IBM - Corporate Employee Attrition Analytics

Introduction

Every year a lot of companies hire several employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees.

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees. The Attrition rate of Employee directly impacts the growth of the company. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

Over past two years, 55% of the American workforce said that they plan on looking for new employment over the next 12 months. Global labour markets have swung dramatically due to the COVID-19 pandemic, and in August 2019, 60% of Americans expect to look for new jobs.

Key factors for Employee Attrition:

For many people, the reasons behind leaving a job can be deeply personal. A clash with a colleague, not receiving a promotion, or a change in life circumstances. However, finding patterns in factors like these is the first step in trying to prevent them in future.

According to Factors influencing employee attrition in Indian ITeS call centres by Neeraj Pandey. There are two types of attrition ‘Drive Attrition’ which is due to Employer and another ‘Drag Attrition’ which is due to Employee i.e., Company policy to lay-off the employees who underperform. Drive Attrition happens because of employer’s policy. The drag attrition is due to insecurities of employees with their career. For Example, only a few employees will get Promoted in an iteration. Some employees will get frustrated and quit their job. It is not sure to happen all department in other industries. Great attention to the employee attrition to the HR is very important. These are the factors that is responsible for Employee Attrition.

Some of the factors mentioned are:

- Higher salary and monetary
- Non-favourable job content and inadequate job enrichment
- Hard to understand customer's accent
- Non-transparent appraisal systems
- Repetitive, mechanical and
- Involves high transaction volumes
- Tightly scripted, heavily monitored and controlled
- Lack of promotions and career advancement opportunities
- Health and psychological ailments
- Problems with client handing
- Uneasy relationships with peers and managers
- Long working hours and work pressure
- Workload and targets
- Shift timings
- Ineffective leadership
- Lack of challenge and opportunity
- Lack of trust in senior management
- Dissatisfaction with the work culture/cross cultural issues
- Insufficient leave and no national holidays
- Non-conducive policies and procedures

Higher salary and monetary

Most of employees in India want transparent performance-linked incentives as a component of their salary structure. The fringe benefits like bonus allowances, social security and leave provisions should be adequately provided as a part of compensation structure for the call centre workforce.

Timing

Flexible working hours, better allocated shift timings and two off days in a week should be given to the employees. Many females left this industry after their marriage because Indian family culture did not promote the night shift work schedules. The length and frequency of breaks should also be adequate.

Career Planning

The study revealed that though the employees had career planning provision, but implementation was not up to the mark. It was virtually non-existent. Due to hectic work schedules and high job pressure, higher studies become a distant dream for the employees. This will also help in attracting the best talent to the company.

Health Problems

Long working hours and long travel hours are increasing the scale of attrition rate. The company may conduct psychometric profiling of applicants to choose employees with high stress-bearing capacity. Employees should also be given counselling regarding stress handling, time management and healthy eating habits.

Appraisal

The appraisal system in a company should be transparent, timely and based on performance-based metrics. HR department should make use of early warning system, which uses RAG analysis (red, amber and green) to identify employees who are likely to quit or stay.

Factors associated with employee retention:

- Employee - centric HR Policies
- Efforts to keep the workforce motivated
- Satisfaction with working
- Security of the job
- Resolution of grievances

Compensation and Benefits:

- Adequate perks
- Post-retirement benefits
- Linking of performance with adequate rewards
- Foreign trips

Data set parameters (with an example)

Dataset sample 1:

Name	Count	Mean	Std	Min	25%
Age	1470.0	36.923810	9.135373	18.0	30.0
Daily Rate	1470.0	802.485714	403.509100	102.0	465.00
Distance From Home	1470.0	9.192517	8.106864	1.0	2.00
Education	1470.0	2.912925	1.024165	1.0	2.00
Employee Count	1470.0	1.000000	0.000000	1.0	1.00
Employee Number	1470.0	1024.865306	602.024335	1.0	491.25
Environment Satisfaction	1470.0	2.721769	1.093082	1.0	2.00
Hourly Rate	1470.0	65.891156	65.891156 20.	30.0	48.00
Job Involvement	1470.0	2.729932	0.711561	1.0	2.00
Job Level	1470.0	2.063946	1.106940	1.0	1.00
Job	1470.0	2.728571	1.102846	1.0	2.00

	coef	std err	z	P>|z|
Intercept	-1.5561	1.120	-1.389	0.165
Age	-0.0103	0.016	-0.630	0.529
BusinessTravel_Travel_Frequently	1.8565	0.510	3.639	0.000
BusinessTravel_Travel_Rarely	1.0758	0.477	2.254	0.024
DailyRate	-0.0003	0.000	-1.271	0.204
DistanceFromHome	0.0351	0.013	2.708	0.007
EducationField_Marketing	0.5042	0.384	1.314	0.189
EducationField_Medical	0.1811	0.251	0.721	0.471
EducationField_Other	0.1935	0.442	0.438	0.661
EducationField_Technical_Degree	1.2845	0.370	3.471	0.001
EnvironmentSatisfaction_2	-0.9721	0.326	-2.985	0.003
EnvironmentSatisfaction_3	-1.0769	0.298	-3.608	0.000
EnvironmentSatisfaction_4	-1.2422	0.305	-4.071	0.000
JobInvolvement_2	-1.1635	0.425	-2.740	0.006

Source: ivpanda.com

Dataset sample 2:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	Age	Attrition	BusinessT	DailyRate	Departme	DistanceF	Education	Education	Employee	Employee	Environm	Gender	HourlyRat	JobInvolv	JobLevel	JobRole	JobSatisfa	MaritalSta	MonthlyIr	MonthlyR	NumCom	Over18	OverTime	PercentSa	Performa	Relations	StandardH	StockOpti	TotalWork Tr
2	41	Yes	Travel_Ra	1102	Sales	1	2	Life Scien	1	1	2	Female	94	3	2	Sales Exec	4	Single	5993	19479	8	Y	Yes	11	3	1	80	0	8
3	49	No	Travel_Fre	279	Research	8	1	Life Scien	1	2	3	Male	61	2	2	Research	2	Married	5130	24907	1	Y	No	23	4	4	80	1	10
4	37	Yes	Travel_Ra	1373	Research	2	2	Other	1	4	4	Male	92	2	1	Laborator	3	Single	2090	2396	6	Y	Yes	15	3	2	80	0	7
5	33	No	Travel_Fre	1392	Research	3	4	Life Scien	1	5	4	Female	56	3	1	Research	3	Married	2909	23159	1	Y	Yes	11	3	3	80	0	8
6	27	No	Travel_Ra	591	Research	2	1	Medical	1	7	1	Male	40	3	1	Laborator	2	Married	3468	16632	9	Y	No	12	3	4	80	1	6
7	32	No	Travel_Fre	1005	Research	2	2	Life Scien	1	8	4	Male	79	3	1	Laborator	4	Single	3068	11864	0	Y	No	13	3	3	80	0	8
8	59	No	Travel_Ra	1324	Research	3	3	Medical	1	10	3	Female	81	4	1	Laborator	1	Married	2670	9964	4	Y	Yes	20	4	1	80	3	12
9	30	No	Travel_Ra	1358	Research	24	1	Life Scien	1	11	4	Male	67	3	1	Laborator	3	Divorced	2693	13335	1	Y	No	22	4	2	80	1	1
10	38	No	Travel_Fre	216	Research	23	3	Life Scien	1	12	4	Male	44	2	3	Manufactu	3	Single	9526	8787	0	Y	No	21	4	2	80	0	10
11	36	No	Travel_Ra	1299	Research	27	3	Medical	1	13	3	Male	94	3	2	Healthcar	3	Married	5237	16577	6	Y	No	13	3	2	80	2	17
12	35	No	Travel_Ra	809	Research	16	3	Medical	1	14	1	Male	84	4	1	Laborator	2	Married	2426	16479	0	Y	No	13	3	3	80	1	6
13	29	No	Travel_Ra	153	Research	15	2	Life Scien	1	15	4	Female	49	2	2	Laborator	3	Single	4193	12682	0	Y	Yes	12	3	4	80	0	10
14	31	No	Travel_Ra	670	Research	26	1	Life Scien	1	16	1	Male	31	3	1	Research	3	Divorced	2911	15170	1	Y	No	17	3	4	80	1	5
15	34	No	Travel_Ra	1346	Research	19	2	Medical	1	18	2	Male	93	3	1	Laborator	4	Divorced	2661	8758	0	Y	No	11	3	3	80	1	3
16	28	Yes	Travel_Ra	103	Research	24	3	Life Scien	1	19	3	Male	50	2	1	Laborator	3	Single	2028	12947	5	Y	Yes	14	3	2	80	0	6
17	29	No	Travel_Ra	1389	Research	21	4	Life Scien	1	20	2	Female	51	4	3	Manufactu	1	Divorced	9980	10195	1	Y	No	11	3	3	80	1	10
18	32	No	Travel_Ra	334	Research	5	2	Life Scien	1	21	1	Male	80	4	1	Research	2	Divorced	3298	15053	0	Y	Yes	12	3	4	80	2	7
19	22	No	Non-Trave	1123	Research	16	2	Medical	1	22	4	Male	96	4	1	Laborator	4	Divorced	2935	7324	1	Y	Yes	13	3	2	80	2	1
20	53	No	Travel_Ra	1219	Sales	2	4	Life Scien	1	23	1	Female	78	2	4	Manager	4	Married	15427	22021	2	Y	No	16	3	3	80	0	31
21	38	No	Travel_Ra	371	Research	2	3	Life Scien	1	24	4	Male	45	3	1	Research	4	Single	3944	4306	5	Y	Yes	11	3	3	80	0	6
22	24	No	Non-Trave	673	Research	11	2	Other	1	26	1	Female	96	4	2	Manufactu	3	Divorced	4011	8232	0	Y	No	18	3	4	80	1	5
23	36	Yes	Travel_Ra	1218	Sales	9	4	Life Scien	1	27	3	Male	82	2	1	Sales Rep	1	Single	3407	6986	7	Y	No	23	4	2	80	0	10
24	34	No	Travel_Ra	419	Research	7	4	Life Scien	1	28	1	Female	53	3	3	Research	2	Single	11994	21293	0	Y	No	11	3	3	80	0	13
25	21	No	Travel_Ra	391	Research	15	2	Life Scien	1	30	3	Male	96	3	1	Research	4	Single	1232	19281	1	Y	No	14	3	4	80	0	0
26	34	Yes	Travel_Ra	699	Research	6	1	Medical	1	31	2	Male	83	3	1	Research	1	Single	2960	17102	2	Y	No	11	3	3	80	0	8
27	53	No	Travel_Ra	1282	Research	5	3	Other	1	32	3	Female	58	3	5	Manager	3	Divorced	19094	10735	4	Y	No	11	3	4	80	1	26
28	32	Yes	Travel_Fre	1125	Research	16	1	Life Scien	1	33	2	Female	72	1	1	Research	1	Single	3919	4681	1	Y	Yes	22	4	2	80	0	10
29	42	No	Travel_Ra	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2	Sales Exec	2	Married	6825	21173	0	Y	No	11	3	4	80	1	10
30	44	No	Travel_Ra	477	Research	7	4	Medical	1	36	1	Female	42	2	3	Healthcar	4	Married	10248	2094	3	Y	No	14	3	4	80	1	24
31	46	No	Travel_Ra	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5	Manager	1	Single	18947	22822	3	Y	No	12	3	4	80	0	22
32	33	No	Travel_Ra	924	Research	2	3	Medical	1	39	3	Male	78	3	1	Laborator	4	Single	2496	6670	4	Y	No	11	3	4	80	0	7
33	44	No	Travel_Ra	1459	Research	10	4	Other	1	40	4	Male	41	3	2	Healthcar	4	Married	6465	19121	2	Y	Yes	13	3	4	80	0	9
34	30	No	Travel_Ra	125	Research	9	2	Medical	1	41	4	Male	83	2	1	Laborator	3	Single	2206	16117	1	Y	No	13	3	1	80	0	10
35	39	Yes	Travel_Ra	895	Sales	5	3	Technical	1	42	4	Male	56	3	2	Sales Rep	4	Married	2086	3335	3	Y	No	14	3	3	80	1	19
36	24	Yes	Travel_Ra	813	Research	1	3	Medical	1	45	2	Male	61	3	1	Research	4	Married	2293	3020	2	Y	Yes	16	3	1	80	1	6
37	43	No	Travel_Ra	1273	Research	2	2	Medical	1	46	4	Female	72	4	1	Research	3	Divorced	2645	21923	1	Y	No	12	3	4	80	2	6
38	50	Yes	Travel_Ra	869	Sales	3	2	Marketing	1	47	1	Male	86	2	1	Sales Rep	3	Married	2683	3810	1	Y	Yes	14	3	3	80	0	3

WA_Fn-UseC_-HR-Employee-Attriti

Source: https://assets.researchsquare.com/files/rs1833481/v1/e3f6fc84483f97f1d20b2079.zip

Dataset sample 3:

	table id	name	phone number	Location	Emp. Group	Function	Gender	Tenure	Tenure Grp.	Experience (YY.MM)	Marital Status	Age in YY.	Hiring Source	Promoted/Non Promoted	Job Role Match	Stay/Left
0	1	sid	9876544345	Pune	B2	Operation	Male	0.00	< =1	6.08	Single	27.12	Direct	Non Promoted	Yes	Left
1	2	sid	9876544345	Noida	B7	Support	Male	0.00	< =1	13.00	Marr.	38.08	Direct	Promoted	No	Stay
2	3	sid	9876544345	Bangalore	B3	Operation	Male	0.01	< =1	16.05	Marr.	36.04	Direct	Promoted	Yes	Stay
3	4	sid	9876544345	Noida	B2	Operation	Male	0.01	< =1	6.06	Marr.	32.07	Direct	Promoted	Yes	Stay
4	5	sid	9876544345	Lucknow	B2	Operation	Male	0.00	< =1	7.00	Marr.	32.05	Direct	Non Promoted	Yes	Stay

Source: https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/

To study about the factors that lead to employee attrition people and companies have used datasets that contain certain common features like Age, Employee Role, Daily Rate, Job Satisfaction, Years at Company, Years in Current Role etc in their data set. As we can see in most of the references listed below the data is collected and maintained by the HR team to analyse employee attrition and also to analyse other aspects as well. Below are some of the sample datasets used in some of the references mentioned below.

Methodologies:

Logistic Regression model and CART:

Logistic regression model and CART is used to determine the probability of a certain employee to fall into the condition of Attrition and thus its high risk of leaving the company. Then different parameters were tested and probability threshold using confusion Matrixes, Area under the Curve and Gini Coefficient to determine which of the three models is the best predictor and will recommend its use in practice.

Logistic Regression:

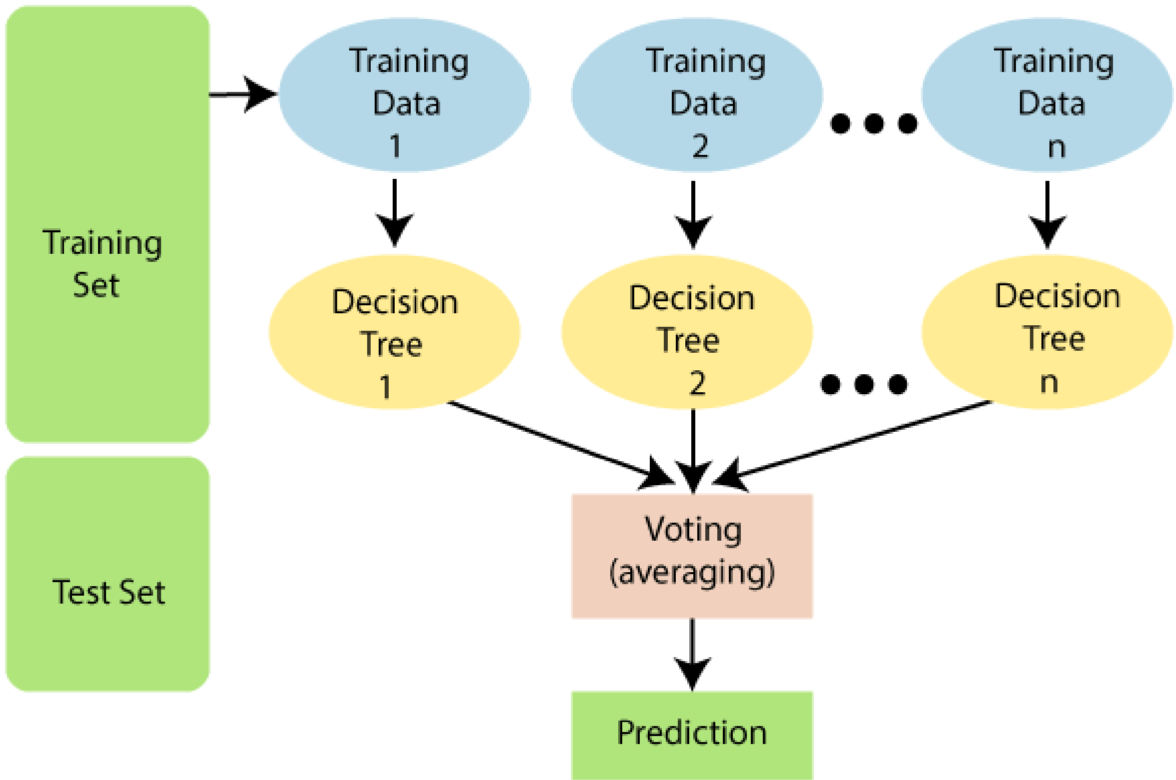
Logistic Regression is a method similar to linear regression except that the dependent variable is discrete (e.g., 0 or 1). Linear logistic regression estimates the coefficients of a linear model using the selected independent variables while optimizing a classification criterion. Given a set of independent variables, the output of the estimated logistic regression (the sum of the products of the independent variables with the corresponding regression coefficients) can be used to assess the probability an observation belongs to one of the classes. Specifically, the regression output can be transformed into a probability of belonging to, say, class 1 for each observation. The default decision is to classify each observation in the group with the highest probability.

CART:

CART is a widely used classification method largely because the estimated classification models are easy to interpret. This classification tool iteratively “splits” the data using the most discriminatory independent variable at each step, building a “tree” on the way. The CART methods limit the size of the tree using various statistical techniques in order to avoid overfitting the data. For example, using the rpart and rpart.control functions in R, we can limit the size of the tree by selecting the functions’ complexity control parameter cp.The leaves of the tree indicate the number of estimation data observations that “reach that leaf” that belong to each class. A perfect classification would only have data from one class in each of the tree leaves. However, such a perfect classification of the estimation data would most likely not be able to classify well out-of-sample data due to overfitting of the estimation data. One can also use the percentage of data in each leaf of the tree to get an estimate of the probability that an observation (e.g., customer) belongs to a given class. The purity of the leaf can indicate the probability that an observation that "reaches that leaf" belongs to a class. In our case, the probability our validation data belong to class 1.

Random Forest Algorithm:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. The bagging method is used to increases the overall results by combining weak models. In the case of Classification problem, it takes the mode of the classes, predicted in the bagging process. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



The random forest works quite well even with the default parameters. That's one of reason we used RF for this problem. Though this can be improved by tuning hyper parameters of Random Forest classifier. Random forest also doesn't over fit easily because of its randomness feature. One of the best features is that Random Forest model provides the importance of variables/features in the data/model. For this HR Analytics problem, we are interested in knowing which feature/factor contribute the most in the Attrition and RF's one function can give us this information. This is just another reason why we have used RF.

Advantages:

Until late 90's Employee attrition rate analysis was not taken seriously, it was considered as a matter of concern only when companies faced shortage of talent and employees leaving the organization within six months to a year was no joke. Thus, HR managers began giving importance to employee attrition analysis and improvising core company values. The major part of analysing employee attrition is to predict when and why the employee will leave the company.

To analyse that, some of the papers that cited below give some clear understanding of how to analyse the HR data of employees for prediction, in each paper they tried and used machine learning algorithms like random forest algorithm, CART, Logistic regression, Multilayer perceptron classifier (MLP), some authors merged two to three algorithms (Ensemble models) to bring out better results and predictions (combining logistic regression and CART is one example).

All the above-mentioned models and algorithms are very much helpful to analyse certain aspects of employee attrition at a high accuracy level.

Disadvantages:

However, findings revealed that no model up until now could be considered ideal and perfect for each case of business context. Yet, the models chosen in the references mentioned below were pretty much optimal as per their requirements and adequately satisfied the intended goal.

Most of the papers we referenced had these limitations in common the method of data gathering and processing that was used, the low diversity of data sources, the reliability of the information provided, and the overall sample size.

Conclusion:

In reference with the papers below, most of the papers conclude that the future work in this field should be carried out by having a vision in mind to check the validity of the analysis and predictions made by the models and generalizing those models.

Generalizing is important because most of the models and methodologies that exist and are stated in the reference part of this review are applicable only to analyse certain aspects of attrition for certain business context but not a generalized one. As much as predicting, analysing and generalizing are important, Validating the results with real world outcomes are also equally important as to get a clear view of the performance of the proposed methodologies. The future of employee attrition rate analysis could be made more reliable and flexible by researching and implementing the above-mentioned features in the methodologies that will be proposed in the future.

Reference:

1. Factors influencing employee attrition in Indian ITeS call centres -Neeraj Pandey National Institute of Industrial Engineering (NITIE), Vihar Lake, Powai, Mumbai, Maharashtra, India
2. Factors Influencing Employee Retention: An Empirical study with Reference to IT Industry - Prof. Bhavani V1, Assistant professor, MVM Group of institutions, Bengaluru, Amanjot Kaur, Assistant Professor, Bhai Gurdas Institute of Management and Technology, Sangrur.
3. https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/
4. http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html#step_2:_set_up_the_dependent_variable
5. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
6. https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/
7. https://towardsdatascience.com/people-analytics-with-attrition-predictions-12adcce9573f