AFRL-RI-RS-TR-2021-194

TWORAVENS: INTUITIVE STATISTICAL EXPLORATION, MODEL EXTRACTION, AND CURATION

HARVARD JOHN A. PAULSON SCHOOL OF ENGINEERING AND APPLIED SCIENCES

NOVEMBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-194 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
GENNADY STASKEVICH
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2021	FINAL TECHNICAL REPORT	MARCH 2017	MAY 2021

4. TITLE AND SUBTITLE

TWORAVENS: INTUITIVE STATISTICAL EXPLORATION, MODEL EXTRACTION, AND CURATION

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-17-2-0114	N/A	62702E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
D3MP	S0	08

6. AUTHOR(S)
James Honaker and Vito D'Orazio

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Harvard John A. Paulson School of Engineering and Applied Sciences
Science and Engineering Complex
150 Western Ave, Boston MA 02134

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/RISC
525 Brooks Road
Rome NY 13441-4505

DARPA/I2O
675 N. Randolph St.
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)
RI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RI-RS-TR-2021-194

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The DARPA Data Driven Discovery of Models (D3M) program automates the methods in data science to create empirical models of real, complex processes. TwoRavens, a D3M system, is a Web-based platform for automated machine learning and statistical analysis. The goal is to allow the domain expert, in concert with our system, to complete a high quality, predictive and interpretable model without a statistical expert. To do so, the system facilitates intuitive machine learning and model interpretation, model discovery, and data exploration, as researchers impart substantive knowledge about their data and own research questions to guide the automated generation of AI assistance for data analysis through human-guided machine learning.

15. SUBJECT TERMS
Automated Machine Learning, Human-guided Machine Learning, Machine Learning Pipelines, Model Interpretation.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	28

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
GENNEDY R. STASKEVICH	**N/A**

PREVIOUS EDITION IS OBSOLETE.

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1 SUMMARY

TwoRavens quickly brings machine learning insight from data.

- It allows a domain expert to quickly complete high quality, predictive and interpretable models without any statistical or machine learning expertise.
- It helps a data scientist to rapidly search through the entire machine learning catalog to build quality baseline models for insight and further elaboration.

To do so, the system facilitates intuitive machine learning and model interpretation, model discovery, and data exploration. 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.

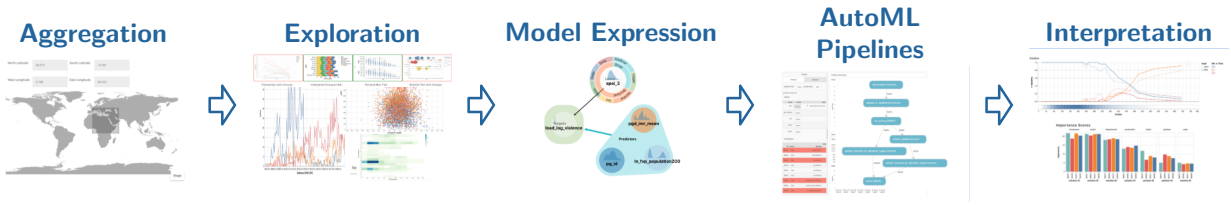


Figure 1: *Representative workflow across TwoRavens components*

2 INTRODUCTION

The DARPA Data Driven Discovery of Models (D^3M) program automates the methods in data science to create empirical models of real, complex processes (Shen, 2016). D^3M 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 is a D^3M system.

TwoRavens, a D^3M system, is a Web-based platform for automated machine learning and statistical analysis (Honaker and D'Orazio, 2014; D'Orazio *et al.*, 2018). The goal is to allow the domain expert, in concert with our system, to complete a high quality, predictive and interpretable model without a statistical expert. To do so, the system facilitates intuitive machine learning and model interpretation, model discovery, and data exploration. 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 (Gil *et al.*, 2019).

Philosophy

In the very best research settings, where there are collaborations between domain experts and data scientists or statistical experts, exploration into the data is a joint venture where statisticians drive the computational machinery of analysis, but are directed to the interesting features by the knowledge of the domain expert. Our belief is that you can automate much of what the statistician brings, but the domain expert remains central to the task. Thus, our goal is to augment the domain expert, leading to the construction of high quality, impactful and interpretable models.

Thus, the fundamental goal of the TwoRavens interface is to extract from an expert all of the substantive and domain knowledge they have about a problem, in order to better inform the construction of a statistical model.

Often the domain expert does not even realize what knowledge is most valuable, or what its implications are for tailoring the statistical choices, but a good statistical collaborator can untangle and extract that information through discussion and patient collaboration.

Features

TwoRavens revolves around *modes* to step through a workflow. Different modes signify different steps or goals a user might be attempting in the current part of their analysis. From these modes, the features available to the user are swapped or modified to better work for that intended task.

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.

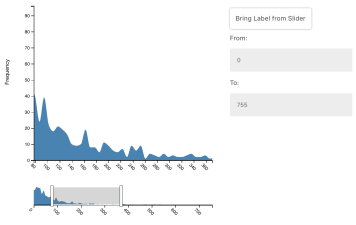


Figure 2: *Data aggregation in dataset mode*

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

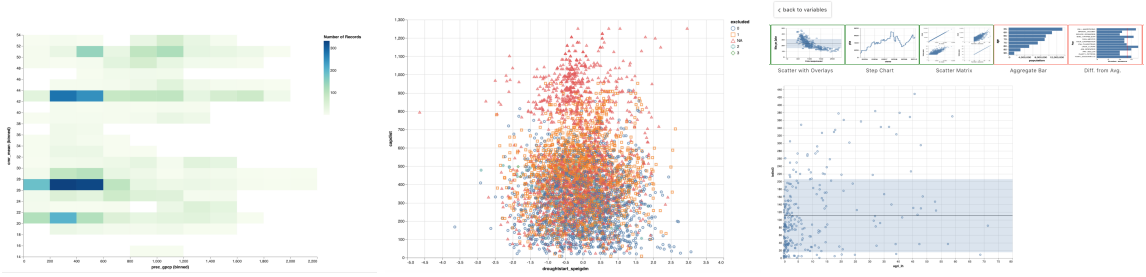


Figure 3: *Representative explore mode visualizations*

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.

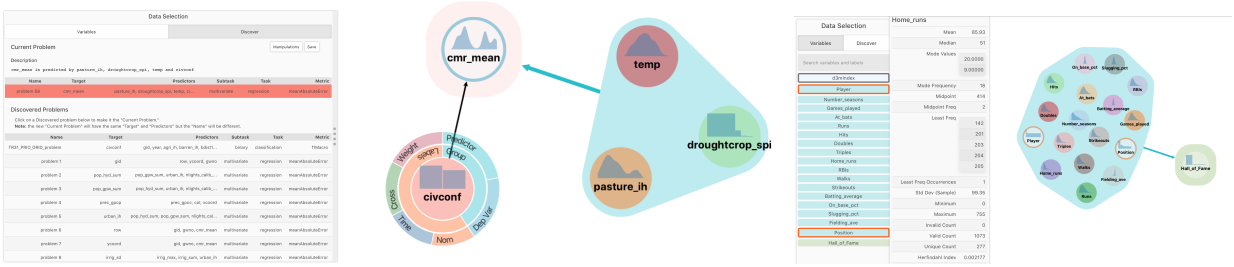


Figure 4: *Representative model mode features*

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.

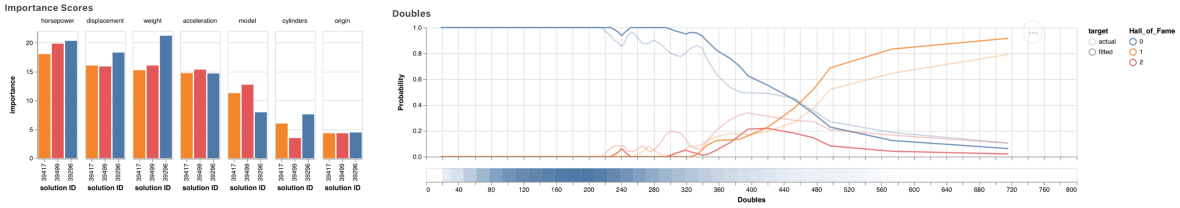


Figure 5: *Representative model mode features*

3 METHODS, ASSUMPTIONS, AND PROCEDURES

We describe here the methods available through the TwoRavens interface, and the procedures a user follows, to explore and understand their data, specify the domain information they know, and set up a search for a machine learning model matched to their needs.

3.1 User Login and Account

Users are given their 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.

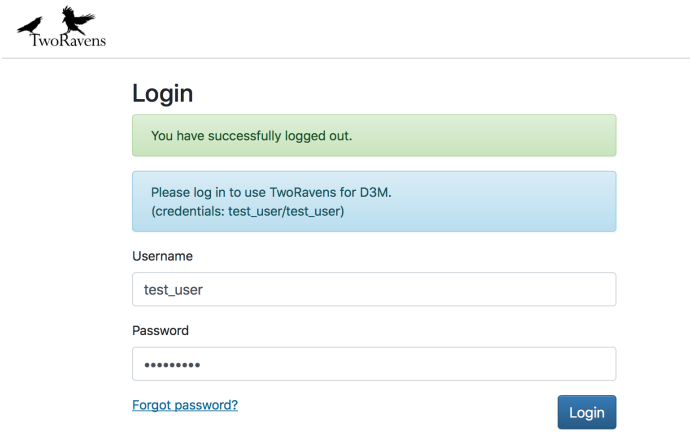


Figure 6: *User login window*

3.2 Help

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

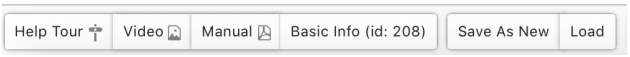


Figure 7: *Available help buttons*

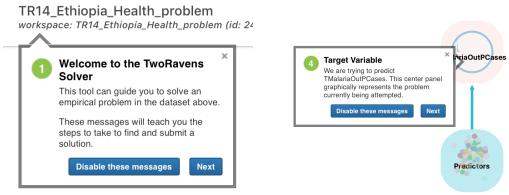


Figure 8: *Signposted help tour*

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.

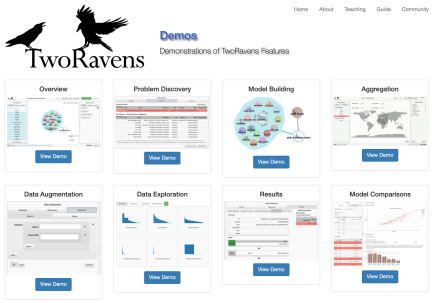


Figure 9: *Page with available help video thumbnails*

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.

3.3 Header and Footers

3.3.1 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).

Modes: A core feature of TwoRavens is the concept of a Mode. Users switch between Modes by clicking one of the four Mode buttons in the Header. More information on Modes is in Section 3.4.

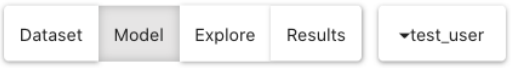


Figure 10: *Mode buttons*

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.

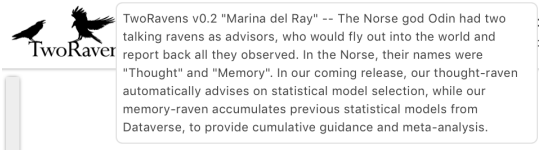


Figure 11: *TwoRavens version information*

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. Clicking the name will switch to Dataset, where more information is provided. The workspace and problem id are also listed here.

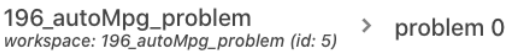


Figure 12: *Dataset and workspace information*

3.3.2 Footer

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

Figure 13: *Footer help buttons*

On the left of the footer are buttons for accessing the help materials (the tour, the videos and the manual) already described in section 3.2. 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. The external link button, on the right, opens a new tab to view the data.

Figure 14: *Dataset download and peak functions*

3.4 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: The four mode buttons in the top right of the header allow switching between modes.

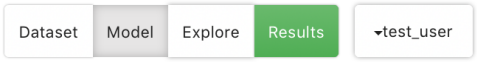


Figure 15: *Mode selection buttons*

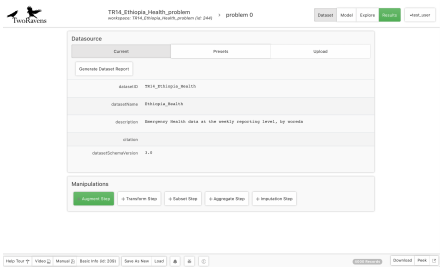


Figure 16: *Dataset mode*

Dataset Mode gives different ways for a user to bring data into the system. Here a user can upload new datasets, pull in data from the Web via URL or OpenML ID, switch between datasets already available, and access the Datamarts to augment current datasets with additional data.

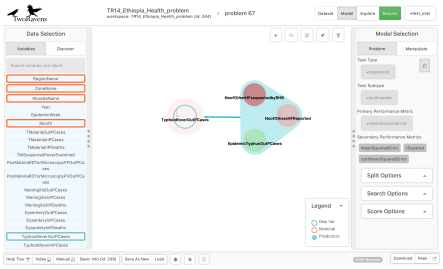


Figure 17: *Model mode*

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.

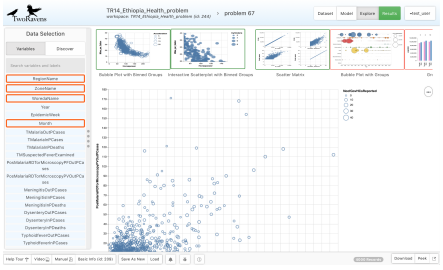


Figure 18: *Explore mode*

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

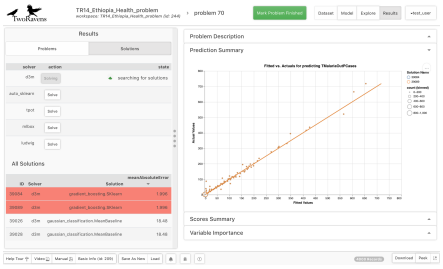


Figure 19: *Results mode*

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.

3.5 Dataset Mode

1. **Dataset Report and Annotations** provides an overview of the data and options to input descriptions and tag features in the data. For example, add variable descriptions or tag a feature as *nominal*.
2. **Datasource** allows the user to load new data. Details below.
3. **Manipulations** includes features to transform the data and then use that transformed data for the remainder of the session. See Section 3.6 for more information.

Datasource

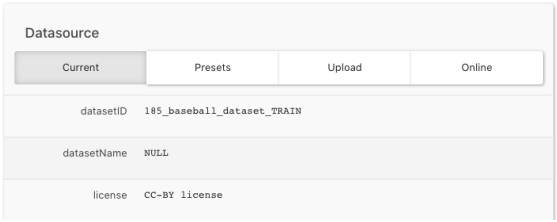


Figure 20: *Dataset summary tab*

Current Tab gives a quick summary of high-level 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.

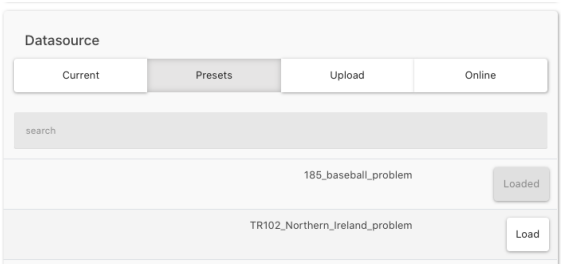


Figure 21: *Presets tab*

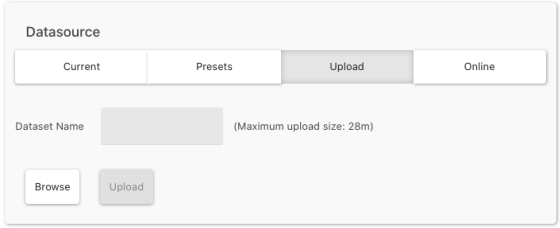


Figure 22: *Upload tab*

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 compute an automated analysis of the data. When a new dataset is uploaded, it will also be loaded as the current dataset in TwoRavens.

Online Tab provides an input box for either a URL or an OpenML ID. These are used to pull data into TwoRavens directly from online sources.

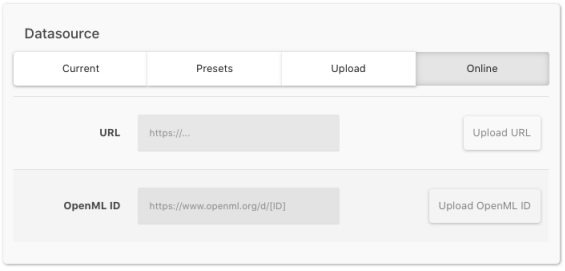


Figure 23: *Online tab*

3.6 Augment with Datamart

Datamart enables the user to find new, relevant data and to merge it into the existing dataset.

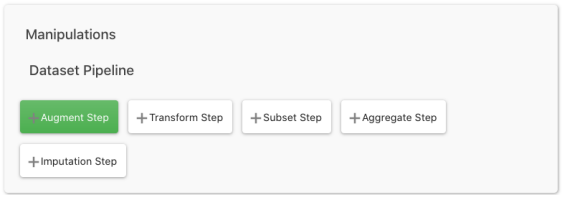


Figure 24: *Dataset manipulations control*

Under Manipulations, select "Augment Step" highlighted in green.

Enter a keyword that you'd like to use to begin your search for new data. Then, using the NYU Datamart, select Search. This will open a new tab, and from here you'll use the NYU Datamart system to identify and join appropriate data.

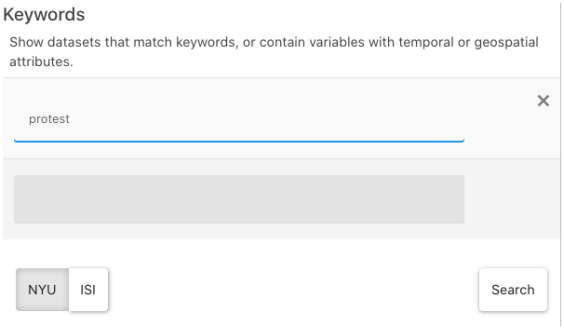


Figure 25: *Datamart keyword search*

Documentation for the NYU Datamart system, *Auctus*, can be found here: https://docs.auctus.vida-nyu.org/. Please refer to their documentation for the details of their system.

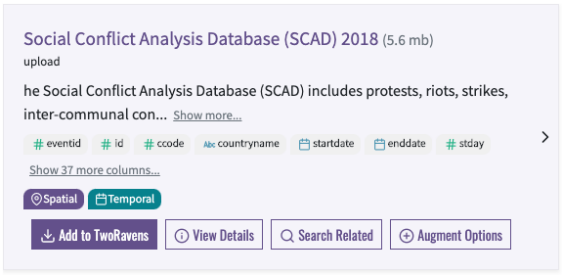


Figure 26: *NYU Auctus Datamart system*

To add an entire dataset using the columns that Auctus identifies as relevant, click the "Add to TwoRavens" button.

Auctus also provides users with many options to control the data merge, which are set by selecting the "Augment Options" button.

When the join is complete, switch back to the TwoRavens tab in your browser. The TwoRavens system is now analyzing a new dataset, which consists of the original data plus the augmented data.

3.7 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.

3.7.1 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.

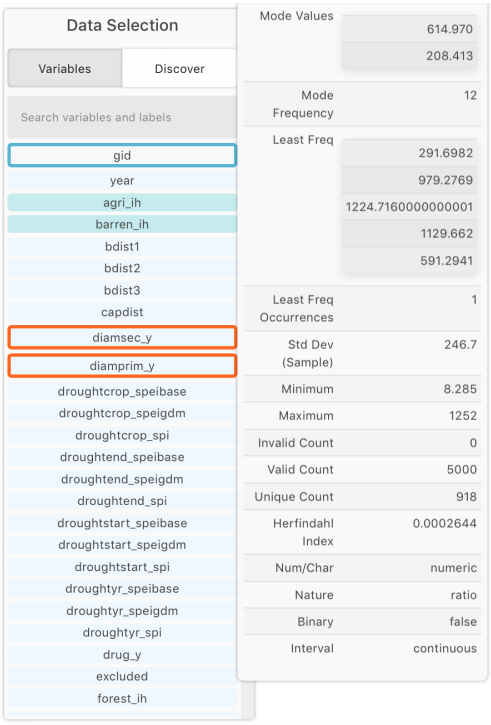


Figure 27: *Variable list*

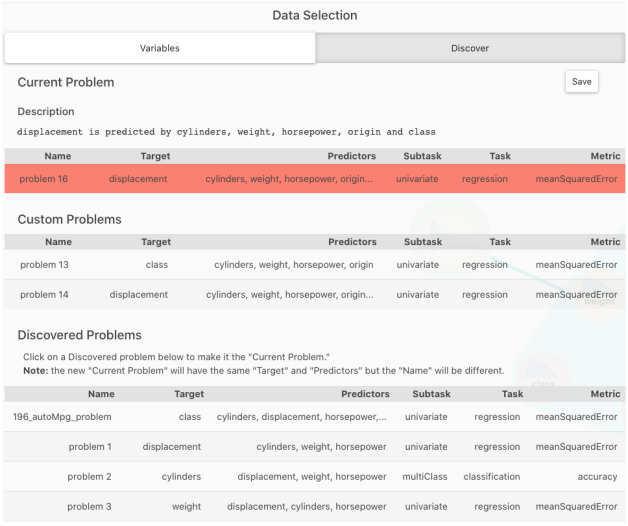


Figure 28: *Problem discovery*

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*.

Add New Problems:

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 create a *custom problem*.

First make sure that TwoRavens is in "Model" mode by selecting the "Model" button in the header, and select the "Variables" 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 "Target" arc that surrounds it.

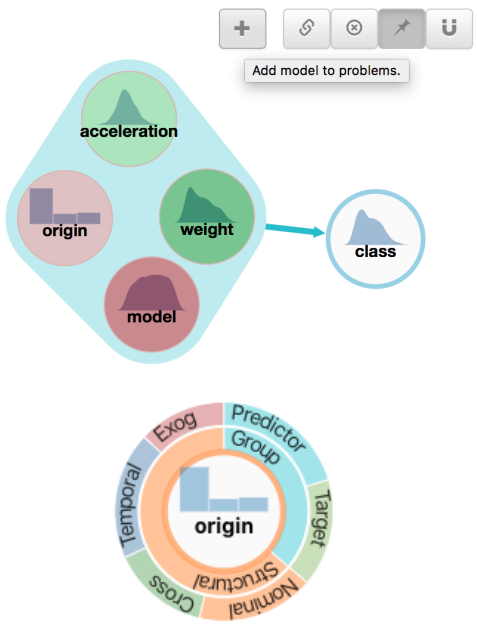


Figure 29: *Problem control surface*

When you have constructed a possible relationship you are interested in adding to the discovered problems table, click the "Discover" tab in the left panel controls and then select "Save." These will appear as "Custom Problems" in the Discovered Problem tab.

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.

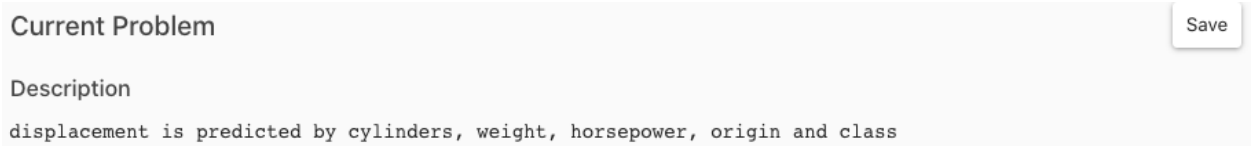


Figure 30: *Problem description text*

For more information on using the center panel to describe possible relationships between variables, see section 3.7.3 on model building.

3.7.2 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*.

3.7.3 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 target, 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.)

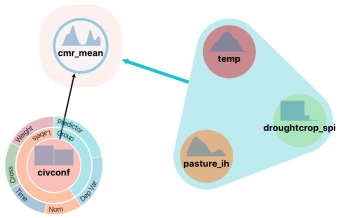


Figure 31: *Control panel directed graph*

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 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.

3.7.4 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 as the current problem in the discovered problem tab.

Figure 32: *Add button*

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.

Figure 33: *Pin button*

Link Button

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

Figure 34: *Link button*

Unlink Button

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

Figure 35: *Unlink button*

Wipe Button

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

Figure 36: *Wipe button*

Model Legend

When tags have been applied to the pebbles of variables, to describe their attributes, they are given either a colored halo or a grouping fill color as a visual identifier of the attributes of that variable for the problem. These colorings appear in the legend box, together with a description of their meaning, to aid in interpretation of the relationship that has been described for the problem. In the example to the right, a *dependent variable* (Target) has been selected, and the legend shows this has a green fill, while a categorical variable has also been tagged appropriately as *nominal* and the legend identifies this with an orange halo. The legend can be minimized by clicking the down arrow in the legend title bar.

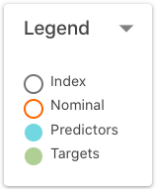


Figure 37: *Control panel pebble legend*

3.7.5 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.

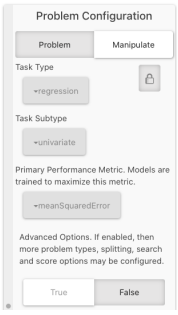


Figure 38: *Problem configuration panel*

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.

By selecting True under Advanced Options, the user has access to additional configurations to control data splits and the automated machine learning engine.

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

- *transform* the data by constructing a new feature as a user constructed function of other features,
- *subset* the data by selecting a set smaller set of the current observations that the model should be run on.
- *impute* missing data values

All steps to manipulate the dataset 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.)

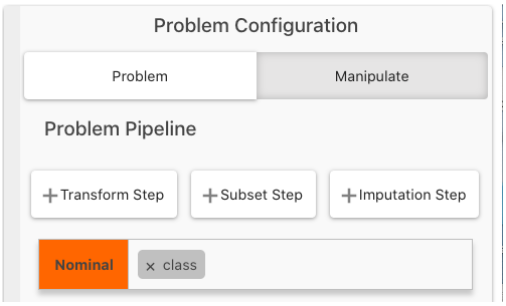


Figure 39: *Data manipulation panel*

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.

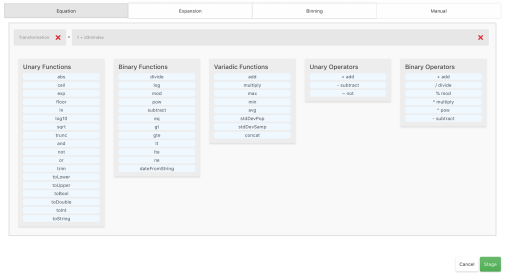


Figure 40: *Feature transformation formulation*

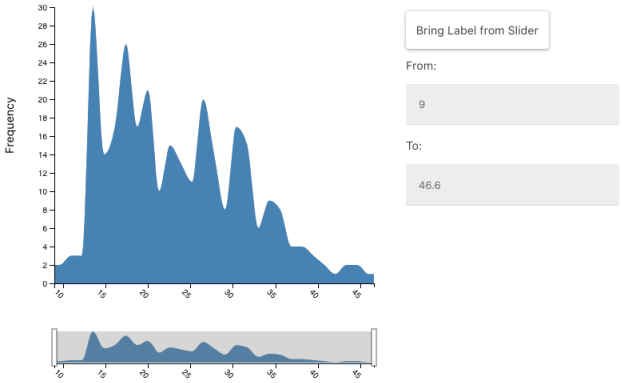


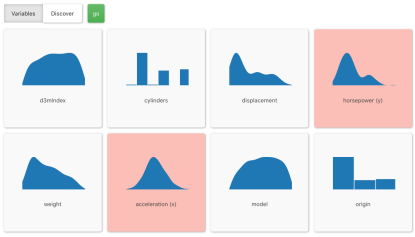
Figure 41: *Feature subset window*

The "Subset Step" allows a user to select observations in the data that meet specified criteria. The appears of visualization in the Subset Step will vary based on the type of data. Here, a continuous variable is shown, and the user may brush a region of the plot, or set values using the input boxes. When finished, click the green "Stage" button in the bottom right part of the screen.

3.8 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.

To enter Explore Mode, select *Explore* button in the header.



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

Figure 42: *Explore mode feature tiles*

To explore the data, click a tile to place it on an axis of the forthcoming plot. First variable selected is placed on the x-axis, the second on the y-axis, and the third on the z-axis. The z-axis is useful to plot data with a sorting variable.

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.

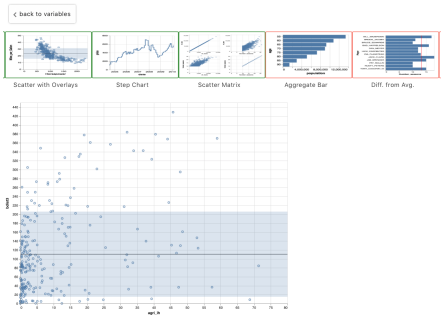


Figure 43: *Explore mode graph control*

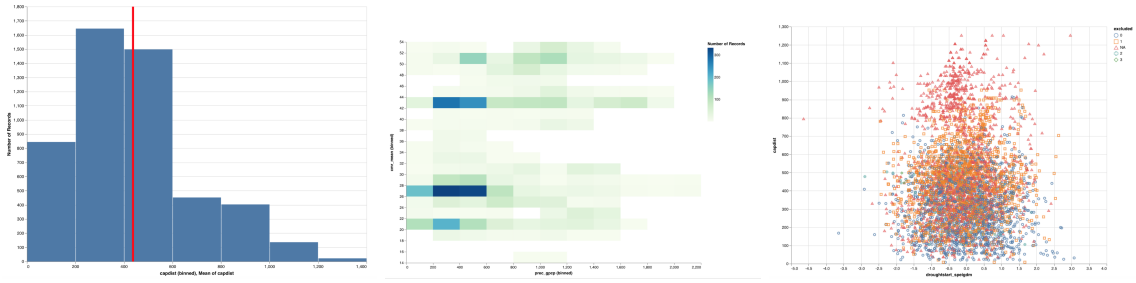
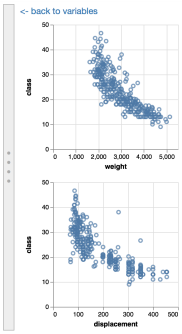


Figure 44: *Explore model representative visualizations*



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.

Figure 45: *Two-way relationship visualizations*

4 RESULTS AND DISCUSSION

The end product of the methods and procedures discussed in the previous section are automated pipelines presenting machine learning solutions. We describe here how these results are accumulated, compared and intuitively interpreted to the user, to provide machine learning predictive pipelines.

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.

To begin searching for solutions, the user must click “Solve” for one of the Solvers in the left panel.

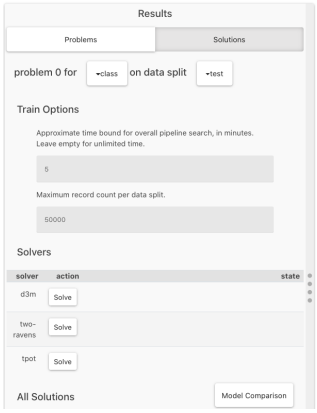


Figure 46: *Solver controls*

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 “two-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 summary representation of how well the solution fits. For regression problems, this is a scatter plot of predicted versus the actual values, as seen here.

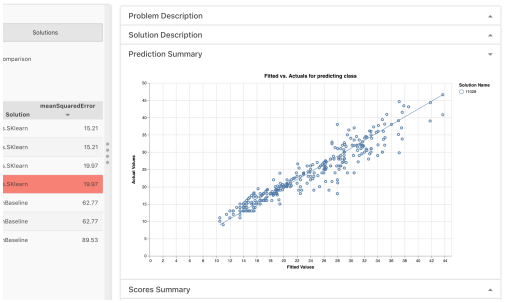


Figure 47: *Regression prediction summary*

Figure 48: *Multiple model comparision*

You can compare the predictions of multiple pipelines on the graph by clicking the box marked ”Model Comparison” and selecting the models you want to compare.

For classification problems, the fit is summarized using a confusion matrix. All observations that have the correct prediction will show up on the diagonal of the table, so models with most observations on the diagonal are making good predictions. Mouseover on any cell of the table will provide information about predictions for that cell.

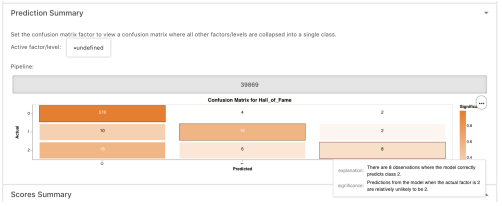


Figure 49: *Classification prediction summary*

Figure 50: *Results accordion*

Exploration of solutions is navigated using the Results Accordion. The summary plots discussed above are in the Prediction Summary tab.

Variable Importance describes which features in the model are the most important in constructing the predictions for the solution. A bar graph 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.

If Model Comparison is checked, then multiple models can be selected, and the 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.

Figure 51: *Feature importance diagnostics*

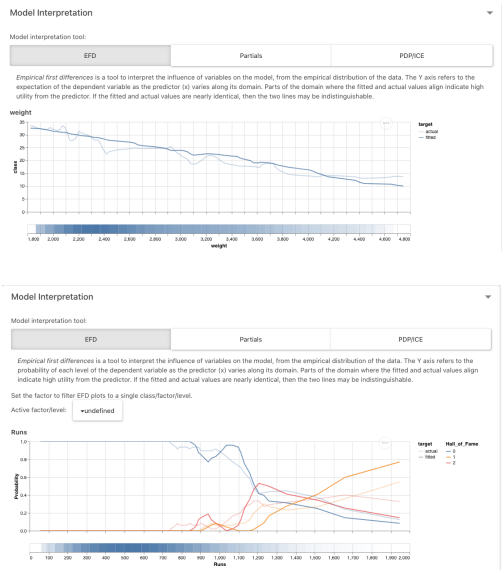


Figure 52: *Feature effect interpretation*

Model Interpretation shows the model's predicted value of the target variable (y-axis) across the range of the feature (x-axis). These graphs are arranged in order of feature importance. Empirical first differences show 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.

When the modeled problem is a classification problem, Empirical First Differences show multiple sets of lines, color coded, each of which represent the average probability of each class being the predicted class.

In addition to EFD, users may use the Partials and PDP/ICE tabs for model interpretation.

Opening the "Solution Description" allows users 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. You may also download input data and predictions.

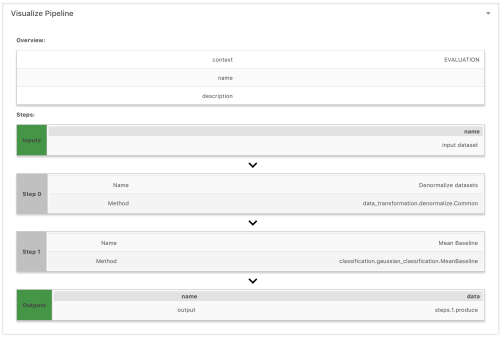


Figure 53: *Pipeline solution component visualization*

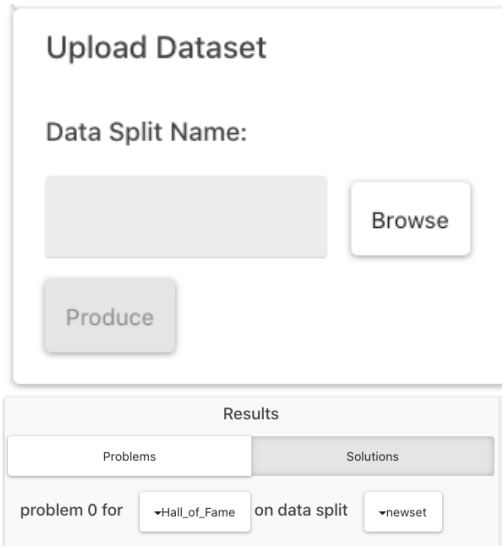


Figure 54: *Upload prediction dataset control*

Upload Dataset provides the option to generate predictions on a new dataset. Select Browse, upload your new data, and click Produce. Once complete, you can click "Select." To go back to a different dataset, choose a different option for "data split" in the Results panel on the left.

5 CONCLUSIONS

TwoRavens is a platform for machine learning that allows a domain expert, in concert with our system, to complete a high quality, predictive and interpretable model without any statistical or machine learning expertise. To do so, the system allows a user to easily upload and profile new data, append data from the related D^3M Datamart system. This data can then we explored through automated low dimensional graphs automatically constructed, and lower dimensional relationships automatically discovered by the system. These can form the foundation for user defined models through our directed graph control panel to intuitively represent the relationships of interest in the data. From this, the system uses our own and other D^3M autoML engines to search for end-to-end pipelines to accomplish the desired predictive task. Competing solutions are intuitively represented to back to the user, along with interpretations of the relative feature importance of variables in these models, and their predictive accuracy.

Most analysts are heavily invested in the collection of their data, understand their variables well, and possess expert substantive knowledge about the plausibility of various relationships. What they may not know is the set of plausible machine learning models appropriate to their objectives, or how to evaluate their performance, or interpret their results. TwoRavens provides the automation to deliver these insights and quickly produce highly performant machine learning pipelines.

6 REFERENCES

References

D'Orazio V, Deng M, Shoemate M (2018). "TwoRavens for Event Data." In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 394–401. IEEE.

Gil Y, Honaker J, Gupta S, Ma Y, D'Orazio V, Garijo D, Gadewar S, Yang Q, Jahanshad N (2019). "Towards Human-guided Machine Learning." In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 614–624. ACM, New York, NY, USA.

Honaker J, D'Orazio V (2014). "Statistical Modeling by Gesture: A graphical, browser-based statistical interface for data repositories." In *Extended Proceedings of ACM Hypertext 2014*. ACM.

Shen W (2016). "Data-driven discovery of models (d3m)." *Defense Advanced Research Projects Agency, Arlington, VA*.

A PUBLICATIONS AND PRESENTATIONS

A.1 Publications

1. Bessa, Aline, Sonia Castelo, Rmi Rampin, Acio Santos, Mike Shoemate, Vito D'Orazio, and Juliana Freire. "An Ecosystem of Applications for Modeling Political Violence." In Proceedings of the 2021 International Conference on Management of Data, pp. 2384-2388. 2021.

2. D'Orazio, Vito. "Conflict Forecasting and Prediction." Oxford Research Encyclopedia of International Studies. Published July 2020.

3. D'Orazio, Vito, Marcus Deng, and Michael Shoemate. "TwoRavens for Event Data." In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 394-401. IEEE, 2018.

4. D'Orazio, Vito, James Honaker, Raman Prasad, and Michael Shoemate. "Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning" IEEE Big Data: 3rd International Workshop on Big Data Analytics for Cyber Intelligence and Defense (BDA4CID), Los Angeles, CA, pp 4714-4723. Published December 2019. DOI: 10.1109/BigData47090.2019.9005963.

5. Gil, Yolanda, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. "Towards human-guided machine learning." In Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 614-624. ACM, 2019.

A.2 Presentations

1. Meeting Name: Privacy Tools for Data Sharing Workshop: Lessons Learned and Directions Forward
   Purpose: Workshop on tools, software and approaches for sharing privacy sensitive data and metadata.
   Start Date December 11, 2017
   End Date December 12, 2017
   Location: Cambridge, MA. Harvard John A. Paulson School of Engineering and Applied Sciences
   Attendees: Vito D'Orazio, James Honaker
   List titles of presentations that were made: TwoRavens and Software Development on Academic Projects

2. Meeting Name: Invited Guest Lecture
   Purpose: Describe trends and technologies in browser based data visualization.
   Start Date December 14, 2017
   End Date December 14, 2017
   Location: Boston, MA. Harvard T. H. Chan School of Public Health
   Attendees: James Honaker
   List titles of presentations that were made: Interactive Data Manipulation with Data Driven Documents (d3) and Javascript

3. Meeting Name: IEEE International Conference on Information Reuse and Integration (IRI) 2018
   Purpose: Present invited paper on TwoRavens work on Time-series event data.
   Start Date July 7, 2018
   End Date July 9, 2018
   Location: Salt Lake City, UT.
   Attendees: Vito D'Orazio
   List titles of presentations that were made: TwoRavens for Event Data

4. Meeting Name: Meeting of Southern Political Science Association
   Purpose: Present paper on TwoRavens work human guided machine learning in conflict data.
   Start Date January 17, 2019

End Date January 19, 2019
Location: Austin, TX
Attendees: Vito DOrazio List titles of presentations that were made: 1. Modeling and Forecasting Armed Conflict: AutoML with Domain Expertise

5. Meeting Name: Meeting of the International Studies Association Purpose: Present TwoRavens software advances and research contributions to applied researchers working on quantitative modeling of conflict and political instability.
   Start Date: March 27, 2019
   End Date: March 30, 2019
   Location: Toronto, Canada
   Attendees: Vito D'Orazio
   List titles of presentations that were made: Modeling and Forecasting Armed Conflict: AutoML with Domain Expertise

6. Meeting Name: University of North Texas, Invited Talk
   Purpose: Present TwoRavens software advances and research contributions to applied researchers working on quantitative modeling of conflict and political instability.
   Start Date: April 24, 2019
   End Date: April 24, 2019
   Location: Denton, TX
   Attendees: Vito D'Orazio
   List titles of presentations that were made: Modeling and Forecasting Armed Conflict: AutoML with Domain Expertise

7. Meeting Name: Harvard Business School Engineering Sciences Program, Invited Talk
   Purpose: Overview TwoRavens system, and novel technological advances to members of joint engineering and business program as case study for potential technology transition and use case exploration.
   Start Date: October 15, 2019
   End Date: October 15, 2019
   Location: Cambridge MA
   Attendees: James Honaker
   List titles of presentations that were made: TwoRavens Automated Machine Learning and Discovery

8. Meeting Name: IEEE Big Data: 3rd International Workshop on Big Data Analytics for Cyber Intelligence and Defense (BDA4CID)
   Purpose: Present research paper overviewing TwoRavens system, and automated machine learning solutions to conflict forecasting, with testing of D3M TA2 systems and commercial packages.
   Start Date: December 9, 2019
   End Date: December 12, 2019
   Location: Los Angeles, CA
   Attendees: Michael Shoemate
   List titles of presentations that were made: Modeling and Forecasting Armed Conflict: AutoML with Human-Guided Machine Learning

9. Meeting Name: Texas Triangle International Relations Conference
   Purpose: Present paper on TwoRavens work in conflict data, and methodological research agenda.
   Start Date: January 25, 2020
   End Date: January 25, 2020
   Location: Houston, TX
   Attendees: Vito D'Orazio

List titles of presentations that were made: Conflict Forecasting and the Role of Model Selection

10. Meeting Name: Bureau of Conflict and Stabilization Operations, Department of State
    Purpose: Present capabilities and demonstration of TwoRavens system for conflict data modeling to potential transition partners.
    Start Date: March 2, 2020
    End Date: March 2, 2020
    Location: Department of State, Washington D.C.
    Attendees: James Honaker, Vito D'Orazio
    List titles of presentations that were made: TwoRavens Interface for Human-Guided Machine Learning

11. Meeting Name: Dataverse 2020 Community Meetings
    Purpose: Present overview of TwoRavens capabilities to data archivists and librarians, and show integration of automated machine learning tools with data repositories.
    Start Date: June 17, 2020
    End Date: June 19, 2020
    Location: Harvard University, Cambridge MA. (Web conference because of Covid)
    Attendees: Vito D'Orazio, James Honaker, Raman Prasad, Michael Shoemate
    List titles of presentations that were made: The TwoRavens Project: Exploration and Analysis of Conflict in Nigeria

12. Meeting Name: Microsoft AI for Good
    Purpose: Present overview and live demonstration of TwoRavens capabilities with emphasis on potential for using as interface for Microsofts own AutoML engine.
    Start Date: August 17, 2020
    End Date: August 17, 2020
    Location: New York (Web conference because of Covid)
    Attendees: Vito D'Orazio, James Honaker
    List titles of presentations that were made: TwoRavens Interface for Human-Guided Machine Learning

13. Meeting Name: Central Intelligence Agency
    Purpose: Present demo and overview of capabilities of TwoRavens system to potential transition partners.
    Start Date: October 28, 2020
    End Date: October 28, 2020
    Location: Arlington, VA (Web conference because of Covid)
    Attendees: Vito D'Orazio, James Honaker
    List titles of presentations that were made: TwoRavens Interface for Human-Guided Machine Learning

14. Meeting Name: USAID Conflict Prevention and Stabilization Bureau
    Purpose: Present demo and overview of capabilities of TwoRavens system to potential transition partners.
    Start Date: November 12, 2020
    End Date: November 12, 2020
    Location: Washington D.C. (Web conference because of Covid)
    Attendees: Vito D'Orazio, James Honaker
    List titles of presentations that were made: TwoRavens Interface for Human-Guided Machine Learning

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

AI	Artificial intelligence
AutoML	Automated machine learning
CSV	Comma-separated Values
D^3M	Data driven discovery of models
DARPA	Defense Advanced Research Projects Agency
EFD	Empirical first differences
ICE	Individual conditional expectation
NYU	New York University
PDP	Partial dependence plot