International Journal of Electrical Electronics & Computer Science Engineering
Special Issue - NCSCT-2018 | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
March, 2018 | Available Online at www.ijeecse.com

# Predict Employee Retention Using Data Science

Dr. Anil Kumar Dubey[1], Ila Maheshwari[2], Ashutosh Mishra[3]
[1]Assosciate Professor, [2-3]UG Scholar, Dept. of CSE, Poornima Institute of Engineering & Technology, Jaipur, India
[1]anil.dubey@poornima.org, [2]2014pietcsila@poornima.org, [3]2014pietcsashutosh026@poornima.org

***Abstract:*** ***Now a day's data science predictions are used in IT industries, for the improvement in market investment, employee management etc. Retention of valuable employees within an organization has become an important issue as it is hard to find out the reasons that why employees are leaving an organization and keep them satisfied is a big challenge, for this a report ismade to predict the retention of an employee in an organization using the python programming with data science methods. The main idea of this report is to find out that which valuable employee will leave the company and the features which are affecting him/her to making this decision like salary level, no. of hours spending in a week, promotion, no. of work accident etc. The application was developed in python programming language and prediction are made with the help of data science and machine learning models.The design criteria and the implementation details are presented in this report.***

***Keywords: Data Science, Preprocessing Techniques, Machine learning, Supervised Learning, Logistic Regression.***

## I. INTRODUCTION

Data mining is the next big in the world of Information Technology, usage of data extraction is increasing day by day. Data science is the process of mining of useful insights from larger amount of data to use it for the development purpose. To extract data several algorithms, methods and analyzing processes are used depending upon the kind of data we have and what the analyst intended to do with the data. The data we get is in the form of raw data, it needs to get preprocessed to make it in the form to apply algorithm on it. Preprocessing techniques includes collection, noise removal, data reduction, transformation etc. data science methodologies are mainly classified in two categories as making prediction and pattern discovery, prediction making is the process of producing estimate result by analyzing previous results known as regression or supervised learning and pattern discovery is that method when we apply different approaches to find out similarities and dissimilarities in the given data by assigning class notations which is known as clustering or unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.A Data Analyst as a rule clarifies what is happening by handling history of the information. Then again, Data Scientist not exclusively does the exploratory investigation to find bits of knowledge from it, yet in addition utilizes different propelled machine learning calculations to recognize the event of a specific occasion later on. A Data Scientist will take a gander at the information from numerous edges, at times edges not known before. Along these lines, Data Science is essentially used to settle on choices and forecasts making utilization of prescient causal examination, prescriptive investigation (prescient in addition to choice science) and machine learning.

We know that larger companies contain more than thousand employees working for them, so taking care of the needs and satisfaction of each employee is a challenging task to do, it results in valuable and talented employees leave the company without giving the proper reason. This paper provides solution for the given problem as it gives a prediction model that can be used to predict which employee will leave the company and which will not leave. It also helps in finding the exact reasons which are motivating the employees for shifting companies like lower salary, less promotions or heavy work load etc. To find the result in the form of yes or no, we have used logistic regression method, which predicts result in binary values that are 0 or 1, 0 means employee will not leave the company and 1 means he/she will.

## II. PREVIOUS WORK

Retention of valuable employee within an organization is a major issue in the companies, so several efforts are made to find out the proper employee management policies in the companies, we are discussing some work from them –

Piotr Płoński (MLJAR) et.al [1] proposed the analytic methods those can improve Human Resources (HR) management for companies with large number of employees by providing approaches to predict employee attrition with machine learning. They used 1200 employee's data for training datasets, which contains description, but the retention is unknown, which is predicted using binary classification.

Le Zhang and Graham Williams et.al [2] proposed that employee retention is the biggest challenge for a company, so it is important for company to recognize behavioural patterns to understand their employees better. They used R for predictions by feature extraction methods as word-to-vector, term frequency, or term frequency and inverse document frequency, R packages such as tm etc. They finally concluded that ensemble techniques can be deployed to effectively boost model performance.

Ashish Mishra et.al [3] proposed that it is first important to recruit right person to do talent management, the easily available data source for

present and past candidates is their resume. This paper provides a method to calculate the employee score using his educational and business experience scores. They concluded that information like number of years of education, number of organizations worked for, number of positions held in the past, and age can be easily translated into a score for every employee which can be used for predicting retention.

Rupesh Khare, Dimple Kaloya and Gauri Gupta et.al [4] proposed that a risk equation can be develop, which can be used assess attrition risk with current set of employees that a company is having. They concluded by stating that among the various attrition predictive techniques available in the market, Logistic Regression and Discriminant Analysis are the closest to give a solution which produced highly accurate results.

Randy Lao et.al [5] states that a company which make healthy environment and provide equal opportunities for employees to glow, grows rapidly. Their goal is to create a model that help in improving retention strategies on targeted employees. He used R programming language and, they concluded by saying that employees having higher satisfaction and evaluation rate will have fewer chance to leave the company.

## III. MACHINE LEARNING

Machine learning is the process of making the machine tolearn itself through patterns and training data sets. Training data sets are data which is given to machine for understanding the hidden patterns within data and make relations for own understanding. It helps in working of machines efficiently by making them processed like a human brain. Pattern recognition is the most challenging task for developers to use such algorithms that allows different machines to work according to the requirement.

This paper emphasizes on making prediction of retention of an employee within an organization such that whether the employee will leave the company or continue with it. It uses the data of previous employees which have worked for the company and by finding pattern it predicts the retention in the form of yes or no. It uses various parameters of employees such as salary, number of years spent in the company, promotions, number of hours, work accident, financial background etc.

Considering new processing innovations, machine adapting today isn't care for machine learning of the past. It was conceived from design acknowledgment and the hypothesis that PCs can learn without being customized to perform assignments; specialists intrigued by manmade brainpower et.al [6] needed to check whether PCs could gain from information. The iterative part of machine learning is essential claiming as models are presented to new information, they can freely adjust. They gain from past calculations to deliver solid, repeatable choices and results. It's a science that is not new – but rather one that is increasing crisp energy. While numerous machine learning calculations have been around for quite a while, the capacity to naturally apply complex scientific computations to huge information again and again, quicker and speedier is a current advancement.

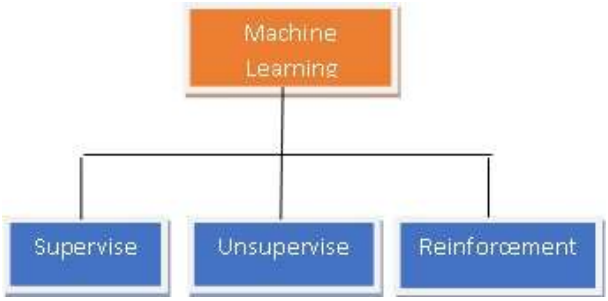



Fig. 1. Types of Machine Learning

Machine learning algorithms are differentiated as supervised or unsupervised.

*A.* Supervised machine learning calculations can apply what has been realized in the past to new information utilizing marked cases to anticipate future occasions. Beginning from the examination of a known preparing dataset, the learning calculation creates a surmised capacity to make expectations about the yield esteems. The framework can give focuses to any new contribution after adequate preparing. The learning calculation can likewise contrast its yield and the right, planned yield and discover mistakes to adjust the model appropriately.

*B.* In differentiate, unsupervised machine learning calculations are utilized when the data used to prepare is neither grouped nor named. Unsupervised learning contemplates how frameworks can induce a capacity to portray a concealed structure from unlabeled information. The framework doesn't make sense of the correct yield; however, it investigates the information and can attractderivations from datasets to depict concealed structures from unlabeled information.

*C.* Semi-directed et.al [7] machine learning calculations fall some place in the middle of regulated and unsupervised learning, since they utilize both marked and unlabeled information for preparing – ordinarily a little measure of named information and a lot of unlabeled information. The frameworks that utilization this strategy can significantly enhance learning precision. For the most part, semi-administered learning is picked when the procured named information requires gifted and significant assets to prepare it/gain from it. Something else, obtaining unlabeled information by and large doesn't require extra assets.

*D.* Reinforcement machine learning calculations is a learning technique that interfaces with its condition by creating activities and finds mistakes or rewards. Experimentation seek and postponed compensate are the most pertinent attributes of fortification learning. This technique enables machines and programming operators to naturally decide the perfect conduct inside a setting to augment its execution. Basic reward input is required for the specialist to realize which activity is ideal; this is known as the support flag.

## IV. TECHNOLOGY

We have utilized Python programming dialect, which is a translated, progressively written dialect and least difficult in grammar. Python is utilized for every one of the applications like in IOT advancement, information science field, web improvement, scripting reason and so forth. Consequently, now it is being utilized generally over the globe.

Python contains various number of libraries accessible in it, this makes it simple to use for each application like for web rejecting delightful cleanser, for GUI improvement TKinter, for web network urlib2, for machine learning sklearn et.al [8], numpy, pandas and so on. Python is one of the for the most part utilized dialect for Data Science applications since it gives libraries, for example, Pandas, nltk which can oversee substantial number of datasets into fitting way, it gives representation libraries like Matplotlib, Bokeh, Seaborn and so on that are exceedingly expressive regarding charts and plots portrayals.

The sklearn library is one which gives bigger number of machine learning calculations, for example, direct and various relapse, polynomial relapse, choice tree characterization and so on., to make expectations, bunching and grouping of information in number of billions.Machine learning is a branch in software engineering that reviews the outline of calculations that can learn. Run of the mill errands are idea learning, work learning or "prescient demonstrating", bunching and finding prescient examples. These undertakings are found out through accessible information that were seen through encounters or directions, for instance. The expectation that accompanies this teach is that including the experience into its assignments will in the end enhance the learning. However, this change needs to occur such that the learning itself ends up programmed with the goal that people like ourselves don't have to meddle any longer is a definitive objective.

Scikit-learn is the most helpful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a great deal of efficient devices for machine learning and factual displaying including arrangement, relapse, bunching and dimensionality lessening. Scikit-learn gives a scope of directed and unsupervised learning calculations through a reliable interface in Python. It is authorized under a lenient disentangled BSD permit and is circulated under numerous Linux appropriations, empowering scholastic and business utilize.

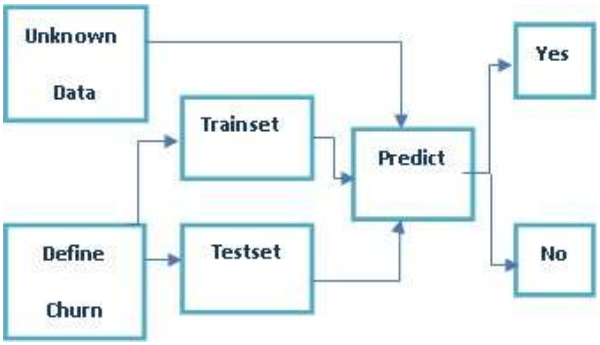



Fig. 2. Prediction Methodology

## V. PREPROCESSING TECHNIQUES

In straightforward words, pre-preparing et.al [9] alludes to the changes connected to the information before nourishing it to the calculation. In python, scikit-learn library has a pre-assembled usefulness under sklearn. pre-processing. The information we get from client is as crude information, so it needs to get perfect, change and decrease to make it proper for applying strategies on it, this procedure is known as preprocessing. require scientific sandbox in which you can perform examination for the whole term of the task. You have to investigate, preprocess and condition information preceding demonstrating. Further, you will perform ETLT (remove, change, stack and change) to get information into the sandbox. It enhances the general nature of the information and effectiveness of the model to deliver comes about. There are numerous more alternatives for pre-preparing as –


Fig. 3. Preprocessing Techniques

### *A. Feature Scaling:*

Highlight scaling is the strategy to restrict the scope of factors with the goal that they can be thought about on basic grounds. It is performed on constant factors.

### *B. Label Encoding:*

Sklearn gives an extremely proficient device to encoding the levels of an all-out highlights into numeric esteems. Name Encoder encode names with an incentive about 0 and classes-

*C. One-Hot Encoding:*

One-Hot Encoding changes each clear-cut component with n conceivable esteems into n parallel highlights, with just a single dynamic. Most of the ML calculations either take in a solitary weight for each component or it figures remove between the examples.

## VI. METHODOLOGY USED FOR PREDICTION

Utilizing this expectation demonstrate, which intends to foresee whether a representative will proceed or leave the association based upon the investigation of the information of past workers. The expectation factors incorporate fulfillment level, last assessment, normal month to month hours, compensation, work mischance, advancement, time spent at the organization and division, in view of these parameters, diverse machine learning models like calculated relapse, choice tree order and so forth are connected to foresee which worker will leave straightaway and the variables that are most huge in this choice.

In measurable demonstrating, relapse investigation is an arrangement of factual procedures for assessing the connections among factors. It incorporates numerous systems for displaying and dissecting a few factors, when the emphasis is on the connection between a reliant variable and at least one free factors (or 'indicators'). More particularly, relapse examination causes one to see how the run of the mill estimation of the needy variable (or 'model variable') changes when any of the free factors is fluctuated, while the other autonomous factors are held settled.

Most regularly, relapse investigation evaluates the restrictive desire of the needy variable given the autonomous factors – that is, the normal estimation of the reliant variable when the free factors are settled. Less regularly, the attention is on a quantile, or other area parameter of the restrictive conveyance of the reliant variable given the autonomous factors. In all cases, a component of the free factors called the relapse work is to be evaluated. In relapse investigation, it is additionally important to portray the variety of the needy variable around the forecast of the relapse work utilizing a likelihood conveyance. A related however particular approach is Necessary Condition Analysis (NCA), which gauges the most extreme (instead of normal) estimation of the needy variable for a given estimation of the autonomous variable (roof line as opposed to focal line) to recognize what estimation of the free factor is important yet not adequate for a given estimation of the reliant variable.

Relapse investigation is broadly utilized for expectation and estimating, where its utilization has considerable cover with the field of machine learning. Relapse examination is likewise used to comprehend which among the autonomous factors are identified with the needy variable, and to investigate the types of these connections. In confined conditions, relapse investigation can be utilized to induce causal connections between the autonomous and ward factors. However, this can prompt figments or false connections, so alert is advisable; for instance, relationship does not demonstrate causation.

Numerous strategies for completing relapse investigation have been created. Well-known techniques, for example, straight relapse and common minimum squares relapse are parametric, in that the relapse work is characterized as far as a limited number of obscure parameters that are evaluated from the information. Nonparametric relapse alludes to strategies that permit the relapse capacity to lie in a predefined set of capacities, which might be endless dimensional.

Through this expectation show an organization can choose its arrangements to keep great representatives from leaving the organization. Information science part that utilized as a part of this venture is to take crude information from csv record and then apply distinctive preparing system to settle on information valuable in settling on choices from it like arrangement of dataset, LabelEncoding, OnehotEncoding and highlight scaling. Relapse is the most widely recognized technique utilized for making expectation utilizing python programming dialect. Relapse examination likewise enables us to look at the impacts of factors estimated on various scales, for example, the impact of value changes and the quantity of limited time exercises. These advantages help economic specialists/information experts/information researchers to dispose of and assess the best arrangement of factors to be utilized for building prescient models.

*A. Linear Regression:*

Coordinate backslide is the path toward finding the association between two ward factors using a straight condition. It is the most principal kind of making figures using backslide that is known as coordinated learning, in it a planning dataset is used to set up the machine with the objective that when we ask for to impact desires it to will can make comes to fruition using the association between the components. It can be used for most prominent two elements for various variable conjectures polynomial backslide is used. It produces data as some motivating force after associated distinctive preprocessing methods. It is the most broadly perceived system used for fitting a backslide line. It figures the best-fit line for the watched data by constraining the aggregate of the squares of the vertical deviations from each datum point to the line. Since the deviations are first squared, when included, there is no counterbalancing among positive and negative regards.

*B. Polynomial Regression:*

Polynomial backslide is the methodology in which association between no less than two variables ought to

be find in a polynomial condition shape, later this condition is used for making desire for test dataset. It is the refreshed shape if coordinate backslide, as it can be used for finding association between more than two variables. While there might be a motivation to fit a higher degree polynomial to get cut down botch, this can realize completed the process of fitting. Consistently plot the associations with see the fit and focus on guaranteeing that the curve fits the possibility of the issue. Especially pay uncommon personality to twist towards the terminations and see whether those shapes and examples look good. Higher polynomials can end up conveying wired results on extrapolation.

*C. Logistic Regression:*

Backslide is the route toward making desire the association state of two ward factors. the minimum complex kind of the backslide condition with one dependent and one free factor is portrayed by the condition et.al [10]

$y = m + c*x$

where y = assessed subordinate variable score, m = enduring, c =regression coefficient, and x = score on the self-sufficient variable.

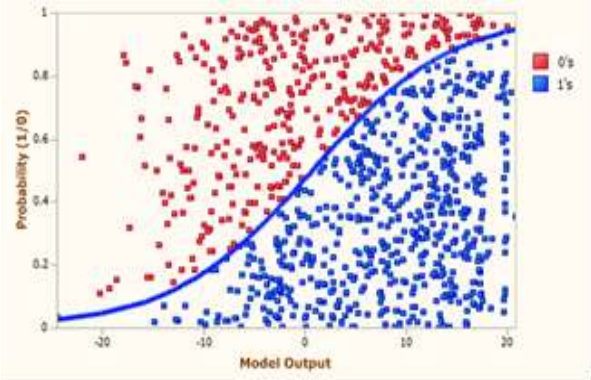



Fig. 4. Logistic Regression

Computed backslide is the one of a kind sort of backslide where desires are made as yes or no as twofold regards. Here we should predict whether specialist will leave or not, so it is the best proper technique for making desires using backslide. It is extensively used for arranging issues; Logistic backslide doesn't require straight association among poor and self-sufficient components. It can manage various types of associations since it applies a non-straight log change to the foreseen chances extent, to keep up a vital separation from over fitting and under fitting, we should consolidate each critical variable. A respectable method to manage ensure this preparation is to use a phase clever system to assess the ascertained backslide, it requires considerable illustration sizes since most outrageous likelihood checks are less extraordinary at low case sizes than standard scarcest square, the free factors should not be associated with each other i.e. no multi collinearity. Regardless, we have the contrasting options to join affiliation effects of full scale factors in the examination and in the model. If the estimations of ward variable are ordinal, by then it is called as Ordinal ascertained backslide, if subordinate variable is multi class then it is known as Multinomial Logistic backslide

*D. Lasso Regression:*

Rope (Least Absolute Shrinkage and Selection Operator) furthermore rebuffs undoubtedly the traverse of the backslide coefficients. Also, it can reduce the capriciousness and upgrading the accuracy of direct backslide models. Tie backslide contrasts from edge backslide in a way that it uses preeminent regards in the discipline work, as opposed to squares. This incite rebuffing (or indistinguishably obliging the aggregate of the aggregate estimations of the examinations) values which makes a part of the parameter assessments turn out absolutely zero. Greater the discipline associated, encourage the evaluations get contracted towards add up to zero. This results to variable assurance out of given n factors.

## VII. RESULT AND DISCUSSION

This report expects to foresee whether a worker will proceed or leave the association in view of the examination of the information of past representatives. The expectation factors incorporate fulfillment level, last assessment, normal month to month hours, pay, work mischance, advancement, time spent at the organization and division, in view of these parameters, distinctive machine learning models like strategic relapse, choice tree characterization and so on are connected to anticipate which worker will leave straightaway and the components that are most critical in this choice.

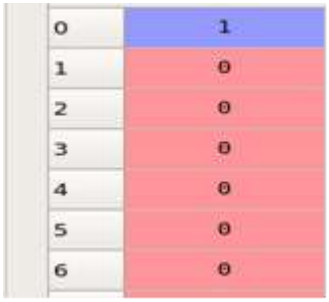
Through this paper an organization can choose its strategies to keep great representatives from leaving the organization. Information science part that utilized as a part of this report is to take crude information from csv document and then apply diverse handling component to settle on information helpful in settling on choices from it like arrangement of dataset, Label Encoding, Onehot Encoding and include scaling.

It at that point applies diverse relapse models to anticipate whether the worker will leave the organization or not as 0 and 1. If 0 comes in the outcome that implies that the worker will proceed with the organization, however if 1 comes then the representative will leave the organization.

Here is given the example information that we utilized for making expectations, it is in an unthinkable frame which contains segments as fulfillment level, last assessment, number of undertakings, normal month to month hours, years spent in the organization, work mischance, advancement, office and pay.

Fig. 5. Dataset for Prediction

When the accuracy of the result is being calculated from the previous analysed data with the help of confusion matrix and the accuracy score, this result is being compared with the available data to find the result accuracy and 97% of the predictions are made correct.



| | |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 0 |
| 6 | 0 |

Fig. 6. Result

The figure contains the result in the form of 0 or 1 as o representing the employee who will not leave the company and 1 representing as employee who will going to leave the company.

## VIII. CONCLUSIONS

In this investigation, we become more acquainted with that maintenance of a representative inside an association can be discover utilizing strategic relapse procedure, which delivers an outcome with 97% exactness. It can likewise help in discovering the components that are influencing the representatives in the association like pay level, work stack, advancements and so forth.

The future extent of information science is brilliant; consequently, this procedure can be utilized as a part of any association for better worker administration and for their fulfillment. This paper can be additionally reached out as it requires information as .csv records just, so this impediment can be expelled.

## ACKNOWLEDGMENT

This examination is guided by Dr. Anil Kumar Dubey, we thank our guide from Poornima Institute of Engineering and Technology, Jaipur who gave understanding and aptitude that enormously helped the examination for this paper.

## IX. REFERENCES

[1] Piotr Płoński (MLJAR), "Human-first Machine Learning Platform," Human Resource Analytics Predict Employee Attrition.

[2] Le Zhang and Graham Williams (Data Scientist, Microsoft), "Employee Retention with R based Data Science Accelarator".

[3] Ashish Mishra (Data Scientist, Experfy), "Using Machine Learning to Predict and explain Employee Attrition".

[4] Rupesh Khare, Dimple Kaloya and Gauri Gupta, "Employee Attrition Risk Assessment using Logistic Regression Analysis," from 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence.

[5] Randy Lao, "Predicting Employee Kernelover," Kaggle.

[6] Sandra W. Pyke & Peter M. Sheridan, "Logistic Regression Analysis of Graduate Student Retention," from The Canadian Journal of Higher Education, Vol. XXIII-2, 1993.

[7] Prof. Dr. Vjollca Hasani and Prof. Dr. Alba Dumi, "Application of Logistic Regression in the Study of Students' Performance Level," Journal of Educational and Social Research Italy.

[8] Dr. Jonathan Erhardt, "Artificial Intelligence: Opportunities and Risks," Policy paper by the Effective Altruism Foundation.

[9] Sofia Stromberg's, "Binary Logistic Regression and its application to data from a study of children's recognition of their own recorded voices" term paper in statically method.

Anish Talwar and Yogesh Kumar, "Machine Learning: An artificial intelligence methodology," from International Journal of Engineering and Computer Science ISSN:2319-7242 Volume 2 Issue 12, Dec.2013PageNo.3400-3404.