National College of Ireland

Predicting the Best Fit Employees to a Project by Human Resources

MSc Data Analytics (MSCDA_B)
Research in Computing

Tejas Phopale
Student ID: x17162301

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Tejas Phopale
Student ID:	X17162301
Programme:	MSc. In Data Analytics **Year:** 2018-19
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	20/12/2018
Project Title:	Predicting the Best Fit Employees to a Project by Human Resources
Word Count:	7393 **Page Count:** 25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 20th December 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Predicting the Best Fit Employees to a Project by Human Resources

Tejas Phopale
x17162301

Abstract

Talent management and strategic HR are the vital factors of Human Resource Management (HRM). The technology evolution leads to the demand for skilled and knowledgeable individuals with the ever-changing requirement for product and services. HRM needs to manage the expertise and talent in challenging enterprises. It is challenging to allocate the skilled employee for a project by manually screening their resume due to time constraints. Maintaining employee satisfaction is the perennial challenge for the company. The approach outlined in this research will help to utilise employee's details along with keeping budgetary constraints into mind, allocating best employees for the project and finding the attrition pattern to retain the crucial employees. Additionally, finding the low performers and re-skill them to their target skills is also necessary. The Predicting Best Employee Selection (PBES) solution will provide the resolution of all the complications in that by entering the budgetary constraints, HRM team or Manager can get a list of best-fit employees to choose. Then, the list of the low performers can help the HRM team to assign training on the domain. Furthermore, it helps in identifying the attrition pattern to retain the crucial players. The employee's grading clusters achieved successfully with K-mean model along with completion of SVM model will successful find the attrition pattern. The Integration with existing HR application remains uncovered.

Keywords: Clustering; Classification; Predicting Best Employee Selection; Re-skilling; Attrition

1 Introduction

Human Resources is the central department of an organisation (Hausknecht, Rodda, & Howard, 2008) which has responsibilities such as searching perfect candidates by screening their resume, on-boarding and project allocation and proper utilisation of the skillsets. The department is also responsible for many objectives such as socialisation, managing training, benefit programs, compensation and appraisal cycle. The research provides predictive analysis and clustering mechanism which helps to accomplish the required outcome. This paper will try to answer the research questions:

1) Can high performing employees be selected to the project and upskilling be improved by detecting low performers?

2) Can an employee's attrition pattern be detected?

The methodology works in such a way that the PBES solution will provide the list of a best-performing employee along with grades after computing the provided data and then the user has the option to analyse budgetary constraints which includes the range of employee salary. The combinational methodology of narrating the low performing employee and assign

trainings, boot camps to them to increase the domain knowledge which will result in increasing the productivity for upcoming projects. Another module of the PBES solution helps to provide the attrition pattern which states the important reasons why the employees wish to exit the organisation. The computation analysis performed on employee's data who were terminated or resigned will be helpful to achieve analytics on current employees. It helps to address the critical issues faced by the organisation for affective HR management. The data mining with the help of a new analytics solution is going to be cooperative regarding time management and productivity. It provides real-time information on currently available data which makes the realisation of current state statistics within the organisation and help to take rapid actions. The computational machine learning theories and pattern recognition will provide predictive analytics. This analytic transforms into an actionable intelligence which enables quick decision making. The PBES solution enables the method for selecting highly skilled and suitable candidate with substantial evidence. Eventually, it increases the productivity of the project and endeavours smooth project delivery within the timeline. Also, upskilling of employees can be achieved by getting a list of data from the projected forecast and allocating the sessions or training. The below figure (1) describes the key analytics of HR department.

The dataset taken from Kaggle published by IBM HR.[1] The dataset originates for Kaggle challenge to uncover the critical analysis based on seven main reasons for employee turnover identified by the author (Yazinski, 2009) retrieved from website.[2]

Figure 1: Key Responsibility of HR Department

In everyday life, HRM performs a manual intervention for the resource allocation. The paperwork makes work fussy and time consuming. Even with lot of efforts and manual

[1] https://www.kaggle.com/dgokeeffe/ibm-hr-dataset-exploratory-data-analysis/data

[2] http://hr.blr.com/whitepapers/Staffing-Training/EmployeeTurnover/Strategies-for-Retaining-Employees-and-Minimizing-Turnover

screening, it will not provide the highest probability that they have selected the best one among them. The project team or HR organise many technical and non-technical boot camps for knowledge upskilling and increasing the productivity, but it does not guarantee that right candidate has got that opportunity to attend it. The candidate recruitment may need manual intervention for screening resume and hiring them because HR team is unaware about candidate's skills and performance matrix, but it is possible to apply automated analytics within the available pool of employees. Selecting correct low scored employees and appeal for reskilling is overlooked till date. It is imperative to fill this gap and give valuable benefits to the organisation by implementing this solution. The most critical aspect is that it is a less time-consuming process. The manual intervention involves screening the individual's profile, discussion with the project manager and then decision-making meetings even though there might be chances on missing best employee.

2 Related Work

2.1 Introduction

Human Resource is a vital department (Cao et al.,2011) for the organisation to accomplish staffing and to manage sufficient skills. The failure in staffing service management and organisational engagement can lead to inconvenience in project allocation and organisational management failure. The lack of management can surge bench count which reduces the productivity. Moreover, it roots to the disappointment of clients and surges in employee turnover thereby this process diminished the organisational tasks.

The formation of non-routine employers has been raised from 16% to 33% by 2010 (Varshney et al. 2014) which categorised as creative and experts. The conventional generation needs innovative, unimaginable skills and productivity which needs to conserve and manipulate by HR team. The further section helps to identify loopholes in the process and management skills.

2.2 A Review of Current Scenario

The market needs hot skills persons (Trevor et al., 1997) with limited pool size. Therefore, HR department need to play a vital role by addressing changing skills demand without bothering about intake candidates influences of salary growth and in-house employee's promotion. Also, HR department (Anderson et al., 1980) should keep the assurance to hire sufficient numbers of candidates to meet companies demand.

The market of every organisation needs skill persons (Trevor et al., 1997) to familiarise their system to the world and for that HR department has to play a crucial role to manage the skill on demands and allocate the employees who matched with their profile. Also, HR team manages the employee upkeeps, salary, proper utilisation of skillsets and in-house employee's promotion. To achieve it (Anderson et al., 1980), they need to assure the sufficient numbers of employees to meet the demands and immediate desires.

The essential factor of HR department (Bersin et al., 2013) is that transforming the support functionalities into the strategic outcomes and there are some analytics stated by Bernstein who is vice president of HR department in eQuest of Big Data. These advanced analytics includes operational and advanced reporting. The predictive analytics works (Singh

et al. 2012) by developing a model with the help of strategic planning which matches the organisational agenda, but only 4% of organisations follow those agendas. The organisation should plant the platform for strategic planning with an analytical solution to manage talent workforce, appropriate allocation at matching fields, justify positions and other happenings for HR department.

2.3 An Investigation for Selecting Best Fit Employee

The IBM design approach (Brown et al., 2009) anticipated for deployment of an innovative system planning which should be one-time approach and apply to the sales organisation worldwide without a user interface or fixed framework. The approach is such a way that the paradigms that user stories are segregated based on job role and creating a detailed prediction task. This process helps to capture the required elements to administrate specific profiles. It could have user interface design approach. The architecture is consisting of in-line text analytics using logistic regression and data warehouses. The logistic regression in pair with k cross-validation helps in validating the generated model.

By considering only two parameters (R Fan et al., 2008) such as job role and skillsets, the study performed on supervised multi-categorical classification. The feature sets were found and classification was applied which elaborate to the comprehensive study. The accuracy for this logistic regression performed by using "Liblinear" implementation. It helps to evaluate the data and perform the computation by filtering based on the job title.

Another study has been performed on multicategory dataset (Balog et al., 2007) wherein medium-level job role classification performed. The following method does estimate on coarse-level data (Han et al., 2016) but enterprise data.

2.4 A Review of Technique, Identified Gaps and Different Methods

The present demand need empowerment with a newly advanced methodology which fulfils the necessities of skillsets and accomplishes the actual plan of an organisation. It helps in bridging the previous employment gaps by establishing the solid project plan and eventually meet the required expectations. For example, there can be other solutions which could be a similar type of procedure like resume matching. The best candidates judged after analysing the skillsets, experience in the field, total experience in the industry, job satisfaction rating, manager rating, previous hike and other factors. It will be too messy to analyse with manual intervention because it is a vast and complicated process. The study regulates little different approach by moving workloads (Wei et al., 2015) to high pressure as per the current position within the organisation, necessity of trainings for reskilling and base clearance expertise. However, this approach reflects higher component optimisation and performance would not be up to the mark. It demonstrates mathematical representations which leads to the wrong direction undoubtedly with regards to profiling and scoring scheme.

Furthermore, there is the use of collaborative filtering (Fang et al., 2013) to investigate next level of the cold-start regime using matrix completion. However, collaborative filtering will not be that useful to gather and analyse required parameters like job role, specialities and other aspects of predicting approaches. The testing approach for correcting noisy labels (Natarajan et al., 2013) was fade away after deployment to maintain the simplicity else it

creates enormous complexity. There are around 4% of the organisations understands the necessity of predictive capacity analytical reports, if the employees (Bersin et al.,2013) and these organisations are mainly focusing on the proactive measure on retention of talents and resource re-skilling (Ramamurthy et al., 2015). The talent management with strategic and tactical skills should be engaged (Ilgen et al., 1991) to retain the skilled employees and maintain the track for up-to-date information. Most of the organisation neglects it by considering the theoretical terms which produce difficulties to create taxonomies on hierarchical expertise.

There is a study using ordinal regression and K-mean clustering (Jain, A.K., 2010). The cluster creates a different band of data into a particular space. The clusters levels can be minimal or more profound but measurable. Those boundaries called as hyperplanes which are parallel, and weighted slope of the vector can measure the depth. The study by using supervised multi-category classification (Varshney et al., 2014) implemented and analytics performed against only two essential parameters such as job role and specialised skills. This paradigm helps to measure the upscales and downscales employee. So, there is an alternate analytically driven approach (Ramamurthy et al., 2013) which is used for re-skilling the employees, found in low scale measures by evaluating the dataset and high performers listed into the critical and valuable project allocations. However, the solution does not seem justifiable by considering only two features because there might be other aspects of evaluating the scaling and grouping priority may change with those factors. The author (Li et al., 2017) explains the essential role of KDD methodology over others. In this, predictive analytics has performed on state-of-the-art baseline and ranking constraints methods. It evaluates future prediction based on employees' performance, organisation turnover and profession progress.

Another approach for prediction by Matrix-Query fusion techniques (Horesh, Varshney and Yi, 2016) wherein ordinal regression and clustering are performed. It provides the detailed machine learning evidence on the expertise of the employee in the particular skills and data distributions.

2.5 Limitations

The organisation and HR department can misjudge the existing skilled and talented employees and may ask them to leave the organisation due to manual intervention. Inefficient HR management can lead to low work satisfaction, lack of exhibition skills, lack of confidence, absence growth vision and instabilities in resource management.

It is necessary (Mojsilović et al., 2015), to manage the talent for an organisation and own skills and progression in a career for an employee. In the era of big data, it is essential to store the employee details and factors which can help to evaluate business management and create future roadmap. The employee's career footprint such as resume, domain knowledge, skills, certifications, role history and other common factors are important to evaluate. There is a public social-networking site like LinkedIn (Xu et al., 2014) where they capture all the particulars but most of the organisations have unnoticed the information management for in-house talents.

2.6 Conclusion for Literature Review

Based on an overall study on literature review, it observed that research on predicting best employees, finding attrition and re-skilling the existing employees' talent are the essential factors for organisation progress and ultimate profits. The grade wise categorisation of employee with budget range is unnoticed. By implementing the PBES solution, HR team can evaluate and read the approach at the employee level. The management can suggest better solution for every small problem and stabilise the organisation position. Also, sudden growth and progressive steps on every interval observed after implementation of the solution. Furthermore, the employee's productivity and outcome level can be improved which eventually moves the organisation in the position way.

3 Research Methodology

3.1 Scientific Methodology

The CRISP-DM methodology (Azevedo, 2008) emerged to be suitable for PBES solution. The reason behind selecting CRISP-DM is that it is iterative, long-term strategy and has structured template. The iterative stage allows the model to go backward if the model goes in the wrong direction. By detecting those mistakes, the model becomes strong due to long-term strategy phase and template helps to find missing or remaining part of the work.

3.1.1 Business Understanding

It is necessary to understand the problem statement by HR department with the project manager regarding the ongoing process and confirming the objective of the research defined in the process. This phase should take a minimal amount of time and money for the initial phase.

3.1.2 Exploratory Data Analysis

The required dataset collected from a data source mentioned above and analysed for different data features were observed by understanding the relevance of it. For differentiating best-performing employees and worse performing employees, the clustering performed by taking relevant features. For finding the attrition pattern, supervised learning classification method performed. The class imbalance observed in attrition feature with proportion as 84/16, so the accuracy cannot be measured with ROC curve because performance measures or standard optimisation criteria may not be effective (Lobo, Jiménez-valverde, & Real, 2008). The detail of the dataset explained in the further section.

The pre-processing data phases such as formatting, cleaning and sampling implemented sequentially. The data formatted by removing blank space to NA and capturing it into the data frame. The cleaning of data performed by omitting entry as there are only 0.00041% of NA into the data frame out of 23299 entries. Furthermore, checked and removed the rows which are constant blank cells and zero standard deviation. Next, the unwanted columns like application id, over 18, employee count and employee source deleted. Then highly correlated data deleted which having cut-off value of more than 0.6 to reduce pair-wise correlations.

For clustering, the numerical data selected, binning performed but normalisation not performed because all use a distance measure to determine if an object 1 is more likely belong to the same cluster as object 2 or in another cluster as object 3, and this measurement affects the scale of the variables. So, the scaling performed to standardise the data and find the number of clusters using the wss Elbow method (Reference) as Silhouette and gap statistic method does not provide a clear indication for this project. Normalising all variables by keeping them into the similar range and weight makes no intelligence. However, for classification, the normalisation performed to transform the data by subtracting the minimum from each value and dividing it with the range of differences. Binning and labelling performed for classification. The essential features selected using variable importance method before running the actual models.

3.1.3 Modelling

In this project, unsupervised and supervised machine learning technics applied to predict likelihood success of PBES solution. By understanding the business goal of HR department, clustering performed using KMean based on employee's overall statistics and evaluated based on some supervised learning. For predicting the attrition pattern of employees, the classification model built on different models such as KNN, SVM, GBM and Random Forest to formulated the best one. For KNN, the Kth value determined before running the final model. The project plan performed by analysing situation assessment of objectives by evaluating the employee's details. The data sampling performed by dividing into training and testing datasets.

3.1.4 Evaluation

The model evaluation performed based on their performance, accuracy, sensitivity, specificity and context of successful business criteria of both clustering and classifications. The detail information is accessible in the evaluation section of the paper.

3.1.5 Deployment

Once the model evaluated, the integration performed with code, results extracted and comparison performed. The graphical representation with the help of the Tableau helps to find best-performing employees availability based on budgetary constraints, re-skilling the low graded employees and finding the attrition pattern.

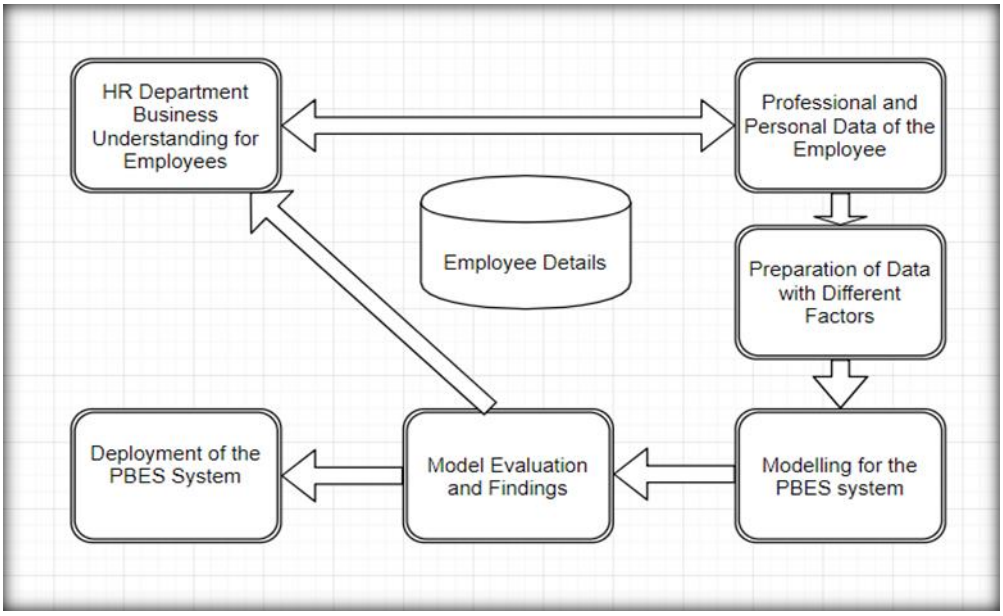

Figure 2: CRISP-DM methodology

3.2 Research Significance

The principal objective of the research is to benefit HR management and project managers with automation techniques. The PBES solution helps them to evaluate the chance of successful, talented employee selection, finding the attrition reasons and the probability of finding the employees who might plan to leave the organisation. It helps to stop manual intervention and saving the massive amount of efforts, money and time.

Also, it helps to grow a standard solution for representing the organisational activity to management and future statistical of internal employee's skill improvement. The organisational growth can be observed based on continue practice of evaluation of solution's reports.

3.3 Research Limitation

The current solution works on the format provided by IBM HR Data. The other data structure might need to change the pre-processing data methods. There might be some constriction on PBES solution, that HR member can wrongly enter or miss the data which reduces the prediction efficiency. So, the assumption for this project is that personal and professional details available of the employees should be accurate and the process should be conducted truthfully. For clustering, the dataset can be considered as the employees of an organisation who are unassigned for any project.

4 Design and Implementation Specifications

4.1 Dataset

The PBES solution helps to evaluate the results based on employee's professional details which are maintained by HR team. The dataset comprises of 23,533 entries and 37 features which consist of both numerical and categorical type. The dataset consists of different essential factors shown in the table (1) below.

Table 1: Structure of the Dataset

Sr. No.	Feature Name	Details
1	Age	In Years
2	Attrition	Current employee; Voluntary Resignation; Termination;
3	BusinessTravel	Travel-Rarely; Travel-Frequently; Non-Travel
4	DailyRate	Payment per day
5	Department	Human Resources; Research & Development; Sales
6	DistanceFromHome	How far the employee lives from work
7	Education	Graded as least 1 and highest 5
8	EducationField	Human Resources; Life Sciences; Marketing; Medical; Other; Technical Degree
9	EmployeeCount	No of employees in this record (Constant count 1)
10	EmployeeNumber	Employee ID
11	Application ID	Application ID
12	EnvironmentSatisfaction	1; 2; 3 ; 4
13	Gender	Male / Female
14	HourlyRate	Payment per Hour
15	JobInvolvement	1; 2; 3; 4
16	JobLevel	1; 2; 3; 4; 5
17	JobRole	Healthcare Representative; Human Resources; Laboratory Technician; Manager; Manufacturing Director; Research Director; Research Scientist; Sales Executive; Sales Representative
18	JobSatisfaction	1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
19	MaritalStatus	Divorced; Married; Single
20	MonthlyIncome	monthly salary
21	MonthlyRate	Charge per month
22	NumCompaniesWorked	No. of previous employers
23	Over18	Y/N
24	OverTime	Yes/No
25	PercentSalaryHike	Last Increment
26	PerformanceRating	3; 4
27	RelationshipSatisfaction	1; 2; 3; 4
28	StandardHours	Contract hours (Default 80)
29	StockOptionLevel	0; 1; 2; 3
30	TotalWorkingYears	Career Age
31	TrainingTimesLastYear	No. of training courses attended last year 0 to6
32	WorkLifeBalance	1 'Bad'; 2 'Good'; 3 'Better'; 4 'Best'
33	YearsAtCompany	No. of years with company
34	YearsInCurrentRole	No. of years in current role
35	YearsSinceLastPromotion	No. of years since last promoted
36	YearsWithCurrManager	Years spent with current manager
37	Employee Source	Recruiting agency

For classification, the Labelling performed on categorical features and normalised the data to increase the model accuracy and kappa with consideration of other confusion matrix parameters such as sensitivity and specificity.

4.2 Algorithm

There are two different segments for this project. One, to create clusters of employees of an organisation with their overall statistical performance and group them to find the best performers and least performers. Second, to find attrition pattern and finding reasons for leaving an organisation with the help of classification methods. So, there are many classifications algorithm used such as KNN, SVM, GBM and Random Forest. For clustering, K-mean algorithm used to get the range of performing employees. To evaluate those clusters, supervised learning classification algorithms applied over clusters.

4.2.1 K-mean clustering

The K-mean is most significant and widely used to find the explicit distance and partition data into the clusters to represent the vector of mean feature for numeric attributes. The K-mean used for creating clusters for best-fit employees and low performing employees. After data pre-processing described above, the decision for number of clusters taken by computing different features such Davies-Bouldin's cluster index (Zhao, Xu, & Fränti, 2009), Elbow method, PCA, Silhouette method and Gap statistical method. It has been observed that the Elbow method and Davies-Bouldin's cluster index shows perfect analysed throughput and represents four optimal number of clusters. The below figure (3) shows the optimal number of clusters and lowest DB index value for four clusters.

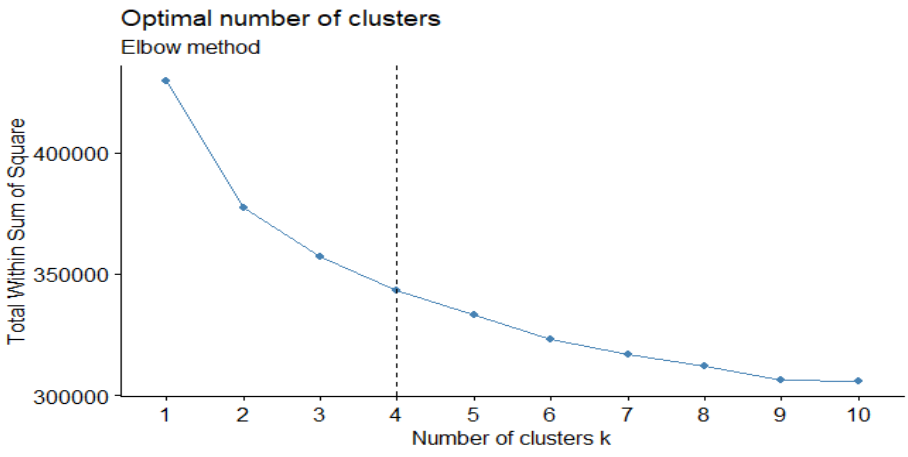

Figure 3: Elbow Method and Davies-Bouldin's cluster index

4.2.2 k-nearest neighbours (KNN)

It is a pattern recognition technique using the non-parametric method for classification and used for finding an attrition pattern. Feature selection performed and

different approaches applied to find k value such as *sjc.elbow*, square root size of the training dataset and model fit after pre-processing. By considering the optimal k value, the model trained and tested against training and testing datasets respectively and observed the confusion matrix and model accuracy. The training control managed after every attempt and resampling results across tuning parameters shows best result occupied on k equals to 5 with threshold 0.95, frequency cut 95/5 and cut-off at 0.9.

4.2.3 Support Vector Machine (SVM)

The SVM is supervised learning and discriminative classifier model which appropriately defines by separating hyperplane. Many computations constructed on trading kernels such as "rbfdot", "polydot", "vanilladot", "tanhdot", "laplacedot", "besseldot", "anovadot", "splinedot", "stringdot" and classification type such as "C-svc", "nu-svc". The final model observed optimal with "laplacedot" Laplacian kernel and "nu-svc" classification type because "nu-svc" type ranges from 0 to 1 and attrition pattern of this project is with the binary classification. The SVM also used for evaluation of clusters which obtained for best fit employee's selection created using KNN. In this case, the model worked optimally with "rbfdot" radial basis Gaussian kernel and "spoc-svc" singer native multi-class type.

4.2.4 Gradient Boosting Model (GBM)

It is an ensemble of stage-wise weak prediction models which typically runs as decision trees and provides an optimisation using cost function. It refits the sub models from a complex model to pseudo residuals. For finding the attrition pattern, the model trained to get important variable using *varImp* function and then by selecting those features the model created with the final training set. The prediction performed on testing data and observed the confusion matrix.

4.2.5 Random Forest

The Random Forest is an extra layer of randomness to bagging Breiman (2001) (Find this paper). It uses different bootstrap sample data to construct a classification wherein each node splits among best predictors subsets and this diverse strategy performs well in comparison to many classifiers. For finding the attrition pattern, the critical feature selection is performed and model built based on those features. The prediction performed on testing data and observed the confusion matrix.

4.3 Architecture

The steps for the flow architecture of Predicting Best Employee- solution mentioned below:

1) The data uploaded from an excel sheet with employee's details for analysis.

2) The machine learning algorithm performed on the data to get optimal results. The first result help to get best-fit employees and low performer with the help of clusters wherein the second result provides the attrition pattern, and analyses report with explanations with the help of classification.

3) The group of clusters along with employee numbers segregated with four types. The first group consist of excels or above grade level (4). The second group consist if

Proficient or at grade level (3). The third Approaching proficiency or approaching grade level (2) and the fourth group Well below proficiency or below grade level (1).

4) The level four and level three graded employees considered as best fit employee for a project. The budgetary constraint applied on the list of employees and budget administration represented by Tableau.

5) The level two and level one graded employees considered as the low performers so HR might request them for re-skilling with some boot camps. The list of employees with employee id in the form of excel and representation by Tableau.

6) The classification helps to find the attrition pattern and with the help of analysing tool such as Tableau, the reason for leaving the organisation investigated.

The below figure (4) explains the business flow for the same.

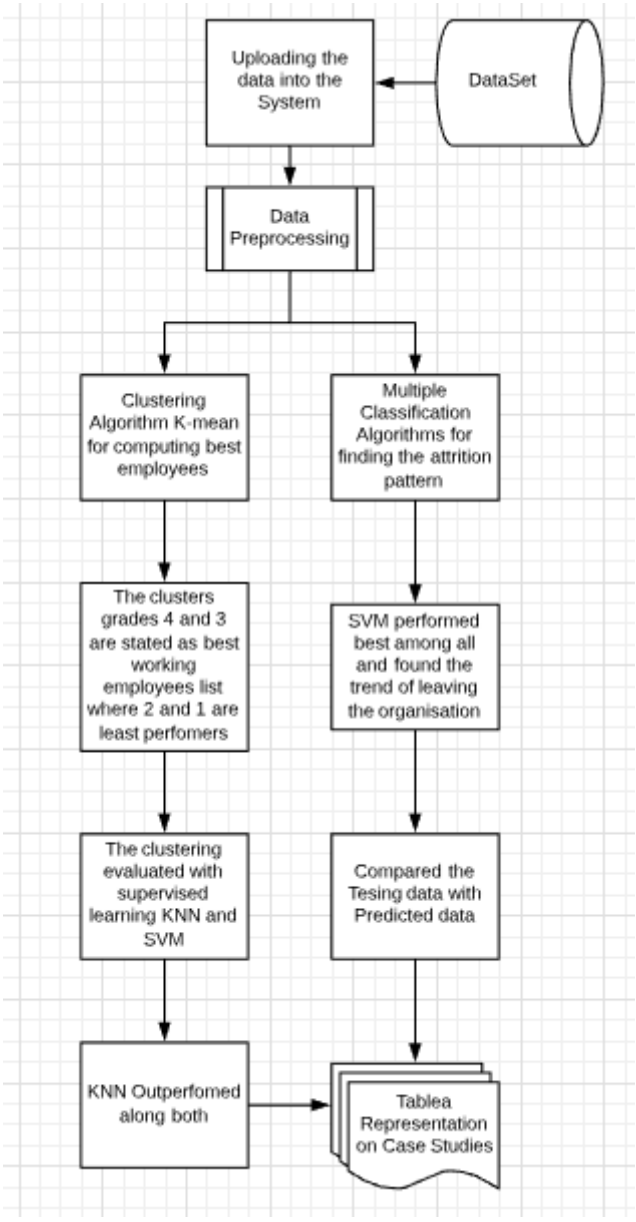

Figure 4: PBES Solution Architecture

5 Evaluation

This section elaborates the determination of results and essential findings of the project. The implications of these findings help in speculative representation and practitioner outlook. For clustering, KNN algorithm used to create the model to predict the likely success of grading employees and categorisation based on their overall performance. The four clusters selected mentioned in below figure (5). The K-mean algorithm executed with multiple parameters but finally, model shows the optimal result with twenty random samples for performance enhancement, hundred maximum iterations and algorithms such as "Hartigan-Wong", "Lloyd", "Forgy" and "MacQueen" together.

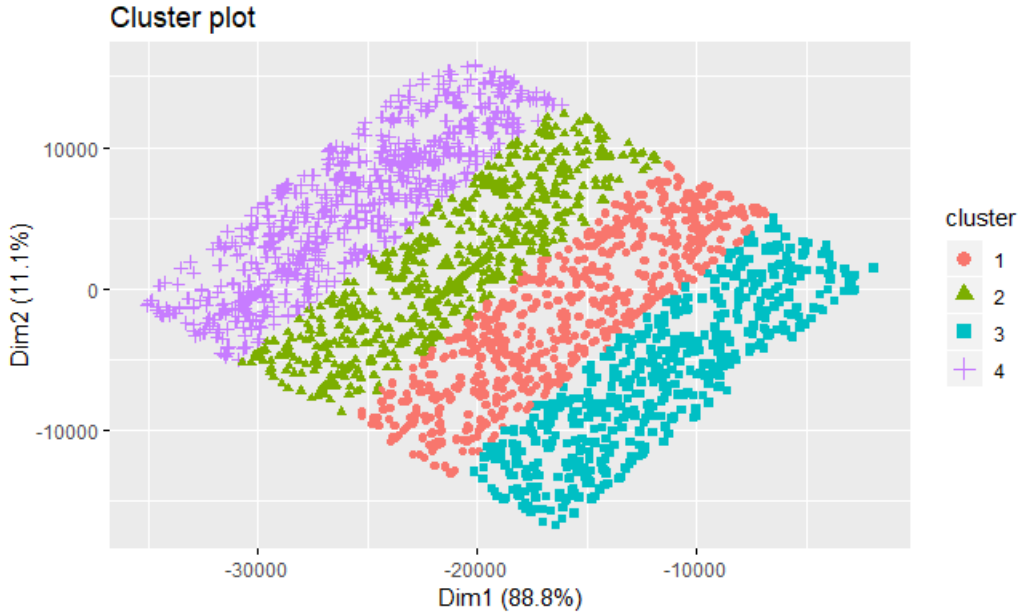

Figure 5: Plot for Clusters based on Overall Factors

By taking the order of clusters, the list of assignment merged with trained data framed for further evaluation. Now, to evaluate the grades of each clusters, labelling performed based on pointwise mutual information. The joint probability distribution of features is performed based on features like the highest percentage hike, performance rating, relationship, environmental and job satisfaction. There are total 27 observation selected in which 13 are optimal, and 14 least performers taken based on the feature as mentioned earlier and checked the cluster number in which those allocated. It's observed that 75% of optimal performers allocated in cluster four, and 70% of least performers allocated in cluster three. By evaluating this result against the distributed clusters, it has concluded that cluster four is of grade four (4) employees, cluster two is of grade three (3) employees, cluster one is of grade two (2) employees and cluster three is of grade one (1) employees. The cluster's evaluation is performed using SVM and KNN to check accuracy for the clustering. The KNN outperformed with an overall average accuracy of 98% with low false negative observations than SVM with an overall average accuracy of 90%. The below table (2) shows the confusion matrix of both the models. The Kth value is determined by *sjc.elbow* method as 5 and computation performed wherein SVM performed on "rbfdot" kernel with type "spoc-svc" that is Singer native multi-class.

Table 2: Evaluation of Cluster by KNN and SVM

Confusion Matrix and Statistics					Confusion Matrix and Statistics				
predicted_knn	1	2	3	4	predicted_svm	1	2	3	4
1	1163	11	6	1	1	1023	6	8	59
2	55	1140	14	9	2	70	1138	4	0
3	15	12	1197	49	3	54	19	1183	12
4	6	0	14	1189	4	92	0	36	1177
Overall Statistics					Overall Statistics				
Accuracy : 0.9607					Accuracy : 0.9262				
95% CI : (0.9548, 0.9659)					95% CI : (0.9186, 0.9334)				
No Information Rate : 0.2557					No Information Rate : 0.2557				
P-Value [Acc > NIR] : < 2.2e-16					P-Value [Acc > NIR] : < 2.2e-16				
Kappa : 0.9476					Kappa : 0.9017				
Mcnemar's Test P-Value : 3.642e-12					Mcnemar's Test P-Value : NA				
Statistics by Class:					Statistics by Class:				
	Class: 1	Class: 2	Class: 3	Class: 4		Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.9387	0.9802	0.9724	0.9527	Sensitivity	0.8257	0.9785	0.9610	0.9431
Specificity	0.9951	0.9790	0.9792	0.9945	Specificity	0.9800	0.9801	0.9767	0.9648
Pos Pred Value	0.9848	0.9360	0.9403	0.9835	Pos Pred Value	0.9334	0.9389	0.9330	0.9019
Neg Pred Value	0.9795	0.9937	0.9906	0.9839	Neg Pred Value	0.9429	0.9932	0.9867	0.9801
Prevalence	0.2538	0.2383	0.2522	0.2557	Prevalence	0.2538	0.2383	0.2522	0.2557
Detection Rate	0.2383	0.2336	0.2452	0.2436	Detection Rate	0.2096	0.2331	0.2424	0.2411
Detection Prevalence	0.2420	0.2495	0.2608	0.2477	Detection Prevalence	0.2245	0.2483	0.2598	0.2674
Balanced Accuracy	0.9669	0.9796	0.9758	0.9736	Balanced Accuracy	0.9028	0.9793	0.9689	0.9539

For finding the attrition pattern with classification, the SVM outperformed with an accuracy of 97% than KNN, GBM and Random Forest. The KNN performed with an overall average accuracy of 93%, GBM with 83% and Random Forest with 90%. The below figure (6) shows the details of the confusion metrics and overall statistics of all algorithms.

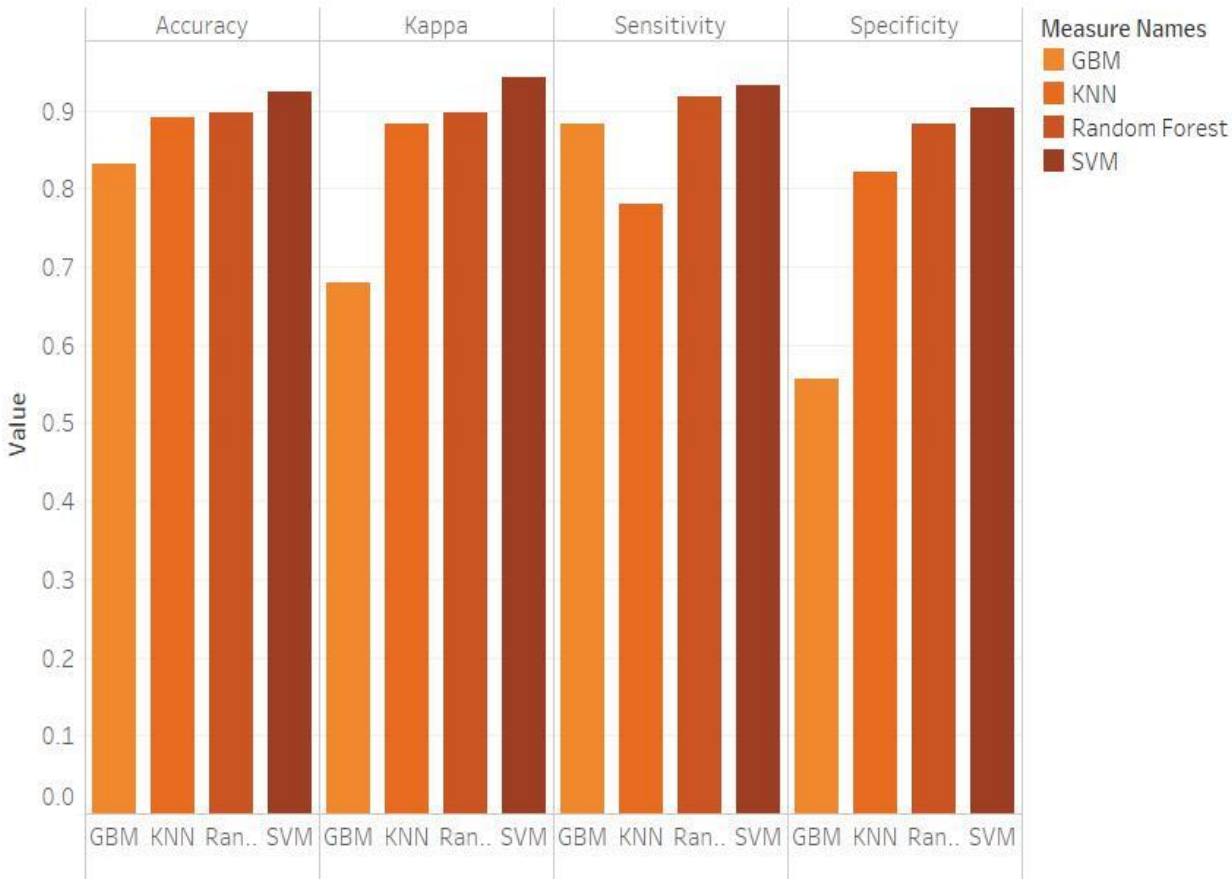

Figure 6: Finding the Attrition Models Comparison

5.1 Case Study 1

The first case study performed to study the overall availability of the employee from low grades (1) to high grades (4) within low (1) to high (4) budgetary constraints. The employee's categorisation performed based on skills and budget. It has been observed that every skill group has above mid-range budgetary employees. Also, by considering only best fits high grade (4) employees then there are around 13.62% employees available for low budget projects wherein around 12% of employees available for high budget projects. The figure (7) elaborates availability of different graded employees at every budgetary constraint.

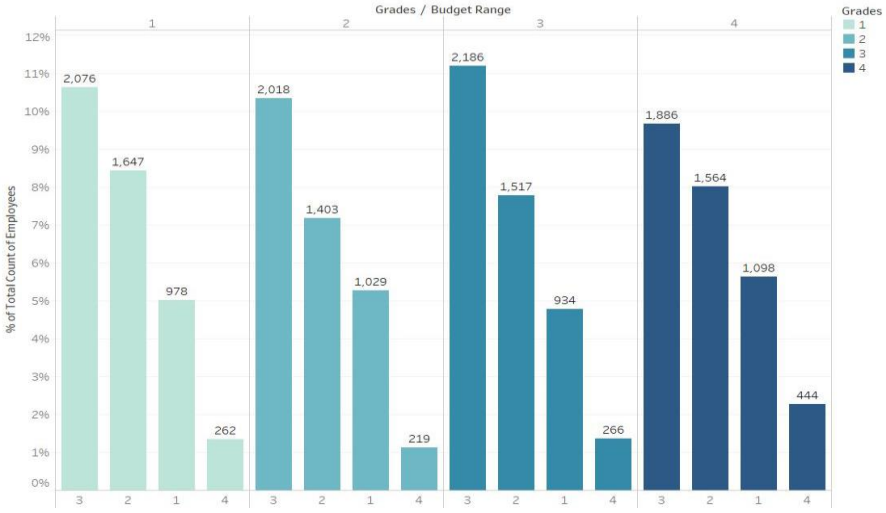

Figure 7: Categorisation of Employees by Grades and Subcategorised on Budget Range

5.2 Case Study 2

The second case study performed to study on employee's marital status. This helps to observe if marital status deflects the concentration from work. The evaluation performed by taking low grade (1) employees. The study says that married employees are more in low performing category than other factors and proves that there are no effects of working overtime over grades. The figure (8) shows that marital status effects on the performance of the employees.

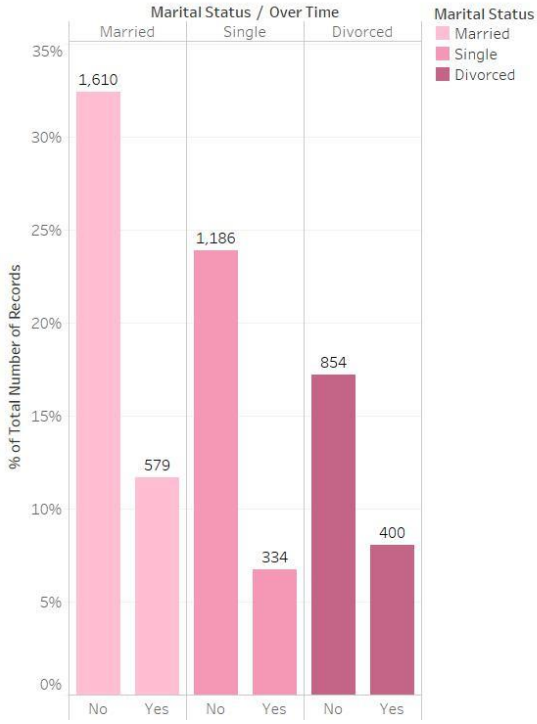

Figure 8: Low-Grade Employees Categorised on Marital Status

5.3 Case Study 3

The fourth case study performed to investigate which employees are leaving the organisation more. It has been detected that employees who rarely travel for the onshore

environment, specifically females from Research & Development department are likely to leave the organisation. The below figure (9) shows investigation on gender wise traveller scenario.

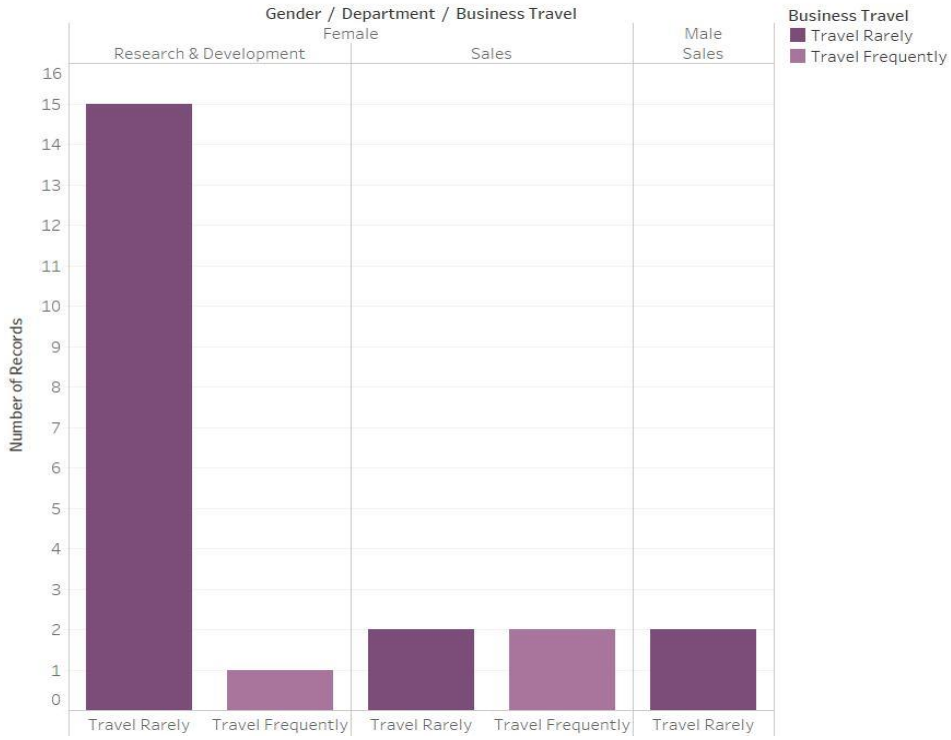

Figure 9: Organisation Leaving Trend by Comparing Gender vs Business Traveller

5.4 Discussion

The above experiment and case studies states that the research product accomplished the target successfully. The experiment tends to be an exception for the management to take quick decisions on selection of employees. The experiment and case studies seem to fulfil the objective for finding the best-performing employee. The list of employees with their employee Id are available in final data for project allocation process. The basic R Shiny UI created mentioned in figure (10) to view the list of employees with grade (4) against the range in which they are categorised. Likewise, user interface for low-grade employees can be created to assign them for knowledge development trainings.

EmployeeNumber	BudgetRange	Grade
17555	3	4
17556	3	4
17557	2	4
17560	2	4
17561	2	4
17562	3	4
17565	2	4
17568	2	4
17569	2	4
17570	3	4
17573	2	4
17574	2	4
17575	2	4
17576	2	4
17577	2	4
17578	2	4
17579	2	4
17580	2	4
17581	2	4

Figure 10: R Shiny UI for Grade (4) Employees

Furthermore, the detection of attrition pattern helps HR team to take prior steps to prevent the employees who might be willing to leave the organisation. Increasing the number of experiments and selection of additional parameters can help to get more accurate results. The literature review based on different algorithms and statistical computation developed roadmap for this research.

6 Conclusion and Future Work

The principal objective of the research is to develop a model for the organisation that can be used by HR department to find the most efficient employees within budgetary constraints, low performers, attrition pattern and monitoring the primary employment behaviour within the organisation to implement into HR analytics. Both supervised and unsupervised learning used as per requirement and multiple models generated using machine learning algorithms. Using K-means four clusters are generated and graded them on performance. The model implementation performed by selection of most important numeric features, binning and standardised them. To evaluate the clusters are well separated, SVM and KNN classification methods are performed. The KNN proved to be the best model with 96% accuracy. The individual and combination of results helps to understand the employee's abilities and HR team to retain key players of the organisation based on statistical analysis. Eventually, after efficiently collating all the parameters as mentioned in the research, there would be a positive impact on every project performance parameter in an organisation.

To find the attrition pattern, the supervised classification methods such as KNN, SVM, GBM and Random Forest are performed. The model implemented with the selection of most important categorical and numeric features. Then, encoding and binning is performed on that data. The SVM proved to be the best performing along them with the accuracy of 92.3%. The data visualizations used for interactive and easy representation of functional knowledge for non-technical users. The overall objective of the research achieved successfully as the system allows to select right employees for the project under the budgetary constraints, low performing employees detected which can be nominated for upskilling. The attrition pattern

detected which helps the HR team to hold the key employees after settling the concerns. Furthermore, it increases the efficiency and effectiveness of HR management skills and proposed technique helps in making better and faster decisions which saves the extra amount of time and money. The limitation of this research is that, the model based on only IBM HR data format. The project will not cover other HR management tasks and their integrations. In future, the models can be trained with more data and system can be enhanced with web applications using different technologies.

7 Acknowledgements

I want to express my sincere appreciation to my supervisor Noel Cosgrave for his academic support throughout my masters and teaching the data mining techniques for the research. It would not have been possible without his guidance and persistent help. Also, I would like to extend my sincere thanks and appreciation for the National College of Ireland for providing me with such an innovative platform on which I could further build my skills. Not to forget, I want to specially thank my parents for supporting me on my decision for Masters after five years of my professional tenure and my elder brother Mr. Gaurav Phopale, project manager in TCS, for encouraging me for this research topic which is the actual industrial necessity in the professional life.

References

Anderson, J.C. and Milkovich, G., 1980. Propensity to leave: A preliminary examination of March and Simon's model. Relations industrielles/Industrial Relations, 35(2), pp.279-294.

Apte, C., Skillicorn, D., Liu, B. and Parthasarathy, S. eds., 2007. OF THE SEVENTH SIAM INTERNATIONAL CONFERENCE. Proceedings of the 2007 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics.

Balog, K. and De Rijke, M., 2007, January. Determining Expert Profiles (With an Application to Expert Finding). In IJCAI (Vol. 7, pp. 2657-2662).

Bentley, R., Appelt, W., Busbach, U., Hinrichs, E., Kerr, D., Sikkel, K., ... & Woetzel, G. (1997). Basic support for cooperative work on the World Wide Web. International journal of human-computer studies, 46(6), 827-846.

Bersin, J., O'Leonard, K. and Wang-Audia, W., 2013. High-impact talent analytics: Building a world-class HR measurement and analytics function. Bersin by Deloitte.

Brown, T. and Katz, B. (2011) 'Change by design', Journal of Product Innovation Management, 28(3), pp. 381–383.

Cao, H., Hu, J., Jiang, C., Kumar, T., Li, T.H., Liu, Y., Lu, Y., Mahatma, S., Mojsilović, A., Sharma, M. and Squillante, M.S., 2011. OnTheMark: Integrated stochastic resource planning of human capital supply chains. Interfaces, 41(5), pp.414-435.

Cappelli, P., 2000. A market-driven approach to retaining talent. Harvard business review, 78(1), pp.103-11.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. and Lin, C.J., 2008. LIBLINEAR: A library for large linear classification. Journal of machine learning research, 9(Aug), pp.1871-1874.

Fang, D., Varshney, K.R., Wang, J., Ramamurthy, K.N., Mojsilovic, A. and Bauer, J.H., 2013, December. Quantifying and recommending expertise when new skills emerge. In Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on (pp. 672-679). IEEE.

Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M. and Ronen, I., 2013, May. Mining expertise and interests from social media. In Proceedings of the 22nd international conference on World Wide Web (pp. 515-526). ACM.

Hausknecht, J.P., Rodda, J. and Howard, M.J., 2009. Targeted employee retention: Performance based and job related differences in reported reasons for staying. Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 48(2), pp.269-288.

Horesh, R., Varshney, K.R. and Yi, J., 2016, December. Information retrieval, fusion, completion, and clustering for employee expertise estimation. In Big Data (Big Data), 2016 IEEE International Conference on (pp. 1385-1393). IEEE.
HR Analytics, https://www.indiamart.com/proddetail/hr-analytics-6465687355.html [Date accessed: 15-08-2018]

Hoshen, J. and Kopelman, R., 1976. Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm. Physical Review B, 14(8), p.3438.

Ilgen, D.R. and Hollenbeck, J.R., 1991. The structure of work: Job design and roles (In MD Dunnette & LM Hough (Eds.). Handbook of industrial and organizational psychology (Vol. 2, pp. 165–207). Palo Alto.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

John, A. et al. (2006) ‘Collaborative tagging and expertise in the enterprise’, in Proceedings of Collaborative Web Tagging Workshop Held in Conjunction With Www 2006.

Li, H., Ge, Y., Zhu, H., Xiong, H. and Zhao, H., 2017, August. Prospecting the career development of talents: A survival analysis perspective. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 917-925). ACM.

Lobo, J.M., Jiménez-Valverde, A. and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography, 17(2), pp.145-151.

Martin, R.L., 2014. The rise (and likely fall) of the talent economy. Harvard business review, 92(10), p.16.

Natarajan, N., Dhillon, I.S., Ravikumar, P.K. and Tewari, A., 2013. Learning with noisy labels. In Advances in neural information processing systems (pp. 1196-1204).

Piramuthu, S., 2004. Evaluating feature selection methods for learning in data mining applications. European journal of operational research, 156(2), pp.483-494.

Ramamurthy, K.N., Singh, M., Davis, M., Kevern, J.A., Klein, U. and Peran, M., 2015, November. Identifying employees for re-skilling using an analytics-based approach. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW)(pp. 345-354). IEEE.

Richter, Y., Naveh, Y., Gresh, D.L. and Connors, D.P., 2008. Optimatch: applying constraint programming to workforce management of highly skilled employees. International Journal of Services Operations and Informatics, 3(3-4), pp.258-270.

Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V. and Kambhatla, N., 2010, October. PROSPECT: a system for screening candidates for recruitment. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 659-668). ACM.

Singh, M., Varshney, K.R., Wang, J., Mojsilovic, A., Gill, A.R., Faur, P.I. and Ezry, R., 2012, December. An analytics approach for proactively combating voluntary attrition of employees. In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on (pp. 317-323). IEEE.

Varshney, K.R., Chenthamarakshan, V., Fancher, S.W., Wang, J., Fang, D. and Mojsilović, A., 2014, August. Predicting employee expertise for talent management in the enterprise. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1729-1738). ACM.

Wei, D., Varshney, K.R. and Wagman, M., 2015, June. Optigrow: People Analytics for Job Transfers. In BigData Congress (pp. 535-542).

Xu, Y., Li, Z., Gupta, A., Bugdayci, A. and Bhasin, A., 2014, August. Modeling professional similarity by mining professional career trajectories. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1945-1954). ACM.

Yazinski, S., 2009. Strategies for retaining employees and minimizing turnover. Retrieved from York, NY.

Zhao, Q., Xu, M. and Fränti, P., 2009, April. Sum-of-squares based cluster validity index and significance analysis. In International Conference on Adaptive and Natural Computing Algorithms (pp. 313-322). Springer, Berlin, Heidelberg.