CASE STUDY 1

ANALYZING AUTOMOBILE WARRANTY CLAIMS: EXAMPLE OF THE CRISP–DM INDUSTRY STANDARD PROCESS IN ACTION [17]

Quality assurance continues to be a priority for automobile manufacturers, including Daimler Chrysler. Jochen Hipp of the University of Tubingen, Germany, and Guido Lindner of DaimlerChrysler AG, Germany, investigated patterns in the warranty claims for DaimlerChrysler automobiles.

1. Business Understanding Phase

DaimlerChrysler's objectives are to reduce costs associated with warranty claims and improve customer satisfaction. Through conversations with plant engineers, who are the technical experts in vehicle manufacturing, the researchers are able to formulate specific business problems, such as the following:

- Are there interdependencies among warranty claims?
- Are past warranty claims associated with similar claims in the future?
- Is there an association between a certain type of claim and a particular garage?

The plan is to apply appropriate data mining techniques to try to uncover these and other possible associations.

2. Data Understanding Phase

The researchers make use of DaimlerChrysler's Quality Information System (QUIS), which contains information on over 7 million vehicles and is about 40 gigabytes in size. QUIS contains production details about how and where a particular vehicle was constructed, including an average of 30 or more sales codes for each vehicle. QUIS also includes warranty claim information, which the garage supplies, in the form of one of more than 5000 possible potential causes.

The researchers stressed the fact that the database was entirely unintelligible to domain nonexperts: "So experts from different departments had to be located and consulted; in brief a task that turned out to be rather costly." They emphasize that analysts should not underestimate the importance, difficulty, and potential cost of this early phase of the data mining process, and that shortcuts here may lead to expensive reiterations of the process downstream.

3. Data Preparation Phase

The researchers found that although relational, the QUIS database had limited SQL access. They needed to select the cases and variables of interest manually, and then manually derive new variables that could be used for the modeling phase. For example, the variable *number of days from selling date until first claim* had to be derived from the appropriate date attributes.

They then turned to proprietary data mining software, which had been used at DaimlerChrysler on earlier projects. Here they ran into a common roadblock—that the data format requirements varied from algorithm to algorithm. The result was further exhaustive preprocessing of the data, to transform the attributes into a form usable for model algorithms. The researchers mention that the data preparation phase took much longer than they had planned.

4. Modeling Phase

Since the overall business problem from phase 1 was to investigate dependence among the warranty claims, the researchers chose to apply the following techniques: (1) Bayesian networks and (2) association rules. Bayesian networks model uncertainty by explicitly representing the conditional dependencies among various components, thus providing a graphical visualization of the dependency relationships among the components. As such, Bayesian networks represent a natural choice for modeling dependence among warranty claims. The mining of association rules is covered in Chapter 10. Association rules are also a natural way to investigate dependence among warranty claims since the confidence measure represents a type of conditional probability, similar to Bayesian networks.

The details of the results are confidential, but we can get a general idea of the type of dependencies uncovered by the models. One insight the researchers uncovered was that a particular combination of construction specifications doubles the probability of encountering an automobile electrical cable problem. DaimlerChrysler engineers have begun to investigate how this combination of factors can result in an increase in cable problems.

The researchers investigated whether certain garages had more warranty claims of a certain type than did other garages. Their association rule results showed that, indeed, the confidence levels for the rule "If garage *X*, then cable problem," varied considerably from garage to garage. They state that further investigation is warranted to reveal the reasons for the disparity.

5. Evaluation Phase

The researchers were disappointed that the support for sequential-type association rules was relatively small, thus precluding generalization of the results, in their opinion. Overall, in fact, the researchers state: "In fact, we did not find any rule that our domain experts would judge as interesting, at least at first sight." According to this criterion, then, the models were found to be lacking in effectiveness and to fall short of the objectives set for them in the business understanding phase. To account for this, the researchers point to the "legacy" structure of the database, for which automobile parts were categorized by garages and factories for historic or technical reasons and not designed for data mining. They suggest adapting and redesigning the database to make it more amenable to knowledge discovery.

6. Deployment Phase

The researchers have identified the foregoing project as a pilot project, and as such, do not intend to deploy any large-scale models from this first iteration. After the pilot project, however, they have applied the lessons learned from this project, with the goal of integrating their methods with the existing information technology environment at DaimlerChrysler. To further support the original goal of lowering claims costs, they intend to develop an intranet offering mining capability of QUIS for all corporate employees.

What lessons can we draw from this case study? First, the general impression one draws is that uncovering hidden nuggets of knowledge in databases is a rocky road. In nearly every phase, the researchers ran into unexpected roadblocks and difficulties. This tells us that actually applying data mining for the first time in a company requires asking people to do something new and different, which is not always welcome. Therefore, if they expect results, corporate management must be 100% supportive of new data mining initiatives.

Another lesson to draw is that intense human participation and supervision is required at every stage of the data mining process. For example, the algorithms require specific data formats, which may require substantial preprocessing (see Chapter 2). Regardless of what some software vendor advertisements may claim, you can't just purchase some data mining software, install it, sit back, and watch it solve all your problems. Data mining is not magic. Without skilled human supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. The wrong analysis is worse than no analysis, since it leads to policy recommendations that will probably turn out to be expensive failures.

Finally, from this case study we can draw the lesson that there is no guarantee of positive results when mining data for actionable knowledge, any more than when one is mining for gold. Data mining is not a panacea for solving business problems. But used properly, by people who understand the models involved, the data requirements, and the overall project objectives, data mining can indeed provide actionable and highly profitable results.

FALLACIES OF DATA MINING

Speaking before the U.S. House of Representatives Subcommittee on Technology, Information Policy, Intergovernmental Relations, and Census, Jen Que Louie, president of Nautilus Systems, Inc., described four fallacies of data mining [18]. Two of these fallacies parallel the warnings we described above.

- ***Fallacy 1.*** There are data mining tools that we can turn loose on our data repositories and use to find answers to our problems.
  - *Reality.* There are no automatic data mining tools that will solve your problems mechanically "while you wait." Rather, data mining is a process, as we have seen above. CRISP–DM is one method for fitting the data mining process into the overall business or research plan of action.
- ***Fallacy 2.*** The data mining process is autonomous, requiring little or no human oversight.
  - *Reality.* As we saw above, the data mining process requires significant human interactivity at each stage. Even after the model is deployed, the introduction of new data often requires an updating of the model. Continuous quality monitoring and other evaluative measures must be assessed by human analysts.
- ***Fallacy 3.*** Data mining pays for itself quite quickly.
  - *Reality.* The return rates vary, depending on the startup costs, analysis personnel costs, data warehousing preparation costs, and so on.
- ***Fallacy 4.*** Data mining software packages are intuitive and easy to use.
  - *Reality.* Again, ease of use varies. However, data analysts must combine subject matter knowledge with an analytical mind and a familiarity with the overall business or research model.