Application of Machine Learning for Prediction of Employees at Risk of Leaving Organization

Phaneendra Akula
REVA Academy for Corporate Excellence
REVA University
Bengaluru, India
Akula.res@reva.edu.in

Abhilash Bodanapu
REVA Academy for Corporate Excellence
REVA University
Bengaluru, India
Abhilash.HR02@reva.edu.in

***Abstract*— This practice presents a scientific approach in identifying the talent at risk enabling the organization to respond proactively. The model takes a group of statistically significant factors that correlate to an employee's decision to leave and further uses Machine Learning algorithms to Predict employee's probability to quit organization with an accuracy of over ~80%.**

The impact of the model has been multi-fold on the organization, this assisted leadership to arrive at a macro level decision to revamp the Rewards, Incentive schemes & Promotion policies. Further with the model's capacity to deliver risk probability with the individual predictor (reasons) for each employee has further helped project managers to act at a micro level resulting in proactive retention of employees.

Keywords-*Employee Attrition, Logistic Regression, Survival Analysis*

INTRODUCTION

NASSCOM states 'Skills are the new currency', with IT and Startup industry becomes skill-centric, attracting and retaining skilled employees have a pivotal role in the sustenance and growth of the organization. Especially with ever-rising attrition impacting business severely in terms cost of hiring, market premium, the time-taken to hire and loss of knowledge capital. Hence, an early prediction of employees at risk can enable organizations to proactively engage and act on their Human capital and curb the possible loss of revenue and knowledge.

There are numerous set of reasons on why employees could be leaving the organization, such as compensation, promotion, work-life balance, culture fitment, etc., while Human Resource as a function has been effective and been enlarging themselves in reacting to the employee issues and engaging them, the pro-activeness in terms of predicting the employee with 'Exit tendencies' and acting upon them has been a challenging aspect. Which brings us to the definite need of a Risk prediction to this function, this can be done with numerous techniques like Machine learning, Engagement survey's and focused individual discussions.

Organizations The investment that the Organizations making in terms of training, upskilling, rewards and recognition, onsite opportunities, premium compensations further demands. At present these interventions are commonly applied in order to manage employee attrition, through this paper, we used Machine learning as a technique to predict the attrition risk thereby enabling the Management to deal with the attrition proactively and more importantly identify the strategic pockets of investment which maximizes the return.

Figure depicting the various application level possibilities of Predictive Analytics in Attrition.

I. JOURNEY OF PREDICTIVE ANALYTICS

1. Objective settings:
Any analytics journey starts with a well-defined problem statement. To arrive at the same, multiple levels of focused discussions have been conducted with Business leaders and HR partners.
Below areas have arrived as an area of interest for the
• High performers
• Experience range
• Skill groups

2. Understanding Data:
A vital aspect of an Analytics study is the ability to gather relevant. Significant time has been spent on collecting, cleaning and preparing the data for further steps.
Initial data gathering started with ~6500 records, however, post-cleanup we arrived at 3500 records for further action.

Sources of Data:
Data collection and validation have happened from multiple teams to accurately arrive at relevant data in sync with HR policies
• COE's
  O Compensation & Benefits
  O Performance Management
  O Shared services

• Business partnering team
• HR Shared Services

3. Analytical Modeling:
Employed 3 techniques to arrive at a Macro and Micro level insights –

A. Descriptive statistics
B. Survival Analysis
C. Logistic regression

A. Descriptive Statistics:

It is applied to describe the basic features of the data. It provides a lucid explanation about the sample and the measures. With simple visualization, they form the basis for virtually every quantitative analysis of data.

Descriptive Statistics is generally applied to present "quantitative descriptions" in an understandable form. There are a lot of measures in any research-based study or it could be about measuring a large number of people on any measure. Descriptive statistics bring a sensible understanding by simplifying the large chunks of data. Each descriptive statistic reduces lots of data into a simpler summary. For example, consider a simple promotion data where the employee is eligible once in a year for a review for promotion, however not necessarily get it. To identify the employees on a Fast track promotion, let us say getting promoted once in every two years consecutively for 2 times defines an employee as a fast track progressive person, then taking the average no of promotions in last 4 years of an employee tenure qualifies them for the eligible list.

B. Survival Analysis:

This is commonly defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. These events can be disease, marriage, death, etc. The measurement till Time of event could be in days, weeks, years, etc. For example, if the event of interest is a heart attack, then the survival time is the time measured in years until a person develops a heart attack.

Both survival & hazard functions are important concepts in survival analysis to explain the distribution of event times. The survival function gives, for every time, the probability of surviving up to that time. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. While these are often of direct interest, many other quantities of interest (e.g., mean survival) can also be estimated from knowing either the survival or hazard function. It is a common interest during the study of Survival concept to explain the relationship of a factor of interest (e.g. treatment) to the time of an event, in the presence of many covariates, such as demographics, age, race, gender, etc. Various set of models are available to analyze the relationship of a set of predictor variables with the survival time. Methods include parametric, semiparametric and nonparametric approaches.

In this study, we used "a nonparametric estimator of the survival function, the Kaplan Meier method is widely used to estimate and graph survival probabilities as a function of time." This is being applied to obtain a "Univariate descriptive statistics" for survival data, which includes the median survival time and compare the survival experiences for two or more groups of subjects. "To test for overall differences between estimated survival curves of two or more groups of subjects, such as males versus females, several tests are available, including the log-rank test."

C. Logistic Regression

"Logistic regression is used for predicting 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." Similar to many other forms of regression analysis, this will make use of many predictor variables that may be either categorical or numerical. Logistic regression is used extensively in the medical & social sciences as well as marketing and Sales applications such as prediction of a customer's inclination to purchase a product or cease a subscription.

For example:
If we have a model which predicts whether a person is "male or female" based on their height (completely fictitious). If a height of 150cm is given to know whether the person is a female or male.

"We arrived at coefficients, b0= -100 & b1= 0.6. With the above equation, we can calculate the probability of male for a given height of 150cm, P(male | height=150)." We will use EXP() for e, because that is what you can use if you type this example into your spreadsheet:

y = e^(b0 + b1*X) / (1 + e^(b0 + b1*X))

y = exp(-100 + 0.6*150) / (1 + EXP(-100 + 0.6*X))

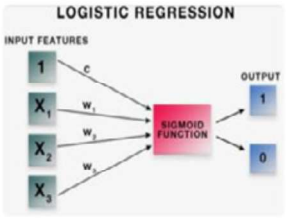
y = 0.0000453978687

Or a probability of near zero that the person is a male.

While practically we can use the probabilities directly. Because this is a classification and we would prefer a brief answer, we can snap the probabilities to a binary class value, for example:
0 if p(male) < 0.5

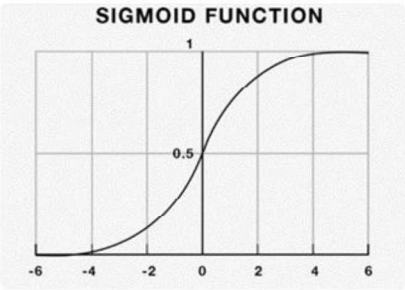
1 if p(male) >= 0.5

"A visual depiction of Logistic regression with weights. The logistic regression model computes a weighted sum of the input variables similar to the linear regression, but it runs the result through a special non-linear function, the logistic function or sigmoid function to produce the output y."



$y = \text{logistic}\ (c + x_1{*}w_1 + x_2{*}w_2 + x_3{*}w_3 + \ldots\ldots + x_n{*}w_n)$

$y = 1 / 1 + e\ [-(c + x_1{*}w_1 + x_2{*}w_2 + x_3{*}w_3 + \ldots\ldots + x_n{*}w_n)]$

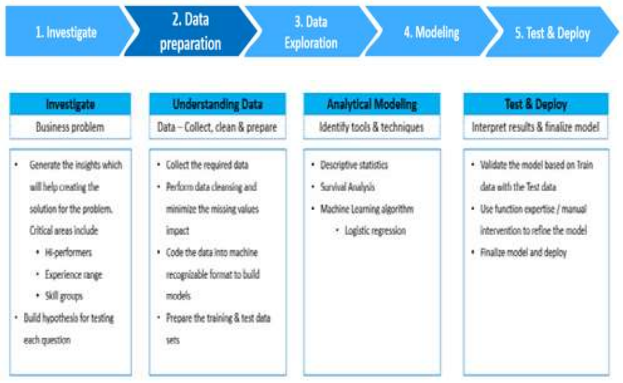


II. APPLICATION OF LOGISTIC REGRESSION

Noname.com (name changed to maintain confidentiality), an MNC company is facing a problem in controlling employee attrition. Some initiatives had been taken internally to control the attrition, however, these initiatives were not adequate and the leadership and management felt the need for more pro-activeness to be better equipped in order to manage employee attrition scenarios better. Noname.com requested HR and Data scientists in the organization to come up with a predictive model which would estimate the attrition risk of employees for a better decision making and initiatives.

Study objective: Predict Employees at risk to Reduce cost and save time.

Overall Project Landscape:



Sample size:

Data collected has been segregated in 2 sets based on the relevance of the data.

Data Set 1: For building the model (Train data)

- Active employees as on Dec 31st 2017 : 1650
- Separated employees : 640
- Total sample size : 2290

Data Set 2: For testing the model (Test data)

- Active employees as on Jun 30th 2018 : 2750
- Separated employees : 570
- Total sample size : 2180

Data gathering & cleaning steps:

1. Metric-level data has been converted into categorical data from some demographical questions like Age, Year of service etc.

2. The missing values information for "performance rating" is replaced/filled with average or root-mean-squared values

3. Parameters like Rating and compensation hike amount were mandatory, data not available cases have been removed

4. Outliers had to be cleaned including the invalid data points and employees whose data was missing

5. Status is the dependent variable name used with values '0' (zero) or '1' (one). The model considered 'Attrition' as a "dependent variable" and rest of the variables were considered as "independent variables"

6. Below is the list of variables used in the model building

General	Profile	Employee Level
Gender	Skill	Time in grade
Age	Total Experience	Last Rating
Location	Organization experience	Hike Amount
Grade		Compa-ratio
		Promotion
		Rewards & Recognition

Developing Equation through Logistic Regression:

'R' programming has been used to run a logistic regression model which gives out the probability of staying or leaving with a probability ranging from 0 to 1.

Logistic Regression considered all parameters and eliminates insignificant variables through an iterative process. "Chi-Square test and Maximum Likelihood Estimates were used to identify coefficients for significance for inclusion or elimination from the model."

The model fitment has been tested post every round of elimination. The analysis has been concluded when no more variables needed to be eliminated from the model and when the model converged.

Following equation has been developed:

Attrition risk estimate(all) = -2.315 + 0.101 * Age - 0.018 * Total experience + 0.129 * TIG + 0.250 * Compa-ratio + 0.387 * Hike amount + 0.043 * promotion + 0.227* RnR - 0.134 * Rating + 0.340 * Skill + 0.221 * Organization experience – 0.076 * Location - 0.173 * Grade + 0.128 * Gender

Parameters like Total experience, promotion and location have come out as insignificant ones and have been removed from the final equation due to its minimal impact.

Revised Attrition risk estimate(only significant) = 2.315 + 0.101 * Age + 0.129 * TIG + 0.250 * Compa-ratio + 0.387 * Hike amount + 0.227* RnR - 0.134 * Rating + 0.340 * Skill + 0.221 * Organization experience - 0.173 * Grade + 0.128 * Gender

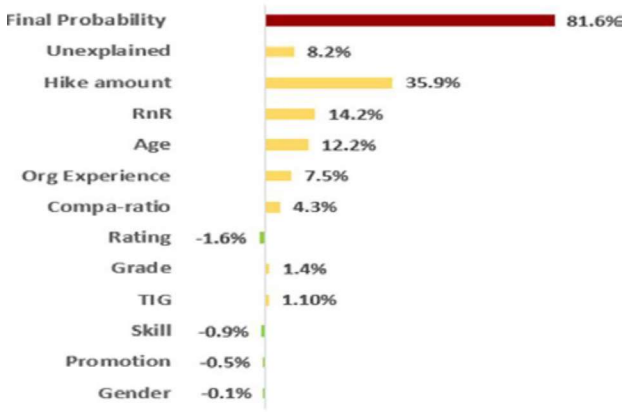
Testing model:

Model built using the previous data has been tested on the test data to validate the equation. Below table shows the Actual vs Predicted exits with an accuracy of about ~80% thus providing confidence to make use of the model results.

Actual	Predicted		Count of employees
	Predicted Value of Active Employee	Predicted Value of Exited Employee	
Actual Active Employee	1789	391	2180
Actual Exited Employee	177	393	570
	1966	784	2750

The model also provides a breakdown plot of parameters individual employee wises there by enabling the organization to act on a specific individual as per their problem areas.

Below graph gives an example of the individual risk prediction:



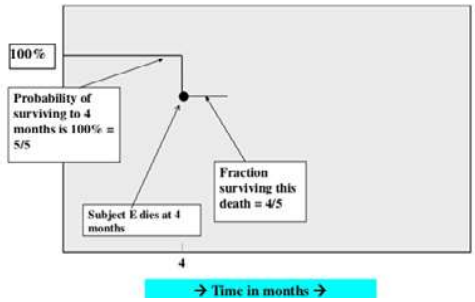
In the above example, ethe mployee has 81.6% chances to exit, out of which increment amount causes the maximum effect (50.1%) followed by RnR (14.2%) and Age(12.2%)

Macro Insights:

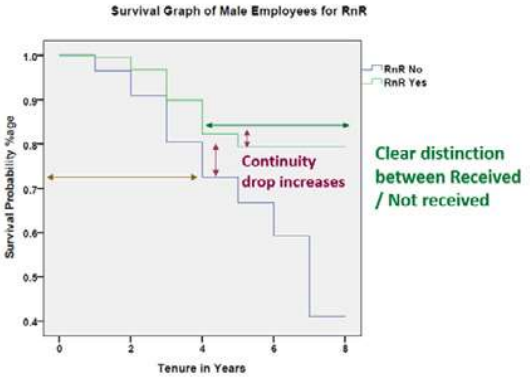
We used Survival graph as the technique to arrive at Macro level insights for the leadership to arrive at practical and strategic initiatives to invest in the right pocket.

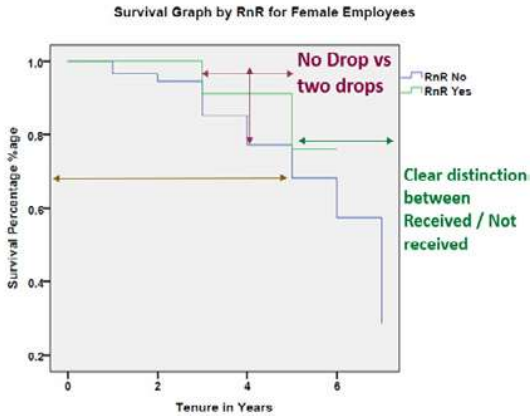
Survival graph: A method to analyze data where the outcome variable is the time until the occurrence of an event of interest. Kaplan-Meier is the technique used to explore.

Corresponding Kaplan-Meier Curve



1. **Survival Graphs on Rewards & Recognition (R&R):** While everyone knows the significance of R&R, we rarely measure the ROI of the same or validate whether its working or whether the organization is investing in the right pockets or not. Below survival, graph is performed on Male and Female employees separately to see the impact of R&R awarded in last one year.



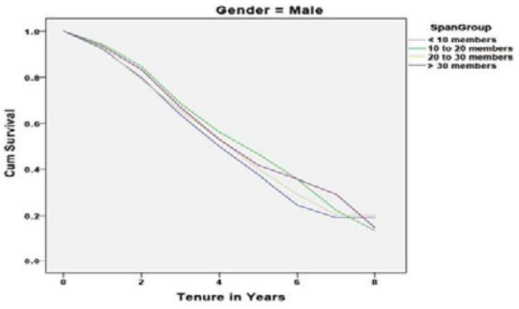


Insights from the above survival graphs:

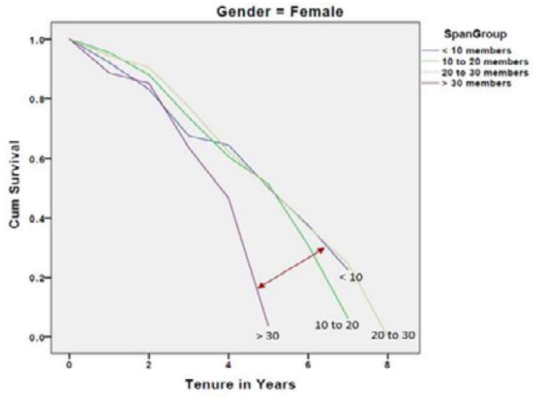
- R&R received shows at least ~10% more probability to stay back in the initial years (up to 4yrs), greater probability to stay back at higher experienced levels (> 4 years)
- Overall, Retention probability drops for R&R not received population as experience increases and drastically drops at higher experienced levels
- From Gender Perspective, Women shows better reception of R&R than Men, i.e., if given award the impact continues for a long duration than Men (No drop vs Two drops scenario in the above pic)

Survival Graphs on Team size:

Similar to R&R, survival graphs have been applied on Span of control or Team size as per the gender.



Span of control not a significant factor for Male Employees



Women may prefer to work in smaller teams compared to Men

Insights from the above survival graphs:

- While Men really not show any specific pattern as per the team size across their tenure, Women shows better probability to stay back in the organization if working in a smaller team (<10 members) over bigger teams (>30 members)

III. APPLICATION OF INSIGHTS FOR BETTER RETENTION

As per the retention probability arrived for each individual, below categorization is what is formulized to arrive at specific actions on the employees

1) Basis the Probability arrived

Risk group (% of Probability to Exit)	Action
> 80%	Immediate personal connect with HR Business partner to identify the key talent and propose corrective actions as per
50% to 80%	Intervene at the employee level based on the Individual predictor analysis. Employees are tapped as per the maximum contributing parameter
<50%	No intervention as there is no immediate set of risk

2) Basis the Survival Graphs

a. Introduced a new R&R category for women employees to maximize the awards in this place

b. Introduced a new R&R category for the experienced employees beyond 4 years of age to maximize returns

c. When women are up for role change, they are being preferred as per the team size they will be working on

IV. CONCLUSION

The practice shown in the paper which predicts the employee attrition with a probability opens up a plethora of opportunities for the HR & Business to take proactive actions. The individual probability risk of each employee and each parameter enables the HR Partners and project managers to address the near to exact concerns of the employees.

Further the insights generated using Survival Graphs enables the leadership to take strategic decisions in terms of revamping the HR policies, especially R&R, Promotion criteria, etc., This study also gives a direction for the HRs to enlarge their scope of Predictive analysis across different functions of HR gradually and also develops the mindset to unravel insights using analytics and apply them on work aspects thereby moving away from traditional approach of decisions just based on experience.

V. REFERENCES

[1] Dr. K. Aparna Rao (2011)"Employee retention-a real-time challenges in the global work environment" - Journal of research in commerce & management, volume no.1, issue no.11 ISSN 2277-1166, P.No: 125-131.

[2] Exploring the role of perceived external prestige in Turnover Intentions, Int. J. of Human Resource Management 15:8 December 2004 13901407. Authored by Oliver Herrbach, Karim Mignonac and Anne-Laure Gatingnon.

[3] R. Shanmugam, A.Anbu, Dr.K.Kalpana (2012) "Retention of employees in IT industries with special reference to Wipro technologies", International Journal Of Management (IJM), Volume 3, Issue 2, May-August (2012), P. No 270-278.

[4] Shaw, J.D., Gupta, N, & Delery, J.E., (2005), "Alternate conceptualizations of the relationship between voluntary turnover and Organizational Performance," Academy of management journal, Vol.48, pp. 50-68.

[5] Research conducted by Aon Hewitt Associates, 2009, "Developing a Predictive Model for Employee Attrition"

[6] J.C., and J. Travis, "Nontraditional Regression Analyses," Trexler, Ecology 74:1629-1637, 1993

[7] Suvro Raychaudhuri, "Manpower Planning and Employee Attrition Analytics"

[8] Menard, Scott, "Applied Logistic Regression Analysis"s, Quantitative Applications in the Social Sciences, No. 106, SAGE Publications,1995

[9] Menard, Scott, "Applied Logistic Regression Analysis, Quantitative Applications in the Social Sciences," No. 106, SAGE Publications,1995

[10] David W. Hosmer and Stanley Lemeshow, Applied Logistic Regression, Wiley, John & Sons, 1989