Predicting Employee Attrition

Rob Englund
rte4@pitt.edu

Shaunak Ghate
ssg27@pitt.edu

Mohamad Sahil
mos52@pitt.edu

1. Abstract

Employee turnover is one of major problems faced by organizations and often looked at as an opportunity to cut costs associated with it. This paper will showcase the application of data mining methods to predict employee attrition and to determine key factors that might contribute in attrition. Four different classification models are presented in this paper. Models performance is evaluated and compared to determine the best classification model.

2. Introduction

Retaining employees and experience has become a challenge for organizations in recent years. Organizations face huge costs due to employee turnover. Ideally organizations would like to eliminate attrition to retain talent and knowledge thus eliminating new employee training. To avoid being blind-sided and improve certain elements of the organization, predictive modeling is viewed as a great tool for human resource departments. This paper presents four different classification methods to predict employee attrition. Each model is evaluated and compared based on certain performance metrics. Models presented in the paper can be scaled and used by organizations today.

3. Related Work

Employee attrition prediction as part of the human resource analytics cloud has been a research topic for a while. Research in employee attrition prediction uses some of the classification models presented in this paper. Depending upon the data attributes and size different models and performance was presented. Threshold's for classification models depend on size and split of the target variable, hence models depend upon number of observations available.

Previous work for this particular dataset provides insight into the model performance[2]. This paper focusses on model comparison as well as the identification of key contributors in attrition.

4. Data Description

The analysis presented in this paper is based on a fictitious Kaggle dataset created by IBM data scientists and contains 1470 observations/employees and 35 variables[1]. The data consists of numerical and categorical variables. Some of the key variables in the dataset are described in Table 1.

Education	1 = Below College	2 = College	3 = Bachelor	4 = Masters	5 = Doctor
Environment Satisfaction	1 = Low	2 = Medium	3 = High	4 = Very High	
Job Involvement	1 = Low	2 = Medium	3 = High	4 = Very High	
Job Satisfaction	1 = Low	2 = Medium	3 = High	4 = Very High	
Performance Rating	1 = Low	2 = Good	3 = Excellent	4 = Outstanding	
Relationship Satisfaction	1 = Low	2 = Medium	3 = High	4 = Very High	
Work Life Balance	1 = Bad	2 = Good	3 = Better	4 = Best	
Business Travel	Rarely	Frequently			
Department	Sales	Engineering	Management	…	…
Attrition	Yes	No			
Marital Status	Single	Married	Divorced		
Overtime	Yes	No			

Table 1: Variable description

To gather additional information about the structure of the dataset, exploratory data analysis techniques were used. Figure 1 shows the attrition split in the dataset. 1233 out of the 1470 observations didn't leave the company.

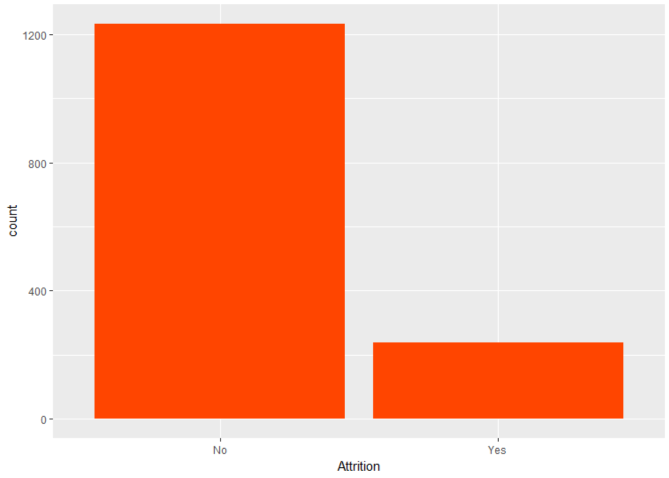


Figure 1: Attrition count in dataset

Correlation in a dataset can be a very important piece of information before building a model. Even though correlation doesn't imply causation, understanding inherent association between variables helps model formulation and evaluation. Figure 2 shows a correlation matrix of the continuous variables in the dataset.

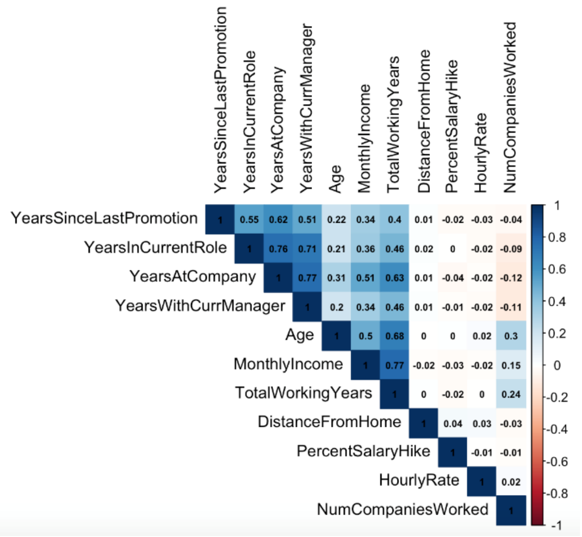


Figure 2: Correlation of continuous variables in dataset

5. Method

5.1 Logistic Regression

Employee attrition prediction being a classification problem, a logistic regression model was built first. Since Logistic regression is one of simpler models in terms of implementation and interpretation, it acts as a baseline model for the analysis. The logistic regression model performed well in terms of ROC and Figure 3 shows the ROC plot of Logistic Regression model. The AUC value is 0. 8421. The point on the curve gives the optimal threshold of 0.169. The first coordinate of the point is the specificity while the second coordinate is the sensitivity at optimal threshold and this is consistent with all the models presented in this paper.

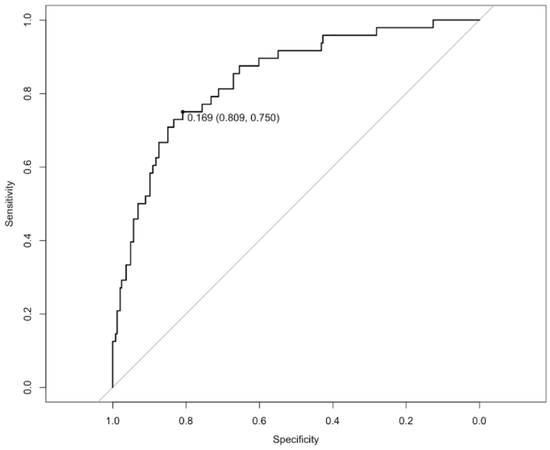


Figure 3: ROC plot for Logistic Regression model

Figure 4 shows the order of importance of attributes as per the Logistic Regression model. The most significant variable is Over Time followed by Environment Satisfaction, Job Satisfaction, Business Travel Frequency among others.

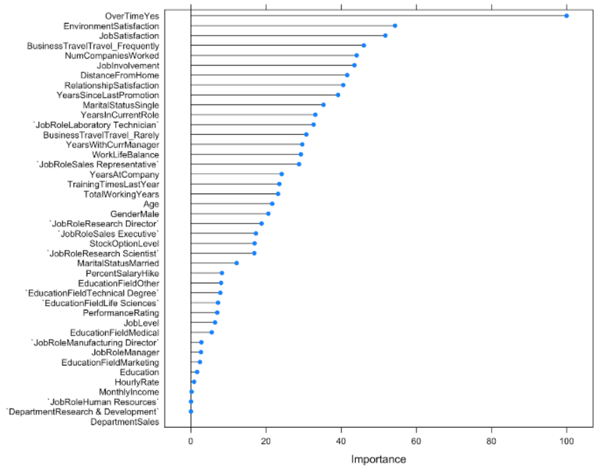


Figure 4: Attribute importance plot for Logistic Regression

5.2 Random Forest

To further investigate the dataset and improve the model performance, the Random Forest model was applied. The Random Forest model is far more advanced when compared to Logistic Regression and one would expect better results. But surprisingly, the Logistic regression model performed better than the Random Forest(RF) model. The RF models were analyzed by varying the tuning parameter "mtry" (number of randomly selected predictors)[3][8]. The optimal model was obtained when mtry was set to 5, AUC obtained was 0.8405 with an optimal threshold of 0.175[6]. The ROC plot of the Random Forest model is shown in figure 5.

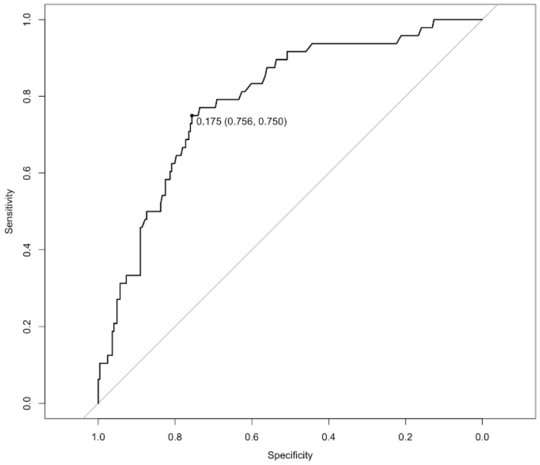


Figure 5: ROC plot for Random Forest model

The order of significance of attributes as per the Random Forest model are shown in Figure 6. The most significant variables are Monthly Income, Age, Total working years, Overtime, Distance from home, Years at company.

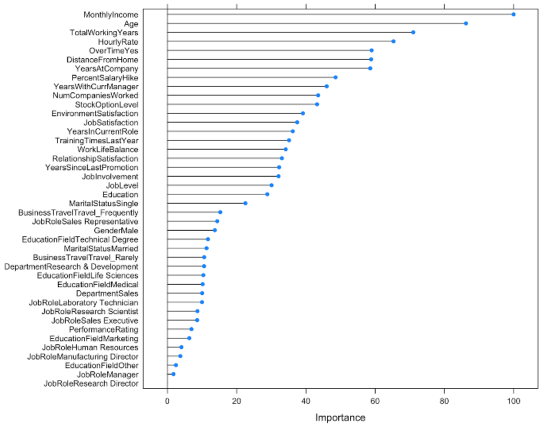


Figure 6: Attribute importance plot for Random Forest

5.3 Support Vector Machine

The SVM-linear algorithm was applied next. The model was tuned by using different values of cost parameter (C) and the optimal model was selected when C was 0.25. The SVM model performance was similar to the Logistic regression with small improvements. The AUC was measured to be 0.8405 and the optimal threshold was 0.169[9]. The ROC plot at optimal threshold is shown in figure 7.

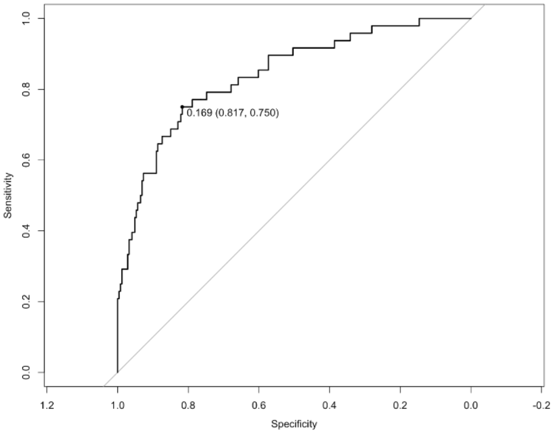


Figure 7: ROC plot for SVM model

Figure 8 shows the variable importance in the decreasing order as measured by the SVM model. Monthly Income, Years at Company and Total Working Years were the key attrition contributors according to the SVM model[3].

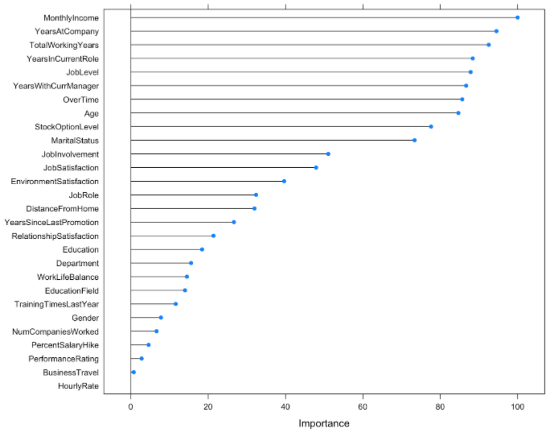


Figure 8: Attribute importance plot for SVM

5.4 Boosting

The last method applied was Gradient Boosting. The boosting model resulted in the best AUC value when compared to all the other models. The model was optimized by using the following values of tuning parameters.

n.trees (Number of trees) = 300
interaction.depth (Maximum nodes per tree) = 1
Shrinkage (Learning Rate) =0.1
n.minobsinnode (minimum number of observations in trees' terminal nodes) = 20.[8]

The AUC for this model was 0.8508 and the optimal threshold calculated to be 0.165[7]. The ROC plot of the boosting model is shown in figure 9.

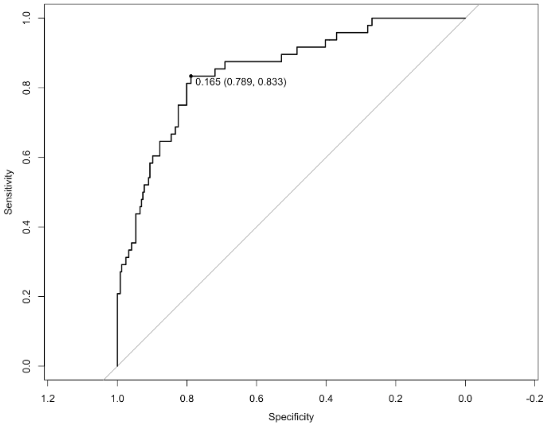


Figure 9: ROC plot for Boosting model

Figure 10 shows the variable importance plot as measured by the Boosting model. Over Time, Monthly Income, Stock Option Level and Total Working Years were the key attributes according to the Boosting model[3].

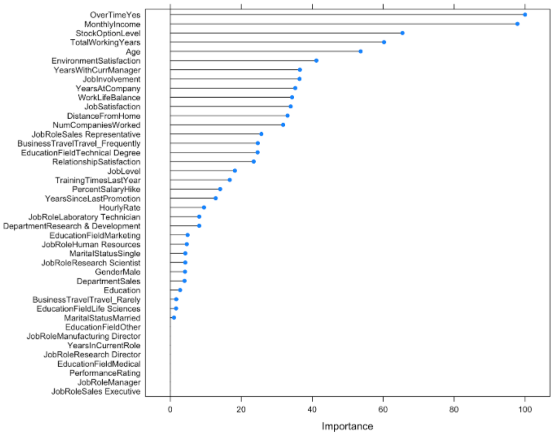
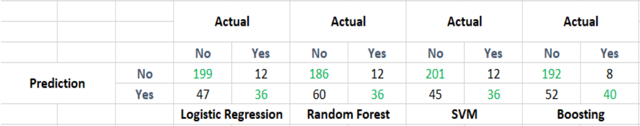


Figure 10: Attribute importance plot for Boosting

6. Evaluation Results

10-fold cross validation method was used to evaluate each model. Considering the size of the dataset, 80% of it was used to train models and remaining 20% was used to test the models. Table 2 shows confusion matrices for the four models.

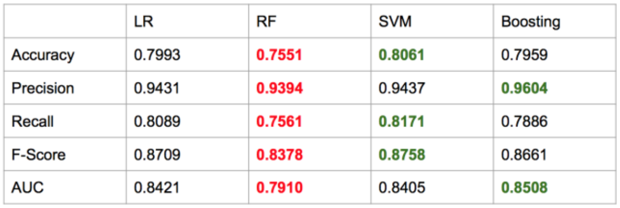
For this problem, sensitivity is the probability of correctly identifying the employees who are likely to leave the company and specificity is the probability of correctly identifying the employees who are going to stay. No model can achieve 100% sensitivity and 100% specificity; hence one can expect a trade-off between sensitivity and specificity. When it comes to employee attrition, sensitivity is as important as specificity. At 0.5 threshold value, all the models give high specificity but very low sensitivity. As a result, the accuracy is very high but the balanced accuracy (average of Specificity and Sensitivity) is low. So, the threshold value was adjusted to get the best-balanced accuracy. Upon decreasing the threshold value, the sensitivity increases at the expense of specificity and at a certain threshold, the model gives the optimized sensitivity and sensitivity. This threshold is called optimal threshold. In this paper, the performance of each model is compared at the optimal threshold[4].



		Actual		Actual		Actual		Actual	
		No	Yes	No	Yes	No	Yes	No	Yes
Prediction	No	199	12	186	12	201	12	192	8
	Yes	47	36	60	36	45	36	52	40
		Logistic Regression		Random Forest		SVM		Boosting	

Table 2: Confusion matrices for the models

Performance metrics such as accuracy, precision, recall, f-score and AUC were used to compare each model. Table 3 presents model performance for each model



	LR	RF	SVM	Boosting
Accuracy	0.7993	0.7551	0.8061	0.7959
Precision	0.9431	0.9394	0.9437	0.9604
Recall	0.8089	0.7561	0.8171	0.7886
F-Score	0.8709	0.8378	0.8758	0.8661
AUC	0.8421	0.7910	0.8405	0.8508

Table 3: Model Comparison

The choice of performance metric is dependent on the problem at hand and can vary. Accuracy is usually considered to be the best measure to evaluate the models, but ROC was the main evaluation metric used to determine the best model because it explains the balanced accuracy of the model i.e. considers sensitivity and specificity[4].

Table 2 shows how each model performed with respect to the different performance metrics stated earlier. The Random Forest model was worst in terms of performance across the five performance criteria. The best performance was split between the SVM and Boosting models depending upon the choice of performance metric. Considering ROC or AUC as the primary metric of evaluation, the Boosting model was the best prediction model.

7. Discussion

The employee attrition problem has been analyzed before, but this paper looks at the problem from a data mining perspective by comparing and contrasting four different models and focusing on contributing factors. Each model presents the key attrition contributing attributes which can provide insight on employee behavior. Since four different models are compared, a more robust solution is obtained for the models evaluated. This by no means is the best solution to the problem, there are many extensions and evolutions of the models presented in this paper that could be applied to this problem to achieve better results. One can explore other classification models to improve performance.

8. Conclusion

Employee turnover has huge implications n organizations and is a non-value add cost. Predictive modeling can help HR departments predict employee attrition. This paper analyzes four classification models to predict employee attrition based on 35 distinct employee attributes. With ROC as the key performance metric, the Boosting model was the best model followed by SVM. The Random Forest model was the worst performing model while Logistic Regression provided consistent yet average results. The attribute importance plots for each model point out the key factors involved in employee attrition.

The classification models discussed in this paper if scaled and tailored can be applied in real life. Other more advanced models like neural networks and adaptive boosting can be applied to this problem.

9. Acknowledgement

Each member of the group contributed in one way or the other to the project. Article selection was a collaborative effort and each team member agreed to model the employee attrition problem. R coding for the four models was again divided between the three team members. The presentation, progress reports and final paper was also a collaborative effort with each team member contributing to the analysis and evaluation. Overall, this project was a collaborative effort between Sahil, Shaunak and Rob.

10. References

[1]*Kaggle.com*.(n.d.).Retrievedfrom https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[2] Dancho, M. (2017). *HR Analytics using Machine Learning.*

[3] Kuhn, M. (2017). *The caret Package*. Retrieved from Github: http://topepo.github.io/caret/index.html

[4] Max Kuhn, K. J. (n.d.). *Applied Predictive Modeling*. Springer.

[5] Neff, D. L. (1993). *Predicting and managing turnover in human service agencies: a case study of an organization in crisis.*

[6] Wiener, A. L. (2002). *Classification and Regression by randomForest.*

[7] Bhalla, D. (n.d.). Retrieved from Listen Data: http://www.listendata.com/2015/07/gbm-boosted-models-tuning-parameters.html

[8] Brownlee, J. (2016). Retrieved from Machinelearningmastery: https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

[9]geekoverdose. (n.d.). *GeekOverdose*. Retrieved from GeekOverdose.wordpress: https://geekoverdose.wordpress.com/2014/07/25/svm-classification-example-with-performance-measures-using-r-caret/