

Chapter 2

A Day in Your Life as a Data Miner

In This Chapter

- ▶ Participating in a data-mining team
- ▶ Focusing on a business goal
- ▶ Framing your work with an industry-standard process
- ▶ Comparing data with expectations

Good morning! Welcome to an ordinary day in your data-mining career.

Today, you will meet with other members of the data-mining team to discuss a project that is already under way. A subject matter expert will help you understand the project's business goals, and explain why they are important to your organization, to make sure that everyone is working toward the same end. Another member of the team has already begun gathering data and preparing it for exploration and modeling. (You're lucky to have a strong team!)

After the meeting, you'll begin working with the data hands-on. You'll get familiar with the data. Although some of the data preparation work has been done, you will still have more data preparation to do before you can start building predictive models. Data miners spend a lot of time on data preparation!

Later today, you'll begin exploring the data. Perhaps you'll begin to build a model that you'll continue to refine and improve in the days to come. And of course, you'll document all your work as you go.

It's just another day in the life of a data miner. This chapter shows you how it's done.

Starting Your Day Off Right

You've had a good night's sleep, and now you wake up early for a little exercise and a good breakfast. This has little to do with data mining, but it is a nice way to start your day.

On your way to work, ponder this: Successful data mining is a team effort. No one person possesses all the knowledge, all the resources, or all the authority required to carry out a typical data-mining project and put the results into action. You need the whole team to get things done. Your coworkers may be charming people with the best of skills and the purest of motivations, or they may have challenging personalities and hidden agendas, but you vow to start your data-mining day right by setting out to treat each person with patience, to listen to everyone with respect, and to explain yourself plainly in terms that other team members can understand.

Meeting the team

Today you'll be meeting with your team: Virginia, your resource for business expertise, and Matt, your data sourcing and programming expert. They are charming people with the best of skills and the purest of motivations.

Virginia will act as the client liaison and explain your organization's business goals. She'll explain the business problem and its impact on the organization. She can point out factors that are likely to be important. And she can answer most of your questions about the workings of the business, or help you reach someone who can.

Matt is very familiar with the data that you'll be using. He has prepared datasets for you to use, derived from public sources and further developed with a few calculations of his own. This simplifies your work and saves you a lot of time. He'll be the person you rely on for information about data sources, documentation, and the details of how and why he has restructured the data.

Virginia and Matt rely on you, too. Matt needs your input to understand what data is most useful for data mining and how to organize data for your use. He needs you to point out any errors (or suspected errors) in the data so that he can investigate and address any problems. Others are depending on the information he provides — not just you — so don't let errors linger! Virginia needs your input about what kinds of analysis you can provide, clear information about your results, and good documentation of your work.

Exploring with aim

Saying that data miners explore data in search of valuable patterns may create a mental image that's a bit magical or mysterious. You're about to replace that image with one that is far more down to earth and approachable. Data mining isn't magical, and its purpose is to eliminate mystery, a little bit at a time, from your business.



You might explore a shopping mall or a quaint little town just for the experience of looking around, but when you're data mining, you're exploring with a specific purpose. The very first thing you'll do in any data-mining project will be to get a clear understanding of that purpose. As you work with data, you will frequently revisit your goals and give thought to whether and how the information you find within the data supports them.

You'll be faced with temptation now and then, temptation to spend time examining some pattern in the data that is not immediately relevant to the goals at hand. As with other temptations, you may be free to indulge a little bit, if you have some time and resources to spare, but your first priority must always be to address the business goals established at the start of the project.

Introducing the real people on your project team

The project described in this chapter is real in every way. It addresses a real business issue that impacts people and businesses in a real community. The data is real. And the people on your team, Virginia and Matt, are also real.

Virginia Carlson is a data strategist. She is principal researcher for data integration at Impact Planning Council (www.impactinc.org/impact-planning-council), a Milwaukee, Wisconsin, based organization devoted to improving lives of community members, and associate professor at University of Wisconsin, Milwaukee. She's an expert in the collection and use of data to support social sector initiatives. She's led significant economic research organizations and projects, and she's the coauthor of *Civic Apps Competition Handbook, A Guide to Planning, Organizing, and Troubleshooting* (published by O'Reilly Media) (<http://shop.oreilly.com/product/0636920024484.do>).

Matt Schumwinger is an independent data analyst. He's the owner of Big Lake Data (<http://biglakedata.com>), a services firm that helps its clients to visualize, analyze, and present quantitative information. Matt

studied labor economics and labor relations at Cornell University, and has devoted much of his career to improving the well-being of Americans by organizing low-wage workers across the United States.

Virginia and Matt share common interests in improving the lives of public citizens and using data to support communities. In that context, they have worked together as a team, bringing together their complementary talents and experiences to work toward common goals.

Your project is an extension of Virginia's and Matt's real work. The example builds on projects that they have done in the past to create something entirely new. As members of your team, they provide expertise in community development and data management. Each of them is capable of data mining, but they have their own jobs to do! Besides, you know things they don't know and have skills they don't have. They need you to bring your own special mix of knowledge and experience to the team, and enrich everyone's knowledge. Together with Virginia and Matt, you can make discoveries that will help build stronger communities.

Structuring time with the right process

Many a would-be data miner has downloaded and installed software, started it up, and wondered, “Now what?” That won’t happen to you today.

You’ll know how to use your time, because you will take advantage of ground-work that data miners from hundreds of organizations have done for you when they developed and published a model process for data mining. The *Cross-Industry Standard Process for Data Mining (CRISP-DM)*, an open standard, provides you with guidelines for organizing and documenting your work. It’s a six-phase process that begins with defining business goals and ends with integrating your results into routine business and reviewing your work for next steps and opportunities for improvement.

Chapter 5 explains the CRISP-DM process in detail. There you will see that each of the six phases calls for several defined tasks, and that each task has one or more deliverables, which may be reports, presentations, data, or models. In this chapter, you won’t see every one of those details, but you will touch on each of the six major phases in the CRISP-DM process.

Understanding Your Business Goals

Virginia explains the data-mining team’s latest project: helping a local planning council. Its mission is to promote economic well-being by encouraging land use that makes the community attractive to businesses and residents. A key part of its work is retaining and attracting businesses that employ local residents and offer good compensation.

Your team’s role is to provide new and relevant information, grounded in data and analysis, that the planning council can use to decide where to focus efforts to make the most of its resources. Virginia and Matt have already been involved in projects supporting these aims. In earlier projects, they’ve produced analyses of factors that impact land use and shared information through consultations and presentations, written reports, and interactive maps.

The council understands that the best opportunity to influence the use of a particular parcel of land comes when the land is about to change ownership. But land owners aren’t going to just drop in and announce their intentions to sell. Many significant real estate transactions are arranged quietly, so the council might not know a thing about the opportunity until after the property has been sold.

So, the council’s business goal is to identify parcels of land that are about to change ownership, and to do so early enough to influence the use of the land.

How will the council decide whether it is successful in meeting that goal? At this stage, the council has only informal (and not entirely consistent) ways of predicting which parcels of land are about to change hands. The stated success criteria simply call for establishing a process to make change-of-ownership predictions in a consistent way. (Future projects will build on this goal and have quantitative success criteria.)



When you're presented with a goal, always discuss and document success criteria from the start. Although you may only be responsible for a narrow part of the work needed to achieve the business goal, understanding how the ultimate results will be evaluated helps you to understand the best ways to contribute to the project's success.

These success criteria may sound simple, but you have doubts. You ask questions like these:

- ✓ **Does the council expect that just one model will work for all types of property?** Industrial, commercial, single-family, multifamily, and so on — it's not realistic to think that you'll find one big equation to address them all.
- ✓ **How many property types exist?** You could have dozens.
- ✓ **Is the council equally interested in all properties?** You'd think large, industrial parcels would be the most important.
- ✓ **Which property types are most important to the council?** You may want to push for modeling just one or two important categories on the first round.



Always ask about recent mishaps. Unspoken goals often include not repeating something that just went wrong.

Asking questions helps you to get more information, of course, but your questions do more than that. They help others on the team (including executives, if you have the opportunity to meet with them) become aware of what's missing, what's going to be challenging, and what's a lot more complicated than they thought it would be! By asking probing questions in the business-understanding phase, you help everyone to clarify thinking, define reasonable goals, and set realistic expectations.

After some discussion, it's agreed (and documented!) that the business goal for this project will be to demonstrate the feasibility of modeling to predict land ownership change — a narrower and less grand goal than the one initially suggested. You're not expected to create a megamodel (no, that's not a technical term) that covers all types of property. If the council finds that even one factor has predictive value for property transfers, that will be satisfactory for the first round. No quantitative criteria will be stated for model performance on this first investigation. The object is just to demonstrate that potential exists to develop a useful model to predict property ownership changes using the available data.



Business goals are determined by the client (external or internal), not the data miner. If you and your team have doubts about a particular goal, don't change it on your own. Clients won't accept that! Instead, enter into a discussion with the client, explain your concerns, and come to an agreement about reasonable business goals for the project.

Based on the business goals, you define data-mining goals. Because the business goal is to demonstrate the feasibility of modeling to predict land ownership change, you will set a data-mining goal of creating a rudimentary predictive model for change of property ownership. Because you have no specific numbers about the performance of the current, informal approach to predicting ownership changes, you'll simply aim to demonstrate that at least one variable has measurable value for prediction. (As with the business goals, future projects will build on this, and you'll set more specific quantitative success criteria at that stage.)

You'll complete this phase of the data-mining process by outlining your step-by-step action plan for completing the work (including a schedule and details of resources required for each step) and your initial assessment of the appropriate tools and techniques for the project.

Understanding Your Data

In the data-understanding phase, you will first gather and broadly describe your data. You won't have to start from scratch to gather data, because Matt has already assembled several datasets for you to use. He's drawn from data used in earlier projects and derived some additional fields that you will need. Then you'll examine the data in a little more depth, exploring the data one variable (field) at a time, checking for consistency with expectations and any obvious signs of data quality problems.

You begin to review the data, making notes for your report as you work.

Describing data

The data is in several text files, each in comma-separated value (.csv) format. The files are somewhat large, 50–100MB, but not too large to handle with the computer and software that you have available. You note the name and size of each file.

Your first concern is to identify the variables in each file and confirm that you have adequate documentation for each of them. Several of the files contain historic public property records; a lengthy document defines those variables. You've also been given notes explaining how derived variables were created. You review each variable in the data, comparing the variable names to the information in the documentation.

You note findings about the data and the documentation, including the following:

- ✓ Most of the fields appear consistent with the documentation that you have.
- ✓ Some of the fields in the property record data files are not explained in the documentation.
- ✓ Some of the fields described in the property record documentation don't appear in the data.
- ✓ One of the property record data files contains many more fields than the others, and those fields are not explained in the documentation.

You write detailed notes about each file and each variable. Using your notes as a reference, you look for information to address the discrepancies. You find that

- ✓ A few of the fields in the data from public sources simply don't match the documentation provided (public data isn't always perfect data).
- ✓ Additional notes are available to explain how some of the derived fields were created.
- ✓ Some of the undocumented data was obtained by *web scraping* (using specialized software to automatically extract information from websites), and you can't find any dependable documentation for it.

You update your notes about the data, revising them with additional documentation. You note which variables are still undocumented. Although some of those fields seem likely to have predictive value for modeling property ownership changes (such as foreclosures), a number of disadvantages exist to using them for predictive modeling, including the following:

- ✓ Some of the data was collected by web scraping. You're not confident that you'll be able to get that data in the future.
- ✓ You don't have details on the scraping process, so you can't be sure that scraped data was defined consistently.
- ✓ You'll have a heck of a time explaining the meaning of data without documentation.

So you decide that on the first attempt to develop a predictive model for property ownership change, you'll use only those fields that have been adequately documented. In a future project, you may seek out alternative sources for some of the other fields.

Exploring data

Now it's time to briefly examine the data for each variable in each file. You must check basics, such as whether the data is string or numeric, that the range of values is appropriate, and that the distribution of values looks reasonable. You'll note any discrepancies from the documentation and your own reasonable expectations.



The procedures you'll use to generate diagnostic information about your data vary with the kind of data that you have, the tools available, and the way that you like to work. You may use highly automated functions or you may work with variables in small groups or one at a time. You'll almost always have a choice of ways to go about it.

For each field, you prepare a brief summary, with a name and description, number of missing cases, and the range of values (low and high). You may also include additional information such as a distribution graph, the average (mean), and most frequently occurring (mode) value of the variable. At this point, you won't try to relate one variable to another.

You start by using software that produces a basic report for each variable in the data, including information such as the range of values, the average for continuous variables, the most common value for categorical variables, and so on (shown in Figure 2-1). This report is a starting point for understanding your data. You use it to identify what data you have and whether the data is consistent with what you were led to expect by the documentation and your colleagues. You add to it by using graphs or other simple methods for adding detail to your understanding of each variable.

Name	Type	Miss.	S...	Filter (76 / 76 attributes):	Filter
TAXKEY	Integer	30			
BI_VIOL	Binominal	0	Least XXXX (162403)	Most XXXX (162403)	
DIV_ORG	Integer	0	Min 0	Max 999	Average 43.951
YR_ASSMT	Integer	0	Min 2010	Max 2010	Average 2010
SUB_ACCT	Integer	0	Min 0	Max 0	Average 0
P_A_LAND	Integer	0	Min 0	Max 19029000	Average 25320.01
NR_UNITS	Integer	0	Min 0	Max 718	Average 1.610
C_A_LAND	Integer	0	Min 0	Max 19029000	Average 25478.03
DIV_DROP	Integer	0	Min 0	Max 0	Average 0
			Min	Max	Average
Showing attributes: 1 - 76 Examples: 162,.... Special Attributes: 0 Regular Attributes: 76					

Figure 2-1:
Variable
summaries.

As you review each variable, you describe it and make note of any concerns and what should be done to address them. In your summaries, you state whether the variable appears ready for use in modeling, needs further preparation, or is in such poor condition that it cannot be used.

Your individual variable summaries read like the examples shown in Table 2-1.

Table 2-1 Dataset: 2010 Property Data	
<i>Variable Name</i>	<i>Description</i>
BI_VIOL	Description: Unknown (No documentation available for this variable) Variable type: String Range: XXXX to XXXX Number of missing cases: 0 Assessment: Not acceptable for modeling. All cases have the same value. Reason unknown. Next steps: Will not use in this project.
TAXKEY	Description: Ten-digit property ID code number Variable type: Identifier (string) Range: 9999000–7369999110 Number of missing cases: 30 Assessment: A small number of cases are missing. Some cases have fewer than ten digits, possibly due to trimming of leading zeros because the variable format was read as integer, rather than string. Next steps: Must clean this variable as well as possible, as it is the unique ID for each property. Change variable type from integer to string. Reassess.
C_A_CLASS	Description: Assessment class code — defines property use. Detailed explanations of codes in Appendix A. Variable type: Nominal Range: 1–9 Number of missing cases: 0 Assessment: Distribution (Figure 2-2) looks appropriate with Class 1 (residential) the most frequently occurring category. No obvious signs of quality issues. Next steps: This field appears ready for modeling use.
DIV_ORG	Description: A control number used by the assessor's office Variable type: String Range: 0–999 Number of missing cases: 0 Assessment: This is used for administration within the assessor's office and does not appear to have any value for modeling purposes. Next steps: None.

Some of these variables won't be useful for modeling. For example, BI_VIOL sounds like it might represent the number or kind of building inspection violations reported for a property. Maybe it was used for that purpose at some time, but in this dataset, every case has the same value, "XXXX." The field is not mentioned in any of the documentation that you have. Building violations could be valuable information for predicting property transfers, but you may have to wait for a future project when you have time to track down another source for that information.

Fortunately, some fields are in much better condition. For example, C_A_CLASS, the assessment class code, identifies the use of the property in major classes such as residential, manufacturing, and commercial. This could be very important for modeling, because you expect different behavior patterns for different property uses. No cases are missing for C_A_CLASS, the range of values is consistent with the documentation, and a bar chart (see Figure 2-2) shows that the distribution of property uses appears reasonable, with the residential class occurring far more often than any other use.

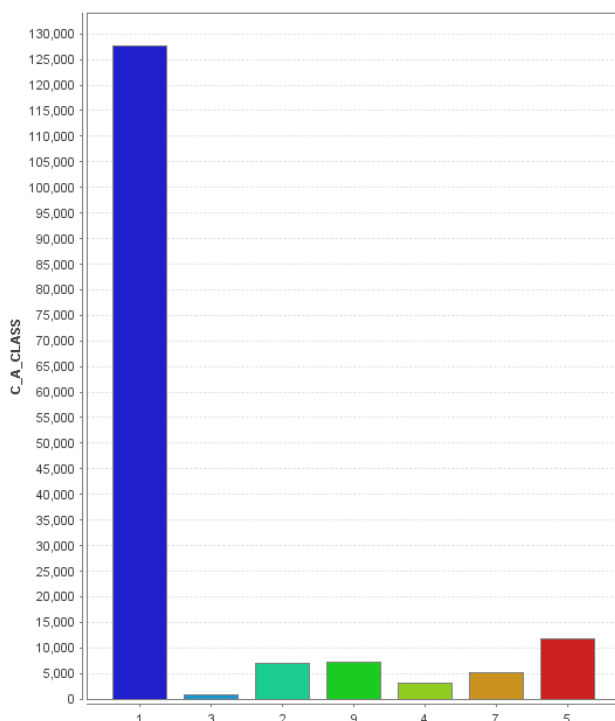


Figure 2-2:
Bar chart for
C_A_CLASS.

You notice that many of the fields that might at first appear valuable for modeling are not in a good state. Many are undocumented, some are not being maintained by the public source (and the documentation says so), and others don't vary or don't seem consistent with your expectations. You have doubts that the remaining data will be sufficient to build a useful model.

Cleaning data

You've explored the data and found that some of the fields that seem to have value for modeling exhibit minor errors or other issues that you want to correct first.

The TAXKEY field is a good example. It's the code number that identifies each individual parcel of property. Strictly speaking, an identifier is not a modeling variable, but your model won't have any value unless you can match your predictions to specific properties. You've noticed two issues in the data:

- ✓ A few cases (a fraction of a percent of the total) are missing identifier codes.
- ✓ Many cases have fewer than the ten digits that the documentation says they should.

You take a moment to ponder the missing cases (30 out of more than 160,000 total). In theory, the public agency that shared the data can fill in those blanks. But you imagine phoning the property assessor's office and explaining the problem, perhaps many times over, looking for someone who understands it and is willing to help. When you reach that person, you have no assurance that willingness to help will translate into success in correcting the flaws in the data. You think that you can do more productive things with that time and decide to live without those 30 cases.

Then some cases have fewer than ten digits in their property codes. This problem occurs frequently, but you suspect that you can find an easy fix for it. Because the code is numeric, your software has interpreted it as an integer, but a string would be more appropriate. Changing the field type to a string would prevent the software from trimming any leading zeros in the property codes. So you import the data into the software again, this time making sure to identify the field as nominal (like a name). Still, you find many cases where the values have fewer than ten digits. Your easy fix did not fix anything.

You peek at the data in a text editor (because this data is in a text format, you could use a word processor or perhaps a spreadsheet to view it) and confirm that the problem has nothing to do with trimming leading zeros. Some of the values are simply shorter than the ten digits that the documentation led you to expect. You make a note of this for your report and decide (for today) to put your faith in the data rather than the documentation.

You go through a similar process for each of the fields that you've found to be potentially useful, but not in perfect condition. As you work, you document your observations and any changes that you make. For each field, you judge whether it is good enough to use in modeling. (You're not deciding whether the variable will be in the final model or work well as a predictor, just whether it is of sufficient quality to test.) Finally, you combine your notes on these observations and actions into a data quality report.

How data miners spend their time

Cooks who serve delicious dinners spend a lot of time chopping vegetables. Runners who win races spend a lot of time stretching and training. Data miners who develop valuable predictive models spend a lot of time preparing data.

People who haven't tried data mining sometimes think that it is a nonstop thrill ride of discovering great insights and developing powerful models. It isn't. Most of your

time goes to doing all the things that must be done right before you can start building models.

Data preparation isn't the most glamorous aspect of the job. It's painstaking work, and you have a lot to do, so much that data miners spend more time preparing data than doing anything else. Yet data preparation is worth the effort, because it makes meaningful discovery possible.

Preparing Your Data

Now that you've gathered data and reviewed the fields one by one to familiarize yourself with the data and check for quality problems, you move forward and prepare the data for modeling. In this phase of your work, you do the necessary tasks to transform the data from its original form into the form you require for modeling, such as

- ✓ Joining datasets
- ✓ Specifying the roles of the fields
- ✓ Sampling the data
- ✓ Splitting your sample into subsets for building and testing models

Many projects call for deriving new fields based on the ones that are already in the data. For example, the indicator variable that you will need to identify properties that changed ownership does not exist in the public data. That must be calculated from other fields. Lucky for you, your colleague Matt has already created that variable and saved you a step on this project. But you will have to derive some other new fields for yourself.

Taking first steps with the property data

In the data-understanding phase, you identified a number of fields that you won't use for modeling. You have ruled each of them out for one of these reasons:

- ✓ **Does not make sense as a predictor:** Includes unique fields such as an address or names, or something that you believe to have no relationship to ownership change
- ✓ **Data quality is poor:** Many missing cases or incorrect values
- ✓ **Does not vary:** All cases have the same value (not necessarily a data quality problem)

You work with specialized data-mining software. Although you can do any of the same operations with other kinds of tools, your data-mining software is designed to help you view the steps of your process easily, and work quickly, by stringing together a sequence of operations represented by small icons. Each icon is a tool with a specific function and its own options and settings. This is called a *visual programming* interface.

The property data file is fairly large, so to begin, you will import the property data file, remove the variables you cannot use, and save the rest in a new (somewhat smaller) file. First, you choose a tool to read the data and place the tool on the main work area of your data-mining software, as shown in Figure 2-3. A *wizard* (a special user interface that simplifies complex tasks) helps you import the data correctly. One step from the wizard is shown in Figure 2-4. After the data has been imported, you can view it and verify that it looks right to you (see Figure 2-5).

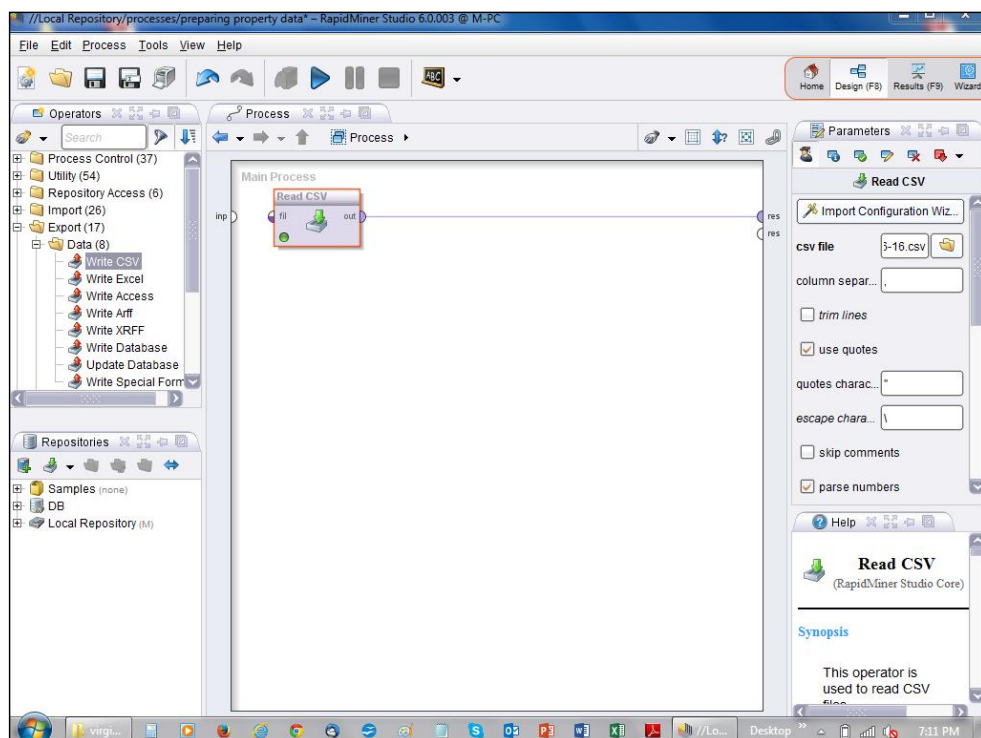


Figure 2-3:
Working
with data-
mining
software.

Data import wizard - Step 2 of 4

This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding:

☐ Trim Lines

☐ Skip Comments:

Column Separation

☒ Comma "," ☐ Space

☐ Semicolon ";" ☐ Tab

☐ Regular Expression:

Escape Character:

☒ Use Quotes:

SDIR	STREET	TAXKEY	BI_VIOL	DIV_ORG	YR_ASSMT	SUB_ACCT	P_A_LAND	NR_UNITS	C_A_LAND
W	COUNTY LIN	10001000	XXXX	8	2010	0	48200	1	48200
N	124TH	10011000	XXXX	69	2010	0	146200	0	150700
N	124TH	10021000	XXXX	21	2010	0	115000	0	115000
N	124TH	10022000	XXXX	21	2010	0	0	0	0
N	124TH	18100000	XXXX	196	2010	0	100	0	100
N	124TH	18101000	XXXX	196	2010	0	100	0	100
N	124TH	19989000	XXXX	0	2010	0	0	0	0

Row, Column Error Original value Message

Figure 2-4:
A wizard
makes com-
plex tasks
easier.

ExampleSet (162403 examples, 0 special attributes, 76 regular attributes)ter (162,403 / 162,403 examples):

Row No.	SDIR	STREET	TAXKEY	BI_VIOL	DIV_ORG	YR_ASSMT	SUB_ACCT	P_A_LAND	NR_UNIT
1	W	COUNTY LIN	10001000	XXXX	8	2010	0	48200	1
2	N	124TH	10011000	XXXX	69	2010	0	146200	0
3	N	124TH	10021000	XXXX	21	2010	0	115000	0
4	N	124TH	10022000	XXXX	21	2010	0	0	0
5	N	124TH	18100000	XXXX	196	2010	0	100	0
6	N	124TH	18101000	XXXX	196	2010	0	100	0
7	N	124TH	19989000	XXXX	0	2010	0	0	0
8	N	124TH	19990000	XXXX	0	2010	0	0	0
9	N	124TH	19991000	XXXX	0	2010	0	0	0
10	N	124TH	19992100	XXXX	389	2010	0	40600	0
11	W	COUNTY LIN	19996100	XXXX	0	2010	0	53400	6
12	W	COUNTY LIN	19996210	XXXX	69	2010	0	0	0
13	W	COUNTY LIN	19998200	XXXX	0	2010	0	47800	1
14	W	COUNTY LIN	19999100	XXXX	8	2010	0	139700	0
15	N	107TH	20032000	XXXX	78	2010	0	495700	0
16	N	107TH	20051000	XXXX	11	2010	0	204100	0
17	N	107TH	20052000	XXXX	11	2010	0	114700	0
18	N	107TH	20071100	XXXX	198	2010	0	0	0
19	W	COUNTY LIN	20072000	XXXX	88	2010	0	40600	0
20	W	COUNTY LIN	20081000	XXXX	144	2010	0	124000	0
21	W	COUNTY LIN	20082000	XXXX	144	2010	0	95900	0
22	N	107TH	29996110	XXXX	0	2010	0	549800	0

Figure 2-5:
Viewing
data after
importing.

You add another tool to the work area, shown in Figure 2-6, to select variables to keep in the data. This is not complicated to set up. The tool displays a list of variables in the data, and you select the ones that you want to keep. Figure 2-7 shows the setup. The list on the right includes all variables selected to keep.

Figure 2-6:
Adding a
tool to select
variables.

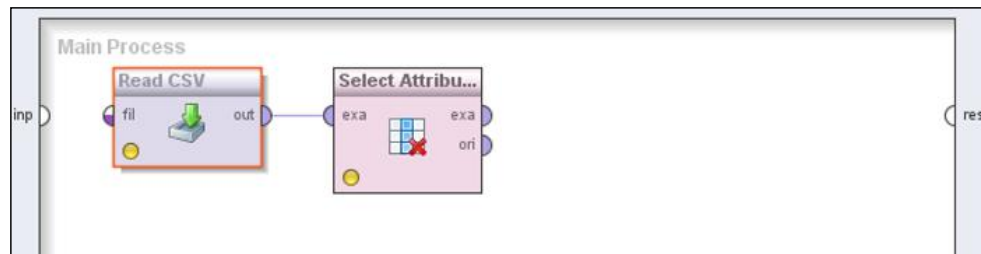
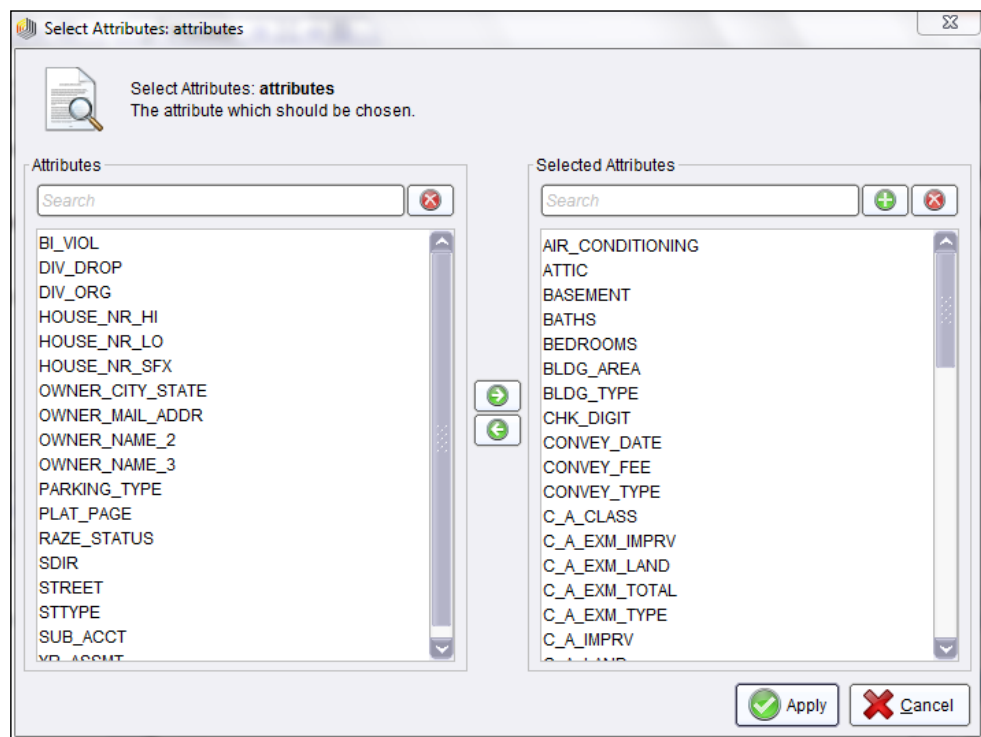


Figure 2-7:
Setting up
variable
selection.

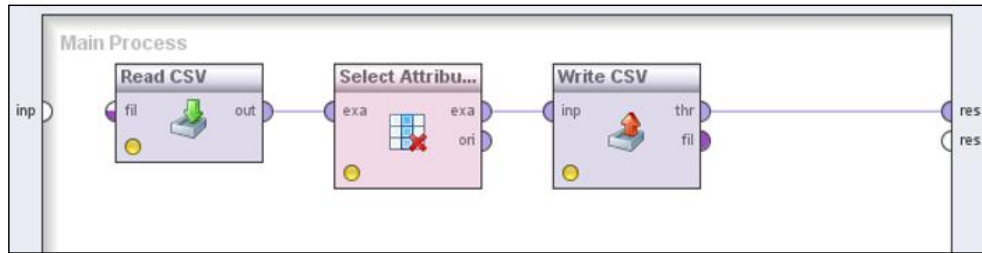


One more tool (see Figure 2-8) lets you save the variables that you have selected in a new file.



The data-mining software used in this example has many of these specialized tools. For example, it has a different tool for each of the file types that it can read and each type that it can save. Not every product takes this approach; others might have a single tool that could save your choice of several types of files.

Figure 2-8:
Saving
selected
data in a
new file.

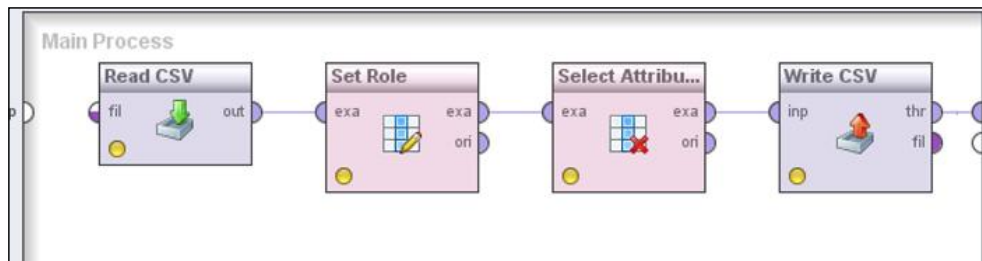


Preparing the ownership change indicator

Matt made your job easier by deriving a variable that indicates which properties have and have not changed ownership. This will be your dependent, or *target*, variable for modeling. You'll still need to do some preparation with this part of the data, particularly selecting the right settings for your data-mining software to identify the target variable.

You create a sequence similar to the one that you used in the previous section for the property data. You'll import the data, select the variable to keep, and write it to a new file. But you can see in Figure 2-9 that this time, another tool exists between data import and data selection. With it, you indicate which variable is the target by setting the tool's properties, as shown in Figure 2-10.

Figure 2-9:
Preparing
the owner-
ship change
data.



Merging the datasets

You have property data in one file and data about which properties changed ownership in another. You need to merge the two.

The process is shown in Figure 2-11. You read in each of the files that you have created earlier. For each, you indicate the name of the property identification variable by setting the properties of the appropriate tool (see Figure 2-12). The identification variable guides the merging of the two files, matching the general data for each property with the results: Did the property change ownership?

Figure 2-10:
Setting a
variable's
role.

Parameters

Set Role

attribute name: Y2change

target role: label

set additional role... Edit List (0)...

Figure 2-11:
Process
for merging
data
sources.

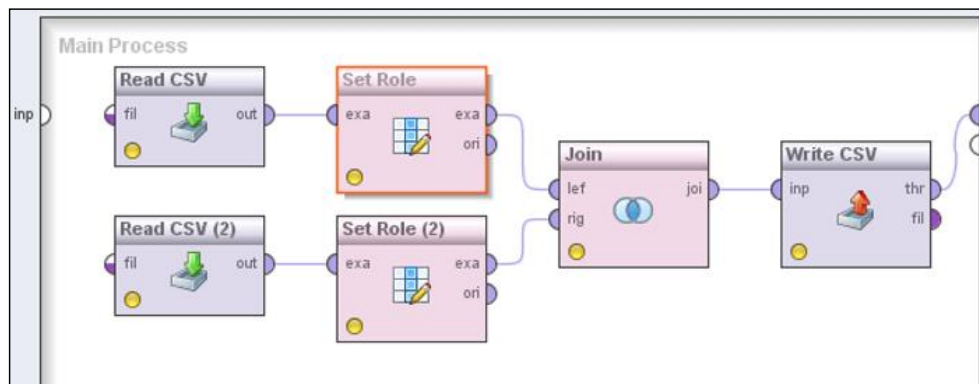


Figure 2-12:
Selecting
the iden-
tification
variable.

Parameters

Set Role

attribute name: TAXKEY

target role: id

set additional role... Edit List (0)...

Deriving new variables

Data-mining software does a lot of work for you, but you have no substitute for your business knowledge. You understand that one variable represents the price paid for a property, another the investments to improve the property, and a third, the assessed value — but the software doesn't. The software only sees numbers and categories and text, not meaning. You understand that you have something of special interest about the relationships among those three variables, apart from any others. The software doesn't. You integrate this kind of business knowledge into your analysis by using it to derive appropriate new variables for modeling.

Choosing a starting point

Virginia and Matt have told you about a number of factors that may be good indicators of imminent changes of property ownership. These suggestions are a result of investigations and interviews that they performed in earlier projects. Some of these factors are

- ✓ The owners do not live in the area.
- ✓ The owner is the local government.
- ✓ Taxes are unpaid.
- ✓ The property is vacant.
- ✓ Zoning and actual use are not matched.
- ✓ The value is not consistent with the assessment.
- ✓ Improvements are low relative to the value of the property.
- ✓ Building code violations are open.
- ✓ Many building code violations have been closed.
- ✓ Many calls for service have been closed.
- ✓ The property is marketed for lease or sale.
- ✓ The property is in foreclosure.

Although you have good reasons to believe that each of these factors is important, no one has yet confirmed their value by building and testing a predictive model. You'd like to investigate each of them — and others as well. But you don't have adequate data for some, and the others will all require effort for data preparation.

Your goal for this project is not to develop the greatest possible model, but to use the data to demonstrate that at least one variable has value for predicting changes in property ownership. The idea is to quickly provide concrete evidence that predictive modeling is feasible. In the interest of speed, you choose a couple of items from this list to try first. (If they don't work, you can come back and try others.)

You choose to look at properties with owners who do not live in the area and properties with unpaid taxes first. Your reasons are simple: You have adequate data for those variables, and the preparation required is reasonably straightforward.

Performing calculations

This part of the process (see Figure 2-13) is more elaborate than the steps you have taken so far. You'll create two new variables, select a subset of cases to use for modeling, and remove any cases that don't have adequate data to use in the modeling process.

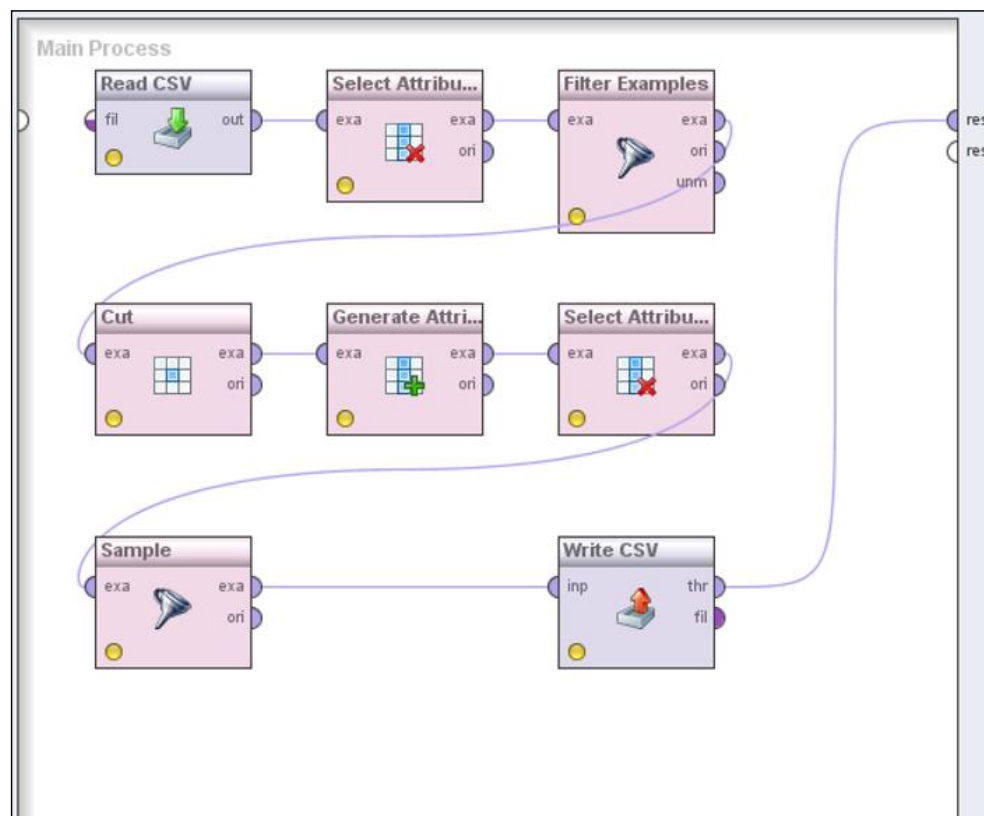


Figure 2-13:
Process
for deriv-
ing new
variables.

Before creating any new variables, you do a little housekeeping. Although you find many variables in the data, you've decided to use just a handful for your first model, so you select just those from the data (see Figure 2-14). Modeling tools, and even some data preparation tools, don't perform well, and may not function at all, if you have missing values in the data, so you filter out cases with missing values. The relevant setup for the filtering tool is shown in Figure 2-15.

Figure 2-14:
Selecting
variables to
use.

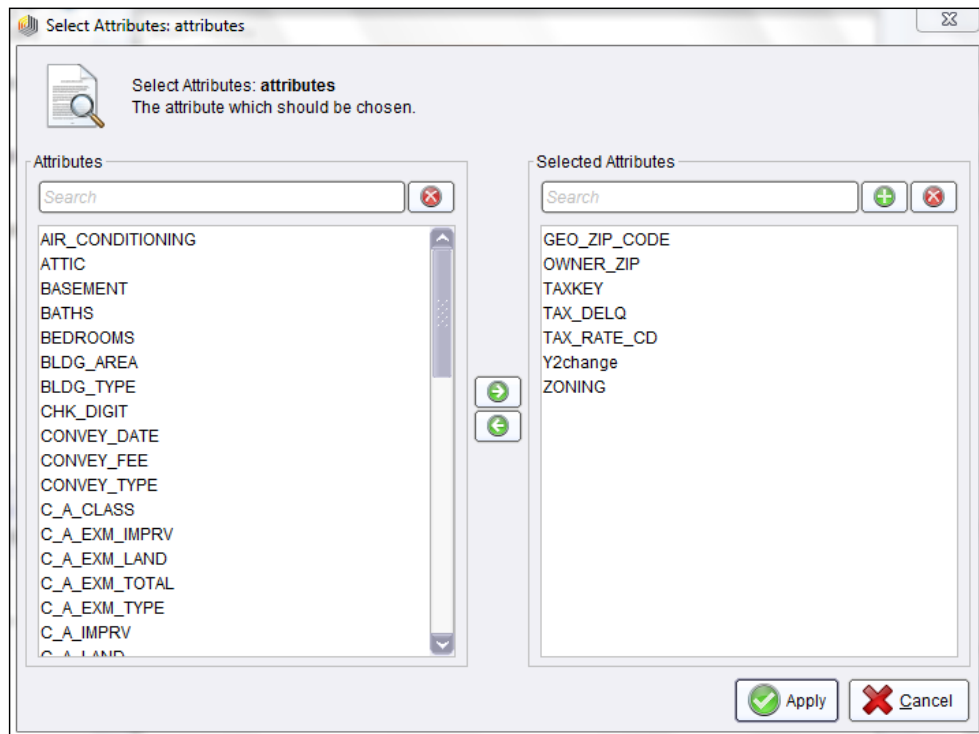
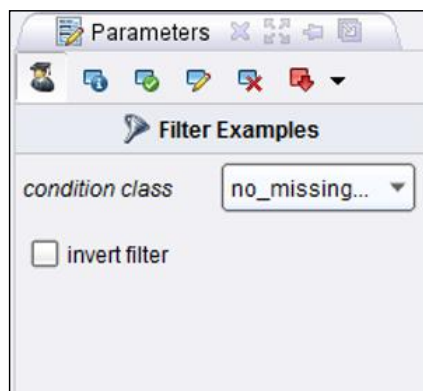


Figure 2-15:
Filtering out
cases with
incomplete
data.



To identify property owners who don't live in, or very close to, their properties, you want to compare the owner's home zip code to the zip code of the property. You have data for each of those, but some challenges exist for comparing the two. Some of the zip codes are recorded as five digits; others are in longer formats. So you'll need to get all the zip codes into a consistent format before you can create an indicator variable for properties whose owners are not local. You set the variable cut to keep the first five characters (see Figure 2-16) of the two zip code variables (see Figure 2-17).

Figure 2-16:
Cutting
variables
down to five
characters.

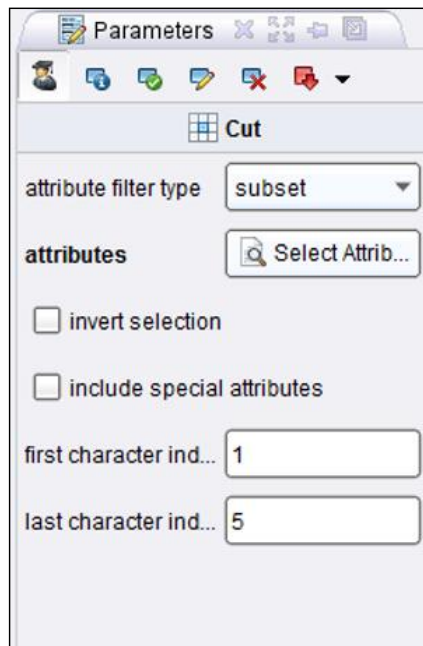
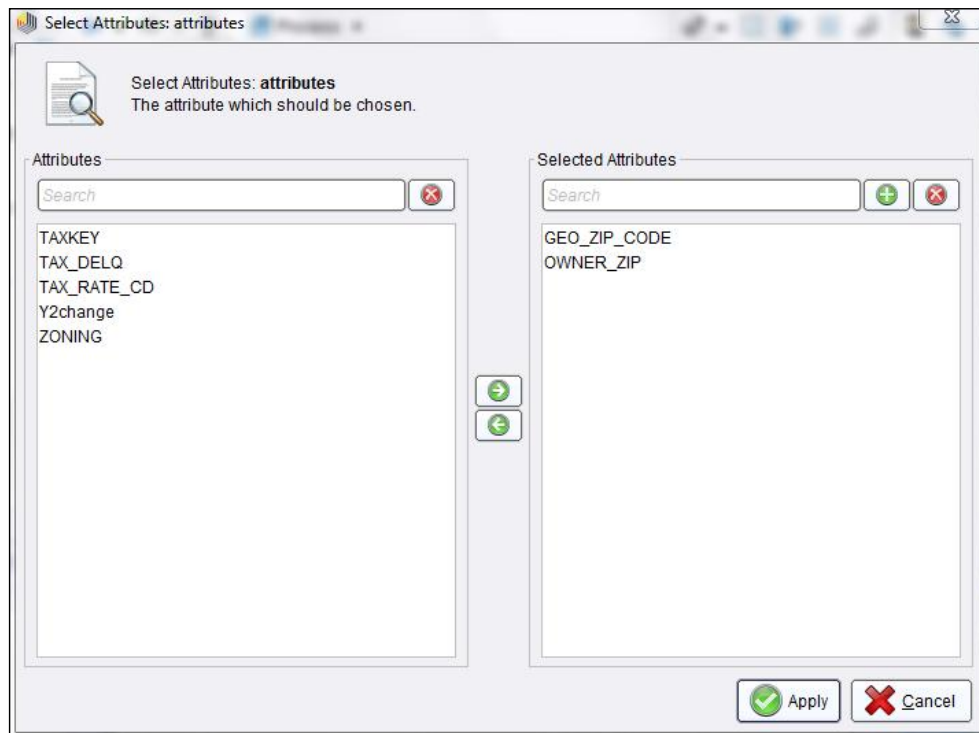


Figure 2-17:
Identifying
the zip code
variables.



A variable already exists in the property data to indicate which properties have unpaid taxes, but it's not in good form to use for modeling. That variable has a value of 1 if taxes are unpaid, but "NA" otherwise. Modeling tools don't like that! So, you'll create a nice, new variable, with a value of 1 if taxes are unpaid and 0 otherwise. The setup for creating both of the new indicator variables is shown in Figure 2-18.

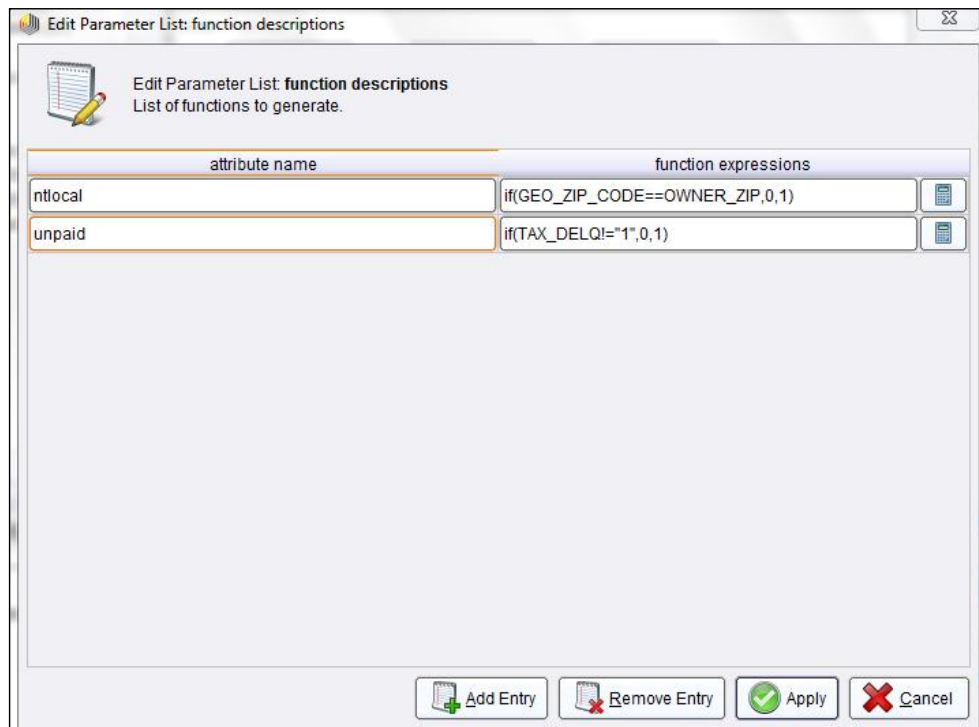


Figure 2-18:
Functions
for generat-
ing new
variables.

You have a few more steps before moving on to the modeling phase. Now that you've derived new variables, you won't need the old ones, so you use the variable selection tool (see Figure 2-19) to keep just what you need. And you'll use a data-sampling tool to *balance* the data and select a sample with roughly equal proportions of properties that did and did not change ownership. Figure 2-20 shows the setup for balancing the dataset. You request about 4,000 cases in each group, but you understand that the actual sample sizes may be a little different.

Wow, what a lot of steps! And the data preparation for this example is simpler than most. That's why the 3rd Law of Data Mining states that data preparation is more than half of every data-mining process (see Chapter 4 to read about the 9 Laws of Data Mining).

Figure 2-19:
Discarding
variables
that are
no longer
needed.

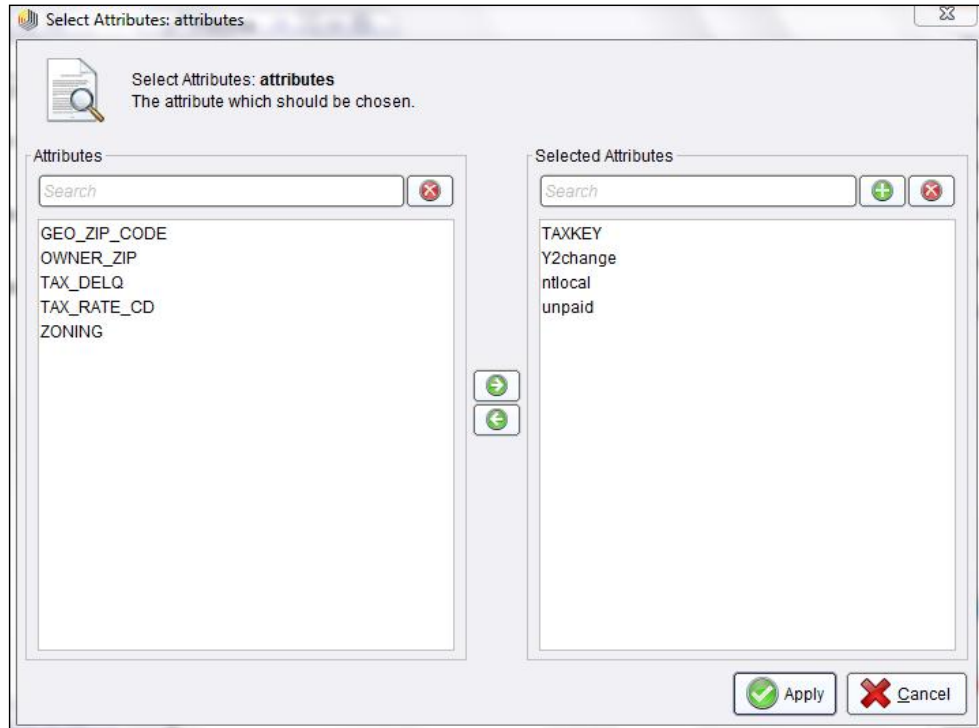
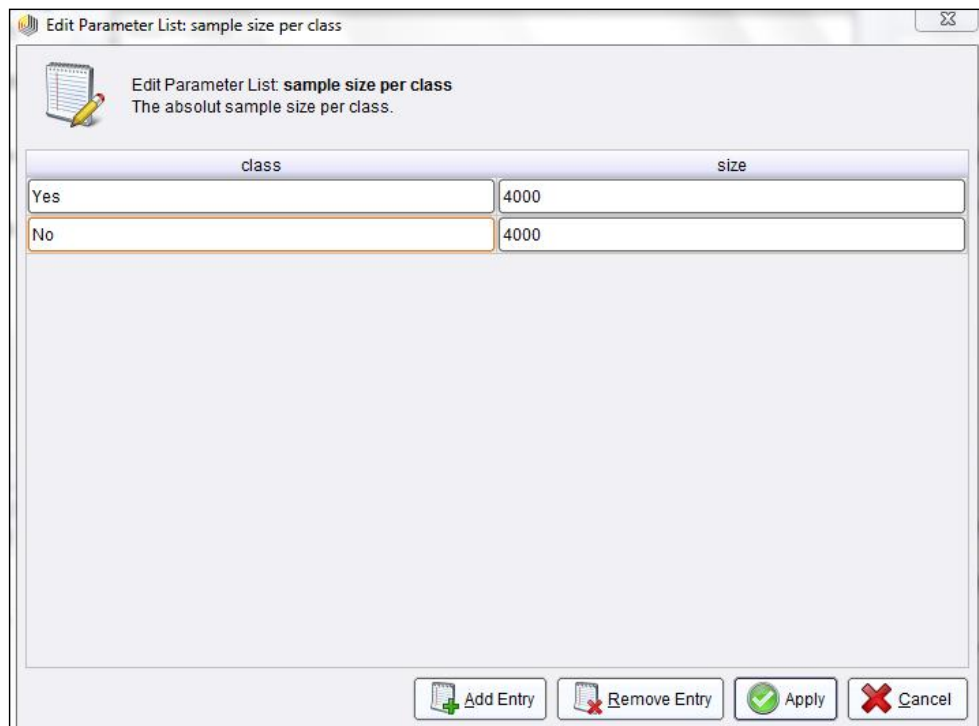


Figure 2-20:
Balancing
the data.



Modeling Your Data

Predictive models are nothing more than equations that help you make educated guesses in methodical, consistent ways, based on data.

People make predictions informally all the time, at home and at work:

- ✓ **Buying groceries:** Estimating consumption based on recent experience and anticipated changes, such as guests in the house or upcoming travel
- ✓ **Budgeting:** Planning for financial needs using information such as past spending, known upcoming events, and estimated requirements for emergency funds
- ✓ **Sales forecasting:** Anticipating future sales based on historical performance, envisioned deals, attitudes about the economy, and perhaps just a bit of wishful thinking

Because these informal predictions are made in an inconsistent, undocumented, and subjective way, it's hard to improve them. As a data miner, you create dependable fact-based predictive models, and you document the process so that you can update and improve your models in the future.

Using balanced data

In the data preparation phase, you took a special type of sample from the property data. The sample was balanced, that is, it included roughly equal numbers of cases for properties that changed ownership within a specific time frame, and properties that did not. Now that you are established as a data miner, you do this as a matter of habit.

Balancing data often seems odd or wrong to data-mining novices. It's not obvious why data miners would use data representing equal proportions of events that don't occur with equal frequency in real life. For example, in any given year, only a small fraction of properties change ownership. Why give this event representation equal to the much more common case where a property remains in the same hands? It's done because the purpose of the model is to differentiate these two events, based on patterns in the data. To construct a model that can differentiate these patterns, you need examples of each, and you give each type of pattern equal importance in modeling by giving it equal frequency in the data.

Splitting data

Some of the machine learning techniques that are widely used in data mining, such as decision trees and neural networks, call for one last bit of data preparation before constructing a model. (Data preparation goes on and on and on, doesn't it?)

Data miners can't always use theory to find one best model from data, as classical statisticians do. So data miners evaluate models by testing, testing, and testing. Some of the testing is hidden within the model-fitting process, automatic and (almost) unnoticed as you work. Some testing is done in the field through small-scale or full-scale deployment. And some of it is done by splitting off a portion of the data (called *test* or *holdout* data) before modeling and using your model to predict results for that data so that you can compare those predictions with what actually happened.

Your work process for splitting the data, building a model, and beginning the evaluation is shown in Figure 2-21. To split the data, you use a special sampling tool and specify two things: the sampling method (see Figure 2-22) and the proportions of data to use for training and testing the model (see Figure 2-23). You specify *stratified sampling*, which maintains the balance of proportions of properties that did or did not change hands in the training and testing samples. And you choose to use 70 percent of the data for training and 30 percent for testing.

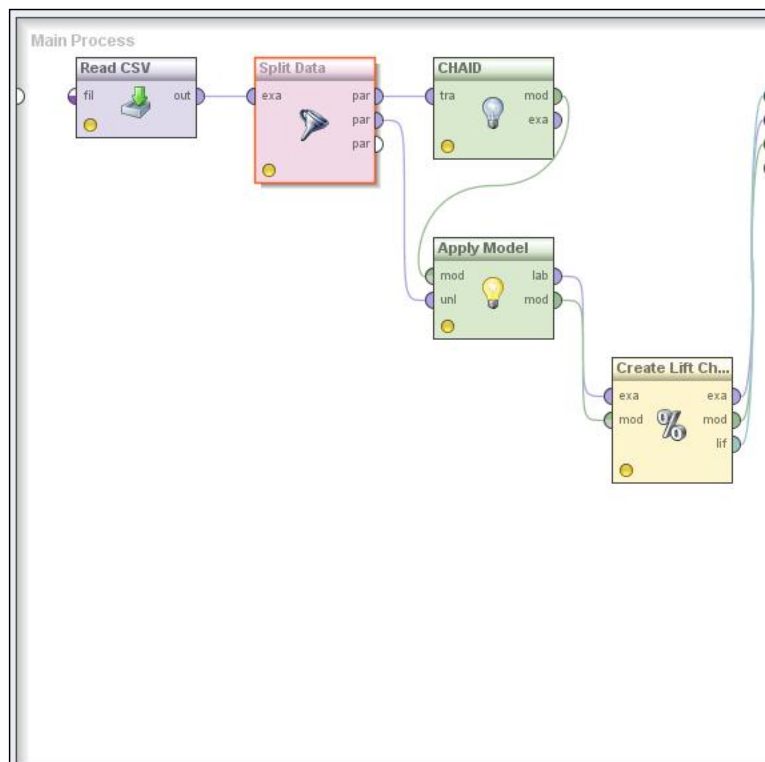


Figure 2-21:
Splitting
data, and
building and
evaluating a
model.

Figure 2-22:
Selecting
the sampling
method.

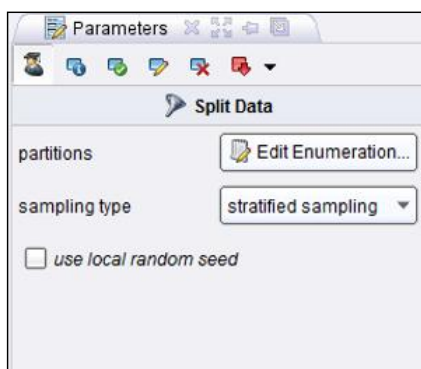


Figure 2-23:
Splitting
data into
training
and testing
subsets.



Building a model

Compared to all the work that you've invested to prepare data, creating your first model for this data takes little effort.

You have only two *predictor* variables ready to try in a model so far. One indicates whether the property owner is local (the owner address has the same zip code as the property) or not. The other indicates whether unpaid taxes exist for the property. Both are categorical variables, which narrows your choice of modeling techniques.

You choose the Chi-squared Automatic Interaction Detector (CHAID) model, a type of decision-tree model, for your first try, because it is a good fit for working with categorical variables. It's easy to use. You just add the tool (see Figure 2-24) to your process and run data through to build a model without even changing any parameters. Later, you may choose to tweak settings, but it's not necessary for your first try.

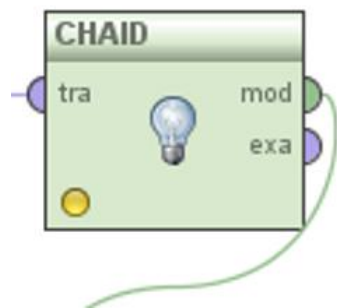


Figure 2-24:
CHAID modeling tool.

Before you run the model, you connect two tools to the data that you split earlier. The CHAID tool will use the 70 percent of the data that you've put in the training partition, and the 30 percent of the data that you set aside for testing connects to another tool. This tool will apply the CHAID model to the testing data.

Finally, you add one last element to your process. A chart will help you visualize the results of the model test. The chart tool requires a little bit of setup (see Figure 2-25). You specify the category that you're most interested in predicting. In this case, it's the "Yes" category.



Figure 2-25:
Setting up a
diagnostic
chart.

Evaluating Your Results

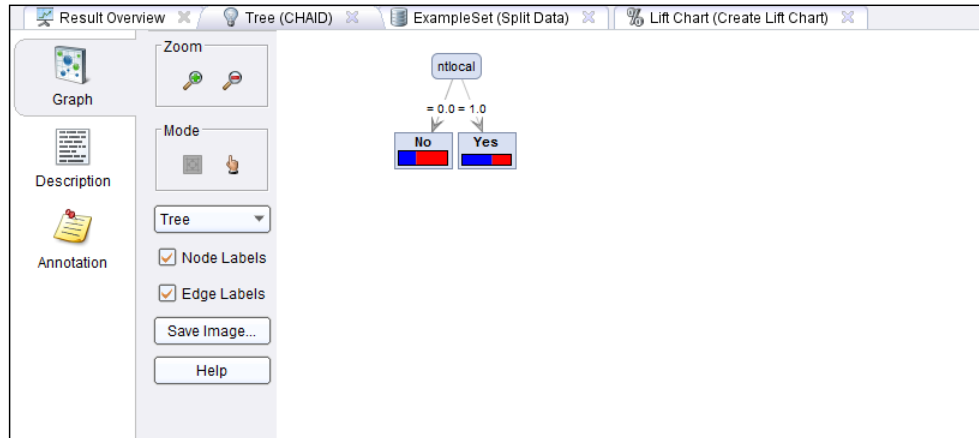
Now that you've created a model, it's time to look the model over, see how it performs, and choose your next steps.

Examining the decision tree

You've tried only two predictor variables on your first modeling attempt, so you're not expecting an elaborate result. The big question is whether you will find that even one of those variables has predictive value.

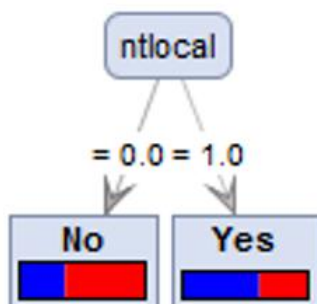
Your data-mining software displays the CHAID model as a decision-tree diagram in an interactive results viewer (see Figure 2-26). At first, only the first branch is displayed. Tools on the left side of the viewer enable you to expand the tree, to zoom in on areas of interest, and to make other changes to the way that the tree is displayed. You also have the alternative of viewing the model another way: written in simple text (see Figure 2-28).

Figure 2-26:
Tree model
viewer.



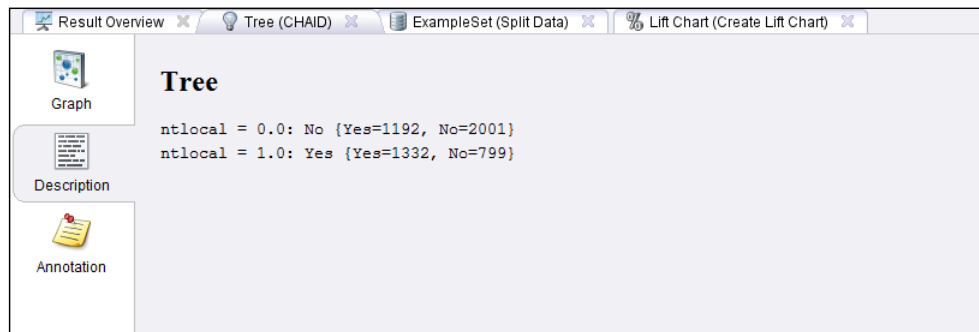
The tree (see Figure 2-27) shows that the local owner variable is the most important predictor. The data branches into two groups. Local owners ($ntlocal=0$) are indicators for the “No” category; most kept their property. Nonlocal owners ($ntlocal=1$) are indicators for the “Yes” category; they were more likely to sell. In this example, most of the properties with nonlocal owners changed hands; you can see that from the tiny bar chart on the tree branch. (But differences don’t have to be that dramatic to form a branch in the decision tree. Much more subtle differences can be detected if a strong-enough pattern exists in the data.)

Figure 2-27:
CHAID deci-
sion tree.



You use the pointer tool and click on the tree branches. They don’t expand. The local owner variable is the only one in the tree. A look at the model description (see Figure 2-28) shows the same thing in a different way.

Figure 2-28:
CHAID
model
description.



Why didn't the second predictor, the unpaid tax variable, show up in the model? Perhaps it really isn't a good indicator of change in property ownership. Perhaps it has some value, but the type of model you've chosen, or the settings you used (you left them all at the default values), were not appropriate for detecting the relationship between unpaid taxes and changes in property ownership. That's all you know for now.

Using a diagnostic chart

Diagnostic charts help you understand how effectively your model makes accurate predictions from the data available. (This is not unique to data mining; classical statisticians also use diagnostic charts.) A variety of diagnostic charts exist. You choose them based on what's available in your data-mining software and your own preferences.

You use a *lift chart* (see Figure 2-29), which compares your model's predictions to random selection. The chart is based on model predictions for the 30 percent of your data that you set aside for testing purposes before building the model. The bar on the left shows the group that the model gives the greatest confidence for a "Yes," a change in ownership. From your examination of the decision tree, you know that this group is the nonlocal property owners. The model predicts that each member of the group will be a "Yes," a change in ownership. For that group, the predictions are correct in 62.5 percent of the cases. (The confidence level noted at the base of each bar is the same as the proportion of correct predictions.)



In this model, you only see two bars in the chart, but lift charts for more complex models often have many bars. The greatest-confidence group is always the first bar on the left, and each subsequent bar has the next-greatest confidence.

By using the model, you can select 909 of the 2,282 cases (909 nonlocal + 1,373 local owners) in your test dataset to predict in the "Yes" category, and 62.5 percent of them, 549 cases, will be true changes in property ownership. The line through the bars shows that choosing 909 cases at random would only turn up about 280 true property changes. So the model is nearly doubling your effectiveness at predicting true property changes.

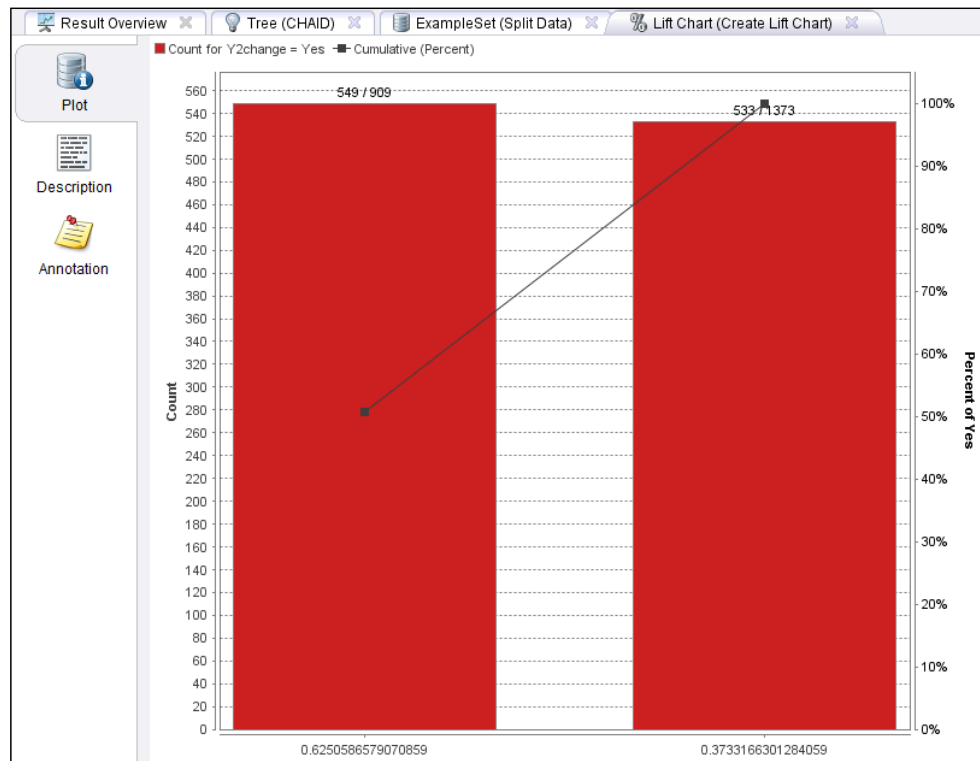


Figure 2-29:
Lift chart.



You'll find several types of lift charts. They all represent the advantages of using a model rather than random selection, but may vary in organization and appearance.

Assessing the status of the model

Your data-mining goal was to demonstrate the feasibility of using predictive modeling for property change ownership by showing that at least one variable has measurable predictive value for that purpose. Strictly speaking, you've met the goal, but if you still have time available before your deadline, you should use that time to improve on what you've done.

You have accomplished the minimum that you set out to provide. But you don't want to do just the minimum, so you keep working. You can try these things:

- ✓ Go back and do the data preparation needed for some more of the factors that Virginia and Matt suggested. (The list is in the "Choosing a starting point" section, earlier in this chapter.)
- ✓ Experiment with alternative model types.
- ✓ Refine the model settings.


You document what you've accomplished so far, and then return to the work to build the best model that you can before the project due date.

Putting Your Results into Action

In one day, you have not built a model that's ready to use in everyday business. That's fine; it was never your goal to do that. But you've already shown that predictive modeling is feasible, and that's pretty darned good for one day.

Because you've shown that modeling is a realistic option, chances are that the client will want you to go on and build the best model you can. When it's ready, you will put it into action making predictions.

You'll start by making lists of properties that are likely to change ownership. In fact, you've made one of these already. It's in the output from the chart tool (see Figure 2-30). A prediction exists for each property listed in the data. In the future, you can take advantage of other options to make predictions like these outside the data-mining software, and even integrate prediction capabilities into ordinary business applications.



Row No.	Y2change	prediction(Y2change)	confidence(Yes)	confidence(No)	TAXKEY	ntllocal
1	No	Yes	0.625	0.375	20032000	1.0
2	No	No	0.373	0.627	30142000	0.0
3	No	No	0.373	0.627	40061000	0.0
4	No	Yes	0.625	0.375	50100000	1.0
5	Yes	Yes	0.625	0.375	50122000	1.0
6	Yes	Yes	0.625	0.375	310311000	1.0
7	Yes	Yes	0.625	0.375	320083000	1.0
8	Yes	Yes	0.625	0.375	340589000	1.0
9	Yes	Yes	0.625	0.375	340615000	1.0
10	No	Yes	0.625	0.375	340920000	1.0
11	Yes	Yes	0.625	0.375	340388000	1.0
12	No	Yes	0.625	0.375	330412000	1.0
13	No	Yes	0.625	0.375	330252000	1.0
14	Yes	Yes	0.625	0.375	340314000	1.0
15	No	No	0.373	0.627	330110000	0.0
16	Yes	Yes	0.625	0.375	340403000	1.0
17	Yes	No	0.373	0.627	330300000	0.0
18	Yes	Yes	0.625	0.375	420152000	1.0
19	No	Yes	0.625	0.375	409973100	1.0
20	Yes	Yes	0.625	0.375	409975110	1.0
21	No	Yes	0.625	0.375	400181000	1.0
22	No	Yes	0.625	0.375	390361000	1.0

Figure 2-30:
A list of
predictions.