12 Case Study: Customer Churn

"There is only one boss. The customer. And he can fire everybody in the company from the chairman on down, simply by spending his money somewhere else."
—Sam Walton

Acme Telephonica (AT) is a mobile phone operator that has customers across every state of the USA. Like every telecommunications company, AT struggles with customer **churn**—customers leaving AT for other mobile phone operators. AT is always looking for new ways to address the churn issue and in 2008 founded a **customer retention** team. The customer retention team monitors the number of calls made to the AT customer support center by each customer and identifies the customers who make a large number of customer support calls as churn risks. The customer retention team contacts these customers with special offers designed to entice them to stay with AT. This approach, however, has not proved particularly successful, and churn has been steadily increasing over the last five years.

In 2010 AT hired Ross, a predictive data analytics specialist, to take a new approach to reducing customer churn. This case study describes the work carried out by Ross when he took AT through the CRISP-DM process[1] in order to develop a predictive data analytics solution to this business problem. The remainder of this chapter will discuss how each phase of the CRISP-DM process was addressed in this project.

12.1 Business Understanding

As is the case in most predictive data analytics projects, AT did not approach Ross with a well-specified predictive analytics solution. Instead, the company approached him with a business problem—reducing customer churn. Therefore, Ross's first goal was to convert this business problem into a concrete analytics solution. Before attempting this conversion, Ross had to fully understand the business objectives of AT. This was reasonably straightforward as AT management had stated that their goal was to reduce their customer churn rates. The only factor left unspecified was what the magnitude of that reduction was expected to

1. See Section 1.6[15].

be. Based on previous projects he had worked on, the current approach to customer retention that AT was taking, and AT's historical data, Ross agreed with AT management that a target reduction from the current high of approximately 10% to approximately 7.5% was realistic and probably achievable. Ross did stress to AT management that until he actually examined the data, he could not know how useful a model he would be able to build.

Ross's next task was to fully assess the current situation within AT. In particular, Ross needed to understand the current analytics capability of the company and its readiness to take action in response to the insights that an analytics solution would provide. AT already had a customer retention team proactively making interventions in an effort to reduce customer churn. Furthermore, this team was already using data from within the organization to choose which customers to target for intervention, which suggested that the team members were in a position to use predictive data analytics models.

Ross spent a significant amount of time meeting with Kate, the leader of the customer retention team, in order to understand how they worked. Kate explained that at the end of every month, a call list was generated, capturing the customers who had made more than three calls to the AT customer support service in the previous two months. These customers were deemed to be at risk of churning in the coming month, so the customer retention team set about contacting them with a special offer. Typically, the offer was a reduced call rate for the next three months, although retention team members had the freedom to make other offers.

Ross also spoke to the chief technology officer (CTO) at AT, Grace, in order to understand the available data resources. Ross learned that AT had reasonably sophisticated transactional systems for recording recent call activity and billing information. Historic call and bill records as well as customer demographic information were stored in a data warehouse. Grace had played a significant role in developing the process that had made information about customer support contacts available to the customer retention team. Ross hoped that this would make his task a little easier because Grace was the main gatekeeper to all the data resources at AT, and having her support for the project would be important. Other parts of the business that Ross spent significant time interviewing included the billing department, the sales and marketing team, and the network management.

Throughout the early stages of the project, Ross had been consciously working on developing his **situational fluency**. Through his discussions with the AT management team, Kate, and Grace, he had learned a lot about the mobile phone industry. The basic structure of the AT business was that customers had a contract for call services that AT provided. These contracts did not have a fixed time and were essentially renewed every month when a customer paid a fixed **recurring charge** for that month. Paying the recurring charge entitled a customer to a **bundle** of minutes of call time that were offered at a reduction to the standard call rate. For different recurring fees, customers received different-sized bundles of call time. When a customer used up all the call time in his or her bundle, subsequent

call time was referred to as **over bundle minutes**. These tended to be more expensive than the minutes included as part of a customer's bundle. At AT, all calls were classified as either **peak time calls** or **off-peak time calls**. Peak time was 08:00 to 18:00 from Monday to Friday, and calls made during peak time were more expensive than calls made during off-peak times.

Based on his assessment of the current situation within AT, Ross developed a list of ways that predictive analytics could help address the customer churn problem at AT. These included finding answers to the following questions:

- **What is the overall lifetime value of a customer?** A model could be built to predict the overall value that AT was likely to receive from a particular customer over the person's entire customer lifecycle. This could be used to identify customers who currently did not look valuable but that were likely to be valuable customers later in their customer lifecycles (college students often fall into this category). By offering these customers incentives now to prevent them from churning, AT would ensure that it received the full value from these customers in the future.
- **Which customers were most likely to churn in the near future?** A prediction model could be trained to identify the customers from the AT customer base who were most likely to churn in the near future. The retention team could focus their retention efforts on these customers. The process that the AT retention team had in place at the beginning of the project to identify customers likely to churn took a single feature approach to this identification—they looked only at how many calls a customer had made to the AT customer support service. It was likely that a machine learning model that looked at multiple features would do a better job of identifying customers likely to churn.
- **What retention offer would a particular customer best respond to?** A system could be built to predict which offer, from a set of possible retention offers, a particular customer would be most likely respond to when contacted by the AT retention team. This could help the retention team convince more customers to stay with AT.
- **Which pieces of the network infrastructure were likely to fail in the near future?** Using information about network loads, network usage, and equipment diagnostics, a predictive model could be built to flag upcoming equipment failures so that pre-emptive action could be taken. Network outages are a driver of customer dissatisfaction and ultimately customer churn, so reducing these could have a positive impact on churn rates.

After discussion with the AT executive team, it was decided that the analytics solution most appropriate to focus on was predicting *which customers are most likely to churn in the near future.* There were a number of reasons this project was selected:

- Ross's previous discussions with Grace, the AT CTO, had established that the data required to build a churn prediction model were likely to be available and reasonably easily accessible.

- The prediction model could be easily integrated with AT's current business processes. AT already had a retention team in place that was making proactive interventions to help prevent churn, albeit using a very simple system to identify which customers to contact. By creating a more sophisticated model to identify those customers, this existing process would be improved.
- Building a churn prediction model was also attractive to the AT executive team, as they hoped that as well as being useful in reducing churn rates, it would help to explain the main drivers behind customer churn. A better understanding of the main drivers of customer churn would be useful to many other parts of the AT business.

By contrast, the other analytics solutions developed suffered from a lack of available data (e.g., AT had no data available on the success or otherwise of various retention offers made); from being too significant a change in the business processes used by AT to be considered achievable at the time (e.g., generating a prediction of the overall lifetime value of a customer); or from not being based on sufficiently well-grounded assumptions (e.g., the fact that customer churn is heavily influenced by network failures).

Once the analytics solution had been defined, the next step was to agree on the expected performance of the new analytics model. Based on a recent evaluation of historical performance, AT management believed at the time of this project that their current system for identifying likely churners had an accuracy of approximately 60%, so any newly developed system would have to perform considerably better than this to be deemed worthwhile. In consultation with the members of the AT executive team and the retention team, Ross agreed that his goal would be to create a churn prediction system that would achieve a prediction accuracy in excess of 75%.

12.2 Data Understanding

During the process of determining which analytics solution was most suitable for the current situation at AT, Ross had already begun to understand the data resources available. His next task was to add much more depth to this understanding, following the process described in Section 2.3[28]. This involved working very closely with Grace to understand what data was available, the formats that the data was kept in, and where the data resided. This understanding would form the basis of Ross's work on designing the **domain concepts** and **descriptive features** that would make up the **analytics base table** (**ABT**), which would drive the creation of the predictive model. This was an iterative process in which Ross moved back and forth between Kate at the AT retention team, Grace, the CTO, and other parts of the business identified as having insight into the data associated with customer churn. It quickly became apparent that the key data resources within AT that would be important for this project were

- The customer demographic records from the AT data warehouse

- The customer billing records stored in the AT billing database, where records stretch back over a time horizon of five years
- The transactional record of calls made by individuals, stretching back over a time horizon of 18 months
- The sales team's transactional database, containing details of phone handsets issued to customers
- The retention team's simple transactional database, containing all the contacts they had made with customers, and the outcomes of these contacts, stretching back to a time horizon of 12 months

Before going any further, Ross had to define the **prediction subject** for the ABT and the target feature. The goal was to develop a model that would predict whether a customer would churn in the coming months. This meant that the prediction subject in this case was a customer, so the ABT would need to be built to contain one row per customer.

Predicting churn is a form of **propensity modeling**,[2] where the event of interest in this case is a customer making the decision to churn. Consequently, Ross needed to agree with the business (in particular the customer retention team) on a definition of churn. The definition would be used to identify churn events in AT's historical data and, consequently, was fundamental to building the ABT for the project. The business agreed that a customer who had been inactive for one month (i.e., had not made any calls or paid a bill) or who had explicitly canceled or not renewed a contract would be considered to have churned. Ross also needed to define the lengths of the **observation period** and the **outcome period** for the model. He decided that the **observation period**, during which he would collect data on customer behavior, would stretch back for 12 months. This was a decision made based on the data available and Ross's expectation that anything farther back than this was likely to have little impact on predicting churn. With regard to defining the **outcome period**, the company agreed that it would be most useful to make a prediction that a customer was likely to churn three months before the churn event took place, as this gave them time to take retention actions. Consequently, the outcome period was defined as three months.[3]

With the target feature suitably defined, Ross's next task was to determine the domain concepts that would underpin the design of the ABT. The domain concepts are those areas that the business believes have an impact on a customer's decision to churn. The domain concepts were developed through a series of workshops with representatives of various parts of the AT business—in particular the retention team, but also sales and marketing and billing. AT believed that the main concepts that affected churn were underlying

2. See Section 2.4.3[36].

3. Obviously, churn events will happen on different dates for different customers; therefore, to build the ABT, the observation and outcome periods for different customers would have to be aligned. This situation is an example of the propensity model scenario illustrated in Figure 2.6[39] in Section 2.4.3[36].

customer demographics (e.g., perhaps younger customers were more likely to churn); customer billing information, and in particular changes in billing patterns (e.g., perhaps customers whose bill suddenly increased were more likely to churn); the details of a customer's handset (e.g., perhaps customers who have had a handset for a long time are more likely to churn); the interactions a customer has had with AT customer care (e.g., perhaps customers who are making a large number of calls to customer care are having difficulties with the AT network and so are likely to churn); and the actual calls the customer is making, in particular, changing call patterns (e.g., perhaps customers who have started making calls to new groups of people are more likely to churn). This set of domain concepts was felt to be extensive enough to cover all the characteristics that were likely to contribute to a customer's likelihood to churn and is shown in Figure 12.1[690].

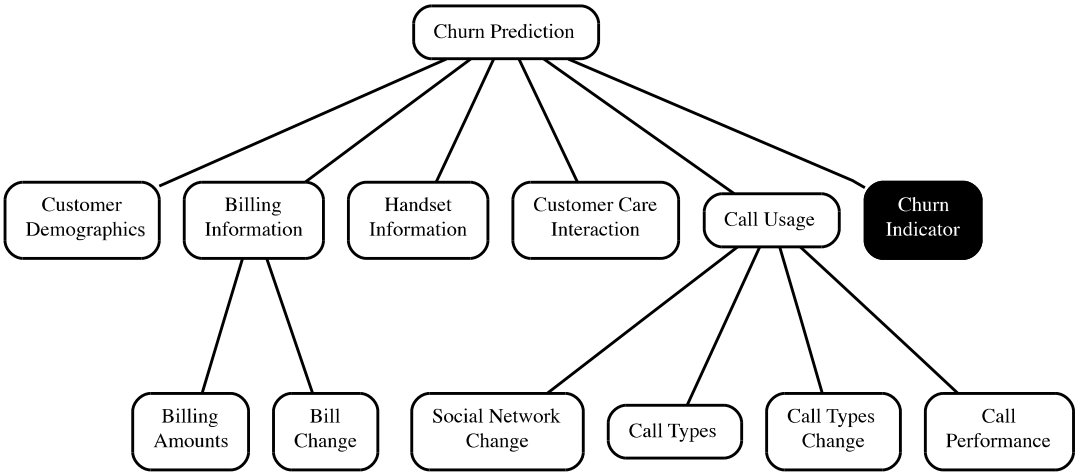


Figure 12.1
The set of domain concepts for the Acme Telephonica customer churn prediction problem.

From these domain concepts, Ross worked on deriving a set of descriptive features. Some of the descriptive features were simply copies of available raw data. For example, the AGE, GENDER, CREDITRATING, and OCCUPATION columns from the customer demographics data warehouse could be directly included as descriptive features in the ABT to capture the CUSTOMER DEMOGRAPHICS domain concept. The more interesting descriptive features were ones that had to be derived from the raw data sources. For example, Ross learned that the retention team believed that one of the main reasons customers churned was the availability of new, high-end handsets at other networks. To try to capture the HANDSET INFORMATION domain concept, Ross designed three descriptive features:

- SMARTPHONE: This feature indicated whether the customer's current handset was a smartphone, which was derived from the customer's most recent handset entry.

- NUMHANDSETS: This was a count of how many different handsets the customer had had in the past three years. This was derived from a count of all the handset entries for a particular customer.
- HANDSETAGE: Based on a customer's latest handset entry, this feature captured the number of days that the customer had had his or her current handset.

In churn analysis, and in any sort of propensity modeling, change is usually a key driver of customer behavior. For this reason, and based on discussions with the AT team, Ross included the BILL CHANGE and SOCIAL NETWORK CHANGE domain concepts. It was understood by the AT retention team that customers often made a decision to churn if their bill increased significantly due to changing call patterns, or when they began to make large numbers of calls to new friends or colleagues on other networks. For these reasons, Ross designed the following descriptive features:

- CALLMINUTESCHANGEPCT: Derived from the raw call data, this feature captured the amount by which the number of minutes a customer used had changed that month compared to the previous month.
- BILLAMOUNTCHANGEPCT: Derived from the raw call data, this feature captured the amount by which a customer's bill had changed that month compared to previous month.
- NEWFREQUENTNUMBERS: Derived from analysis of the actual numbers dialed in the raw call data, this feature attempted to capture how many new numbers a customer has begun calling frequently that month. A frequent number was defined as a number that constituted more than 15% of a customer's total calls.

Often descriptive features that are likely to be very useful cannot be implemented due to the unavailability of data. For example, the AT team felt that a customer beginning to frequently call other networks would be a good indicator of churn, but a suitable data feature could not be extracted to capture this. In its call records, AT did not include information about which network calls are made to, and with the free movement of numbers among operators, numbers themselves were no longer a reliable indicator of network.

The full set of descriptive features Ross developed, along with a short description of each, is shown in Table 12.1[692].

12.3 Data Preparation

With help from Grace to implement the actual data manipulation and data integration scripts using the tools available at AT, Ross populated an ABT containing all the features listed in Table 12.1[692]. Ross sampled data from the period 2008 to 2013. Using the definition of churn as a customer who had not made any calls or paid a bill for one month, Ross was able to identify churn events throughout this time period. To collect instances of customers who had not churned, Ross randomly sampled customers who did not match the churn definition but who also could be deemed active customers. Working with

Table 12.1
The descriptive features in the ABT developed for the Acme Telephonica churn prediction task.

Feature	Description
BILLAMOUNTCHANGEPCT	The percentage by which the customer's bill has changed from last month to this month
CALLMINUTESCHANGEPCT	The percentage by which the call minutes used by the customer has changed from last month to this month
AVGBILL	The average monthly bill amount
AVGRECURRINGCHARGE	The average monthly recurring charge paid by the customer
AVGDROPPEDCALLS	The average number of customer calls dropped each month
PEAKRATIOCHANGEPCT	The percentage by which the customer's peak calls to off-peak calls ratio has changed from last month to this month
AVGRECEIVEDMINS	The average number of calls received each month by the customer
AVGMINS	The average number of call minutes used by the customer each month
AVGOVERBUNDLEMINS	The average number of out-of-bundle minutes used by the customer each month
AVGROAMCALLS	The average number of roaming calls made by the customer each month
PEAKOFFPEAKRATIO	The ratio between peak and off-peak calls made by the customer this month
NEWFREQUENTNUMBERS	How many new numbers the customer is frequently calling this month
CUSTOMERCARECALLS	The number of customer care calls made by the customer last month
NUMRETENTIONCALLS	The number of times the customer has been called by the retention team
NUMRETENTIONOFFERS	The number of retention offers the customer has accepted
AGE	The customer's age
CREDITRATING	The customer's credit rating
INCOME	The customer's income level
LIFETIME	The number of months the customer has been with AT
OCCUPATION	The customer's occupation
REGIONTYPE	The type of region the customer lives in
HANDSETPRICE	The price of the customer's current handset
HANDSETAGE	The age of the customer's current handset
NUMHANDSETS	The number of handsets the customer has had in the past 3 years
SMARTPHONE	Whether the customer's current handset is a smartphone
CHURN	The target feature

Kate, Ross defined an active customer as a current customer who made at least five calls per week and who had been a customer for at least six months.[4] This definition ensured

4. The fact that active customers were defined as *current customers* means that they were all active on the same date—namely, whatever day the ABT was generated. This could be problematic: a model trained on this data might ignore seasonal effects such as Christmas. The alternative is to define active customers as any customer in the AT data that was active at some point. Such a definition, however, has the complication that the same customer could appear in the ABT as both an active and a churn customer, although admittedly the descriptive features for these two instances would be calculated over different periods of time.

that the non-churn instances in the dataset would include only customers with a relatively normal behavior profile and for which there was a long enough data history that realistic descriptive features could be calculated for them.

The final ABT contained 10,000 instances equally split between customers who churned and customers who did not churn. In the raw data, customers who did not churn outnumber those who churned at a ratio of over 10 to 1. This is an example of an **imbalanced dataset**, in which the different levels of the target feature—in this case, churners and non-churners—are not equally represented in the data. Some of the machine learning approaches we have discussed in the preceding chapters perform better when a **balanced sample** is used to train them, and this is why Ross created an ABT with equal numbers of instances with each target level.[5]

Ross then developed a full data quality report for the ABT including a range of data visualizations. The data quality report tables are shown in Table 12.2[694]. Ross first assessed the level of **missing values** within the data. Within the continuous features, only AGE stood out with 11.47% of values missing. This could be handled reasonably easily using an imputation approach,[6] but Ross held off on performing this at this stage. The REGIONTYPE and OCCUPATION categorical features both suffered from a significant number of missing values—74% and 47.8% respectively. Ross strongly considered removing these features entirely.

When he considered **cardinality**, Ross noticed that a number of the continuous features had very low cardinality—for example, INCOME, AGE, NUMHANDSETS, HANDSETPRICE, and NUMRETENTIONCALLS. In most cases, Ross confirmed with Kate and Grace that these were valid because the range of values that the features could take was naturally low. For example, HANDSETPRICE can take only a small number of values—e.g., 59.99, 129.99, 499.99, and so on. The INCOME feature stood out as unusual with only 10 distinct values (the histogram for this feature confirmed this; see Figure 12.2(a)[695]). Grace explained to Ross that incomes were actually recorded in bands rather than as exact values, so this was really a categorical feature. The cardinality of the CREDITCARD and REGIONTYPE categorical features were higher than expected (the histograms for these features are shown in Figures 12.2(b)[695] and 12.2(c)[695]). The issue was that some levels had multiple representations—for example, for the REGIONTYPE feature, towns were represented as *town* and as *t*. Ross easily corrected this issue by mapping the levels of the feature to one consistent labeling scheme.

5. We return to this discussion in Section 12.5[698] and Section 13.4.1[719].

6. See Section 3.4[69].

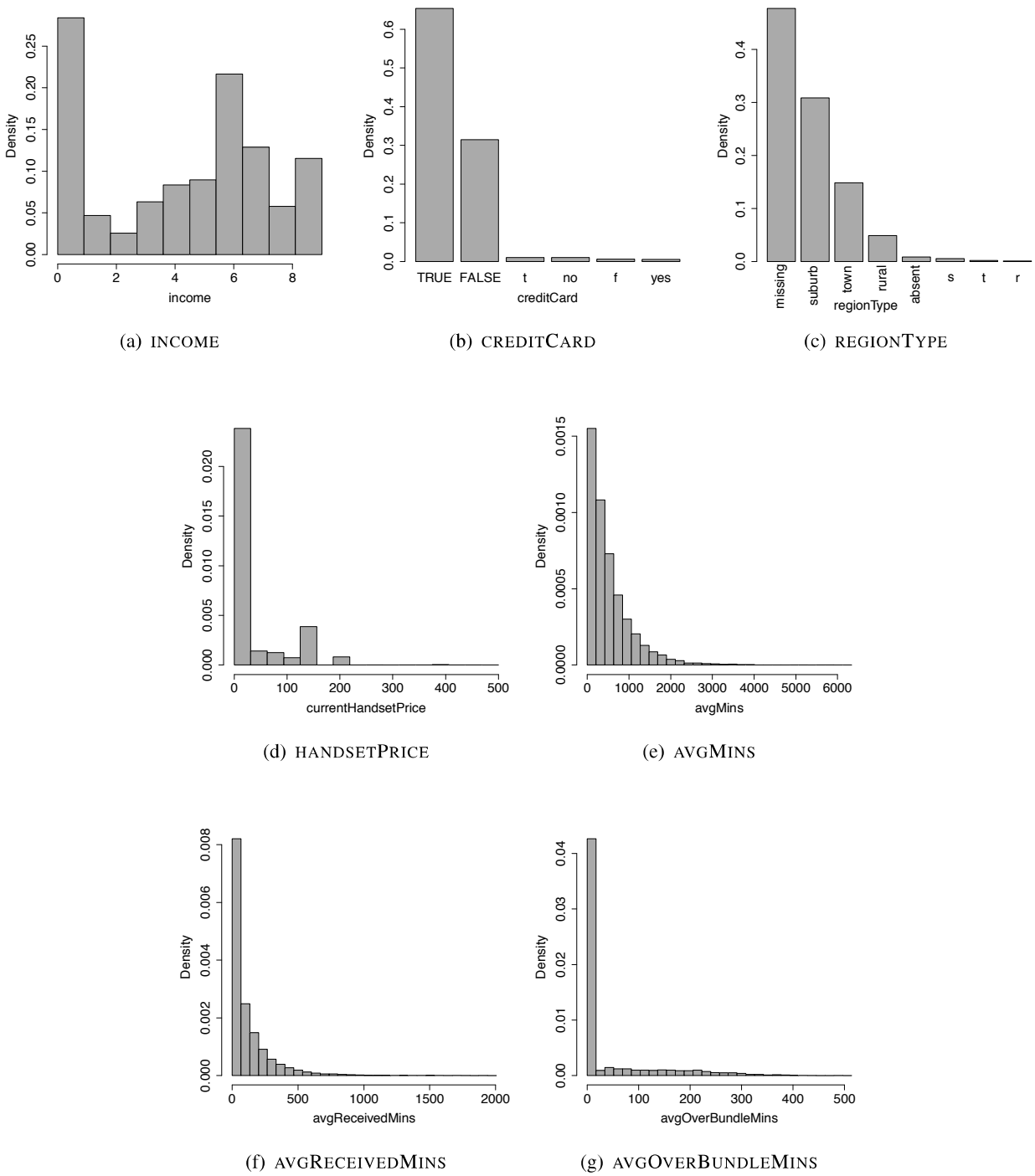
Table 12.2
A data quality report for the Acme Telephonica ABT.

(a) Data quality report for continuous features

Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
AGE	10,000	11.47	40	0.00	0.00	30.32	34.00	48.00	98.00	22.16
INCOME	10,000	0.00	10	0.00	0.00	4.30	5.00	7.00	9.00	3.14
NUMHANDSETS	10,000	0.00	19	1.00	1.00	1.81	1.00	2.00	21.00	1.35
HANDSETAGE	10,000	0.00	1,923	52.00	590.00	905.52	887.50	1,198.00	2,679.00	453.75
HANDSETPRICE	10,000	0.00	16	0.00	0.00	35.73	0.00	59.99	499.99	57.07
AVGBILL	10,000	0.00	5,588	0.00	33.33	58.93	49.21	71.76	584.23	43.89
AVGMINS	10,000	0.00	4,461	0.00	150.63	521.17	359.63	709.19	6,336.25	540.44
AVGRECURRINGCHARGE	10,000	0.00	1,380	0.00	30.00	46.24	44.99	59.99	337.98	23.97
AVGOVERBUNDLEMINS	10,000	0.00	2,808	0.00	0.00	40.65	0.00	37.73	513.84	81.12
AVGROAMCALLS	10,000	0.00	850	0.00	0.00	1.19	0.00	0.26	177.99	6.05
CALLMINUTESCHANGEPCT	10,000	0.00	10,000	-16.422	-1.49	0.76	0.50	2.74	19.28	3.86
BILLAMOUNTCHANGEPCT	10,000	0.00	10,000	-31.67	-2.63	2.96	1.96	7.56	42.89	8.51
AVGRECEIVEDMINS	10,000	0.00	7,103	0.00	7.69	115.27	52.54	154.38	2,006.29	169.98
AVGOUTCALLS	10,000	0.00	524	0.00	3.00	25.29	13.33	33.33	610.33	35.66
AVGINCALLS	10,000	0.00	310	0.00	0.00	8.37	2.00	9.00	304.00	17.68
PEAKOFFPEAKRATIO	10,000	0.00	8,307	0.00	0.78	2.22	1.40	2.50	160.00	3.88
PEAKRATIOCHANGEPCT	10,000	0.00	10,000	-41.32	-6.79	-0.05	0.01	6.50	37.78	9.97
AVGDROPPEDCALLS	10,000	0.00	1,479	0.00	0.00	0.50	0.00	0.00	9.89	1.41
LIFETIME	10,000	0.00	56	6.00	11.00	18.84	17.00	24.00	61.00	9.61
CUSTOMERCARECALLS	10,000	0.00	109	0.00	0.00	1.74	0.00	1.33	365.67	5.76
NUMRETENTIONCALLS	10,000	0.00	5	0.00	0.00	0.05	0.00	0.00	4.00	0.23
NUMRETENTIONOFFERS	10,000	0.00	5	0.00	0.00	0.02	0.00	0.00	4.00	0.155
NEWFREQUENTNUMBERS	10,000	0.00	4	0.00	0.00	0.20	0.00	0.00	3.00	0.64

(b) Data quality report for categorical features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
OCCUPATION	10,000	74.00	8	professional	1,705	65.58	crafts	274	10.54
REGIONTYPE	10,000	47.80	8	suburb	3,085	59.05	town	1,483	28.39
MARRIAGESTATUS	10,000	0.00	3	unknown	3,920	39.20	yes	3,594	35.94
CHILDREN	10,000	0.00	2	false	7,559	75.59	true	2,441	24.41
SMARTPHONE	10,000	0.00	2	true	9,015	90.15	false	985	9.85
CREDITRATING	10,000	0.00	7	b	3,785	37.85	c	1,713	17.13
HOMEOWNER	10,000	0.00	2	false	6,577	65.77	true	3,423	34.23
CREDITCARD	10,000	0.00	6	true	6,537	65.37	false	3,146	31.46
CHURN	10,000	0.00	2	false	5,000	50.00	true	5,000	50.00



(a) INCOME (b) CREDITCARD (c) REGIONTYPE

(d) HANDSETPRICE (e) AVGMINS

(f) AVGRECEIVEDMINS (g) AVGOVERBUNDLEMINS

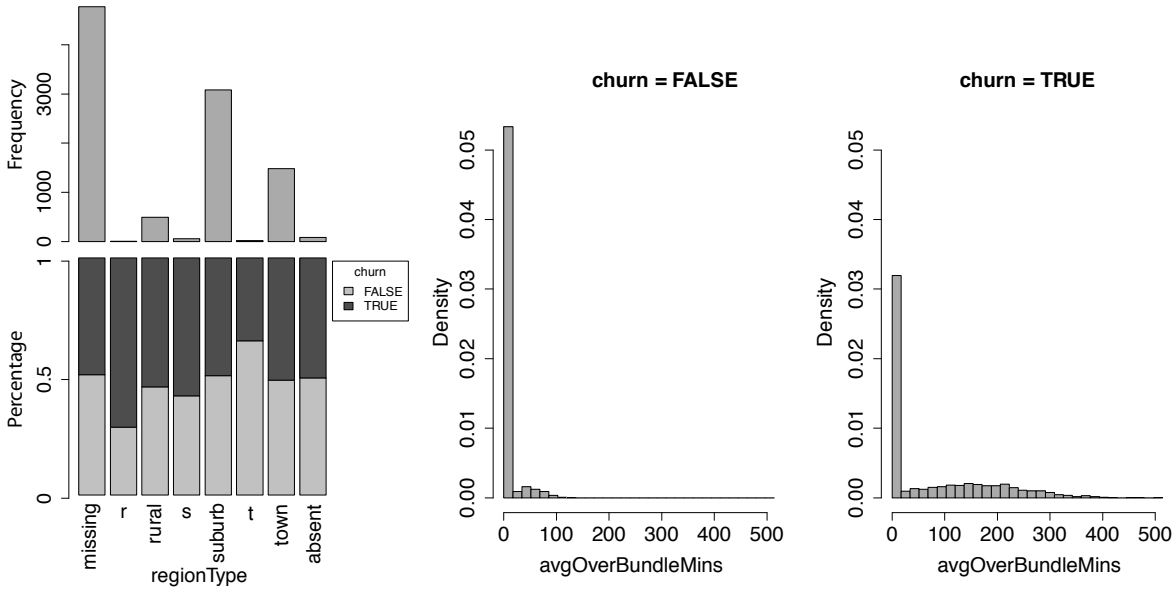
Figure 12.2
(a)–(c) Histograms for the features from the AT ABT with irregular cardinality; (d)–(g) histograms for the features from the AT ABT that are potentially suffering from outliers.

Four continuous features stood out as possibly suffering from the presence of **outliers**: HANDSETPRICE, with a minimum value of 0, which seemed unusual; AVGMINS, with a maximum of 6,336.25, which was very different from the mean and the 3rd quartile values for that feature; AVGRECEIVEDMINS, with a maximum of 2,006.29, which was also very different from the mean and the 3rd quartile values for that feature; and AVGOVERBUNDLEMINS, with minimum, 1st quartile, and median values of 0 compared to a mean of 40. Figure 12.2[695] presents the histograms for these features. Ross confirmed with Grace and Kate that these were valid outliers—for example, some handsets are given away for free, and some customers just make a lot of calls. They did spend some time, however, discussing the AVGOVERBUNDLEMINS. The histogram for this feature has an unusual shape that results in the unusual minimum, 1st quartile, and median values (see Figure 12.2(g)[695]). By examining the data for this feature more closely, they eventually explained this shape by the fact that most customers did not go over the number of minutes in their bundle, which accounts for the large bar for 0 in this histogram. The values above zero seem to follow something close to a wide normal distribution, and the large number of, albeit valid, zero values account for the unusual minimum, 1st quartile, and median values. At this point Ross just made note of these outliers as something he might have to deal with during the modeling phase.

Ross then turned his attention to examining the data visualizations of the relationship between each descriptive feature and the target feature. No individual feature stood out as having a very strong relationship, but the evidence of connections between the descriptive features and the target feature could be seen. For example, Figure 12.3(a)[697] shows a slightly higher propensity of people in rural areas to churn. Similarly, Figure 12.3(b)[697] shows that customers who churned tended to make more calls outside their *bundle* than those who did not.

Once Ross had reviewed the full data quality report in detail, he made the following decisions regarding the problematic features he had identified. First, he decided to delete the AGE and OCCUPATION features because of the level of missing values in each of these features. He decided to keep the REGIONTYPE feature, however, because it appeared to have some relationship with the target. He also applied the planned mapping of the REGIONTYPE values to a consistent labeling scheme: $\{s|suburb\} \rightarrow suburb$; $\{t|town\} \rightarrow town$; $\{missing|absent\} \rightarrow missing$.

Ross further divided this dataset into three randomly sampled partitions—a training partition (50%), a validation partition (20%) and a test partition (30%). The training partition was used as the core training data for the prediction models built. The validation partition was used for tuning tasks, and the test partition was used for nothing other than a final test of the model to evaluate its performance.



(a) REGIONTYPE (b) AVGOVERBUNDLEMINS

Figure 12.3
(a) A stacked bar plot for the REGIONTYPE feature; and (b) histograms for the AVGOVERBUNDLEMINS feature by target feature value.

12.4 Modeling

The requirements for this model were that it be accurate, that it be capable of being integrated into the wider AT processes, and, possibly, that it act as a source of insight into the reasons people might churn. In selecting the appropriate model type to use, all these aspects, along with the structure of the data, should be taken into account. In this case, the ABT was composed of a mixture of continuous and categorical descriptive features and had a categorical target feature. The categorical target feature, in particular, makes decision trees a suitable choice for this modeling task. Furthermore, decision tree algorithms are capable of handling both categorical and continuous descriptive features as well as handling missing values and outliers without any need to transform the data. Finally, decision trees are relatively easy to interpret, which means that the structure of the model can give some insight into customer behavior. All these factors taken together indicated that decision trees were an appropriate modeling choice for this problem.

Ross used the ABT to train, tune, and test a series of decision trees to predict churn given the set of descriptive features. The first tree that Ross built used an entropy-based information gain as the splitting criterion, limited continuous splits to binary choices, and no pruning. Ross had decided, again in consultation with the business, that a simple classification accuracy rate was the most appropriate evaluation measure for this task. The first

tree constructed achieved an **average class accuracy**[7] of 74.873% on the hold-out test set, which was reasonably encouraging.

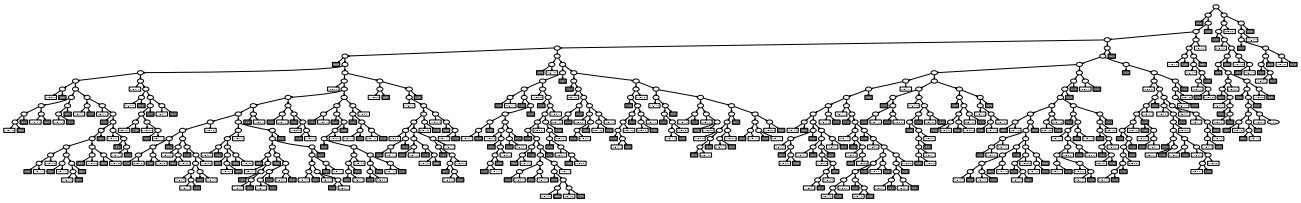


Figure 12.4
An unpruned decision tree built for the AT churn prediction problem (shown only to indicate its size and complexity). The excessive complexity and depth of the tree are evidence that overfitting has probably occurred.

This tree is shown in Figure 12.4[698], and the lack of pruning is obvious in its complexity. This complexity and the excessive depth of the tree suggest overfitting. In the second tree that he built, Ross employed **post-pruning** using **reduced error pruning**,[8] which used the validation partition that was created from the initial dataset. The reasonably large dataset that Ross had to begin with, which in turn led to a reasonably large validation partition, meant that reduced error pruning was appropriate in this case.[9] Figure 12.5[699] shows the tree resulting from this training iteration. It should be clear that this is a much simpler tree than the previous one. The features used at the top levels of both trees, and deemed most informative by the algorithm, were the same: AVGOVERBUNDLEMINS, BILLAMOUNTCHANGEPCT, and HANDSETAGE.

Using pruning, Ross was able to increase the average class accuracy on the hold-out test set to 79.03%, a significant improvement over the previous model. Table 12.3[699] shows the confusion matrix from this test. The confusion matrix shows that this model was slightly more accurate when classifying instances with the *non-churn* target level than with the *churn* target level. Based on these, results Ross was confident that this tree was a good solution for the AT churn prediction problem.

12.5 Evaluation

The model evaluations based on misclassification rate described in the previous section are the first step in evaluating the performance of the prediction model created. The classification accuracy of 79.03% is well above the target agreed on with the business. This is

7. All average class accuracies used in this section use a **harmonic mean**.

8. See Section 4.4.4[153].

9. If data had been more scarce, pruning using a statistical test, such as χ^2, would have been a more sensible route to take.

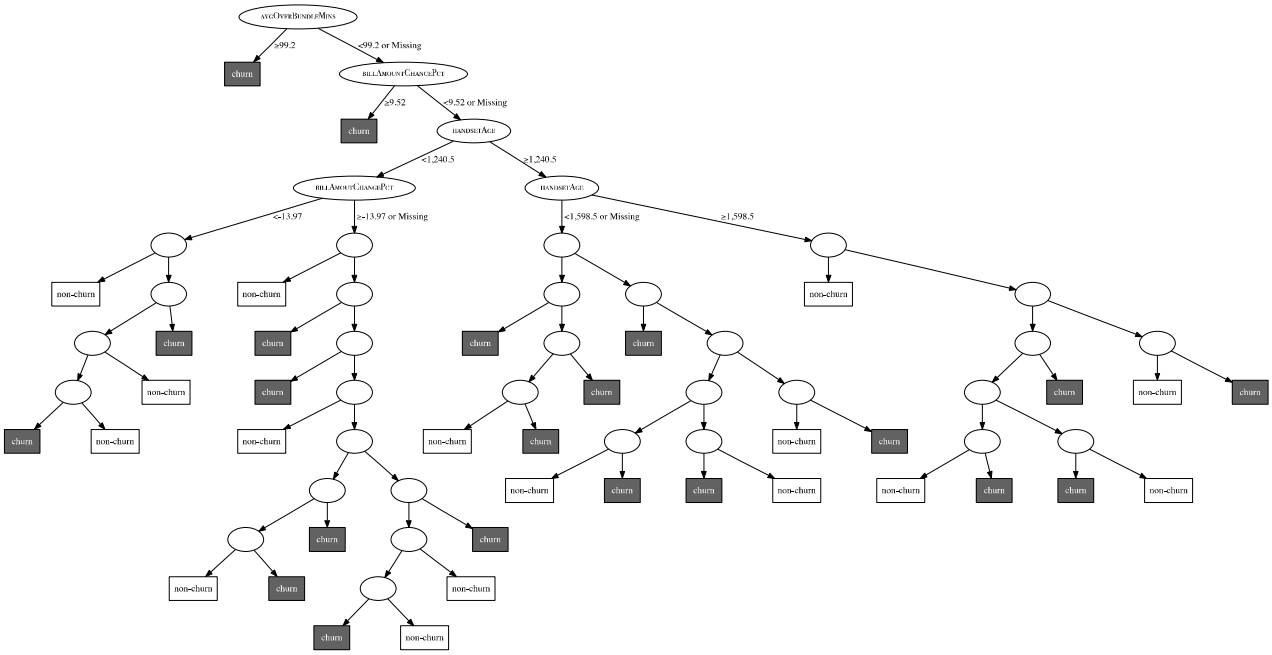


Figure 12.5
A pruned decision tree built for the AT churn prediction problem. Gray leaf nodes indicate a churn prediction, and clear leaf nodes indicate a non-churn prediction. For space reasons, we show only the features tested at the top-level nodes.

Table 12.3
The confusion matrix from the test of the AT churn prediction stratified hold-out test set using the pruned decision tree in Figure 12.5[699].

		Prediction		
		churn	*non-churn*	Recall
Target	*churn*	1,058	442	70.53
	non-churn	152	1,348	89.86

misleading, however. This performance is based on a stratified hold-out test set, which contains the same number of churners and non-churners. The underlying distribution of churners and non-churners within the larger AT customer base, however, is much different. Rather than a 50:50 split of churners to non-churners, the actual underlying ratio is, in fact, closer to 10:90. For this reason, it is very important to perform a second evaluation in which the test data reflect the actual distribution of target feature values in the business scenario.

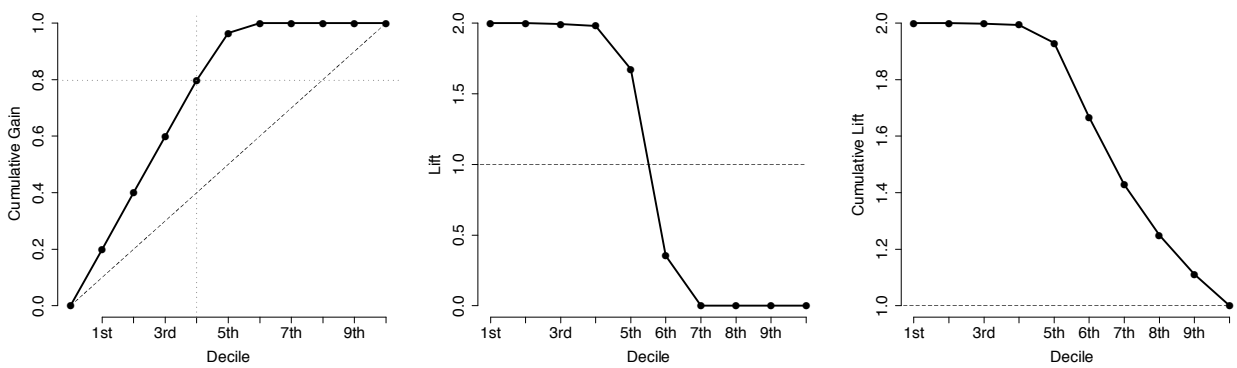
Ross had AT generate a second data sample (which did not overlap with the sample taken previously) that was not stratified according to the target feature values. The confusion

Table 12.4
The confusion matrix from the test of the AT churn prediction non-stratified hold-out test set.

		Prediction		
		churn	*non-churn*	Recall
Target	*churn*	1,115	458	70.88
	non-churn	1,439	12,878	89.95

matrix illustrating the performance of the prediction model on this test set is shown in Table 12.4[700].

The average class accuracy on the non-stratified hold-out test set was 79.284%. Ross also generated **cumulative gain**, **lift**, and **cumulative lift** charts for the dataset.[10] These are shown in Figure 12.6[700]. The cumulative gain chart in particular shows that if AT were to call just 40% of their customer base, they would identify approximately 80% of the customers who are likely to churn, which is strong evidence that the model is doing a good job of distinguishing between different customer types.



(a) Cumulative gain (b) Lift (c) Cumulative lift

Figure 12.6
(a) Cumulative gain, (b) lift, and (c) cumulative lift charts for the predictions made on the large test data sample.

Given these good results Ross decided that it was appropriate to present the model to other parts of the business. This was an important step in gaining credibility for the model. The tree shown in Figure 12.5[699] is reasonably straightforward to interpret, but when taken out to other parts of the business, it may be hard for people to deal with this much informa-

10. **Cumulative gain**, **lift**, and **cumulative lift** are introduced in Section 9.4.3.3[565].

tion, so Ross decided to create a purposefully stunted version of the decision tree, with only a small number of levels shown for the presentation of the model to the business (although he intended to use the larger pruned tree for actual deployment). The idea behind this was that stunting the tree made it more interpretable. The fact that the most informative features occupy berths toward the top of a tree means that stunted trees usually capture the most important information. Many machine learning tools will allow the maximum depth of a tree to be specified as a parameter, which allows for the creation of such **stunted trees**.

Figure 12.7[701] shows the stunted tree Ross generated for the churn problem, where the depth of the tree is limited to five layers. This tree results in a slightly lower classification accuracy on the test partition, 78.5%, but is very easy to interpret—the key features in determining churn are clearly AVGOVERBUNDLEMINS, BILLAMOUNTCHANGEPCT, and HANDSETAGE. It seems, from this data, that customers are most likely to churn when their bill changes dramatically, when they begin to exceed the bundled minutes in their call package, or when they have had a handset for a long time and are considering changing to something newer. This is useful information that the business can use to attempt to devise other churn handling strategies in parallel to using this model to create call lists for the retention team. The business was interested in the features that were selected as important to the tree, and there was a good deal of discussion on the omission of the features describing customers' interactions with AT customer care (these had been the basis of the organization's previous model).

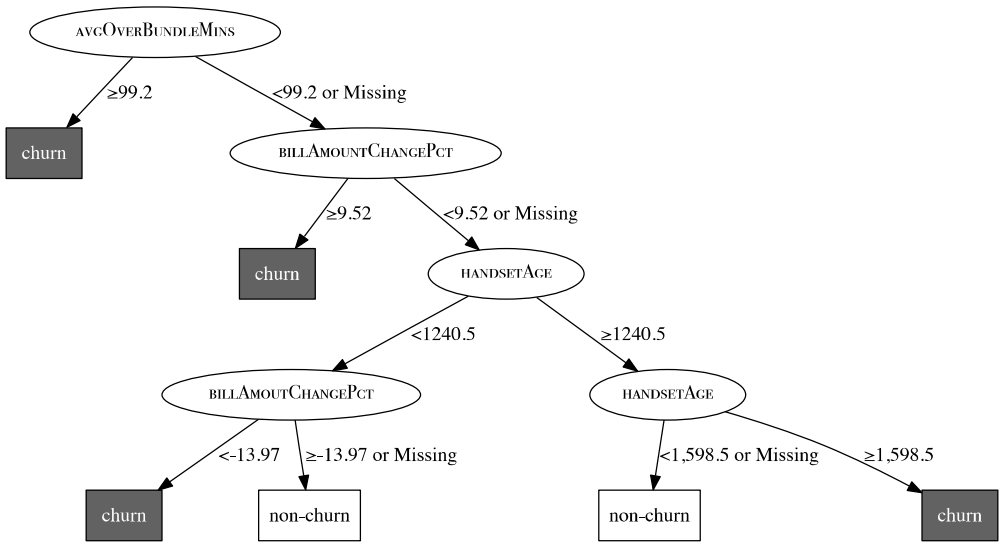


Figure 12.7
A pruned and stunted decision tree built for the Acme Telephonica churn prediction problem.

To further support his model, Ross organized a control group test (see Section 9.4.6[578]) in which for two months, the AT customer base was randomly divided into two groups, and call lists for the retention team were selected from the first group using the old approach based on calls to customer care, and for the second group using the new decision tree model. It was shown after two months that the churn rate within the sample for which the retention team used the new model to build their call list was approximately 7.4%, while for the group using the old model, it was over 10%. This experiment showed the AT executive team that the new decision tree model could significantly reduce churn rates within the AT customer base.

12.6 Deployment

Because AT was already using a process in which its retention team generated call lists based on collected data, deployment of the new decision tree model was reasonably straightforward. The main challenge was a return to the Data Preparation phase to make the routines used to extract the data for the ABT robust and reliable enough to be used to generate new query instances every month. This involved working with the AT IT department to develop deployment-ready **extract-transform-load** (ETL) routines. Code was then written to replace the previous simple rule about customer care contacts with the decision tree when retention call lists were generated.

The last step in deployment was to put in place an **ongoing model validation** plan to raise an alarm if evidence arose indicating that the deployed model had gone **stale**. In this scenario, feedback on the performance of the model implicitly arises within a reasonably short amount of time after predictions are made—churn predictions can be easily compared to actual customer behavior (taking into account interventions made by the business). The monitoring system that Ross put in place generated a report at the end of every quarter that evaluated the performance of the model in the previous quarter by comparing how many of the people not contacted by the retention team actually churned. If this number changed significantly from what was seen in the data used to build the model, the model would be deemed stale, and retraining would be required.