datamine.it

Analysis Report

Prepared for: Raju Chandan, Manager, ACME Corporation

Prepared by: Athina Pandi, Data Engineer, datamine.it

November 7th, 2008

Report number: 000-0001

Executive Summary

Objective

The hereby report summarizes the results of the extended data mining analysis performed for ACME Corporation. The initial data provided regards a survey on web usage for automotive clients, 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:

- people of age 26-35 with high knowledge level, visit frequently automotive webpages
- very knowledge level in either men or women results in only rare visits
- low knowledge level about technology is followed by no visits
- attributes of most informational value, in relation to the target ‘internet use’, were general_tech_knowledge_level, car_tech_knowledge_level and grade_tch_ad_auto

The totality of contents of this report consist a work and property of datamine.it.

Table of Contents

The context

Data, in general

Data stand as the least biased input to decision making, a pure source of insights and knowledge. And data today is generated, stored and, literally, used in an unprecedented rate. However, time spent on consuming these data remains constant and, what's more, the typical tools to serve this task turn out to be incapable; the resulting 'data gap' is today an omnipresent reality. In this context, common and widely used techniques and approaches, like surveys and the way they are analyzed, or statistical reports, clearly cannot respond efficiently to the hurdles data volume and its in depth analysis pose. If all these leave much to be desired, datamine.it and the on-hand report comes to the rescue, at least for the data in focus.

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. Data mining that is, and you can visualize it as the way and process of searching for secrets bared in the sand, or drilling for gold in a mine -thus 'mining'-, but in a truly systematic and efficient way. In our case, stone stands for data and gold for the insights and knowledge hidden within the data set, while the single purpose of this report is to provide you with evidence on the existence and the description of this very treasure.

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 37 attributes (you may visualize it as the number of ‘questions performed’) and 319 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	**37**
nominal	37
numeric	0
target	1
instances	**317**
missing	0
uniques (on average)	6 (2%)

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	age	14	4x4	27	ESP
2	sex	15	ABS	28	Immediate spraying
3	car owner	16	Diesel	29	Karter
4	car value	17	Katalyst	30	Sinemplok
5	general_tech_knowledge_level	18	Immobilizer	31	ECU
6	car_tech_knowledge_level	19	Turbo	32	DSG
7	my_car_knowledge_level	20	Hybrid	33	Wastegate

#	Name	#	Name	#	Name
8	design	21	16v	34	SMG
9	practicality	22	Dynamo	35	grade_tch_ad_auto
10	technology	23	Cruise control	36	extended report
11	value	24	Differential gear	37	internet use {target attribute}
12	brand	25	Spoiler		
13	performance	26	Xenon		

Table 2: Titles of attributes in use

As the target for the analysis performed served the single attribute of ‘internet usage’ (#37). 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
37	internet use	nominal	never, rarely, some-times, frequently, always	1 (0%)	5	0 (0%)

Table 3: Description of the target attribute

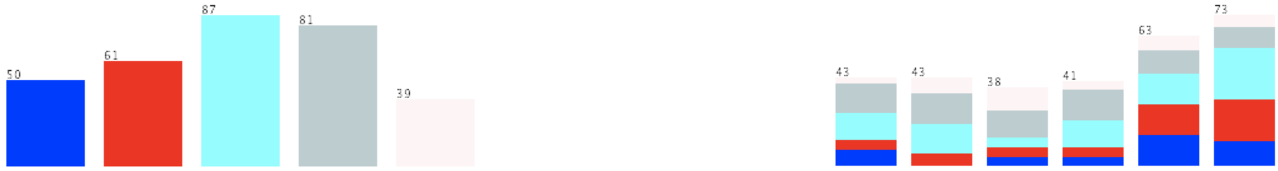


Figure 1: a) Distribution of the target attribute, b) Distribution of attribute ‘performance’, in regard to the target attribute

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 10 most valuable out of these, as occurred by such a process, while Table 5 contributes the ones of least informational value.

#	Name
1	general_tech_knowledge_level
2	car_tech_knowledge_level
3	grade_tch_ad_auto
4	Xenon
5	performance
6	Immediate_spraying
7	16v
8	Cruise_control
9	my_car_knowledge_level
10	practicality

Table 4: Attributes of most informational value

#	Name
1	car_owner
2	value
3	Diesel
4	Sinemplok
5	Wastegate
6	SMG
7	ABS
8	Katalyst
9	4x4
10	extended report

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.

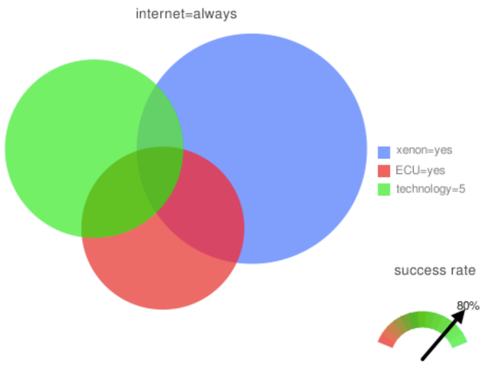
The analysis

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.

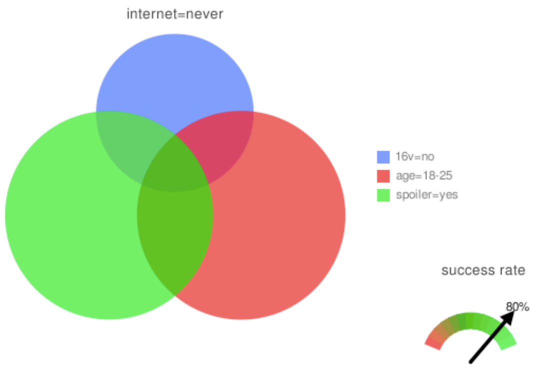
Best rules discovered

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.



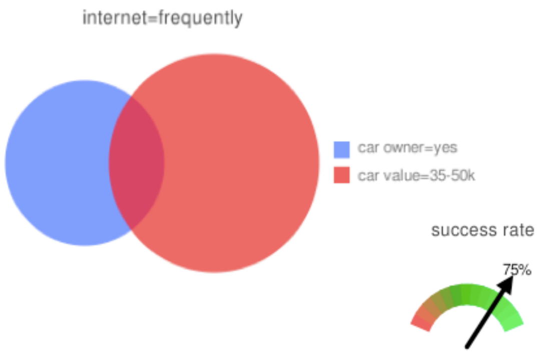
Rule 1: if Xeron=yes & ECU=yes & technology=5 then internet = always (80% success)

Rule 1 indicates that an individual who happens to know the meaning of Xenon and ECU technologies, while she also has a high level of technology knowledge, will use the web as an information resource on automotive news with a certainty of 80%.



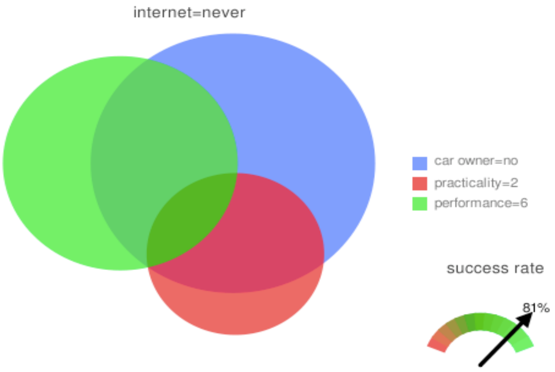
Rule 2: if 16v=no & age=18-25 & spoiler=yes then internet = never (80% success)

Rule 2 suggests, with a 80% certainty, that a young user (aged 18-25) who does understand about spoilers but not about 16v features won't search for relevant information in a website.



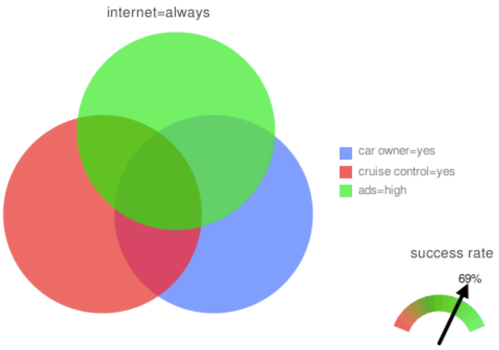
Rule 3: if car owner=yes & car value=35-50k then internet = frequently (75% success)

Rule 3 provides the insight that a typical owner of a car valued between 35 to 50 thousand euros is expected to use the internet frequently for searching relevant to cars information.



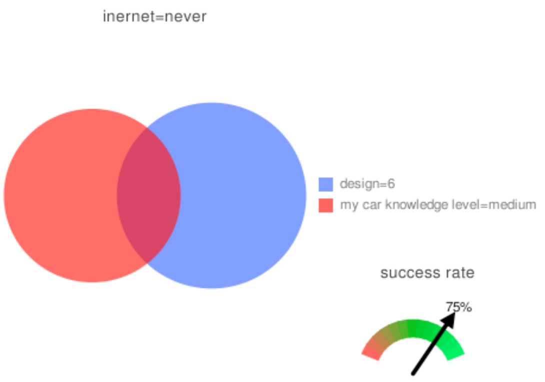
Rule 4: if car owner=no & practicality=2 & performance=6 then internet = never (81% success)

The pattern emerging from this rule indicates that for non car owners with a low record on practicality and a strong call for performance, the web appears not to be their media of choice. The rule is supported by the given data at a 81% rate of success.



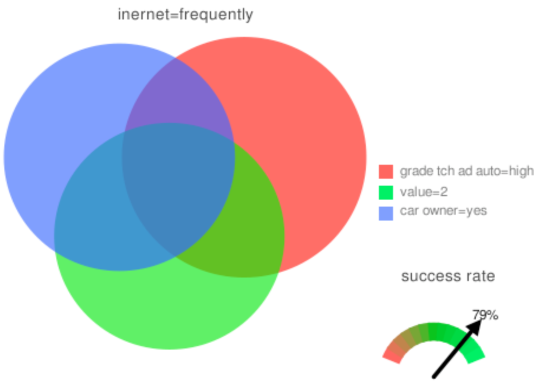
Rule 5: if car owner=yes & cruise control=yes & ads=high then internet = always (69% success)

Rule 5 reveals that a car owner, who does know about cruise control and watches a considerable number of relative to cars advertisements, will always use the web as a medium for his updates. That rule comes with a 69% rate of support.



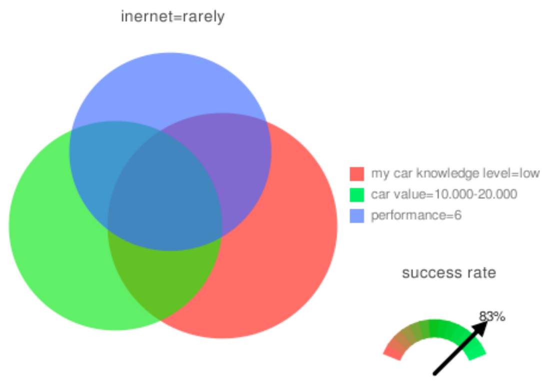
Rule 6: if design = 6 and my_car_knowledge_level = medium then internet = never (75% success)

Rule 6 introduces that a customer putting significant weight into design (answers the relevant question giving a value of 6), while possessing a medium level of knowledge of her car, is never expected to utilize the web for finding relevant to car information.



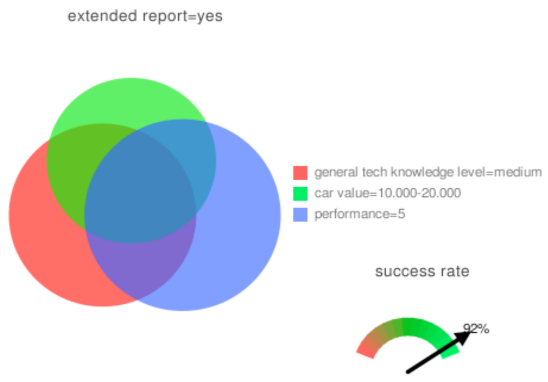
Rule 7: if grade_tch_ad_auto = high AND value = 2 AND car_owner = yes then internet = frequently (79% success)

Rule 7 points out that a car owner of value equal to 2 and great exposure to auto advertisement, will use the web frequently for the under study purpose. The rule has a 79% accuracy in the data set provided.



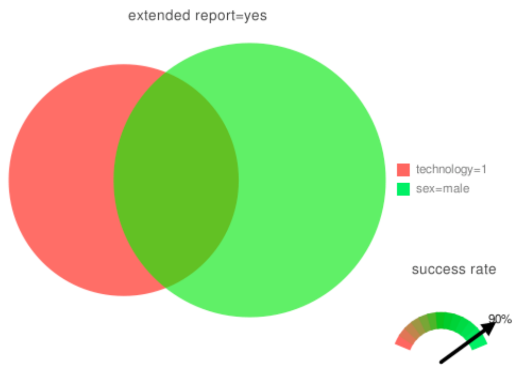
Rule 8: if my_car_knowledge_level = low and car_value = 10.000-20.000 and performance = 6 then internet=rarely (83% success)

Another pattern suggests that a owner of a car valued between 10 and 20 thousand euros, with a low level of knowledge of her car, but a strong tendency to high performance, is rarely expected to use the web as an informational source.



Rule 9: if performance = 5 and general_tech_knowledge_level = medium and car_value = 10.000-20.000 then extended report = yes (92% success)

Moving focus next to the web usage attribute, the analysis has also introduced some other, equally valuable insights. Rule 9 for example shows that owner of a car valued between 10 and 20 thousand euros, with medium tech knowledge level and high focus on performance is expected to read an extended report with a strong probability of 92%.



Rule 10: if technology = 1 AND sex = male then extended report = yes (90% success)

Equally safe and insightful proves to be Rule 10 that suggests a male customer, positive to technology, as being a reader of an extended report.

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 out of which are also deployed hereby:

- Most of the respondent having a very low knowledge level about technology in general, never visit car-industries websites to gain further information about a car seen in an advertisement.
- The same remark stands for people whose level of understanding cars advertisements is low.
- On the other hand, people whose knowledge level is very high always visit these websites.
- As far as the gender is concerned, women and men with very low knowledge level, rarely visit these websites.
- Low budget car owners of low knowledge level about their car and high focus on performance will rarely use the web.
- Finally, people who are 26-35 years old with high knowledge level about cars technology frequently visit these sites.

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	age	nominal	-17, 18-25, 26-35, 36-45, 46-55, 56-65, 66+	0 (0%)	4	0 (0%)
2	sex	nominal	male, female	0 (0%)	2	0 (0%)
3	car owner	nominal	yes, no	0 (0%)	2	0 (0%)
4	car value	nominal	0-10.000, 10.000-20.000, 20.000-35.000, 35.000-50.000, 50.000+	61 (19%)	5	0 (0%)
5	general_tech_knowledge_level	nominal	very-low, low, medium, high, very-high	2 (1%)	5	0 (0%)
6	car_tech_knowledge_level	nominal	very-low, low, medium, high, very-high	3 (1%)	5	0 (0%)
7	my_car_knowledge_level	nominal	very-low, low, medium, high, very-high	50 (16%)	5	0 (0%)
8	design	nominal	1, 2, 3, 4, 5, 6	21 (7%)	6	0 (0%)
9	practicality	nominal	1, 2, 3, 4, 5, 6	18 (6%)	6	0 (0%)
10	technology	nominal	1, 2, 3, 4, 5, 6	20 (6%)	6	0 (0%)
11	value	nominal	1, 2, 3, 4, 5, 6	20 (6%)	6	0 (0%)
12	brand	nominal	1, 2, 3, 4, 5, 6	21 (7%)	6	0 (0%)
13	performance	nominal	1, 2, 3, 4, 5, 6	18 (6%)	6	0 (0%)
14	4x4	nominal	0, 1	0 (0%)	2	0 (0%)
15	ABS	nominal	0, 1	0 (0%)	2	0 (0%)
16	Diesel	nominal	0, 1	0 (0%)	2	0 (0%)
17	Katalyst	nominal	0, 1	0 (0%)	2	0 (0%)
18	Immobilizer	nominal	0, 1	0 (0%)	2	0 (0%)
19	Turbo	nominal	0, 1	0 (0%)	2	0 (0%)
20	Hybrid	nominal	0, 1	0 (0%)	2	0 (0%)
21	16v	nominal	0, 1	0 (0%)	2	0 (0%)
22	Dynamo	nominal	0, 1	0 (0%)	2	0 (0%)
23	Cruise control	nominal	0, 1	0 (0%)	2	0 (0%)
24	Differential gear	nominal	0, 1	0 (0%)	2	0 (0%)

#	Name	Type	Values	Missing	Distinct	Unique
25	Spoiler	nominal	0, 1	0 (0%)	2	0 (0%)
26	Xenon	nominal	0, 1	0 (0%)	2	0 (0%)
27	ESP	nominal	0, 1	0 (0%)	2	0 (0%)
28	Immediate spraying	nominal	0, 1	0 (0%)	2	0 (0%)
29	Karter	nominal	0, 1	0 (0%)	2	0 (0%)
30	Sinemplok	nominal	0, 1	0 (0%)	2	0 (0%)
31	ECU	nominal	0, 1	0 (0%)	2	0 (0%)
32	DSG	nominal	0, 1	0 (0%)	2	0 (0%)
33	Wastegate	nominal	0, 1	0 (0%)	2	0 (0%)
34	SMG	nominal	0, 1	0 (0%)	2	0 (0%)
35	grade_tch_ad_auto	nominal	very-low, low, medium, high, very-high	3 (1%)	1	0 (0%)
36	extended report	nominal	yes, no	1 (0%)	2	0 (0%)
37	internet use {target attribute}	nominal	never, rarely, some-times, frequently, always	1 (0%)	5	0 (0%)

Table x: Analytical description of data set attributes

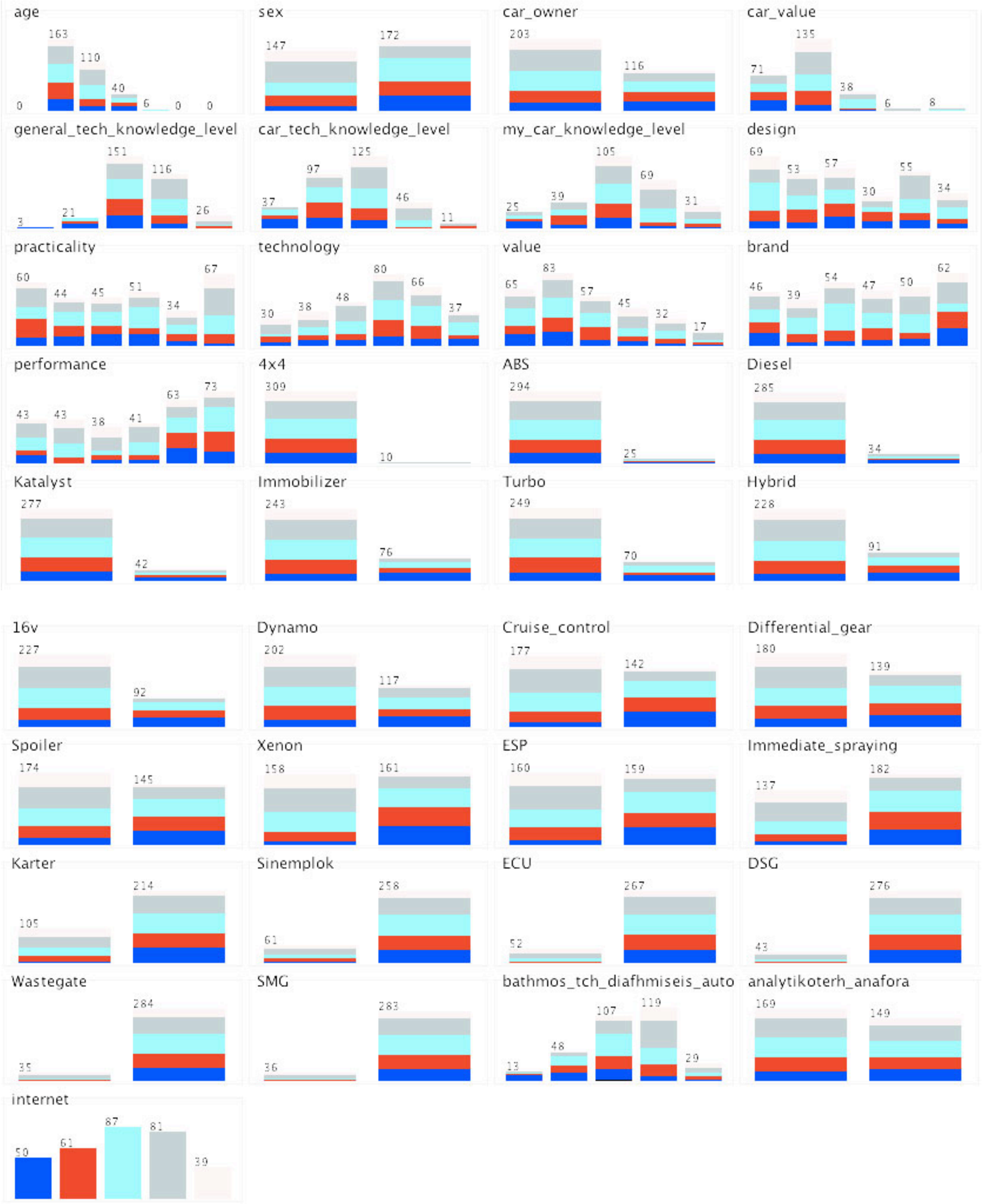


Figure x: Visualization of the data set's distribution, according to variable 'internet'

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 Xenon = 1 & general_tech_knowledge_level = very-high & car_value = 10.000-20.000 & ECU = 1 then always (78% success)
2	if Xenon = 1 & car_tech_knowledge_level = very-high & grade_tch_ad_auto = very-high then rarely (83% success)
3	if Xenon = 1 & ECU = 1 & technology = 3 & Sinemplok = 1 then frequently (5.25% success)
4	if Xenon = 1 & ECU = 1 & technology = 5 then always (79% success)
5	if Xenon = 1 & general_tech_knowledge_level = very-high & car_value = 20.000-35.000 then always (80% success)
6	if 16v = 0 & general_tech_knowledge_level = low & Differential_gear = 1 then never (100% success)
7	if Xenon = 1 & car_value = 20.000-35.000 & Sinemplok = 0 & sex = male & age = 26-35 & my_car_knowledge_level = high then some-times (83% success)
8	if Xenon = 1 & Immediate_spraying = 1 & general_tech_knowledge_level = very-high then frequently (74% success)
9	if Xenon = 1 & Cruise_control = 1 & grade_tch_ad_auto = low & Hybrid = 1 then some-times (6% success)
10	if Xenon = 1 & Cruise_control = 1 & Immobilizer = 0 & performance = 2 then frequently (89% success)
11	if 16v = 0 & grade_tch_ad_auto = very-low & Cruise_control = 0 & extended report = no then never (100% success)
12	if 16v = 0 & age = 36-45 & Spoiler = 1 then never (80% success)
13	if 16v = 0 & age = 18-25 & sex = male & car_owner = yes then rarely (80% success)
14	if 16v = 0 & age = 18-25 & value = 1 & Spoiler = 0 & Cruise_control = 0 then some-times (71% success)
15	if Xenon = 1 & grade_tch_ad_auto = medium & technology = 5 & Sinemplok = 0 then some-times (94% success)
16	if Xenon = 1 & car_value = 50.000+ & Immediate_spraying = 1 then frequently (72% success)
17	if 16v = 0 & age = 18-25 & value = 3 & sex = female then rarely (71% success)
18	if car_tech_knowledge_level = very-low & value = 5 then never (96% success)
19	if car_owner = yes & car_value = 20.000-35.000 & DSG = 0 then some-times (85% success)
20	if car_owner = yes & car_value = 35.000-50.000 then frequently (75% success)
21	if car_owner = yes & car_value = 50.000+ then some-times (75% success)
22	if Xenon = 1 & ECU = 1 & design = 2 & general_tech_knowledge_level = high then frequently (88% success)
23	if car_owner = yes & design = 5 & Sinemplok = 1 then frequently (88% success)
24	if car_tech_knowledge_level = high & extended report = no then some-times (81% success)
25	if car_tech_knowledge_level = high & car_value = 10.000-20.000 & design = 3 then frequently (77% success)
26	if car_tech_knowledge_level = high & design = 6 then frequently (77% success)

#	Rule
27	if car_tech_knowledge_level = very-high then frequently (0.99% success)
28	if Xenon = 0 & design = 2 & Hybrid = 1 & Karter = 0 then frequently (95% success)
29	if Xenon = 0 & age = 36-45 & car_tech_knowledge_level = low & car_owner = yes & 16v = 1 then never (100% success)
30	if car_owner = yes & grade_tch_ad_auto = low & car_value = 0-10.000 then some-times (83% success)
31	if car_owner = yes & grade_tch_ad_auto = low & performance = 6 then rarely (100% success)
32	if car_owner = yes & grade_tch_ad_auto = medium & design = 2 then some-times (83% success)
33	if car_owner = yes & grade_tch_ad_auto = medium & design = 1 & sex = female then some-times (84% success)
34	if car_owner = yes & grade_tch_ad_auto = high & technology = 4 & Immediate_spraying = 0 then rarely (86% success)
35	if car_owner = yes & design = 5 & 16v = 1 & general_tech_knowledge_level = medium & grade_tch_ad_auto = medium then some-times (100% success)
36	if car_owner = yes & design = 5 then frequently (80% success)
37	if Xenon = 0 & age = 26-35 & Spoiler = 0 then some-times (78% success)
38	if Xenon = 0 & age = 36-45 then rarely (100% success)
39	if car_owner = no & practicality = 2 & performance = 6 then never (83% success)
40	if car_owner = yes & car_value = 0-10.000 & Hybrid = 1 & general_tech_knowledge_level = high then frequently (77% success)
41	if car_owner = yes & car_value = 0-10.000 & performance = 6 then rarely (75% success)
42	if 16v = 0 & Dynamo = 1 & ESP = 0 then rarely (74% success)
43	if Spoiler = 0 & performance = 5 then never (75% success)
44	if car_owner = yes & Cruise_control = 1 & grade_tch_ad_auto = high then always (69% success)
45	if performance = 4 & general_tech_knowledge_level = medium then rarely (71% success)
46	if performance = 1 then never (77% success)
47	if value = 4 & Cruise_control = 1 then always (75% success)
48	if grade_tch_ad_auto = low then some-times (65% success)
49	if practicality = 6 & Cruise_control = 0 then rarely (94% success)
50	if practicality = 1 & Immediate_spraying = 0 then rarely (76% success)
51	if practicality = 1 then frequently (87% success)
52	if grade_tch_ad_auto = high then some-times (80% success)
53	if value = 2 & Differential_gear = 1 then never (83% success)
54	if value = 1 & car_owner = yes then some-times (80% success)
55	if technology = 4 then some-times (66% success)
56	if car_tech_knowledge_level = low then frequently (66% success)
57	if { } then always (96% success)

Table x: Analytical description of data set attributes

Contact Information

This report was prepared by Athina Pandi, data engineer. You may contact her directly at athina@datamine.it.

This report was prepared for Raju Chandan, Manager, ACME Corporation.

datamine.it

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

This report remains the property of datamine.it and its content and format are for the exclusive use of the ACME Corporation.