Prediction of Employee Retention using Cassandra and Ensemble Learning

Shubham Karande[1], Ajay Shelake[2], Sivagami M[3], Sharon Sophia[4]
VIT University
India
shubham.sadashiv2017@vitstudent.ac.in
ajaymaruti.shelake@vitstudent.ac.in
sivagami.m@vit.ac.in
sharon.sophia@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 raising among many HR managers. Employees are the most important assets of an organization. Hiring new 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 classify whether an employee will leave or stay. Data is taken from Kaggle. The proposed work will uses ensemble learning to solve the problem, rather than focusing on a single classifier algorithm it will combine weak learning algorithms to get a better ensemble model. We have used Cassandra to store the data in the form of table and retrieving data to perform machine algorithm on them.*

Keywords: Employee Retention, Cassandra, Classification, Ensemble Learning, Machine Learning

Received: 2 June 2019, Revised 20 August 2019, Accepted 19 September 2019

DOI: 10.6025/jio/2019/9/4/134-140

1. Introduction

"You take away our top twenty 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 employees. The assets of the company, and to retain them is now become one of the major task of the Human Resource (HR) Managers of the company. One of the main goals of HR's of the company is to retain their employees and make use of their knowledge for the growth of the company. The approach in which the organizations can compact with this problem is predicting the threat of attrition of employees. Machine Learning Algorithms are frequently used in employee churn study. Implementation of these ideas in ERM (Employee Relationship Management) has now become new trend, as employee turnover shows superior business results. 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 study discusses the problem of employee retention and the algorithms used to resolve this delinquent are conferred. The unique influence of this study is to discover the application of ensemble learning as an advancement of traditional algorithms.

The study is organized as; Review of literature in section 2, Methodology in section 3, 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. (2016), proposed a novel contribution of extreme gradient boosting in prediction of employee turnover, and comparison of XGBoost with six other historically used supervised classifiers is shown. The results showed XGBoost gives significantly higher accuracy, efficient memory and relatively low runtime utilization than the other six. J. L. Cottan et al. (2016), proposed the strongest predictors for voluntary turnover are job satisfaction, overtime, salary, distance from home, marital status and employee's perception of fairness, an dynamic prediction model for forecasting the employee turnover that have left the company has been introduced. Decision tree is realistic to predict the significance of the aspects for turnover. This model is used to decide, whether employee will leave or not. M. Stovel et al. (2017), proposed a model constructed using data from UCI repository to predict the standing of employee turnover has been proposed. It uses three grouping algorithms namely j48, bayesNet and naive Bayes. The model was implemented using Weka and the best performing algorithm was j48 based on the accuracy. B. Holtom et al. (2016), an upgraded risk prediction clustering algorithm, which was multi-dimensional, was implemented to regulate bad assets. In this effort primary and secondary levels of employee retention were used and association regulation was integrated to avoid redundancy. S. L. Peterson et al. (2016), here 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. L. K. Marjorie et al. (2017), 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. D. Alao et al. (2016), 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. G. King et al. (2017), 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. A. Liaw 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.

Our approach used in the paper is ensemble learning. Here applied diverse machine learning algorithms on dataset to predict the employee turnover. Based on the performance of the individual classification, voting is taken, 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. Description of the Dataset

The data used in this paper is taken from Kaggle; the dataset contains details of employee retention, which is used for case study of HR Analytics. The class attribute in the dataset, left represented as 0(employee did not left) or 1(employee left). Exploratory data analysis is done on the dataset and feature importance is done using random forest.

3.2. Variable Importance

Classification trees analysis and Regression tress 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 10 features of classification, here selecting top 5 variables with better variable importance and use them for classification. This minimizes the time required to train the model.

3.2.1 Before using Variable Importance

Variable selection is vital aspect of model designing. Here we have taken a dataset with 10 attributes (variables) these variables are namely: last evaluation, satisfaction level, number of projects, average monthly hours, time spend in company, work accident, promotion last five years, sales, salary and left (Target Variable)

3.2.2 After using Variable Importance

After performing variable importance using CART method, the top five variables are taken into picture for final model training. These variables are as follows: satisfaction level, time spend in company, promotion last five years, number of projects and left (Target Variable)

3.3. Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm. SVM algorithm works by plotting the data points in n-dimensional feature space where n is the number of features. After plotting depending on the number of dimensions, 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 gap that is as wide as possible. New instances are then plotted into that sane space and are classified as belonging to as class based on which side of the gap they fall. In addition to performing linear classification, SVM can proficiently perform a non-linear classification using what is called the kernel trick. Kernels are based on the principle that a non-linearly separable dataset containing 'n' features can made linearly separable when plotted in higher dimensional space. Here, in this paper svmLinear kernel to the classification.

3.4. Multilayer Perceptron

A Multilayer Perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

3.5. 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 predict 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, so the model complexity becomes low.

$$g(E(y)) = \alpha + \beta x1 + \gamma\, x2$$

Here, g () is the link function, $E(y)$ is the expectation of target variable and $\alpha + \beta x1 + \gamma\, x2$ is the linear predictor (α, β, γ to be predicted).

The role of link function is to 'link' the expectation of y to linear predictor.

3.6 Voting Classifier

It's defined as taking the prediction with maximum vote / recommendation from multiple models predictions while predicting the outcomes of a classification problem.

Model 1 SVM	Model 2 MLP	Model 3 Logistic Regression	Voting Prediction
1	1	0	1

3.7 Apache Cassandra

Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data. Cassandra's support for replicating across multiple datacenters is best-in-class, providing lower latency for your users and the peace of mind of knowing that you can survive regional outages. Cassandra is a distributed

database from Apache that is highly scalable and designed to manage very large amounts of structured data. It provides high availability with no single point of failure.

3.9 Architecture Diagram

Here is the Architectural Diagram for Proposed Model,

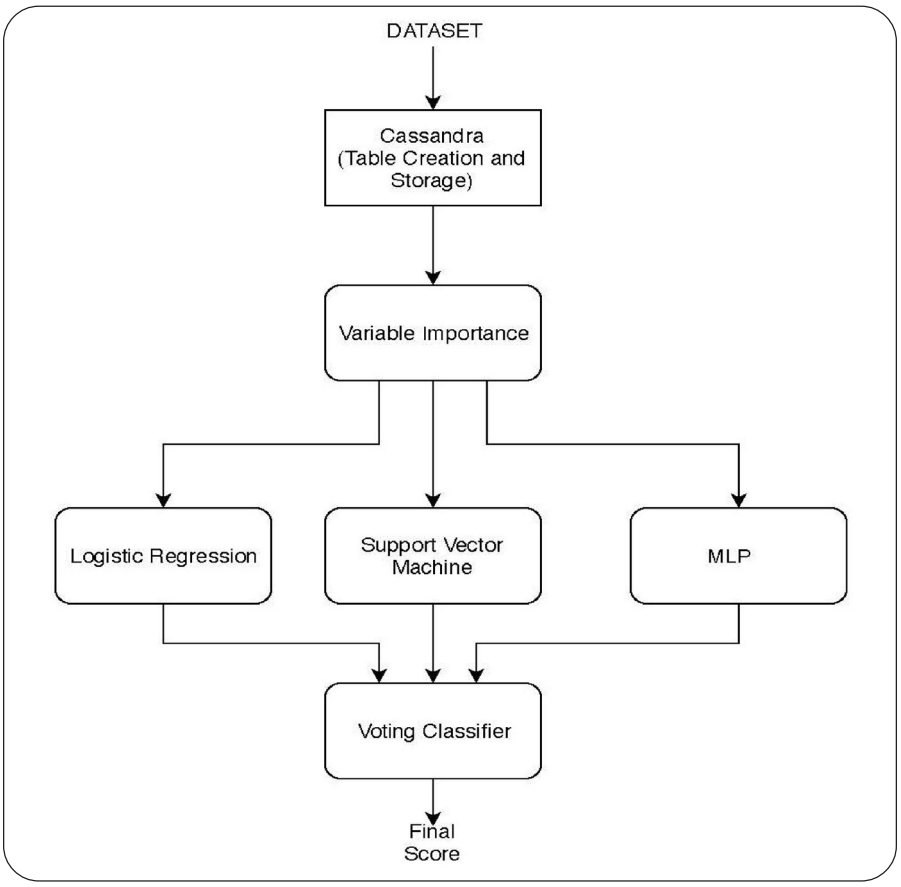


Figure 1. Architecture Diagram of Proposed Model

4. Results

The results are discussed here,

4.1 Dimensionality Reduction

Algorithm	Accuracy before dimensionality reduction	Accuracy after dimensionality reduction
Support Vector Machine	69.0	77.65
Logistic Regression	76.0	80.15
Multilayer Perceptron	77.10	82.05
Ensemble Learning	80.0	84

Table 1. The accuracy measures before and after applying dimensionality reduction

4.3 Confusion Matrices

The confusion matrix shown in Table 1 and these entries used to define and evaluate the performance of the classifiers discussed in this paper.

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

Table 2. Performance Metrics and their Definition

4.3.1 Support Vector Machines

The accuracy obtained by classifying the dataset using SVM is 77.65%. The Confusion matrix is shown below in Table 3.

	Prediction	
Reference	**No**	**Yes**
No	5954	1055
Yes	956	1034

Table 3. Confusion Matrix for Support Vector Machine

4.3.2. Logistic Regression

The accuracy obtained by classifying the dataset using Logistic Regression is 80.15%. The Confusion matrix is shown below in Table 4.

	Prediction	
Reference	**No**	**Yes**
No	6546	463
Yes	1321	669

Table 4. Confusion Matrix for Logistic Regression

4.3.3 Multilayer Perceptron

The accuracy obtained by classifying the dataset using Multilayer Perceptron is 82.05%. The Confusion matrix is shown below in Table 5.

4.3.4 Ensemble Model

The accuracy obtained by classifying the dataset using Ensemble Model is 84%. The Confusion matrix is shown below in Table 6.

	Prediction	
Reference	**No**	**Yes**
No	6233	776
Yes	839	1151

Table 5. Confusion Matrix for Multilayer Perceptron

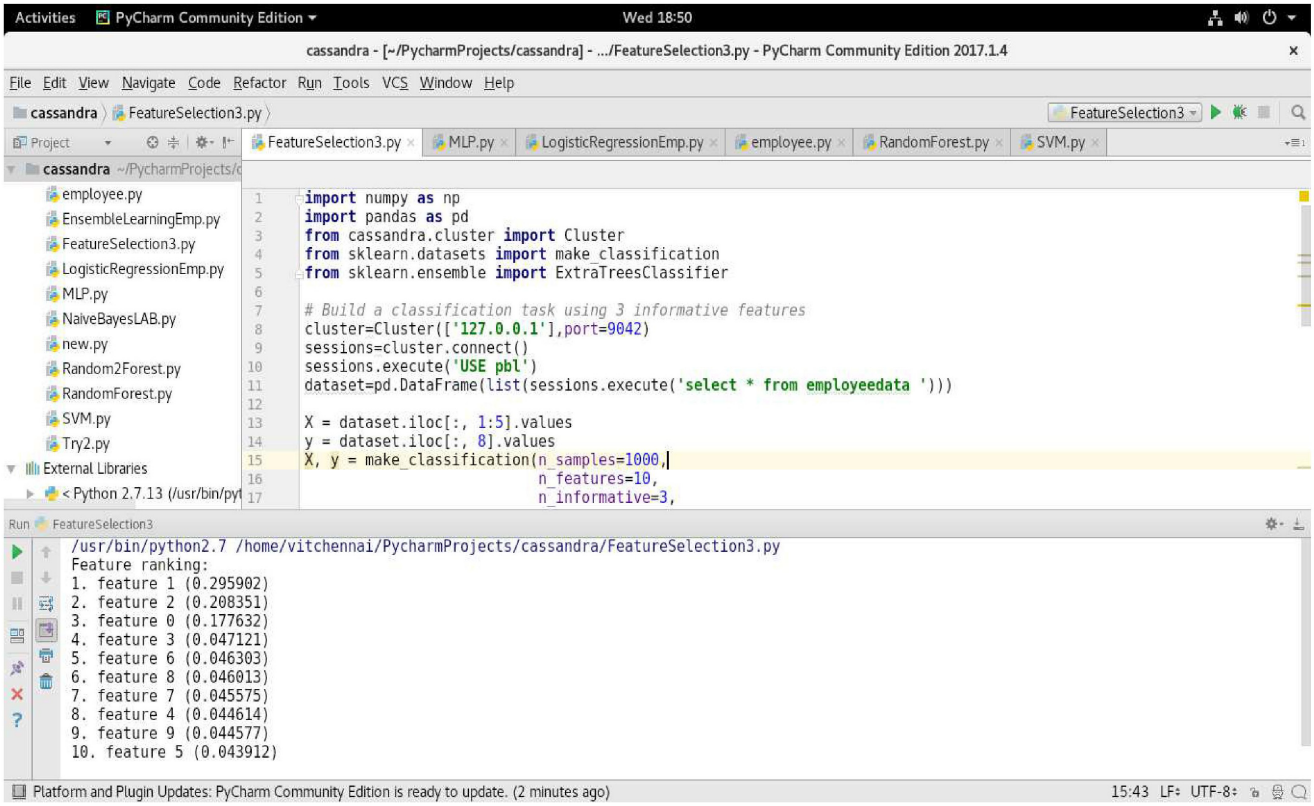
	Prediction	
Reference	**No**	**Yes**
No	6530	479
Yes	961	1029

Table 6. Confusion Matrix for Ensemble Model

4.4 Screenshots

5. Conclusion

The need of predicting employee turnover in organizations and use of machine learning in building these models was presented in this paper. The Key challenge of building an Ensemble Learning Model which is a combination of Support Vector Machine, Linear Regression and Random Forest was highlighted. This Model will be able to predict the turnover of employees more precisely, based on the accuracy obtained from the individual classifications weights are 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.



References

[1] Liaw, A., Weiner, W. (2017). Classification and Regression by Random Forest, *R News*, 2 (3) 18-22.

[2] Holtom, B., Mitchell, T., Lee, T., Eberly, M. (2016). Turnover and Retention research: A glance at past, closer review of the present and a venture into the future, *Academy of Management Annals*.

[3] Alao, D., Adeyemo, A. B. (2016). Analyzing employee attrition using decision tree algorithms, Computing, Information Systems, *Development Informatics and Allied Research Journal*.

[4] King, G., Zeng, L. (2017). Logistic Regression in rare events of data, *Political Analysis*.

[5] Cottan, J. L., Tuttle, J. M. (2016). Employee Turnover: A Meta-analysis and review implications for research, *Academy of Management Review*.

[6] Marjorie, L. K. (2017). Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis, Texas, A & M University College of Education.

[7] Stoval, M., Bontis, N. (2017). Voluntary Turnover: Knowledge management-Friend or Foe? *Journal of Intellectual Capital*, 3(3) 303-322.

[8] Punnooose, Rohit., Ajit, Pankaj. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms, *International Journal of Advanced Research in Artificial Intelligence* (IJARAI).

[9] Peterson, S. L. (2016). Toward a theoretical model of employee turnover: *A Human Resource Development Review*, Elsevier.