Analysis of Employee Attrition Using Classification Models

Boris Chan, Brendon Chae

Abstract

We evaluate three methods: logistic regression, naive bayes, and decision tree classifiers as predictive models for predicting employee attrition based on IBM HR dataset. We used a comparison of measures, k-fold cross validation, and hypothesis testing to determine which method produced the 'best' classifier. Our findings suggest that each model is recommended in different situations depending on which measures value most to the organization.

1. Introduction

In this research paper our goal is to tackle employee attrition using the IBM HR dataset. Attrition is essentially the turnover rate of employees inside an organization. It can happen for a multitude of reasons, which may include: negative working environment, poor management, employee sickness, overworking, and employees looking for better opportunities. These reasons as well as the information from previous research will be kept in mind while conducting our study. The issue of keeping an employee satisfied is a challenge for all employers. To be able to decrease attrition, the managers must understand the causes and know the costs of staff turnover. The attrition rates would generally differ across different organizations, so we will focus on IBM using the HR dataset. We will implement a predictive model using different classifiers in order to determine if an employee is going to quit the organization or not.

2. Literature Review

In the 2009 paper by Yazinski, there were 7 main reasons identified apart from salary that are factors in employee turnover. They are as follows [7]:

1. Employees feel the job or workplace is not what they expected
2. There is a mismatch between the job and person
3. There is too little coaching and feedback
4. There are not enough growth and advancement opportunities
5. Employees feel unrecognized and devalued
6. Employees feel overworked and stressed and have a work/life imbalance
7. There is a loss of confidence and trust in senior leaders

In the paper "People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis" by Hofman and Tadelis, they focus on the managers of the organization in a large firm. They answer: how much does on particular set of skills, namely, people management skills, or interpersonal skills for dealing with one's subordinates matter? [8] It measures the managers

people management skills by taking employee surveys about their manager. Then they see the relation to employee attrition. They find a negative relationship between the manager overall rating and employee attrition. [8]

3. Algorithms

The Naive Bayes classifier is based on Bayes' theorem with strong independence assumptions between predictors. Bayes' theorem provides a way of calculating posterior probability P($X \mid C$) from P(C), P(X), and P($C \mid X$). To reduce the computation cost in evaluating P($X \mid C_i$), the naive assumption of class-conditional independence is made, where the naive bayes classifier assumes that the components of X on given class C are independent, and then estimates each of the univariate distributions separately [1, pp. 351-354] [2].

Decision tree induction is a method which utilizes a tree model of decisions where each branch represents an outcome of the test, and each leaf holds a class label. A trained decision tree will test a given tuple X (where the class is unknown) using its attributes and a path will be traced to a leaf node which holds the class prediction for X. The CART method produces a binary decision tree, which uses the Gini Index as its splitting rule [1, pp. 332-350]. This can be contrasted with the ID3 method which uses information gain as its attribute selection measure.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

Logistic regression is a method to predict the probability of an outcome that can only have two values (e.g. yes or no). The prediction is based on the use of one or several predictors, which can be numerical or categorical. The logistic regression produces a logistic curve, which values are limited between 0 and 1 [3]. The model can be defined with:

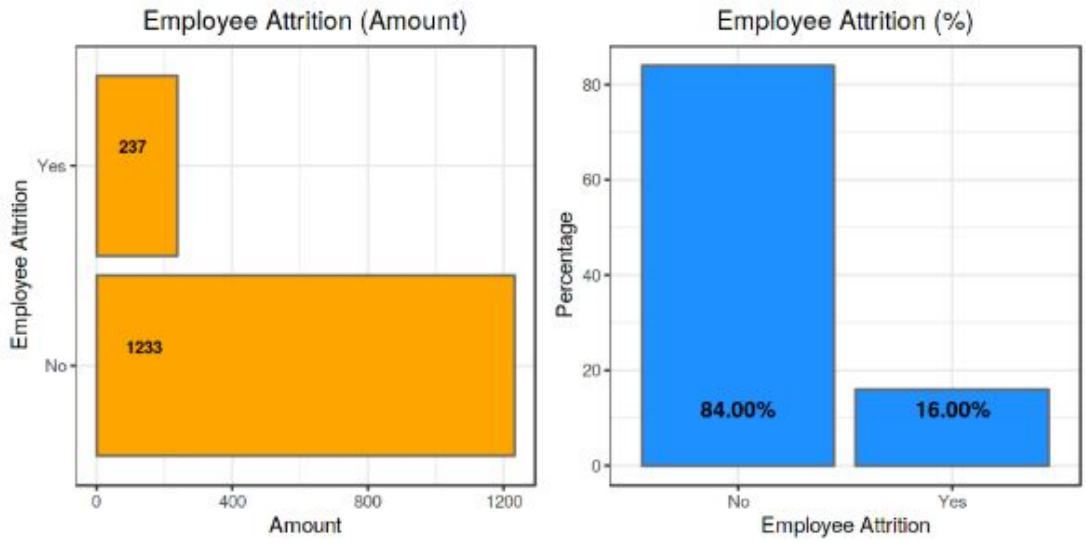
$$P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Where P is the probability that an observation β_0 is the constant and β_1 the slope defines the steepness of the curve. The logistic regression model will be used as the baseline model for comparing the results of the Naive Bayes classifier and the decision tree classifier.

4. Dataset and Preprocessing

The dataset used in this paper contains 1470 tuples with 35 attributes. Although the dataset is relatively small, the 35 attributes are extensive in terms of describing employee information. The Graph 2.1 shows the total Employee Attrition as well as the percentage. We can observe an imbalance in the dataset.

Graph 2.1 Employee Attrition in IBM



There are attributes related to the personal details of the employee, e.g. age, gender, education, etc. in addition to details within the business domain, e.g. job role, involvement, satisfaction, etc. Categorical attributes such as *Education, Environment-Satisfaction, Job-Involvement, Performance-Rating, Relationship-Satisfaction* and *Work-Life-Balance* have been preprocessed with dummy coding, as illustrated in the table 4.1.

Table 4.1 - Categorical attributes mapped to dummy variables.

	Dummy variable				
	1	**2**	**3**	**4**	**5**
Education	Below College	College	Bachelor	Master	Doctor
EnvironmentSatisfaction	Low	Medium	High	Very High	n/a
JobInvolvement	Low	Medium	High	Very High	n/a
PerformanceRating	Low	Good	Excellent	Outstanding	n/a
RelationshipSatisfaction	Low	Medium	High	Very High	n/a
WorkLifeBalance	Bad	Good	Better	Best	n/a

The dataset also contains the binary attribute *attrition*. As this paper focuses on using classification techniques on the dataset, the *attrition* attribute will be used as the class attribute when building the models. The dataset will be split into a training set and testing set with an approximately 70/30 split.

5. Method/Experimental Results

We trained and constructed the logistic regression model, naive Bayes classifier and decision tree classifier using the training set with Python. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [1, p. 365], which can be denoted by:

$$accuracy = \frac{TP + TN}{P + N}$$

Once the initial accuracy scores have been computed, the accuracy scores will be compared to the accuracy scores using k-fold cross validation. The following table illustrates the initial accuracy scores for the logistic regression model, naive bayes classifier and decision tree classifier.

Table 5.1 - Initial accuracy scores from algorithms.

Model	**Accuracy (3 s.f.) (%)**
Logistic Regression	87.243
Naive Bayes	73.457
Decision Tree	76.955

The initial results show that the logistic regression model resulted in the highest accuracy score, followed by the decision tree classifier and lastly the naive bayes classifier. The baseline model outperformed the naive bayes and decision tree classifier by 10%. It is important to note that the accuracy scores are not always the same. This is because the training and testing set are selected at runtime, therefore different samples are selected every time the code is executed and hence affect the model. A technique called k-fold cross validation can be employed to find a mean accuracy score at a 95% confidence interval. K-fold cross validation analysis will be described in the evaluation.

6. Evaluation

6.1 Model selection

A confusion matrix can be used to evaluate the quality of a classifier. The matrix shows the true positives, true negatives, false positives, and false negatives. The measures used to assess a classifier predictive ability are: accuracy, recall, and precision. It should be noted that the accuracy measure might not be reliable when the main class of interest is in the minority [1, p. 386] Precision is a measure of exactness, whereas recall is a measure of completeness [1, p. 386]; these measures can be computed by the following formulas:

$$precision = \frac{TP}{TP + FP} \text{ , } recall = \frac{TP}{TP + FN} \text{ ,}$$

Table 6.1 - Confusion matrix (Logistic Regression)

	Predicted 0	**Predicted 1**
Actual 0	412	4
Actual 1	59	11

Table 6.2 - Confusion matrix (Naive Bayes)

	Predicted 0	**Predicted 1**
Actual 0	318	98
Actual 1	31	39

Table 6.3 - Confusion matrix (Decision Tree)

	Predicted 0	**Predicted 1**
Actual 0	353	63
Actual 1	48	22

Table 6.4 - Precision, recall and F-1 scores for models (5 s.f.)

Model	**Precision**	**Recall**	**F Score**
Logistic Regression	0.73333	0.15714	0.25882
Naive Bayes	0.28467	0.55714	0.37681
Decision Tree	0.25882	0.31429	0.28387

Precision can be used as a measure for comparing models when the cost of false-positive is high. In employee attrition, a false positive means that an employee that is not going to leave (actual negative) has been identified as likely to leave (predicted positive). Similarly, recall can also be used as a measure when the cost of false-negative is high. If an employee that is going to leave (actual positive) is predicted as not going to leave (predicted negative). Both situations can result in negative consequences depending on how the organization uses this information. Table 6.4 illustrates that the logistic regression model offered the highest precision measure, whereas the naive bayes classifier offered the highest recall measure.

The f-score measure is the harmonic mean of precision and recall, where it gives equal weight to precision and recall. When a model has perfect precision and recall, the f score would be 1, and

when a model has the lowest precision and lowest recall, the f score would be 0. The results above show that the Naive Bayes classifier has the highest f-score, followed by the decision tree classifier and lastly the logistic regression model [6]. The f-score should be considered when scoring a model when an organization values both the precision and recall and their respective implications.

6.2 Evaluating models

In k-fold cross validation, the dataset is randomly partitioned into k subsets, $D_1, D_2, \ldots, D_k$. Training and testing is done k times. In iteration i, subset D_i is reserved as the test set, and the remaining subsets are used to train the model. For instance, when $i = 1$, $D_2, \ldots, D_k$ will be used to train the model and D_1 will be used as the testing set; then when $i = 2$, $D_1, D_3, \ldots, D_k$ will be used to train the model and D_2 will be used as the testing set, etc. [1, p. 370].

This paper will use stratified cross-validation, where the folds are stratified to ensure that each fold is a good representative of the whole. For example, in a binary classification problem where each class consists of 50% of the data, it is best to arrange the data so that every fold follows this distribution [4]. A study conducted by Kohavi concludes that using 10-fold stratified cross validation is the best method to use for model selection, in terms of of bias and variance when compared to regular cross-validation [5].

Table 6.5 - Sample mean accuracy for 10-fold cross validation (5 s.f.)

Model	**Accuracy**	**Error Rate**	**Margin of Error**
Logistic Regression	0.85037	0.14963	0.0079624
Naive Bayes	0.75914	0.24086	0.025625
Decision Tree	0.74287	0.25713	0.031776

For a given model, the individual error rates calculated in the cross-validations may be considered as different independent samples from a probability distribution, where they follow a t-distribution with $k - 1$ degrees of freedom (where $k = 10$ from the 10-fold cross validation). Our hypothesis is that two models (Logistic Regression and Naive Bayes) are the same (the difference in mean standard error is 0). If we reject this hypothesis (null hypothesis), then we can conclude that the difference between the two models is statistically significant, in which case we can select the model with the lower error rate [1, p. 372].

$$t = \frac{err(M_1) - err(M_2)}{\sqrt{var(M_1 - M_2) / k}}$$

The t-test for the means from the 10-fold cross validation of the Logistic Regression model and Naive Bayes classifier resulted in a t-statistic of: 7.69080 and a p-value of 0.0000000427642. From

the t-table shows that the critical value 2.262 (at $k-1$ degrees of freedom). Our t-statistic is greater than the critical value, and therefore we can reject the null hypothesis and can select the model with the lower error rate, that is the Logistic Regression model.

The results of the k-fold cross validation show that the mean accuracy measures for the logistic regression model and decision tree classifier had a minor decrease in accuracy (compared to the results in table 5.1), whereas the naive bayes model had a minor increase in accuracy. Overall, the logistic regression model still has the highest accuracy score of all the models.

6.3 Evaluating the dataset

One of the major weaknesses with the dataset is that it does not describe or contain temporal data. Temporal data within the employee attrition domain would be useful to understand how long each employee stayed at the organization for, as well as external contextual factors such as an economic recession or if the organization was going through structural changes during that time period.

With the dataset and focus on attrition we can conclude that it is two-class data. The dataset is imbalanced, as shown in Graph 2.1 we see that in regards to attrition 84% of employees return a 0 - No and 16% return a 1 - Yes. Traditional classification algorithms assume that the cost of false positive and false negative errors are equal, which is not the case for our class-imbalanced data. Some general approaches in improving classification accuracy include oversampling, undersampling, and threshold moving. Given our dataset, we cannot use the oversampling and undersampling methods as it is unrealistic if the tuples are equal. However, threshold moving would be possible if we increase the threshold in which our classifier would return a positive result. Therefore, there would be less chance of a costly false negative error [1, p. 384].

7. Conclusion

In conclusion, the best model depends on the metric that is most relevant to the organization. In hypothesis testing, the logistic regression model was the best choice because it had the lowest error rate. However, with measures for the models, the Naive Bayes was the best model because it had the highest f-score. This shows that the best model depends on the metric used to evaluate them.

There were a couple obstacles found (mentioned in section 6.3) that prevented the predictive model from being accurate to a higher degree. The model should not be used to classify a single employee, but rather identify a group of potential employees that are more likely to leave based on certain characteristics. Our predictive model contains too many weaknesses for it to be effective in real-world situations. However, given more time, we could collect more data, use different metrics, and try different algorithms and models (such as ensemble methods) to improve the predictive capabilities so that it could be considered useful for the future.

8. References

[1] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann; 3 edition (July 6, 2011), 2011.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, and etal., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, 2007.

[3] "Logistic Regression", Saedsayad.com, 2019. [Online]. Available: https://www.saedsayad.com/logistic_regression.htm. [Accessed: 27- Mar- 2019].

[4] "Understanding stratified cross-validation", Cross Validated, 2019. [Online]. Available: https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation. [Accessed: 27- Mar- 2019].

[5] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995.

[6] "F-Score Definition," DeepAI. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score. [Accessed: 27-Mar-2019].

[7] S. Yazinski, "Strategies for Retaining Employees and Minimizing Turnover", *Hr.blr.com*, 2019. [Online]. Available: https://hr.blr.com/whitepapers/Staffing-Training/Employee-Turnover/Strategies-for-Retaining-Employees-and-Minimizing-. [Accessed: 23- Mar- 2019].

[8] M. Hoffman and S. Tadelis, "People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis", NBER, 2019. [Online]. Available: https://www.nber.org/papers/w24360. [Accessed: 23- Mar- 2019].