Analysis Report

Prepared for: Peter Ronson, Political Analyst, PMD Corporation

Prepared by: Lina Massou, Data Engineer, datamine.it

November 27th, 2008

Report number: 000-0003

datamine.it 14 Meletiou Vasileiou Str 11 745 Athens, Greece **T** +30 6937 122 065 go@datamine.it http://datamine.it

Executive Summary

Objective

The hereby report summarizes the results of the extended data mining analysis performed for PMD's think tank. The initial data provided regards a data set which includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA, which served as the input for a bunch of advanced methodologies and algorithms run to reveal underlying structure and patterns that reside as latent across the data. The paragraphs to follow include, among others, a careful selection of the most significant out of these results, in terms of relevance, consistency and accuracy. The results are presented in a comprehensible and easily digestible format, ready to support decision making processes.

Goals

The analysis performed served a single goal: To extensively study the given data set in order to search for and find out the most important of the rules and patterns hidden within the data. The study, eventually, contributes the shaping of these patterns into usable knowledge, while putting focus on the given variables of specific interest.

Means

The tools and approaches used for extracting the underlying patterns out of the available data set lie in the conjunction of Artificial Intelligence / Machine Learning and Statistics, an area commonly called Data Mining. The datamine.it team leverages on extended research experience on the topic to utilize state-of-the-art tools and techniques and provide you with the most insightful of the results, while yet in an absolutely familiar way.

Outcomes

Among the vast number of results occurred and the most significant out of them to be appeared throughout the report, a sneak peek of the insights gained is provided here:

- Representatives who voted for the physician fee freeze bill and voted against the synfuels corporation cutback were mainly republicans.
- Votes for the physician fee freeze and immigration bill came primarily from the Republican Party.
- Congressmen who paired for physician fee freeze and El-Salvador aid while they opposed to the adoption of the budget resolution were in minority democrats.
- Those who voted against to the aid to Nicaraguan contras and the religious groups in schools bill while supported the bill concerning crime were in their majority democrats.

Table of Contents

The context

Data, in general

Data stands as the least biased input to decision making, the purest source of insights and knowledge. Today, data is generated, stored and used at an unprecedented rate and volume. Typical tools available to interpret data generated by commonly used tools and techniques such as statistical reports and surveys cannot respond efficiently to the hurdles today's volume of data and required in-depth analysis pose. Datamine.it presents a solution to this problem.

Data Mining, in general

Where classical approaches prove to be ineffective of the scale, speed and simplicity needed, artificial intelligence comes to join statistics and provide the much needed solution. That solution is Data Mining. You can visualize data mining as a process of searching for treasure buried in the sand or digging up rock to mine for gold - thus 'mining' -, but the tools we use do it in a truly systematic and efficient way. In our case, the rock stands for data and the gold are the insights and knowledge hidden within the data set.

That said, a miner with a mattock in his hand is a very rough way to conceptualize the complexity and state-of-the-art of the processes executed. A diverse and extended set of exploration and filtering algorithms, next to a variety of learning and meta-learning techniques, were utilized, optimized and evaluated, while the problem is a computationally intensive one and demands a highly customized approach.

Data Mine.it, in specific

The paragraphs to follow aim at providing insight on the patterns that emerge from the extended -in both width and depth- data mining analysis of the given data set. A bunch of sophisticated machine learning algorithms were run and fine-tuned by one or more datamine.it engineers to end up on extracting outcomes and patterns that make perfect sense for your dataset and really provide you with insights you never imagined before, or never thought them as being well proven; we like to call it "a tale of discovery, from your data to the report on hand". What's more, rest assured we've worked really hard to separate the wheat from the chaff, all the peculiar terminology included. And if you were used to concern a pie chart or a histogram as the most insightful thing you could expect from a data analysis, get ready to be astonished on the pages to follow.

The content

Analysis of the data set

The initial dataset consisted of 17 attributes (you may visualize it as the number of 'questions performed') and 435 instances (the number of 'samples collected'). The analytical description of attributes is provided in the Appendix I, while Table 1 that follows gives a very sneak peek.

Description	Quantity
attributes	**17**
nominal	17
numeric	0
target	1
instances	**435**
missing	15
uniques (on average)	0 (0%)

Table 1: Data set at a glance

Let's take a deeper view. Table 2 provides the titles of all attributes, which consist the data set. These are referred here to provide you with a broader view of the data in focus that are potentially utilized in the results of the following pages. Again, you may find a more detailed description of the submitted attributes in Appendix I.

#	Name	#	Name	#	Name
1	Handicapped-infants	7	Anti–satellite–test-ban	13	Superfund–right–to-sue
2	Water project–cost-sharing	8	Aid–to-Nicaraguan-contras	14	Crime
3	Adoption–of-the budget-resolution	9	Mx-missile	15	Duty–free-exports
4	Physician–fee-freeze	10	Immigration	16	Export–administration–act-south Africa
5	El-Salvador-aid	11	Synfuels–corporation-cutback	17	Vote {target attribute}
6	Religious–groups–in-schools	12	Education-spending	-	

Table 2: Titles of attributes in use

As the target for the analysis performed served the single attribute of 'vote' (#17). In other words, the analysis performed attempt to extract relationships and insights of all other attributes in regard to this one. Table 3 provides more details on this attribute, next to the distribution of its values in the given data set in Figure 1. Figures of all the attributes are given in the Appendix I.

#	Name	Type	Values	Missing	Distinct	Unique
17	vote	nominal	Democrat, Republican	0 (0%)	2	0(0%)

Table 3: Description of the target attribute

Figure 1: a) Distribution of the target attribute, b) Distribution of attribute 'education-spending', in regard to the target attribute

Due to the sample's complexity and size, various advanced filtering techniques were repeatedly utilized to firstly rank these attributes according to their correlation and informational value in regards to the analysis' target, and then put focus on the ones that matter the most. Table 4 presents the 7 most valuable out of these, as occurred by such a process, while Table 5 contributes the ones of least informational value.

#	Name
1	physician-fee-freeze
2	adoption-of-the-budget-resolution
3	el-salvador-aid
4	education-spending
5	crime
6	aid-to-nicaraguan-contras
7	mx-missile

Table 4: Attributes of most informational value

#	Name
1	anti-satellite-test-ban
2	religious-groups-in-schools
3	handicapped-infants
4	synfuels-corporation-cutback
5	export-administration-act-south africa
6	immigration
7	water-project-cost-sharing

Table 5: Attributes of low informational value

Given the rough description of the submitted data set and the analysis framework deployed before, the next paragraph stands as the core of this report, moving to the actual results of the knowledge discovery process.

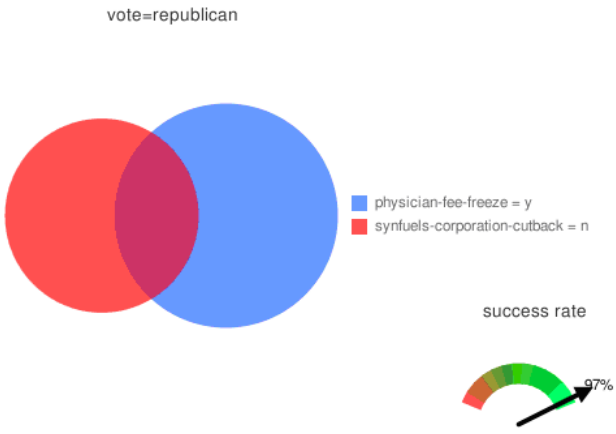
The analysis

Introduction

As referred above, the analysis performed utilized an extended variety of advanced data mining techniques and machine learning algorithms, next to the outcomes of the data set's analysis, to finally extract the best and brightest of its latent patterns. Significant effort was also put into transforming these patterns and analysis results into some direct, tangible and easily comprehensible outcomes.

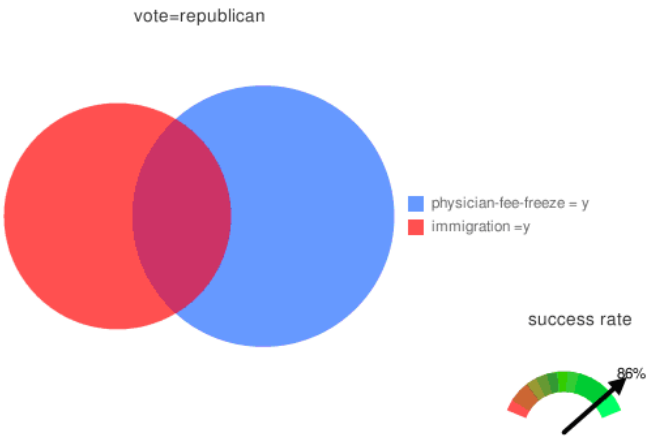
Best rules discovered

The pages to follow describe in words and figures the most significant out of the rules discovered, in other words the most distinguishable of the patterns emerged out of the extensive mining processes performed. Each pattern is also described by the number of cases that validates it across the data set, as well as its success rate. Apart from the rules presented here, Appendix II provides an extended list of (less or more) significant rules discovered, essentially contributing to the formation and understanding of the latent knowledge in the given data set.



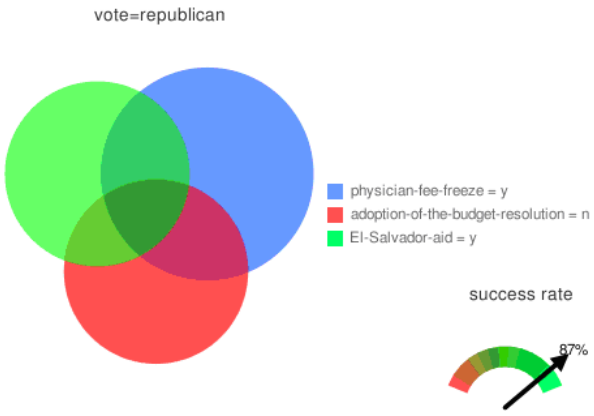
Rule 1: if physician-fee-freeze = y & synfuels-corporation-cutback = n then vote=republican (138 cases, 97% success)

Rule 1 indicates that congressmen who paired for physician fee freeze while they opposed to the synfuels corporation cutback were republicans with a certainty of 97%.



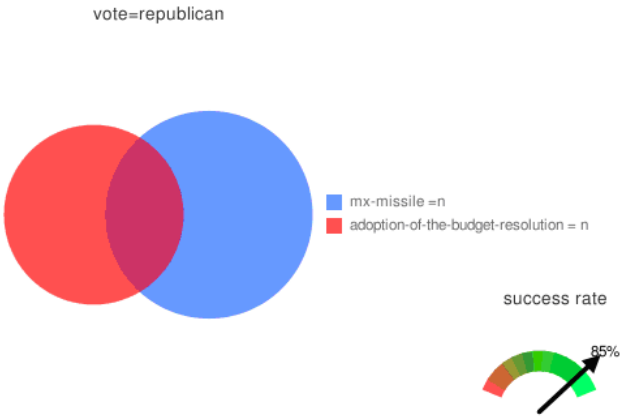
Rule 2: if physician-fee-freeze = y & immigration =y then vote=republican (118 cases, 86% success)

Rule 2 suggests with a certainty of 86% that votes for the physician fee freeze and the immigration bill came from republicans.



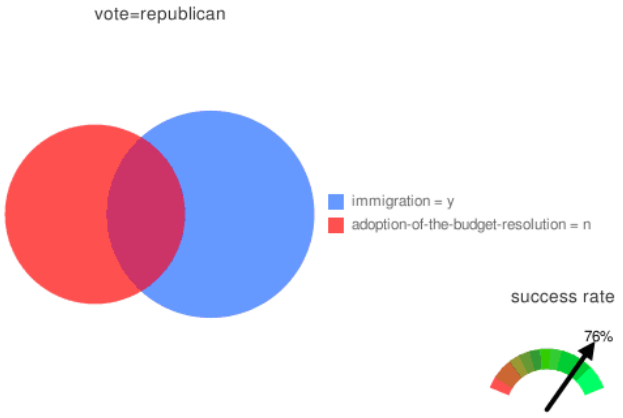
Rule 3: if physician-fee-freeze = y & adoption-of-the-budget-resolution = n & el-Salvador-aid = y then vote=republican (161 cases, 87%success)

Rule 3 provides the insight those representatives who paired for the physician fee freeze bill and the El-Salvador aid while they didn't adopt the budget resolution, were members of the Republican Party. The rule comes with an 87% rate of support.



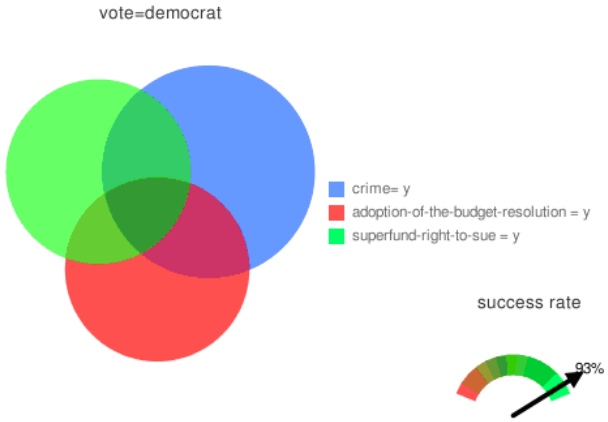
Rule 4: if mx-missile = n & adoption-of-the-budget-resolution = n then vote= republican (23 cases , 85% success)

The pattern emerging from this rule indicates that republican Congressmen voted down the bill concerning the mx-missile and the adoption of the budget resolution, with a certainty of 85%.



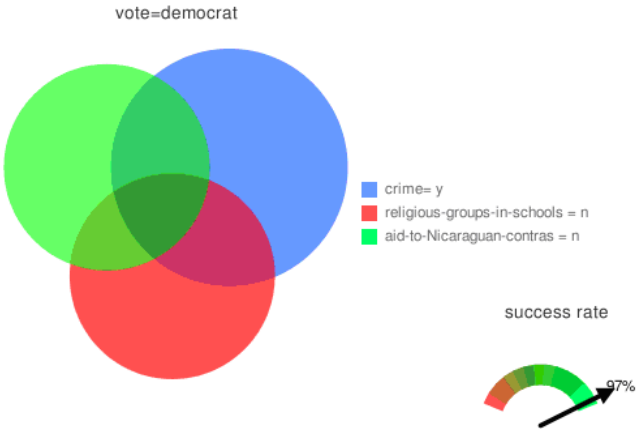
Rule 5: if immigration = y & adoption-of-the-budget-resolution = n then vote= republican (14 cases, 76 % success)

Rule 5 reveals that a congressman who gave his vote for the bill about immigration whereas opposed to the adoption of the budget resolution was republican. The rule is supported by the given data at a 76% rate of success.



Rule 6: if crime = y & adoption-of-the-budget-resolution = y & superfund-right-to-sue = y then vote= democrat (7cases, success 93%)

Rule 6 introduces that votes for the bills concerning crime, adoption of the budget resolution and superfund right to sue, came from the democrat party, at a 93% rate of support.



Rule 7: If crime = y & religious-groups-in-schools = n & aid-to-Nicaraguan-contras = n then vote= democrat (6 cases, 97% success)

Rule 7 points out those congressmen who upheld the crime bill whilst paired against the religious groups in schools and the aid to Nicaraguan, were democrats. The rule has 97% accuracy in the data set provided.

Again, the rules demonstrated here are a small part from the best of the rules found, while a much more extended set of them can be found at Appendix II.

General outcomes

The extended analysis performed and the numbers of results presented in the previous pages, as long as in the Appendix II, clearly shaped out a number of outcomes, the most significant of which are also deployed hereby:

- The physician fee freeze bill was more likely to be voted down by democrats.
- Democrats who were in favor of the anti-satellite test ban and the aid to Nicaraguan Contras bill were expected to be against the El Salvador aid bill.
- On the other hand, bills concerning crime, adoption of the budget resolution and superfund right to sue, were mostly supported by the Democratic Party.
- Representatives who paired for the physician fee freeze bill and the el-Salvador aid while didn't adopt the budget resolution were typically republicans.
- Those who were supportive of the mx-missile bill whereas opposed to the education-spending and crime bills were mainly democrats.

While the results found are presented at full extent in the Appendixes below (including the attributes analytical description and plots, most valuable -information wise- attributes and a really big list of rules extracted), it is by now clear that the on hand analysis has contributed deep insights, yet simple descriptions, on the patterns and knowledge that were lying unveiled through the submitted data set. This tale of discovery, from your data to the report on hand, seemed to reach its end, at least on the part of maximizing the value of your data input. We do believe you'll come to validate this, while we continuously remain at your request for shaping the next episode of your data tales.

Appendix I: Data set attributes

Description of data set attributes

The list of attributes of the given data set is provided here.

#	Name	Type	Values	Missing	Distinct	Unique
1	handicapped-infants	nominal	Y(yes), n(no)	12(3%)	2	0(0%)
2	water-project-cost-sharing	nominal	Y(yes), n(no)	48(11%)	2	0(0%)
3	adoption-of-the-budget-resolution	nominal	Y(yes), n(no)	11(3%)	2	0(0%)
4	physician-fee-freeze	nominal	Y(yes), n(no)	11(3%)	2	0(0%)
5	El-Salvador-aid	nominal	Y(yes), n(no)	15(3%)	2	0(0%)
6	religious-groups-in-schools	nominal	Y(yes), n(no)	11(3%)	2	0(0%)
7	anti-satellite-test-ban	nominal	Y(yes), n(no)	14(3%)	2	0(0%)
8	aid-to-Nicaraguan-contras	nominal	Y(yes), n(no)	15(3%)	2	0(0%)
9	mx-missile	nominal	Y(yes), n(no)	22(5%)	2	0(0%)
10	immigration	nominal	Y(yes), n(no)	7(2%)	2	0(0%)
11	synfuels-corporation-cutback	nominal	Y(yes), n(no)	21(5%)	2	0(0%)
12	education-spending	nominal	Y(yes), n(no)	31(7%)	2	0(0%)
13	superfund-right-to-sue	nominal	Y(yes), n(no)	25(6%)	2	0(0%)
14	crime	nominal	Y(yes), n(no)	17(4%)	2	0(0%)
15	duty-free-exports	nominal	Y(yes), n(no)	28(6%)	2	0(0%)
16	export-administration-act-south Africa	nominal	Y(yes), n(no)	104(24%)	2	0(0%)
17	vote	nominal	Democrat, Republican	0(0%)	2	0(0%)

Table 6: Analytical description of data set attributes

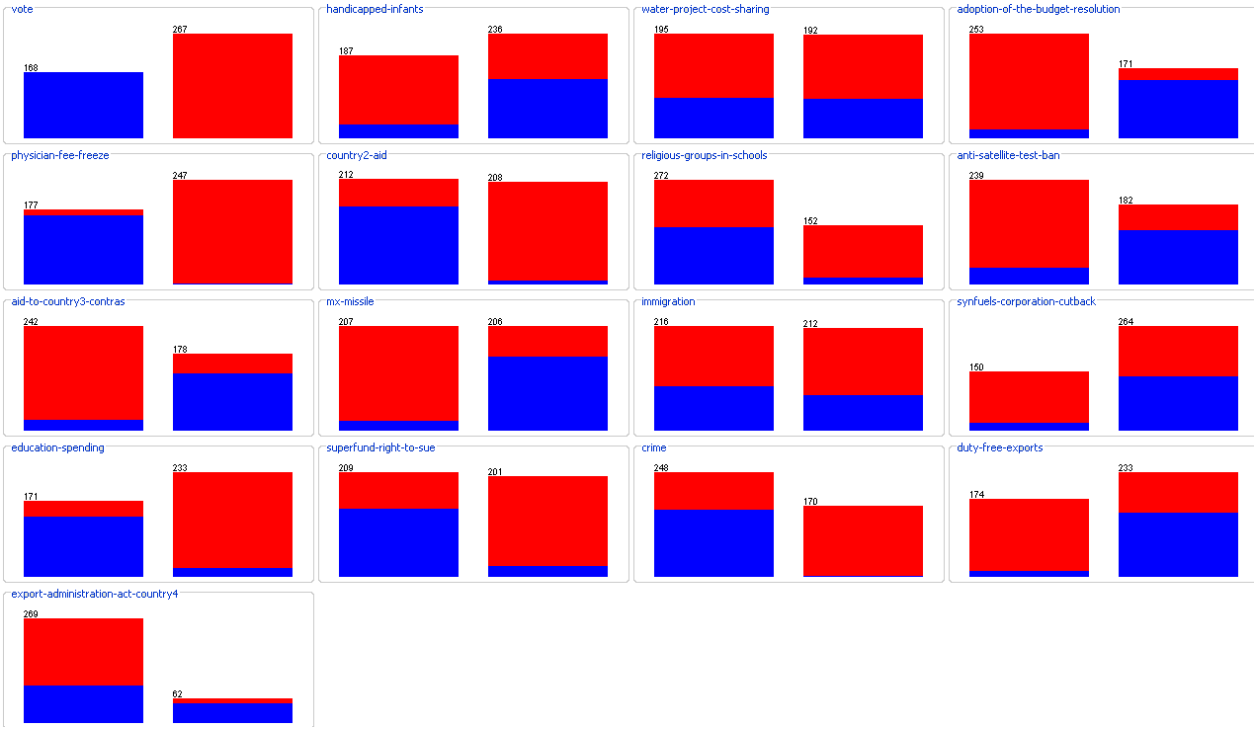


Figure 2: Visualization of the data set's distribution, according to variable 'vote'

Appendix II: Rules discovered

List of significant rules discovered

Apart from the most significant rules that were referred to in the analysis section and out of the huge bulk of rules that were found during the study of the given data set, a number of other rules are definitely worth or mentioning. These are referred to in the Table XX that follows.

#	Rule
1	If crime = y & anti-satellite-test-ban = y then vote = republican (92%success)
2	If physician-fee-freeze = n then vote = democrat (95% success)
3	If synfuels-corporation-cutback = n & education-spending = y then republican (98% success)
4	If crime = y & anti-satellite-test-ban = y: republican (88% success)
5	If el-Salvador-aid = n & crime = n then vote = democrat (99% success)
6	If education-spending = y & el-Salvador-aid = y & synfuels-corporation-cutback = n then vote = republican (97% success)
7	If mx-missile = y & religious-groups-in-schools = y then vote = democrat (68% success)
8	If el-Salvador-aid = n & immigration = y & synfuels-corporation-cutback = y then vote = democrat (89% success)
9	If crime = n & immigration = n then vote = democrat (97% success)
10	If synfuels-corporation-cutback = n & duty-free-exports = n then vote = republican (81% success)
11	If education-spending = n & synfuels-corporation-cutback = y then vote = democrat (87% success)
12	If mx-missile = n & anti-satellite-test-ban = n & duty-free-exports = n & superfund-right-to-sue = y then vote = republican (73% success)
13	If immigration = n & anti-satellite-test-ban = n then vote = democrat (95% success)
14	If el-Salvador-aid = y & education-spending = y then vote=republican (74% success)
15	If el-Salvador-aid = y & mx-missile = n & synfuels-corporation-cutback = n & crime = y & duty-free-exports = n then vote=republican (92% success)
16	If crime = y & religious-groups-in-schools = n & el-Salvador-aid = y then vote=republican (83% success)
17	If education-spending = y & aid-to-Nicaraguan-contras = n then vote=republican (64% success)
18	If crime = y & synfuels-corporation-cutback = n & mx-missile = n then vote=republican (89% success)
19	If crime = y & education-spending = y & export-administration-act-south Africa = y then vote=republican(80%success)
20	If crime = y & religious-groups-in-schools = n & (aid-to-Nicaraguan-contras = n) => vote=republican (75 %success)

#	Rule
21	If el-Salvador-aid = y & mx-missile = n & synfuels-corporation-cutback = n & crime = y & duty-free-exports = n then vote=republican (90% success)
22	If education-spending = n & crime = n then vote = democrat (87 %success)
23	If synfuels-corporation-cutback = n & mx-missile = n & education-spending = y: republican (95%success)
24	If duty-free-exports = y & mx-missile = y then vote = democrat (91% success)
25	If synfuels-corporation-cutback = y & education-spending = n then vote = democrat (90% success)
26	If superfund-right-to-sue = n & immigration = y & aid-to-Nicaraguan-contras = y& mx-missile = y then vote = democrat (84% success)
27	If handicapped-infants = n & anti-satellite-test-ban = y then vote = republican (85% success)
28	If aid-to-Nicaraguan-contras = y & mx-missile = y then vote = democrat (83%success)
29	If duty-free-exports = y & water-project-cost-sharing = y then vote = democrat (95% success)
30	If duty-free-exports = n & superfund-right-to-sue = y & water-project-cost-sharing = y then vote=republican (80% success)
31	If duty-free-exports = n & water-project-cost-sharing = n & synfuels-corporation-cutback = y then vote=democrat (69% success)
32	If superfund-right-to-sue = n & immigration = y & mx-missile = y then vote= democrat (79%success)
33	If synfuels-corporation-cutback = n & duty-free-exports = n then vote= republican (84%success)

Table 7: Extended list of significant rules discovered

Contact Information

This report was prepared by Lina Massou, data engineer. You may contact her directly at lina@datamine.it.

This report was prepared for Peter Ronson, Political Analyst, PMD Corporation.

datamine.it

14 Meletiou Vasileiou Str
11 745 Athens, Greece
T +30 6937 122 065
go@datamine.it
http://datamine.it