Paper to be presented at the 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence

Indian Institute of Management, Ahmedabad, January 8-9, 2011

Employee Attrition Risk Assessment using Logistic Regression Analysis

Rupesh Khare♠, Dimple Kaloya♣, Chandan Kumar Choudhary♥ & Gauri Gupta♦

♠ **Unit Head, Consulting Operations, Aon Hewitt, Gurgaon, India**
e-mail: rupesh.khare@hewitt.com; Tel (+91) 9810512408

♣ **Senior Project Manager, Consulting Operations, Aon Hewitt, Gurgaon, India**
e-mail: dimple.kaloya@hewitt.com; Tel (+91) 9560508561

♥ **Project Manager, Consulting Operations, Aon Hewitt, Gurgaon, India**
e-mail: chandan.kumar.choudhary@hewitt.com; Tel (+91) 9990983737

♦ **Senior Associate, Consulting Operations, Aon Hewitt, Gurgaon, India**
e-mail: gauri.gupta@hewitt.com; Tel (+91) 9818512143

Key Words: Employee Attrition, Logistic Regression, Attrition Risk model, Predictive Techniques

Table of Contents

I. ABSTRACT ---------- 3

II. INTRODUCTION ---------- 4

1. EMPLOYEE ATTRITION – A PREDICAMENT ---------- 7

2. AN OVERVIEW OF VARIOUS PREDICTIVE TECHNIQUES ---------- 9

3. A PREVIEW OF LOGISTICS REGRESSION ---------- 12

4. APPLICATION OF LOGISTICS REGRESSION ---------- 14

4.1 PROJECT BACKGROUND ---------- 14

4.2 OBJECTIVE OF THE RESEARCH ---------- 14

4.3 METHODOLOGY ---------- 15

4.3.1 DATA COLLECTION ---------- 16

4.3.2 DEVELOPMENT OF ATTRITION RISK EQUATION USING LOGISTIC REGRESSION ---------- 19

4.3.3 TEST THE ATTRITION RISK EQUATION ---------- 20

4.3.4 RETENTION PLAN ---------- 21

4.3.5 IMPLEMENTATION OF REAL TIME IT SOLUTION --- 22

III. CONCLUSION ---------- 23

IV. APPENDIX ---------- 24

A. TANGIBLE & INTANGIBLE COSTS ---------- 24

B. OVERVIEW OF BEHAVIORAL MODEL ---------- 26

C. MODEL OUTPUT ---------- 27

C.1 TESTS OF GLOBAL MODEL FIT

C.2 PARAMETER ESTIMATES

D. MODEL EQUATIONS BY DEPARTMENTS ---------- 30

V. ACKNOWLEDMENT ---------- 32

VI. REFERENCES ---------- 33

I. ABSTRACT

This paper presents the application of logistic regression technique to predict employee attrition risk in an organization based on demographic data of separated employees. The paper is based on a real life project executed with one of our clients. In this project the team used demographic information of separated as well as existing employees. This data was used to develop a risk equation, which was later applied to assess attrition risk with current set of employees. Post this assessment, high risk cluster was identified and focus group discussions were initiated to find out the reasons and their requirements and hence action plan was created to minimize the risk.

This paper details out the overall approach taken to create the attrition risk model, process flows, data streams involved and the output attained from the model. It further aims to recommend the thrust areas and best practices on employee retention at different stages of the employee's association with an organization.

II. INTRODUCTION

"…take our 20 best people and virtually overnight we become a mediocre company"

- Bill Gates

According to Businessdictionary.com, employee attrition is an *Unpredictable and uncontrollable, but normal, reduction of work force due to resignations, retirement, sickness, or death*. However there are some statistical procedures or techniques, through which employee attrition can be predicted. This paper is an attempt to show how attrition can be predicted using logistic regression thus can be controlled.

Attrition is a big problem in industries like IT, BPO and KPO etc. This problem can be attributed to dissatisfaction to various aspects of a job, for example career aspirations, work location, salary, performance management, job satisfaction, managers and many more.

Employee attrition control is critical to the long term health and success of any organization. Staffing costs are one of the largest expenses regularly charged to the budget of organizations. With salaries, benefits, bonuses, training and other personnel costs, companies invest a great deal of resources in their employees. To reduce the cost of attrition, organizations need to ensure that employees' aspirations are met. It is a known fact that retaining the best employees ensures customer satisfaction, increased revenues and satisfied colleagues and staff.

Organizations invest a lot of money on training, giving employees onsite opportunity, offering compensations above market level to retain employees. However, currently these methods are being generically applied in order to control employee attrition. Through this paper, the authors introduce the concept of Logistic Regression as a technique of predicting attrition risk attached with each employee and highlights the importance of attrition risk assessment using predictive technique.

Application of Logistic Regression enables organizations to employ a more targeted approach towards their employee retention strategies. The following diagram illustrates the different dimensions for the application of the model.

Business Performance	Decision Making
▪ Quantify ROI of talent investments ▪ Benchmark ability to retain pivotal talent versus peers ▪ Provide input to long term business planning and investments	▪ Identify workforce segments/ individuals most at risk of leaving ▪ Determine targeted actions to take to improve retention and engagement ▪ Develop business case for talent investment with a more complete cost/benefit rationale
Workforce Planning	**Talent Management and development**
▪ Set practical targets for talent retention across business and specific workforce segments ▪ Predict future turnover within specific workforce segments ▪ Track pivotal retention across business units	▪ Hold managers accountable for talent retention and engagement ▪ Better manage pivotal employees and critical workforce segments; e.g. input to compensation design, succession planning, diversity efforts, etc.

The Logistic regression modeling technique also finds applicability in various other risk prediction analyses, like; churn risk attached with credit card users, subscribers of television channels or magazines, infant survival probability etc.

The first section of the paper emphasizes the fact that employee attrition is a predicament for organizations across the globe. The second section scopes and compares various statistical predictive techniques available, and summarizes Logistic Regression as a relevant technique to apply in a real life project. The third section gives an overview of the concept of Logistic Regression by explaining the statistics of the model. Subsequently, the fourth and the last section present the background of a real life attrition risk assessment project and underscore the application of Logistic Regression, along with the results and recommendations from the study.

1.0 EMPLOYEE ATTRITION – A predicament for all organizations

Employees who enjoy the work and the work environment are more likely to remain employed with their company. Retention strategies are important because they help create a positive work environment and strengthen an employee's commitment to the organization.

Creating retention strategies in today's world is a difficult task as organizations continue to struggle with high attrition rates. There are high turnover costs associated with attrition which ultimately impact the bottom line of all organizations. The turnover costs are generally classified under two main heads: tangible and intangible. The tangibles can further be classified as recruitment cost, employee's salary and benefits during the training process, the cost of advertising for the position (in the newspaper, or via the Web training), the cost of recruiting a new employee (time spent interviewing all applicants, checking references, etc.), the cost of training the employee to an acceptable level (trainer's salary, for example) and administration cost.

Intangibles on the other hand include unquantifiable factors like negative impact on employee morale, loss of knowledge, productivity etc. A detailed, but not exhaustive, list of tangible and intangible cost metrics is mentioned in Appendix A.

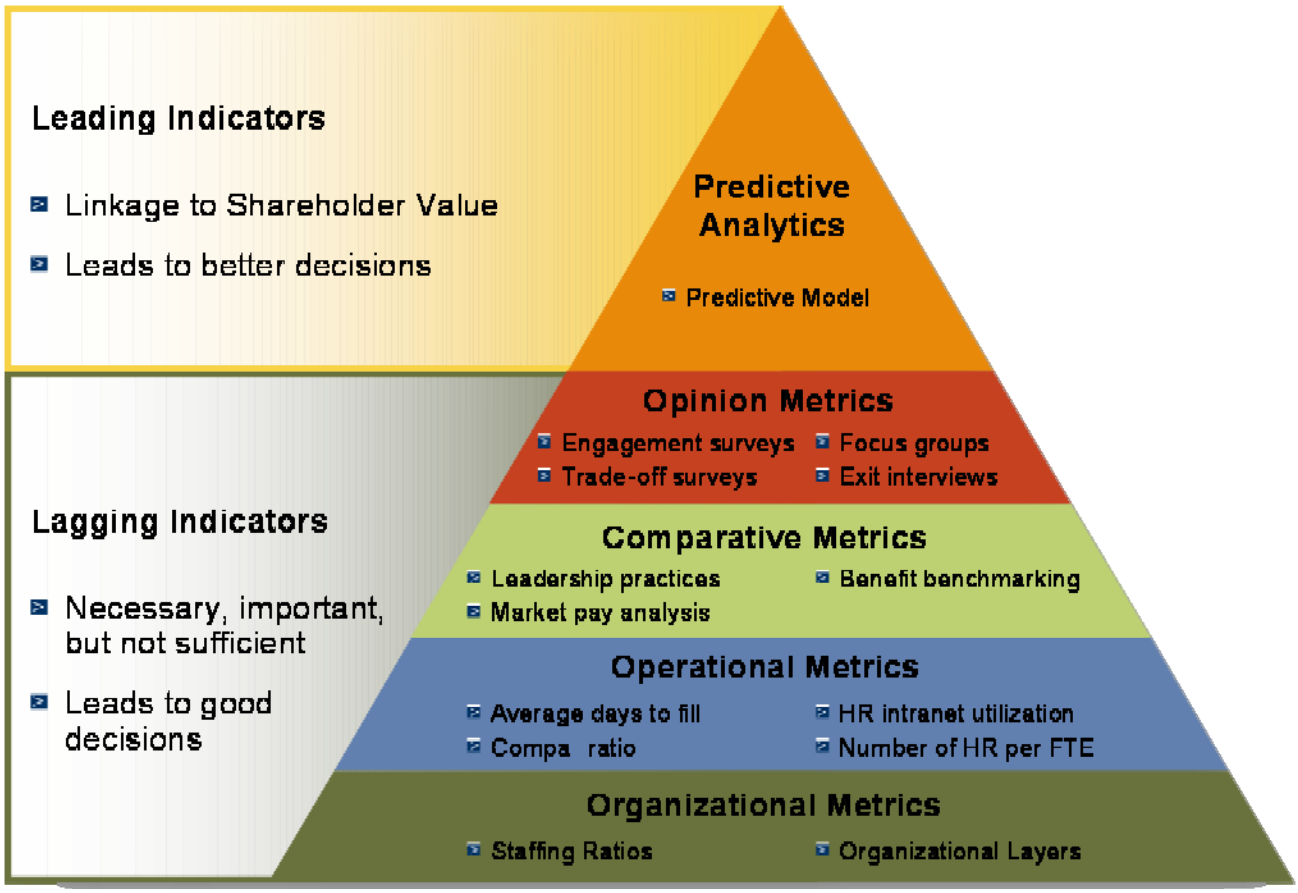
Some organizations calculate an employee turnover cost at 150 percent of the yearly salary of the existing employee or one third of a new hire's annual salary. For managerial and sales positions, the cost can go up to 200-250 percent of the annual salary of an employee[1]. Therefore, failing to retain a key employee is a costly proposition for any organization.

Understanding the nuts and bolts of human capital analytics can prove to be beneficial for all organizations struggling with attrition. The chart below shows "Hierarchy of Human Capital Analytics" with leading and lagging indicators and their importance. Both of these indicators are created based on perception data, exit data, benchmarking data, engagement model, compensation surveys etc. Lagging indicators suggest solutions for problems that have occurred

[1] Source: Mahindra Special Services Group, (Curtis and Wright, 2001)

in the past; this may or may not be relevant to the current situation. Leading indicators are indicators which can predict attrition in the future using trends and patterns emerging out of historical data.

The Hierarchy of Human Capital Analytics



In most of the organizations, managing attrition is a reactive exercise, where attrition is analyzed by comparing exit interview data of employees from different units and profiles. The organizations are feeling the need of "in time" attrition risk assessment which would better equips the management to deal with attrition. Proactive prediction of attrition through predictive models will lead to improved decision making to deliver shareholder values and in turn save revenue due to loss in human capital.

2.0 AN OVERVIEW OF VARIOUS PREDICTIVE TECHNIQUES

Predictive analytics aids in extracting information from data and using it to predict future trends and behavioral patterns. Predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting it to predict future outcomes.

There are multiple statistical predictive techniques, such as ANOVA, Linear Regression, Discriminant Analysis and Logistic Regression that are used industry wide for various predictive requirements.

ANOVA (analysis of variance) is a statistical method for making simultaneous comparisons between two or more means; a statistical method that yields values that can be tested to determine whether a significant relation exists between variables.

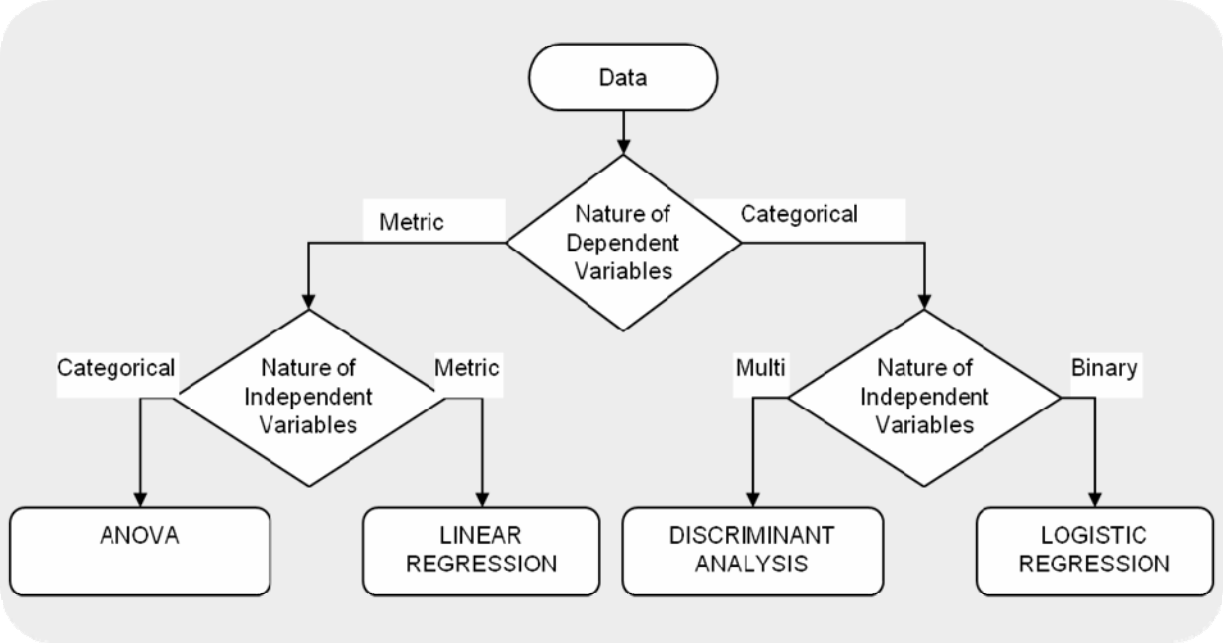
Linear Regression refers to any approach to modeling that defines the relationship between one or more variables denoted Y and one or more variables denoted X, such that the model depends linearly on the unknown parameters to be estimated from the data. Such a model is called a "linear model".

Linear Discriminant analysis (LDA) method is used in pattern recognition and to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification.

Logistic regression (sometimes called the logistic model or logit model) is used for prediction of the probability of occurrence of an event by fitting data to a logit function logistic curve. It is a generalized linear model used for binomial regression. Like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. Logistic regression is used extensively in the medical and social sciences as well as marketing

applications such as prediction of a customer's propensity to purchase a product or cease a subscription.

The flowchart below shows the relevance of each technique with respect to the nature of dependent and independent variables to be used in the model.



The chart illustrates that in techniques like ANOVA and regression analysis, the dependent variable is metric or interval scaled, whereas in Discriminant Analysis or Logistic Regression it is categorical. The independent variables are categorical in the case of ANOVA but metric in case of regression and Discriminant Analysis. Henceforth, in this section we distinguish between Discriminant Analysis and Logistic Regression as unique predictive techniques.

Logistic Regression commonly deals with the issue of how likely an observation is to belong to each group. On the other hand, Discriminant Analysis deals with the issue of which group an observation is likely to belong to.

Discriminant Analysis and Logistic Regression are widely used multivariate statistical methods for analysis of data with categorical outcome variables. While both techniques are appropriate for the development of linear classification models, Discriminant Analysis is based on more assumptions about the underlying data.

The table below presents a comparison of underlying assumptions of the two techniques:

Assumptions	Discriminant Analysis	Logistic Regression
Multivariate Normality	Yes	No
Homoscedasticity (Homogeneity of variances/covariances)	Yes	No
Non-Multicolinearity (Low Correlation between independent variables)	Yes	Yes
Absence of Outliers	Yes	Yes
Large Sample Size*	No	Yes
Predictor Variables can be Categorical, Continuous or Discrete	Yes	Yes

Above table indicates that in cases when the assumptions for Discriminant Analysis are violated, the technique should be avoided and Logistic Regression should be employed to analyze the data in order to give more robust results.

**Based on ten (10) independent variables, Discriminant analysis would need minimum of 50 cases and Logistic regression would need minimum of 250 cases.*

3.0 A PREVIEW OF LOGISTIC REGRESSION

Logistic regression predicts the outcome of a dependent variable through a set of predictors. It is applicable where dependent variable is categorical and dichotomous and independent variables are categorical, continuous or mixed. The dependent variable in logistic regression takes the value of 1 (one) with probability of success of an event, or the value of 0 (zero) with the probability of failure of an event.

Logistic Regression models the probability of 'success' as:

$$\text{logit}\left[\theta(\mathbf{x})\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

The equation above shows the relationship between, the dependent variable (success), denoted as (θ) and independent variables or predictor of event, denoted as x_i. Where α is the constant of the equation and, β is the coefficient of the predictor variables.

Minimum valid sample size required for LR Model:

Based on the work of Peduzzi et al. (1996)[15] the following guideline has been defined for a minimum number of cases to be included in Logistic Regression.

$$N = 10\ k / p$$

Where,

N = Minimum sample size required for model

k = Number of independent/Predictor variable

p = the smallest of the proportions of negative or positive cases

In Logistic Regression, each outcome can assume only two values and the procedure used to achieve this is called the maximum likelihood method.

Logistic regression modeling can be performed in two ways:

1. Stepwise Regression – A method in which independent variables are entered one by one and their significance in the model is checked
2. Backward Stepwise Regression – A method in which all independent variables are used and insignificant variables are removed stepwise in an iterative process to ensure that the model adequately fits the data

From the results of Logistic Regression modeling, the co-efficient are tested for significance using several tests like Wald Chi Square Test, Likelihood-Ratio Test, and Deviance test in order to validate the model. A Wald Chi-square test is used to test whether two (or more) variables are independent or homogeneous. The chi-square test for independence examines whether knowing the value of one variable helps to estimate the value of another variable. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model over the maximized value of the likelihood function for the simple model.

4.0 APPLICATION OF LOGISTIC REGRESSION

4.1 Project Background:

Anonymous Inc. (name changed to maintain confidentiality), a global IT company was facing a challenge in managing employee attrition. Some initiatives had been taken internally to control the attrition, however, these initiatives were not sufficient and the organization felt the need to be better equipped in order to handle situations of employee attrition.

Anonymous Inc. approached Aon Hewitt to develop a predictive model which would proactively predict attrition risk of employees, and subsequently enable it for effective decision making and management.

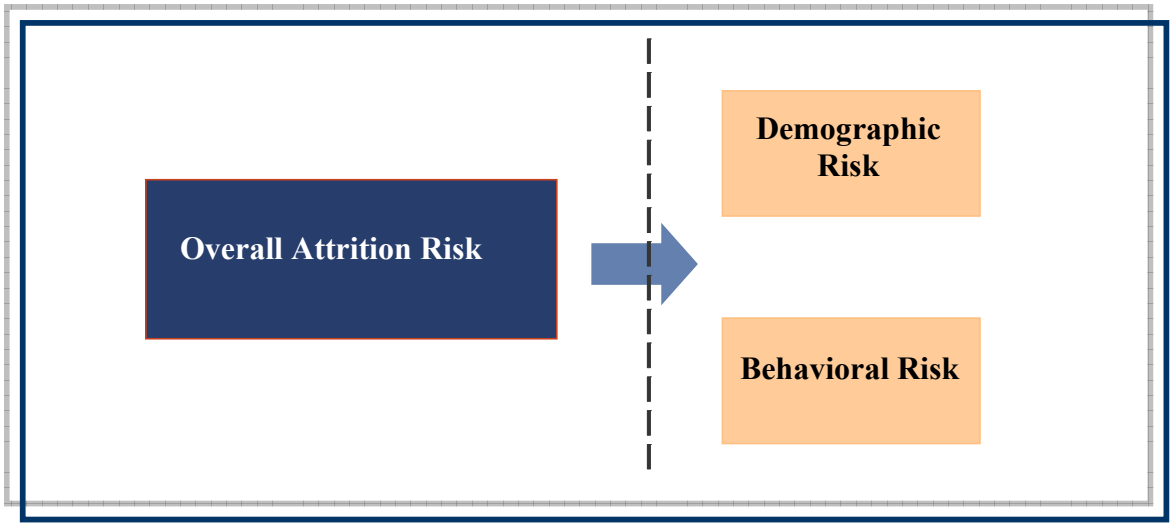
4.2 Objective of the research:

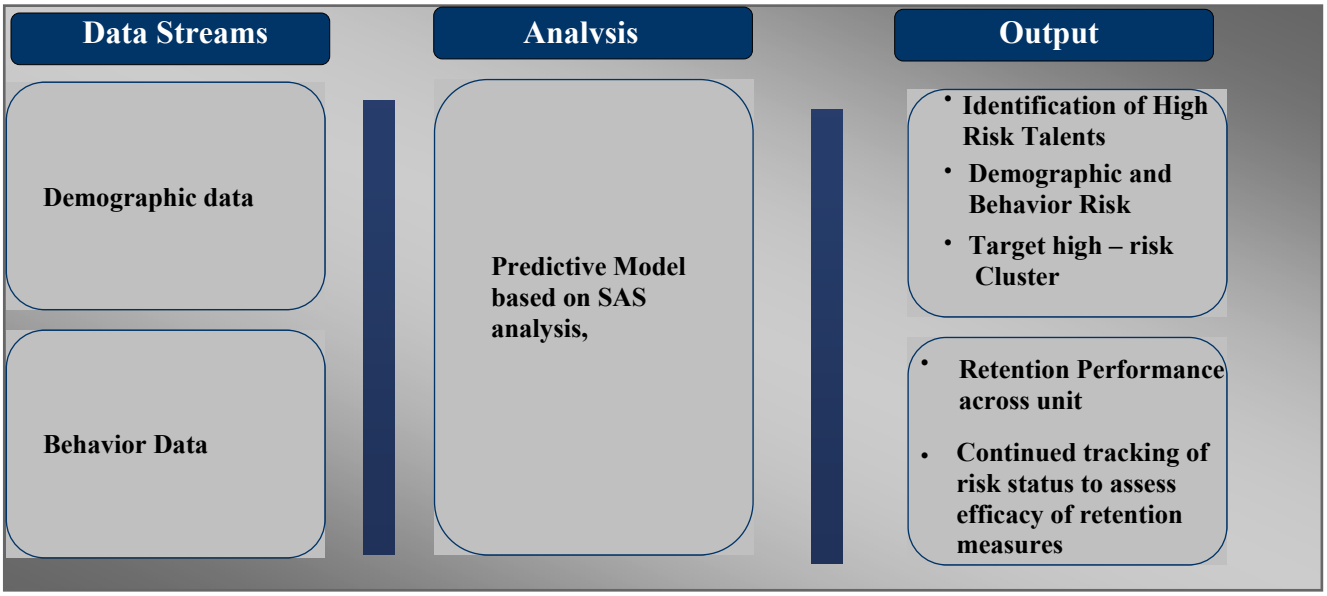
The client wanted to minimize attrition by improving its retention strategies by developing a real time solution to target high risk employees and accordingly take better decision. The research was split into specific objectives shown below:

1. Develop a model to predict employee membership towards risk categories at an overall as well as at a department level
2. Create four clusters bases on risk measured from the risk model
3. Identify and target high-risk talents
4. Charting a retention plan targeted at specific categories
5. Measure retention performance across units
6. Provide inputs to reshaping talent sourcing strategies
7. Implement a real time IT infrastructure to indicate high risk category

4.3 Methodology:

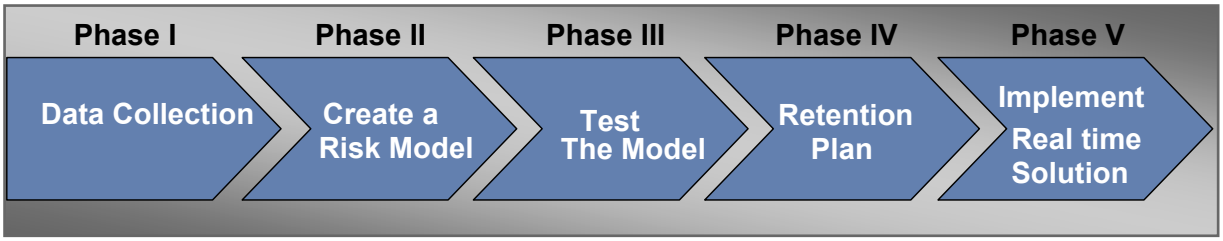
The team used two modeling techniques to predict overall attrition risk, as shown in the diagram below. Behavioral risk* modeling based on the online survey data and attrition risk modeling based on the demographic data.





**Behavioral risk model is out of the scope of this paper. However a brief overview of behavioral model is presented in Appendix B.*

Below is the process flow used for conducting and implementing attrition risk modeling.



4.3.1 Data Collection:

Data collection process was undertaken with:

- Understanding of organization structure
- Collecting of data around employee demographics, attrition, drivers of employee satisfaction and retention through various sources

4.3.1.1 Sources of Data:

The diagnostic was done through a multi-faceted approach, which involved collecting information from different sources to accurately arrive at the status of HR policies.

BUSINESS PARTNER DISCUSSIONS	Discussion with Heads(COO, GTS, Marketing)
FOCUS GROUP DISCUSSIONS	Discussion with employees at different location
ATTRITION DATA	Demographic data on separated employee
HR PROCESS ANALYSIS	Discussion with HR teams to understand the HR processes
EXIT INTERVIEW DATA	Data from Exit interviews

Logistic regression modeling is based on the attrition data only and rests of the datasets were used for other qualitative analysis of this project. Hence, for the purpose of this paper, we will refer to attrition data only.

The HR function of Anonymous Inc. maintains employee database for all current as well as previous employees with their demographic information. The project team collected the two different samples of data. One sample was utilized for running the model and creating the attrition risk equation, while the other sample was utilized for validating the model.

4.3.1.2 Sample size:

Minimum sample required for our model with 10 independent variables and approximately 40% of separated employee data can be calculated as: (As per the formula defined by Peduzzi et al. (1996)[15])

$$N = 10 * 10/(0.4) = 250$$

The details of the sample of data collected for actual research and for testing the model were much more than minimal required sample of 250:

Data Set #1:

- Separated Employees from year 2006 – 2008 (Sample Size=3271)
- Existing Employees (Sample Size=5208)

Data Set #2:

- Separated Employees from year 2006 – 2008 (Sample Size=1619)
- Existing Employees (Sample Size=1937)

While gathering data, the research team ensured that the data has good sample representation from each department as the organization had more than seventeen departments. Their research perspective was to do the analysis at the department level as well to identify the unique equations for different departments.

The team used stratified random sampling technique and significant testing tool to select the sample. Stratified random sampling was relevant in this case because sub-groups within the population were heterogeneous. Through stratification, grouping of members of the population was done to get them into relatively homogeneous subgroups before sampling. The team also ensured that there was a reasonable representation from different departments in the overall sample.

Data preparation and cleaning was done after data collection. This involved the following steps:

1. Conversion of metric data into categorical data from some demographical questions like Age, Year of service etc.
2. The missing data for performance rating was replaced with root mean squared value. Performance rating was an important variable for analyzing attrition and thus, the team could not afford to lose out on individuals for whom the performance rating was missing.
3. Data cleaning step involved cleaning of outliers, cleaning of invalid data points and removal of individuals whose information was missing.
4. A variable named 'Attrition' was created in the data set. This variable contained the option of '0' (zero) or '1' (one) depending on whether the employee was existing or separated respectively. The model treated 'Attrition' as a dependent variable while demographic variables were treated as independent variables.

The following independent demographic variables were used in the model:

- Gender
- Marital Status
- Age
- Education
- Tenure in the organization
- City
- Salary Grade
- Designation

- 2006 Performance Rating and
- 2007 Performance Rating

4.3.2 Development of Attrition Risk Equation using Logistic Regression

The team used Statistical analysis software (SAS) to run the logistic regression model. The analysis was done to find out the probability of occurrence of an event (probability of leaving or not leaving the organization) by fitting data into a logistic curve.

The analysis was done at two levels:

- Regression Modeling first done to identify the coefficients of the Master Equation, at an overall organization level
- Later, Regression Modeling done for each of the Departments

At the outset, Logistic Regression model included all demographic variables and subsequently eliminated insignificant variables through an iterative process. Wald Chi-Square test and Maximum Likelihood Estimates were used to identify coefficients for significance for inclusion or elimination from the model.

The fitment of the model was tested after each round of elimination. The analysis was concluded when no more variables needed to be eliminated from the model and the model converged. Refer to the appendix B for the output and the statistics.

Following overall equation was developed:

Overall Organization Level Demographic Attrition Risk = **0.000935** * ***Designation*** - **0.0247** * ***PR (2006)*** + **0.0259** * ***PR (2007)*** + **0.2697** * ***Gender*** - **0.2117** * ***Marital Status*** - **0.0571** * ***Age*** - **0.2321** * ***Education*** + **0.1400** * ***Tenure*** - **0.2601** * ***City*** - **0.1065** * ***Salary Grade*** + **2.3537**

The analysis revealed that demographic designation and performance ratings were not as significant as the others. Hence these variables were removed and the modified equation developed was:

Demographic Attrition Risk (Significant factors) = **0.2695** * ***Gender*** - **0.2120** * ***Marital Status*** - **0.0569** * ***Age*** - **0.2317** * ***Education*** + **0.1403** * ***Tenure*** - **0.2606** * ***City*** - **0.1056** * ***Salary Grade*** + **2.3610**

Here values like 0.2695 represent the coefficients of independent variable Gender and the constant 2.3610 represents the effect of all uncontrollable variables. This constant represents the value of dependent variable when all independent variables are made equal to zero.

The analysis revealed that independent variables like department and the performance ratings were not as significant as other. The risk equations by departments are given in Appendix C.

4.3.3 Test the Attrition Risk Equation on the Data #2 (test data) set

The attrition risk equation was tested on the Data #2 (test data) check its accuracy. The model created on Data #1 (main data) threw out similar results on Data #2 (test data), thus validating the equation.

Table below shows actual employee count versus predicted employee count from the data set #2:

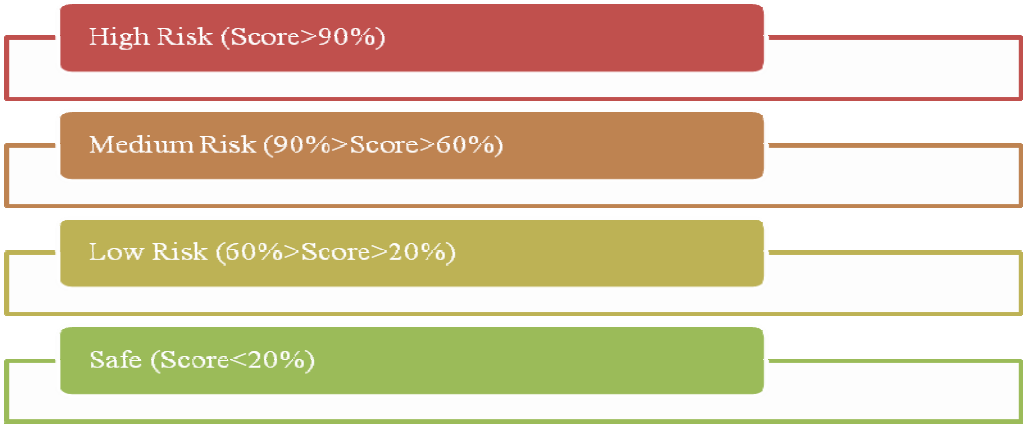
Actual vs. Predicted Employee counts			
	Actual Predicted Value of Existing Employee	Actual Predicted Value of Separated Employee	Actual Employee Count
Actual Existed Employee	**62%**	**38%**	**1937**
Actual Separated Employee	**20%**	**80%**	**1619**
Predicted Employee Count	**1523**	**2033**	**3556**

From the table above we can say that,

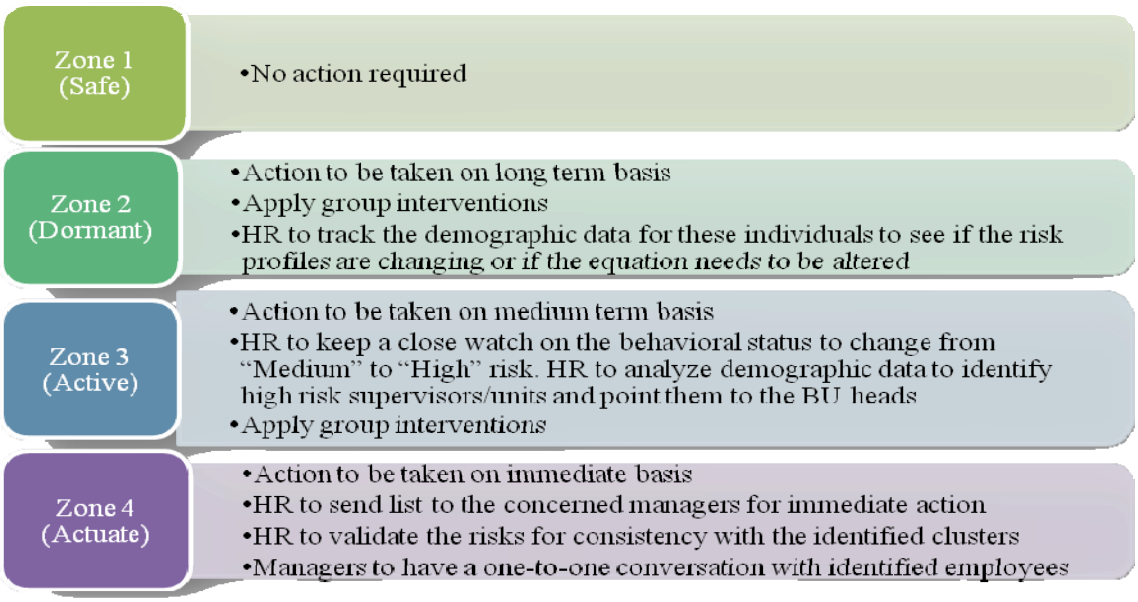
- Model predicted 80% of actual separated employee (N=1619) correctly.
- Model predicted 62% of existed employee (N=1937) correctly.

4.3.4 Retention Plan

Based on the model results, four levels of employee risk buckets were identified and have been shown below:



The retention plan charted using identified risk bucket for Anonymous Inc. has been presented below:



A roll out plan spread over a time span of 4 weeks was suggested to Anonymous Inc. with the roles defined for each of the stakeholders and the nature of their involvement.

The following action plan was devised for each of the zones identified above:

- Safe Zone – No action will be taken. Employees in this zone are engaged.
- Low Risk Zone – No action will be taken. Employees are at a low risk of attriting.
- Medium Risk Zone – A discussion to be scheduled by the manager with the employee. During this discussion, the manager would probe on the employee's level of engagement by seeking to understand his/ her concern areas.
- High Risk Zone – A discussion to be scheduled by the manager with the employee. During this discussion, the manager would probe on the employee's level of engagement by seeking to understand his/ her concern areas. If the employee is a high-performer or a high-potential, a further discussion will be scheduled by the skip-level manager with the employee. The focus of the discussion would be to understand employee's immediate concerns.

4.3.5 Implementation of real time IT solution

The model equation form Behavioural risk and Demographic attrition risk were implemented online by the IT department of Anonymous Inc by linking the model parameter to their database. Managers were provided access to employees of his/her team to check the risk association and according chalk out the retention strategy.

III. CONCLUSION

The approach shown in the paper to predict employee attrition using 'Logistic regression' predictive technique is based on separated employee's demographic data for particular organization. This technique to predict employee attrition can be applied to every organization based on employee demographic data.

This predictive technique to define risk attached with each employee should be modified and remodelled bi-yearly to refine coefficients based on current data.

The motive of this approach is to help organizations proactively predict attrition in real time and therefore take the necessary steps to prevent it, or plan the manpower inventory accordingly. Instead of trying to retain everyone, an organization should identify precisely who needs to be kept on board, and how the company can continue to appeal the high potential employees.

Employee Attrition Risk Assessment is receiving significant attention and opening a scope of focused research initiatives. An analytical approach to this assessment aids in prediction of attrition risk and subsequent action planning. Among the various statistical predictive techniques available, Logistic Regression and Discriminant Analysis come the closest to give a solution. Logistic Regression in this case would give more robust results as it does not assume conditions of multivariate normality and homoscedasticity. In the case presented, Logistic Regression has been employed to predict employee attrition risk based on demographic information and a retention plan has been charted out to target the risk categories derived.

IV. APPENDIX

A. Tangible & Intangible Costs

Highlighted categories in red are some important losses to an organization.

TANGIBLE COSTS
Termination
Exit Interviews – HR staff time
Severance Pay
Accumulated leaves
Separation Processing – Administrative support
Vacancy
Overtime for co-workers
Temp agency services
Recruitment
Writing job ad
Running job ad
Third party recruiter fees
Other (e.g. Referral bonus)
Selection and Hiring
Application Screening
Interviewing
Reference Check
Finalizing employee contract
Relocation
Other (e.g. Signing bonus)
Orientation and Training
New Hire Processing
Orientation
Orientation Material

Training cost (Trainer cost + trainer materials)
INTANGIBLE COSTS
Lost productivity of incumbent prior to departure
Lost productivity of co-workers or subordinates
Lost productivity / time of supervisor during vacancy, orientation and training
Lost productivity of new hire during orientation and training
Lost productivity of new hire during transition
Increased defects / operating errors during vacancy or transition

B. Overview of behavioral model

Based on the interactions with employees through FGDs and HR Business Partner meetings, behavioral factors which are key precursors to attrition to the Anonymous Inc. context were identified. Lists of factors leading to attrition were identified and a survey was administered.

Based on the responses to the behavioral questions a factor was calculated for each employee and four risk categories were identified as below:

C. Model Output

The output had two main tests as below:

C.1 Test of global model fit

C.1.1. Model Fit Statistic

C.1.2. Testing Global Null Hypothesis

C.2 Parameter Estimates

C.2.1 Analysis of Maximum Likelihood Estimates

C.2.2 Odds Ratio Estimates

C.1. TESTS OF GLOBAL MODEL FIT

The likelihood ratio test is a global test of fit. The null hypothesis is that none of the predictor variables are related to the outcome (ALL the betas=0). If the likelihood ratio test has a significant p-value, this means that at least one of the predictor variables is significantly related to the outcome (beta not equal to 0)

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	11220.005	10975.046
SC	11227.045	11052.483
-2 Log L	11218.005	10953.046

The likelihood ratio test comes directly from the likelihood equation in Maximum Likelihood Estimation.

When the model is fit with only the intercept (no predictors), the value of the likelihood equation -2LogLikelihood (-2LogL) of 264.9589. When the model is fit with the intercept and the 10 predictors, the value of the likelihood equation -2LogLikelihood is 10953.046 at its maximum value.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	264.9589	10	<.0001
Score	261.8686	10	<.0001
Wald	253.5399	10	<.0001

If we subtract the -2LogL of a reduced model (intercept only) from the -2LogL of a full model (intercept and number of predictors), this has a chi-square distribution with K degrees of freedom under the null hypothesis (ALL Betas=0). Here we get a value of 264.9589 for a chi-square with 10 degrees of freedom (highly significant, so reject the null!). Where K is the number of predictor variables.

If the null hypothesis rejects, this means that at least one of the predictor variable is important. Something in our model is predictive!

The "Analysis of Maximum Likelihood Estimates" table lists the parameter estimates, their standard errors, and the results of the Wald test for individual parameters.

C.2. PARAMETER ESTIMATES

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.4947	0.0228	468.9943	<.0001
Designation	1	0.00304	0.0239	0.0161	0.8989
PR (2006)	1	-0.00576	0.0241	0.0571	0.8111
PR (2007)	1	0.00882	0.0246	0.1288	0.7197
Gender	1	0.1341	0.0234	32.8351	<.0001
Marital_Status	1	-0.1830	0.0260	49.4498	<.0001
Age	1	-0.0855	0.0315	7.3421	0.0067
Education	1	-0.1683	0.0246	46.7029	<.0001
Year of Service	1	0.2094	0.0290	52.0260	<.0001

City	1	-0.2178	0.0241	81.7257	<.0001
Salary Grade	1	-0.1457	0.0317	21.1911	<.0001

Overall Organization Level Demographic Attrition Risk = **0.000935** * ***Designation*** - **0.0247** * ***PR (2006)*** + **0.0259** * ***PR (2007)*** + **0.2697** * ***Gender*** - **0.2117** * ***Marital Status*** - **0.0571** * ***Age*** - **0.2321** * ***Education*** + **0.1400** * ***Tenure*** - **0.2601** * ***City*** - **0.1065** * ***Salary Grade*** + **2.3537**

Variables like Gender, Marital Status, Age, Education, Year of service, City and Salary grades are significant and variables are Designation, performance rating are insignificant.

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Designation	1.001	0.987	1.015
PR (2006)	0.976	0.797	1.194
PR (2007)	1.026	0.891	1.182
Gender	1.310	1.194	1.436
Marital Status	0.809	0.763	0.858
Age	0.944	0.906	0.984
Education	0.793	0.742	0.847
Year of Service	1.150	1.107	1.195
City	0.771	0.729	0.816
Salary Grade	0.899	0.859	0.941

Interpretation: For every 1 unit increase in Year of Service there is an estimated 15 % increase in odds of employee leaving the organization.

D: Model equations by departments

Department	Overall Equation	Equation of Significant factors
Department 1	1.9279 – (0.6271*PR (2006)) + (1.4275*PR (2007)) – (0.1674*Gender) – (0.4937*Marital Status) + (0.3889*Age) + (0.279*Education) + (0.8665*Tenure) – (0.5880*City) – (0.1189*Salary Grade)	No Significant Factors
Department 2	2.8146 – (0.0585*PR (2006)) + (0.3612*PR (2007)) – (0.0455*Gender) – (0.1991*Marital Status) + (0.0437*Age) - (0.3648*Education) - (0.5339*Tenure) – (0.7228*City) – (0.1007*Salary Grade)	No Significant Factors, except the Intercept
Department 3	1.3263 – (0.467*PR (2006)) + (0.584*PR (2007)) – (0.279*Gender) – (0.7246*Marital Status) - (0.4217*Age) - (0.4986*Education) + (0.4298*Tenure) + (0.4123*City) + (0.0381*Salary Grade)	No Significant Factors, except the Intercept
Department 4	1.7741 – (0.0197*PR (2006)) + (0.035*PR (2007)) + (0.0673*Gender) – (0.1274*Marital Status) + (0.1149*Age) + (0.1706*Education) + (0.00901*Tenure) – (0.4357*City) – (0.3101*Salary Grade)	1.7741 + (0.1706*Education) – (0.4357*City) – (0.3101*Salary Grade)
Department 5	1.4277 + (0.3178*PR (2006)) + (0.4259*PR (2007)) + (0.3401*Gender) – (0.2889*Marital Status) + (0.942*Age) + (0.2541*Education) - (0.1034*Tenure) + (0.1275*City) – (1.0851*Salary Grade)	No Significant Factors except Intercept
Department 6	1.568 – (0.0524*PR (2006)) + (1.1885*PR (2007)) + (0.1747*Gender) – (0.0197*Marital	1.568 + (0.4937*Tenure) – (0.2766*City)

	Status) - (0.1967*Age) + (0.1506*Education) + (0.4937*Tenure) – (0.2766*City) – (0.2363*Salary Grade)	
Department 7	1.6002 + (0.1186*PR (2006)) - (0.0443*PR (2007)) + (0.1492*Gender) – (0.1662*Marital Status) + (0.0667*Age) - (0.124*Education) + (0.0425*Tenure) – (0.0798*City) – (0.0719*Salary Grade)	No Significant Factors except Intercept
Department 8	1.6726 – (0.1983*PR (2006)) + (0.2024*PR (2007)) + (0.5563*Gender) – (0.1875*Marital Status) + (0.2714*Age) + (0.3518*Education) + (0.2317*Tenure) – (1.9874*City) – (0.7207*Salary Grade)	No Significant Factors
Department 9	1.6452 – (0.0473*PR (2006)) - (0.2742*PR (2007)) + (0.0677*Gender) – (0.1077*Marital Status) - (0.2593*Age) - (0.00529*Education) + (0.2774*Tenure) + (0.1970*City) – (0.0846*Salary Grade)	No Significant Factors except Intercept
Department 10	16.5583 + (9.5741*PR (2006)) - (59.8548*PR (2007)) + (16.7982*Gender) + (43.5132*Marital Status) + (79.5301*Age) + (4.9771*Education) + (43.5473*Tenure) – (15.9377*City) – (123.2*Salary Grade)	No Significant Factors
Department 11	30.8236 + (3.2992*PR (2006)) - (30.3507*PR (2007)) – (0.2457*Gender) + (28.2294*Marital Status) - (3.9096*Age) - (4.772*Education) + (19.324*Tenure) – (37.2366*City) + (32.291*Salary Grade)	No Significant Factors

V. ACKNOWLEDGEMENT

We would like to thank Aon Hewitt for the help and support it provided during the execution of the live project and at the time of writing the research paper.

We would also like to thank our leaders, Rahul Malhotra and Kabir Pandit, for granting the permission to write the paper. Their concern for identifying attrition risk through predictive techniques provided the quest for writing this research paper.

In addition, we would also like to express our gratitude to all of those, who supported us with enthusiasm and had belief on the paper.

VI. REFERENCES

1. *Developing a Predictive Model for Employee Attrition*
   A Primary Research conducted and published by Aon Hewitt Associates, 2009
2. Agresti, Alan, *An Introduction to Categorical Data Analysis, John Wiley and Sons, Inc.* 1996
3. David W. Hosmer and Stanley Lemeshow, *Applied Logistic Regression*, Wiley, John & Sons, 1989
4. Menard, Scott, *Applied Logistic Regression Analysis, Quantitative Applications in the Social Sciences, No. 106*, SAGE Publications,1995
5. Barbara and Linda Fidell, *Using Multivariate Statistics, Tabachnick, Third edition*, 1996
6. J.C., and J. Travis, *Nontraditional Regression Analyses*, Trexler, Ecology 74:1629-1637, 1993
7. Connor, E.F., Adams-Manson, R.H., Carr, T.G., and M.W. Beck, *The effects of host plant phenology on the demography and population dynamics of the leaf-mining moth, Cameraria hamadryadella* (Lepidoptera: Gracillariidae), Ecological Entomology 19:111-120, 1994
8. Suvro Raychaudhuri, *Manpower Planning and Employee Attrition Analytics*
9. Sudipta Dev, *Calculating employee attrition,* Express Computer Online
10. Mentor Cana, *What is the difference between Logistic Regression and Discriminant Analysis?,* 2003
11. John Poulsen and Aaron French, *Discriminant Function Analysis (DA)*
12. George Antonogeorgos, Demosthenes B. Panagiotakos,Kostas N. Priftis, and Anastasia Tzonou, *Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods*,
13. Naresh Malhotra, *Marketing Research- An Applied Orientation*, Fifth edition
14. *en.wikipedia.org*
15. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR, *A simulation study of the number of events per variable in logistic regression analysis,*, Journal of Clinical Epidemiology 49:1373-137, 1996