

Prediction of Employee Turnover Using Ensemble Learning

Shubham Karande¹, L. Shyamala²

¹ School of Computing Science and Engineering,
VIT University, Chennai Campus
Chennai - 600 127, Tamil Nadu India.

shubham.sadashiv2017@vitstudent.ac.in

² School of Computing Science and Engineering,
VIT University, Chennai Campus
Chennai - 600 127, Tamil Nadu India.

shyamalal@vit.ac.in

Abstract. Employee turnover is now becoming a major problem in IT organizations, telecommunications and many other industries. Why employees leave the organization is the question rising amongst many HR managers. Employees are the most important assets of an organization. Hiring new employees will always take more efforts and cost rather than retaining the old ones. This paper focuses on finding the key features of voluntary employee turnover and how they can be overcome well before time. The problem is to predict whether an employee will leave or stay based on some metrics. The proposed work will use the application of ensemble learning to solve the problem, rather than focusing on a single classifier algorithm. Each classification model will be assigned with some weight based on the individual predicted accuracy. The ensemble model will calculate the weightage average for the probabilities of the individual classification and based on this weightage average, an employee can be classified. Accurate prediction will help organizations take necessary steps towards controlling retention.

Keywords: Employee Turnover, Classification, Ensemble Learning

1 Introduction

“You take away our top ten employees and we the Microsoft will become one of the mediocre companies”, this statement by Bill Gates is enough to prove the importance of the employee turnover. To retain employees is now become one of

the major tasks of the Human Resource Managers (HR) of the company. One of the main goals of HR's is to retain their employees and make use of their knowledge for the growth of the company. The way companies can deal with this issue is predicting the risk of employee churn. Machine Learning Algorithms are frequently used in employee churn study. Implementation of these ideas in Employee Relationship Management (ERM) has now become new trend. Employee Turnover can be divided into two categories: voluntary turnover, where employee chooses to leave the company or retirement and involuntary turnover, where employer decides to let go the employee. Retirement is something which won't be needing prediction as it is legally enforced. We are focusing on voluntary turnover therefore involuntary turnover is out of the scope of this paper. The novel contribution of this paper is to explore the application of ensemble learning as an advancement of traditional algorithms. The objective behind the work is to provide an improved system to tackle the employee churn problem and give HR's of the companies a heads-up, so that they can plan some strategies to overcome the turnover(Kotsiantis, S. B., 2007). The proposed work is comparing traditionally used classification algorithms with an ensemble learner which is combination of these same weak learners and by weighted average we can give weights to these algorithms which will be the novelty if this paper.

The rest of the paper is organized as; we have discussed the related work in section 2, Methodologies in section 3 and Results in the section 4 and Conclusion in section 5 which follows references at the end.

2 Literature Survey

The related work on the employee turnover is discussed here, at the first Rohit Punnoose et al. proposed a novel contribution of extreme gradient boosting in prediction of employee turnover, and comparison of XGBoost with six other historically used supervised classifiers(Rohit Punnoose and Pankaj Ajit ,2016). The results showed XGBoost gives higher accuracy, relatively low runtime and efficient memory utilization than the other six. In the proposed work, the strongest predictors for voluntary turnover are job satisfaction, overtime, salary, distance from home, marital status and employee's perception of fairness, an effective prediction model for predicting the employee turnover that have left the company has been introduced (J. L. Cottan et al.,2016). Decision tree is applied to predict the relevance of the attributes for turnover. This model can be used to decide employee will leave or not. A model is built using data from UCI repository to predict the status of employee turnover has been given. It uses three classification algorithms namely j48, bayesNet and naive Bayes. The model was implemented using Weka and the best performing algorithm was j48 based on the accuracy(M. Stovel et al. ,2017). An improved risk prediction clustering algorithm, which was multi-dimensional, was implemented to determine bad assets. In this work primary and

secondary levels of employee retention were used and association rule was integrated to avoid redundancy (B. Holtom et al., 2016). Two data mining models were developed for employee turnover to assist in decision. In this work, based on the accuracy obtained, regression model was found to outperform radial function model (S. L. Peterson et al., 2016). Three ensemble models were built and their performance in classifying the turnover as good risk group or bad risk group was analyzed. The ensemble models were built using Adaboost, Bagging, Random Forest combined with three learning algorithms (L. K. Marjorie et al., 2017). Ensemble machine learning algorithms were used to evaluate and decide the features which play a crucial role in predicting the risk involved in leaving the Company. Here Tree based classification was used and the algorithms were improved to favor the potential (D. Alao et al., 2016). An improved ensemble algorithm based on automatic clustering and under-sampling was proposed. In this method, clustering was done based on the weight of the samples and then a balanced distributed dataset was built which had a certain proportion of the majority class and all the minority class from each collection. By using Adaboost algorithm these datasets are used to build an ensemble classifier (G. King et al., 2017). A methodology for improving the performance of the classification through ensemble learning was proposed. Here classification was done using 3 different classifiers and the final classification was done by taking the majority voting from the classifiers (A. Liaw et al., 2017). Proposal of a customer churn problem in telecommunications industry, customer retention is given by (Namhyoung Kim et al., 2012). Here combination of SVM and PCA were used to not only get the higher accuracy but to boost the reliability of the model. All these references encourage us to build a model which will better the results, a model which is focusing on all the parts of architecture such as variable importance, algorithm selection and performance matrices. Ensemble learner models can give the solution with weighted average at its tail. The approach used in the paper is ensemble learning model. Here, applied diverse machine learning algorithms on dataset to predict the employee turnover. Weighted average is calculated from the probabilities of individual classification, using which the final classification is done.

3 Methodologies

The various classifiers and techniques used in this paper are described in this section.

3.1. Support Vector Machines

Support Vector Machine is a supervised machine learning algorithm. SVM algorithm works by plotting the data points in n -dimensional feature space where n denotes the number of features. After plotting, depending on the number of di-

mensions, a line or a plane or a hyper plane is drawn separating the data points such that the data points in one side belong to one class and the data points on the other side belongs the another class making it as non-probabilistic binary linear classifier. The separating line is drawn in such a way that they are divided by a clear margin that is as wide as possible. New instances are then plotted into that same space and are classified as belonging to as class based on which side of the gap they fall.

3.2. Random Forest Classifier

Decision trees are of two types: classification and regression trees. A decision tree can be viewed as a flow chart like arrangement in which the internal node denotes test on a feature, each branch denotes the result of the test and each leaf node denotes the decision taken after calculating all features. However the error rate is large for decision trees and they tend to over fit their training sets. A random forest is a meta-classifier that fits numerous decision trees classifiers on several sub-samples on the dataset and use averaging or majority voting to increase the predicted accuracy and reduce over-fitting.

3.3. Logistic Regression

Logistic Regression is another algorithm for Predictive Analysis borrowed from statistics. Despite the name Logistic regression, it is used from classification also. Unlike the other regression models, logistic regression does not try to give the value of numerical variable with given set of inputs instead it gives probability that the point belong to which class. Overfitting is very less for this model; therefore the model complexity becomes low. Eqn. (1) gives general logistic regression formula below,

$$g(E(y)) = \alpha + \beta x_1 + \gamma x_2 \quad (1)$$

The role of link function is to ‘join’ the expectation of y to linear predictor.

3.4. Variable Importance

Classification trees analysis and Regression trees analysis can be collectively called as Classification and Regression Trees (CART) analysis. CART analysis produces a predictor ranking also known as variable importance on the basis of contribution predictors make to the building of the tree. Importance is decided by playing a role in the tree, either as a main splitter or as a surrogate. In this paper, random forest is used to calculate the variable importance. Instead of using all 34 features of classification, here selecting top 10 variables with better variable importance and used them for classification. This minimizes the time required to train the model.

3.5 Under Sampling

When training the model with an imbalanced dataset, the model tends to be biased towards the majority class. In this paper, the class attribute has 83.87% of one class and 16.23% of other class. Hence by using under sampling we are making the proportion of the classes as 60% and 40% respectively. By using this method the biasing can be reduced while building the model and increase the sensitivity and specificity of the ensemble model.

3.6 Weighted Average Prediction

By making use of the accuracy from individual predictions of the classifiers, we are assigning weights to each classifier. In this paper, SVM was assigned 0.40, Logistic Regression was assigned 0.30 and Random Forest was assigned 0.30 as weights. By taking the product of probability that an instance will be assigned to a particular class for each classifier and the weights assigned to the respective classifier, we can predict the final classification. If the resultant of the product is greater than 0.5 it will be assigned to the class whose probability we took while taking the product otherwise it will be assigned to the other class. By using this approach we were able to increase the accuracy of the prediction for ensemble model. Example is given in Fig.1.

	Model1	Model2	Model3	WeightAveragePrediction
Weight	0.4	0.3	0.3	
Prediction	45	40	60	48

Figure 1: **Weighted Average Prediction**

3.7 Architecture Diagram

The proposed Ensemble Learning e Architectural is presented in Fig.2. in this model, the dataset is taken and the features with variable importance are identified and collected. All these features are separately executed with different algorithms those are discussed earlier to get the individual score. Then to model the ensemble learning model weighs are assigned for each model and given to the ensemble model. The classifier classifies and the accuracy is calculated.

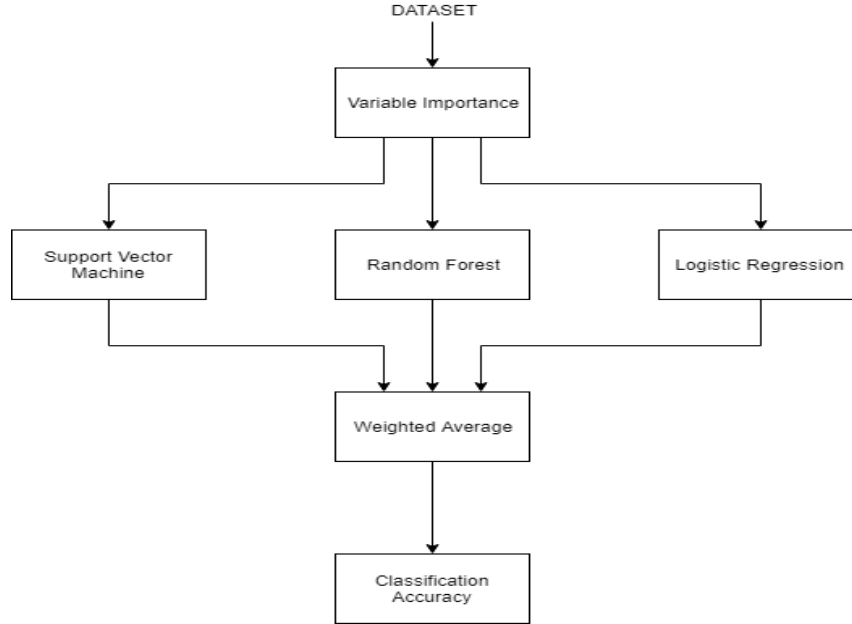


Figure 1: Architecture Diagram of Ensemble Model

4 Results

To validate the proposed method, the implementation is carried out on system with the configuration of Processor Intel Corei7,HDD 500GB, RAM 8 GB, with Windows OS. The Tool used to implement the code is JetBrains Pycharm 2017.3.2.

4.1 Description of the dataset

The dataset (<https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employeeattrition>) contains details of employee turnover from IBM Data, which is used for case study of HR Analytics. The class attribute in the dataset, attrition represented as 0(employee did not leave) or 1(employee left). Exploratory data analysis is done on the dataset and it is reveals that the dataset is imbalanced in terms of target variable. 16.23% (employee will leave) of the total instances have the class variable 'yes' and 83.87% of the total instances have class variable as 'no' (employee will not leave).The results of the individual classifiers and the ensemble model are discussed in this section. Below given Table 1 represents the confusion matrix.

Table 1: Confusion Matrix Metrics and their Definition

Actual Class	Predicted Class		
		No	Yes
	No	True Positive	False positive
	Yes	False negative	True Negative

4.2 Performance Metrics

The performance metrics used to validate the methodologies are given in Table 2 and these entries used to define and evaluate the performance of the classifiers discussed in this paper.

Table 2: Performance Metrics and their Definition

Metric	Equation	Definition
Accuracy	$(TP+TN)/(P+N)$	Ratio of the total number of predictions that are correct
Precision	$TP/(TP+FP)$	Ratio of the predicted positive cases that are correct
Sensitivity	$TP/(TP+FN)$	Ratio of the positive cases that are correctly identified
Specificity	$TN/(FP/TN)$	Ratio of negative cases that are correctly identified

By making use of the confusion matrix, performance metrics such as accuracy, precision, sensitivity and specificity for the classifiers and the ensemble model are calculated and their results are shown in the below Table 3,4.

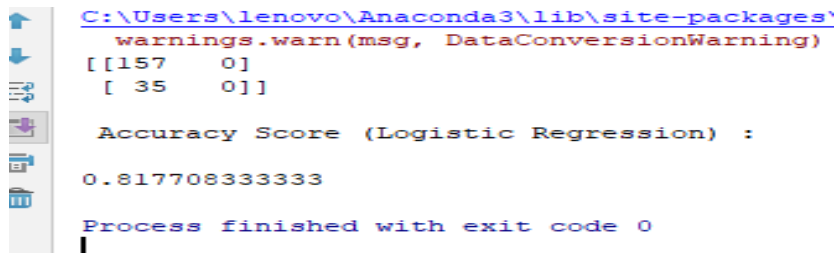
Table 3: Confusion Matrix for different methodologies used

Methodology	Reference	Prediction	
		No	Yes
Support Vector Machine	No	5954	1055
	Yes	956	1034
Logistic Regression	No	6546	463
	Yes	1321	669
Random Forest	No	6233	776
	Yes	839	1151
Ensemble Model	No	6530	479
	Yes	961	1029

Table 4: Performance Metrics

Methods	Accuracy	Precision	Sensitivity	Specificity
Support Vector Machine	77.65%	84.95%	86.16%	49.50%
Logistic Regression	81.77%	93.39%	83.21%	59.10%
Random Forest	82.64%	88.93%	88.14%	59.73%
Ensemble Model	83.87%	93.17%	87.17%	68.24%

From the Table 3,4 the result shows that the Ensemble model has more True positive and True negative compared to the other model. That means the prediction on the employees of their job interest to continue or not has been well predicted by this method than the other methods. Ensemble model out performs in accuracy, precision and specificity over other models. But its sensitivity of the data is slightly less than Random forest but good than other two techniques. From this, we conclude that the prediction accuracy of ensemble model is better than other models. Fig.2 shows some screen shot of implementation.



```

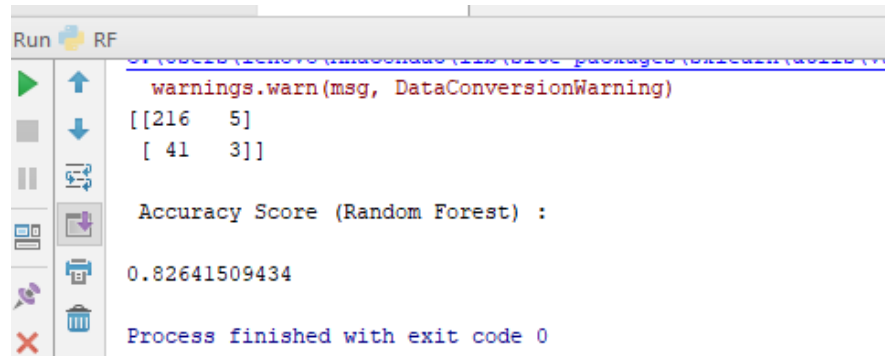
C:\Users\lenovo\Anaconda3\lib\site-packages\
warnings.warn(msg, DataConversionWarning)
[[157   0]
 [ 35   0]]

Accuracy Score (Logistic Regression) :

0.817708333333

Process finished with exit code 0

```

```

Run RF
warnings.warn(msg, DataConversionWarning)
[[216  5]
 [ 41  3]]

Accuracy Score (Random Forest) :
0.82641509434

Process finished with exit code 0

```

Figure 3. Screen shots of implementation

5 Conclusion

The need of predicting employee turnover in companies and use of machine learning algorithms in building these models was represented in this paper. The main challenge of building an Ensemble Learner Model which is a combination of Support Vector Machine, Logistic Regression and Random Forest was highlighted. This Model will be able to predict the employees turnover more precisely, based on the accuracy obtained from the individual classifications weights were assigned, and calculated the weighted average. Based on the weighted average the final classification is done which gives improved performance which is more superior to the results given by individual classifiers. In future, it can be fine-tuned more refrained feature to improve the sensitivity of the data.

References

1. Ajit, P. Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, 4(5), C5(2016).
2. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*. 4, (2013).
3. Cotton, J. L., Tuttle, J. M.: Employee turnover: A meta-analysis and review with implications for research. *Academy of management Review*, 11(1), 55-70 (1986).
4. Holtom, B. C., Mitchell, T. R., Lee, T. W., Eberly, M. B.: 5 turnover and retention research: a glance at the past, a closer review of the pre-

- sent, and a venture into the future. *The Academy of Management Annals*, 2(1), 231-274 (2008).
5. King, G., Zeng, L.: Logistic regression in rare events data. *Political analysis*, 9(2), 137-163 (2001).
 6. Kotsiantis, S. B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications* Kotsiantis, S. B., Zaharakis, I., Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24. ns in computer engineering, 160, 3-24 (2007).
 7. Kane-Sellers, M. L.: Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis. Texas A&M University.(2007).
 8. Kim, N., Lee, J., Jung, K. H., Kim, Y. S.: A new ensemble model for efficient churn prediction in mobile telecommunication. In 2012 45th Hawaii International Conference on System Sciences ,pp. 1023-1029. (2012)IEEE.
 9. Liaw.A, and Weiner.W.: Classification and Regression by Random Forest, *R News*, 2(3), 18-22 (2017).
 10. Peterson, S. L.: Toward a theoretical model of employee turnover: A human resource development perspective. *Human Resource Development Review*, 3(3), 209-227 (2004).
 11. Stovel, M., & Bontis, N.: Voluntary turnover: knowledge management–friend or foe?. *Journal of intellectual Capital*, 3(3), 303-322 (2002).