User Login and Account

You will be given your own login credentials when provided a link to the system for training and Experiment. In addition to the login using D^{3}M credentials which give you access to the server, the TwoRavens platform has an additional login system. As a test user, you can log in simply as:

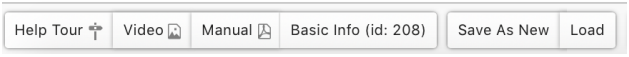
Username `test_user`
Password `test_user`

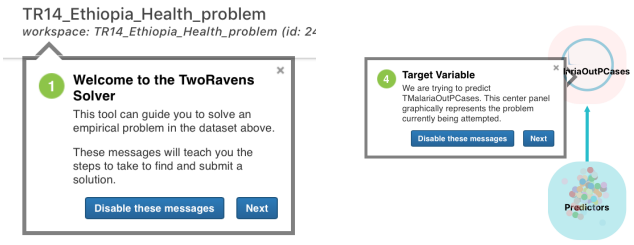
These credentials are also noted in the blue box on the login screen if forgotten. Enter these credentials to be a test user, and press the blue "Login" button on the bottom right.



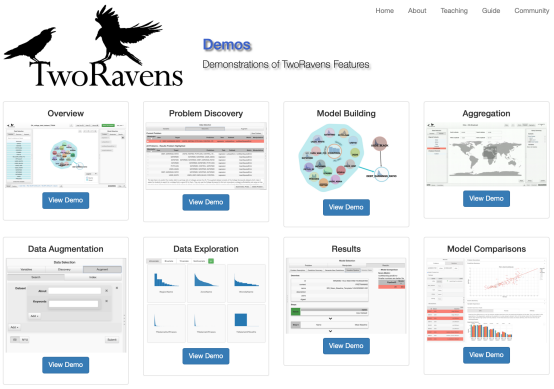
2.2 Help

Help Buttons: In the footer of the page, are three buttons to bring up additional help.





Help Tour: The first help button brings up a guided tour of the interface with a series of help signposts describing and pointing to features. This tour can be turned off or started again at any time.



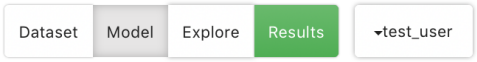
Help Video: The third help button brings up a page of guided videos demonstrating different available portions of the TwoRavens workflow.

The third help button brings up the latest version of the full manual, which you are currently reading.

3 Modes

The core concept of TwoRavens is *mode*. Different modes signify different main goals a user might be attempting in the current part of their analysis. When the system is in different modes, the features available to the user will be swapped or modified to better work for that intended task. The following outlines the available modes for users, and their uses.

Mode Buttons: There are four mode buttons in the top right of the header. These allow switching between modes.



Dataset Mode gives different ways for a user to bring data into the system. Here a user can upload new datasets, switch between datasets already available, and access the Datamarts. Datamarts are a way to augment current datasets with additional relevant data to a problem.

Model Mode allows a user to construct models of interest, specify a target to predict, explore the automatically discovered problems suggested by the system, override any metadata created for features and change the problem characteristics of the model.

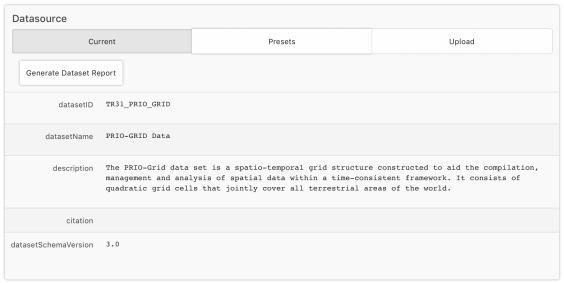
Explore Mode allows a user to perform exploratory data analysis by intuitively generating visualizations of relationships between variables.

Results Mode allows a user to automatically generate machine learning solutions for a user constructed model, explore previously constructed models to gain insights into model performance, and understand the relative importance of features within and across different machine learning solutions.

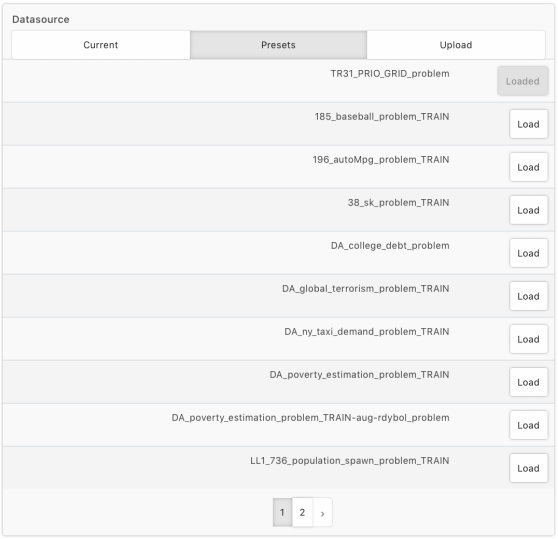
Typically a workflow moves across these modes in order, but a user is free to move between them as needed. We now detail features available in each mode.

4 Dataset Mode

Dataset mode has three possible panels to help you understand the current data or load in new datasets.



Current Tab gives a quick summary of any available information about the current dataset loaded into TwoRavens.



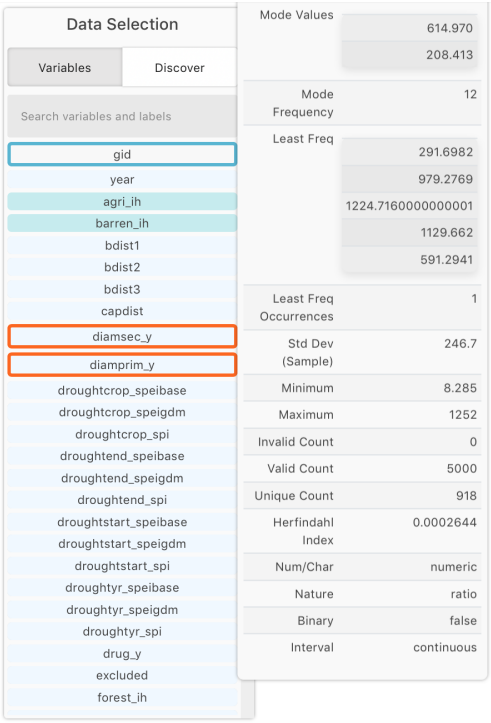
Presets Tab brings up a list of all the other available datasets that TwoRavens currently has available. The current dataset will be signified as "Loaded" while the user can switch to any of the other datasets by pressing their "Load" button.

Upload Tab allows the user to add a dataset into the TwoRavens system from a local file. Currently, this works for files formatted as `.csv`. When a file is uploaded, the TwoRavens metadata profiler will make automated judgements about the nature of the variables present, compute summary statistics, and attempt to discover low dimensional relationships that may be of interest to the user. When a new dataset is uploaded, it will also be loaded as the current dataset in TwoRavens.

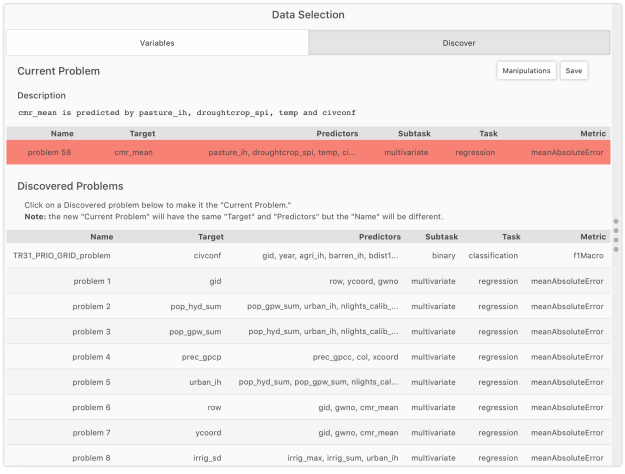
5 Model Mode

In model mode, there are left and right panels, and a central control surface. The left panel facilitates learning about the variables, the central control surface lets a user build up a directed graph among the features to highlight information they know about the data and build relationships among the variables to model, while the right panel lets them specify exactly how that model should be tailored.

5.1 Model Mode: Left Panel



Variables Tab: The default panel shows a list of the features available in the current dataset. Mouseover of any feature name will give a table of summary statistics for that variable. Variables that have special types of metadata will be outlined in specific colors, while those that are included in the present model will be have darker backgrounds.

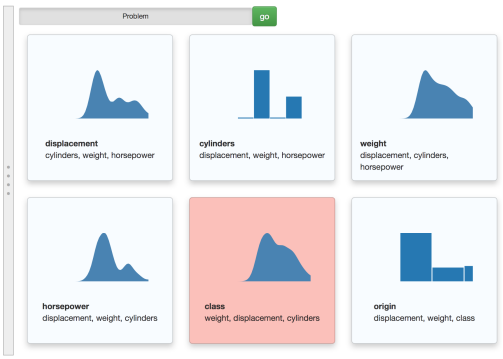


Problem Discovery: The "Discovery" panel suggests relationships between features in the data that appear to have explanatory power.

This window is reached by opening the "Discovery" tab in the left panel.

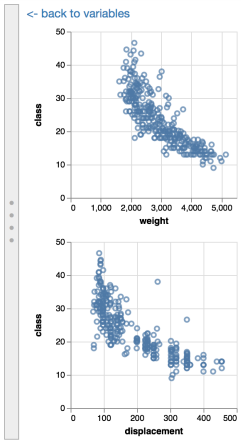
We call these *discovered problems* in the sense that they are relationships we could attempt to build a machine learning algorithm to model and use to predict the target variable. We also sometimes refer to these as *discovered relationships*.

Explore Relationships: If you click the "Explore" button in the footer while the Discovery table is open then all the available discovered problems will be available as tiles for exploration.



If the left panel is blocking explore tiles, you always click the sidebar on the outer edge of the panel to minimize it (click it again to maximize it).

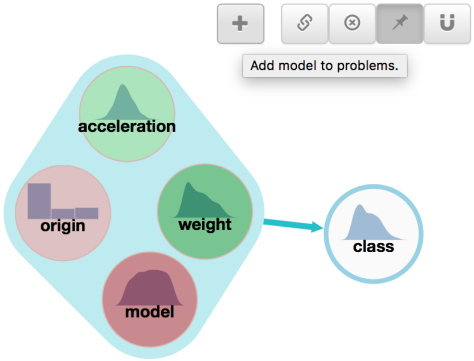
By default, the discovered problem selected in the table will also be selected in the set of tiles, but you can click any tile. Press the green "Go" button at the top of the screen and visualizations to show the relationships in that discovered problem will be constructed.



Clicking the green "Go" button will construct visualizations of all of the predictor variables in the discovered problem, against the target variable, to show the two-way relationships.
Press the "back to variables" link to return to the previous page to examine a different explored problem.

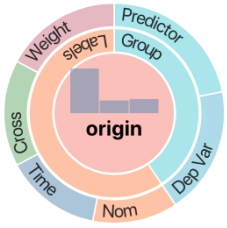
Add New Relationships:

The discovered problem table is populated with potential relationships that are automatically discovered by TwoRavens. However, if the user is interested in adding a new relationship of their own construction into this list, they can do this from the center panel.



First make sure that TwoRavens is in "Model" mode by selecting the "model" button in the footer, and select the "variable" tab in the left panel to list the available variables.

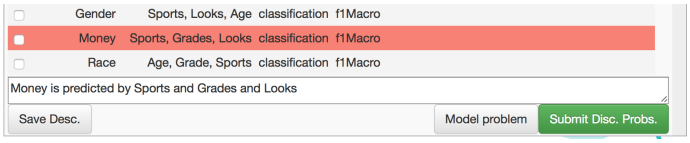
The center panel can then be used to add or remove variables of interest by clicking on their name in the left panel, and by selecting a target dependent variable by selecting a pebble and clicking the "Dep Var" arc that surrounds it.



When you have constructed a possible relationship you are interested in adding to the discovered problems table, click the "+" button in the center panel controls to add it to the table.

For more information on using the center panel to describe possible relationships between variables, see section 5.2.1 on model building.

Edit Problem Descriptions When a problem is highlighted, a short sentence description of that problem is written below the discovered problem table. If you think that sentence could be improved, you can select the text and edit it in any manner you wish.



When done, press the "Save Description" button and the revised text description of the problem will be saved as part of the submission.

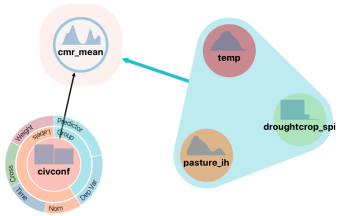
5.2 Model Mode: Control Surface

The central *Control Surface* is where variables, represented as *pebbles*, can be arranged into a directed graph to represent possible relationships to explore among the variables. The panel is made up *pebbles* that represent all the information in a variable, and can be arranged and connected to form a possible network of relationships between variables, called a *directed graph*.

5.2.1 Model Builder

The model builder in the control surface is the heart of the TwoRavens interface. Here, every variable that has been selected in the variable selection panel is represented as a circular icon called a *Pebble*. A pebble is more than just the name of a variable, it should be thought of as a container for all the information in that variable. Pebbles typically have graphs of the distribution of their variable, to emphasize that one is manipulating all the observations of a variable, and not just a name. These graphs also help make the display in this panel more informative and intuitive.

On the mouseover of a pebble, options for that pebble will appear as *arcs* or tabs around the border. These will contain possible attributes about that the user can assign to that variable. For example, the user can make a variable the outcome, or dependent variable, of the analysis, or state that a variable is nominal (categorical). Variables that have been assigned attributes will be given colored *halos* to represent this information, and a legend will build that explains the meaning of these colors. (These colors will also map back to the variables names in the variable list in the Data Selection Panel.)



Pebbles can be connected by arrows. Arrows are initiated by a two-finger or right-click. If this is dragged to another variable an arrow will be constructed between those two variables. Arrows represent possible causal relationships, that is, an arrow from A to B may mean A causes B or the event of A leads to B. Arrows may simultaneously point is both directions, for example an arrow from A to B and also an arrow from B to A. In such situations these are created as two separate arrows. Together, the set of all arrows is called a *directed graph*. Clicking on any arrow, deletes it from the graph.

5.3 Control Buttons

The control button in the model builder allow shortcuts and other control points for constructing models.

Add Model Button

The add model button takes the currently constructed model in the builder and adds it to the list of problems in the discovered problem table.

Pin Button

By default, the graph in the model moves like a force diagram, that is, it acts as though the pebbles have some repulsive force keeping them apart, and the arrows act like springs. This generally moves the pebbles into a useful array. However, if more precise control of the pebble location is desired, pressing the pin button will lock all pebbles in place. Afterwards, any pebble can be dragged to any location in the panel, and it will remain in that new location. Reclicking the pin button will revert to the force effect where the pebbles adjust themselves automatically.

Link Button

The link button is a shortcut that when clicked will link every pebble to the target dependent variable.

Unlink Button

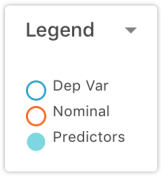
The link button is a shortcut that deletes all links in the model builder.

Wipe Button

The wipe button, when clicked will remove all pebbles from the exploration panel, leaving a blank panel.

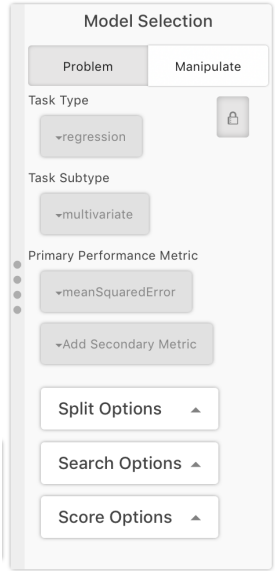
Model Legend

When tags have been applied to the pebbles of variables, to describe their attributes, they are given a colored halo as a visual identifier of the attributes of that variable. These colored halos will then appear in the legend box, together with a description of their meaning, to aid in interpretation of the relationship that has been described in the exploration panel. In the example to the right, a *dependent variable* (Dep Var) has been selected, and the legend shows this has a blue halo, while a categorical variable has also been tagged appropriately as *nominal* and the legend explains such variables are identified by an orange halo. The legend can be minimized by clicking the down arrow in the legend title bar.



5.4 Model Mode: Right Panel

The right panel of model mode can fine tune the underlying task that model that is going to be constructed is solving, and also has features for manipulating the dataset to subset rows, or construct new features.

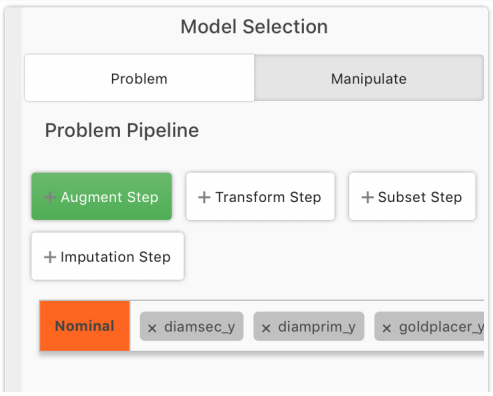


In the right panel, the "Problem" tab will allow a user to specify the exact class of problem that is being attempted to solve, and the *metric* or measure of the performance of any given solution. In current user testing, these are automatically set correctly by the system, and do not normally need to be adjusted. If a user does want to change them, clicking the lock symbol would allow each of these values to be changed by a drop down menu.

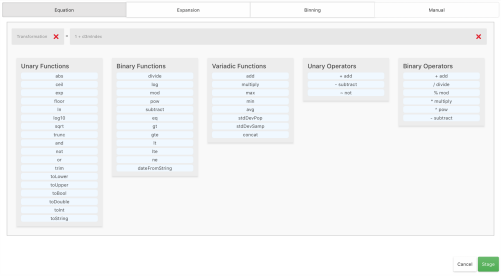
The "Manipulate" tab in the right panel allows a user to manipulate the data by a number of means, including:

- *augmenting* the data by searching for new data to combine to the present dataset,
- *transforming* the data by constructing a new feature as a user constructed function of other features,
- *subseting* the data by selecting a set smaller set of the current observations that the model should be run on.

Each of these stages manipulations to the dataset that will be performed before the dataset is searched for machine learning solutions. (In the language of machine learning pipelines, these stages are the earliest steps in the end-to-end pipeline solution.)



The "Transform Step" allows a user to specify a formula by which to construct a new features as a function of any other current features.



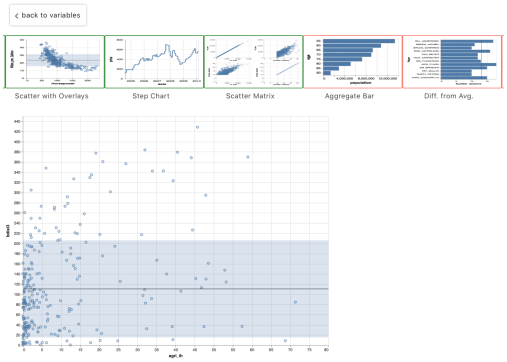
The "Augment Step" allows a user to search for additional data that can be *augmented* or added to the current dataset, so as to construct a better model. Available datasets are listed to the right. A search using the current dataset is conducted automatically when TwoRavens opens a new dataset, so search results will always be present, however, this search can be combined with any keywords the user is interested in for a better tailored search. If a dataset is selected to augment, TwoRavens will reload as if the combined dataset is a completely new datset, including recomputing summary statistics and renewing the search for discovered relationships to show in the discovery panel.

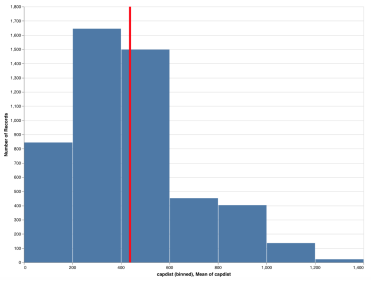
6 Explore Mode

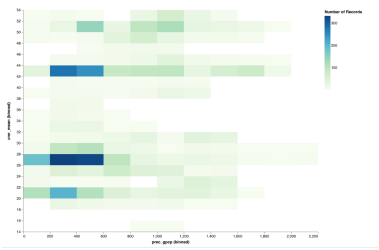
Explore mode facilitates the speedy construction of plots of variables to let a user quickly learn and understand the features present in the dataset.

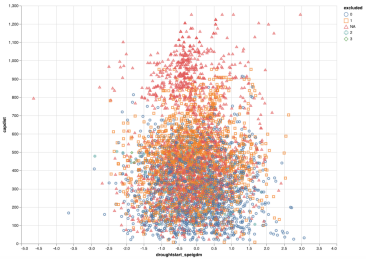
When in explore mode, the center panel will fill with tiles that can be used to select variables (if the left panel has the "Variables" tab selected), or discovered problems (if the left panel has the "Discovery" tab selected).

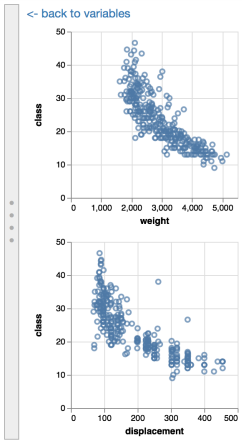
For exploring variables, select if you want univariate (distributions of individual variables), bivariate (relationships between two variables), or trivariate (three-dimensional relationships) by selecting the appropriate tab at the top. Then click on one, two, or three variables as appropriate. When you press the green "Go" button, all the possible visualizations for the selected variable(s) will be constructed and you can scroll between them using the legend in the top. Visualizations that seem to be appropriate to those variables, as judged by the system, will be outlined in green, while visualizations that are unlikely to be useful (or may be failing) are outlined in red.









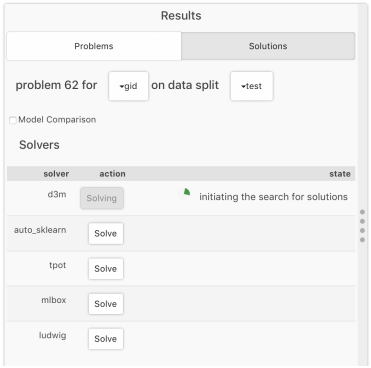


For exploring discovered problems, each tile represents the target and predictor variables, and clicking the green "Go" button will show the two-way relationship between each predictor with the target, to show the strength of the relationship in the discovered problem.

You can switch back into "Model" mode at any time using the same footer tabs, to return to the standard functionality described elsewhere in this guide.

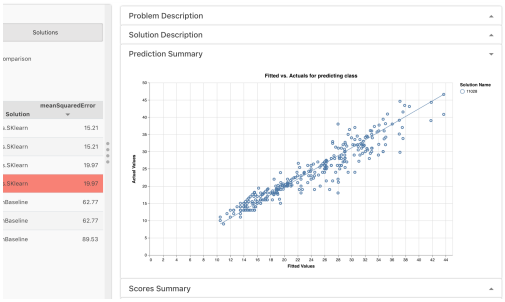
7 Results Mode

Results mode is where TwoRavens searches for solutions to the current user constructed model, and presents intuitive ways of understanding the performance, predictions and substantive meanings of these solutions.



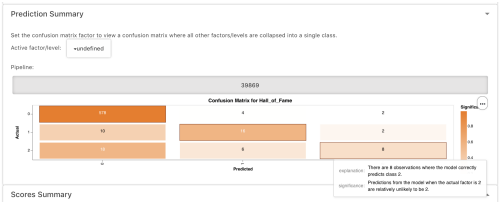
When you open results mode a list of solvers will be available in the left panel. These are each engines that search for machine learning solutions to the problem that has been constructed in model mode. You can click any or all of these solvers to begin finding solutions. For the purposes of this experiment, we suggest you use the "d3m" solver and the "ravens" that have been created in the DARPA D^{3}M program. As solutions are found by these solvers, they will accumulate in a list to the bottom of this panel.

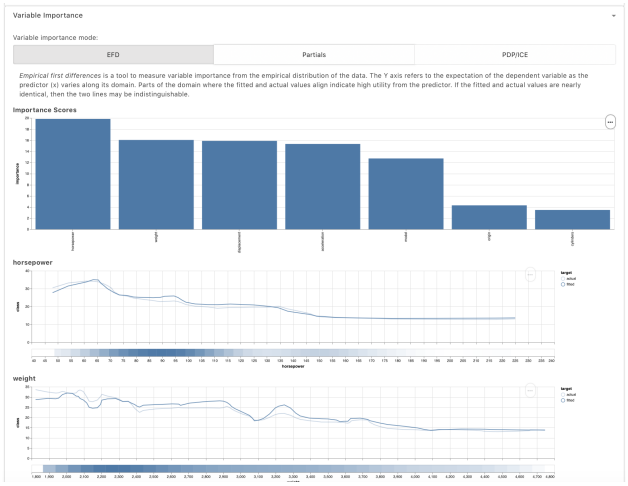
When solutions are complete, they will be presented as a list. Click on any solution to see a simple representation of how well the solution fits. Here a scatter plot of predicted versus the actual real values lines up very well.



You can compare the predictions of multiple pipelines on the graph by clicking the box marked "Make Comparisons" and selecting the models you want to compare.

Here the problem is to classify observations into types and the fit of the solution is shown as a table. In classification, all the observations that have the correct prediction will show up on the diagonal of the table (from top-left to bottom-right), so models with most observations on the diagonal are making good predictions. Mouseover on any cell of the table will give you more information about the number of meaning of observations in that part of the table.

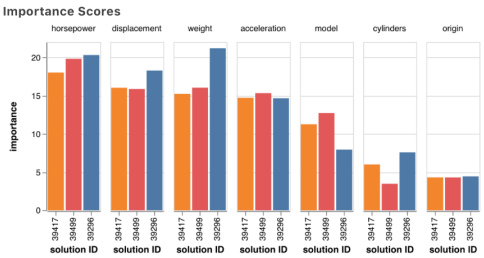


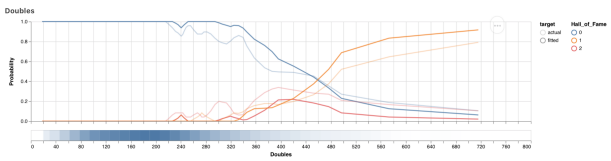


Below the graphs of model fit, the user can open up a view titled "Variable Importance" which describes which features in the model are the most important in constructing the preditions. At the top of this section is a bar graph that shows the relative impact of each variable in the model towards the final prediction; features with higher values have a greater impact on the outcome target of interest.

Below this graph are individual plots of the model's predicted value of the target variable (y-axis) for across the range of the feature (x-axis). These graphs are arranged in order of feature importance. As a precise definition, the prediction lines shown are the average target value across all observations in the test data that are close to that value in the feature dimension. A shaded line is also plotted that shows the actual empirical average of all the target values when the feature is at that value in the dimension. These lines should be close (in good models, they will often be on top of each other) or else that means there are regions in the feature where the model is making systematically bad predictions.

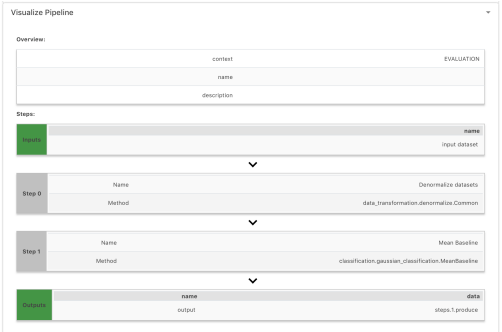
If the model comparison box is checked, then multiple models can be selected, and the feature importance bar chart will show the feature importances for all selected models. This is an easy visual diagnostic that tells whether the models agree on the relative importance of the variables in the model.





When the modeled problem is a classification problem, then there will be multiple sets of lines, color coded, each of which represent the average probability of each class being the predicted class.

Finally, opening the "Visualize Pipeline" view at the bottom allows you to see the constituent parts that have been composed together in a sequence to build a machine learning solution for this problem. In simple models, this may only have one or few steps, while complex pipelines often have a dozen or more components.

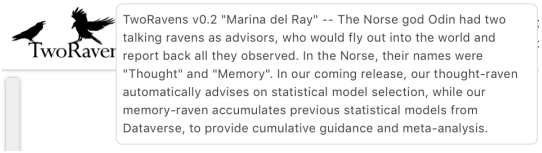


8 Header and Footers

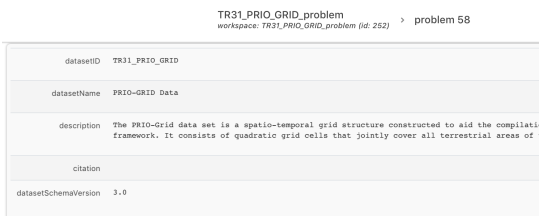
8.1 Header

The *Header* bar across the top of TwoRavens contains some preliminary information about TwoRavens and the dataset that has been opened. The header contains abstract information that is true regardless of the exploration and analysis performed, such as the name and citation to the data used, and the state of the software (for example, the version of TwoRavens being used, and the readiness of the server to run estimation).

About Image: A TwoRavens icon can be found on the top left of the interface. On mouseover, a message will describe the current version number and release name of this instance of TwoRavens.



Dataset Name: The name given to the dataset is shown in the center of the header. This name is read from the metadata file of the dataset in the data repository. If the dataset has a defined citation or a digital object identifier (DOI) these will appear on mouseover of the name (or when the name is clicked on a touch device), alongside any other available metadata about the dataset.



8.2 Footer

The footer provides quick access to help materials, logs of any system alerts or errors, and the raw data.

On the left of the footer are buttons for accessing the help materials (the tour, the videos and the manual) already described in section 2.2. Next to these are buttons that will bring up logs of any system alerts and system errors.

On the right of the footer is denoted the number of rows in the current dataset. If the current dataset is very large, at startup this number will grow as our system reads in batches of observations into our database. The "Download" button allows you to copy the current dataset into a local file. The "Peek" button selects a random set of rows from the dataset to display, for the user to explore. Only the variables in the current model will be displayed in the dataset table.