2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362933

Predicting Employee Attrition along with Identifying High Risk Employees using Big Data and Machine Learning

Apurva Mhatre[1], Avantika Mahalingam[2], Mahadevan Narayanan[3], Akash Nair[4], Suyash Jaju[5]
Department of Computer Engineering
SIES Graduate School of Technology
Nerul-400706, India
apurva.mhatre16@siesgst.ac.in[1], avantika.mahalingam16@siesgst.ac.in[2], mahadevan.narayanan16@siesgst.ac.in[3],
akash.nair16@siesgst.ac.in[4], suyash.jaju16@siesgst.ac.in[5]

***Abstract*— "It takes a lot of time and energy to build a great employee and only a second to lose one." Employee turnover is a perennial challenge faced by all the major companies across the globe, performance of a company is directly proportional to the quality of employees retained by them. Whenever a good employee quits the organization it leads to financial losses, gaps in company's execution capability, re-recruiting costs and loss of productivity. The success of a company lies not only in impeding the attrition rate but also in retaining the right talent. According to NASSCOM, the global employee churn rate as of 2019 is 18-20 percent, which is what makes it necessary to alleviate the business risks associated with the turnover using statistical analysis. This research aims to foresee potential attrition (specifically in the B.P.O. sector) by mining turnover trends amongst employees and use supervised classification techniques to cluster out vulnerable employees."**

Keywords— Big data; turnover; data mining; patterns; supervised learning; machine learning; concept drift.

I. INTRODUCTION

Attrition refers to an employee leaving a company due to some reason. Attrition has a negative effect on both the organization as well as the individual. The organization suffers a very high financial cost; service quality may suffer as a result of which it may lose its customer-client base. As for the individual, he/she may have some financial difficulties, relocation costs, loss of social network. The first step in tackling attrition is identifying the drivers that cause this problem:

A. Lack of Employee Engagement

As per Harvard Business Review, most of the workers in the lower bands of the work force don't have a direct line of communication with their superior. Lack of engagements prevents trust building between co-workers and their superiors.

B. Monotonous Work Life

Due to prosaic and repetitious job profile, most of the workforce loses interest in doing their job. Also, no new challenges faced by the employees leads to lower working efficiency and reduced creative/critical thinking ability, indirectly stopping them from learning something new and bringing more value to the organization.

C. Pool of Opportunities

Due to the exponential growth of B.P.O. in the last decade, a lot of options are available for the employees to switch jobs. Studies show that people migrate to companies that give them a relatively higher pay than the market average.

D. Low career prospects

Due to lack of skill and creativity involved in the job the experience gained does not hold much weightage. Moreover, the opportunities to climb up the corporate latter are quite sparse.

In the absence of structured and robust analytics, businesses are having mixed successes in addressing these risks. This project aims to leverage technology to be able to provide proactive alerts to help them manage these risks better; we intend to do this by:

- Predicting potential attritions patterns in the form of employee groups, location, type of work, levels, etc.
- Identifying the ideal successful employee profile pattern through leveraging data on successful employees.

These analytics are not static; one-time exercise but they continuously learn based on new scenarios.

II. RELATED WORK

Many exhaustive studies have been made pertaining to Employee turnover in the literature and many effective data

ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE

mining problems have been attested. Although there are studies of voluntary as well as involuntary attrition, this study is mainly focused on voluntary attrition.

As per Kransdorff A. in his study, Voluntary attrition has been accelerating in the past decade and according to his study, employees tend to switch the respective organization every six years for their personal growth. Senior management has to be alerted precociously; else the organization will face serious repercussions when their best performers leave [1]. Johnson, Julie T., Rodger W. Griffeth, and Mitch Griffin have proved that functional turnover (i.e. good performers leave and bad performers stay) can actually help reduce optimal organizational performance and excessive turnover can cause detrimental effects to the productivity and yield of the firm. This mostly causes loss in the business and relationships which brings jeopardy for the actualization of the firm's objectives [2][3]. Also, a theory of dysfunctional turnover has been proposed by Abbasi, Sami M., and Kenneth W. Hollman. Dysfunctional turnover (i.e. good performers stay and bad performers leave) can damage the organizational structure of the company through lethargic implementation of new programs, less innovation and reduced productivity [4][3]. This can radically affect a firm's ability to burgeon in today's competitive economy. A study by Dolatabadi, Sepideh Hassankhani, and Farshid Keynia, throws light into the relationship between employee attrition and the cost involved with it being directly proportional. More turnover means more economical loss to the firm. He states that a churn as low as 5% could cost approximately cost 1.5 times the annual income of an employee [5].

Abbasi, Sami M., and Kenneth W. Hollman have also identified and highlighted few vital reasons for voluntary employee turnover. They include hiring patterns, managerial style, lack of acknowledgement of work, lack of competitive rewarding systems and toxic workplace environments. Further this study concluded that style of managerial work, workplace environments, and good hiring practices can be a stimulus for attrition in an organization [4]. A meta-analytic study done by Cotton, J.L. and Tuttle, J.M., provides evidence of voluntary turnover being age, tenure, pay, overall job satisfaction, and employee's perceptions of fairness [6]. Allen and Griffeth and D. Liu, T. R. Mitchell, T. W. Lee, B. C. Holtom, and T. R. Hinkin, display similar research findings that are focused on salary levels, working conditions, job satisfaction, supervision under managers, advancement, growth potential in the company etc. [7][8].

Organizations in the BPO sector have devised technological efforts to combat attrition. One such example is an information management company wherein combating attrition is a strategic imperative measure in order to continue providing the so-called 'best in-class' services at a comprehensive rate to its clients [9]. Their deployed framework, called the "Early Warning System" (EWS) with RAG (Red, Amber, Green) indicators, has been successful in limiting the attrition rate to a significant percentage. M. Singh et. al, with the IBM Watson team has brought out a framework which realizes the reasons of employee attrition and identifies potential attrition. The model works on the basis of cost of attrition and comparing the difference between Expected Cost of Attrition Before the retention period (EACB) and Expected Cost of Attrition After the retention period (EACA) [10].

The main and vital objective of this project work is to provide comprehensive description along with demonstration and assessment of machine learning approaches towards detecting attrition of a reputable revenue cycle management company in the health sector by giving a robust technical solution by foreseeing the probable reasons for it. Keeping emphasis on voluntary turnover, the results thus can be then used by the HR department to look into the major reasons leading to attrition and thus mitigate them.

TABLE I. SUMMARY OF PREVIOUS WORK

Problem Studied	Data Mining Techniques used	Recommended model	Dataset used
Employee Attrition Prediction [11]	KNN, Naive Bayes, MLP Classifier, Logistic Regression	KNN (Accuracy: 94.32%)	Kaggle sample dataset
Employee churn prediction using SVM following the feasibility study of 4 other algorithms [12]	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees and Random Forests	SVM (Accuracy: 80%)	Particular Sample dataset from HR department of three companies in India
Demonstration of XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy [13]	Naïve Bayes, Support Vector Machines, Logistic Regression, Linear discriminant analysis, Random Forests, KNN, XGBoost	XGBoost (Accuracy: 88%)	Dataset of a certain level of employees in a particular leadership team of a global retailer
Study of turnover prediction models: Logit and Probit models [14]	Logistic regression model (logit), probability regression model (probit)	Logistic regression model (logit) (Accuracy: 90.5%)	Custom dataset drawn from a motor marketing company in Taiwan
Comparison of various decision trees for the analysis of the turnover of employees [15]	ID3 Decision tree, CART Decision tree	CART Decision Tree (Accuracy: 90%)	Kaggle sample dataset
Behavioral comparison of Random Forest and Naive Bayes [16]	Naive Bayes and Random Forest	Naive Bayes (Accuracy: Up to 85%)	Sample dataset of sales agents

TABLE I. above briefly identifies and documents the literature review findings. Subsequent sections include the studies of different papers indicating the inadequacy of certain models and their recommended models which demonstrate successful results.

III. RESEARCH METHODOLOGY

In the present project work, six supervised machine learning algorithms are evaluated for a dataset obtained from a revenue cycle management company in the health sector which comprises data of about 12,000 employees with over 24 features across various working bands, designations and departments. In the light of the literature survey that is conducted, the dataset is split into two parts i.e. lower bands (A1, A2, A3) and higher bands (Manager, P1, P2). Since both sets of bands were found to behave differently, predictions are done separately and confined them into one module in the later stage.

The proposed system, (EWS) has the ability to provide predictions of two classes of bands along with the indication of high risk and low risk employees done by deep analytical study abutting the business perspective of the company and also providing the graphical visualization for the obvious patterns. This section includes in-depth description of the various steps involved in developing the EWS for the organization.

A. Data Cleaning

The data cleaning/pre-processing is the first and vital step that is taken towards building a predictive model. This helps us identify the features that accounts for the labelling of the classes that are to be predicted.

The first important step in data cleaning is to comprehend the structure of the data. The dataset consisted details of roughly 12000 employees across 24 columnar attributes. Below is a snippet of selected attributes from the dataset.

TABLE II. STRUCTURE OF DATASET

Feature	Data Type	Description
Employee Code	Int64	Numerical
Salary	Int64	Numerical
Interim Manager	Object	Text/String
Rating	Int64	Numerical
Exit Interview - Comments	Object	Text

B. Standardizing the data:

The data provided was captured using a native system, ARMS. Most of the data was entered without proper validation owing to a lot of typos, missing fields and inconsistent values. Discrepancies were removed after converting the text format to UTF-8, standardizing the nomenclatures, rectifying any ambiguous entries, removing faulty data and filling the missing values.

C. Feature Engineering & Exploratory data analysis

Irrelevant or ambiguous fields can negatively impact the model. In this stage we identify the features (attributes) which contribute most to the prediction variable.

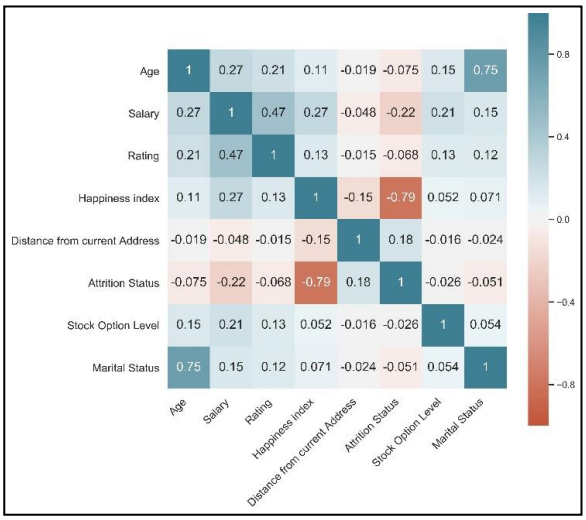


Fig. 1. Correlation Heatmap*

Some Observations:

- Happiness Index α Rating
- Age α Salary
- Rating α Salary

Note: The correlation between the attributes is underrated as these variables were standardized into integer ranges [1-5] by the data provider to protect privacy.

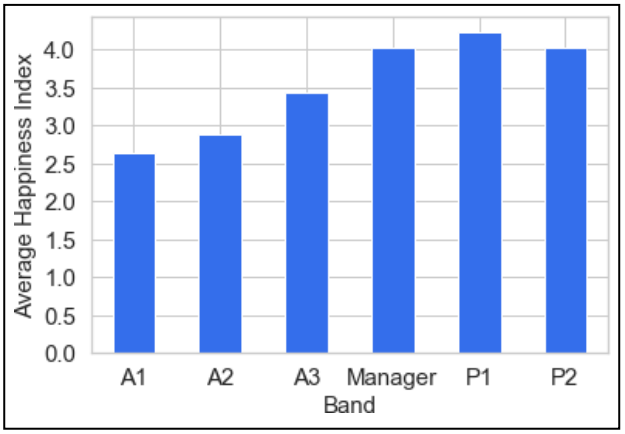


Fig. 2. Happiness Index (Average) across various bands

From Fig. 2., it can be concluded that the employees of the higher bands have a better work life balance in general than the lower bands.

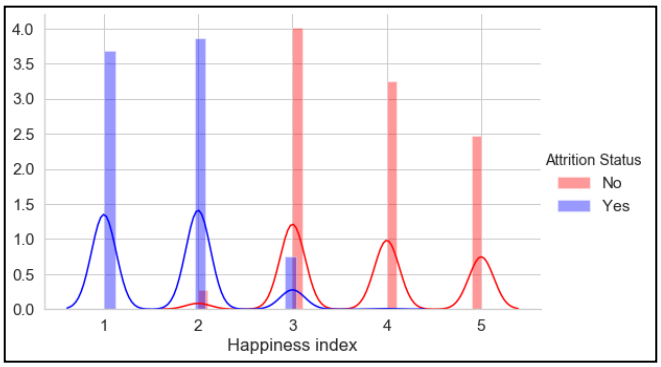


Fig. 3. Probability Distribution of Happiness Index across Attrition Status

Fig. 3. Helps us identify that majority of the employees attrite if they have a happiness index lower than 3.

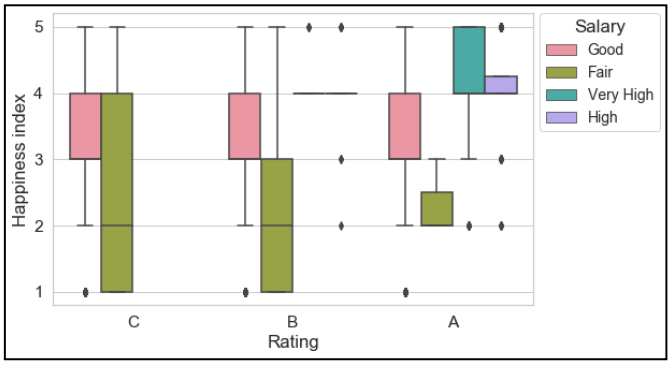


Fig. 4. Box plot (Happiness Index vs Rating vs Salary)

The Box plot in Fig. 4. Helps us understand that employees having higher rating have a better PayScale, also people with higher rating have a better Happiness Index than the lower rated employees.

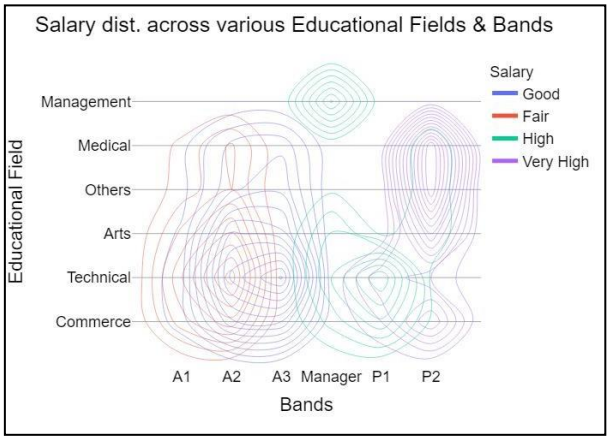


Fig. 5. Contour plot of Salary vs Designation vs Bands

Fig. 5. shows the Contour plot highlights that people having a technical and commerce background have a higher PayScale.

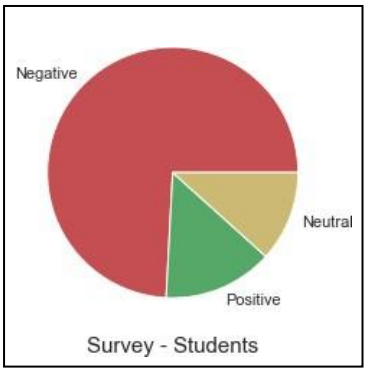


Fig. 6. Exit Interview Comments

Fig. 6. shows results exit interview comments which were analyzed using Natural Language processing and the following results were postulated.

Feature Engineering allows us to select the most relevant fields pertaining to the output variable. In a nutshell, correlation matrix gives us an overall picture on how different attributes affect each other. Using these results, we can remove redundant attributes. Further, data analysis helps us to mine trends/patterns in the data which may not be directly visible. These findings are used to develop a prediction model for the output variable.

IV. MODELLING

This section details and outlines the theory and logical explanation behind various Supervised algorithms that are compared from the list of recommended models that has been studied extensively in the literature review. With the class labels being identified as 'Yes' or 'No' for attrited and nonattrited employees respectively we use various classifier algorithms for the prediction.

A. Data pre-processing

Most machine learning algorithms require numerical input and output variables. One hot encoding is used to convert categorical data to integer data. This not only simplifies computation but enables the model to learn at an improved rate.

B. Model Validation

The organization's dataset has been split 70:30 into training and testing sets. The training of the model has been done in its optimal configurations. The picked-up pattern from the trained model from each algorithm was then used to predict on the test set.

Model validation has been done on the basis of the technique of comparing the accuracies according to the area under the receiver operating characteristics (ROC-AUC) of each model. The discrimination ability of a binary test which are used for predictive purposes when applied to a scored dataset is commonly judged by ROC Analysis [17]. Lessmann, Stefan, and Stefan Voß, state that the AUC characteristic is a general measure of comparing the 'predictiveness' assessment of various models [18]. Furthermore, since AUC measures the probability of a classifier that indicatively ranks an arbitrarily chosen positive instance of these data points higher than an arbitrarily chosen negative one, which is equipollent to the

Wilcoxon test of ranks than other alternative indicators such as error rate [19].

Additionally, model run-time, accuracy, F1-score and recall are also used to compare the performance of the classifiers. Determining models on these two measures are important as they construct a perspective based on the suitability for implementing the algorithms for real-life business quandaries, essential for scalability and performance.

C. System Environment Specification

All the models except XGBoost are used from the scikit package in the latest Python configuration. The XGBoost model was used from the XGBoost standalone package. The algorithms were run on a 16GB Windows 10 version with a GPU of NVIDIA 1070.

Below is a detailed explanation of all the supervised algorithms used to develop a model.

Logistic Regression:

Cox, David R. proposed that Logistic Regression is a customary classification algorithm involving linear discriminants [20]. It is a branch of regression used for predicting binary variables or categorically dependent variables based on probability which creates a linear boundary for separation [21]. To avoid over-fitting, is it used with L1-norm or L2 norm. A regularized L2-norm has been applied in this study.

The model from the Sklearn module has been used for the training and testing of the dataset. The Logistic Regression in the Sklearn package implements regularized logistic regression

Decision Tree:

As proposed by Morgan and Sonquist, Decision trees are nothing but a form of classification or regression models which builds a tree-like structure for prediction [22]. It is usually not stable with high variance in the specified datapoints [23] and large effects can be observed on the tree-like structure even if there are small-variations in the input data [24].

In this study, the CART Decision tree has been implemented because it constructs the tree based on a criterion that applies numerical splitting which is then recursively applied to the data. Scikit uses an optimized version of the CART Decision tree and produces optimal results for any dataset.

KNN:

K-Nearest Neighbors is a type of nonparametric supervised algorithm used for classification problems. The conception behind KNN is to identify the K data points in the trained data that are closest to the new instance data points and classify these data points by a majority vote of its K neighbors. Various other distance measures like Euclidean distance, Manhattan distance, Minkowski distance etc. are used in KNN [25].

The scikit module KNN package uses the default Minkowski metric system.

Naïve Bayes:

Naïve Bayes algorithm uses Bayes algorithm for probabilistic approach. It performs with a base assumption that all variables are independent of each other conditionally. Naïve Bayes classifiers first learn on the basis of joint probability distribution of their inputs by utilizing the assumption. Then for a given input, by means of Bayes Theorem the methods produce an output by calculating the maximum posterior probability [25].

In this study we have implemented Gaussian Naïve Bayes classifier in the scikit module for optimized performance.

SVM:

Support vector machine incorporates the principles of statistical learning theory [26]. It can solve linear as well as nonlinear binary classification problems. The theory behind SVM is that for achieving class separation it constructs a hyper-plane or set of many hyper-planes in a higher dimensional space which are divided into two classes wherein, this hyper-plane maximizes the said geometric distance to the nearest data points. These are called as support vectors [21]. For this reason, it is called maximum margin classifier [13].

This study is implemented using SVM from the Sklearn module with RBF as the kernel. Since this dataset yields a nonlinear boundary, RBF creates nonlinear combinations of all the features to uplift the said data points onto a higher dimensional space where a decision boundary that is linear is used to separate the classes [27].

XGBoost:

Chen, T., Guestrin, C. introduced Extreme Gradient Boosting which is an end-to-end tree boosting method. Gradient boosting trees involve fitting a succession of weak learners in the model on altered data. The predictions from each iteration are then integrated through a weighted majority sum to give the final prediction. Each iteration consists of assigning higher weights to the examples that went through misclassification in the previous step. As the iterations progress, the data points that are difficult to predict receive more concentration from the model [28].

When compared to other gradient boosted methods, XGBoost uses a better than usual regularized-model formalization to control issues like over-fitting. It has a more scalable and accurate implementation which yields a better prediction and much faster computational run-times.

XGBoost package is a separate module which is used for the prediction.

D. Results

The quality results from all the classifiers correctly justify that the chosen features are the main causes for voluntary attrition in this company. Since the dataset is of a workforce in the company distributed across various bands and processes, it's intuitive to assume that an improved rule-based method or a tree-based model that is coherent in terms of scalability and accuracy will be best suited for the EWS system. It can be gauged from Table 2 that XGBoost out performs all the other models in all the aspects considerably.

TABLE III. CLASSIFICATION OF EMPLOYEE DATA

Algorithm	AUC	Accuracy	F1-Score	Recall	Run-time (seconds)
XGBoost	**0.95**	**96%**	**0.96**	**0.95**	**1.76**
Logistic Regression	0.71	86%	0.62	0.59	0.23
Decision Tree	0.71	78%	0.61	0.61	0.030
KNN	0.56	74%	0.58	0.56	0.098
Naive Bayes	0.67	62%	0.56	0.67	0.027
SVM	0.53	85%	0.51	0.53	11.79

E. Risk Analysis

The area under the receiver operating characteristic curve (AUC), measures the discrimination component but not the calibration of the predictive rationality. The AUC actually does not capture the gap between the predictions done by the model and the actual observed risk [29]. So, we further segregated the churned employees into three categories – high, optimal and low risk. This was achieved by incorporating various parameters like no. of years in the company, salary, designation. Using this we were able to compute the cost to the organization.

$$\text{Cost per employee} = f(\textit{Salary, experience, designation, rating}) \quad (1)$$

Setting up a threshold mark based on the patterns prevalent in the firm, we were able to cluster out the employees into the overmentioned groups.

F. Early Warning System

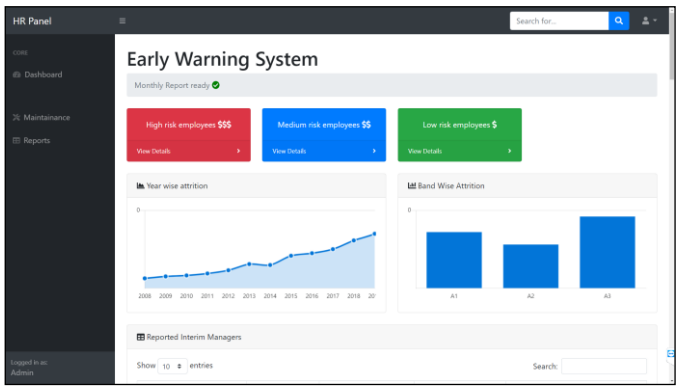


Fig. 7. EWS- Dashboard

Early Warning System (EWS) is a unique type of node-based web application which acts as a centralized system in an organization for segregating high, low risk employees and prediction of the potentially attriting employees. This system is handled by the HR team to cluster out the employees and measure the range of risk monitoring on the basis of the live data.

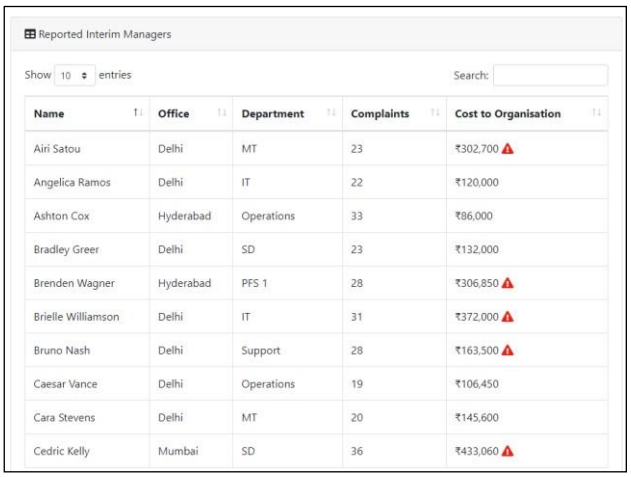
With reference to the patterns in the attrition data acquired in the previous month, EWS predicts and alerts the HR team about the employees who could potentially leave in the following month. EWS system generates a short report of all the vulnerable employee codes with a graphical analysis of the overall features. The data of the employees is indicated along with the potential reason for the departure from the organization.

Predicted Employees with Risk segregation

Employee Name	Employee Code	Band	Rating	H. Index	Salary
Tiger Nixon	610068	A2	C	2	₹200,000 p.a.
Garrett Winters	320078	A2	C	1	₹170,000 p.a.
Ashton Cox	320086	A1	B	3	₹135,000 p.a.
Cedric Kelly	810034	A2	C	2	₹200,000 p.a.
Airi Satou	210045	A3	B	1	₹300,000 p.a.
Brielle Williamson	310056	A3	A	2	₹300,000 p.a.
Herrod Chandler	410045	A1	C	2	₹130,000 p.a.
Rhona Davidson	210078	A2	C	1	₹320,000 p.a.
Colleen Hurst	510066	A2	B	2	₹200,000 p.a.

Fig. 8. Clustering of employees in different risk categories

EWS does this by extracting and analyzing features from a set of employees having similar patterns. To prepare the company for the upcoming detrimental economic effects owing to the probable attrition, EWS also estimates the total turnover cost for that particular month. During the testing phase on the live data acquired from the organization for the month of January, it was observed that the model attained an accuracy of 80% which is optimal considering the dynamic nature of the data. Since the model was trained on cleaned and standardized historical data, we recognize certain changes in trends over the time when tested with a new set of data after a certain significant interlude. Due to these variations in patterns; when the accuracy considerably decreases, EWS is built to handle this problem of concept drift. Concept drift is a term introduced to realize the problem of learning in dynamic environments [30]. When accuracy drops below a set threshold, it obtains the new pattern from the data, relabels new data points and retrains the model incorporating these patterns. The risk analysis is implemented in the EWS by the previously postulated mathematical formula. This segregates the present employees in the organization into three parts using RBG (Red, Blue and Green) indicators with Red being the most vulnerable cluster and Green indicating the safe cluster. The key here is to retain the high-risk employees that included talented people with high experience. These set of employees will be difficult to replace, thereby the cost per employee will be higher. In this way, the organization can devise suitable measures for the vulnerable cluster in the present month itself.



Reported Interim Managers

Show 10 entries Search:

Name	Office	Department	Complaints	Cost to Organisation
Airi Satou	Delhi	MT	23	₹302,700
Angelica Ramos	Delhi	IT	22	₹120,000
Ashton Cox	Hyderabad	Operations	33	₹86,000
Bradley Greer	Delhi	SD	23	₹132,000
Brenden Wagner	Hyderabad	PFS 1	28	₹306,850
Brielle Williamson	Delhi	IT	31	₹372,000
Bruno Nash	Delhi	Support	28	₹163,500
Caesar Vance	Delhi	Operations	19	₹106,450
Cara Stevens	Delhi	MT	20	₹145,600
Cedric Kelly	Mumbai	SD	36	₹433,060

Fig. 9. Reported Interim Managers

Every month the system captures the data pertaining to the interim managers and computes the total number of employees leaving which were formerly working under him/her. Based on the above data, cost to organization is calculated, if the cost exceeds a certain threshold a flag is raised which aids HR to identify managers who contribute to a higher cost to the firm.

In a nutshell, rather than just predicting the potentially vulnerable employees this dashboard also segregates them in different risk categories which enables the HR to go a step further in retaining employees by preventing the more talented employees from attriting. Furthermore, the python backend algorithms fetch data from the employee database and generate real time interactive graphs. Also, a monthly report is generated which helps to compare the performance based on attrition cost.

V. Conclusions & Results

By judging the data, we concluded that Salary, Rating and Happiness Index are closely corelated with each other. We performed an external survey where we contacted 200 attrited employees (refer Fig. 6). The survey mainly focused on obtaining the reason for employees to leave the organization. After performing sentimental analysis on the comments, we found out there were issues with the upper management. The results highlighted the fact that there was lack of direct communication between the workers of lower bands and upper management. Moreover, the upper management's evaluations while giving ratings and salary incentives were not up to the mark. We later web scrapped all the employee reviews from Glassdoor, Naukri.com, etc. A thorough sentimental analysis highlighted the same management issue.

The developed system was tested for a period of about 45 days between January and February, EWS predicted an attrition of around 45%. Using the risk segregated clusters given by the algorithm, the HR team was able to retain the skilled employees thereby preventing loss of talent and experience. The overall attrition thereby reduced to around 30%.

Churning of talented employees have direct and indirect impacts across a company. Most companies today have understood the importance of reducing attrition and they do so by leveraging captured employee data to find patterns. However, these methods do not always assure success especially in organizations with a maturity of high level. However, one domain where mature organizations can converge their efforts is minimizing the impact of attrition. Assuming a balanced rate of attrition that cannot be reduced, organizations can try to assess the direct and indirect results of attrition, and turning up opportunities for minimization. Processes can then be put into place to exacerbate these impacts of aptitude loss, and ascertain that transition periods in the organization transpire smoothly and truncating dependencies on critical personnel.

Acknowledgment

We are grateful to the concerned organization to provide us with the dataset and for complete support and mentorship in pursuing this research.

References

[1] Kransdorff, Arnold. "Succession planning in a fast-changing world." *Management Decision* (1996).

[2] Johnson, Julie T., Rodger W. Griffeth, and Mitch Griffin. "Factors discriminating functional and dysfunctional salesforce turnover." *Journal of business & industrial marketing* (2000).

[3] M. Stoval and N. Bontis, "Voluntary turnover: Knowledge management – Friend or foe?", Journal of Intellectual Capital, 3(3), 303-322, 2002.

[4] Abbasi, Sami M., and Kenneth W. Hollman. "Turnover: The real bottom line." *Public personnel management* 29.3 (2000): 333-342.

[5] Dolatabadi, Sepideh Hassankhani, and Farshid Keynia. "Designing of customer and employee churn prediction model based on data mining method and neural predictor." *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2017.

[6] Cotton, J.L. and Tuttle, J.M., 1986. "Employee turnover: A metaanalysis and review with implications for research" Academy of management review, pp.55-70.

[7] D. G. Allen and R. W. Griffeth, "Test of a mediated performance – Turnover relationship highlighting the moderating roles of visibility and reward contingency", Journal of Applied Psychology, 86(5), 1014-1021, 2001.

[8] D. Liu, T. R. Mitchell, T. W. Lee, B. C. Holtom, and T. R. Hinkin, "When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover", Academy of Management Journal, 55(6), 1360-1380, 2012.

[9] Sachdeva, Sujata. *Times of India Online,* 27 Nov. 2007, www.timesofindia.indiatimes.com. Accessed 6 May 2020.

[10] Singh, Moninder, et al. "An analytics approach for proactively combating voluntary attrition of employees." *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, 2012.

[11] Yedida, R., Reddy, R., Vahi, R., Jana, R., GV, A., & Kulkarni, D. (2018). Employee Attrition Prediction. *arXiv preprint arXiv:1806.10480*.

[12] Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999-2006.

[13] Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, *4*(5), C5.

[14] W. C. Hong, S. Y. Wei, and Y. F. Chen, "A comparative test of two employee turnover prediction models", International Journal of Management, 24(4), 808, 2007.

[15] Gao, Ying. "using decision tree to analyze the turnover of employees." (2017).

[16] Mauricio A. Valle & Gonzalo A. Ruz (2015) Turnover Prediction in a Call Center: Behavioral Evidence of Loss Aversion using Random Forest and Naïve Bayes Algorithms, Applied Artificial Intelligence, 29:9, 923-942, *DOI: 10.1080/08839513.2015.1082282.*

[17] Metz, Charles E. "Basic principles of ROC analysis." *Seminars in nuclear medicine*. Vol. 8. No. 3. WB Saunders, 1978.

[18] Lessmann, Stefan, and Stefan Voß. "A reference model for customer-centric data mining with support vector machines." *European Journal of Operational Research* 199.2 (2009): 520-530.

[19] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters 27 (8), 861–874, 2006.

[20] Cox, David R. "The regression analysis of binary sequences." *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958): 215-232.

[21] Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

[22] Morgan, J.N., Sonquist, J.A.: Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415–434 (1963)

[23] Efron, B.S., Hastie, T.: Computer Age Statistical Inference. Cambridge University Press, Cambridge (2016)

[24] Géron, A.: Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media (2017)

[25] Zhao, Yue, et al. "Employee turnover prediction with machine learning: A reliable approach." *Proceedings of SAI intelligent systems conference*. Springer, Cham, 2018.

[26] Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. 20, 273–297 (1995)

[27]"How do I select SVM kernels?," Dr. Sebastian Raschka, May-2020. [Online]. Available:https://sebastianraschka.com/faq/docs/select_svm_kernel.html. [Accessed: 14-May-2020].

[28] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, ACM (2016).

[29] Singh, Jay P. "Predictive validity performance indicators in violence risk assessment: A methodological primer." Behavioral Sciences & the Law 31.1 (2013): 8-22.

[30] I. Žliobaite, "Learning under concept drift: an overview," Computing Research Repository by Cornell library, pp. –1–1, 2010.