Arrest Attrition with Analytics

Industries such as Banking and Finance, Healthcare, Hospitality and Technology are currently experiencing high attrition rates. You, as a HR manager or senior management executive, may know some apparent causes for attrition in your company. And you also know the high costs of attrition and the negative impact it has on customer facing scenarios. But how often have you felt that if you had known earlier you could have prevented one good staffer leaving? How many of us are reactive about attrition and do damage control after a separation notice has been mailed?

A certain amount of attrition is unavoidable and essentially unpredictable because you may never be able to get all the data that went into each decision. Nevertheless, usually there are some clear patterns in attrition. This goes beyond some known causes, and subjective perspectives employees offer in exit interviews. We have found that solutions that use targeted analytics can automatically derive novel and actionable insights from attrition data. These insights, along with data-driven predictive models, can be used to design effective plans for reducing attrition, improving retention, reducing attrition costs and mitigating attrition effects.
About the Authors

Girish Keshav Palshikar

Girish Keshav Palshikar is a Principal Scientist and leads the Machine Learning R&D group in Systems Research Lab (SRL), at Pune, a part of TCS Innovation Labs - TRDDC in Tata Consultancy Service Limited. His research interests are in machine learning, data mining, text mining and theoretical computer science. He has more than 60 publications in international journals and conferences.

Rajiv Srivastava

Rajiv Srivastava leads the iWork program in SRL aimed at building tools and systems related to HR analytics. iWork includes a sub-system called iRetain, which is an as-is state analysis and predictive modeling system for attrition. iRetain is being tested for various users in TCS HR.

Sachin Pawar

Sachin Pawar is part of the iWork program. He has been working on building predictive models for attrition, among other things.
Table of Contents

1. Attrition 4
2. Importance of Attrition 4
3. Attrition Costs 5
4. Analysis of Attrition 5
5. Root-causes for Attrition 6
6. Retention Strategies 7
7. Understanding As-Is State of Attrition 8
8. Predictive Models 9
9. A Case Study 10
10. Comparison of Customer Churn and Attrition 11
    10.1 The Customer Churn Problem 11
    10.2 Comparison 13
11. Conclusions 13
**Attrition**

Attrition (also called employee churn) is a serious issue for all organizations, but particularly for high-tech industries (IT, telecom, manufacturing), service organizations (banking, finance, insurance) and support organizations (service desks, call centers, BPO). Regular attrition is often distinguished from infant attrition, where an employee leaves within, say, six months of joining.

While managing and controlling attrition is an important responsibility of HR, attrition affects other organizational functions as well. However, the highest impact and effects of attrition are felt in the business functions of the organization: services, projects, production, delivery and so on. As illustrated in Figure 1, understanding attrition is important across all functions of HR, such as talent acquisition and talent management.

![Figure 1: Attrition affects HR, other organizational functions as well as business and delivery.](image-url)

**Impact of Attrition**

Losing an employee is a problem for various reasons, such as:

- It is difficult to find suitable replacements for lost employees, particularly those with high experience and special skills.
- It takes time, effort and money to recruit new employees having required skills and experience, to train them and to help them reach the levels of performance and quality of work, which are comparable to the lost employees.
- Loss of an employee adversely affects ongoing projects and services, which leads to dissatisfaction among customers and other stakeholders.
- It takes time and effort for new employees to achieve the same levels of expertise and productivity.
- Loss of an employee costs money, which is spent in recruitment, training and salaries for the new employees, which are often higher than what the lost employees were getting.
- Employee churn rates can be as high as 12–15% annually. Moreover, attrition happens throughout the year. Thus, high churn organizations are continuously engaged in “fire-fighting” attrition.
Attrition Cost – a ballpark figure

It is common knowledge in the HR domain that losing good people is expensive, because of high costs (and long time periods) involved in finding appropriate people to replace lost employees. Here is a hypothetical example to get a rough idea of the attrition costs: Assuming an average cost of $10,000 to replace an employee and assuming 5000 cases of attrition per year leads to a total attrition-related expenditure of $50 million per annum. Reducing attrition by 20% will lead to a saving of about $10 million, in addition to improved levels of customer satisfaction and project success rates.

Analysis of Attrition

A certain amount of attrition is unavoidable and essentially unpredictable (random) because of the limitations of the data (many factors relevant to attrition may not be measured) as well as human factors involved in attrition. Nevertheless, usually there are some clearly visible broad patterns in the attrition data, when seen over a reasonable time period. Further, there are often well-understood root causes that lead to attrition. This opens up the possibilities of designing targeted analytics to automatically derive novel and actionable insights from attrition data. For example:

(a) Systematically deriving an understanding of the “as-is” state of the attrition phenomenon faced by the organization
(b) Building data-driven predictive models for attrition, so as to predict future attrition over time, both at aggregate levels (for example, location-wise, designation-wise and so on) as well as for identifying individuals with high risk of attrition
(c) Identifying the most-likely root-cause(s) for each predicted case (instance) of attrition, so that appropriate remedial actions and incentives can be chosen to retain the employee
(d) Preparing an optimal retention plan for each employee predicted as having a high risk of attrition
(e) Preparing a comprehensive and optimal attrition-handling plan, to mitigate the after-effects of predicted attrition. Such a plan includes recruitment of suitable replacement candidates for lost (or high attrition risk) employees.

This document discusses the phenomenon of attrition and focuses more on understanding and predicting attrition (tasks (a), (b)) and less on retention (task (c), (d) and (e)).

Understanding attrition, particularly by combining domain-knowledge with novel data-driven insights into attrition derived from targeted analytics, can help HR executives in several ways. It can help to:

(i) Reduce attrition;
(ii) Retain good employees;
(iii) Reduce costs of attrition;
(iv) Make HR processes sensitive and pro-active towards attrition factors
(v) Reduce after-effects of attrition (for example, maintain team quality; minimize impact on projects and client satisfaction)
Fortunately, attrition is predictable to a large extent, though there always remain some “unpredictable” and even “unexplained” cases of attrition, due to the complex human factors involved. A number of statistical and machine learning techniques are now available as part of Business Intelligence (BI) packages, to automatically discover a predictive model from past data. These techniques can be used to discover a predictive attrition model from detailed historical data about both resigned and non-resigned employees. Effective use of these BI techniques to build accurate attrition prediction models require analytics expertise and extensive experimentation, for tasks such as data cleaning, feature selection, dimensionality reduction, exploratory analysis, algorithm selection, parameter tuning and use of domain knowledge. Once learned, this model can be used to predict (for example, for the next quarter) individual employees with high risk of attrition. Having a highly accurate predictive attrition model at their disposal can help the HR in the above tasks. For example, HR executives can proactively identify and address issues related to candidates predicted by the model as having high propensity (or “red flags”) for attrition. They can also work out an individual-specific retention plan for such employees.

Alternatively, advanced data mining techniques, such as subgroup discovery can identify logically related groups of employees having unusually high risk of attrition. A hypothetical example of such an interesting group of employees might be: “candidates with BE and MBA degrees AND age between 28-30 AND gender = male AND number of job changes > 3 AND major_skill = Microsoft_Technologies”. This group of 170 employees might show an attrition of 19% whereas the overall attrition levels might be 7%. One may design target analytics to also identify top root-causes of attrition for such a group of employees: For example, the top root-causes might be related to (a) role; (b) salary or (c) supervisor. Then HR executives can design an optimal retention plan to reduce the attrition levels within this specific group, by taking into account the root-causes of attrition in it.

Insights derived from attrition data can also help during recruitment:

- Prefer recruiting candidates with high propensity for stability (“green flags for attrition”)
- Identify and address (with remedial measures) candidates with high propensity for attrition (“red flags for attrition”). Such candidates are likely to have issues that may escalate to attrition in near future.
- One can design targeted analytics to identify these kinds of insights from historical data.

**Root-causes for Attrition**

As part of the domain and organization-specific knowledge, a large variety of reasons (also called root-causes) are known to HR executives regarding why employees churn. These known root-causes can be divided into broad categories such as personal, work-related, facilities-related and supervisor-related (as shown in Figure2). There could be other ways of classifying these root-causes. Also, there is a need to include a catch-all root-cause such as “unknown” to explain all other cases of attrition. HR executives often assign one (or sometimes more than one) of these known root-causes to “explain” why a particular employee has left. While exit interviews can provide an idea of the correct root-cause of why an employee is leaving, there is a need to gather additional evidence (including from analysis of past work history) that corroborates or refutes the root-cause stated by the employee when leaving. HR executives can also use other means, such as interviews with colleagues and supervisor, to gather more evidence.
Thus an important task for HR analytics is to infer the most likely root-cause for each case of attrition using that particular employee’s work history. In the simplest case, each root-cause can be thought of as a binary variable, which either applies (“explains”) or does not apply (“does not explain”) to a specific case of attrition. The critical issue is how can we check and decide whether a particular root-cause applies or does not apply in a specific given case of attrition. It should be noted that, as often happens for human factors, one can also assign a degree of strength to each root-cause for explaining the reasons in a particular case of attrition.

**Retention Strategies**

When an employee formally indicates a desire to resign, the HR executives can choose from many different retention strategies to prevent the resignation of that employee. They can offer:

- A higher salary
- A promotion
- Overseas deputation
- Transfer to a location of choice
- A change of role, or one with higher responsibility
- Financial assistance or loans
- A change of project
- Training and other competency building initiatives
- Redressal to any specific grievances

![Diagram showing typical reasons for attrition](image-url)
Apart from such targeted (individual-specific) retention efforts, there are other, broader means at the disposal of HR to reduce impact of attrition. Some of these are:

- Pro-active identification (in advance) of employees at high risk of attrition
- Training and deployment for “back-up” team members for critical tasks and core employees
- Improved and effective knowledge transfer mechanisms
- Creation and implementation of a succession plan for leadership positions etc.
- Preparation of a comprehensive and optimal attrition-handling plan, to mitigate the after-effects of predicted attrition.

**Understanding As-Is State of Attrition**

Various key performance index (KPI) parameters can be defined to summarize information about the current state of attrition. The KPI parameters are defined for a particular time period $T$ (like a quarter or a year) and also either for the overall organization or at a finer level such as for a specific geography, business unit, and designation or experience band. The most basic KPI parameter is attrition count $\Delta(T)$, which is the number of employees who submitted their resignation during a given time period $T$ (even though the last day of work may fall after $T$ for some of them). The other basic KPI parameter is attrition rate, which is defined as $a(T) = \frac{\Delta(T)}{N(T)}$, where $\Delta(T)$ = the attrition count for the time period $T$ and $N(T)$ = the number of currently working (non-resigned) employees at the start of $T$. Many other KPI parameters can be defined (Fig. 3). A similar set of KPI parameters can be defined for monitoring infant attrition. An Attrition Dashboard usually displays various charts to depict the evolution of each KPI parameter over time, across locations etc.

An exploratory slice-and-dice analysis is also often done manually to drill-down into attrition levels in specific subsets of employees (defined in terms of combinations of attributes and values). Such analysis can capture easy-to-understand reports and charts that summarize attrition at specific organizational levels: location-wise, business-unit-wise, designation-wise, age-wise, education-wise, experience-wise, gender-wise and more.

Several kinds of statistical analyses of attrition data, such as correlation, regression, association and dependency analysis, can be used to discover useful and novel insights. Time-series techniques can help in understanding temporal characteristics of attrition (for example, trends, seasonality, peaks, cross-correlations) as well as in building predictive models.
Predictive models are now widely used in many business applications to predict occurrences of events. The problem of supervised learning consists of learning (discovering) classification rules by generalizing from given labeled training examples. This supervised learning problem is stated informally as follows: Given a set of records, each labeled with one of $k$ distinct class labels, learn(discover) a function $f$ which classifies every (new) record into one of the $k$ classes. The set of labeled records is called training data. The new records for which the class label is to be predicted are usually not part of the training data; they are called test data. The method used to find function $f$ is referred to as a learning algorithm. Several learning algorithms are known: decision tree, logistic regression, neural network, nearest neighbor, random forests, naïve Bayes, support vector machines (SVM) and so on. Most of these learning algorithms are non-parametric in the sense that they do not assume any particular distribution for attribute values. Since attrition data is typically dynamic (that is, varies over time), statistical techniques such as time-series modeling and survival analysis may also be useful.

In the present context, given a set of employee records labeled with churned or not churned, the supervised learning task is to discover a rule which can then be used to classify a new employee record into one of the two categories. The learned function $f$ is thus a predictive model. Validation of the learned function is carried out using several techniques, including cross-validation and the use of a separate validation data set. In cross validation, the training data is split into $m$ equal sets and function $f$ is learned using $m-1$ sets as training data and effectiveness is tested against the set which was left out. The overall accuracy is the average of the accuracy seen in these $m$ experiments.

The performance accuracy of the classification function $f$ learned from training data is typically measured using three quantities precision $P$, recall $R$ and $F$-measure $F$ computed on its predictions on the test data. A particular prediction of churn made by $f$ is called true positive (TP) if that employee has indeed churned in the test data; otherwise it is a false positive (FP). A particular prediction of no-churn made by $f$ is called true negative (TN) if that employee has indeed not churned in the test data; otherwise it is a false negative (FN). Precision, recall and $F$-measure are defined as follows:

$$P = \frac{\#TP}{\#TP + \#FP}$$
$$R = \frac{\#TP}{\#TP + \#FN}$$
$$F = \frac{2PR}{P + R}$$
Precision $P$ indicates the fraction of the predicted churned employees entities that are correct. Recall $R$ indicates the fraction of the actual churned employees that are correctly predicted. Low precision indicates high value of #FP, that is, lot of errors in prediction of churned employees. Low recall indicates high value of #FN, that is, lots of actually churned employees are predicted as non-churned. The F-measure computes an overall accuracy by combining precision and recall. There are other ways of measuring the performance of a predictive model, though $P$, $R$ and $F$ are most commonly used.

One crucial aspect is related to the quality and availability of the various kinds of data that is used for attrition analysis, particularly for building predictive attrition models. Sometimes the given data contains wrong or missing values. Extensive validation checks should be made on the given data to detect erroneous values and possibly replacing them with sensible values. Missing values are typically handled by replacing them with sensible values derived using imputation techniques (for example, average or median value in case of a numeric attribute). Moreover, there is often a need to add new attributes, which are derived from the original “raw” data; for example, given the last and current year’s performance rating, we can add a derived attribute that indicates the sign and extent of the change in rating. To some extent, the derived attributes help in encoding our “domain knowledge”. We need to take care of ethical and privacy issues as well as legal and policy compliance issues when selecting the attributes that are used for attrition data; for example, such issues will arise if attributes related to medical events were used.

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal</td>
<td>Age, Gender, Marital Status, number of dependents</td>
</tr>
<tr>
<td>Education</td>
<td>Degree, Specialization, College/university, Grade, Marks, and so on</td>
</tr>
<tr>
<td>Current Project</td>
<td>Project location in India, Type of project, Role in project, Turnkey project flag, Domain, Functional area, Business unit, Business sub-unit, Client, HW platform, SW platform, SW sub-platform, Employee home location, Duration spent overseas, Total duration in project, Number of role changes in project, Promoted in last year, and so on</td>
</tr>
<tr>
<td>Past project(s)</td>
<td>Similar data as above for previous project(s)</td>
</tr>
<tr>
<td>Experience</td>
<td>Current designation, Current grade, Current role, Previous experience, TCS experience, Total experience, Number of job changes, Average job duration and so on</td>
</tr>
<tr>
<td>Performance</td>
<td>Date of last designation change, Date of last grade change, Last performance rating, current performance rating, last gross salary, current gross salary, percentile salary fit against employees in same grade, average gross salary expected in market, awards received, current appraisal disagreement raised, papers published, conferences attended, patents applied, and so on.</td>
</tr>
<tr>
<td>Leaves Related</td>
<td>Current leave balance for various types of leaves, leaves in current quarter, leaves in last quarter, and so on.</td>
</tr>
</tbody>
</table>

### A Case Study

We now discuss a small real-life case study. There were 3622 employees at the start of the 3-months training dataset; 261 (7.2%) resigned in this period and 368 new employees joined, yielding a total of 3729 employees at the end of the period. The test data consisted of the next 3 months, during which 179 (4.8%) employees resigned. We ignored the employees who joined during the test period. The predictive model identified 352 employees as resigned, out of which 72 had actually resigned. The accuracy figures
for the predictive model were: \( P = \frac{72}{352} = 20\% \), \( R = \frac{72}{179} = 40\% \) and \( F = 26\% \). For this study only a few attributes (26), both numeric as well as categorical, were used for each employee. Increasing the number of relevant attributes will offer higher prediction accuracy. We used a subgroup discovery algorithm [12] to identify several subsets of employees in the training data, which shows unusually high attrition (as compared to the global average); see Table 2. See [14] for more details on the problem of employee churn prediction.

### Table 2. Some groups of employees having unusually high attrition

<table>
<thead>
<tr>
<th>Group Descriptor</th>
<th>#Total</th>
<th>#Attrition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Designation = 'PA' AND Grade = 'O1' AND Previous_exp Between (4, 7)</td>
<td>37</td>
<td>12 (29.7%)</td>
</tr>
<tr>
<td>Project_work_type = 'J' AND Percentile_fit_against_same_grade&gt; 90 AND Previous_exp Between (4, 7)</td>
<td>31</td>
<td>13 (41.9%)</td>
</tr>
<tr>
<td>Designation = 'PA' AND Project_work_type = 'J'</td>
<td>679</td>
<td>130 (19.2%)</td>
</tr>
</tbody>
</table>

Comparisons based on information gain as well as correlations showed that attributes related to type of work and leave are (very mildly) important for prediction. It is likely that there are no subsets of attributes that are strong predictors of attrition.¹

## Comparison of Customer Churn and Attrition

We are often told that employee churn is similar to customer churn and that the same tools can be applied to analyze both issues. Customer churn and employee churn are two similar but not identical problems, though both of which are of concern for organizations. Let us explore the characteristics of customer churn.

### The Customer Churn Problem

The phenomenon of customer churn is commonly observed in volatile consumer service markets such as mobile phones, insurance, subscription services (for example, magazines) and banking. A customer "churn" is observed when the customer discontinues using the services, although the customer may continue to use other products or services. Customer churn is a serious problem for several reasons, such as:

1. Acquiring a new customer is much harder and more expensive.
2. Process-related costs for terminating a customer's service are high. For example, in 2001, customer churn rates were around 40\% for wireless telephone carriers and churn-related costs for them were around US $10 billion.
3. Losing a customer leads to loss of revenue and a negative impact on the bottom line.
4. Churned customers affect the brand-value and may influence prospective customers.

¹ We have only presented a simple case study to illustrate the value of the approach. In actual usage scenarios more attributes and better data sets are used, and prediction accuracy is high.
A voluntary churn happens when a customer voluntarily terminates the services of a company and usually switches to the services of a competitor company. An involuntary churn happens when a company on its own terminates the services of a customer, typically for non-payment of dues. The time period for which a customer stays with the company (for a specific service) is called the customer’s lifetime. Primary reasons for voluntary churn are availability of more competitors that offer better products/services, more flexibility/convenience, better pricing, better customer support and so on. In addition, availability of many competitors makes the customer switch companies for even small reasons for example, customer support related issues. Involuntary churn happens because of inadequate screening at the timing of acquiring a customer (low entry barriers), lack of flexible payment options, lack of early warnings (as with inaccurate or untimely billing) so that customers can attempt to improve the usage and so on.

Several strategies can be designed to reduce customer churn; examples for mobile phone company:

1. Improve services offered, such as wider coverage, better voice quality, cheaper call or activation rates, and innovation for competitive differentiation.
2. Offer discounts or other promotional benefits, improve billing services (accuracy, frequency), and offer more payment options.
3. Improve customer support service by improving service time and minimizing the waiting time.

There is very little one can do to make a customer change his/her mind, when the customer has already decided to churn. On the other hand, if there is an early warning (a “red flag”) for a particular customer then it may be possible to take some pro-active actions (depending on the nature of the red flag) to prevent the churn. For example, if the red flag is “customer duration <=63 months AND current monthly payment is delayed” then the customer can be offered a discount for the next payment. For this purpose, each red flag must be a strong predictor of the impending churn event. Thus there is a need for a predictive model, which can be applied to individual customers (for example, every month) to predict accurately and in time whether or not the customer will churn. The model must indicate the “red flags” that led to the prediction of that customer’s churn. Ideally, the model must also help in identifying the correct root-causes for each predicted to churn customer, even though they may not be directly available in the data (for example, as reasons stated by the customer). See Table 3 for some of the published case studies involving predictive models for customer churn.

Clearly, not all customers are equally valuable (or profitable) for a company. For example, for a telecom company, customers who make many international calls are usually more valuable than customers who make only a few local calls. Given the fact that it may not be possible to stop customer churn altogether, one area of focus then could be to prevent the churn of high-value customers. For this purpose, there is a need for an accurate customer lifetime value model, which can associate a measure of value with each customer; for example, expected profits over the customer’s lifetime with the company. Thus, a predictive churn model should also take into account the value of a customer and should deliver more accurate churn predictions for more valuable customers.
Predictive churn models often need to be supplemented with the following:

(a) Models that identify logically related groups of customers exhibiting unusually high churn levels
(b) Methods to detect significant changes in customer behavior
(c) Methods to detect new/emerging needs of the customers
(d) Methods to detect new/emerging segments of customers with distinctive needs.

Table 3. Some published work related to predictive models for customer churn

<table>
<thead>
<tr>
<th>Domain</th>
<th>Predictive modeling techniques used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile telephony</td>
<td>SVM [1], decision tree [7], lifetime value models [13], statistical [15],</td>
</tr>
<tr>
<td></td>
<td>neural network [5], evolutionary [2], neural network, logistic regression,</td>
</tr>
<tr>
<td></td>
<td>decision tree [8]</td>
</tr>
<tr>
<td>Insurance</td>
<td>SVM [10]</td>
</tr>
<tr>
<td>Subscription services</td>
<td>SVM [4]</td>
</tr>
<tr>
<td>Retail</td>
<td>Logistic regression, decision tree [3]</td>
</tr>
<tr>
<td>Banking</td>
<td>Regression [6], random forests [9]</td>
</tr>
</tbody>
</table>

**Comparison**

We observe, therefore, there are significant differences between attrition and customer churn. As a consequence, retention strategies used for customers and employees differ significantly; effective employee retention strategies tend to be much more individual-centric. Measuring the effectiveness of customer retention strategies is easier (as it is directly reflected in costs and revenues) while it is much harder for employees (a retained employee may have lower happiness levels or productivity). Further, most employees do not directly bring in any revenue; their value is in terms of productivity, the quality of work and other intangible contributions (such as leadership, knowledge-sharing, communication skills, initiative, expertise levels, team spirit to name but a few). This value is hardly ever reflected in data and hence it is harder to use in prediction and even harder to project into future. Despite these differences, the common threads in customer and employee churn are: the need to understand as-is state, need for predictive models, need for identifying root causes, need for effective retention strategies and need for an effective mitigation plan.

**Conclusions**

We believe that attrition can be largely predicted. You have to design targeted analytics to bring to light patterns that were not apparent before. Statistical and predictive modeling techniques help us effectively understand and control attrition. The methodology consisting of as-is state understanding, building a predictive model, making predictions of high attrition risk individuals along with likely root causes provides novel and actionable insights into attrition that can help you design an informed and targeted retention plan and a longer-term attrition mitigation plan. Savings to business from this will be considerable.
References


About Tata Consultancy Services (TCS)
Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India’s largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com