Employee Attrition

# DSC 324/424

Nikki Allen

Goutham Selvakumar

# Abstract

Employee attrition refers to the gradual reduction in employees over time. It occurs when employees retire, resign, or are replaced. The result of a high attrition rate is that employees are leaving a company faster than they are hired. The IBM HR Analytics Employee Attrition & Performance dataset is a fictional dataset generated by IBM data scientists that can be used to compare different factors with attrition. These factors include demographic information, such as age and marital status, wage information, such as daily rate and monthly income, satisfaction metrics, such as job and environmental satisfaction, education information, and performance metrics. With this bevy of information, we are working towards identifying factors from all sectors of an individual’s life, from education to salary to satisfaction, that are most important in determining attrition for a company. This is important to measure not only to maintain a stable working environment, but also overall company growth.

# Exploratory Analysis of the Data

The original dataset on employee attrition contained a total of 1470 observations and 35 distinct variables. This dataset was determined to have no incomplete observations (i.e., no missing values), thus we were able to proceed with the analysis. As we are concerned with the reasoning behind employees leaving a company, we created dummy variables for our categorical variables such that we are able to use them in later predictive model building algorithms. Since the “fastDummies” package creates a dummy variable for each level, we had to remove one level of each variable to avoid perfect correlation.

After removing unnecessary columns that contained identification variables or variables with no variance, we examined each of the continuous variables to ensure their distributions were approximately normal. This was done as many model building techniques require the assumption of approximately normally distributed variables. A few of our continuous variables required a log transformation to satisfy the normality assumption. These variables are MonthlyIncome and DistanceFromHome. However, some of the variables, when transformed, resulted in gaps in the histograms used to visualize their distributions. This stems from the relatively limited range of values these variables can assume. Therefore, we must take into account that these variables are discrete with a limited number of possible values in preceding with our machine learning. The variables that fall into this category are NumCompaniesWorked, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrentManager.

Chart, scatter chart

Description automatically generatedFollowing this, we generated a correlation plot on the transformed dataset. From the plot, it is clear that many of our variables are correlated with one another (Figure 1). Predictive modeling techniques require minimal levels of multicollinearity in order to correctly determine the contributions of predictors to the variation in the dependent variable. Thus, because one of the goals of this analysis is to determine what influences an individual’s decision to leave a company, we will need to take a closer look at these variables later. The bottom right corner of the dataset shows significantly high levels of correlation. YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager are all highly correlated. To address this, we will need to consider either eliminating, or combining the variables when building predictive models.

Figure 1

# Analysis Techniques

Because we want to determine the major factors influencing employee attrition rates, we have elected to run two scaled Principal Component Analyses (PCA) as a means of both dimensionality reduction and feature extraction, followed by a Linear Discriminant Analysis (LDA) to find the separation between observations with individuals who have left the company versus those who have not.

# Scaled Principal Component Analysis

Before building a predictive model, we have utilized Principal Component Analysis as a means of dimensionality reduction. Principal Component Analysis simplifies the complexity in high-dimensional data while retaining trends and patterns by creating new variables that are linear combinations of the original variables in the dataset. Because we determined a difference in units in our continuous variables as part of our exploratory analysis, we have elected to run a scaled Principal Component Analysis. Rather than utilizing the covariance matrix to project the dataset, scaled PCA utilizes the correlation matrix, which will allow us to see the direct relationship between the variables and the components. Since our dataset contains 44 predictor variables, PCA generated 44 components to account for 100% of the variation in the attrition variable. We have reduced the Employee Attrition Dataset to 18 PC’s since each of them have an average variance that is greater than one. Figure 2 shows a portion of the scree plot which shows the variance of the first Chart, histogram

Description automatically generated10 components being greater than one. The first 18 components also account for 72.763% of the variation in the attrition variable. Because dimensionality reduction always results in a loss of information due to distortion, we want to select components that minimize this loss, and retain a maximum amount of the variation.

Through use of dimensionality reduction, we were able to increase the parsimony of our model, which we will use in our Linear Discriminant Analysis.

Figure 2

However, PCA may also be used for feature extraction, which can be applied to this dataset as a way to understand the types of employees that work at a company, regardless of attrition. For feature extraction, we are willing to sacrifice total captured variance in order to boost interpretability. Thus, we will be using the “knee test” to determine how many components to extract. To do this, we look at where the scree plot appears to “bend at the knee”. After applying this test, we elected to use a total of three components. From the output below, it is clear that we have sacrificed quite a bit of variance, totaling only 25.296%. While this is a much smaller proportion than we yielded through use of dimensionality reduction, it reveals that the company is comprised of a wide mix of individuals. This is in fact profitable for a company, as it is then able to gain differing opinions and insight from their employees to yield the best decisions for the organization. To understand the importance of each component, or each type of employee, it is necessary to identify the most significant variable loadings in each of the components. The output of the loadings for the first three components is included here.

A screenshot of a computer

Description automatically generated with low confidenceText

Description automatically generated

Figure 3

From this output, we can see that the most significant component loadings for PC1 are for MonthlyIncome, JobLevel, and TotalWorkingYears. All of the loadings have negative values which indicate that most company workers do not attain such attributes. Thus, this component reveals that within the company there are not many experienced workers who are relatively high in the organizational hierarchy and on the pay scale. This component intuitively makes sense as workers who would have positive values for these variables are often high-level managers or C-Suite members, and these seats in the company are very limited. For the second component, the loadings that are the most significant are TotalWorkingYears, PerformanceRating, and DailyRate. However, these loadings are all positive, indicating that a significant portion of workers do attain these attributes. This information suggests that the company is comprised of quite a few experienced and productive workers, which is comprehensible since a company strives for its workers to be productive and increase their profit margins. The third PC is influenced most heavily by NumCompaniesWorked, YearsInCurrentRole, and YearsSinceLastPromotion. NumCompaniesWorked has a negative loading, suggesting that the there are many employees in the company who have worked for relatively few organizations. On the other hand, both YearsInCurrentRole and YearsSinceLastPromotion have positive loadings indicating that these employees have worked in the same role at the same level of pay for a significant portion of their career.

Overall, through our use of scaled PCA, we were able to reduce our dataset from 44 dimensions to 18 while losing less than 30% of the information in the original dataset. In order to understand the types of employees working at the company, we used feature extraction to understand over 25% of the company’s workforce. We will use the results of this exploratory analysis later to generate a predictive model.

# Linear Discriminant Analysis

Following our Principal Component Analysis, we generated a Linear Discriminant Analysis model which we can use to predict whether or not an individual will leave a company based on specific variable values. As discovered through the exploratory analysis of our dataset, our data does have a significant level of multicollinearity. We did this because the interpretation of an LDA model can be inhibited by significant multicollinearity. To determine the variables that are collinear with one another, we first decided to run a regression model with the Variance Inflation Factor. Variables with VIFs greater than 10 indicate a multicollinearity issue. When running a full regression model, the variables with high VIFs were our dummy variables. Because the categorical variables associated with the variables with high VIFs (Department\_Sales, Department\_Research & Development and JobRole\_Sales Executive) have more than three levels, these VIFs are artificially high due to the number of each of these categories in the original variable.

Chart, line chart

Description automatically generatedAfter addressing the multicollinearity, we build a model on the full dataset. This was done to yield a preliminary model that we will use to evaluate whether the dimensionality reduction from scaled PCA will produce a statistically significant and better model. After running the full LDA, we used the test set to predict attrition and then used those results to yield a confusion matrix. The model was able to correctly classify an individual’s attrition 86% of the time. To evaluate the usefulness of this model we generated an ROC curve and printed the AUC value (Figure 4). Our AUC value was 0.788 which suggests that the model has a 78.8% chance of correctly distinguishing between positive and negative classes.

Figure 4

In analyzing the model, we have concluded that for management purposes, we would prefer to have a type two error over a type one error. Essentially, we would rather misclassify an individual as leaving the company who does not as opposed to misclassifying someone as staying who leaves. This conclusion was drawn because managers will want to focus on increasing the level of retention of employees and target those areas that are most influential in an individual’s decision to leave a company. While it is undoubtably important that managers continue to please individuals who will remain in the company, to lower attrition rates, they will want to prioritize increasing the probability that their employees that are contemplating leaving remain in the company.

Chart, line chart

Description automatically generatedTo determine if our dimensionality reduction techniques produce better results and a higher level of accuracy, we have generated a model on a data set created from the first 18 principal components. We combined the principal components with the Attrition variable, and then created a new 75-25 training-testing split for the new dataset. Following this, we generated the model, and used it to make predictions on the testing set. From the confusion matrix, we found that this model has an accuracy level of 85% on the testing data. This is lower than our original LDA, however this decrease in accuracy makes sense because the first 18 principal components only account for around 72.763% of the variability in the attrition variable, while the full LDA model accounts for a larger portion of the variation. We once again created an ROC curve and the AUC value for the new LDA model (Figure 5). Our results showed that this model has an AUC of 0.824 which suggests that the model has an 82.4% chance of correctly distinguishing between positive and negative classes, which is higher than the full model. This indicates that the dimensionality reduction from scaled PCA yields better classification between someone leaving versus staying at a company.

Figure 5

The output of the LDA shows that the component with the largest positive effect on attrition is PC18, with a value of 0.411225789, and the component with the largest negative effect is PC14, with a value of - 0.433281260. From the component loadings in our scaled PCA, we have determined that PC18 is characterized by JobInolvement, RelationshipSatisfaction, and TrainingTimesLastYear. All have negative loadings which suggests that employees who are heavily involved with their jobs, participated in job training, and are satisfied in their personal relationships, have low probabilities of leaving a company. While management cannot control an employee’s outside relationships, they may consider instituting more training which will in turn increase the level of involvement one has in their occupation and lower the likelihood of attrition. On the other hand, PC14 is influenced primarily by MonthlyRate, suggesting that the higher monthly rate a person has, the less likely they are to leave a company. This intuitively makes sense, and management should continuously evaluate how they are compensating their employees to ensure they are given proper monetary reward for their efforts.

Overall, the LDA on the components attained from scaled PCA yielded better results than our LDA on the original dataset. Thus, by combining the predictors into new components, we are able to better understand the most important factors influencing attrition rates.

While from our original 44 variables, using 18 components does greatly reduce the dimensions of our original set of variables, to increase interpretability, we would like to build a model using the components that we have interpreted through use of our feature extraction from PCA. Thus, this model will be built from the three principal components we identified through use of the knee test on our scree plot. The model equation is provided below.

***Attrition*** = 0.3680094***PC1*** - 0.3395980***PC2*** - 0.1041073***PC3***

Chart, line chart

Description automatically generatedUsing this equation in combination with our previous interpretation of the feature extraction, we can determine the types of employees that are more or less likely to leave a company. The most significant loading is the positive loading of PC1, which suggests that higher level, well-compensated employees, possibly CEOs or CFOs, are less likely to leave a company. The most significant negative loading is that of PC2 which suggests that experienced and productive employees are more likely to leave an organization. This may be a result of these individuals feeling that their contributions to the company are not being adequately compensated, and thus they desire to find employment elsewhere. Another possibility is that these employees, since they are characterized as having a higher number of working years, are retiring, and thus leaving the company not to find new employment, but to stop working completely. Lastly, PC3 has a slightly less significant loading, but it is also negative. This suggests that employees with little experience elsewhere, who have worked in the same role and at the same level for many years, are more likely to leave a company. This may be because they feel that they have plateaued in the company, and still desire to grow their careers. The results from the confusion matrix resulted in an 84.42% rate of correct classification. The AUC value of the ROC curve (Figure 6) was 0.727 which indicates this model has a 72.7% chance of correctly classifying between attrition and non-attrition. This value is about 10% lower than our first LDA model generated from our scaled PCA data. However, the decrease in complexity of the model was 15 dimensions, and thus we are willing to sacrifice some predictability power in favor of attaining a better interpretation of the model.

Figure 6

To gain a better understanding of the most influential variables in employee attrition, we have elected to run two more LDA models with a reduced selection of the original variables in our dataset. The results from our exploration through scaled PCA have shown that the largest amount of variation in employee attrition depends on factors influencing income, such as MonthlyIncome, JobLevel, and TotalWorkingYears (all of which are most influential in the first scaled component). Because of this, we have decided to run an LDA with a set of variables all relating to income. The model equation is provided below.

***Attrition*** = -.000104***DailyRate*** + .377242***JobLevel*** -1.7439***MonthlyIncome*** -.0000276***MonthlyRate*** -.062965***PercentSalaryHike*** +.28380***PerformanceRating*** + .10896***NumCompaniesWorked*** -.03034***YearsInCurrentRole*** - .07399***YearsWithCurrManager***

Chart, line chart

Description automatically generatedFrom the coefficients of the linear discriminants, the most significant variable decreasing the probability of attrition is MonthlyIncome, which confirms the results of our other methods; that one’s level of monetary compensation is one of the leading factors in the individual decision to leave a company. On the other hand, this model suggests that the factor that is the most influential in increasing the probability of attrition is JobLevel. As discussed earlier, individuals attaining higher level roles must be adequately recognized and compensated for their contributions to the company. If they are not, they have higher incentive to find an employer who recognizes their skills, and therefore, they have a higher probability of leaving the company. To evaluate the usefulness of this model, we have repeated the generation of a confusion matrix and an ROC/AUC curve (Figure 7). Our confusion matrix resulted in an 83.95% rate of correct classification. However, our AUC curve resulted in a chance of only 64.5% in correctly classifying between attrition and non-attrition. This suggests that the reduced selection of variables lacks a significant portion of the variance that explains attrition rates. It is nearly 20% lower than the LDA model generated from scaled PCA.

Figure 7

The next grouped selection of variables we considered were those influencing employee satisfaction. The idea behind this selection is the hypothesis that employees who are satisfied in their jobs and their environments are less likely to consider leaving. The resulting model equation is represented below.

***Attrition*** = -0.4218***JobSatisfaction***– 1.2748***Monthly Income*** -0.0494***RelationshipSatisfaction*** – 0.3468***EnvironmentSatisfaction*** - .0426***PercentSalaryHike*** + 0.2265***PerformanceRating*** – 0.3979***WorkLifeBalance***

Chart, line chart

Description automatically generatedThe only variable in this model with a positive coefficient is PerformanceRating. Once again, this may be a result of the individual lacking the belief that their performance is being adequately recognized. However, this variable as a smaller loading, and is less influential to the probability of attrition. Once again, satisfaction resulting from higher levels of income is supported by the large negative loading on the MonthlyIncome variable. Also important to recognize, is that all other variables have negative loadings, supporting the hypothesis that higher levels of satisfaction, pertaining to not only the job itself, but also to the work environment and outside relationships, effectively decrease the probability of attrition. Once again, we evaluate the usefulness of this model. This model had a slightly lower correct classification rate from the confusion matrix at 83.4%. However, the AUC of this model was slightly higher than the income variable model at 66.4% (Figure 8).

Figure 8

From the results of these models, we can see that while they both explain a similar amount of the variation in attrition rates, the linear combinations of these original variables provide the best model for predicting attrition for employees.

# Conclusion and Practical Significance

Our models may potentially be used by corporations to help them define the areas that most influence their attrition rates. This will then grant them the ability to cater to the needs of particular types of employees, specifically through use of the LDA model built on the extracted features from scaled PCA. Our models also enables organizations to use the most influential factors in attrition rates to determine the areas in need of improvement in their own organization using not only the full LDA model, but also those built from the variables influencing income and employee satisfaction.

However, one limitation of this study is that the two final models we built are not as powerfully predictive on their own, but rather in combination. Thus, the data analytics department of any organization may use our models as a base to build their own employee attrition models with the factors that are most relevant in their own organization.

# appendix

**Appendix A – Link to dataset**

<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

**Appendix B – Supplemental Charts, Graphs, and Tables**

Table

Description automatically generated with low confidenceAppendix A.1 – VIF Values from Regression Model

Appendix A.2 – Confusion Matrices from LDA Models

Text

Description automatically generated with medium confidenceA picture containing text

Description automatically generatedA picture containing text

Description automatically generatedFull LDA Model Income Metrics LDA Satisfaction Metrics LDA

A picture containing table

Description automatically generatedA picture containing text

Description automatically generated18 Components from scaled PCA LDA 3 Extracted Features from PCA LDA