Machine Learning-Based Prediction of Employee Attrition

Dr. Prakasam S
Associate Professor, Department of Computer Science and Applications,
SCSVMV University,
Enathur, Kanchipuram, India.

Abstract

Every organization has a unique productivity and strength that is derived from its workforce. In today's cutthroat economy, maintaining regular employees is a major difficulty for any firm. One of the largest business issues in HR analytics is employee attrition. Businesses heavily invest in staff training because they see the potential returns on that investment. An employee's departure represents a loss of potential for the business. These studies analyze the attrition rate of employees by looking at associated factors such as job role, overtime, and job level, which have a significant impact on attrition. This study evaluates a number of classification methods, including logistic regression, LDA, SVM, KNN, and Random Forests, to estimate the likelihood that a new hire will attrite.Consequently, using Random Forest for training a balanced dataset produced the second-best results, with a 0.269 F1-score, but our algorithms produced the maximum accuracy.

KEYWORDS: Employee attrition; Support vector machine; random forest; K nearest neighbours; Feature selection;Attrition Rate; HR; Classifier; Preprocessing; Employment Features.

Introduction

Numerous studies have demonstrated that employees are an organization's most significant resource and asset. The attrition rate is now determined by enhanced employee competency requirements and heightened competitiveness. For corporations, staff churn is regarded as a major problem. Hiring and training new staff comes at a very significant expense. Employers must look for, select, and onboard new personnel. Losing seasoned employees, particularly top achievers, is challenging to handle and has a detrimental impact on an organization's ability to succeed and function. The elements that may help limit the employee attrition rate are the main focus of the study. The issue of staff turnover has become a significant concern in organizations due to its negative effects on workplace efficiency and self-esteem. In order to address this issue, companies use machine learning techniques to forecast the likelihood that employees will leave, enabling them to take proactive measures to increase retention.

Literature Review

This research analyzes the employee attrition rate at different levels by collecting data using customized methodologies that use different data mining techniques. The following are the results of a thorough literature analysis of the works of numerous researchers as well as a study on data mining for the purpose of obtaining the employee attrition rate utilized in various models;

Data mining techniques were used by Qasem A, A. Radaideh, and Eman A Nagi to create a classification model that predicts employee performance [3]. In their work, they used the CRISP-DM data mining methodology [4]. The primary data mining method utilized to construct the classification model, from which several classification rules were produced, was the decision tree. They used actual data gathered from multiple companies to run multiple trials and validate the created model. The model is meant to be used to forecast the performance of incoming candidates.

Using actual manufacturing plant data, Amir Mohammad EsmaieeliSikaroudi, [5] RouzbehGhousi, and Ali EsmaieeliSikaroudi et al. applied knowledge discovery techniques. They debate a variety of personnel attributes, including age, skill level, and experience. The Pearson Chi-Square test was used to determine the significance of the data attributes.

Human resource professionals can improve the performance appraisal process of their human capital by refocusing on human capability criteria thanks to a prediction model for employee performance forecasting that was proposed by John M. Kirimi, Christopher Moturi, et al. [6].

The usage of Extreme Gradient Boosting (XGBoost) technique, which is more resilient due to its regularization formulation, was investigated [7] by Pankaj Ajit et al. and Rohit Punnoose. [8] XGBoost's much improved accuracy for staff turnover prediction is demonstrated by comparing it to six supervised classifiers that have been in use in the past using data from the HRIS of a major store.

Authors	Issue Investigated	Techniques Studied	Suggest
Jantan, Hamdan and Othman[9]	Data Mining techniques for performance prediction of employees	C4.5 decision tree, R.andom Forest, Multilayer Perceptron(MLP) and Radial Basic Function Network	C4.5 decision tree
Nagadevara, Srinivasan and Valk[10]	Relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover	Artificial neural networks, logistic regression, classification and regression trees (CART), classification trees (C5.0), and discriminant analysis)	Classification and regression trees (CART)
Hong, Wei and Chen[11]	Feasibility of applying the *Logit* and *Probit* models to employee voluntary Turnover predictions	Logistic regression model (logit), probability regression model (probit)	Logistic regression model (logic)
Marjorie Laura Kane Sellers[12]	To explore various personal, as well as work variables impacting employee voluntary turnover	Binomial logit regression	Binomial logit regression
Alao and Adeyemo[13]	Analyzing employee attrition using multiple decision tree algorithms	C4.5, C5, REPTree, CART	C5 decision tree
Saradhi and Palshikar[14]	To compare data mining techniques for predicting employee churn	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees and Random Forests	Support Vector Machines

In order to extract significant features like Monthly Income, Last Promotion Year, Salary Hike, and other factors that are relatively normal for employee attrition, the data is first obtained from Kaggle and pre-processed. The variables that aid in determining the components that are mostly dependent on employee-related variables are known as dependent or predicted variables. For instance, there is no correlation between the attrition rate and the employee ID or employee count. The first step in the analysis process is called exploratory data analysis, where you can summarize data features to forecast which employees would leave the company and when. The random forest technique is used by the system to create a prediction model. Rather than using a single decision tree for categorization, this ensemble learning technique uses many decision trees.

To assess employee churn, the methods use word formation vector and dependent variable analysis. We can therefore greatly lessen this issue by raising employee confidence and creating a pleasant work atmosphere.

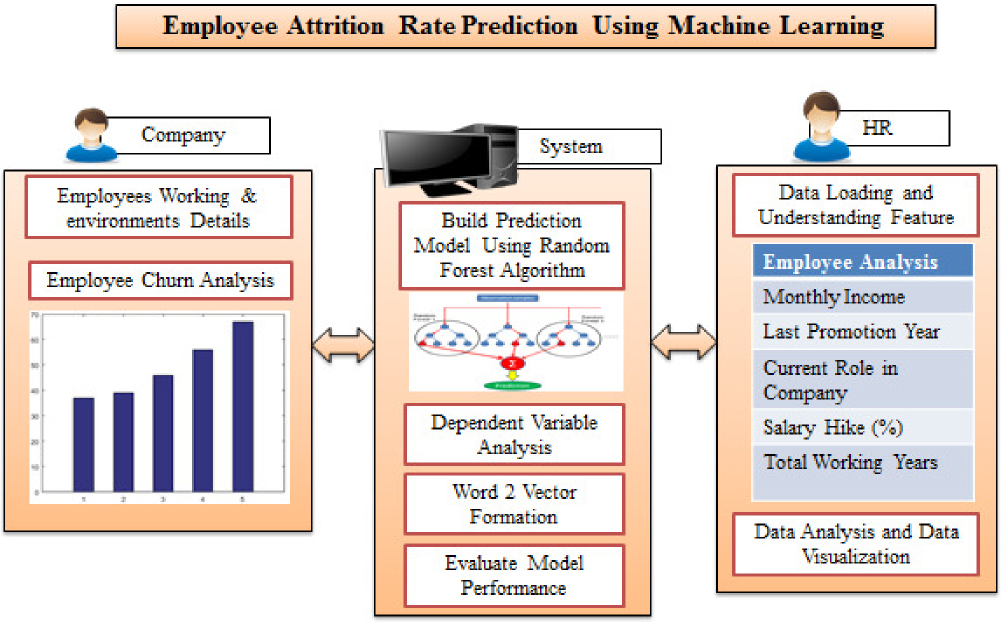

Employee Attrition Rate Prediction Using Machine Learning
Company
Employees Working & environments Details
Employee Churn Analysis
System
Build Prediction Model Using Random Forest Algorithm
Dependent Variable Analysis
Word 2 Vector Formation
Evaluate Model Performance
HR
Data Loading and Understanding Feature
Employee Analysis
Monthly Income
Last Promotion Year
Current Role in Company
Salary Hike (%)
Total Working Years
Data Analysis and Data Visualization

Fig 1. Flow of Architecture

Dataset

To aid in the solution of this issue, we have collected this dataset from IBM HR Analytics. Attrition is the dependent attribute out of the total 35 attributes in this dataset. We've concluded that this dataset may be able to assist us in solving this issue. These characteristics are included in our dataset.

- **Age- Numeric Discrete**
- **Attrition-Categorical**
- **Business Travel- Categorical**
- **Daily Rate- Numeric Discrete**
- **Department- Categorical**
- **DistanceFromHome- Numeric**
- **Education- Categorical**
- **Education Field- Categorical**
- **Employee Count- Numeric Discrete**
- **Employee Number- Numeric Discrete**
- **Environment Satisfaction- Categorical**
- **Gender- Categorical**
- **Hourly Rate- Numeric Discrete**
- **Job Involvement- Categorical**
- **Job Level- Categorical**
- **Job Role- Categorical**
- **Job Satisfaction- Categorical**
- **Marital Status- Categorical**
- **Monthly Income- Numeric Discrete**
- **Monthly Rate- Numeric Discrete**
- **NumCompaniesWorked- Numeric Discrete**
- **Over18- Categorical**
- **OverTime- Categorical**
- **PercentSalaryHike- Numeric Discrete**
- **PerformanceRating- Categorical**
- **RelationshipSatisfaction- Categorical**
- **StandardHours- Numeric Discrete**
- **StockOptionLevel- Categorical**
- **TotalWorkingYears- Numeric Discrete**
- **TrainingTimesLastYear- Numeric Discrete**
- **WorkLifeBalance- Categorica**
- **YearsAtCompany- Numeric Discrete**

- **YearsInCurrentRole- Numeric Discrete**
- **YearsSinceLastPromotion- Numeric Discrete**
- **YearsWithCurrManager- Numeric Discrete**

Feature Selection

When it comes to machine learning, feature selection is seen to be the most important theory. It plays a big part in how well your model actually performs. These features have a huge impact on performance and are easily used to coach your model.

Unimportant and irrelevant features may negatively affect the model's performance. Your model's initial and most important stage should be feature selection and data cleaning. The process of automatically or manually choosing the characteristics that most significantly contribute to your dependent variable or output variable of interest using several strategies such as the Univariate Selection Feature Importance Correlation Matrix is known as feature selection.

We determined that the features Employee Count, Employee Number, and Over 18 have no direct bearing on our outcome variable Arttrition after personally analyzing the information. Thus, before using any feature selection techniques, these features have been totally disregarded. Each feature in your data is given a score based on feature importance; the greater the score, the more significant or pertinent the feature is to your output variable. With Tree Based Classifiers, feature importance is an inherent class. To extract the dataset's top features, we will use Extra Tree Classifier.

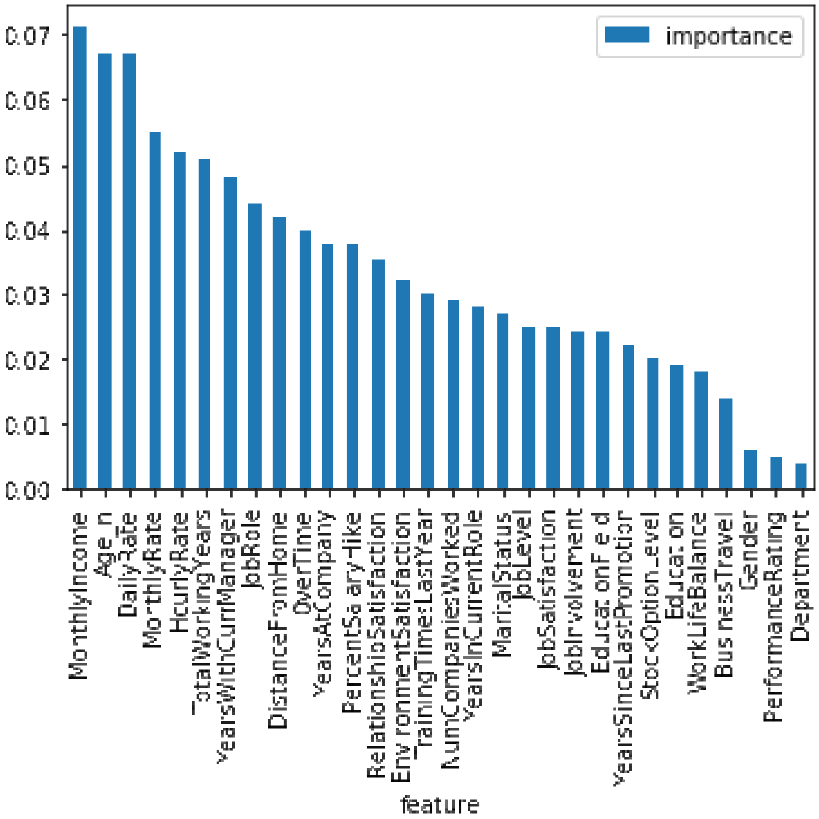


Fig 2. Importance Feature

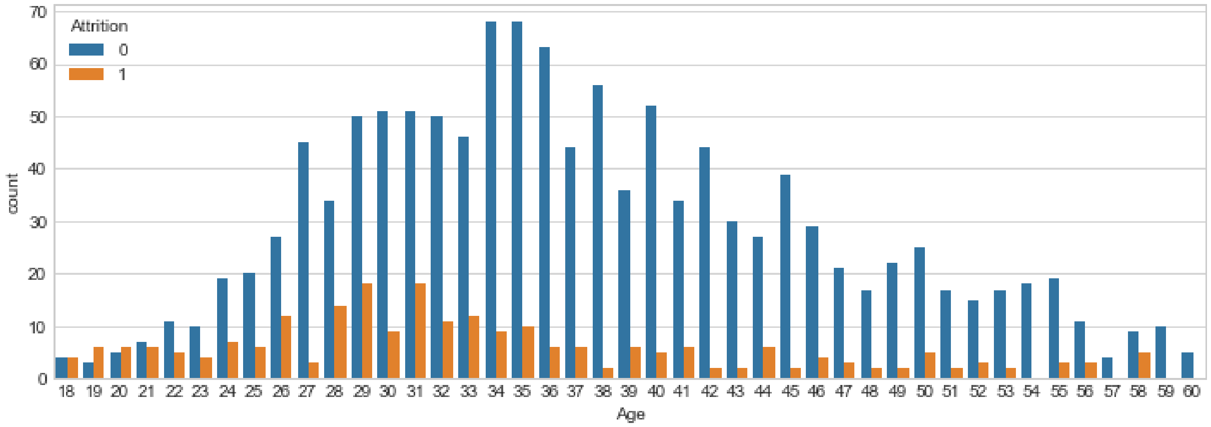
With the aid of this feature importance approach, we were able to analyze and determine which aspects, such as monthly income, age, daily rate, hourly rate, etc., are some of the significant qualities. The diagram above shows the feature importance of each feature in our dataset. We also concluded that the features with the least influence on our output variable, Atrrition, are Business travel, gender, department, and performance rating. We can therefore ignore these aspects in advance.The following attributes are chosen for model design when feature selection techniques are applied.

- Monthlyncome
- Age
- Monthlyrate
- Hourlyrate
- Totalworkingyears
- Yearswithcurrmanager

- Jobrole
- Distancefromhome
- Overtime
- Yearsatcompany
- Percentsalaryhike
- Realationshipsatifaction
- Environmentsatisfaction
- Trainintimelastyear
- Nocompaniesworker
- Yearcurrentrole
- Yearsincurrentrole
- Martialstatus
- Joblevel
- Jobsatisfaction
- Jobinvolvement
- Educationfield
- Yearssincelastpromotion

Investigative Data Analysis

We will examine the relationship between the qualities and the output variable in this section. We are unable to display every attribute relationship because there are so many attributes at our disposal. In order to keep things simple, we will demonstrate the relationship using the Age attribute as an example.



The aforementioned figure makes it evident that the age group of 29 to 31 has the greatest attrition rate of any age group. The likelihood of attrition is lowest for individuals above the age of fifty.

Unbalanced Dataset

90% of the records in the dataset have the class YES labeled, whereas the remaining 10% have the class NO labeled. These kinds of datasets are referred to as unbalanced datasets, and they might negatively impact the model's performance by tilting it in favor of the output variable's majority class. As a result, managing an unbalanced dataset becomes essential for this kind of issue statement.

- ❖ **Random Under Sampling**
- ❖ **Random Over Sampling**
- ❖ **Custer Based Over Sampling**

The oversampling method is being used for our dataset in order to address its unpredictability. Just 237 records had the class YES label prior to the oversampling, out of the 1233 records that had the class NO label. We compared the amount of records for both groups to 1233 records after carrying out oversampling, as the diagram below illustrates**.**

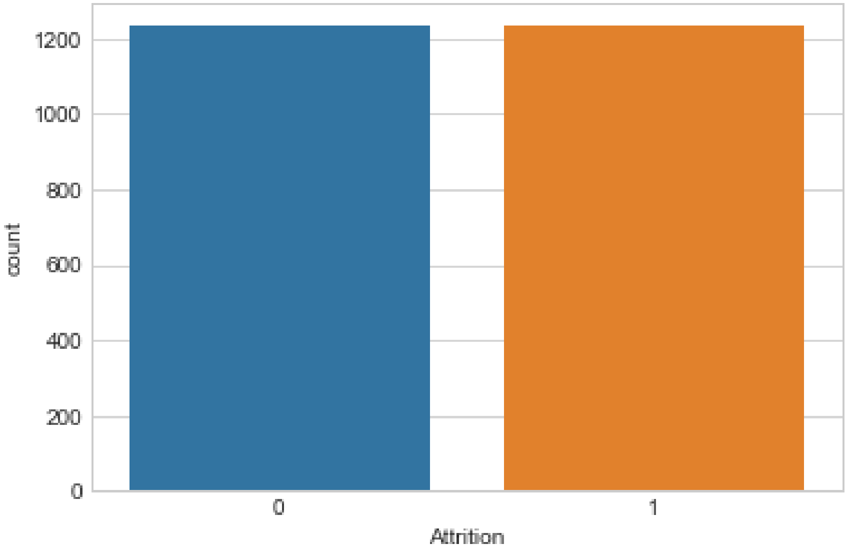


Fig 4. Distribution of Data following Oversampling

Comparative Evaluation

The issue statement for employee attrition prediction falls under the machine learning classification category. Such a categorization problem can have several possible solutions. Choosing the optimal solution for the problem statement is the subject of comparative analysis. In this section, we display the model results according to their accuracy, precision, recall, and F1 score.

In this case, we are contrasting several classification methods using assessment metrics including F1 score, Accuracy, Precision, Recall, and K-Nearest Neighbor (KNN), Logistic Regression (LR), and Random Forest (RFA).

All trained models were evaluated by measuring their accuracy, precision, recall and F1 score which are described below:

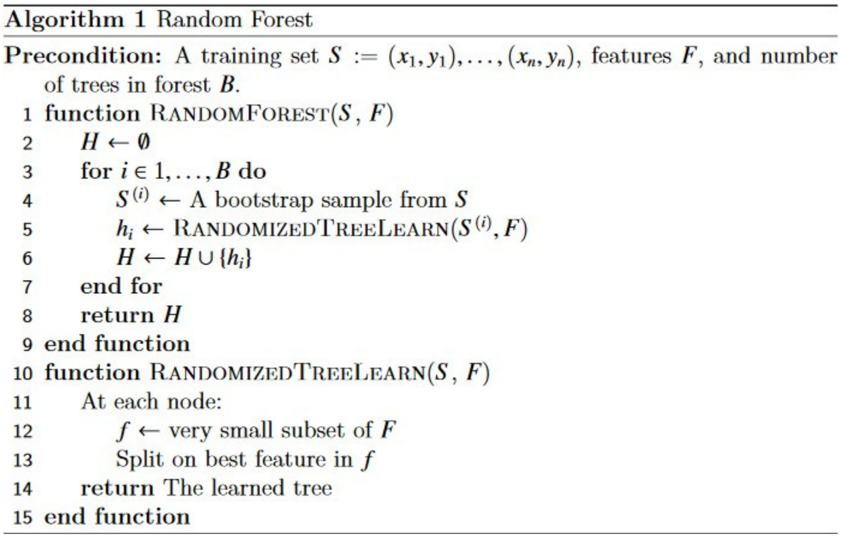
- $\text{Accuracy} = \frac{\text{TP+TN}}{TP+TN+FP+FN}$
- $\text{Precision} = \frac{\text{TP}}{TP+FP}$
- $\text{Recall} = \frac{\text{TP}}{TP+FN}$
- $\text{F1 Score} = 2 * \frac{\text{Precision}* \text{Recall}}{\text{Precision}+ \text{Recall}}$

Random Forest (RFA)

An algorithm for supervised classification is the Random Forest algorithm. One tool for decision help is the decision tree.

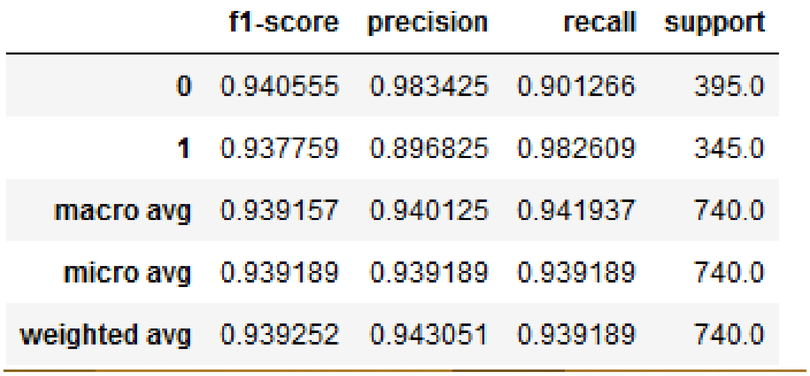
One of the most effective supervised machine learning methods for producing regressions and classifications is random forest (RF). RF trains data using multiple decision trees [9]. The RF model determines which class received the most votes from the decision trees when each tree casts a vote for a classification label for a particular dataset [10].

It illustrates the potential outcomes with a graph that resembles a tree. The decision tree will create a set of rules if you feed it a training dataset that includes targets and features. Predictions can be made using these Rules.There are two stages in Random Forest algorithm, one is random forest creation, and the other is to make a prediction from the random forest classifier created in the first stage.



```
Algorithm 1 Random Forest
Precondition: A training set $S := (x_1, y_1), \dots, (x_n, y_n)$, features $F$, and number
    of trees in forest $B$.
function RandomForest($S$, $F$)
    $H \leftarrow \emptyset$
    for $i \in 1, \dots, B$ do
        $S^{(i)} \leftarrow$ A bootstrap sample from $S$
        $h_i \leftarrow$ RandomizedTreeLearn($S^{(i)}, F$)
        $H \leftarrow H \cup \{h_i\}$
    end for
    return $H$
end function
function RandomizedTreeLearn($S$, $F$)
    At each node:
        $f \leftarrow$ very small subset of $F$
        Split on best feature in $f$
    return The learned tree
end function
```

Performance of RFA Model



	f1-score	precision	recall	support
0	0.940555	0.983425	0.901266	395.0
1	0.937759	0.896825	0.982609	345.0
macro avg	0.939157	0.940125	0.941937	740.0
micro avg	0.939189	0.939189	0.939189	740.0
weighted avg	0.939252	0.943051	0.939189	740.0

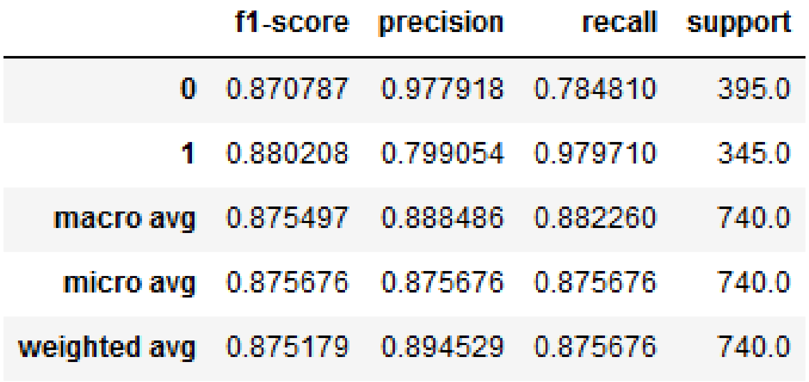
Decision Tree

The decision tree algorithm is a member of the supervised learning algorithm family. The decision tree approach, in contrast to other supervised learning algorithms, is also capable of handling regression and classification issues.

By learning straightforward decision rules from historical data (training data), a Decision Tree can be used to develop a training model that can be used to predict the class or value of the target variable.

With decision trees, we begin at the root of the tree when attempting to forecast a record's class label. We contrast the root attribute's values with those of the record's attribute. We follow the branch that corresponds to that value and go on to the next node based on the comparison.

Performance of Decision Tree Model



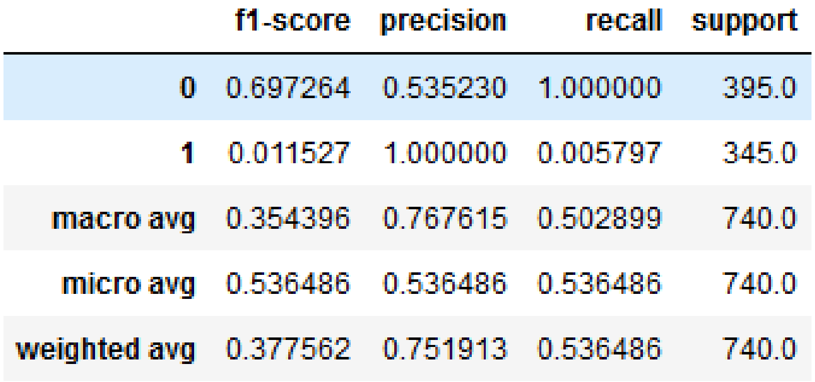
	f1-score	precision	recall	support
0	0.870787	0.977918	0.784810	395.0
1	0.880208	0.799054	0.979710	345.0
macro avg	0.875497	0.888486	0.882260	740.0
micro avg	0.875676	0.875676	0.875676	740.0
weighted avg	0.875179	0.894529	0.875676	740.0

Stochastic Gradient Descent (SGD)

A straightforward yet effective optimization technique called stochastic gradient descent (SGD) is used to determine the parameters and function coefficients that optimize a cost function. Stated differently, it is applied to the discriminative learning of linear classifiers under convex loss functions, namely logistic regression and support vector machines. Because the update to the coefficients is carried out for every training instance rather than at the end of examples, it has been effectively used to large-scale datasets.

The Stochastic Gradient Descent (SGD) classifier essentially carries out a simple SGD learning procedure that accommodates different classification penalties and loss functions. To implement SGD classification, Scikit-learn offers the SGDClassifier module.

Performance of SGD Model



	f1-score	precision	recall	support
0	0.697264	0.535230	1.000000	395.0
1	0.011527	1.000000	0.005797	345.0
macro avg	0.354396	0.767615	0.502899	740.0
micro avg	0.536486	0.536486	0.536486	740.0
weighted avg	0.377562	0.751913	0.536486	740.0

Conclusion

Any organization's primary pillar is its human resource base. The strength of the workforce clearly influences both the rate of growth and the degree of market penetration. Nowadays, any corporation can achieve enormous success because of the growing population and the presence of highly competent individuals. However, attrition is the main problem that any business often addresses. In addition to being a major duty, retention is also a huge challenge. This essay examines the many approaches and strategies employed by different studies for employee prediction strategies.

References

[1] S. Jahan, "Human Resources Information System (HRIS): A Theoretical Perspective", Journal of Human Resource andSustainability Studies, Vol.2 No.2, Article ID:46129, 2014.

[2] C. Cortes and V. Vapnik, Support-vector networks. Machine learning, 20(3), 273-297, 1995

[3] RenukaAgrawal, Jyoti Singh, and Zadgoankar .S, "Formative Assessment For Performance Evaluation Of Faculty Using Data Mining", *International Journal Of Advances In Electronics And Computer Science*, ISSN: 2393-2835.

[4] HosseinAlizadeh, Buinzahra Branch and Islamic, 2016 ,"Introducing A Hybrid Data Mining Model ToEvaluate Customer Loyalty", *Engineering, Technology & Applied Science Research Volume. 6,No.6,1235-1240.*

[5] Amir Mohammad EsmaieeliSikaroudi ,Rouzbehghousi and Ali Esmaieelisikaroudi, 2015 "A Data Mining Approach To Employee Turnover Prediction" (Case Study: Arak Automotive Parts Manufacturing), *Journal*

Of Industrial And Systems Engineering Volume. 8, No. 4.

[6] Anjali A. Dudhe and SachinSakhare .R, January 2018, “Teacher Ranking System To Rank Of Teacher As Per Specific Domain” ‚*Journal On Soft Computing ICTACT*, Volume: 08, Issue: 02, Issn: 2229-6956.

[7] Rohit Punnoose and Pankaj Ajit, 2016 “Prediction Of Employee Turnover In Organizations Using Machine Learning Algorithms”, *International Journal Of Advanced Research In Artificial Intelligence*(IJARAI) Volume. 5, No. 9.

[8] Dilip Singh Sisodia, SomduttaVishwakarma, AbinashPujahari“ Evaluation of Machine Learning Models for Employee Churn Prediction”, Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE
[13] Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.

[9] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Towards Applying Data Mining Techniques for Talent Managements”, 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011

[10] V. Nagadevara, V. Srinivasan, and R. Valk, “Establishing a link between employee turnover and withdrawal behaviors: Application of data mining techniques”, Research and Practice in Human Resource Management, 16(2), 81-97, 2008.

[11] W. C. Hong, S. Y. Wei, and Y. F. Chen, “A comparative test of two employee turnover prediction models”, InternationalJournal of Management, 24(4), 808, 2007.

[12] L. K. Marjorie, “Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis”, Texas, A&M University College of Education, 2007.

[13] D. Alao and A. B. Adeyemo, “Analyzing employee attrition using decision tree algorithms”, Computing, Information Systems, Development Informatics and Allied Research Journal, 4, 2013.