CeADAR – Centre for Applied Data Analytics Research

Enterprise Ireland Data Analytics Technology Centre

Regression Model Visualisation – Technical Specification

Document Type:	Technical Specification
Project Title:	CeADAR
DDN Theme:	1 – Intelligent Analytic Interfaces
Theme Leader:	Sarah Jane Delany (DIT)
DDN Sub-Theme:	Ease of Interaction
Authors:	Caroline Maillet, Susan McKeever, John Kelleher, Ingo R. Keck
Document Version:	0.1
Date of Delivery to ISG:	1st May 2015
Number of pages:	10

ABSTRACT

This document provides a high-level technical specification of ModelVis, a regression model visualisation tool designed to help data analysts in communicating their regression results to a non-expert audience. This web tool is being developed as part of the CeADAR Intelligent Analytic Interfaces: Ease of Interaction theme.

CeADAR is a research partnerhip comprising University College Dublin, University College Cork, and Dublin Institute of Technology.

http://www.ceadar.ie

Contents

1 Description of Industry Needs

Data analysis is an increasingly important task for industry; one they must typically perform before making a business decision. Linear and logistic regression are commonly used as predictive data analytic models.

The key characteristics of a regression model are:

- the regression *weights*, also known as regression coefficients, regression estimates or parameter estimates,
- the *sign* of each weight,
- a *statistical significance* associated with each weight.

Given this, we want to develop a regression model visualisation system, called ModelVis, that provides an intuitive understanding of how these three results for each feature (i.e. the predictor), contribute to the overall outcome of the regression model (i.e. the prediction decision or the target). ModelVis will therefore be an easy-to-use and easy-to-understand web-based interface to visualise regression results.

This document provides a high-level technical specification of the ModelVis system that seeks to solve the problem of communicating regression results to a non-technical audience.

2 System(s) Involved

ModelVis will provide a web-based interface to allow data analysts to easily visualise their regression model results. For this reason, ModelVis' main systems will involve:

- a web browser, so as to be easy to use without any required software deployment on the end user's computer;
- a remote web application, to compute the appropriate regression visualisation.

The CeADAR web application will include access to a statistical tool such as R[1] to allow users to visualise new regression models developed on the fly.

3 Approach

This section provides information on the assumptions underlying ModelsVis, as well as its methods of use.

3.1 ModelVis assumptions

ModelVis assumes that the input data for regression has undergone appropriate collection and cleaning. This implies that the cleaned data has been at least:

- checked for interpretation (e.g., to ensure that "0" represents zero and not missing values) and to confirm that any special values such as 99 or 999 are documented;
- transformed to resolve missing values.

3.2 ModelVis' Usage Models

ModelVis users will visualise regression results. Users will have the option to either upload their own regression model results or to use ModelVis' in-built regression modelling.

[1] The R Project for Statistical Computing: `http://www.r-project.org/`

3.2.1 Development of regression models

ModelVis users will have a choice between:

- ModelVis regression modelling: users will upload their preprocessed cleaned data as a CSV file[2] to develop their regression model via ModelVis' embedded statistical tool:
  - users will have to select the subset of features on which the regression model will be built;
- or running their own regression models outside of ModelVis, and uploading the results of these models to ModelVis by either:
  - keying the model results in through a web form,
  - or uploading the model results as a CSV file.

The formats of both CSV files are detailed in Section 4.

3.2.2 Visualisation of regression models

For each regression model, ModelVis will visualise:

- regression *weights* to show how much a feature contributes to the overall regression output,
- *signs* of each weight to show if a feature has a negative or a positive impact on the overall regression output,
- a *statistical significance* associated with each weight to show how much a feature contribution can be trusted.

ModelVis will display these characteristics via a force-directed graph, that is easy to understand by a non-technical audience.

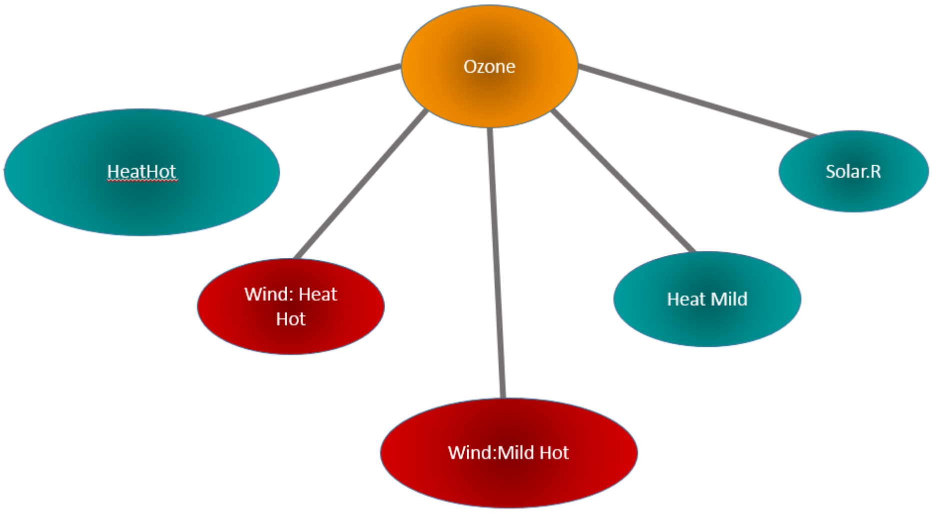


Figure 1: Example of a force directed graph representing the regression model *Reg2*. Details are given in Figures 7 and 8.

As an example, Figure 1 show a force-directed graph visualising a regression model that seeks to predict the Ozone levels in the air according to the solar radiation, the wind and the temperature levels. This regression model was trained on the air quality data obtained from the New York State Department of Conservation and the National Weather Service

[2] CSV (Comma-Separated Values) files are widely used by data analysts.

(meteorological data) [1]. This regression model, called *Reg2*, is detailed in Figure 7 and in Figure 8 at the end of this document.

ModelVis will then let users interact with the regression visualisation. For example, users will be able to drill down through the regression features to display extra information such as text or plots. Plots will be uploaded as .PNG files[3], seeing as common tools used by data analysts, such as like R, SAS[4], Excel[5], allow easy export of plots in a PNG format.

Finally, ModelVis will allow users to compare two different models developed on the same underlying dataset. ModelVis will then highlight most important differences between these two models.

4 Appendix: requirements for ModelVis inputs

This appendix defines the expected format for ModelVis inputs. Apart from plots uploaded as PNG files, all ModelVis inputs will be USA/UK CSV files, where the decimal separator is a dot and the value separator is a comma.

4.1 Cleaned input data format

If users wants to use ModelVis'embedded statistical tool, they will have to upload their cleaned input data as a CSV file, as shown by Figure 2, for example.

```
"Ozone","Solar.R","Wind","Temp","Month","Day","Heat"
41,190,7.4,67,5,1,"Cool"
12,149,12.6,74,5,3,"Mild"
18,313,11.5,62,5,4,"Cool"
36,118,8,72,5,2,"Mild"
32,92,15.5,84,9,6,"Hot"
71,291,13.8,90,6,9,"Hot"
98,11.5,80,6,28,"Mild"
```

Figure 2: Example of CSV file containing cleaned input data.

ModelVis will assume that the first line of such an uploaded CSV file contains:

- the target name (in Figure 2 *Ozone* is the target name),
- followed by the names of the features (in Figure 2 *Solar.R*, *Wind*, *Temp*, *Month*, *Day*, *Heat* are the feature names).

The following lines of the file contain target and features values for different observations, where each line is associated with a single observation. In Figure 2 [41, 190, 7.4, 67, 5, 1, "Cool"] represents one observation, where 41 represents the target value and the other values are the values for the other features respectively for that observation.

[3] PNG - Portable Network Graphics: `http://www.w3.org/TR/PNG/`

[4] SAS - Statistical Analysis System: `http://www.sas.com`

[5] Microsoft Excel: `http://products.office.com/en-ie/excel`

4.2 Regression model results format

If ModelVis users want to visualise regression models they developed with their own statistical tool, they will have to upload their regression model results as a CSV file.

Figure 3 represents a template of the expected format for such an uploaded CSV file, displayed in a spreadsheet. Figure 4 represents the exact same template displayed in a text editor. Figure 5 (displayed in a spreadsheet) and Figure 6 (displayed in a text editor) show a template-example. Finally, Figure 7 (displayed in a spreadsheet) and Figure 8 (displayed in a text editor) detail a concrete example based on the air quality data obtained from the New York State Department of Conservation and the National Weather Service (meteorological data) [1] to predict the Ozone levels according to solar radiation, wind and temperature levels.

As shown by Figures 3, 4, 5, 6, 7 and 8, ModelVis will assume that the first line of such an uploaded CSV file contains the headers describing the expected fields. Each following line will contain field values. The following describes the expected values for each field:

- **Model Name**: text used to differentiate regression models, such as *Regression 1* or *Model 2*;
- **Type**: "target", "feature" or "interaction" to specify what the current line is referring to;
- **Name**: text used to name either the target, a feature or an interaction term;
- **Weight**: a positive or negative number representing the regression weight associated with a feature or an interaction term (leave blank for the target);
- **Statistical Significance**: a number between 0 and 1 representing the statistical significance, associated with a feature or an interaction term (leave blank for the target);
- **Text Info**: extra text information to display in HTML[6] when users interact with the regression visualisation (leave blank if unnecessary);
- **Plot 1**: the file name of a .PNG file representing a plot to be displayed when users interact with the ModelVis visualisation (leave blank if unnecessary);
- **Plot 2**: the file name of a .PNG file representing a plot to be displayed when users interact with ModelVis visualisation (leave blank if unnecessary);
- ...;
- **Plot** n: the file name of a .PNG file representing a plot to be displayed when users interact with the ModelVis visualisation (leave blank if unnecessary).

Even if different model names are entered, all models should refer to a unique business problem. In Figures 7 and 8, *Reg1* and *Reg2* are two different regression models both trying to predict Ozone levels.

Data analysts might wish to perform some binning preprocessing to create binned features. In that case, ModelVis will handle features and binned features in the same way. For this reason, binned feature should be declared as "feature".

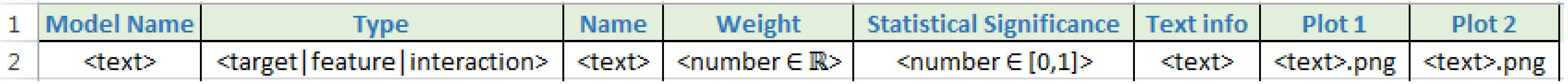
The "interaction" category represents interaction terms, ie. combined features for the purposes of regression

[6] HTML - HyperText Markup Language: `http://www.w3.org/html/`

For example, *Wind:HeatMild* from the regression model *Reg2* in Figures 7 and 8 represent an interaction term between the *Wind* feature and the *HeatMild* binned feature. This binned feature represents the *Mild* category from the *Temperature* feature.

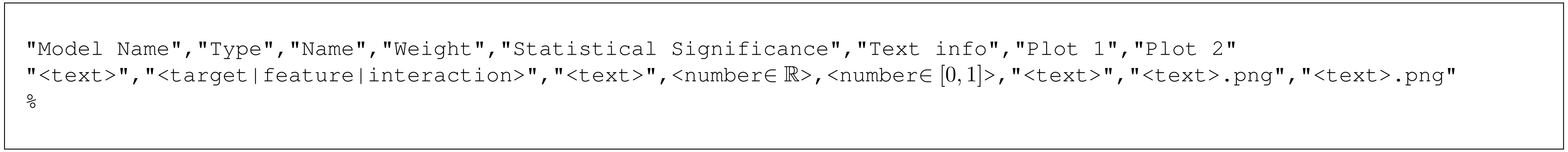
ModelVis users will be able to upload plots with the same name as described in the CSV file. This will allow ModelVis to display uploaded plots when users interact with the regression visualisation to show the correlation between a feature and the target, for example.

For information, ModelVis'statistical significance refers to 2-tailed p-values used in testing the null hypothesis, and is often written as $Pr(> |t|)$.



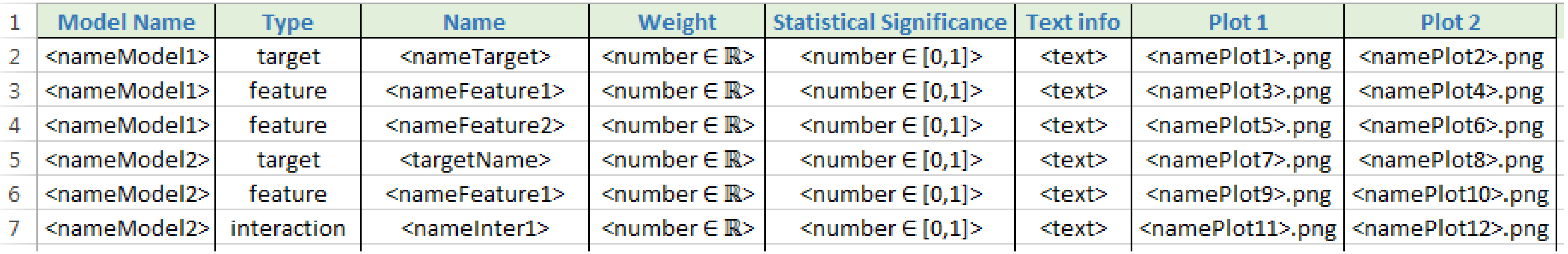
1	Model Name	Type	Name	Weight	Statistical Significance	Text info	Plot 1	Plot 2
2	<text>	<target|feature|interaction>	<text>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<text>.png	<text>.png

Figure 3: Template regression results CSV file displayed in a spreadsheet.



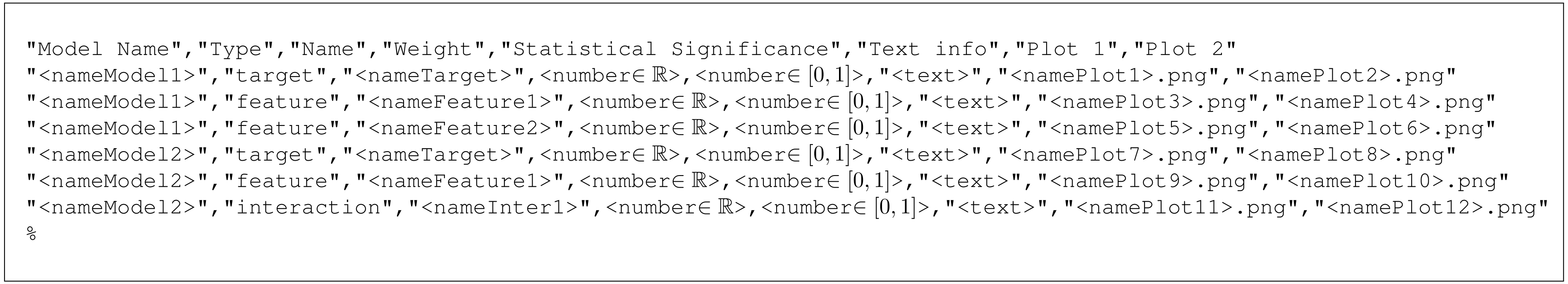
```
"Model Name","Type","Name","Weight","Statistical Significance","Text info","Plot 1","Plot 2"
"<text>","<target|feature|interaction>","<text>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<text>.png","<text>.png"
%
```

Figure 4: Template regression results CSV file displayed in a text editor.



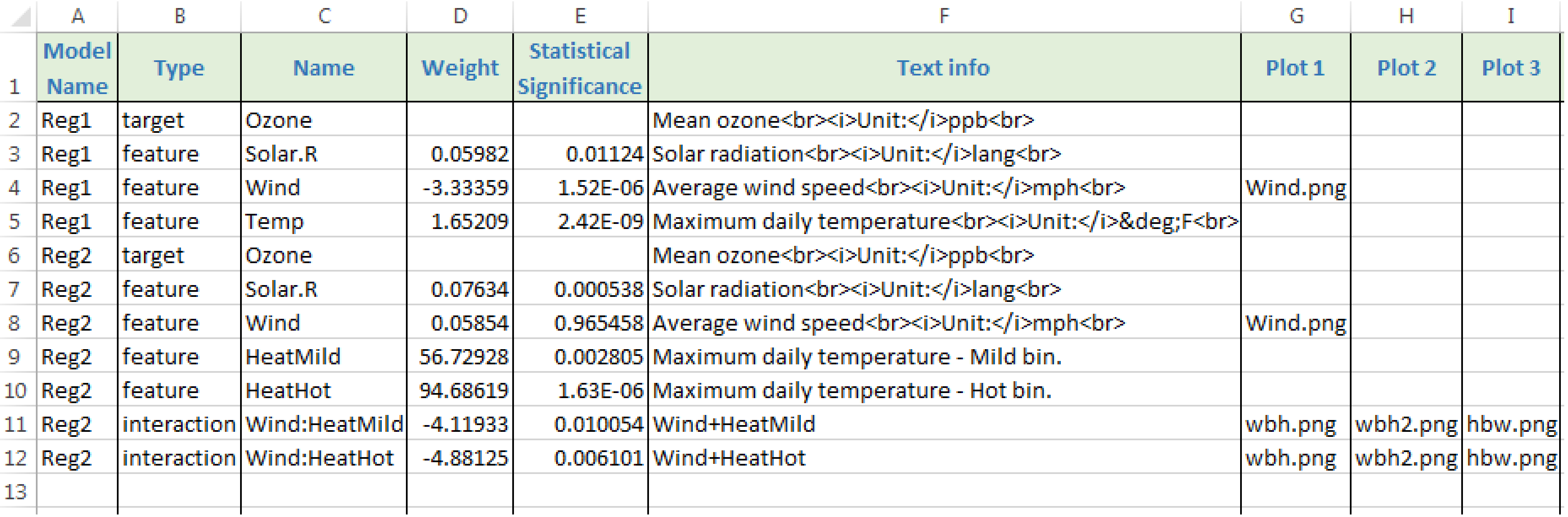
1	Model Name	Type	Name	Weight	Statistical Significance	Text info	Plot 1	Plot 2
2	<nameModel1>	target	<nameTarget>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot1>.png	<namePlot2>.png
3	<nameModel1>	feature	<nameFeature1>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot3>.png	<namePlot4>.png
4	<nameModel1>	feature	<nameFeature2>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot5>.png	<namePlot6>.png
5	<nameModel2>	target	<targetName>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot7>.png	<namePlot8>.png
6	<nameModel2>	feature	<nameFeature1>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot9>.png	<namePlot10>.png
7	<nameModel2>	interaction	<nameInter1>	<number ∈ ℝ>	<number ∈ [0,1]>	<text>	<namePlot11>.png	<namePlot12>.png

Figure 5: Template-example regression results CSV file displayed in a spreadsheet.



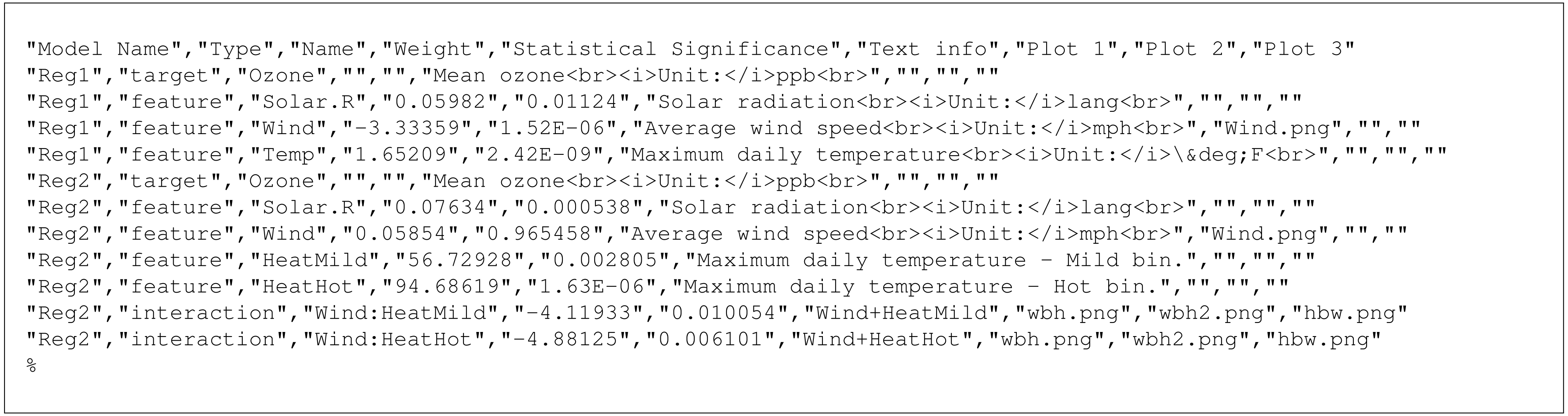
```
"Model Name","Type","Name","Weight","Statistical Significance","Text info","Plot 1","Plot 2"
"<nameModel1>","target","<nameTarget>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot1>.png","<namePlot2>.png"
"<nameModel1>","feature","<nameFeature1>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot3>.png","<namePlot4>.png"
"<nameModel1>","feature","<nameFeature2>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot5>.png","<namePlot6>.png"
"<nameModel2>","target","<nameTarget>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot7>.png","<namePlot8>.png"
"<nameModel2>","feature","<nameFeature1>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot9>.png","<namePlot10>.png"
"<nameModel2>","interaction","<nameInter1>",<number∈ ℝ>,<number∈ [0,1]>,"<text>","<namePlot11>.png","<namePlot12>.png"
%
```

Figure 6: Template-example regression results CSV file displayed in a text editor.



	A	B	C	D	E	F	G	H	I
1	Model Name	Type	Name	Weight	Statistical Significance	Text info	Plot 1	Plot 2	Plot 3
2	Reg1	target	Ozone			Mean ozone<br><i>Unit:</i>ppb<br>			
3	Reg1	feature	Solar.R	0.05982	0.01124	Solar radiation<br><i>Unit:</i>lang<br>			
4	Reg1	feature	Wind	-3.33359	1.52E-06	Average wind speed<br><i>Unit:</i>mph<br>	Wind.png		
5	Reg1	feature	Temp	1.65209	2.42E-09	Maximum daily temperature<br><i>Unit:</i>&deg;F<br>			
6	Reg2	target	Ozone			Mean ozone<br><i>Unit:</i>ppb<br>			
7	Reg2	feature	Solar.R	0.07634	0.000538	Solar radiation<br><i>Unit:</i>lang<br>			
8	Reg2	feature	Wind	0.05854	0.965458	Average wind speed<br><i>Unit:</i>mph<br>	Wind.png		
9	Reg2	feature	HeatMild	56.72928	0.002805	Maximum daily temperature - Mild bin.			
10	Reg2	feature	HeatHot	94.68619	1.63E-06	Maximum daily temperature - Hot bin.			
11	Reg2	interaction	Wind:HeatMild	-4.11933	0.010054	Wind+HeatMild	wbh.png	wbh2.png	hbw.png
12	Reg2	interaction	Wind:HeatHot	-4.88125	0.006101	Wind+HeatHot	wbh.png	wbh2.png	hbw.png
13									

Figure 7: Example regression results CSV file displayed in a spreadsheet.



```
"Model Name","Type","Name","Weight","Statistical Significance","Text info","Plot 1","Plot 2","Plot 3"
"Reg1","target","Ozone","","","Mean ozone<br><i>Unit:</i>ppb<br>","","",""
"Reg1","feature","Solar.R","0.05982","0.01124","Solar radiation<br><i>Unit:</i>lang<br>","","",""
"Reg1","feature","Wind","-3.33359","1.52E-06","Average wind speed<br><i>Unit:</i>mph<br>","Wind.png","",""
"Reg1","feature","Temp","1.65209","2.42E-09","Maximum daily temperature<br><i>Unit:</i>\&deg;F<br>","","",""
"Reg2","target","Ozone","","","Mean ozone<br><i>Unit:</i>ppb<br>","","",""
"Reg2","feature","Solar.R","0.07634","0.000538","Solar radiation<br><i>Unit:</i>lang<br>","","",""
"Reg2","feature","Wind","0.05854","0.965458","Average wind speed<br><i>Unit:</i>mph<br>","Wind.png","",""
"Reg2","feature","HeatMild","56.72928","0.002805","Maximum daily temperature - Mild bin.","","",""
"Reg2","feature","HeatHot","94.68619","1.63E-06","Maximum daily temperature - Hot bin.","","",""
"Reg2","interaction","Wind:HeatMild","-4.11933","0.010054","Wind+HeatMild","wbh.png","wbh2.png","hbw.png"
"Reg2","interaction","Wind:HeatHot","-4.88125","0.006101","Wind+HeatHot","wbh.png","wbh2.png","hbw.png"
%
```

Figure 8: Example regression results CSV file displayed in a text editor.

References

[1] John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, 1983.