Predicting Employee Attrition Using Cost Sensitive Classification

Presented By:

Varun Agarwal

Nikitha Udaykumar Chettiar

Overview

Attrition is defined as the gradual reduction of the volume of employee in an organization, and not replacing their positions [1]. This could occur due to employees resigning or retiring, and is seldom associated with layoffs, too. For a company, attrition rate being high is unfavourable for reasons since it leads to loss of talent, poor ROI, and resources spent on training employees. Also, lesser volume of employees leads to low throughput and performance of the company. Hence, organizations strive to reduce the attrition rate. It is important to analyse the factors that influence the attrition in the company, and thereby, curb these factors and reduce attrition rate.

Objective of the Study

- Understand the influence of the various factors or predictor variables on the predicted variable, the degree of influence and the relation between different attributes.
- Implement a classification model for the dataset with good prediction results.

Data

For this study, we have chosen the IBM HR Analytics Employee Attrition and Performance Attrition dataset, which has been obtained from Kaggle [2]. It is a fictional dataset created by IBM data scientists, which contains data related to employee performance measures and attrition. The dataset contains several predictor variables, having varying influence on the predicted variable Attrition which signifies whether an employee left the company or not.

A breakdown of the variables and their types is as follows:

- Predicted Variable: Attrition (Yes or No)
- Predictor Variables: 24 continuous variables, and 8 categorical variables
- Total instances: 1470 rows

One of the challenges this dataset puts forward is the imbalance in the class distribution. Out of the total 1470 tuples in the dataset, only 237 tuples have their predicted class as “Yes”.

Pre-processing

The dataset contains three columns, namely *Employee Count*, *Over 18* and *Standard Hours*, which have the same values throughout the data. We remove these features from our dataset, since they do not provide any value to our prediction. Apart from this, the dataset is uniform throughout, and we have no missing values or nulls for any of the columns.

Visualizing the dataset

We will explore our dataset visually, to understand our data in a more intricate manner. Firstly, let us visually address the class imbalance problem in the data.

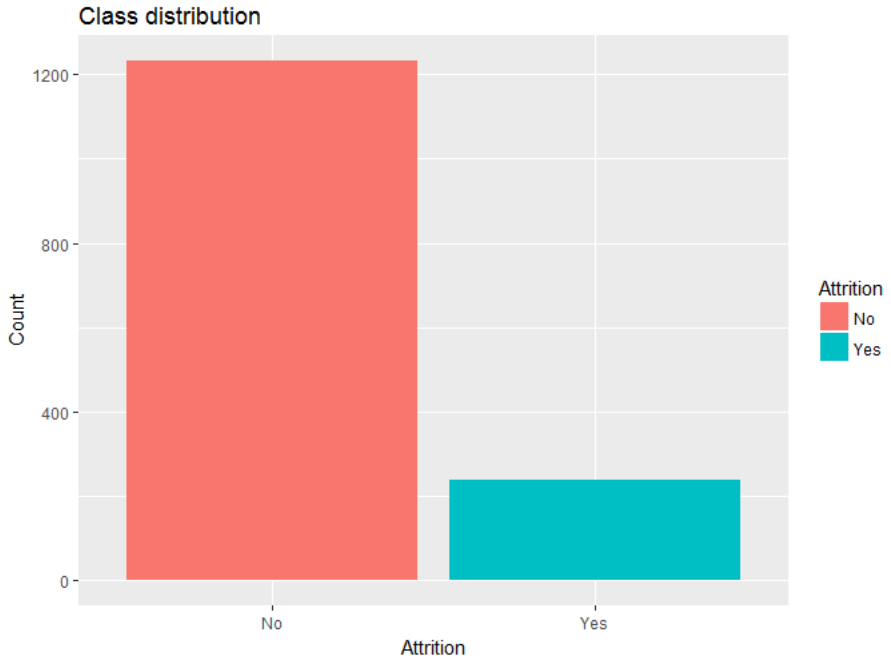

Figure 1

In the Figure 1, we can clearly see the class imbalance in the retrieved data. This imbalance is a major concern in our study, and affects our modelling choice.

Let us also check the correlations between the continuous variables in the model.

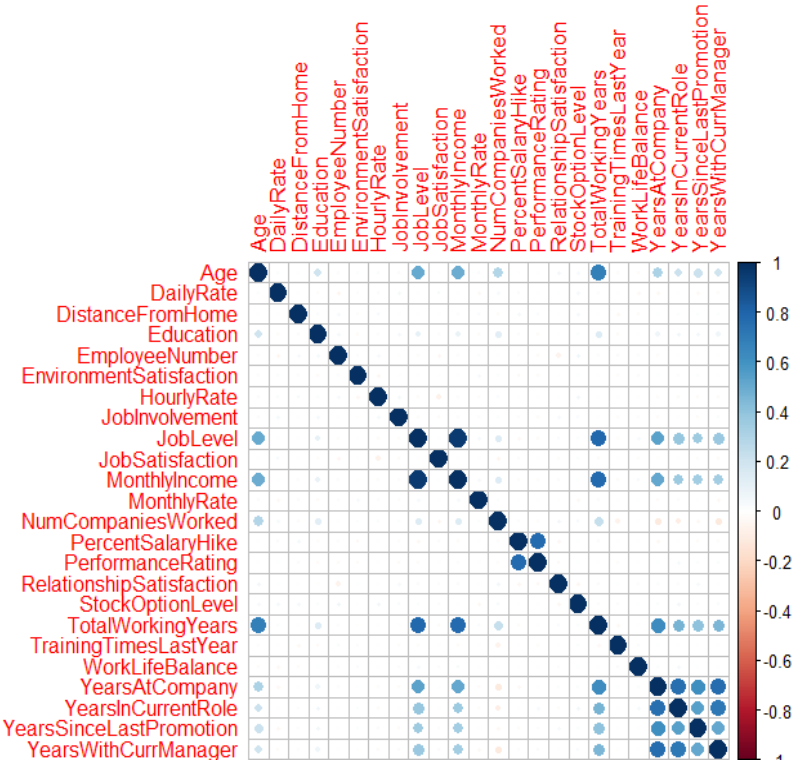

Figure 2

In the correlation plot, we see that variables *Job Level* and *Monthly Income* are correlated. Other than these two variables, there does not appear to be significant correlation among any other predictor variables of the continuous type.

For further exploration, we are going to look at the distribution of class with respect to a few variables, which we perceive to be the most important predictor variables in our study.

We have plotted the density distribution of attrition according to different age groups, set on a bin width of 10 years.

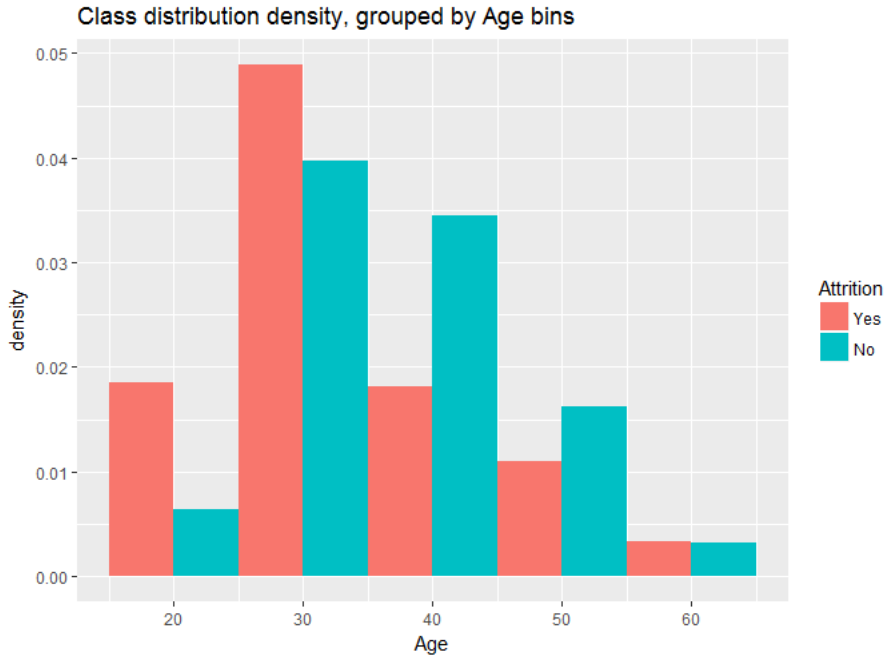

Figure 3

In Figure 3, we can see that the distribution of class according to age is right skewed. We see that the maximum density of attrition is focussed on the age group of 30 – 40 years. So, age seems to be a potentially important predictor variable for our task.

We have plotted the density distribution of attrition according to different Monthly Incomes groups, set on a bin width of $5000.

Figure 4

In Figure 4, we again see a right skewed distribution of Attrition being positive. This tells us that employees who have low monthly salaries are more likely to leave the company. Specifically, employees who have a monthly salary around $5000, are most likely to leave the company. On the other hand, employees who have a higher monthly salary are less likely to leave the company. Hence, monthly income of the employees will be important for our predictions.

Experimental Setup

Most of the classification models focus on the loss function and they do not take the data distribution into consideration. [3] The emphasis of the classification models is to reduce the error rate, which is not affected much by the minority class in the dataset. As determined earlier, our dataset suffers from class imbalance. Thus, our selection of the model to be implemented must be careful, one which can account for the imbalance in the dataset. Naïve implementation of classification techniques such as Random Forest or Gradient Boosting would not be the best techniques for this dataset. The rationale is that, although the resultant accuracy for these models would be high, this would be done at the cost of increasing false negatives (Positive class being *Attrition = 'Yes'*).

Following are the methods that are generally used to handle imbalances in datasets: [4]

1. Under sampling the majority class in the dataset
2. Oversampling the minority class in the dataset.
3. Creating synthetic data to increase number of tuples for the minority class such as SMOTE.
4. Cost sensitive classification.

We analysed each of these methods to select the method which would be apt for the dataset. Undersampling could sometimes lead to loss of information, which could be vital for classification, while oversampling could lead to high generalization error rate.

On the other hand, cost sensitive classification met more with our objective of associating different costs to the predicted values to reduce the overall risk. The optimization using cost sensitive function emphasizes on reducing the cost incurred by the misclassification.

For our study, we have built the following cost matrix:

	Predicted Positive	Predicted Negative
Actual Positive *(Attrition = 'Yes')*	$C(TP) = \mathbf{0.6}$	$C(FN) = \mathbf{0.9}$
Actual Negative *(Attrition = 'No')*	$C(FP) = \mathbf{0.2}$	$C(TN) = \mathbf{0.0}$

Table 1

To reduce the possibility of false negative predictions, $C(FN)$ is associated with the highest cost. The constraint on the cost matrix for this dataset is such that $C(FN) > C(TP) > C(FP) > C(TN)$. The reasoning behind this is that, a company would be at a higher risk if it were to classify an employee as unlikely to resign from the company, when he is likely to resign. Cost of true positives are set lower than the cost of false negatives. This is because the company would still have a chance to talk to employees likely to contribute towards attrition rate, and try to change their views on resigning from the company. In case of false positives, the HR representatives would incur the cost of conversing with the employee classified in this manner, only to find out that he is not at a risk of resigning.

Since we are implementing cost-sensitive classification technique, we use F_1 scores as our performance metric, since accuracy is not sensitive to the risk function. F_1 score is the harmonic mean of precision and recall. F_1 score focusses on balance of the precision and recall and hence is an appropriate performance metric when there is a class distribution is imbalanced [5].

To implement our case of cost-sensitive classification, we will use logistic regression. It is a widely-used method in data mining problems. Logistic Regression is appropriate for this dataset as it works for binary predicted variable. It outputs the probability of the predicted variable class for a given test case.

We will test two cases of logistic regression, one with default probability threshold (0.5), and one which uses empirical thresholding technique to optimise F_1 scores, as given in link [6]. The test and training set for both the models have been split in a 70-30 ratio, obtained through random sampling.

Results

Resultant performance measures without parameter tuning:

Threshold = 0.5	**F_1 Score**	**Misclassification rate**
Training Set (70%)	0.6148148	0.1009709
Test Set (30%)	0.6086957	0.1022727

Table 2

Resultant performance measures after threshold parameter tuning:

Threshold = 0.4576097	F_1 Score	Misclassification rate
Training Set (70%)	0.6263345	0.1019417
Test Set (30%)	0.6065574	0.1090909

Table 3

From Tables 2 and 3, we can see that parameter tuning does not have a very significant effect on the performance of our model. In fact, we see a very slight decrease in the F_1 score after tuning the threshold parameters. The misclassification rate is about 6% lower than that for random guessing (since, the majority class is approximately 84% of the data).

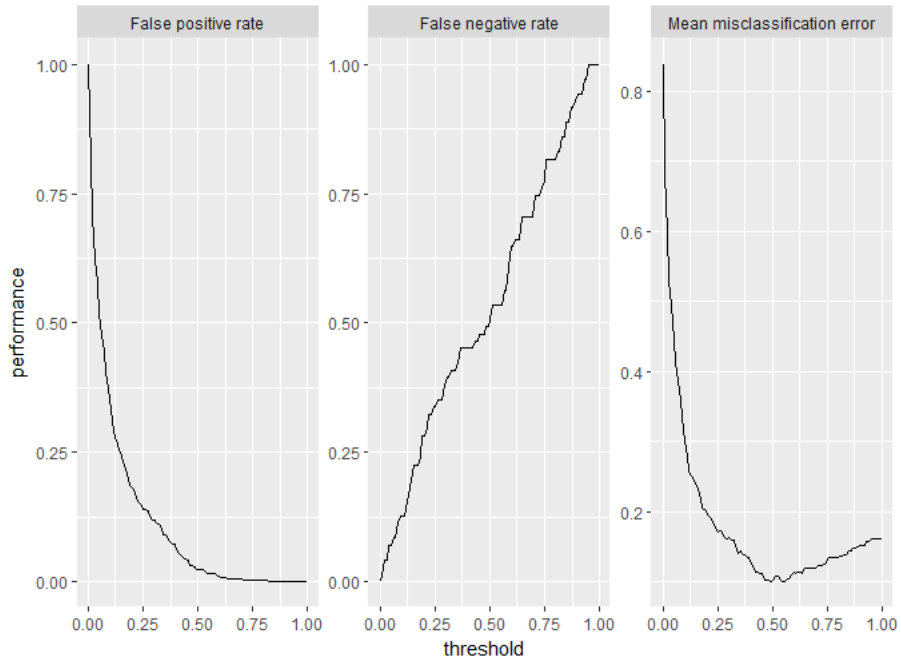

Figure 5 – Performance of the model based on FPR, FNR and Mean misclassification error for different values of thresholds

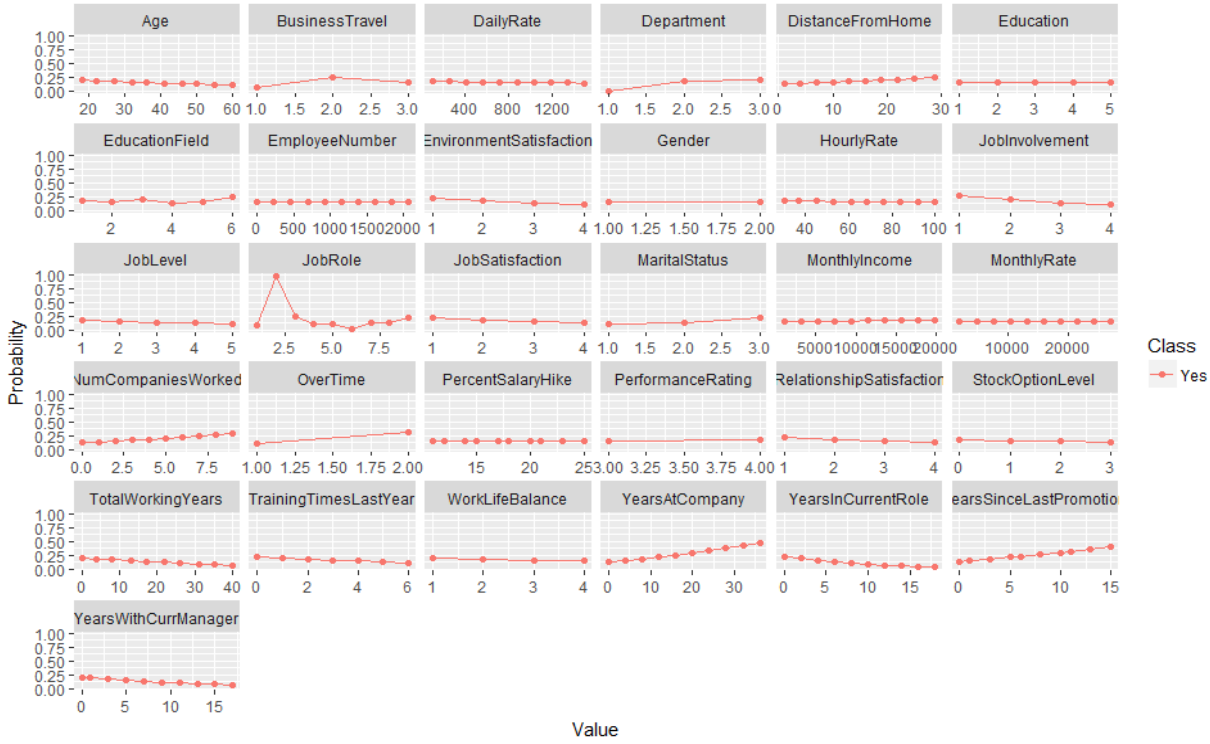

Figure 6 – Partial dependence of features

The Figure 6 shows the partial dependence of features on our predictive power of the model with empirical thresholding. Overall, we see that there is not much variation in the probability of attrition of an employee based on the various features we have selected in the model. Although, we do see that the number of years at the company and the number of years since last promotion matter the most in predicting the probability of attrition. An unexpected find is that age and monthly income parameters do not contribute to much change in the probability of attrition.

Conclusion

Our model seems to be performing reasonably well, for the given number of data points. The dimensionality of the data certainly affects the performance, and can be improved by collecting more data. The misclassification rate is about 6% better than that of random guessing, which is reasonably well for unbalanced classes.

References

[1] Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Attrition.

[2] P. Subhash, "IBM HR Analytics Employee Attrition & Performance," [Online]. Available: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[3] Wiki.UNL. [Online]. Available: http://cse-wiki.unl.edu/wiki/index.php/Cost_Sensitive_Learning.

[4] A. V. C. Team, "Analytics Vidhya," [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/.

[5] N. U. Y. G. Shameem A. Puthiya Parambath, "Optimizing F-Measures by Cost-Sensitive Classification".

[6] mlr, "Cost-Sensitive Classification," [Online]. Available: https://mlr-org.github.io/mlr-tutorial/release/html/cost_sensitive_classif/index.html.