A Study on Employee Attrition Analytics

Abstract

Employee attrition is one of the most important challenges long faced among various business organizations. With the advancement in machine learning, numerous research papers brought into light the various reasons why associate employees leave. In any knowledge-based companies, employee turnover is a significant problem. If an employee leaves, they take with them confidential information, which can be a source of competitive advantage for other companies. With the ultimate goal in mind, a business must minimize employee attrition to stay competitive and retain its workforce. The employee churn/attrition forecast model is discussed in this article using several Machine Learning approaches. The optimal strategies for employee withholding at various stages of an employee's connection with a business are then outlined and tested using model yields. This research has the potential to improve employee retention by laying out improved retention designs and increasing employee satisfaction. This study considers key machine learning systems that have been used to construct predictive churn models and combines and condenses the ability to leverage information and provide information-based experiences, choices, and projections. And also, massive investment in employee skills coaching has been adopted by ample organizations in reaction to the speedy evolution of the world trends and technology adoption. Unfortunately, targeting employee retention after unsatisfactory coaching provides a negative return on investment. Prediction of target candidate call before coaching and understanding the options that influence the candidate decision can greatly contribute to candidate choice and decision feature improvement method for accrued increased retention. This study focused on employee retention methods. Employees are the assets of the organization. To retain skillful and committed workers within the organization, management ought to pay attention of employee satisfaction. To determine the explanations of employee turnover and overcome this. The purpose of this study is to demonstrate, however, that employee retention is essential in our time, and if the organizations don't seem to be aware of the case and immediate actions don't seem to be taken to it result, what repercussions lay ahead and the way they might have an effect on the organization and also the trade.

Keywords Human Resource; Employee attrition, Machine Learning, Tableau, Linear Regression, Random Forest.

Introduction

Employee attrition or voluntary turnover presents a key issue for organisations because it affects not only their productivity and work sustainability but also their future growth strategies. On this path, employee retention may be a major challenge for recruiters and employers alike, since employee attrition means not only the loss of skills, experiences and personnel but also the loss of business opportunities. within the era of massive Data, people analytics help organisations and their human resources (HR) managers to scale back attrition by changing the way of attracting and retaining talent. During our research, HR analytics is taken into account as a must have capability for the HR management and profession and ‘‘a tool for creating value from people and a pathway to broadening the strategic influence of the HR function’’ .So, it represents the quantification and therefore the systematic identification of the people drivers of the business outcomes with the aim of constructing better decisions. There are interchangeable terms used for HR analytics that are talent analytics, people analytics, and workforce analytics. because of people analytics, HR managers gain the power to know their departments and their employees, by providing more accessible and interpretable data about employee attributes, performance and behaviours. Thus, HR analytics plays a big role in every aspect of the HR function in organisations including recruiting, training and development, retention, engagement and compensation. within the context of HR Analytics, employee attrition analysis has caught more and more attention within the business world. In fact, a way to use analytic methods to predict whether employees will leave or not can help the organisation improve the HR management and save the value thereon .Therefore for the HR managers, it's crucial to possess a much better idea of what kind of employees will tend to go away and what kind of features will influence them to depart.Most typically, organisations desire to create sure the proper employees are within the right place at the correct time and identifying employees’ intention to depart by means of analytics .Descriptive analytics are accustomed summarise or turn data into relevant information so investigate what has occurred.In other words, descriptive analytics have some meaningful impact by explaining what has already happened however, they're not much helpful in predicting what's going to happen or may happen within the future.

On the contrary, predictive analytics are proposed and accustomed forecast what is going to happen within the future. within the field of HR, predictive analytics result in achievement of organisational benefits and help surely in better decision-making within the organisation with no biases, especially with the foremost prosperous trend of the large data era and data science basing on machine and deep learning techniques. In fact, data is taken into account which is mandatory ingredients to the people's analytics team requires to be effective. Otherwise, HR is ready to fail in handling Big Data challenges since Big Data focuses on capturing each piece of accessible information and collecting every suitable and unsuitable data.But, in HR analytics context, the difficulty must move from the scale of the information to its smartness and making better use of information to make and capture value, being a necessary prerequisite to the more advanced sorts of big data analysis . Additionally, highlighted the boundaries of the appliance of huge Data within a contextual HR case study, whilst also noting the necessity to shift the main target from a quantitative to an analysis of HR data. during this context, the concept of deep data was born to cater to collecting only relevant and specific information and excluding information that may be unusable or otherwise redundant. Thus, during this paper, we mainly target a functional dimension. From a functional dimension, we aim to check, compare and choose the simplest accurate predictive model which will early detect employee attrition. We also aim to interpret the positive attrition to seek out reasons behind it and then to support HR managers to create retention plan. We also aim to create it easy for HR to form fast and advance decisions so we introduce the visualisation of its data and its predictions using machine learning pipeline.Machine learning seems to be an ever progressing field that shows no signs of slowing down, with the monumental leaps within the way we recognise, process and find patterns in data, lost dreams seem close. Machine learning has taken care of predicting employee attrition. Using these predictive models made and machine learning as our foundation, we tried to answer the question of what specific changes would an employee have to prevent them from leaving the corporate and thus reducing the rate of attrition. In fact, recent related works commonly that specialise in finding the most effective predictive models with high performances to predict employee attrition using generally benchmarks and simulated open data like HR IBM1 datasets. But, during this paper, we argue that except for model's performances, the HR data must be constructed and filtered to offer relevant and rapid prediction without biases.

Literature Survey

Abhay Patro et.al [1] proposed a machine learning pipeline that not only predicts employee attrition but also suggests a minimum cost approach for the company so that the employee does not leave.and also introduce a Machine Learning Pipeline, which business organizations can utilise by choosing their own slack features and willingness factors, complemented with a large database, to drastically reduce employee attrition.**Guru Renuka et.al** [2] proposed Light Gradient Boosting Machine calculations to achieve better results to predict employee attrition based on their features. **Sai Chandan P Reddy et.al** [3] proposed that Employees are the main asset to the company and the company can't run without them and they play an important role in shaping the company and sending the company to next heights and to help HR managers need to know what needs to be improved if an employee is planning to quit or find a new job and this way they would not lose a credible employee to another competing company. **Aseel Qutub et.al** [4] predicts the accuracy of five base models and then combines them to create a more powerful predictor model, which is known as ensemble learning. The research uses a combination of decision trees and linear regression to achieve an accuracy of 86.39 percent, outperforming Adaboost and Random Forest, as well as SVM. **Setiawan et.al** [5] through their work found variables that have a major impact on employee attrition. **M Marchington et.al** [6] discussed various researchers have evidenced the worth of human resource management (HRM) in many settings for example, working situations, production, management and relationships identification with productivity The impact of HRM on productivity is evidenced to have positive outcomes on business intensity and capital growth. **Xiang Gao et.al** [7] developed social network metrics for oscillation and average response time to identify changes in the communication behaviors of managers who were about to quit their jobs. They proposed an improved RF algorithm, the WQRF based on the weighted F-measured. It has also been applied in the prediction of employee turnover in industries such as education, medical, finance, and other fields. Different employees of companies will have similar characteristics but also particular attributes, which do not affect the use of this algorithm. By using the WQRF approach, HR administrators can predict better employee turnover and take timely action. **Evy Rombaut et.al** [8] stated that researchers only focused on work-specific factors that influence employee turnover. Generally, most of the research focuses on employee attrition and retention for already hired employees. No work has been done about prediction of employee retention before getting hired or trained and that is the research gap we are tackling. **Sarah Alduayj et.al** [9] proposed employee attrition using machine learning models such as support vector machine (SVM) with several kernel functions, random forest and K-nearest neighbour (KNN). The main objective of this research was to use machine learning models to predict employee attrition based on their characteristics. They also mentioned Losing high-performing employees is considered a major loss for companies, specifically those that invest in their employees. Finding replacements with a similar level of performance is considered difficult and can cost the company both money and time.**Srivastava et.al** [10] presented a framework that predicts employee churn by analyzing the behaviors of employees and attributes with the help of machine learning techniques.**Manisha Purohit et.al** [11] Employee turnover refers to the number of employees who are leaving the organization over a particular time span which is generally expressed in the percentage of the total number of employees in an organization.

Méthodology

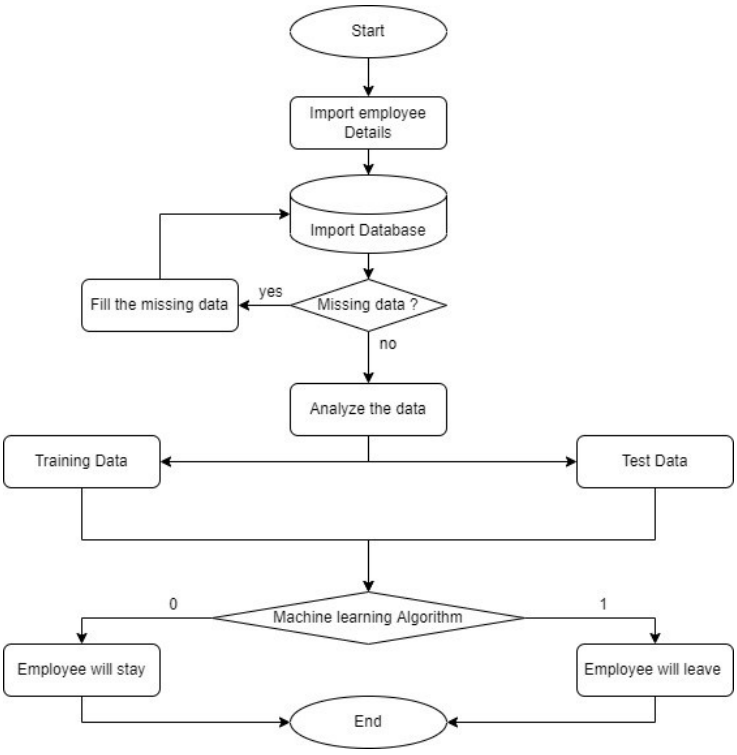

Fig 1. System Architecture

Machine learning is appealing in support of investment firm strategies through speculative HR applications. Our proposed ML application is usually limited to targeted decision-making and interpretation of factors that affect post-training staff decisions. We propose the ML classification of factors that may influence employee final decisions based on predictable analysis of employee data. The proposed automated and intelligent decision-making process is carried out using algorithms that separate the learning machine in order to use the possible predictor of targeted decision-making by a clear definition of models that reflect the factors affecting the candidate decision.

Machine Learning Algorithms are of three types:

Unsupervised Learning: There is no target or outcome variable to estimate in this algorithm. It used to group population in different clusters, which is widely used for segmenting different groups of customers for specific intervention. Apriori algorithm, K-means are the examples of this algorithm.

Reinforcement Learning: Here the machine is trained to make specific decisions. The machine is open to an environ ment where it trains itself continually using trial and error. For this machine to learn. it will consider past experience and will absorb the best knowledge to make accurate business decisions. Markov Decision Process is one of the examples of this algorithm.

Supervised Learning: They comprise of a target (outcome) variable (or dependent variable) to be estimated from a given set of predictors (independent variables). Using these set of variables, a function that maps inputs to desired outputs is created. The training process continues until the model achieves a preferred accuracy level on the training data. Decision Tree, Random Forest, KNN, Logistic Regression, etc. among others are some of these types of learning

To predict the possibility of an employee leaving the company, authors in made a comparison between a decision. tree algorithm J48 Naive Bayes classifier and Naïve Bayes classifier while concentrated on using Machine Learning Techniques to predict Employee attrition. We focused on using Machine learning techniques and analysis of machine learn ing classification algorithms for retention predictive decision making amongst target employee candidates and the features that affect these retention decisions.

Initially the data downloaded from Kaggle is pre-processed so that we can extract important features such as Monthly Income, Final Year Promotions, Salary Increases and more. which is really natural in reducing workload? Dependent variables or predictive variables are the ones that help determine the most dependent factors for employee-related variables. For example, an employee ID or employee number has nothing to do with the expiration date.

Data Test Analysis is the first process of analysis, where you can summarize data features so that you can predict who, and when an employee will terminate a service. The system creates a predictive model using a random forest process. It is one of the learning ensembles that combines fewer trunks than a single deciding decision tree.

Strategies that do not rely on dependent analytics and word processor vectors to assess employee churn. Therefore, by improving staff validation and providing a desirable working environment, we can significantly reduce this problem.

DATASET

The data used for this purpose was "IBM HR Analytics Employee Attrition & Performance". It contains 34 features and 1470 practical examples. The type of feature and the number of unique values in them are given in Fig. 2. The number of employees, Over18 and normal hours are excluded from the list of factors because they all have the same value in all training models. Employee number was also removed from the feature list because it was simply a unique identifier for each employee. In addition, ‘Average Day’, ‘Average Hour’, ‘Month Average’ are also excluded due to their relationship to Monthly Income. This has reduced the list of 27 features. Then the features 'BusinessTravel', 'Gender', 'OverTime', 'Marital Status' were labeled and the 'Department', 'EducationField', 'JobRole' features were coded.
The data was then randomly assigned to have an 80:20 split. training and testing in sequence.

Feature Selection is considered to be the most important theory in the field of machine learning that has a major impact on the actual performance of your model structure. These features can easily be used to train your model and have a great impact on performance. Small and unrelated features can adversely affect the performance of the model. Feature selection and cleaning Data should be the first and most important step in designing your model. Feature Selection is a process where you automatically select or perform those features on the basis of a variety of strategies such as the Univariate Selection Feature Importance Correlation Matrix which contributes significantly to the dependent or output variables you are interested in.

After personally analyzing the data, we came to the conclusion that these features Employee Number, Employee Number, Over18 have no direct effect on our output Arttrition. Therefore, these features were completely ignored before using any of the features.

In the database 90% of the records are labeled YES in class and the remaining 10% records are labeled NO class. These types of databases are called unequal data sets and can have a negative impact on the performance of the model making the model biased in many stages of output variability. Managing an unequal database is therefore a necessary function of this type of problem statement.

In our database we use a sampling method to manage data inequalities. Prior to sampling 1233 records were marked NO and only 237 were marked YES. After conducting several samples, we compared the record for both categories with 1233 records as shown in the diagram below.

Apache Spark

Spark is a Distributed Computer Framework based on Hadoop Map Reduce algorithms. Introduce Hadoop Map Reduce points of interest, yet not quite the same as Map Reduce, spark can keep memory in the middle of the results, called Memory Computing [3]. Memory Computing improves data computing production. Spark is best suited for repetitive applications, for example, Data Mining and Machine Learning.

Spark provides APIs in Java, Scala, Python and R, a customized engine that supports graphical performance graphs. It also supports a large system of high value devices including Spark SQL SQL, MLlib machine learning, GraphX configuration chart, and Spark streaming.

Spark Core incorporates a standard used spark plug engine all that is needed to be useful based on the required method. Provides a built-in memory computer and reference data sets stored on external storage

Spark allows designers to quickly compose code with the help of advanced operators. While it takes a considerable measure of lines of code, it takes fewer lines to compose a similar code in Spark Scala.

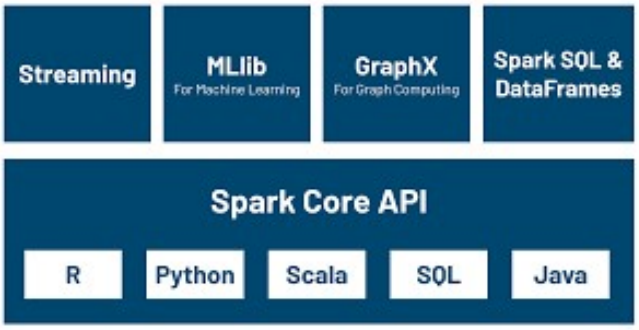

Fig 2. Spark API

Figure showed above Spark's core technologies and components. Each core component of Spark is explained in the following sections of the document.

Spark Streaming
This part enables Spark to process real-time streaming data. It gives an API to control data streams that matches with RDD API. It enables the developers to comprehend the task and switch through the applications that control the data and giving result continuously. Like Spark Core, Spark Streaming endeavours to influence the framework to blame tolerant and adaptable

MLlib (Machine Learning Library)
Apache Spark is outfitted with a rich library known as MLlib. This library contains a wide exhibit of machine learning calculations, classification, clustering and collaboration, and so on. It additionally incorporates few lower-level primitives. Every one of these functionalities enable Spark to scale out over a bunch

Tableau

Tableau is a software company that provides integrated software to display data for organizations that work with business information statistics. Organizations use the Tableau to visualize data and reveal analytical patterns in business intelligence, making the data more understandable. Tableau is an end-to-end data analysis platform that allows you to organize, analyze, share, and share your big data. Tableau is at the forefront of self-help analysis, which allows people to ask new questions of big, governed data and easily share that information across the organization.

Data Processing

STRING INDEXEX

Cord pointer in Python is based on zero, so the first letter on the phone can have a 0, 00:30 point and the next one will be 1, and so on. The character of the last letter will be the length of the character unit minus one. The character of the character in Python is based on zero: the first letter in the series has point 0, the next one has point 1, and so on. The last character index will be the unit length minus one. In any empty series s, s [len (s) -1] and s [-1] both return the last letter.

ONE HOT ENCODER

Enter a category code as a hotline list.
The input to this converter must match the whole values or units of the character, indicating the values taken by the (different) category elements. Features are coded using a hot-text coding scheme — but (aka 'one-of-K' or 'dummy'). This creates a binary column for each category and returns a scattered matrix or similar dense members (depending on the scattered parameter). By default, the encoder receives categories based on different values for each feature. Alternatively, you can re-specify the categories in person. This coding is required to supply phase data to multiple scikit readings, especially straightforward and SVM models with standard characters.

VECTOR ASSEMBLER

Vector Assembler is a converter that combines a given list of columns into a single vector column. It is useful to combine immature features with features produced by different feature converters into a single feature vector, in order to train ML models such as retrofitting objects and cutting trees.

EXPLORATORY DATA ANALYSIS

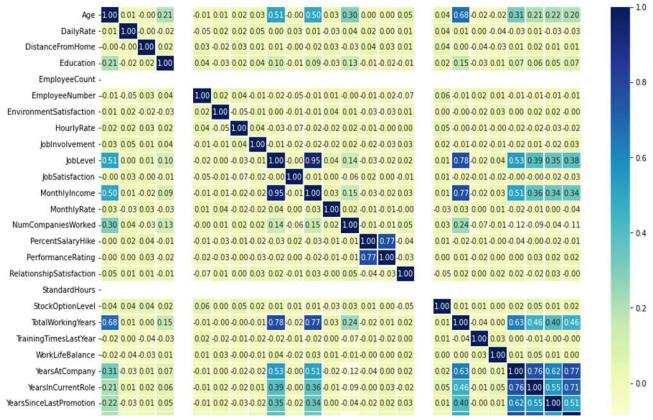

Fig 3. Correlation Matrix

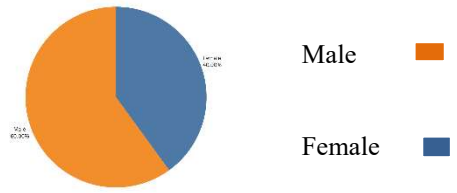

Fig 4. Attrition vs Gender

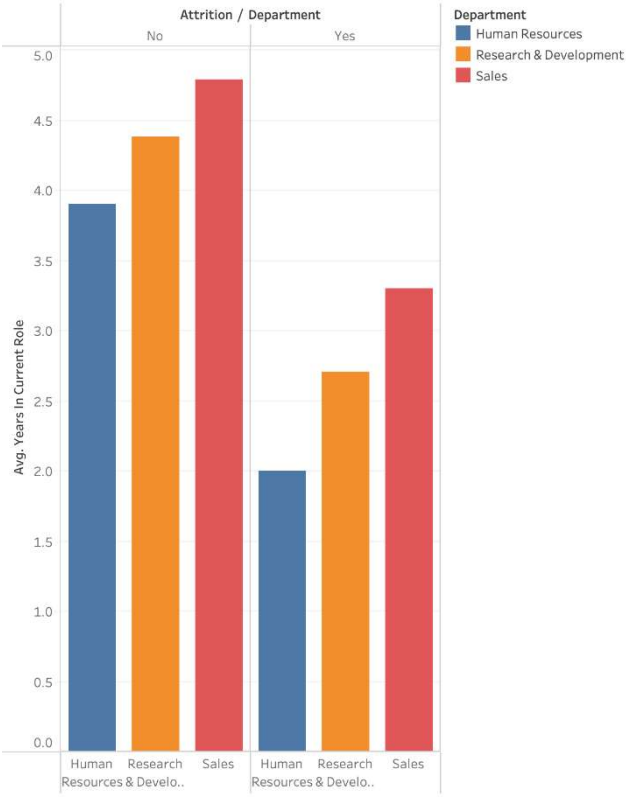

Fig 5. Attrition vs Department

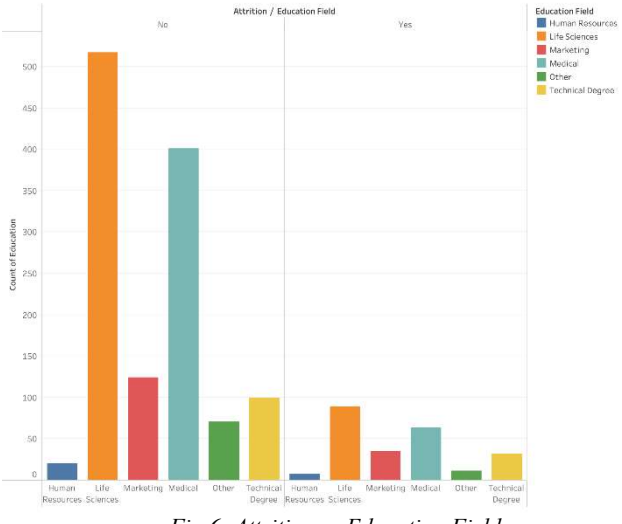

Fig 6. Attrition vs Education Field

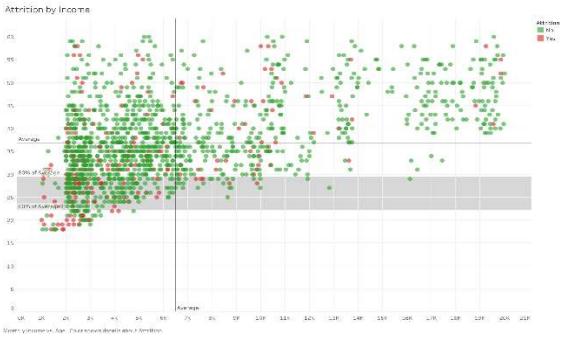

Fig 7. Attrition vs Income

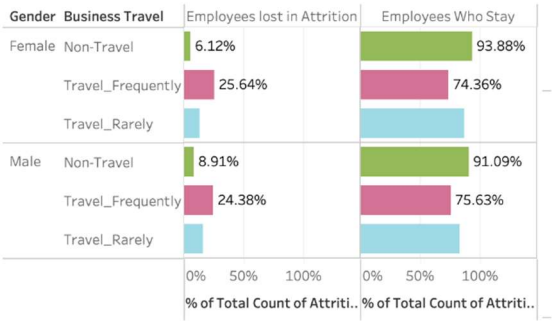

Fig 8. Attrition vs Business Travel

Results and Discussion

Logistic Regression

This is a separation method that is part of a group of line dividers and is almost identical to the polynomial and linear regression. It's fast and basic, and it's worth interpreting the results. Besides being an important method of binary separation, it can be helpful in multi-classroom problems.

Naive Bayes

It is a well-known method of differentiation based on the Bayes' view that assumes independence between predictions, assuming that the existence of a particular element in the class does not depend on the existence of any other factor. These models are easy to build and are particularly useful for very large data sets, known for surpassing even the most complex, fast-moving, simple and generally suitable high-density data sets.

Random Forest Classifier

a supervised machine learning algorithm that combines learning in which different types of algorithms or the same algorithm are often combined to create a more powerful guessing model that can be used in both reversal and segmentation tasks. Although this algorithm is very stable, it works well with a combination of categories and numbers it works best when the data is missing or misplaced, requires a lot of calculating resources, due to the large number of decision trees put together and their complexity. it makes them spend more time training than other comparable algorithms.

Decision Tree Classifier

Model & Their Accuracy

Model	Accuracy
LogisticRegression	89%
NaiveBayes	56.90%
RandomForestClassifier	78.90%
GBTClassifier	84.50%
DecisionTreeClassifier	78.80%
FMClassifier	82.72%

It is a white box type of supervised machine learning algorithm that is widely used for classification problems that works on both phase and continuous phase variables and faster training time compared to those of neural network algorithms. It is a non-distributional or non-distributional method that does not depend on distribution opportunities. More speculation can also resolve high-volume data with good accuracy.

FMClassifier (Equipment)

Factorization equipment is able to measure the interaction between features even in large-scale problems (such as advertising and recommendation program). The implementation of spark.ml supports automation systems for binary split and retrofitting.

GBTClassifier

Gradient-Boosted Trees (GBTs) are a combination of cutting trees. GBTs repeatedly train cutting trees to minimize losses. The use of spark.ml supports GBTs in binary split and retreat, using both continuous and segmental features.

Conclusion & Future Work

Staff retention is a major problem for many employers; effective management teams should recognize the importance of retaining highly productive employees. Higher income leads to the loss of important employees whose replacement is costly.

The departure of employees from an organization may be the subject of a number of exceptions; However, while some are avoided, some aspects may be too strong for the employer to control. It is important that employers identify these factors and develop strategies to prevent their occurrence.

In order to reduce the risk of high profit margins, employers need to use specific strategies to improve job satisfaction as well as retention. Essentially, an employer should review his or her compensation packages, labor relations, job opportunities and developments, and support for work to facilitate the retention of high-quality employees. Employers should offer competitive compensation packages depending on the skills and knowledge of their employees and the length of time they have worked.

In addition, employers may allow high-level employees to design their work schedule or flexible working hours. Otherwise, employers should compensate employees for any additional hours. Therefore, employers need to develop strategies that promote internal cohesion. In addition, a good working environment is well stocked with goods and services while employees are trained to make the best use of their resources.

Taking all of these points into consideration, we have implemented various models which consider the employee's satisfaction, education level, work experience, etc and our best performing model is logistic regression with an accuracy of 89.49%

References

[1] Abhay Charan Patro, Saad Aziz Zaidi, Aaradhya Dixit, Manish Dixit," A Novel Approach to Improve Employee Retention Using Machine Learning ",10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), pp. 680, 2021.

[2] Guru Renuka, Kankipati Anitha, Kella Lavanya, Kethavarapu Naga Syamala Gowthami, Mondru Sion Kumari," EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING", International Research Journal of Modernization in Engineering Technology and Science, pp. vol:04,2022.

[3] Sai Chandan P Reddy, Priyanshi Thakur," Employee Future Prediction", International Journal of Innovative Science and Research Technology, pp. vol. 6, Issue 11,2021.

[4] Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, Hanan S. Alghamdi," Prediction of Employee Attrition Using Machine Learning and Ensemble Methods", International Journal of Machine Learning and Computing, pp. vol. 11, No. 2, 2021.

[5] Setiawan, I., et al. "HR analytics: Employee attrition analysis using logistic regression." IOP Conference Series: Materials Science and Engineering, pp. vol. 830. No. 3. IOP Publishing, 2020.

[6] M Marchington et al., Human resource management at work: The definitive guide., London, England: Kogan Page, 2020.

[7] Xiang Gao, Junhao Wen, Cheng Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover", Mathematical Problems in Engineering, pp. vol. 2019, Article ID 4140707, 2019.

[8] Rombaut, E. and Guerry, "Predicting voluntary turnover through human resources database analysis", Management Research Review, Vol. 41 No. 1, pp. 96-112,2018.

[9] Sarah S. Alduayj; Kashif Rajpoot, "Predicting Employee Attrition using Machine Learning",International Conference on Innovations in Information Technology (IIT),pp.93, 2018.

[10] Srivastava, Devesh Kumar, and Priyanka Nair. "Employee attrition analysis using predictive techniques." International Conference on Information and Communication Technology for Intelligent Systems, pp. SIST, vol:83,2017.

[11] Manisha Purohit," A Study on - Employee Turnover in IT Sector with Special Emphasis on Wipro and Infosys", IOSR Journal of Business and Management (IOSR-JBM), pp. Vol.18, Issue 4.Ver. I, 2016.