Customer Survey Analysis

Brief Description

Customer Survey Analysis explores how data mining tools can be used for analyzing marketing data. Suppose you're a marketing manager attempting to optimize an advertising campaign of your product by making the directing campaign to your best potential customers. You can increase the effectiveness of your marketing campaign by precisely determining the best potential customers and wisely placing the product in relevant advertisements.

The data and the objective of the present investigation are real. However, for privacy reasons names of people, companies and products involved are disguised.

Case Study: To better understand the lifestyles of potential purchasers of the Discovery SUV produced by Land Rover, marketing research manager Paul Montopoli commissioned a study of consumers' attitudes, interests, and opinions. A questionnaire was designed with 30 statements covering a variety of dimensions, including consumers' attitudes towards risk, foreign versus domestic products, product styling, spending habits, self-image, and family. The questionnaire included a final question of attitude towards purchasing the Land Rover Discovery.

The respondents used a nine-point Likert scale, where a value of "1" meant that they definitely disagreed with a statement, and "9" meant that they definitely agreed. A total of 400 respondents were obtained from the mailing lists of Car and Driver, Business Week, and Inc. magazines, who were then interviewed at their homes by an independent surveying company.

Responses: Marketing managers wish to answer two basic questions based on the responses to the questionnaire.

1. Who is the typical sport-utility vehicle customer?
2. How should Land Rover position, feature, and advertise Discovery?

The first question can be answered by analyzing data solely from the responses. After discovering the profile of a typical sport-utility vehicle customer, Land Rover can direct relevant advertising toward specific consumers. A relevant ad campaign includes decisions like selecting TV shows in between which to display Land Rover Discovery commercials.

The second question requires analyzing additional data from external sources with questions like:

Given the age, sex, marital status, income, etc. for each customer what is the probability that the customer will respond to the promotion by purchasing the Discovery?

To answer this question, we need existing external demographic data. Usually that data is available from independent vendors and contains some demographic and financial characteristics of people. This task is beyond the scope of the present lesson and will not be considered here.

Survey Data: The survey consisted of 31 statements as listed below:

1. I am in very good physical condition.
2. When I must choose between the two, I dress for fashion, not comfort.
3. I have more stylish clothes than most of my friends.
4. I want to look a little different from others.
5. Life is too short not to take some gambles.
6. I am not concerned about the ozone layer.
7. I think the government is doing too much to control pollution.
8. Basically, society today is fine.
9. I don't have time to volunteer for charities.
10. Our family is not too heavily in debt today.
11. I like to pay cash for everything I buy.
12. I pretty much spend for today and let tomorrow bring what it will.
13. I use credit cards because I can pay the bill off slowly.
14. I seldom use coupons when I shop.
15. Interest rates are low enough to allow me to buy what I want.
16. I have more self-confidence than most of my friends.
17. I like to be considered a leader.
18. Others often ask me to help them out of a jam.

1. Children are the most important thing in a marriage.
2. I would rather spend a quiet evening at home than go out to a party.
3. Foreign-made cars can't compare with American-made cars.
4. The government should restrict imports of products from Japan.
5. Americans should always try to buy American products.

6. I would like to take a trip around the world.
7. I wish I could leave my present life and do something entirely different.
8. I am usually among the first to try new products.
9. I like to work hard and play hard.
10. Skeptical predictions are usually wrong.
11. I can do anything I set my mind to.
12. Five years from now, my income will be a lot higher than it is now.

The 31st first statement: I would consider buying the Discovery made by Land Rover.

The file contains 31 fields with 400 records. Each record belongs to one respondent. The first field is the 31st question and is labeled Attitude. The rest of the fields are abbreviated as Q1, Q2, and so on. For this tutorial you will need to refer to the above list for the full question.

TOP▲

Step 1: Import the Data

1. From the toolbar click the 🗋 icon to create a new project. The Import Wizard dialog box appears.
2. The data for this tutorial is stored in a file labeled *Lrover.csv*.
4. Click **OK**

TOP▲

Step 2: Using Decision Tree Analysis to Direct Exploration

Case Study: Determining the relation between the Attitude of a person and his or her responses to other statements can reveal which statements are influential and significant. Strongly correlated statements help in profiling the potential customer. To find and analyze these influential attributes we will use the following *PolyAnalyst* exploration engines: **Decision Tree** analysis, **Stepwise Linear Regression**, **Find Dependencies**, and **Find Laws**.

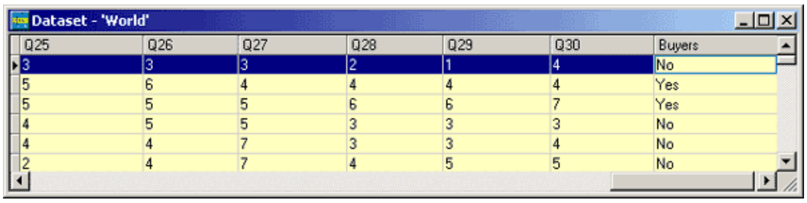
Decision Tree Analysis provides a quick and intuitive method of exploring the dataset. **Decision Tree** exploration engine requires a target variable that is of Boolean or categorical data type. We first create a Boolean attribute of Attitude that divides those surveyed into two groups: those with a high attitude and those with a low attitude toward purchasing Discovery.

1. From the main menu select **Create Object | Create Rule…**
2. Name the rule *Buyers*.
3. Using the **Rule Assistant**, create the following rule:

   Attitude>7

4. Click **OK**. This rule is a conditional statement that when applied to a dataset will either return a value of "*yes*" or "*no*". It will return a "yes" when the value of the Attitude attribute is greater than 7. Conversely, the rule will return a value of "*no"* when the Attitude value is less than or equal to 7. Our next step is to apply the rule to the ***World*** dataset.
5. Right click on the *Buyers* rule in the **Workspace** panel. From the context menu select **Apply to** | **World**. "*Buyers*" is now the last attribute (column) in the dataset ***World***.

We are ready to use **Decision Tree** analysis.

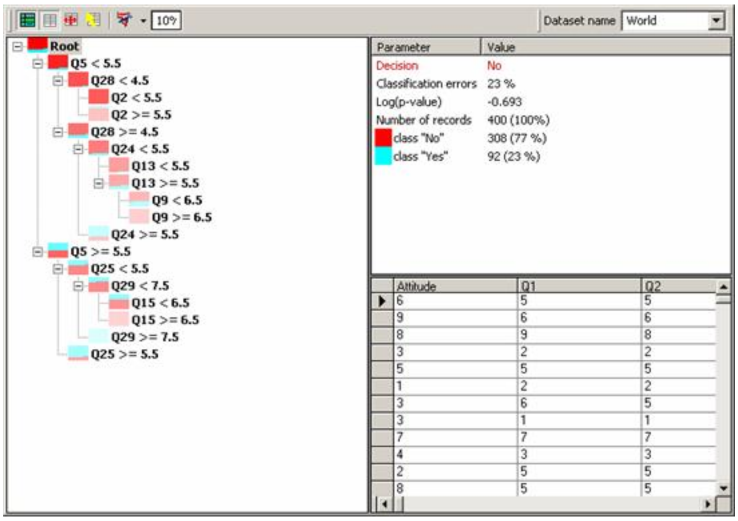


Q25	Q26	Q27	Q28	Q29	Q30	Buyers
3	3	3	2	1	4	No
5	6	4	4	4	4	Yes
5	5	5	6	6	7	Yes
4	5	5	3	3	3	No
4	4	7	3	3	4	No
2	4	7	4	5	5	No

6. Right click on the ***World*** dataset and select **Explore** | **Decision tree…(DT)**. The **Decision Tree – World** dialog box appears.
7. Name the process "*DT_Buyers*".
8. Double click the *Buyers* attribute to select it as the target attribute.
9. Right click the *Attitude* attribute to deselect it. We do not wish to include *Attitude* in our exploration because the *Buyers* attribute is a form of the *Attitude* attribute and is already included.
10. Click **OK**. The Decision Tree report appears in a few seconds.

Note: **Decision Trees** is a probabilistic exploration engine, thus your results may vary from the results below.

Analyzing the results: There are two parts of the report, a text section and the tree object. The text section displays the characteristics of the exploration and the confusion matrix. The total classification error is 13.75%, pretty good for real world data. The classification probability is very high at 86.25%.



- Root
  - Q5 < 5.5
    - Q28 < 4.5
      - Q2 < 5.5
      - Q2 >= 5.5
    - Q28 >= 4.5
      - Q24 < 5.5
        - Q13 < 5.5
        - Q13 >= 5.5
          - Q9 < 6.5
          - Q9 >= 6.5
      - Q24 >= 5.5
  - Q5 >= 5.5
    - Q25 < 5.5
      - Q29 < 7.5
        - Q15 < 6.5
        - Q15 >= 6.5
      - Q29 >= 7.5
    - Q25 >= 5.5

Parameter	Value
Decision	No
Classification errors	23 %
Log(p-value)	-0.693
Number of records	400 (100%)
🟥 class "No"	308 (77 %)
🟦 class "Yes"	92 (23 %)

Attitude	Q1	Q2
6	5	5
9	6	6
8	9	8
3	2	2
5	5	5
1	2	2
3	6	5
3	1	1
7	7	7
4	3	3
2	5	5
8	5	5

The decision tree looked at the value of the "yes" or "no" *Buyers* attribute for each record. The confusion matrix reveals that out of 292 *no* classifications there were 16 *yes's*, and that out of 53 *yes* classifications there are 39 *no's*.

The attributes included in the decision tree are: *Q5, Q28, Q25, Q2, Q24, Q29, Q13, Q15,* and *Q9*. The decision tree diagram begins by splitting the root of the tree with the attribute *Q5* at the threshold of 5.5. Click on the two nodes of the first split, >=5.5 and <5.5. Our target customers are those where the decision is yes as these are customers with a high attitude towards purchasing an SUV. The exploration reveals several characteristics of the Land Rover target customer. We see that the customer is a risk-taker, likes to travel, is self-confident, and adventurous. Our exploration seems very true for the real world SUV drivers.

TOP▲

Step 3: Using Find Dependencies to Exclude Outliers

There are two modifications of the **Find Dependencies** algorithm, **strict** and **liberal**. **Strict** dependencies will return the small list of the most influential and significant attributes. **Liberal Find Dependencies** finds a broader set of significant attributes. Liberal Dependencies helps exclude those records that are not significant.

1. Right click on the ***World*** dataset and select **Explore** | **Find Dependencies…(FD)**.
2. Name the process *FD_liberal*.
3. Set the time limit to 2 minutes.
4. Select *Attitude* as the target attribute.
5. Deselect the *Buyers* attribute.
6. Click **OK**.

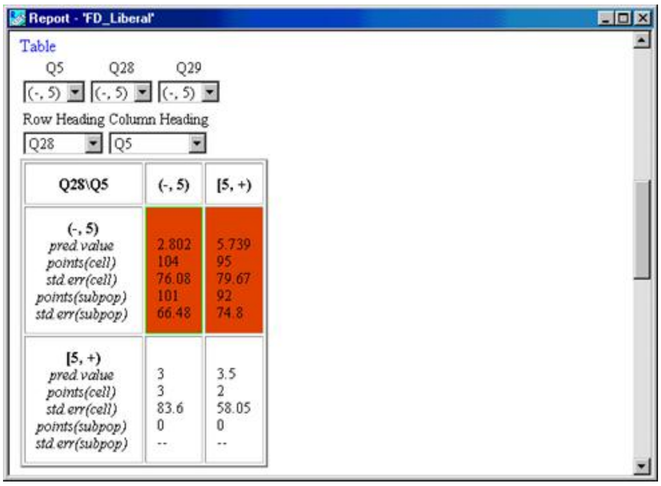
Analyzing the Results: The liberal dependencies exploration found *Q5, Q28* and *Q29* to be the most influential attributes. Reassuringly, these attributes were also found to be significant by the **Decision Tree** analysis.

Q5 – Life is too short not to take some gambles.

Q28 - Skeptical predictions are usually wrong.

Q29 - I can do anything I set my mind to.

Find Dependencies finds the best potential customer to be a risk-taker, optimistic, and self-confident. These characteristics too seem very true for the real world SUV drivers!



Table

Q5 (-, 5) Q28 (-, 5) Q29 (-, 5)

Row Heading Q28 Column Heading Q5

Q28\Q5	(-, 5)	[5, +)
(-, 5) *pred value* *points(cell)* *std err(cell)* *points(subpop)* *std err(subpop)*	2.802 104 76.08 101 66.48	5.739 95 79.67 92 74.8
[5, +) *pred value* *points(cell)* *std err(cell)* *points(subpop)* *std err(subpop)*	3 3 83.6 0 --	3.5 2 58.05 0 --

One can note that **FD** split the data into groups using exactly the mean values of the corresponding attributes on a 9-point scale. If the values of all three independent attributes *Q5*, *Q28*, and *Q29* are less than 5, then we see a large group of people - 101 of the respondents - which are not going to buy Discovery any time soon: for this group *Attitude* equals 2.8 on average - these people are not the best potential customers for Land Rover. Thus, the marketing campaign should try to avoid wasting marketing efforts on this group. These people can be characterized by their unwillingness to take risks, lack of self-confidence, and a skeptical approach to life. Since this group comprises around 25% of the total number of respondents, one would save significant marketing resources if these people can be identified correctly beforehand. As has been mentioned before, this would require a simultaneous analysis of additional demographic data drawn from some external sources. Such an analysis is beyond the scope of the present lesson. Here we simply obtained a clue on how we should search for our worst potential customers in order to stay away from them in the marketing campaign.

There is also a directly opposite group of people who gave the statements *Q5, Q28, Q29* scores above 5. This group embraces 93 people, or about 23% of the total number of respondents. An average value of *Attitude* is more than 7.2 for this group. These potential customers comprise the best group to target in the marketing campaign.

Next we create a complement to **Find Dependencies** that we can then compare to the liberal dataset in a snake chart.

7. Right click on the ***FD_liberal*** dataset and select **Create Complement**.

8. Name the dataset "*Exceptions*".

9. Click **OK**.

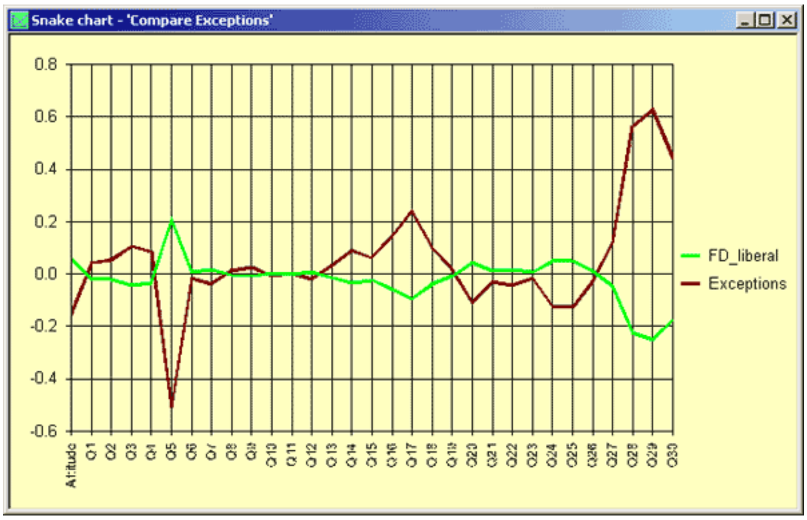
10. From the main file menu, select **Create Object** | **Create Snake-chart…**

11. Name the chart "*Compare Exceptions*".

12. Select the datasets **Exceptions** and **FD_liberal** to be included.

13. Make sure the box **Normalize by dispersion** is checked. This will allow us to better visually compare the two datasets.

14. Click **OK**. The chart is created.



Notice where the red line dips far away from the green line, or where there are major differences between the ***Exceptions*** dataset and the ***FD_liberal*** dataset. The datasets contrasts severely over *Q28, Q29*, and *Q5*. These are the same as the most influential attributes! Based on whether respondents scored these three statements highly or not we can determine whether they are good potential customers. **Find Dependencies** liberal excluded customers who scored the three attributes with a low number.

Stepwise **Linear Regression** can show the linear relationships, if one exists, between the attributes and the target *Attribute* attitude. This exploration will let us know what attributes are influential and by how much.

TOP▲

Step 4: Using Stepwise Linear Regression to find Significant Attributes

PolyAnalyst's **Stepwise Linear Regression** includes only significantly independent attributes, rather than blindly calculating the regression coefficients for all existing attributes. The user can set a critical F-Ratio minimum value to discard corresponding regression terms below that value. Also, the user can set the maximum percentage of missing value.

1. Right click on the **World** dataset and select **Explore** | **Linear Regression…(LR)**.

2. Name the report *LR_Attitude*.

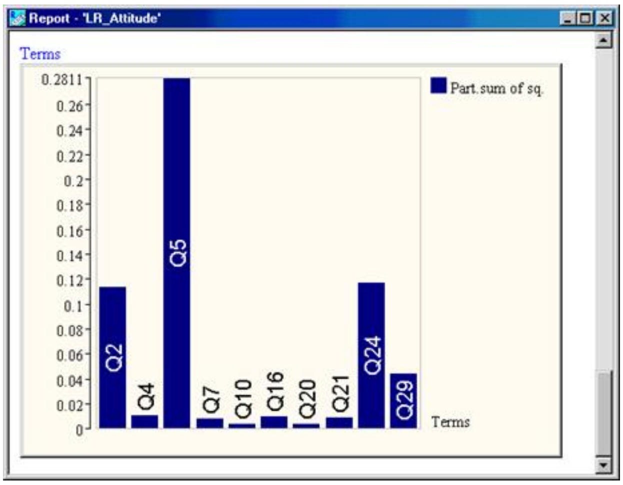
3. Set the *Attitude* attribute as the dependent variable by double clicking on it so that it is highlighted in red.

4. Deselect the *Buyers* attribute.

5. Click **OK**. The report opens in the **Results** window.

Analyzing the Results: The standard error of 0.6413 is not too high nor is the R-squared 0.5887, but this is quite typical of the accuracy of real world data. The high significance is an indicator that the found relation reflects actual dependencies in the data rather than simply fitting the data using a large number of free parameters.

When looking at the **Terms** table, we see the attributes used in the regression equation. All of the attributes have positive coefficients indicating that there is a positive correlation between the independent variable and the dependent variable. The coefficients column also indicates which independent attributes have the most influence on predicting the dependent variable. The larger the coefficient the attribute has, the more effect it has the dependent variable. For this model, Attribute *Q5* has the highest coefficient and is therefore the most influential attribute in predicting the dependent attribute *Attitude*.



The Terms histogram provides a visual representation of the independent attributes and their coefficient weight. We can see from the Terms histogram that four independent attributes have a major influence on the dependent attribute: *Q2, Q5, Q24*, and *Q29*. The large F-Ratios from the terms chart informs us that the four attributes are highly independent of each other.

Here is the four independent attributes listed in order of significance.

Q5 – Life is too short not to take some gambles.

Q24 – I would like to take a trip around the world.

Q2 – When I must choose between the two, I dress for fashion, not comfort.

Q29 – I can do anything I set my mind to.

The people that indicated a high attitude toward the above four attributes are inclined to purchase a Land Rover Discovery SUV. Land Rover's best potential customer is a risk-taker, likes to travel, likes fashion, and is a confident person. Notice that **Stepwise Linear Regression** did not pick out *Q28* as being a significant attribute like **Find Dependencies**, but did pick up a strong linear correlation with *Q2*. **Linear Regression** detects linear relationships only, where as Find Dependencies is both linear and non-linear relationships. *PolyAnalyst* provides the ability to explore data from several angles with several different data-mining exploration engines.

Next we will use **Find Laws** to develop an empirical model that is nonlinear and multidimensional. The exploration engine will also perform rigorous checks of statistical significance of this model to ensure that the model is not over fitting. This will help us find the explicit form of the relationship between *Q28* (among other attributes as well) and *Attitude*.

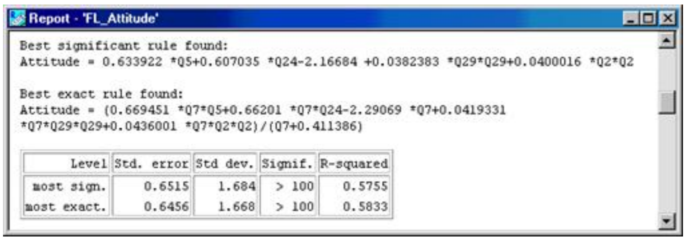
TOP▲

Step 5: Using the Find Laws Exploration Engine

1. Right click on the ***World*** dataset and select **Explore** | **Find Laws…(FL)**.
2. Set *Attitude* as the target attribute.
3. Deselect the *Buyers* attribute.
4. Set the time limit to 5 minutes.
5. Click **OK**.

During the process stage, the report will continuously be updated with better rules. The user sees a live picture of the **Find Laws** exploration engine evolving rules that gradually become more and more accurate. Due to the complexity and depth of the algorithm **Find Laws** requires a slightly longer time to run than some of the other algorithms. Depending on the size of the data and the hardware characteristics, a real exploration with **Find Laws** can take anywhere from several minutes to several hours.

Analyzing the Results: The text report contains two rules. The first rule passed all the tests for statistical significance of the whole rule expression (global significance), as well as for all separate terms entering this expression (local significance). This rule is called the **best significant rule**. The second rule is a next possible improvement from the best significant rule, which was still running when the exploration engine was timed-out by the user, which is called the **best exact rule**. This rule usually has a better accuracy, however, it has not passed important statistical significance tests yet and can be insignificant as a whole or contain some terms whose presence is not statistically justified. This rule also is usually more complicated than the best significant rule. Very few rules survive to be transferred from the "**best exact**" category to the "**best significant**" category. In the majority of practical cases the user should use the first, **best significant rule** as a predictive model for the investigated system.



```
Best significant rule found:
Attitude = 0.633922 *Q5+0.607035 *Q24-2.16684 +0.0382383 *Q29*Q29+0.0400016 *Q2*Q2

Best exact rule found:
Attitude = (0.669451 *Q7*Q5+0.66201 *Q7*Q24-2.29069 *Q7+0.0419331
*Q7*Q29*Q29+0.0436001 *Q7*Q2*Q2)/(Q7+0.411386)
```

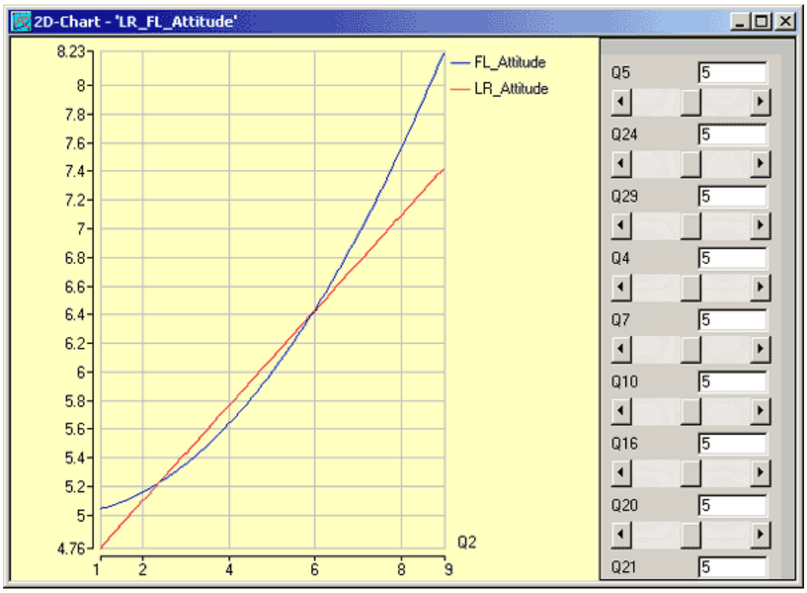
Level	Std. error	Std dev.	Signif.	R-squared
most sign.	0.6515	1.684	> 100	0.5755
most exact.	0.6456	1.668	> 100	0.5833

The table contains statistical characteristics of the built models. We can see that statistical significance of both developed models is very high (>100). As a measure of statistical significance a special index calculated on the basis of randomized testing is used here. A rule of thumb is that if the significance index is larger than 3, the developed model is quite trustworthy and does not get its predictive power as a result of over fitting the training data or statistical fluctuations in the data.

We see that the **FL** engine selected as the most influential attributes the same attributes – *Q2, Q5, Q24*, and *Q29* – as did **Linear Regression**. If one recalls that all the exploration engines we have used to process the data – **Decision Tree** Analysis, **Linear Regression**, **Find Dependencies**, and **Find Laws** – are based on completely different data analysis algorithms, then the evidence that the mentioned attributes above are the ones to pay close attention too.

It is interesting to note that the squares of the attributes *Q2* and *Q29* enter the **Find Laws** generated equation. These attributes acquire larger relative importance as their values increase. At the same time we can see that the *Attitude* values predicted by the model are very close to that predicted by the linear regression equation. In order to see this better let us visualize the two obtained predictions on a graph.

1. From the main file menu, select **Create Object** | **Create 2-D Chart…**
2. Name the chart *LR_FL_Attitude*.
3. Set *Q2* as the attribute on the x-axis. We wish to compare a non-linear attribute.
4. Click **Add Rule Graph**. Select *LR_Attitude* and *FL_Attitude*. Click **OK**.
5. Click **OK** to create the chart.



Q5 5
Q24 5
Q29 5
Q4 5
Q7 5
Q10 5
Q16 5
Q20 5
Q21 5

Find Laws produced a more accurate expression to model the correlation. This is not always the case, but is a good example of how **Find Laws** can detect attributes that are non-linear. To see how the **exact rule** differs from the **most significant** rule, right click and edit the chart to show *FL_Attitude_EX* instead of *FL_Attitude*. Interact with the graph by moving the slider bars to see how certain attribute values affect the lines.

TOP▲

Step 6: Using Discriminate Analysis for Prediction

As a final step of our exploration let us utilize *PolyAnalyst's* **Discriminate** exploration engine, which classify records in different classes. This engine develops a rule that allows assigning a record to one or another class and can be driven by the **Find Laws**, **Linear Regression**, or **PolyNet Predictor** exploration engines.

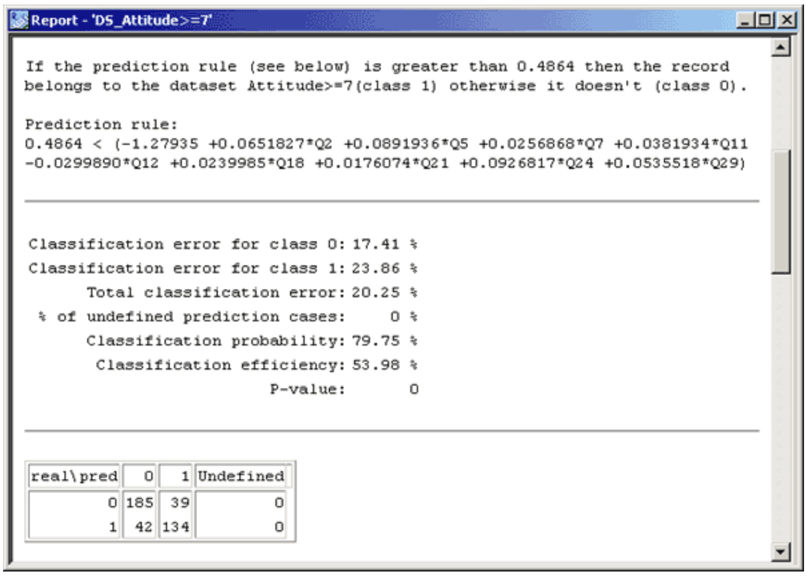
To start our classification analysis we first separate records with an *Attitude* value of 7 or higher.

1. Right click on the **World** dataset and select **Split** | **To equal intervals**…
2. Set the **Step** equal to 6 and the **From** value at its default value of 0.
3. Click **OK**. Two new datasets appear: ***Attitude_1***, which contains records where *Attitude* is below 7, and **Attitude_2** with records where *Attitude* is 7 or higher.
4. Right click on the dataset **Attitude_1** and select **Delete**. A confirmation message box appears. Click **Yes**. We will not need this dataset for future analysis because we are only interested in records that contain potential buyers.
5. Rename the dataset ***Attitude_2*** to ***Attitude>=7*** to make it easily recognizable.
6. Right click on ***Attitude>=7*** and select **Edit**. Right click on the *Attitude* and *Buyers* attributes to exclude it from the dataset. Then Click **OK**

   Now the **Discriminate** engine will show us what combination of independent attributes can be used for classifying an arbitrary record to the investigated dataset. We will use **linear regression** as our chosen process because it works well with the data.

7. Right click on the ***Attitude>=7*** dataset and select **Explore** | **Discriminate…(DS)**.
8. Select **Linear Regression** as the process type. Then, click **OK**. In about one minute the exploration terminates and we can analyze the finished report.

   The report contains a formula modeling the belonging function, a threshold for classification, and some numerical characteristics of the classification. The predicting formula is linear in the present case because we have selected the **Linear Regression** algorithm for its creation.

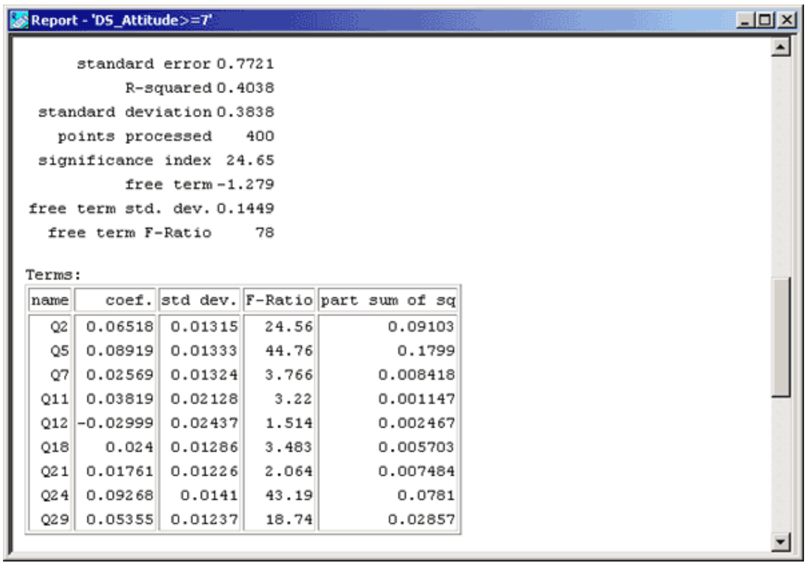


```
If the prediction rule (see below) is greater than 0.4864 then the record
belongs to the dataset Attitude>=7(class 1) otherwise it doesn't (class 0).

Prediction rule:
0.4864 < (-1.27935 +0.0651827*Q2 +0.0891936*Q5 +0.0256868*Q7 +0.0381934*Q11
-0.0299890*Q12 +0.0239985*Q18 +0.0176074*Q21 +0.0926817*Q24 +0.0535518*Q29)

 Classification error for class 0: 17.41 %
 Classification error for class 1: 23.86 %
       Total classification error: 20.25 %
  % of undefined prediction cases:     0 %
       Classification probability: 79.75 %
        Classification efficiency: 53.98 %
                          P-value:       0
```

real\pred	0	1	Undefined
0	185	39	0
1	42	134	0



```
      standard error 0.7721
           R-squared 0.4038
  standard deviation 0.3838
    points processed    400
  significance index  24.65
           free term -1.279
free term std. dev. 0.1449
   free term F-Ratio     78
```

Terms:

name	coef.	std dev.	F-Ratio	part sum of sq
Q2	0.06518	0.01315	24.56	0.09103
Q5	0.08919	0.01333	44.76	0.1799
Q7	0.02569	0.01324	3.766	0.008418
Q11	0.03819	0.02128	3.22	0.001147
Q12	-0.02999	0.02437	1.514	0.002467
Q18	0.024	0.01286	3.483	0.005703
Q21	0.01761	0.01226	2.064	0.007484
Q24	0.09268	0.0141	43.19	0.0781
Q29	0.05355	0.01237	18.74	0.02857

Analyzing the results: The prediction is that a record belongs to the dataset *Attitude >= 7* if and only if the predicting formula produces a value larger than 0.4864 when one plugs in all the values for the independent attributes involved. In practice, this action is performed automatically. The total classification error is 20.25% -- that indicates a correct classification in 79.75% of cases. At the same time, we see that the classification errors for the two classes differ one from another. We can accurately predict which people are not going to buy the Discovery. The classification error for this class is 17.41%. This result constitutes a very important practical achievement since it allows us to reduce the amount of correspondence if we were to perform a direct mailing marketing campaign.

We should note a high statistical significance of this result. Thus, the practical recommendation that we have arrived at is very trustworthy and used to create a very efficient marketing campaign.

Congratulations! You have successfully completed the Customer Survey Analysis Tutorial.

TOP▲

Previous Lesson | Home | Next Lesson

Contact us at info@megaputer.com • Phone: +1.812.330.0110