A System for Analysis and Remediation of Attrition

Neil Brockett
Eaton
Dublin, Ireland
n.brockett@gmail.com

Catriona Clarke
Eaton
Dublin, Ireland
catrionaclarke@eaton.com

Michele Berlingerio
Eaton
Dublin, Ireland
michele.berlingerio@gmail.com

Sourav Dutta
Eaton
Dublin, Ireland
souravdutta@eaton.com

***Abstract*—With the increasing impetus on globalization, workplace modernization, and employee welfare, modern organizations are focusing more and more resources in developing their pool of human capital. Employee attrition poses a major challenge for organizations – be it in terms of operational cost or loss of talent. We present *CLARA* (CLustering for Analysis and Remedial of Attrition), a deployed end-to-end system applying descriptive, predictive, and prescriptive analytics, providing as output *actionable remedial actions* to be used by HR departments to improve employee retention. We propose a coupling between clustering and frequent pattern based scoring measure to identify candidate employees at a high risk of attrition, and subsequently recommend suggestive actions to improve their retention. Using publicly available IBM human resource (HR) dataset, we show that *CLARA* demonstrates comparable performance (compared to state-of-the-art techniques) in identifying such "high risk" employees. We further validate the quality of the recommendations provided by the framework in reducing the overall rate of human capital loss, and discuss the real-life implementation of the framework within our organization. *CLARA* attains up to 65% precision, on IBM HR dataset, in predicting employee attrition and, thanks to the remedial actions, up to 22.5% reduction in the predictive attrition score for the top-5 identified employees.**

I. Introduction

Employee turnover or *attrition* currently poses a significant challenge for enterprises and forms an active area of research in HR analytics [1]. Attrition has been identified as one of the major costs incurred by an organization [2] and also a negative factor for workplace performance [3]. As of 2012, a separated employee might cost an organization roughly 16% to 213% of the employee's annual salary, based on the job position [4], in terms of direct costs relating to advertising, replacement and onboarding time.

We focus on preventing *voluntary attrition* (i.e., resignation), to help reduce the cost incurred by the organization and provide efficiency in budgeting and recruitment planning.

Problem Statement: Specifically, this paper aims at an Artificial Intelligence answer to the question "`How do we prevent people from leaving?`", and goes beyond descriptive (why do employees leave) and predictive (which employees are at risk of leaving) analytics into the space of *prescriptive analytics*: "`What should be done to retain people?`". We consider each employee to be represented by an associated feature vector based on the employee characteristics like age, gender, highest education level, salary band, salary, length of service, length of previous experience, team and time since last promotion to name a few.

Our aim is then to identify employees at a high risk of attrition, and, for each employee compute the *least disruptive* remedial action that would reduce the risk. Note that the trivial solutions of just promoting every employee, or increase all salaries, are neither valid nor practical. Promoting an employee too early may propagate the risk of attrition to the other team members, who may become upset or lose confidence in the company's leadership. Similarly, increasing salaries for all employees is typically not feasible due to the budgetary constraints of an organization.

Formally, given a set of employees $\mathcal{E}$, a set of employee features f (like salary band and team size), and a number k of employees to retain (considered as the "budget"), we define,

Definition 1. *The* ***attrition prevention*** *problem involves finding the* least disruptive *set of remedial actions on f maximizing the retention of k active employees with the highest risk of attrition.*

Note that more complex definitions of budget could be considered and that a full-fledged optimization problem could be modeled, if provided with more data (in terms of employee features) incorporating further constraints like total yearly monetary cost of salary modifications, minimum and maximum team size; desired diversity in a team, and so on. We consider taking the above into account for the next release version of our system.

In the literature, several techniques have been proposed to analyze organizational trends [5], [6], and characteristics of attrition [7], [8], leading to the development of enterprise HR analytics tools like IBM HR Analytics [1] and SAP Workforce Analytics [9].

In this work, we aim at filling the gap of state-of-the-art approaches by proposing *CLARA* (CLustering for Analysis and Remedial of Attrition), a novel *end-to-end attrition analytics* system for not only predicting and identifying employees at a high risk of voluntary attrition, but also providing possible personalized remedial actions that might help retain them. CLARA uses *clustering* and *frequent pattern mining* to detect attrition trends based on employee features (like salary band, experience, etc.). It then identifies employees that might be more prone to voluntary attrition and subsequently recom-

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

mends remedial steps, based on an aggregated *predictive scoring* measure, that might help in retaining employees – thereby reducing the overall attrition rate within an organization.

II. System Architecture

We next briefly describe the working of CLARA. Figure 1 provides a pictorial overview of the system architecture, hinging on 6 modules.

1) Data Importer: Employee data, collected during the onboarding process or voluntarily disclosed by an individual, typically resides in an enterprise's encrypted data store. The data importer module in CLARA provides an interaction layer between the proposed framework and the raw storage. This layer is particularly dedicated to protect the security and sensitivity aspects of the raw information, so that unauthorized access or leakage can be prevented. To this end, the data importer module performs the following steps: (i) reads the raw data from storage, (ii) decrypts the data, (iii) removes sensitive information pertaining to employees, and (iv) anonymizes employee details with random but unique identifiers. Additionally, data cleaning involving duplicate removal and handling of corrupt/missing information is also performed. Note that the input data is historical, and contains employee records of current employees (hereafter "active") as well as of employees who have separated (hereafter "leaver") from the organization.

2) Feature Engineering: Here, the different *categorical features* are identified and are represented appropriately for uniformity and computational ease in the subsequent modules. Further, *continuous features* depicting a range of real values are also suitably categorized. CLARA then performs an analysis of the employee features across the entire dataset to evaluate the "goodness" of each of the obtained features based on its discriminative power for partitioning the data into (dis-)similar groups. Specifically, we use the *Pearson correlation coefficient* to identify and retain attributes, that are un-correlated and demonstrate a high variance, for extracting meaningful patterns to predict individual employee attrition chances. Note that other feature selection techniques, such as Principal Component Analysis (PCA), can easily be adopted in this module, making the system dynamic to several different application domains and requirements.

3) Predictive attrition scoring: This algorithmic module implements a novel core methodology to assess and predict the probability of voluntary attrition for an active employee. To this end, for each active employee, CLARA utilizes an aggregated scoring function based on (a) *clustering* – giving the employee's proximity to other employees, and (b) *frequent pattern mining* – giving the similarity between any employee and the traits of a typical leaver. The score is then used to predict the chances of attrition for each individual, and recommend remedial steps to improve retention.

The two major components in this module are as follows.

A. Employee Clustering – Based on the employee feature values, each employee record (active or leaver) is represented as a high-dimensional vector, and CLARA performs clustering (on employee data) to extract clusters or employee groups demonstrating a high degree of feature similarity. We use the k-means clustering algorithm based on the Euclidean distance (L_2 norm) to obtain the cluster groups. Alternatively, the k-modes [10] algorithm could have been used to deal with categorical features, but in our experiments k-modes did not provide any significant further improvements. For each active employee, a clustering based score is computed by considering the total Euclidean distance between its feature vector and the vectors of other employees assigned to the same cluster. Observe, active employees placed in a cluster with a large number of "leavers" exhibit a high clustering score – demonstrating a larger tendency towards attrition.

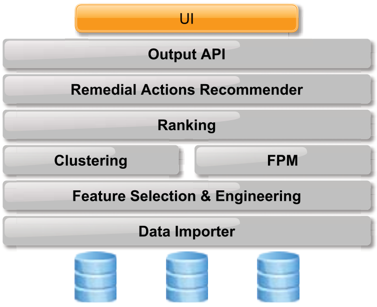


Fig. 1. Modular System Architecture of CLARA

B. Frequent Feature Pattern Mining – As a second predictive cue for CLARA, we analyze the frequent patterns in the feature vectors present among employees that have already left the organization. Intuitively, an active employee having a high similarity to such extracted frequent patterns are at a higher risk of attrition – and should thus be captured by this measure. CLARA next extracts frequent patterns of employee features (for active and leaver) separately using the Eclat's algorithm [11], with the support threshold for the pattern finding algorithm set at 5%, and the minimum itemset (pattern) size considered to be 3.

The commonality of frequent patterns between the active and leaver employees provides an important connecting factor in predicting active employee attrition. Hence, we consider the frequent feature patterns that are present in both the classes as the candidate discriminative feature patterns for our prediction model. However, observe that certain patterns are more informative than others, and hence should contribute with a higher weight in the process of prediction. To take into account this importance factor (of patterns), CLARA ranks the frequent patterns and assigns weights based on its *relative frequency* defined as its frequency of occurrence over the total number of records. Similar to the clustering approach, for each employee, another score based on matching the employee features with the frequent patterns, taking into account the ranks assigned to the frequent patterns.

4) Employee Ranking: The final *predictive attrition score* assigned by CLARA to an active employee is computed by a linear combination of the clustering score and the pattern matching score of the predictive scoring module. Note, that other combinations of the scoring functions can easily be used depending on the application requirements. The employee predictive attrition score provides a proxy for the probability of voluntary attrition of an active employee, and is larger if the predicted chances of separation are higher. Hence, based on the total predictive scores CLARA generates the *attrition rank list*, an ordered list of active employees at the highest risk of attrition. It is for these employees that our framework subsequently recommends remedial actions to help improve their retention.

5) Remedial Actions: The major novelty of our proposed CLARA framework is the recommendation of personalized remedial steps on an individual level to improve the rate of attrition – deviating from predictive to remedial HR analytics. In this regard, for an active employee with a high predicted attrition score, CLARA searches for the *closest* active employee that belongs to an *active cluster*. Since employees in an active cluster are possibly less prone to attrition, comparing the associated employee features provides avenues to improve the retention of the current employee. Also note that this closest possible vector difference provides the ***least disruptive*** change, optimizing operational cost of an organization. This multi-dimension view of the computed feature vector difference is translated to a natural language based recommendation. For example, a salary difference in the feature vectors provides a simple remedial action of "Increase salary", while a difference in the features `salary` and `job grade` might correspond to an "Increase responsibility" recommendation.

Figure 2 shows a toy example of a remedial action. For simplicity, we use a bi-dimensional space here: Salary Band, Team Size. We have 3 active clusters and 1 leaver cluster. Letters 'A' mark active employee in a cluster, while we use 'L' for leavers. In this image, CLARA needs to create a remedial action recommendation for the active employee in the bottom left cluster where the three vectors start from. CLARA considers all the active people in all the active clusters, and find the closest active to compute the remedial vector. We depict here two valid remedial actions, however one of them is the least disruptive (as it's the shortest vector), the other one requires bigger changes in either dimension. We also depict with a red dot-dashed line a non-valid remedial action: this vector would bring the active employee to an active in a leaver cluster, which is not our desired output.

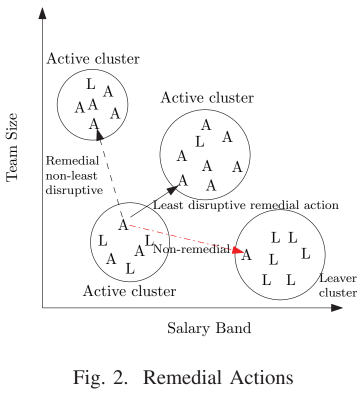


Fig. 2. Remedial Actions

This multi-dimension view of the computed feature vector difference is translated to a natural language based recommendation. For example, a salary difference in the feature vectors provides a simple remedial action of "Increase salary", while a difference in the features `salary` and `job grade` might correspond to an "Increase responsibility" recommendation.

6) Output API: The final module of the CLARA framework provides the external interface API, which can be as simple as presenting to the HR and management a list of employees at a high risk of attrition along with the suggested remedial steps.

III. Performance Assessment

In this section, we present the performance evaluation of our proposed framework on a use case scenario utilizing the publicly available *IBM HR Analytics Employee Attrition & Performance* dataset generated for the Kaggle 2015 competition (obtained from www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset). The dataset contains 1470 anonymized employee records with 34 individual employee features such as demographics, job satisfaction, time since last promotion, tenure in the company, etc. It also consists of categorical features (like job level, frequent travel, etc.) as well as continuous employee features (like monthly salary, hourly rate, etc.), along with the binary information whether an employee has left the company or not. The continuous variables were discritized into almost-equi-populated bins, and all categorical variables were one-hot encoded. Since the IBM data pertains to one year only, to study the performance of CLARA, temporality was introduced by randomly allocating leavers to either the first half of the year, or the last half. The randomization was repeated for five different runs, and all results reported in this section are averaged across the runs. All experiments were conducted in Python on an Intel i7 2.9 GHz processor with 24GB RAM running Windows 10 OS.

We benchmark the performance of CLARA, for our use case, against Random Forest, XGBoost Trees, Support Vector Machines (SVM), Spectral Clustering, standalone k-Means clustering and standalone Frequent Pattern Mining (FPM) state-of-the-art techniques. We empirically set the number of clusters in CLARA at 23 as a good trade-off between cluster purity and computational efficiency.

Figure 3(a) presents the precision of employee attrition prediction performance across the different competing baselines. We observe CLARA to outperform the existing techniques, providing around 15% improvement in performance (with a precision of around 70%) for the top-20 employees identified at the highest risk of attrition. This is important, as a high precision rate towards the top of the prediction list can accurately identify the employees almost at the point of attrition, providing an opportunity for pre-emptive remedial steps to increase retention. However, the approaches demonstrate similar precision (around 30%) towards the tail, with spectral and stand-alone clustering approach performing poorly.

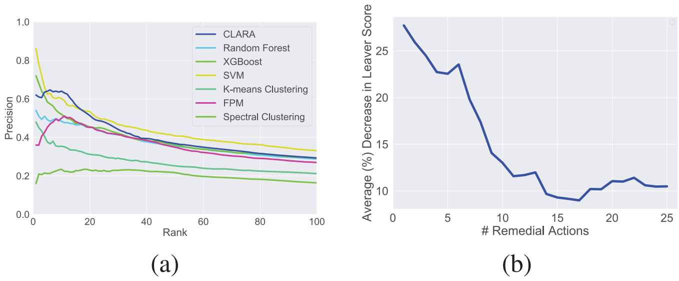


Fig. 3. (a) Performance comparison of precision on predicting Employee Attrition, and (b) Average score reduction for the top x *employees* after applying the remedial actions.

Effect of Remedial Actions. To study the "goodness" of the suggested remedial actions, we compared the effects of CLARA by applying the actions on the top x ranked employees. Figure 3(b) depicts a 22.5% reduction in the attrition score if we apply the top 5 remedial actions, and a 13% reduction with the top 10 actions.

Averaging in Ranking: CLARA's precision as we have reported earlier involves: (i) computing the predicted attrition score for each employee; (ii) ranking the employees; (iii) averaging the precision score for CLARA across the randomized runs. However, as an alternate strategy, we now (i) compute the predicted attrition score for each employee for the different randomized execution of CLARA; (ii) obtain employee ranking based on the average attrition score of employees across the different runs; and (iii) compute the precision. We saw an improvement of upto 15% in the precision (for the top-20 ranking) of CLARA with the latter approach. This can be attributed to the additive nature of errors (possibly w.r.t. sub-optimal clustering and patterns with low support) across the different runs on different data partitions.

A. Deployed System Highlights

A simple version of CLARA (with a basic Web-based UI design) is shown in Figure 4. Given a dataset of employees, CLARA ranks them by the highest predicted risk of "attrition", along with *natural language* based recommended remedial actions. The HR personnel might select potential employees that the organization would want to retain, and would like to consider the effects of the remedial actions (if acted upon for the selected candidates) on reducing the overall attrition rate or on their attrition probabilities. This is demonstrated by the *"What if?"* scenario button, wherein CLARA produces the result of such decision on every individual based on two KPIs: (i) the difference in Predicted Attrition Score $\mathcal{PAS}$ (negative is better); and (ii) the disruption of the operation (lower is better). The disruption is a cost computed on a budget like additional annual salary, total group change operations, etc.

IV. Conclusion

We have presented CLARA, an end-to-end system for recommending remedial actions for employees predicted to be at risk of attrition. CLARA's main novelties are: an *attrition scoring function* based on clustering and frequent pattern mining, and the computation of *least disruptive* personalized remedial actions to apply on employees to reduce attrition rate.

On the IBM dataset CLARA's predictions are up to 65% accurate (comparable to the state-of-the-art), and the expected overall predictive attrition score reduction is 22.5% on the top-5 employees at risk of attrition. Our internal performance validation has achieved higher precision and risk reductions, due to better characteristics provided by the real data. Experimental results proved CLARA to be robust and applicable to diverse enterprise environments.

Acknowledgments. We are very thankful to Leona Farrell, Snehal Khirid and Della Thomas for helping us in the acquisition of organization data, the validation of the recommendations, and expert insights regarding this work.

References

[1] IBM, "HR Solutions, IBM Talent Management," 2018. [Online]. Available: https://www.ibm.com/talent-management/hr-solutions

[2] P. S. Morrison, "Who cares about job security?" *Australian Journal of Labour Economics*, vol. 17, no. 2, pp. 191–210, 2014.

[3] J. Hancock, D. Allen, and F. Bosco, "Meta-analytic review of employee turnover as a predictor of firm performance," *Journal of Management*, vol. 39, no. 3, pp. 573–603, 2013.

[4] H. Boushey and S. J. Glenn, "There are Significant Costs to Replacing Employees," 2012. [Online]. Available: https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf

[5] S. Chitra and P. Srivaramangai, "A Study on Analytics of Human Resource Management in Big Data," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, pp. 58–68, 2018.

[6] H. Chang, "Employee turnover: A novel prediction solution with effective feature selection," *WSEAS Transactions on Information Science and Applications*, vol. 6, no. 3, pp. 417–426, 2009.

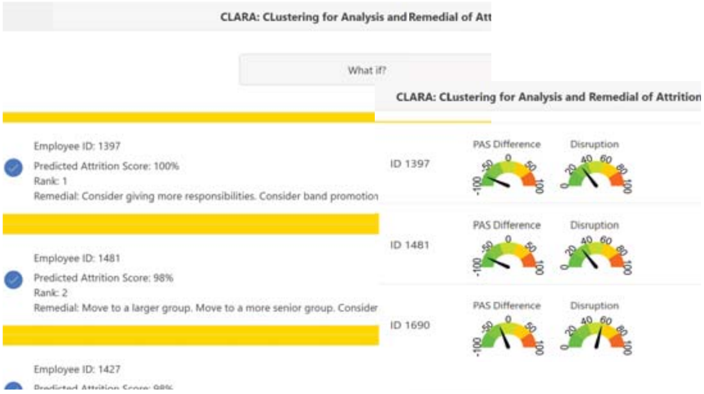
[7] N. Zhou, W. M. Gifford, J. Yan, and H. Li, "End-to-end solution with clustering method for attrition analysis," in *IEEE International Conference on Services Computing (SCC)*, 2016, pp. 363–370.

[8] E. Sikaroudi, A. Mohammad, R. Ghousi, and A. Sikaroudi, "A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *Journal of Industrial and Systems Engineering*, vol. 8, no. 4, pp. 106–121, 2015.

[9] SAP Inc., "Applying predictive analytics to manage employee turnover," 2017. [Online]. Available: https://blogs.sap.com/2017/07/20/applying-predictive-analytics-to-manage-employee-turnover/

[10] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes Clustering," *Journal of Classification*, vol. 18, no. 1, pp. 35–55, 2001.

[11] M. J. Zaki, "Scalable algorithms for association mining," *IEEE Trans. on Knowledge and Data Engineering*, vol. 12, no. 3, pp. 372–390, 2000.



CLARA: CLustering for Analysis and Remedial of Att

What if?

CLARA: CLustering for Analysis and Remedial of Attrition

Employee ID: 1397
Predicted Attrition Score: 100%
Rank: 1
Remedial: Consider giving more responsibilities. Consider band promotion

ID 1397

Employee ID: 1481
Predicted Attrition Score: 98%
Rank: 2
Remedial: Move to a larger group. Move to a more senior group. Consider

ID 1481

ID 1690

Employee ID: 1427
Predicted Attrition Score: 98%

Fig. 4. CLARA System UI Highlights