

Employee Risk Attrition Assessment

using logistic regression analysis

Randy Lao | Capstone Project 1 | August 8, 2017

# TABLE OF CONTENT

## ABSTRACT

## Introduction

## objective

## overview of logistic regression

## o.s.e.m.n. pipeline

## obtaining the data

## data preperation/cleaning

## exploratory data analysis

## modeling

## test evaluation

## retention plan

## CONCLUSIOn

## future work

## ABSTRACT

This paper will cover the usage of logistic regression as a modeling algorithm to predict employee attrition risk within a company based on employee data. In this project, I will be covering my analysis and approach through different process flows in the data science pipeline. The main goal is to understand the reasonings behind employee turnover and to come up with a model to classify an employee’s risk of attrition. A recommendation for a retention plan was created, which incorporates some best practices for employee retention at different risk levels of attrition.

## INTRODUCTION

**“You don’t build a business. You build people, and people build the business.” -Zig Ziglar**

Long-term success, a healthy work environment, and high employee retention are all signs of a successful company. In a sense, it’s the employees who make the company. It’s the employees who do the work. It’s the employees who shape the company’s culture. But when a company experiences a high rate of employee turnover, then something is going wrong. This can lead the company to huge monetary losses by these innovative and valuable employees.

When companies experience a high rate of turnover, their investments on their employees are all drained away, which includes a loss of salary, benefits, bonuses, training, and other expensive resources. Companies that maintain a healthy organization and culture are always a good sign of future prosperity. Recognizing and understanding what factors that were associated with employee turnover will allow companies and individuals to limit this from happening and may even increase employee productivity and growth. These predictive insights give managers the opportunity to take corrective steps to build and preserve their successful business.

## OBJECTIVE

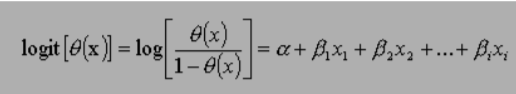
The company wants to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees. Overall, the implementation of this model will allow management to create better decision-making actions.

## OVERVIEW OF LOGISTIC REGRESSION

Logistic Regression commonly deals with the issue of how likely an observation is to belong to each group. This model is commonly used to predict the likelihood of an event occurring. In contrast to linear regression, the output of logistic regression is transformed with a logit function. This makes the output either 0 or 1. This is a useful model to take advantage of for this problem because we are interested in predicting whether an employee will leave (0) or stay (1).

Another reason for why logistic regression is the preferred model of choice is because of its interpretability. Logistic regression predicts the outcome of the response variable (turnover) through a set of other explanatory variables, also called predictors. In context of this domain, the value of our response variable is categorized into two forms: 0 (zero) or 1 (one). The value of 0 (zero) represents the probability of an employee not leaving the company and the value of 1 (one) represents the probability of an employee leaving the company.

Logistic Regression models the probability of ‘success’ as:



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

## O.S.E.M.N. Pipeline

I will be following a typical data science pipeline, which is called “OSEMN” (pronounced awesome):

1. **O**btaining the data is the first approach in solving the problem.
2. **S**crubbing or cleaning the data is the next step. This includes data imputation of missing or invalid data and fixing column names.
3. **E**xploring the data will follow right after and allow further insight of what our dataset contains. Looking for any outliers or weird data.
4. **M**odeling the data will give us our predictive power on whether an employee will leave.
5. I**N**terpreting the data is last. With all the results and analysis of the data, what conclusion is made? What factors contributed most to employee turnover? What relationship of variables were found?

## OBTAINING THE DATA

The data was found from the “Human Resources Analytics” dataset provided by Kaggle’s website: [*https://www.kaggle.com/ludobenistant/hr-analytics*](https://www.kaggle.com/ludobenistant/hr-analytics)*.*

This dataset is a simulation of a hypothetical company, which means that the features and observations used are all made up to mimic a real-world scenario. The number of observations given from the dataset contains 14,999 employee information.

I will be using **Python** as the programming language for the analysis:

#Read the analytics csv file and store our dataset into a dataframe called "df"

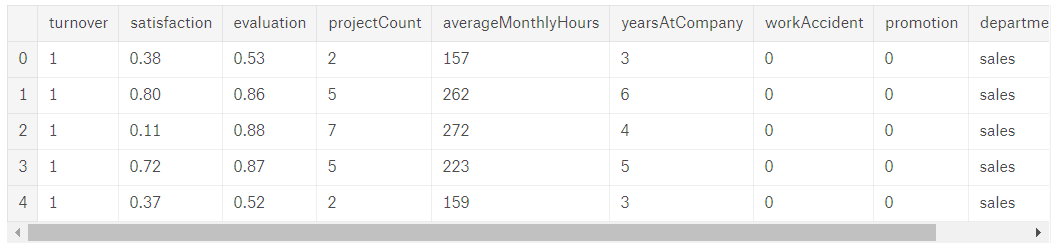
df = pd.DataFrame.from\_csv('../input/HR\_comma\_sep.csv', index\_col=None)

## DATA PREPERATION/CLEANING

Typically, data preparation/cleaning requires a lot of work and can be a very tedious procedure. This dataset from Kaggle is clean and contains no missing values. But still, I will have to examine the dataset to make sure that everything else is readable and that the observation values match the feature names appropriately. This involved:

1. Conversion of data into categorical data, such as the “department” and “salary” features
2. Renaming of features for better readability
3. Check to see if there are any missing values in the dataset

The feature “**turnover**” is considered our **dependent variable**, whereas the rest of the features were our independent variables.



The following independent variables were used in the model:

• **Satisfaction**: An employee’s level of satisfaction in percentage

• **Evaluation:** An employee’s evaluation score in percentage

• **Project Count:** The amount of projects the employee has done

• **Average Monthly Hours:** The total monthly hours an employee worked

• **Years At Company:** The number of years an employee was at the company

• **Work Accident:** Whether an employee had an accident or not. Where 0 (zero) means no and 1 (one) means yes

• **Promotion:** Whether an employee had a promotion within the last five years. Where 0 (zero) means no and 1 (one) means yes

• **Department:** The type of department an employee worked under. Which includes sales, accounting, hr, technical, support, management, IT, product management, and marketing.

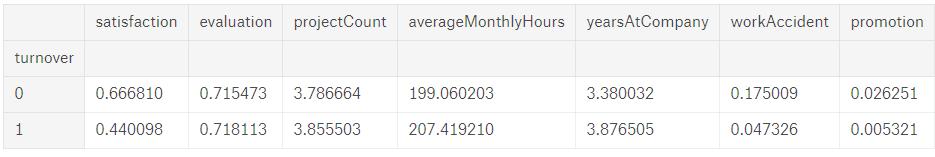
• **Salary:** The type of salary an employee got, which ranges from low, medium, or high.

## EXPLORATORY DATA ANALYSIS

1. **Statistical Overview**

Here are some important numbers to keep in mind of the dataset:

* There is 14,999 employees and 9 independent variables
* Turnover rate: 24%
* Mean satisfaction: 0.61



1. **Correlation Matrix & Heatmap**

**Summary:**

From the heatmap, there is a **positive (+)** correlation between the variables: **projectCount**, **averageMonthlyHours**, and **evlatuion**. Which means that the employees who spent worked more hours and did more projects had higher evaluations.

For the **negative (-)** relationships, the most important feature that correlated with our target variable (turnover) is **satisfaction**. This should support our initial intuition that employees who tend to quit would normally have lower satisfaction level.

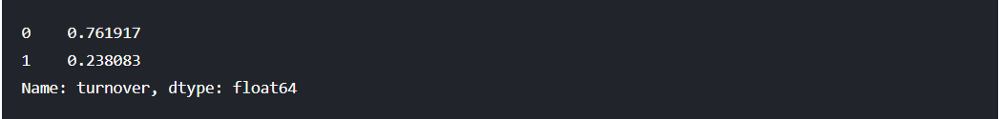
**Questions to Think About:**

What features affect our target variable (turnover) the most?

What features have strong correlations with each other?

Can we do a more in-depth examination of these features?





## MODELING

## TEST EVALUATION

1. **Histogram (Satisfaction / Evaluation / Average Monthly Hours)**

**Summary:**

Let’s examine the distribution on some of the employee’s features:

-**Satisfaction**: There are three distributions for employee satisfaction in the dataset. One group falls within satisfaction level of **(0–0.3),** another within satisfaction level of **(0.3-0.5),** and one from **(0.5-1).**

-**Evaluation:** There are three distributions for employee evaluation. One group falls within the lower spectrum of **(0-0.55),** the middle spectrum of **(0.55-.7),** and the higher spectrum of **(.7-1).**

**-Average Monthly Hours:** There are three distributions for employee average monthly hours. Those that work **100-150 hours**, another group that works **150-250 hours,** and a third group who works **250-300 hours.**

**-**The **evaluation** and **average monthly hours** features both share a similar distribution, which indicates high collinearity of the features.

**Questions to Think About:**

Is there a reason for these distributions on these graphs?

Could employees be grouped distinctively with these features?

1. **Salary V.S. Turnover**

**Summary:**

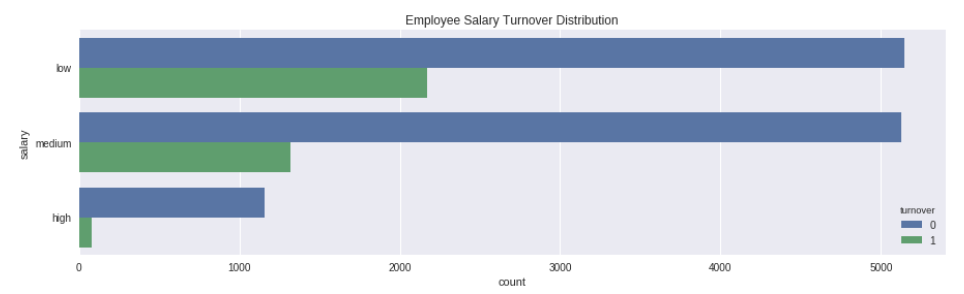
-Majority of employees who left either had **low** or **medium** salary

-Only a few employees left with **high** salary

**Questions to Think About:**

What is the work environment like for each salary level?

What made employees with high salaries leave the company?

What fraction of employees who left were contracted or part time?

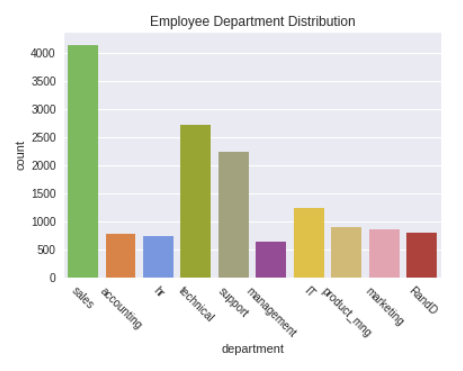
1. **Department V.S. Turnover**

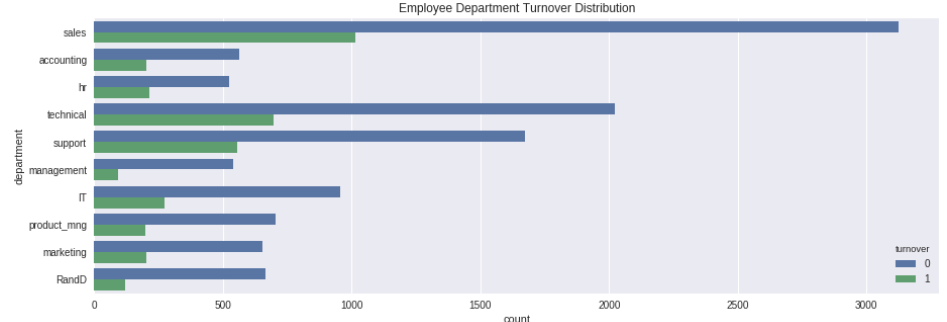
**Summary:**

-The top three departments with the most employees are **sales, technical, and support**.

**Questions to Think About:**

Can we pinpoint a more direct cause for employee turnover in each department?





1. **Project Count V.S. Turnover**

**Summary:**

-More than half of employees with **2,6,** and **7 projects** left the company

-Majority of the employees who **did not leave** had **3,4,** and **5** **projects**

-**All employees** with **7 projects** left the company

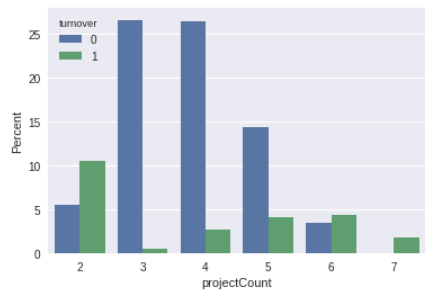
-There is an increase in turnover as project count increases

**Questions to Think About:**

Why are majority of employees leaving at the lower/higher spectrum of project count?

Does this mean that employees with project counts 2 left the company because they were underworked?

Are employees with 6+ projects overworked, thus leaving the company?



1. **Evaluation V.S. Turnover**

**Summary:**

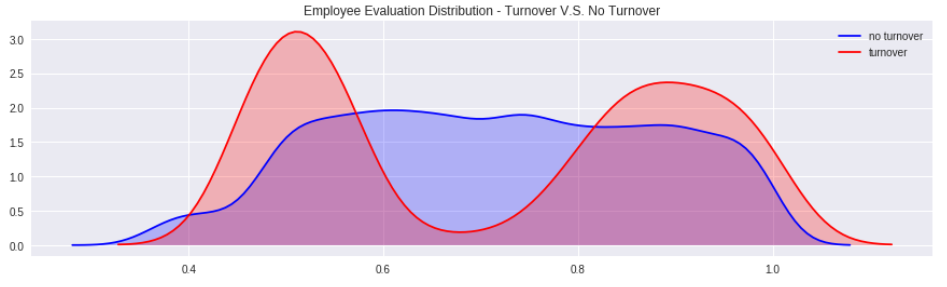
-There is a bimodal distribution for employees that left the company

-Employees with **low evaluation levels (0.2-0.6)** and **high evaluation levels (0.8-1)** were the bulk of employee turnover

-Employees with **evaluation levels (0.6-0.8)** had the smallest turnover rate

**Questions to Think About:**

Why are high evaluated employees leaving the company? Is it because of no promotion?



1. **Average Monthly Hours V.S. Turnover**

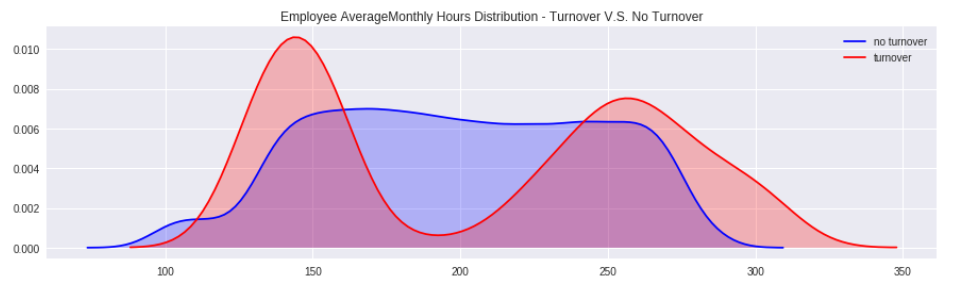
**Summary:**

-There is another bimodal distribution for employees that left the company

-Employees who had **less** hours of work **(~150 hours or less)** left the company more

-Employees who had **more** hours of work (**~250 hours or more**) left the company more

-Employees who left generally were **underworked** or **overworked**



1. **Satisfaction V.S. Turnover**

**Summary:**

-There is a **tri-modal distribution** for employees that left the company

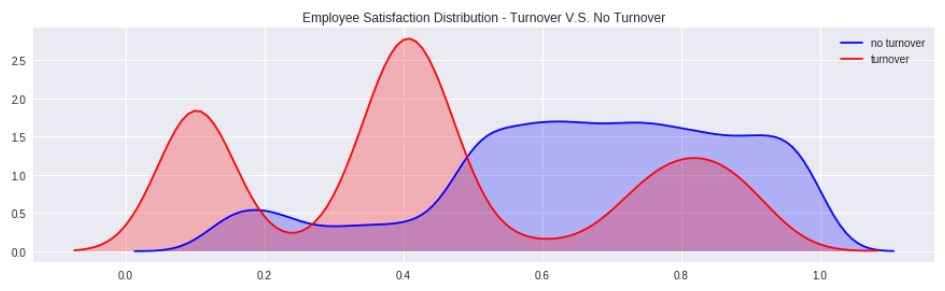
-Employees left with really **low satisfaction levels of (0-0.2)**

-Employees left with **low satisfaction levels of (0.3-0.5)**

-Employees left with **high satisfaction levels of (0.7-1)**

**Questions to Think About:**

Why are employees with high satisfaction levels leaving the company?



1. **Evaluation V.S. Satisfaction**

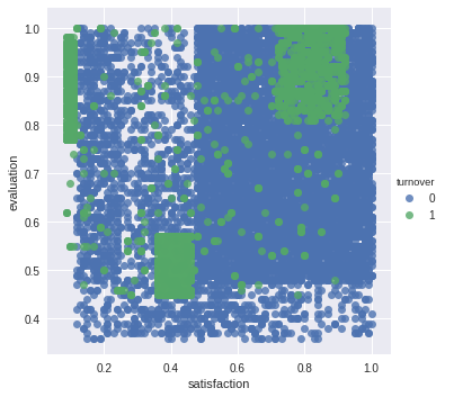
**Summary:**

-There are **three distinct clusters** for employees who left the company

**Cluster 1 (“Overworked” Employees):** These employees had **satisfaction scores below 0.2** and **evaluation** **scores above 0.75**. Employees here were evaluated highly and felt bad at work.

**Cluster 2 (“Under Performing Employees”)**: These employees had **satisfaction scores between (0.35-0.5)** and **evaluation scores below 0.6.** Employees here were evaluated poorly and felt bad at work. This is a typical reason why employees leave.

**Cluster 3 (“Ideal Worker”):** These employees had **satisfaction scores between (0.7-1)** and **evaluation scores of (0.8-1)**. Employees here were evaluated highly and felt satisfied at work.



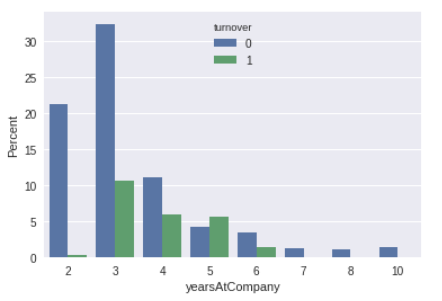
1. **Years at Company V.S. Turnover**

**Summary:**

-More than half of the employees with **4 and 5 years** left the company

**Questions to Think About:**

Why are employees leaving mostly at **3-5 years?**



## Modeling

## Test evaluation

***Precision and Recall / Class Imbalance***

This dataset is an example of a class imbalance problem because of the skewed distribution of employees who did and did not leave. More skewed the class 🡪 accuracy breaks down. In this case, evaluating our model’s algorithm based on accuracy is the wrong thing to measure. We would have to know the different errors that we care about and correct decisions. Accuracy does not measure an important concept that needs to be taken into consideration in this type of evaluation: **False Positive and False Negative errors.**

In this problem what type of errors do we care about more? False Positives or False Negatives?

**False Positives (Type I Error):** You predict that the employee will leave, but do not. Consider the employee turnover domain where an employee is given treatment by Human Resources because they think the employee will leave the company within a month, but the employee actually does not. This is a false positive. This mistake could be expensive, inconvenient, and time consuming for both the Human Resources and employee, but can be seen as a good investment for relational growth.

**False Negatives (Type II Error):** You predict that the employee will stay, but leave. Compare this with the opposite error, where Human Resources does not give treatment/incentives to the employees and they do leave. This is a false negative. This type of error is more detrimental because the company lost an employee, which could lead to great setbacks and more money to rehire.

Depending on these errors, different costs are weighed based on the type of employee being treated. For example, if it’s a high-salary employee then would we need a costlier form of treatment? What if it’s a low-salary employee? The cost for each error is different and should be weighed accordingly.

In our employee retention problem, rather than simply predicting whether an employee will leave the company within a certain time frame, we would much rather have an estimate of the probability that he/she will leave the company. **We would rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.**

## retention plan

## conclusion

## future work

This problem is about **people decision.** When modeling the data, we should not be using this predictive metric as a solution decider. But, we can use this to arm people with much better relevant information for **better decision making.**

We would have to conduct more experiments or collect more data about the employees in order to come up with a more accurate finding. I would recommend gathering more variables from the database that could have more impact on determining employee turnover and satisfaction such as their distance from home, gender, age, and etc.

**Modeling (Train Data)**

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 - 20 –

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.

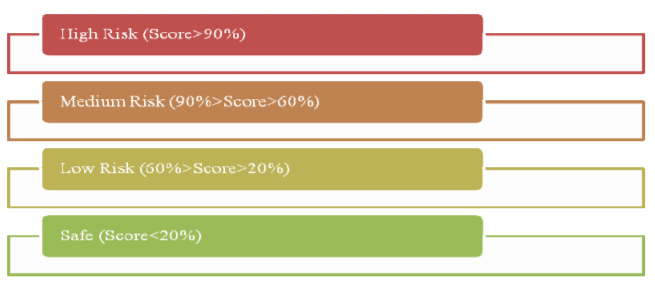
**Test the Attrition Risk Equation on the Data #2 (test data) set**

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 - 21 –

NOTE: Create Confusion Matrix to display HERE

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

**Retention Plan:**

****

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.

**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.