

Prediction of Employee Turnover using Machine Learning

A H Shanthakumara¹, Divya J², Harshitha H T³, Pallavi L V⁴ and Spoorthy B C S⁵

¹⁻⁵Siddaganga Institute of Technology/Department of Computer Science & Engineering, Tumkur, India
Email: ahskondajji@gmail.com, divyajanappa1234@gmail.com, harshithaht03@gmail.com, pallavivivek765@gmail.com, spoorthybcs@gmail.com

Abstract—IT Industry is very competitive and committed to finding new practices for the future. It is important to overcome business challenges to sustain in the competition. The employee turnover is a systematic issue in most industries from past few years. As employees tend to leave the company when they find better opportunities incurs a loss in terms of money, time and resources and must be replaced with their equivalent skills. The productivity and the profitability of the organizations depends on trained and skilled employees, hence turnover rates need to be monitored

In this work, we forecast the employee turnover rates of the organization using machine-learning algorithms. Data collected on employee work profile. Well-reviewed and Industry accepted algorithms like Naive Bayes, random forest and logistic regression are applied. Result taken from the model giving highest accuracy. Our model demonstrates a strong feedback accumulated from employees, which act as a token of exercise to evaluate individual's interests. With help of such predictive models, organization can overcome the business challenges and achieve competitive success.

Index Terms— Employee Turnover, Naïve Bayes, Random forest, Logistic regression.

I. INTRODUCTION

The Information Technology (IT) Sector in India has grown too wide degree to cover many features of technology and computing. Each company has some way to keep track of their assets and are investing in human resource development to manage the employees[1]. Despite the high-quality presentation by industries, they have to counter universal issue of high employee turnover. Employee turnover occurs when workers choose to depart their current jobs voluntarily in search of better opportunities and urges replacement with new employees. Trade-off of erudite employees affects the organizations in terms of hiring charges, training and development of the new employee. It leads to delay in product delivery and customer resentment towards the company. In this study, we forecast the employee turnover rates of the organization using machine-learning algorithms and compare various methods to examine which method has the best performance. Prediction of employee turnover helps organizations to overcome the business challenges and achieve competitive success [3].

The objectives of this work are:

1. To forecast rate of turnover of employees.
2. To recognize the reasons of employees turnover.
3. Assist manager to maintain employee records and progress of the company.

The organisation of the paper:

Section 1 introduction: It gives description about problem statement on which the project depends. It also list the main objectives we tend to meet at the end of the project.

Section 2-literature survey: it provides the background study done to develop the project.

Section 3 proposed work: this section deals with the architecture of the system

Section 4 & 5 Methodology & models: It gives detailing about the steps taken during the day development.

Section 6 results: it gives the outcome of turnover after application of machine listening models

Section 7 & 8 conclusion and future scope: It presents the conclusion from the implementation of the project and the scope of the project in future endeavours in IT industry.

II. LITERATURE SURVEY

The employee turnover is a systematic issue in most industries from past few years. As employees tend to leave the company when they find better opportunities incurs a loss to previous company in terms of money, time and resources and a benefit to its competitor. An employee, who leave the company, has to be replaced by their equivalent skills or better skills only. A lot of high indirect cost goes in finding new replacements and training and development programs of new recruits [1]. Meanwhile companies have to plan career development programs for the trained and skilled employees to retain them since they are the assets. This would lead to poor performance in competitive IT sector.

Itemized beneath are the research documents that we had considered:

1. Title of the work: Employee Turnover Prediction based on Survival Analysis Authors: Qianwen Zhu, Jiaxing Shang, Xinjun Cai, Linli Jiang, Feiyi Liu Publication Details: 2019 IEEE International Conference.
Description: The prediction of employee turnover is very crucial for operation of the company. The departure of key employees will bring great loss. This paper ignores the historical events of turn over and focuses on events and third perspective with algorithm named CoxRF. The method, it is based on survival analysis. The main results found that gender and economic factors of external environment has a great impact. The rate of turnover in females found to be greater than male in the same working time and choice has an effect in this. This could be different in different industry groups but, it is most predominantly seen in IT industry. This paper also summarizes the fact to include the characteristics of both the past turnover history and traits of his current job. It states that comprehensive collection of the historical events of turnover, the current job, the period in which the prediction to be considered. Time variant information need to be collected to make the predictions of turnover. 3
2. Title of the work: A Model Based Prediction of Desirable Applicants through Employee's Perception of Retention and Performance Authors: Eduardo B. Santiago Jr and Glenn Paul P. Gara Publication Details: 2018 IEEE International Conference.
Description: The recruitment of proper skilled employees is very important for organizations productivity and performance. Picking the right employee benefits company in the end. The usual tendency of an employee to quit the job is common and hence, it becomes the primary responsibility of the human resource. The manager to recruit only the desirable applicants who give confidence to continue in the same company for a long time. This paper makes use of 12 dimensions of retention, which may applied with Nevis classifier to help the decision-making in the hiring process.

III. PROPOSED WORK

The proposed system gives the total turnover result with profit and loss percentage of the company by comparing various machine-learning algorithms, which gives result that is more accurate. The overview of the project for better understanding of the work is depicted using use case diagram as shown in Figure 1. The system divided into 2 modules;

1. ADMIN MODULE: Login form made for security purpose. Admin can login by giving user id and password. It contains Add employee, employee details, add salary, view graph, view feedback and predict result, it also contains several graphs of different parameters. Prediction of employee turnover has calculated in predict result page. The feedback from the employees presented in the admin module, which helps in analyzing the turnover.
2. EMPLOYEE MODULE: Login form made for security purpose. My profile, add feedback, view graph and salary report these are the four options for user. User can view his/her profile, which contains personal info like

emp id, name, DOB, address, mobile no, designation and email id. In salary report, employee can check salary details with respect to date. An employee can give the feedback, which shows the reasons to leave the company, and it is stored in database that can be view by admin.



Figure 1.Usecase diagram

IV. METHODOLOGY

The implementation workflow is as follows: there is initial data collection and gathering from the online platform and other related documents, which is stored in the database. There is two different modules that is admin module and employee module. Admin enter parameters to predict turnover. The model building done with the help of machine learning algorithms that is Navie Bayes, random forest, and logistic regression. The result of implementation is octane from the model, which gives the highest accuracy. Data visualization provided with the help of graphs. On the other hand employee as a model to give feedback/reasons for leaving the company to the HR manager.

1. Dataset Description: We operate on dataset holding 15400 dossier of employee made up of 18 attributes as shown in the below Figure3. Data retrieved by simple query execution. It will allow machine-learning method to study from real information rather than unambiguous programming. This is repeated and timely conducted on significant samples, the forecast generated in the output will be precise.

- 1.Satisfaction level: it is the data taken by the employee towards the work culture of the company
- 2.Work accident: It is any misfortune or disaster occurring at the workplace.
- 3.Department: it is the different departments Like sales ,marketing ,engineering.
- 4.Salary: It is the expected salary level of the employee
- 5.Job role: It is the different rules provided by the company like human resource, technical, software engineer...
- 6.Age: It is the age of the employee ranging from 21 to 60
- 7.Gender: the gender is either male or female
- 8.Education: It is the highest qualification of the employee
- 9.Business location: it is the job location provided by the company to the employee.
- 10.Environmental satisfaction is basically the work culture of a company which is considered as one of the important parameters in predicting the turnover.
- 11.Average monthly hours: An Average monthly hours spent in the company by an employee.

12. Company spend time : Total number of hours spent by an employee in a company.
13. Number of projects: Total number of Projects which are undertaken or completed by an employee.
14. Satisfaction level: Satisfaction level is the job satisfaction level refers to an employee's overall positive attitude and feelings towards a job.

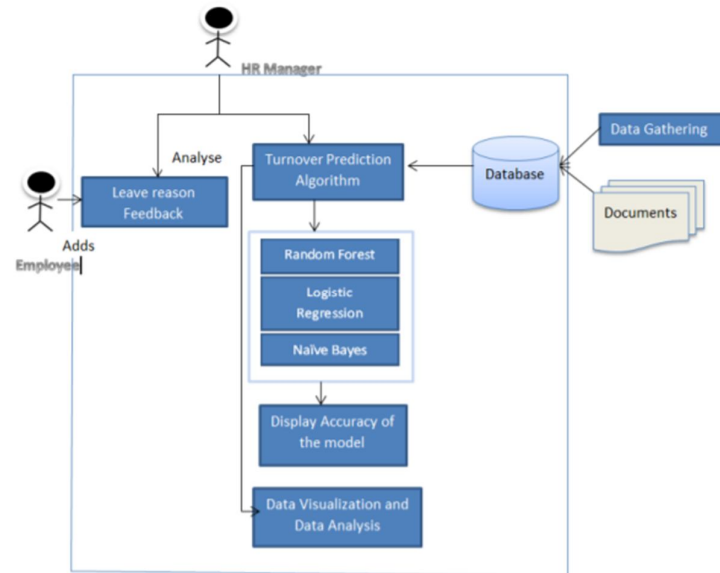


Figure 2. Architecture flow of the web application

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
satisfactio	last_evalu	number_gi	average_n	time_spen	Work_acc	left	sales	salary	age	Job_Role	Business_1	Gender	Over_Time	Distance	J Education	Environme
0.38	0.53	2	157	3	0	1	sales	low	30	HealthCan	Tumkur	Male	1	9	belowcoll	5
0.8	0.86	5	262	6	0	1	sales	medium	31	HumanRes	Banglore	Female	2	3	college	2
0.11	0.88	7	272	4	0	1	sales	medium	36	Technician	delhi	Female	3	6	belowcoll	3
0.72	0.87	5	223	5	0	1	sales	low	38	SoftwareD	mumbai	Male	5	7	BE	1
0.37	0.52	2	159	3	0	1	sales	low	40	SoftwareT	chennai	Male	3	1	phd	4
0.41	0.5	2	153	3	0	1	sales	low	42	ASE	pune	Male	5	3	phd	2
0.1	0.77	6	247	4	0	1	sales	low	27	SE	hyderabad	Female	4	8	BE	1
0.92	0.85	5	259	5	0	1	sales	low	25	TeamMan	mysore	Female	2	3	college	5
0.89	1	5	224	5	0	1	sales	low	31	SoftwareT	tiptur	Male	5	5	belowcoll	4
0.42	0.53	2	142	3	0	1	sales	low	27	SE	tumkur	Male	1	4	college	2
0.45	0.54	2	135	3	0	1	sales	low	27	HealthCan	banglore	Male	2	7	belowcoll	3
0.11	0.81	6	305	4	0	1	sales	low	53	HumanRes	mysore	Female	3	3	BE	5
0.84	0.92	4	234	5	0	1	sales	low	72	Technician	mangalore	Female	5	12	phd	0
0.41	0.55	2	148	3	0	1	sales	low	53	SoftwareD	bombay	Female	4	2	BE	1
0.36	0.56	2	137	3	0	1	sales	low	63	SoftwareT	koilkatta	Male	3	1	belowcoll	2
0.38	0.54	2	143	3	0	1	sales	low	30	ASE	delhi	Female	2	25	college	3
0.45	0.47	2	160	3	0	1	sales	low	64	SE	chennai	Female	4	3	belowcoll	4

Figure 3. The Dataset having 15400 dossier

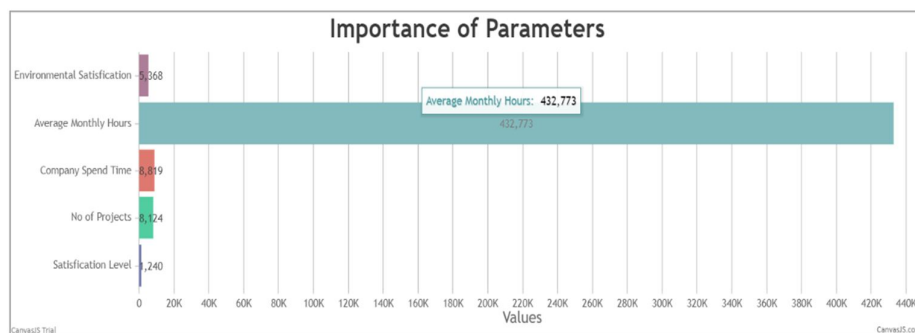


Figure 4 .Important parameters

2. *Data cleaning*: The less noteworthy attributes will be removed. The null and an identified numbers are recognised and removed, since it can adversely affect the right training of the model and give imprecise outcome. Missing data identified before proceeding with the implementation of algorithms.

3. *Model building*: The model developed with help of different machine learning techniques. The aim is to spot the better classifier for the analysis predicament. Hence, each classifier need to be trained with feature set and classification results have to be used for production. Naive bayes, random forest and logistic regression are the algorithms implemented. Training set (consisting of 1029 records) shall be used as already classified target and about 441 of the test sets are used to train the model. To assess the performance of a model, two self-governing evaluation tests are used:

- A. *Hold out*: It is very important to keep the same sharing of target variables in both test and training data set. Target is binary values with 84% no and 16% yes hence both the data sets have to keep similar amount after split.
- B. *Cross validation*: It is a technique needed in order to avoid the over fitting problems and shorten the model. The training set is divided into five parts for validation. In each iteration, a different part taken an average prediction error has taken to predict the average errors in key iterations.

V. MODELS

A. Naive Bayes Classification

Naive Bayes algorithm is used for classification. It is algorithm based on Bayes theorem where every pair of feature used for classification is self-sufficient of each other. The working procedure is as shown in the below steps:

In this system we gauge the prior probability of each given class labels in the training set. Next step is to find a chance probability with each trait of each class. The main traits considered are satisfaction level, number of projects handled, the work accident and the salary level of an employee. Then with the help of Bayes formula, we calculate posterior probability. Finally, the class having higher probability is the value for target variable of test set.

The Bayes' theorem formula is:

$$P(A / B) = \frac{P(B / A)P(A)}{P(B)}$$

Where,

$P(A|B)$ is Posterior probability: Probability of hypothesis A on the observed event B.

$P(B|A)$ is Likelihood probability: Probability of the evidence given that probability of a hypothesis is true.

$P(A)$ is Prior Probability: Probability of the hypothesis is before observing the evidence.

$P(B)$ is Marginal Probability: Probability of Evidence.

B. Random Forest

This is a supervised learning algorithm, implemented for both classification as well as regression. This algorithm used as hyper parameters to give better results. Here each tree is created from a different sample of rows and at each node, a different sample of feature is elected for splitting. Each tree makes its own personal prediction that are later for averaged to give a one result as a turnover value. We have considered the subset of variables at each step and referred to different models, therefore it will increase the amount of randomness to the training data. Decision obtained from citing different models will be more accurate.

It has the following advantages:

1. It takes minimum training time as compared to other algorithms.
2. It predicts output with high accuracy; even we have taken large dataset though it runs efficiently.
3. It can also maintain the results and gives high accuracy when a large proportion of data is missing.

C. Logistic regression

This algorithm uses the sigmoid logic function and it finds suitable hypothesis to predict the results from the input data. It construct the cost function to represent their between the predicted and actual result and finally minimises to find optimal results. Test result is obtain after fitting logistic regression. It will predict the output of a categorical dependent variable. The outcome must be a categorical value or discrete value. It can be either Yes or No .This type of classifier is good for maximising the entropy classification.

VI. RESULTS

In this work, we intend to assemble a model to forecast and tackle the trouble occurring in information industry. Here, we make use of classification tools for prediction and comparing the accuracies. Learning Algorithm i.e Naïve Bayes and logistic regression both are giving the accuracy of 91% and the random forest is giving accuracy of 93%. It disclose that random forest is a finer algorithm for this case in terms of notably higher accuracy.

Model No	Model Name	Accuracy
1	Naive bayes	91%
2	Random forest	93%
3	Logistic regression	91%

The demographics increase the ability of the manager to address the problem [4]. It helps to isolate the aspects, which lead to higher turnover and allows organizations to prioritize the retention of employees.

As shown in Figure 5, admin can view the turnover versus job role, which helps to analyze the turnover of particular job. Therefore, that admin can anticipate the maximum turnover of a job role.

After the prediction of total turnover of the company as shown in Figure 6. The profit and loss of an organization can be estimated. The profit & loss helps the managers to utilize the human resources in future that leads to profit for an organization.

The feedback was given by the employees can be viewed by the managers. It helps to know the major cause of attrition as shown in Figure 7.

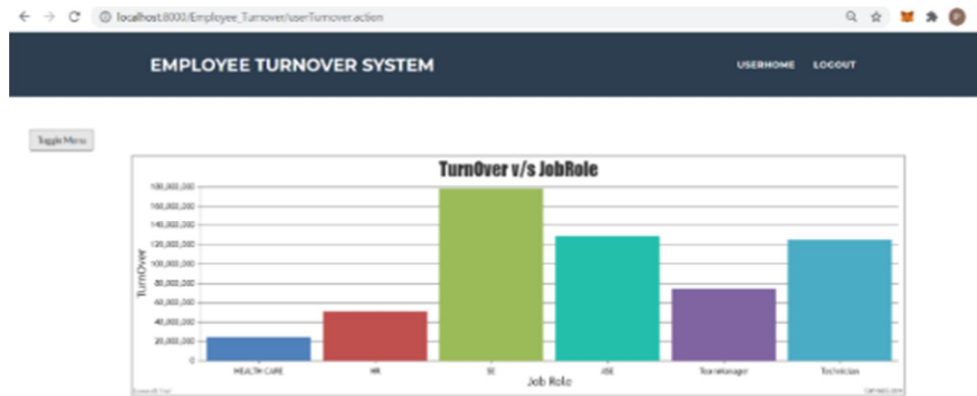


Figure 5. Turnover v/s job role

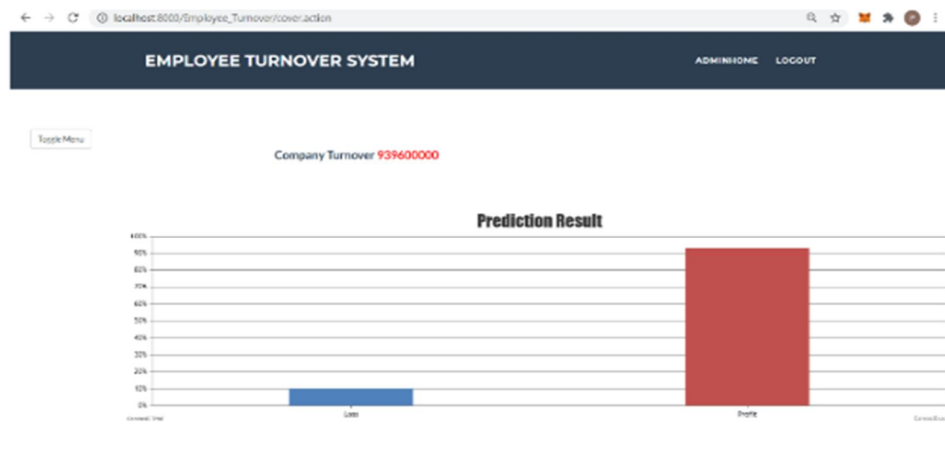


Figure 6.: Total company turnover graph

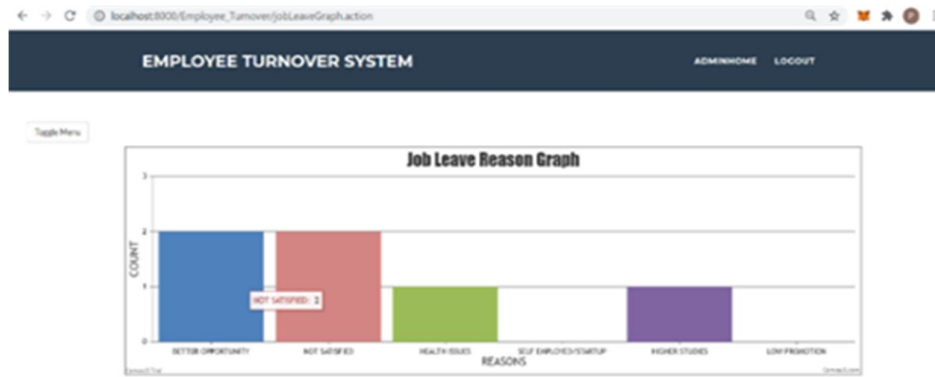


Figure 7. Job leave reason graph

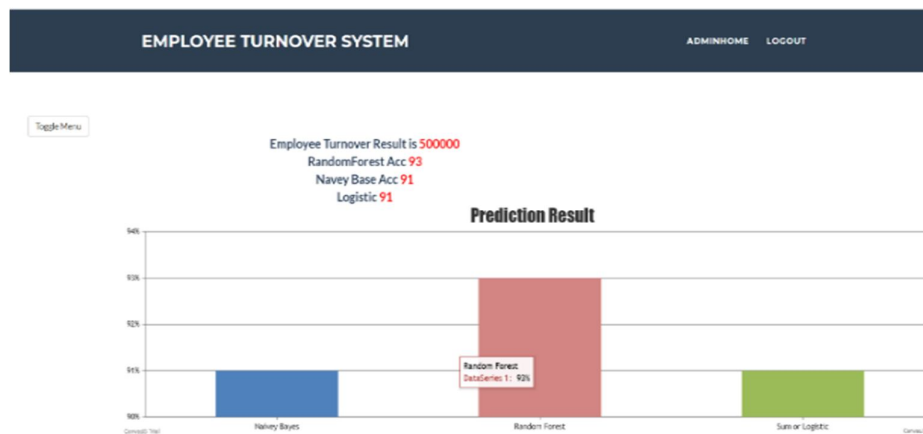


Figure 8. Prediction result graph with accuracy of each ML algorithm

VII. CONCLUSION

Employee turnover is one of the main crisis in the IT industry and this incident is occurring from the last decade. Here, we make use of classification tools using machine learning for prediction and comparing the accuracies. We intend to encourage organizations of different sectors to adopt it. The profile and various parameters, which define the employee, used in order to develop model. Few less important parameter such as the business travel, gender, and number of companies to employ work done are over looked in the process of study. The number of companies and employee works in represents his experience only. The model predicts to find whether an employee shall continue to work at the same place or desires to leave in order to find better opportunities.

Prediction of the model has an accuracy rate of about 70 to 93%, which also suggest that, it can be adopted by organizations in the country. The productivity and profitability of company highly depends on the skilled and talented employees of the organization. The turnover result makes organizations to prioritize the retention of employee and feedback gives a constructive criticism to improve the work culture. Hence, it becomes very important to maintain employee profile and monitor the employee turnover rate. With help of such predictive models, organizations can overcome the business challenges and achieve competitive success.

FUTURE SCOPE

This is first of its kind study, which is applying machine-learning classification tools in order to address the turnover problem of the industry. Human resource is an asset, which can affect the company in positive and negative ways. The only way to attract and retain high performing talent is to build growth into work. The results of employee turnover and the feedback acts as a token of exercise which helps to evaluate individual's interests and expectations and values which may or may not match the current position in the job. It helps to isolate the important aspects, which lead to such behaviors within company, and makes organizations to prioritize the

retention of employees. They track individual team member progress in the end to capture precise snapshot of health of company's culture and overall engagement-Which is ultimately a predictor of turnover. Human resource manager can adopt these models in order to analyze and widen approaches to improve the job fulfillment among the employees. It helps to focus on what really matters the most in the business to bring most profitability into the organization.

REFERENCES

- [1] Qianwen Zhu, Jiaxing Shang, XinjunCai, Linli Jiang, Feiyi Liu and BaohuaQiang, "CoxRF: Employee Turnover Prediction based on Survival Analysis" IEEE 2019
- [2] Eduardo B. Santiago Jr and Glenn Paul P. Gara, "A Model Based Prediction of Desirable Applicants through Employee's Perception of Retention and Performance" IEEE 2018
- [3] T.T. Weeramanthrie, C. N. Thilakumara, K.N.A.C. Wijesiri, N. I. Fernando, S. Thelijjagoda, A.Gamage, "A Web-Based Employee Turnover Analysis Tool for Effective Human Resource Management in Large-Scale Organizations" IEEE 2017
- [4] Rachna Jain and AnandNayyar, "Predicting Employee Attrition Machine Learning Approach" International Conference on System Modeling and Advancement in Research Trends, 23rd–24th November, 2018.
- [5] ZhenkuiJin, Li Zhen and JianjunShen, "Prediction Model of Employee Flow Based on Semi-Markov Chain and its Application Analysis" IEEE 5th Information Technology and Mechatronics Engineering Conference 2020 .
- [6] Human Resource Practices and Employee Turnover July 2020 Journal of Advanced Research in Dynamical and Control Systems
- [7] 19th International Scientific Conference; Economics and Management 2014, ICEM 2014, 23-25 April 2014, Riga, Latvia The impact of human resource management practices on employee turnover IvetaOzolins-Ozolins * a Riga Technical University, Latvia