Employee Attrition for Healthcare

P14 - Boscosylvester John Chittilapilly, Shlok Vivek Naik, Saksham Pandey
Department of Computer Science, North Carolina State University,
Raleigh, NC, 27695
bchitti, snaik2, spandey5@ncsu.edu

https://github.ncsu.edu/spandey5/ALDA2022_EmployeeAttrition_P14

1 Introduction

Employee attrition is one of the major issues faced by companies in recent times due to unpredictable markets and an unyielding workforce. It can be described as the departure of employees from an organization due to reasons other than termination of employment by the organization. It can occur due to a multitude of reasons such as incompetent environment, unsatisfactory compensation, physical factors such as duration of work, location, not enough opportunity for personal and professional growth etc. A high attrition rate would signify that more employees leave the organization than those being recruited. This can be harmful for the organization in several ways.

1. Reduction in employee morale due to insecurity and confusion surrounding mass departure.
2. Hit to the brand reputation of the organization for not being able to satisfy their employees.
3. Transfer of resources to prevent high attrition rates when they were originally meant for business functions
4. Decrease in quality of workforce due to departure of premium and senior employees.

2 Background

Since the problem of High Employee Attrition Rates has been around for a while, there was an opportunity to take inspiration from the abundant amount of research work being done in this domain. In the following paper: Shankar, R. S., Rajanikanth, J., Sivaramaraju, V. V., Murthy, K. V. S. S. R. (2018, July). Prediction of employee attrition using data mining. In 2018 ieee international conference on system, computation, automation and networking (icscan) (pp. 1-8). IEEE. The authors have used Decision Tree, Logistic Regression, SVM, KNN, Random Forest, Naïve Bayes as base models to implement feature selection on human resource data in order to predict Employee Attrition Rates. In another paper, Srivastava, D. K., Nair, P. (2017, March). Employee attrition analysis using predictive techniques. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 293-300). Springer, Cham. The authors have used clustering along with Artificial Neural Networks to classify employees who might potentially leave the organization and have also done extensive pre-modelling analysis to deduce the reasons behind high attrition rate. Taking inspiration from this, we decided to use SVM, Naive Bayes and XGBoost along with PCA enhanced versions of these models in order to determine which one will be able to give the most accurate results. We have also used a cost matrix, used in conjunction with the confusion matrix to determine the model which generates the best cost and will result in the most appropriate prediction outcomes according to every organization's specific use case.

3 Methodology

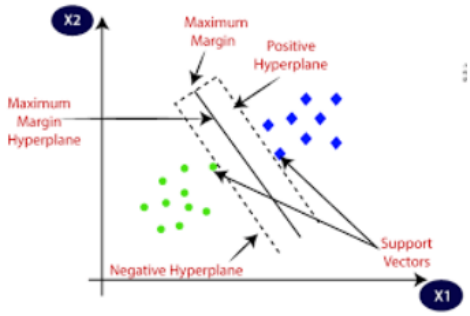
This section highlights the methodology and the detailed information on some of the classification models we have used for our employee attrition dataset. We have implemented three different models - Naive Bayes, Support Vector Machine (SVM), and XgBoost. For the SVM and XgBoost models, we have also implemented a variation with applying PCA to the data and using dimensionality reduction to compare the results with the other results obtained.

3.1 Naive Bayes

The naive bayes classifier is based on the Bayesian theorem and is usually quite helpful when the number of input dimensions is large. A naive-bayes classifier assumes strong independence among the various attributes of data, but still manages to produce great results. This classifier generates a trained model with efficiency using very few data points given its simple design and oversimplified assumptions. The independence between attributes makes it easy to compute only variances instead of an exhaustive covariance matrix, and this estimation allows us to perform classification.

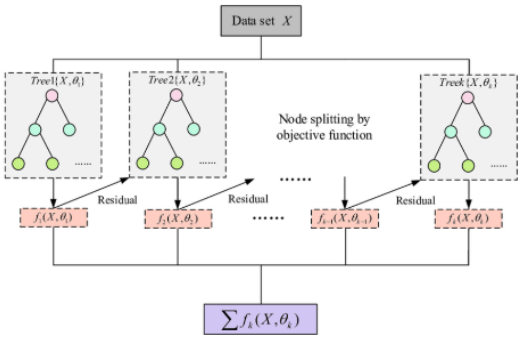
3.2 Support Vector Machine

Support Vector Machine (SVM) is another type of supervised learning method which is typically used for classification as well as regression problems. This method is known to be very efficient for a dataset with high dimensionality. The SVM algorithm attempts to locate a hyperplane in an n-dimensional space where n is the number of features which can distinctly classify any data point. The aim is to find such a plane while maximizing the margin so that a data point can be classified with a high level of confidence going forward.



3.3 XgBoost

XgBoost is an implementation of the gradient boosted trees algorithm. In this supervised learning algorithm, we try to predict a target variable by combining the predictions made by multiple simpler, weaker models. Here, the weak learners are regression trees, and each regression tree maps an input data point to one of its leafs that contains a continuous score. XgBoost minimizes a regularized objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity. The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.



3.4 Support Vector Machine (with Principal Component Analysis)

For SVM, we observe that the accuracy of the model increases when used after PCA of data. With 27 components and a linear kernel used, SVM is able to provide a higher accuracy than with the original data. This was due to there being a few features with low variance present in the original data, which we have taken care of with PCA.

3.5 XGBoost (with Principal Component Analysis)

Similar to the implementation of SVM with PCA, here too PCA tried to reduce the number of attributes by making sure most of the variance from the original dataset is restored. However, unlike SVM, in the case of XgBoost we see that the optimal model with PCA was obtained while using all 30 components and still the cost value is much lower than just applying XgBoost without PCA. Hence, as per this analysis, we see here that applying PCA with XgBoost could not improve the model performance.

4 Experiment setup

In this section, we discuss some of the concepts we have utilized to assist in performing classification with the models described in the previous section, and the reasons behind using these concepts for our specific use-case.

4.1 Dataset

The dataset that was chosen, contains 35 different columns and over 1600 rows. This dataset contains employee and company data useful for supervised ML, unsupervised ML, and analytics. Attrition - whether an employee left or not - is included and can be used as the target variable. The data is synthetic and based on the IBM Watson dataset for attrition. Employee roles and departments were changed to reflect the healthcare domain. Also, known outcomes for some employees were changed to help increase the performance of ML models.

4.2 Hypothesis

With the objective of ensuring accurate prediction of attrition cases, we have penalized, using our cost matrix, false positives and false negatives. The comparison of our different models will help us analyze model performance for our dataset using which we will go on to derive valuable insights and recommendations for the organizations and employees.

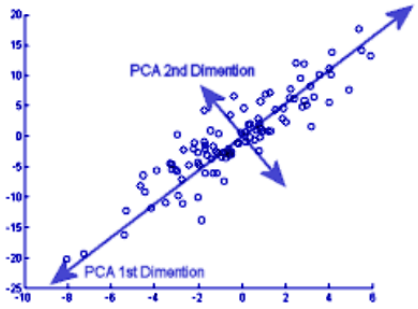
4.3 Preprocessing

The class label for our dataset is the "Attrition" attribute which contains binary classes (Yes or No) for each employee indicating whether the employee has left the company or not. Across the dataset, some of the columns contained discrete categorical data which we have preprocessed using the Label

Encoder to convert those categories into numerical data variables. Certain columns of our dataset had nominal data such as employee IDs, and some attributes simply had a single value across all records. These attributes do not impact the attrition in any way, and are therefore unimportant in determining attrition for an employee, so we have removed these attributes. Missing attributes have been filled as null, if any, and the remaining attributes have been normalized in order to offer more relevance to the values present in these records and to enable accurate classification.

4.4 Principal Component Analysis

Principal Component Analysis, or PCA, is an effective technique used to reduce the dimensionality and in turn avoid use of attributes which do not contribute much to the decision making of how a particular record should be classified. Principal component analysis is generally used to reduce dimensionality in data and provide the optimum number of components to be used in order to reduce variance in the predictive model. PCA can be an advantage (Increased accuracy) as well as a disadvantage (Lossy process) when it is used in conjunction with a model. We have implemented both SVM and XgBoost with PCA and have obtained a certain number of components for which our scoring works best. We ran our models with varying numbers of components from 1 to the total number of columns, and computed the costs after cross validation for each such run. The run which corresponded to the maximum calculated cost, was considered to be the ideal configuration of number of components for that particular model. Upon implementing PCA for our models, we were able to see a slight improvement in our performance with an added reduction in the number of dimensions for the model. The dimensionality reduction, in turn, also helps us avoid overfitting in certain cases which is useful to get improved accuracy and cost for our testing data.



4.5 Cross Validation

After splitting the dataset into training data and testing data (80-20), we have performed 10-fold cross validation using the training data to determine the model that performs best with our given dataset, and we have then used that optimal model to generate relevant metrics and results using the testing data. For our cross validation, we used 10 folds as a standard practice and each fold will contain about 135 data points. So, for each run of every model, we will have nine folds used for testing and one fold used for validation. This gives us a size big enough to train the models and determine the accuracy of these models.

Given our objective, as highlighted earlier, of accurately predicting employees who have left the company, we believe that the scoring for these models after cross validation cannot simply be dependent on the accuracy of our model. We designed a cost matrix which better depicts the importance of true positives, true negatives, false positives and false negatives. Our cost matrix is shown below:

		Predicted	
		Class = Yes	Class = No
Actual	Class = Yes	+10	-25
	Class = No	-100	+150

4.6 Next steps

Based on the performance of our classification models, we plan on using parameter tuning and improving our preprocessing techniques to get better and more refined results. This will further contribute to our objective of providing utility to this project by predicting at-risk employees and generating suggestions for reducing employee attrition. One of the methods, for our roadmap, is to do a correlation analysis among the attributes for our dataset to understand the relationship between them. Recommendations generated from our system would involve using regression for some attributes such as monthlyIncome, yearsSinceLastPromotion, etc. which could be useful characteristics to study, when trying to determine a 'safe zone' for employees that the organizations could target.

5 Results

The results of the training step for the 5 models are as follows. Each of these models are running 10 fold cross validation using the training data (80% of dataset) and the results denote the numbers obtained from that subset.

Model	Naive Bayes	SVM	XgBoost	SVM with PCA	XgBoost with PCA
Defined Cost	8820	17260	18245	17260	16495
Accuracy	86.72%	91.12%	91.57%	91.12%	90.82%

Based on the implementation, user defined cost is the best metric for model evaluation. Here we can see that XgBoost has the maximum defined cost of 4510 and hence this will be selected as the optimal model. Below are the results of testing our XgBoost model:

Optimal Model	XgBoost
Confusion Matrix	[[286, 12],[15, 23]]
Defined Cost	4735
Precision	0.96
Recall	0.95
Accuracy	91.96%
F1-Score	0.955

6 Conclusion

As per the results above we can see that XgBoost without PCA gives the maximum cost based on our defined cost matrix. Continuing on the obtained results, we intend to derive utility from this dataset by providing insights into at-risk employees and providing suggestions to ensure that the employees are fairly managed in order to reduce overall attrition.

REFERENCES

[1] Shankar, R. S., Rajanikanth, J., Sivaramaraju, V. V., Murthy, K. V. S. S. R. (2018, July). Prediction of employee attrition using data mining. In 2018 ieee international conference on system, computation, automation and networking (icscan) (pp. 1-8). IEEE.

[2] Srivastava, D. K., Nair, P. (2017, March). Employee attrition analysis using predictive techniques. In International Conference on Information and Communication Technology for Intelligent Systems (pp. 293-300). Springer, Cham.

[3] Yadav, S., Jain, A., Singh, D. (2018, December). Early prediction of employee attrition using data mining techniques. In 2018 IEEE 8th International Advance Computing Conference (IACC) (pp. 349-354). IEEE.

[4] Alao, D. A. B. A., Adeyemo, A. B. (2013). Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4(1), 17-28.