Paper 357-2008

It's 9:00am - Do You Know Where Your Critical Talent Is?

Retention Analytics for Human Capital Management

Lynn Creecy, SAS Institute Inc., Cary, NC
Brad Klenz, SAS Institute Inc., Cary, NC

ABSTRACT

Employee retention is an increasingly serious issue in many business sectors. Understanding which factors cause employees to leave and which actions retain them is an important Business Intelligence application. This paper demonstrates analytic methods to address this problem.

Data mining and predictive modeling can be used to improve retention of critical employees. The predictive modeling approach yields two valuable outputs that can be used to gain insight into this problem:

1) A *retention score* for each employee, which measures the probability that an employee will leave over a given time period (for example, in the next year).

2) A statistical model that quantifies the relationship of the input data to the probability of retention. Examples of input data include an employee's job-action history, the length of time in a position, and salary history.

The business user can use the retention analysis to generate reports that show how the loss of critical skills would affect an organization. Reports identify job groups, geographical regions, or organizational areas that have higher risk for employee voluntary termination. Additionally, the influential drivers to high-risk groups are identified to suggest the best course of action to reduce the risk.

Organizations gain insight and knowledge about workforce dynamics through these reports, which leads to better human-capital management and less risk of losing critical talent.

INTRODUCTION

It is now routine to build predictive models using data about individuals to determine the likelihood of various events. Examples include:

- Product recommendations
- Credit scoring
- Product or service cancellation

Within a company, much data is available on factors that influence an employee's past and present relationship with that company. Additionally, the Human Resource (HR) management systems have records of all job status changes, including voluntary terminations.

Similar to the applications above, we have the ability to build predictive models on the likelihood of an employee to voluntarily terminate.

Knowing the likelihood of an employee to voluntarily terminate becomes the foundation of an employee-retention program that can yield significant benefits to a company, particularly in the area of critical talent. Companies that have implemented such a program, such as Banca Carige in Italy (SAS Institute, 2007), have reaped these benefits.

The predictive model will create a *retention score,* to quantify the likelihood of an employee to voluntarily terminate. Using the method described in this paper, this retention score is the probability that an employee will voluntarily terminate in the next year. (How this score is built is described in more detail later in the paper.) Every employee will have a retention score. For example, Lynn will have a 7% chance of leaving in the next year, and Brad will have a 16% chance of leaving in the next year.

Having individual-level scores provides much information that can be useful in improving individual retention. You could take the list of the highest risk individuals and notify their managers. Managers could then decide on actions that might be taken to retain the individuals.

The individual scores by themselves, however, do very little to guide changes in a workforce-retention strategy for the company. Here are some important strategy questions individual scores do not answer:

- Are the high-risk individuals in critical roles?
- Do the high-risk individuals have skills that are difficult to replace?
- Are there specific areas in the company (location, departments, or job titles) that are above or below average risk to voluntary terminations?
- What is influencing groups that have a high risk of terminations?

Knowing the answers to these questions would allow new strategies and policies to improve the overall risk to the company.

GAINING INSIGHT INTO AN ORGANIZATION

Organizations can gain valuable insight by looking at the employee attributes for high-risk employees. Looking at the aggregate scores and finding trends for employees with high-termination probability helps pinpoint the reasons for termination (Figure 1). If organizations know why their employees are likely to leave, they can develop effective policies and strategies for employee retention. They can identify and respond to a problem before it affects the bottom line. Organizations can also use turnover predictions to refine forecasts for resources that are necessary to meet future strategic goals. The results are more accurate budgeting and capacity planning.

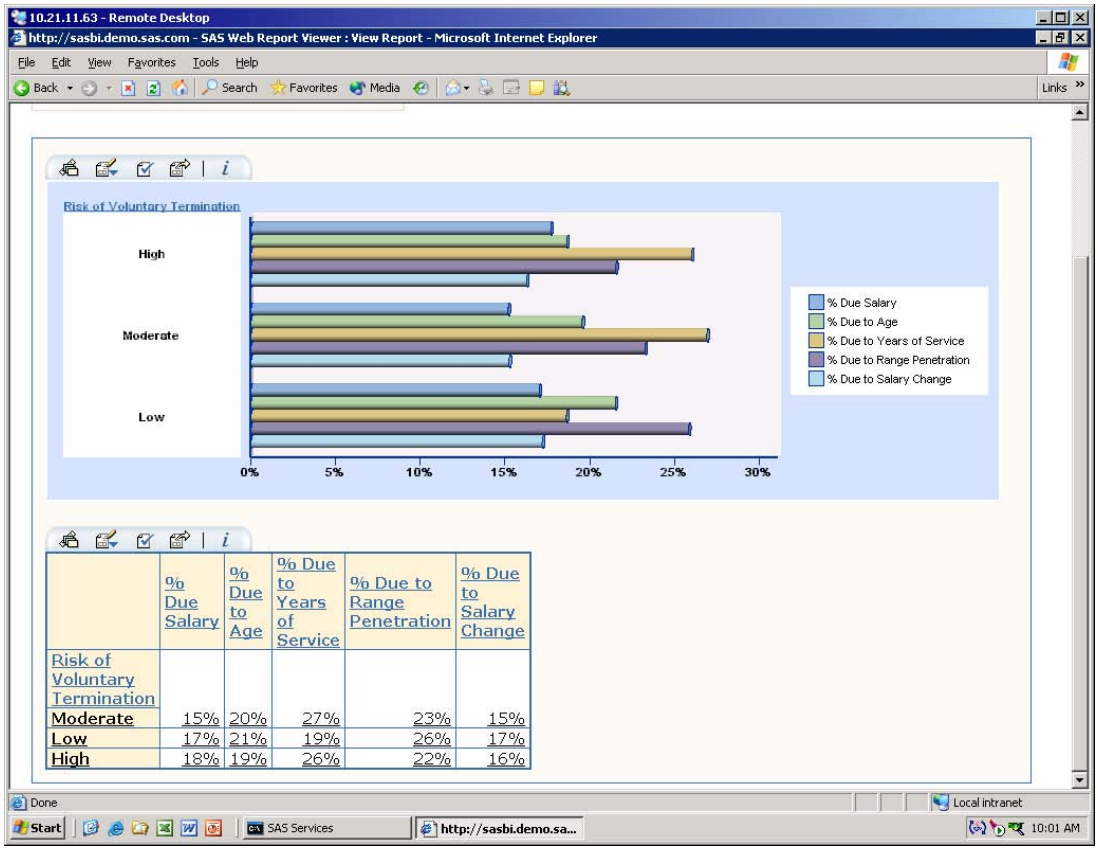

Figure 1. Reasons for voluntary termination

Consider these questions:

- Where are the employees located organizationally, for example, in the tech-support call center or in the sales organization?
- Where are they located geographically, for example, San Diego or Boston?
- To whom do they report?
- What type of occupation or job title is most affected, for example, first line supervisors or java programmers?
- How long have the employees been employed, less than five years or more than ten years?
- How were they recruited, for example at a job fair or from an Internet job website?
- How are employees compensated?
- Is the amount of training for these employees below average for the organization or for industry standards?

When you find the answers to questions like these, you can build a profile of employees at high risk to terminate (Figure 2).

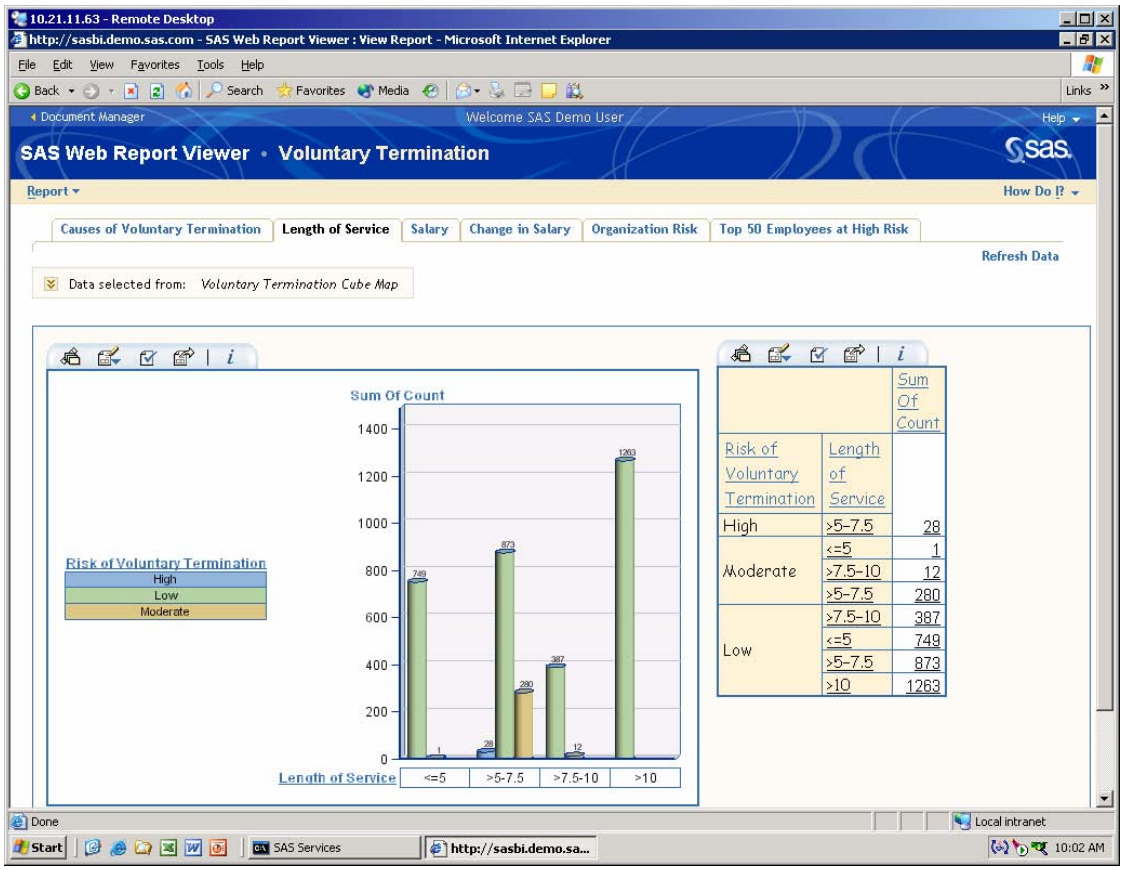

Figure 2. Profile for employee risk groups

Organizations can also build the same profile for low-risk employees and compare the two profiles to gain insight. The profiles can be refined to look at high-performing employees. By comparing the low-risk profile and the high-risk profiles, decisions can be driven by data that reflects what is actually occurring in the organization.

Employee attributes change over time and that causes their retention scores to change over time. For example, Lynn is in the high-risk group in January because of her salary level. In March, she receives a promotion, which gives her a score that puts her in the low-risk group. During the year employees move in and out of the risk groups, but organizations need to measure the overall risk. Predictive metrics can be developed with defined threshold levels that will alert managers when the overall level becomes a risk to the organization. These metrics can be developed to look at the organization in specific ways using the profile developed for high-risk employees.

BUILDING THE PREDICTIVE MODEL

You can use several methods to build predictive models. In this paper, we use SAS® Enterprise Miner™, which captures each step of a predictive modeling project. These steps include:

- Identify and connect to data sources.
- Perform necessary data transformations.
- Use a variety of algorithms to create predictive models.
- Evaluate the performance of the predictive models.
- Score new data to produce retention scores.

DATA SOURCES

The primary data source for building the predictive model will be the HR management system. The HR management system contains data that captures several aspects that influence an employee's decision to voluntarily terminate. Examples of the data available include:

- Demographic data
- Organizational data (the employee's place in the organizational hierarchy, and attributes about that place in the organization)
- Job title, skills, and experience
- Job performance data
- Compensation and benefits
- Job history (including employee-initiated changes)

The one requirement to build the predictive model is historical data that includes records for employees that have voluntarily terminated. This data is referred to as the *training data* for the predictive model. The training data provides past examples from which to learn when building the model.

DATA STRUCTURE

The preferred structure for predictive modeling is to have one record of data for each subject that you are observing. For our example, the employee is the subject. Our input data will have one record for each employee. The input data will have variables containing the values for each employee.

One required variable is an indicator if the employee has voluntarily terminated. This is referred to as the *target* variable. When you are building the model, the other variables are measured against the target variable to determine their influence on the target outcome (whether the employee voluntarily terminated or not).

An important decision that we made during the design of this analysis was to try to predict the likelihood that an employee would terminate in a future time period, instead of trying to predict the date an employee would terminate. Trying to predict the date an employee would terminate requires a much more complex analysis. The analysis would also require input data that is more rich and complete than is typically found in HR management systems. For the analysis in this example, we build a model to predict how likely an employee is to terminate in the next year. Previous project experience has shown that this approach satisfies the objective of the project, which is to identify individuals at high risk of voluntary termination while actions can still be taken to retain them. Additionally this approach uses a less complex analysis suitable for the data typically found in HR management systems.

Another factor in the design of the data structure for the analysis is the recognition that actions or changes from previous time periods will influence the likelihood for an employee to terminate in a future time period. The data structure we use includes variables for the current time period and previous time periods. An example would be to

include data for the percentage salary increase for the current year and the previous two years. Each salary increase would go into a separate variable. The resulting table would be similar to this:

Employee ID	Current salary increase	Previous year salary increase	Previous year 2 salary increase	Voluntary termination indicator
5429	4.3%	5.3%	3.7%	N
9352	1.2%	0.6%	5.2%	Y
3825	6.2%	4.8%	3.9%	N

This data structure allows us to have one record for each employee, which is the preferable format for building predictive models. The input data also contains an indicator variable of previous voluntary terminations that we can use as a target variable when building the model. By including variables with previous year's data, we can build a model that recognizes the potential influence of previous data on voluntary termination decisions in the current period.

BUILDING THE MODEL

Now that we have the data structured properly, we can use SAS® Enterprise Miner™ to build the model. The project flow for this example (Figure 3) contains these steps:

1. Identify the data source and designate the target variable.
2. Perform data transformations.
3. Partition the data for training, validation, and testing.
4. Model the data using three different algorithms (decision tree, regression, and neural network).
5. Compare the models and select the model with the best performance.
6. Score new data to predict the likelihood of voluntary termination for current employees.

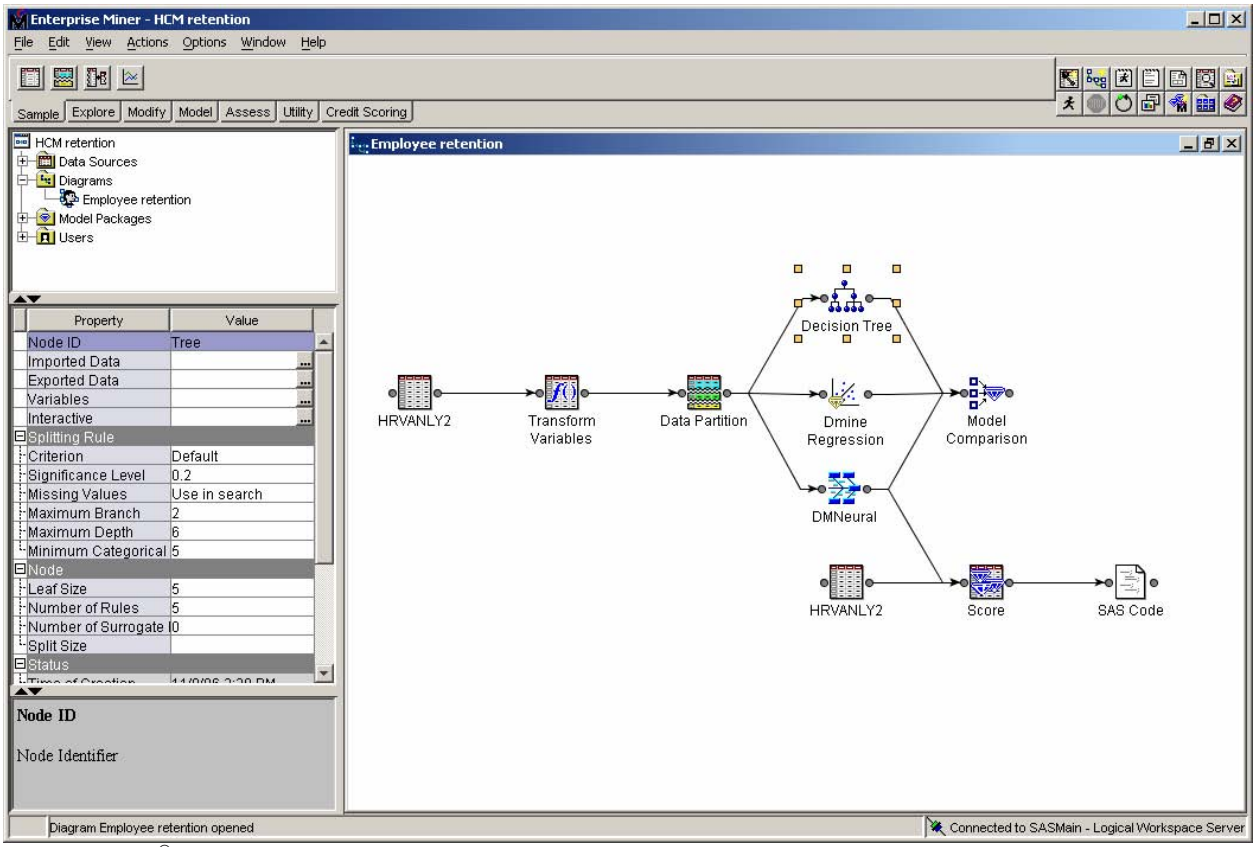

Figure 3. SAS® Enterprise Miner™ project flow

The project flow starts with the input data source, which is structured as described above. The variables in the input data source are given roles in the project. One role mentioned earlier is the target variable. For our example, the target variable is an indicator flag variable with values **1** or **0**. The variable is given a value of **1** on records for employees that have previously voluntarily terminated (recall that the input data is historical data). For employees that have not voluntarily terminated (either currently employed or non-voluntary terminations), the flag value is given a value of **0**. Additional variable roles include ID variables, model-input variables, and date/time variables. Some examples of model-input variables from our HR management system would be:

- Job title
- Department
- Length of service (at current position and total time at the company)
- Age at hire
- Indicator of recent position change
- Current salary
- Salary position with job pay range
- Amount of previous increases
- Previous job performance rating

Within SAS® Enterprise Miner™, you can also explore the data distributions in the input data source (Figure 4).

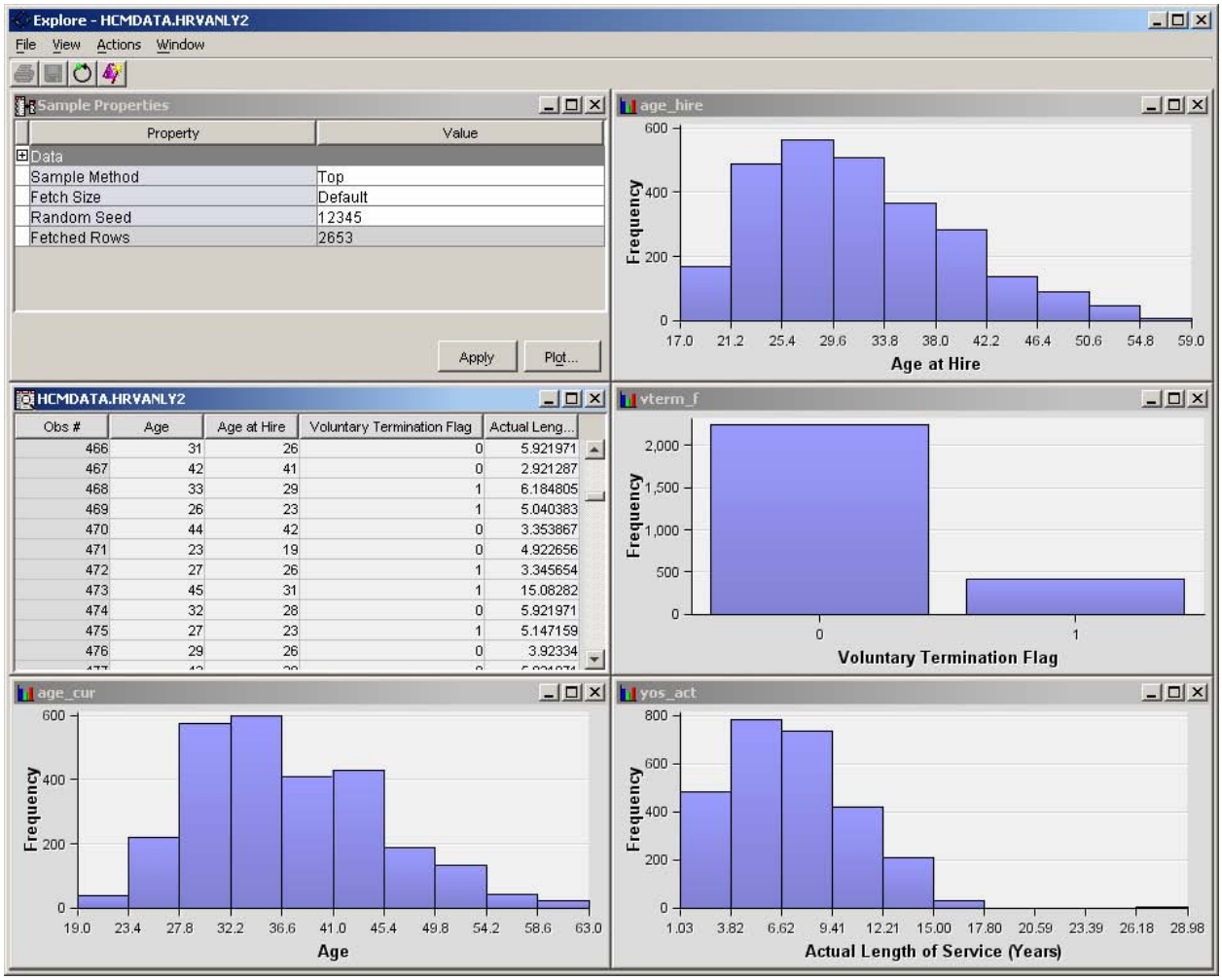

Figure 4. Data distributions from input data source

After identifying the input data source, our next step in the project flow is to prepare the data. Data transformations are applied at this point. In some cases, a calculated variable is more helpful as a model input than the variables in the input data source. One example is a length-of-service calculation. The input data source will probably have a

hire-date variable. This hire date can be used in a calculation with the current date to determine the length of service, which is more helpful in matching employees with similar characteristics.

At this point in the project flow, we can begin building predictive models. SAS® Enterprise Miner™ has pre-built modeling tasks for common algorithms such as regression, decision tree, and neural network. A common practice is to use all of these algorithms to build a model and evaluate which algorithm performs the best. The pre-built task already contains good default settings for adjustments on the algorithms. As each modeling node is connected to the input data, the default settings are displayed on the left side of the project (Figure 5). We can use these settings as is, or adjust them to achieve better model performance.

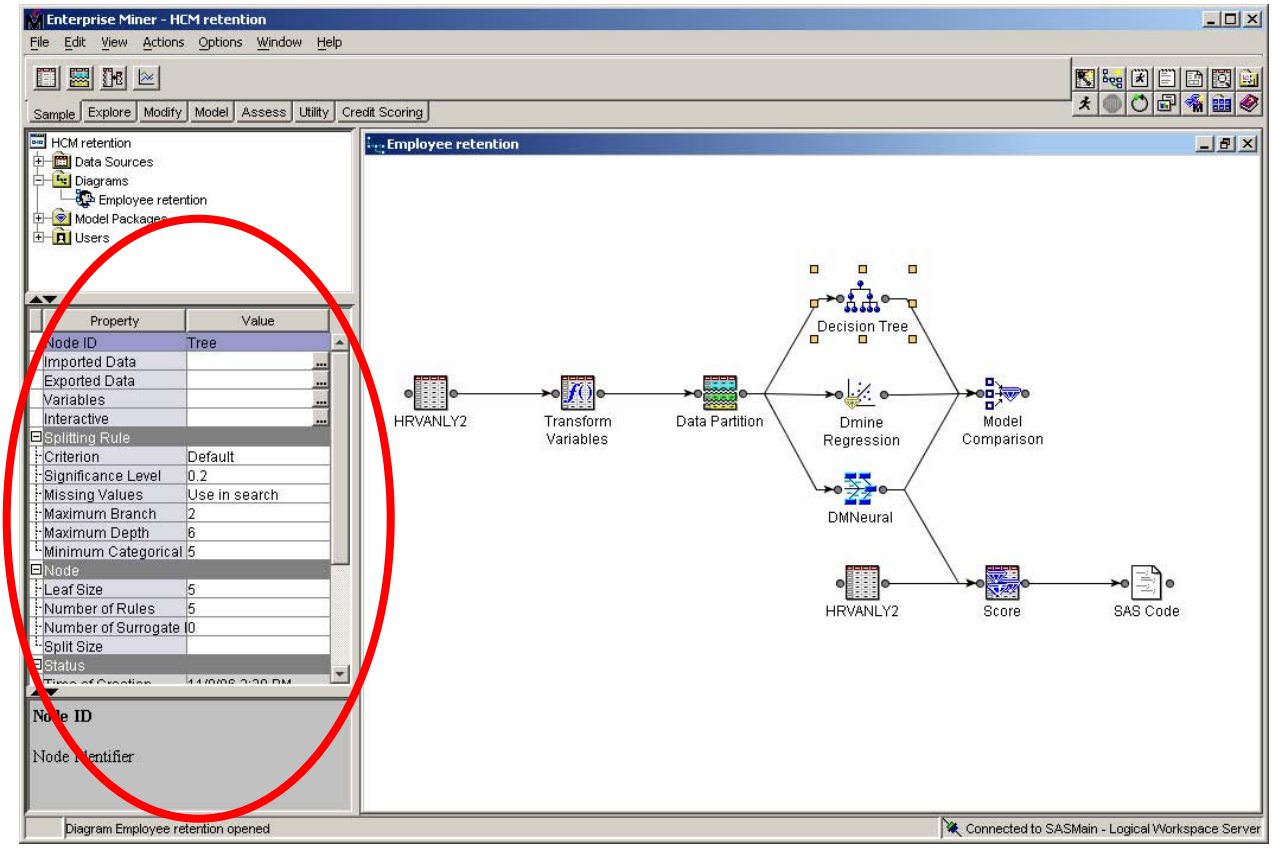

Figure 5. Default model settings.

As we run each model algorithm on the input data source, diagnostic information is available to assess the model performance (Figure 6). Diagnostic information is provided for both the training data and the validation data. The diagnostic information includes:

- A description of the model that was built. The description will vary depending on the type of modeling algorithm. For a regression model, the description includes a list of effects chosen for the model. For a decision tree, the tree can be displayed in a graphical format (Figure 7).
- Statistics on how well the model fits the input data source. Which statistics are displayed depends on the type of the modeling algorithm.
- Expected model performance measures such as cumulative lift. These measures are displayed for both the training data and the validation data to allow comparisons for consistency.

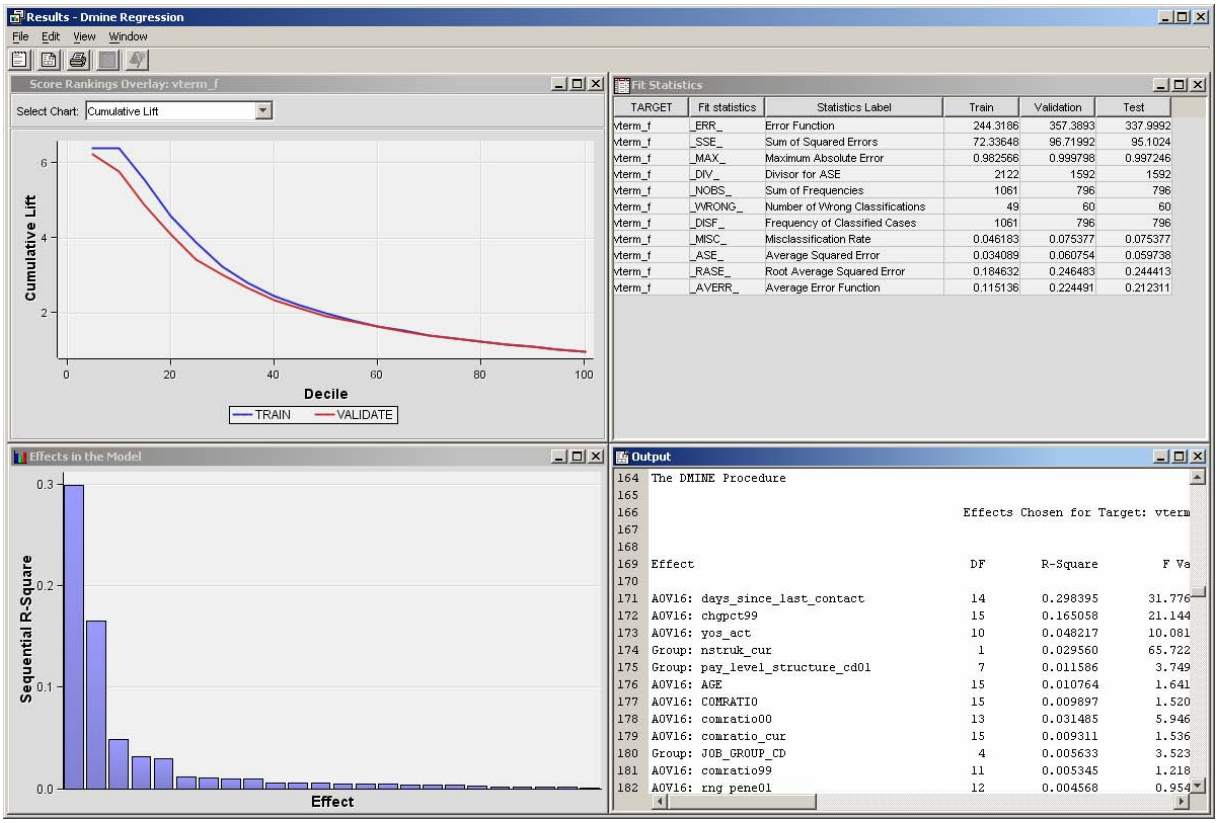

Figure 6. Diagnostic information for a regression model

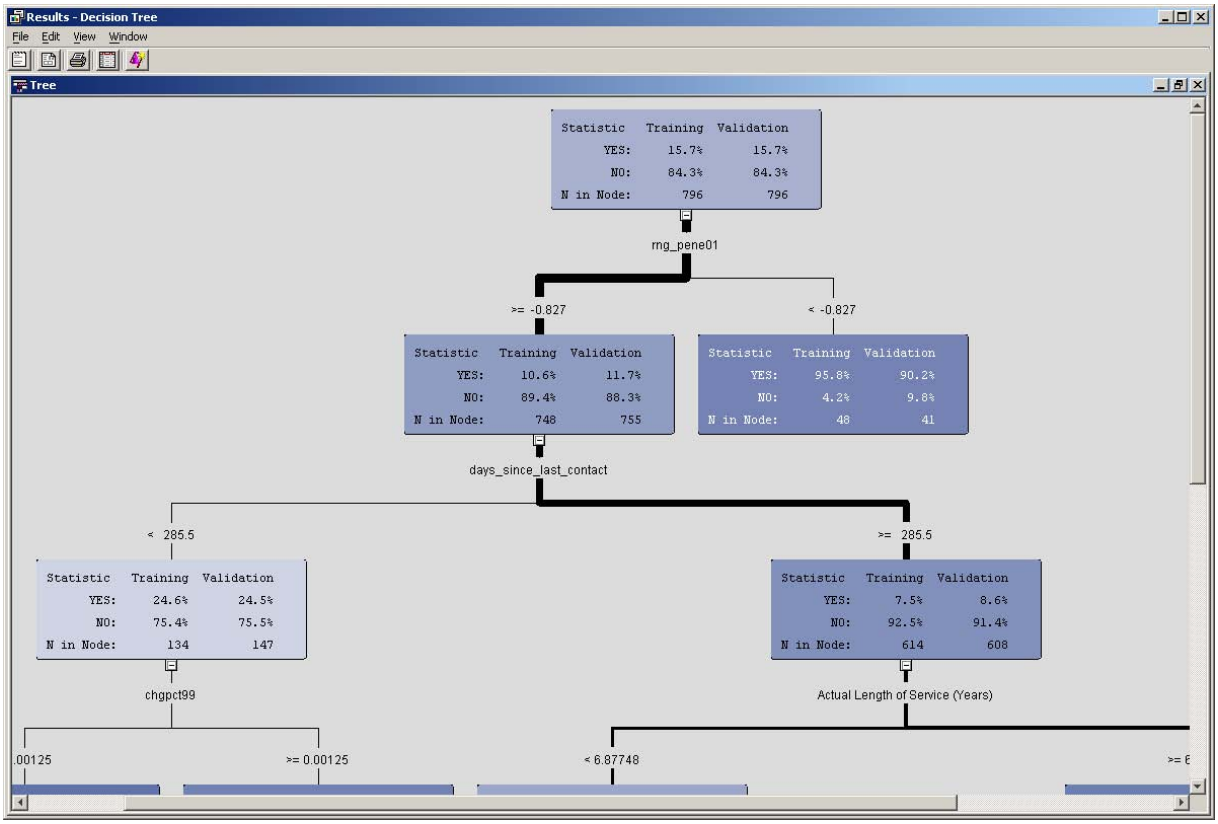

Figure 7. Graphical depiction of a decision tree

After building models of the various types, we can compare them to determine which model has the best performance (Figure 8).

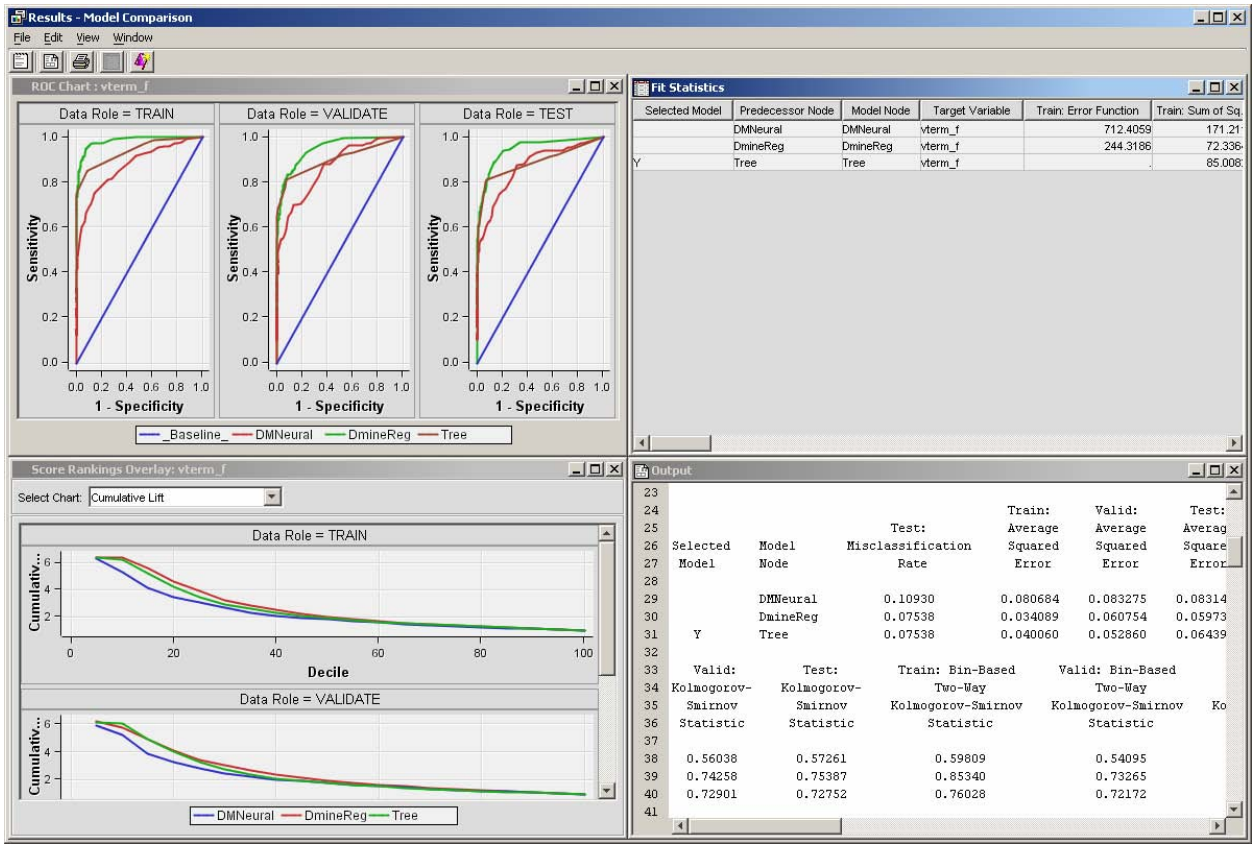

Figure 8. Comparison of model performance

USING THE MODEL TO CREATE NEW PREDICTIVE SCORES

Once the model is built, the main benefit is to use the model to create predictive scores on new data that indicate how likely an employee is to voluntarily terminate. One item created after the model is built is a scoring function. We can apply this function to new data to create a predictive score for that data. In our case, the new data is the current employee file (recall that the model was built on historical data). The predictive score will be a value between **0** and **1** that indicates how likely an employee is to voluntarily terminate in the next year. A score near **0** indicates the employee has little likelihood of voluntarily terminating. A higher score suggests a higher likelihood of voluntary termination.

The predictive score is assigned for every employee in our current employee file. While the score itself is helpful in learning which individuals are most likely to voluntarily terminate, some additional interpretation of the score allows us to provide the analysis discussed earlier in this paper. In particular, how do we know which scores are *high risk* and which scores are *medium risk* or *low risk*? The first step of this interpretation is plotting the distribution of the scores across the entire employee file (Figure 9).

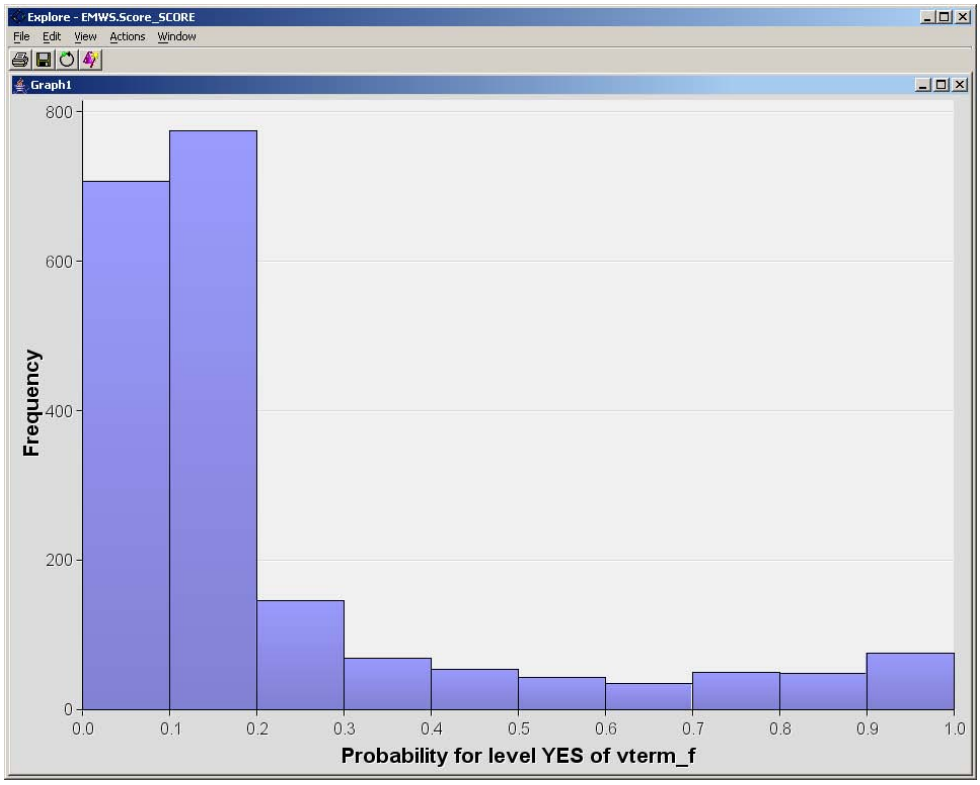

Figure 9. Distribution of predictive scores

We see that most employees have scores on the low end of the scale (less than 0.2). This is good because it indicates only a small portion of our employees have scores that indicate a high likelihood of voluntary termination. As shown by the analysis earlier in the paper, it is beneficial to categorize employees into high, medium, and low risk. Based on the distribution of scores shown above, a good set of criteria for these categories would be the following:

- High-risk employees have a score of 0.25 or higher. Although this category covers a large portion of the range of scores, it represents a relatively small group of employees whose scores are significantly above the normal score.
- Medium-risk employees have a score between 0.10 and 0.25. This represents a larger number of employees whose risk is on the high side of the distribution.
- Low risk employees have a score less that 0.10. This represents the employees on the low side of the distribution.

Assigning each employee to a risk group based on these criteria enables us to more easily interpret the predictive score and identify areas in the company that have a large number of high-risk or low-risk employees.

LIMITATIONS OF THE PREDICTIVE MODEL

There are limitations to the predictive model. One primary limitation is that the model is built using data that is available internally within the company. This data will capture a large portion of an employee's work experience, which influences how likely they are to voluntarily terminate. The data doesn't directly capture influences from outside the company that might cause an employee to voluntarily terminate. One example is an employee who terminates to move to a new city for family reasons.

One category of external data that you might be able to acquire is data related to economic conditions or job-market conditions. You will need to find a source for this data, but it can be added to the data from the internal HR management system. The data can even be acquired for employee categories. An example would be job-market data by job skill or geographic region.

CONCLUSION

Retaining high-performing employees, especially those with critical skills, is a major concern of many organizations. The loss of these employees' skills can put the organization at risk for not meeting strategic goals and can have an impact on the bottom line. Additionally, there are significant costs in replacing employees, which have been well documented (Bliss, 2007). Assume that the average salary of employees in a given company is $50,000 per year. A typical cost of turnover ratio is 150% of salary. The cost of turnover then is $75,000 for each employee that leaves the company. For the mid-sized company of 1,000 employees who has a 10% annual rate of turnover, the annual cost of turnover is $7.5 million. Using the techniques from this paper will significantly reduce this cost of turnover.

REFERENCES

Bliss, William G. 2007 "Cost of Employee Turnover" Seneca, SC: Bliss & Associates, Inc. Available at http://www.blissassociates.com/html/articles/cost_of_turnover15.html.

SAS Institute, Inc. 2007. "How to predict if your best employees are about to walk out the door." Cary, NC: SAS Institute Inc. Available at http://www.sas.com/success/carige.html.

ACKNOWLEDGMENTS

The authors would like to acknowledge SAS Consulting for previous work on this subject.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Lynn Creecy
SAS Institute Inc.
500 SAS Campus Dr.
Cary, NC 27513
919-677-8000
lynn.creecy@sas.com

Brad Klenz
SAS Institute Inc.
500 SAS Campus Dr.
Cary, NC 27513
919-677-8000
brad.klenz@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.