End-to-End Solution with Clustering Method for Attrition Analysis

Nianjun Zhou, Wesley M. Gifford, Junchi Yan, Hongfei Li

IBM T. J. Watson Research Center, 1101 Kitchawan Road
Yorktown Heights, NY, USA
Email: {jzhou, wmgifford}@us.ibm.com, yanjc@cn.ibm.com, liho@us.ibm.com

Abstract— We study a general attrition problem using unsupervised clustering and statistical approaches. The studied problem comes from retention problem in service industries. Our research provides an end-to-end solution from identifying hot job category to analyze the effectiveness of an incentive program applied to the selected categories. One of the barriers of studying the attrition problem is the lack of detailed features of an individual employee due to the confidentiality restriction. Different from the typical attrition approach that requires detailed individual information, we only use the aggregated attrition data and the internal business need data as the base, and cluster the job categories to give a recommendation. We converted the clustering results in a score for the recommendation. To avoid the monthly fluctuation, we apply exponential decay moving average multiple neighboring months on the snapshot scores to ensure consistent recommendation. The end-to-end solution also includes the impact analysis. By comparing the two general groups, we apply an approach similar to A/B test. We score the selected job categories with an effective score.

We can apply this research to large consulting/service companies, and government agencies. For those enterprises or institutes, attrition avoidance is a major consideration as their main assets are their top performance employees. There also exist well-defined job roles and skill categories allowing to us to apply this approach.

Keywords— Clustering analysis, K-means, impact analysis, attrition rate, A/B test, partitioning around medoids.

I. INTRODUCTION

Along with the arrival of a knowledge-based economy, talented person's strategy becomes the source of enterprise core competencies. Enterprise relies on more than ever before on their top performers to innovate and provide better services that differentiate itself from its competitors. In other words, enterprises are reliant upon their human assets to survive and thrive [1].

However, quite a significant number of companies, especially in the service industry, feel that they are struggling with retaining top talents. Voluntary attrition occurs when an employee leaves the company to pursue another career opportunity voluntarily [2]. A high rate of voluntary turnover among top performers could have far-reaching effects across all aspects of an enterprise's business performance. In addition to the downtime or delay of business, a high attrition rate takes more effort to screen, onboard and train backfill employees. The impacted business unit could suffer significant setbacks due to lost institutional knowledge, a decline in team morale, deterioration productivity, and overworked employees absorbing redistributed work left in the wake of a departure [4].

A common practice to keep top performers is to have retention incentives. Retaining an employee with unusually high or unique qualifications is essential. The employee could leave the service in the absence of satisfaction or incentive. The business requires a way to attract and retain these experienced employees with critical growth skills [9]. Typically, the retention payment options [3] are 1) in installments after the completion of specified periods of service during the full-service period (biweekly, monthly, quarterly, or yearly); or 2) as a lump-sum payment after the completion of a full period of service required by a service contract.

However, how to carry out a successful retention program? As a research topic, we need to answer the following questions. First, how to identify the top performers or hot job roles and skills? Second, how effective is the incentive program in term of retention? Finally, whether the attrition avoidance from the program can cover the cost of the stimulus, i.e. is it a valuable investment to its employees? Traditionally, we apply surprised learning approaches to identify the reason that may cause attrition and the effectiveness of incentive programs in achieving retention.

However, with the confidentiality restriction, in many situations, the individual employee information is not available for study. Therefore, we have to invent a new approach. Instead of identifying the reason that may cause attrition, we focus on associating the attrition with the collective attributes, such as the rate of attrition with the job roles and skills using unsupervised approach. The validity of this approach is that the assumption of the cause of attrition coming from the seeking high payment for an experienced employee with hot job skill. This assumption is reasonable if there is a scarcity of highly demanded of certain hot skills. This is typically true for high-tech companies in IT consulting and FinTech areas.

The advantage of this method is that provides an effective solution for attrition prevention based on job characteristics. At the same time, we do not require the typically confidential personal information and make the program execution becomes

978-1-5090-2628-9/16 $31.00 © 2016 IEEE
DOI 10.1109/SCC.2016.54

easier being implemented and provide a relative more fair game to the employees. Another advantage is that this method reduces the influence of personal preference of any individual managers on the decision of incentive payment. The limitation of the approach is that the validity of method relies on our assumption.

We complete our solution in four steps. The first step is to identify the *job roles and skills* needs to pay incentives. The criteria of justification by filtering out those job categories tending to have high attrition or high demand inside enterprise in the future or necessary for project deliveries and success. The second step is to select top performers from those identified categories selected. The third step is to have impact analysis to analyze the effectiveness of the incentive for a given job role and skill category, and finally, the fourth step is to study the cost and benefits trade-off. By combining all the four steps, we can have an end-to-end loop continuously to adjust the program to adapt to the changes in the continuous changing job market and internal and external business environment. Finally, due to confidentiality considerations, we have used synthesized data to illustrate our results and calculations herein.

We arrange this paper as following. In Section II, we discuss objectives of the research and available data. In Section III, we discuss our clustering method to identify candidate hot job categories recommended for the incentive program. Section IV, we carry out an impact analysis to valid the improvement of business performance due to attrition avoidance using a statistical approach. In Section V, we will have a general discussion of a return of investment study for this problem. In Section VI, we discuss related work. Finally, in Section VII, we conclude the paper with a summary and a discussion.

II. PROBLEM FORMULATION AND SYSTEM DEVELOPMENT

We study the problem of reducing voluntary attrition with an incentive program. In this section, we discuss the research requirements and the data used for studies followed by advanced machine learning techniques to identify deep skills in next section.

A. RESEARCH REQUIREMENTS AND APPROACHES

We summarize the research requirements as follows. The following flowchart (Figure 1) outline our end-to-end method to address the incentive program. From the system aspect, we need to ensure this study becomes as an analytic framework for a repeatable, analytics-based and controlled process.

1. Interpretability – The main challenge of applying clustering or prediction is the interpretability of the results. To facilitate business decisions, we need to incorporate the practices, rules, and comments from business practitioners for clustering models. This combination ensures that outcomes from analytics can interpret the business implication of clustering result, and be directly applied to the company decision;

2. Consistency and Stability – To ensure the relative consistency of the outcomes from the modeling while keep using the latest attrition and business data, the model updated monthly to capture the most recent information. We need to consider the outcomes from previous datasets and make incremental changes to keep the stability and sustain the existing business decision for incentive pay to maintain relative stable of the selected job categories;

3. Dynamics – To take into consideration of the trending, we develop models to ensure the separation of seasonality and noise fluctuation patterns over time being minimized from analysis;

4. Adaptive over time – To cope with the seasonality and the dynamicity of labor market, we have to develop an approach adaptive to the change. Therefore, we have to dynamically to allow new data and make the change to the incentive program in term of candidate job role and skill categories. Therefore, the pool of the hot job categories is dynamic. We allow the job role and skill category to join and leave the pool based on the new business analysis.

5. Positive financial outcome - As any business decision, the intention of any retention practice is also looking for a positive financial outcome, but to justify such outcome is not an easy task. The main challenge comes from how to quantifying the results of retention practice financially. Some of the impacts could be long term; some of the impacts might not be measurable using monetary calculation. Such as, in general, an incentive program increases the employees' morale and improve productivity, but very hard to convert this to a solid quantitative analysis.

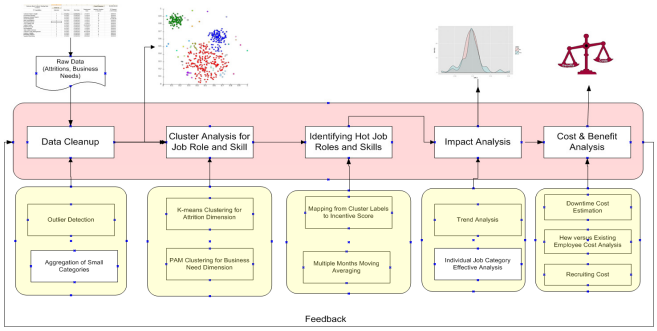


Figure 1. End-End Solution of Incentive Program Analysis

To make an easy alignment with business execution, we developed indicators and metrics to quantify the incentive needs and the outcome of influence attrition behavior. After the initial clustering phase, we assign each job category with following "*very high*" , "*high*" , "*medium*", and "*low*" indictors. We assign the indicators twice for each category; one is for the

attrition dimension and another business need for this category. Later, we convert the indicators as incentive scores used for candidates selection for the incentive program.

Using statistic approach, we evaluate the effectiveness of incentive program in term of attrition reduction. To make this evaluation more scientific sound and robust, we use benchmark comparison and noise reduction to extract the influence on the employee behavior. This evaluation provides a label with values of "highly effective", "moderately effective" or "no effective". The label is used utilized to consider changing the selected category. Finally, a business metrics called cost and benefits ratio is used to quantify the financial benefit of incentive for a given studied job role and skill category. It is a good metrics as a reference for business decision and program adjustment from a financial perspective.

B. RAW DATA AND DEDUCED ATTRIBUTES

Business data is arguably the most valuable asset that an organization owns. With the proper usage of data, we not only can identify the impact of previous business strategies for our clients but also have the potential for improving the business performance of the organization internally.

To make the proposed study possible and complete, we need to have the data for clustering and performance assessment. We have multiple features for each job category. A job category is corresponding to a job role and a seniority level. A typical seniority separation is "*Entry Level*", "*Staff Level*" and "*Senior Level*". We list the raw data as follows.

Table 1. List of Attributes for Job Role and Skill Category Analysis (Monthly)

Attribute Name	Description	Usage
Headcount	Headcount for a given category	
Attrition	Monthly Attrition for a given category	Clustering, Attrition Dimension
Past due	Number of Employee cannot complete their tasks in time	Clustering, Business Need Dimension
Total offer	Total external offer provided in this month	Clustering, Business Need Dimension
Job offer rejected	Number of offers rejected	Clustering, Business Need Dimension
Future Demand	Number of the job required for this position in the following month	Clustering, Business Need Dimension
Incentive Payment	Incentive payment, a zero value means that a category is not selected	Performance Improvement Assessment

We have the monthly collected attrition and business needs metrics. Based on the natures of the attribute features, we separated them two dimensions – attrition dimension and business need dimension. For attrition dimension, we have the following two sets of attributes. The first data set contains monthly headcount and attrition number used to calculate the *attrition rate* that the main feature employed in this study. The second data set contains business-need dimension attributes. We have three metrics with a monthly value of 1) *past due of tasks*; 2) *number of job offer rejected* and is 3) the *future demand* for a given job role and skill category.

The data are time-dependent and consist of several time series variables. The current and previous time snapshots become the base to identify the candidate categories for the incentive. The historical horizon of attrition rate deduce from headcount and attrition is used for an impact analysis of the effectiveness of the program in selected categories. Through comparing the different behaviors of categories with and without incentive, we can have a result in a similar fashion with an *A/B test*.

Finally, we have the monthly incentive payment for each selected job role and skill category. It provides the base for us to study the cost of the program. The dataset used for cost and benefits analysis – this is used to assess the cost and benefit of the program.

III. IDENTIFYING HOT JOB SKILLS

In this section, we use clustering methods to identify hot *job role and skill* (JRS) categories. We first use clustering analysis to identify candidates using snapshots of attrition and business data. Then, a scoring method combining multiple snapshots is used to give a final recommendation. Managers in those selected categories can take further action on incentive payment for the employee based on individual performance.

A. CLUSTERING PROCESS

Cluster analysis is an unsupervised machine learning method. As there are only the business features at the category level, clustering is a more appreciated method. The outcomes rely on us to provide semantic interpretation. As a data-reduction technique, we use clustering to uncover subgroups of job categories within our dataset. We reduce a large number of categories (observations) to a much smaller number of clusters. Applying clustering twice, we can have clusters with data from attrition and business dimension.

With the assumption that the job categories of identified clusters more similar to each other than they are to the job categories in other groups. We accomplish the task of inferring a function to describe hidden structure from unlabeled data. Then, we have to assign the business interpretation into the unsupervised learning outcomes.

Table 2. Deduced Attributes for Clustering Analysis

Deduced Attribute	Formula	Clustering Method
Attrition Rate	Attrition/Headcount	K-Means
Reject Rate	(Offer reject)/(Total offer)	PAM
Past Due Rate	(Past Due)/Headcount	PAM
Future Demand Rate	Future Demand/(Headcount)	PAM

The clustering method starts from choosing appropriate attributes. We only have limited attributes preprocessed and selected by the domain experts. To avoid the variables with the largest range have the greatest impact on the results. This is often undesirable, and analysts scale the data before continuing. The most popular approach is to standardize each variable to a mean of 0 and a standard deviation of 1. In our study, we use the deduced attributes as shown in Table 2. Those different rates consider the needs of normalization.

Most clustering techniques are sensitive to outliers, distorting the clusters obtained. With the rates as feature attributes, we need the attribute values to be reliable and stable. In our case, we use the *headcount* as the denominator for three rates (Table 2). Therefore, the rates are sensitive to the headcount, especial for those job categories with small values of headcount. To minimize the possible impact of outliers, we first only select those job categories with headcount larger than a threshold e.g. 30. Then, we aggregate the small job categories together to large categories based on their similarity.

From technique side, to obtain a final cluster solution, we need to decide how many clusters are present in the data. Once the number of clusters has been chosen, we perform the clustering to extract that number of subgroups and to interpret the clusters.

B. ANALYSIS OF ATTRITION DIMENSION

We apply the *K-means* clustering on attrition rates (number of attritions relative to job headcount) to determine appropriate segmentation and number of segments. The number of clusters is specified before clustering. We loop over a different number of outcome clusters. The output contains the boundaries and the number of job categories in each cluster. From the technique perceptive, the principle of selecting number cluster is to see the value of *within groups of sum squares* being reduced and stabled. Typically, the final number of clusters is among five to eight. In the following illustrated case, the first stable cluster number is the value of six. Therefore, we can choose any number that is >= 6.

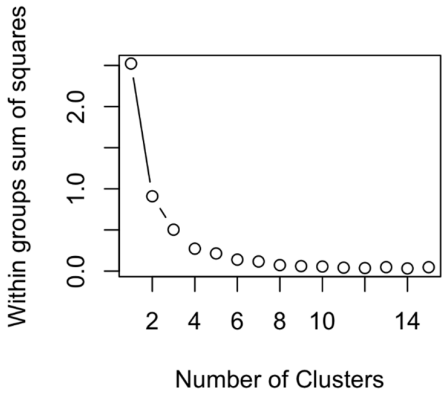


Figure 2. Sum of Squares of the job role and skill (JRS) category Clustering within Group

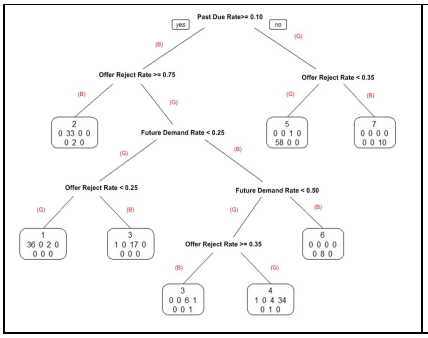
Our final goal is to regroup the clusters into four segments ("*very high*", "*high*", "*medium*", and "*low*"). At this time, we need to utilize some business rules to help to make the final mapping. In this case, an example business rule is that 15% attrition rate is upper boundary of the *medium*, and 25% is the lower boundary of the *very high*. Then, we have to select the value of the cluster allowing us to have those conditions satisfied, i.e. the boundaries of clusters align with the criteria specified by business rule. When all the boundaries defined for the final segments, all the clusters within two constrained lower and the upper boundary will have the same label (Table 3).

Table 3. Illustrated Example of Selected Segments

Classification	Minimum Attrition	Maximum Attrition	Upper Boundary	# of Categories
L	0.00	0.09	0.10	75
M	0.10	0.14	0.15	100
H	0.15	0.24	0.25	25
VH	0.25	0.50	1.00	5

C. ANALYSIS OF BUSINESS DIMENSION

The clustering method for the business need dimension is PAM "*Partitioning Around Medoids*" with outcome partitioning as a decision tree. Different from a Euclidean distance used in *K-means*, *PAM* uses Manhattan distance as the cost of clustering.



Predicted Cluster	Proposed Label
1	L
2	VH
3	H
4	M
5	L
6	VH
7	M

Figure 3. Illustrative Example of Business-Need Dimension Clustering

Similar to the attrition rate dimension, we need to the final clusters from the PAM with "*very high*", "*high*", "*medium*" and "*low*" labels. We used a heuristic approach to label the cluster. Therefore, it looks a little tricky in above Figure 3. We first annotate each edge of the cluster tree with a symbol ('*G*' or '*B*'). We notice that there is a business condition at each node. If the condition is favorable to the company, we annotate the left edge as '*G*', and right edge as '*B*'. In the context of our paper, we consider that lower values of *past due rate*, *offer reject rate*, and

future demand rate mean better for the business of the studied enterprise.

The generated clusters are the leaf nodes of the clustering tree. Those clusters with all or majority edges having "*B*" of their paths are assigned to "*VH*" group, those clusters with all or majority edges having "*G*" are labeled as "L". If only a few edges have "*B*", then the cluster is assigned to "*H*" group, and all the remaining are assigned to "*M*" group.

D. SCORING AND INCENTIVE JOB ROLE AND SKILL CATEGORY SELECTION

With the completion of labeling the job categories, we need to combine the two labels into a numerical score. We use the mapping in Table 4 to provide such conversion. When repeatedly applying this method to other months, we will get a series of scores for a given category.

Table 4. Mapping from Label into Incentive Score

	Business Need			
Attrition	**L**	**M**	**H**	**VH**
VH	30	50	70	80
H	20	40	60	70
M	10	30	40	50
L	10	10	20	30

As we said before, we want to maintain certain consistent and stability of the suggested candidates for job categories. To achieve that, we apply a time moving average to incorporate that impact of previous months into the final suggestion. For example, we use exponential moving average to give the final score with impact from the previous months score exponentially decay over time. If we only choose six months, then, the final score will be:

$$s = \frac{(1-a)\sum_{n=1}^{6} a^{n-1} s(n)}{(1-a^7)} \qquad (1)$$

We provide the final recommendation based on the final scores and recommend those categories with higher (say, a cutoff $s >= 50$). Using the calculation of (1), we can ensure a relative consistency of the recommendation while allows us to utilize the latest attrition and business data.

With the hybrid model, we minimize the impact of the monthly fluctuation, and only capture the consistent attrition signal and put into the final recommendation.

IV. IMPACT ANALYSIS

We carry out the impact analysis to answer two questions for the incentive program after its execution. At the macro level, we want to know whether the incentive program is effective and have attrition avoidance. At the micro level, if the program overall is effective, we want to know whether it is effective in a particular job category.

A. TREND ANALYSIS

Typically, the common practice to identify a method (i.e. incentive program) is effective or not is to use A/B test. From the definition of Wikipedia, A/B test is a way to compare two versions of a single studied variable (i.e. attrition rate) by testing a subject's response to experiment (i.e. employees' response to incentive program). However, in this application, we cannot use A/B test, as it will create inequality to the employees.

Fortunately, the incentive program only applies to some selected categories. Therefore, we can separate the non-selected into *Group A* as the base, and those selected for incentive program as *Group B*. That gives us an opportunity to compare the effectiveness the program similar to an *A/B test*. For *A/B test*, when the two groups (A and B) are compared, they should be identical except for one variation (incentive) affecting the employee's behavior. *Group A* plays as the *control group* as *A/B test* without incentive, while group B played as *treatment group* (with incentive). We assume that all other factors have the same influences on *job roles and skills* categories. To extract impact of the incentive program, we need to filter out seasonality and external labor market influence. The comparison is between the changes of attrition rates of two periods – before and after the program implementation.

Mathematically, we can convert into the trend analysis of the difference between the attrition rate at a different time to the difference between the two groups.

$$[a(t_2) - a(t_1)] \; versus \; [b(t_2) - b(t_1)]$$
$$<=>$$
$$[a(t_1) - b(t_1)] \; versus \; [a(t_2) - b(t_2)] \qquad (2)$$

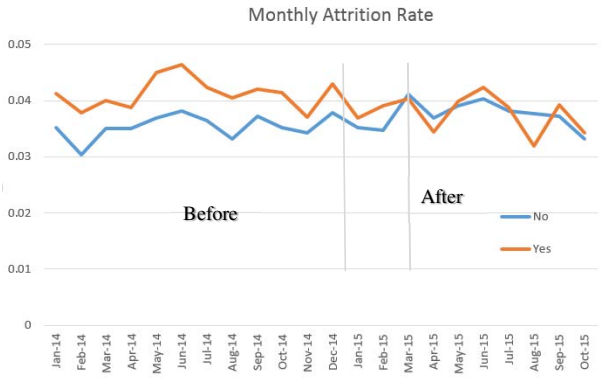


Figure 4. Trend Analysis of Program Effectiveness

To illustrate this analysis, from the chart above, we can see an obvious narrowing of the difference in the attrition rate after

incentive program. The time between the two bars is a transition period introduced as considering the warm-up of the project. With the trend analysis given, we can have the attrition avoidance as:

$$m = n(t_2)\left(b(t_2)\frac{a(t_1)}{b(t_1)} - a(t_2)\right)\frac{T_2}{12} \quad (3)$$

As a short conclusion, we use the above method that first filters out other impacts, such as the market fluctuation and other micro factors influencing employee performance, with the selection the attrition difference of same period of two consecutive years and averaging over multiple months. Then we identify the attrition avoidance performance through a method similar to A/B test approach, which helps us to get the numerical value of the attrition avoidance as shown as follows.

Table 5. Illustration Example of Attrition Avoidance

Averaged Attrition rate (Normalized as yearly rate)	Incentive Group (B)	Non-Incentive Group (A)	Difference
January – December 2014 (Before Incentive Program)	30.00%	20.00 %	10.00%
April – September 2015 (After Incentive Program)	25.00%	25.20%	-0.20%
Total Headcount in September 2015	10,000	20,000	

Using the Estimated attrition counts avoided in 6 months (April – September 2015) are:

$$m = 10{,}000\left(25.20\%\frac{30.00\%}{20.00\%} - 25.00\%\right)\frac{6}{12} = 640$$

If we want further to distinguish the effectiveness of on the attrition improvement on the seniority factors. We can apply the same method of above but only selected those categories belongs to same seniority (say, junior, staff, or senior).

B. ANALYSIS FOR INDIVIDUAL JOB CATEGORY

At the micro level, we will use a similar idea as before by comparing the two groups of with/without the incentive program. Again, taking into consideration of the seasonality of attritions, we compare two period times of two consecutive years before and after the incentive program. Considering the fluctuation of the attrition, we first average the monthly attritions together and normalized into yearly attrition, and create the two yearly attrition rates before and after incentive program execution. The difference, say *c*, of two attrition rates, becomes the studied variable.

As a benchmark, we first select all the job categories without receiving any incentive, then calculate the difference between attrition rates for each category. From the Q-Q chart of the random variable *C* (attrition rate difference), we can conclude that random variable *C* follows a Gaussian normal distribution.

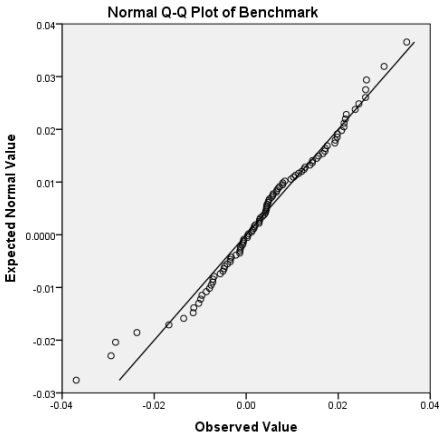


Figure 5. Q-Q Chart of the Attrition Rate Difference (Categories without incentive – Base)

Using this probability distribution as the base, we can justify the effectiveness of incentive program for a selected JRS category. In statistics, the standard score is an observation above the mean. A positive standard score indicates a datum above the average. As the base follows the normal distribution, we can use one-side Z-score. For a selected Z-critical value, we have the area under the normal curve (a percentage or quantile). In this study, we define two quantiles, and correspond Z-scores.

Table 6. Criteria to Define Improvement

Percentile	Z-Score	Label
[0.80,1.0]	[0.84,+∞)	Significant Improvement
[0.60,0.80]	[0.25,0.84)	Moderate Improvement
[0,0.6]	(-∞,0.25)	Not effective

Based on the percentile (or Z-score) of each job category of *group B*, if its percentile is larger or equal to 80%, we will assume that there is a significant improvement in this job category. If the percentile falls from 60% to 80%, then we will say that the improvement in this job category is marginal.

The histograms (Figure 6) of the distributions of attrition rate difference shows normality centered around zero. The distribution of job category with incentive shows heavier tails that indicate more likely larger differences of attrition rates to showing the improvement in some job categories. In this chart, the *"incentive=yes"* refers the attrition rate difference distribution of the job categories included into the incentive program, and "*incentive=no*" are those labeled/suggested as "no incentive".

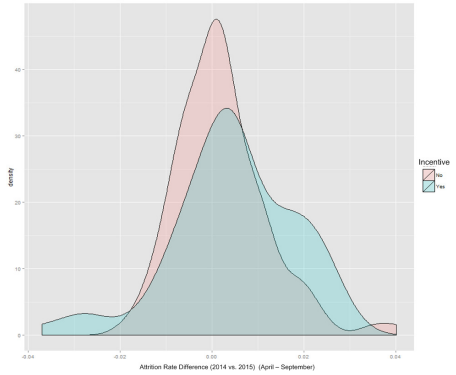


Figure 6. Distributions of *c* of all job role and skill categories with incentive (yes) and without (no)

V. COST AND BENEFIT ANALYSIS

Cost and benefit are the last steps in our end-to-end analysis. Ideally, if we can find a simple positive ROI ratio, then we can clearly claim a victor of an incentive program. Furthermore, we can use this ROI calculation to help us estimate the kind of savings we can expect to get through implementing a strategic incentive program.

However, although we can easily obtain the cost of the incentive payment from accounting practice, the benefit is not so simple to estimate. The benefit comes from the cost saving from the attrition avoidance. Therefore, the challenge becomes how accurately to estimate the damage of attrition. The main reason is that there is no simple quantitative method of estimating negative impact of attrition to the enterprise.

In this section, we focus on the discussion of three main factors to contribute the benefits calculation. They are 1) performance deterioration due to staff downtime, 2) recruiting cost including interview and advertisement, and 3) incurred expense due to the internal and external salary difference.

A. BENEFIT FROM DOWNTIME AVOIDANCE

The deterioration performance roots from the downtime after an employee leaves the company. It typically takes six to eight weeks for a replacement being able to pick up the leftover tasks. It usually takes even longer time to replace a senior position.

We can estimate the monetary loss from the contribution of a similar employee during the downtime period. By assuming contribution is proportional to overall payment (including salary and employee benefits), we can estimate the final impact of performance deterioration as payment multiplied by an adjusted factor. The factor considers a) the lack of institute knowledge of the replacement and b) the ratio of the contrition versus the payment. Finally, we have:

$$c = k\frac{d}{56}\,s \qquad (4)$$

Here, k is the adjusted factor, which is a number larger than 1. The exact number depends on the nature of the position replaced and the industry. For typical retail industry, the value could be slightly higher than *1*, but for more sophisticated industry or management position requiring specific skills, it could become much higher. The s is the overall yearly payment, and the d is the downtime measured in weeks.

B. OTHER SAVING FROM ATTRITION AVOIDANCE

Another obvious factor needed to consider is the saving from recruiting cost from attrition avoidance. The cost includes a) the headhunting agency costs; b) advertising costs, and c) interview and moving expenses. The search and firm fees can account up to 35% of a new employee's annual base salary. The advertising cost is a significant cost reduction by placing all job positions in the enterprise career.

Finally, for growth job role and skills in technical or engineering fields, the external market value increases faster than enterprise salary levels, the enterprise must offer more to recruit such candidates. The difference is another factor to calculate the benefit of attrition avoidance.

VI. RELATED WORK

Voluntary Attrition belongs to a typical attrition problem in human resource management. Attrition itself has been well studied in the past [5] [6]. The main application of attrition is the client service termination in service industries, likes telecom, entertainment, financial and retail. The attrition problem itself belongs to a predictive problem with a dichotomous outcome (0 or 1).

Mathematically, two widely approaches used for dichotomous outcome predication are a) survival analysis model [5] or more advanced approach like Hawkes point process model [11] [13], and b) temporal horizontal predictive models, such as neural network, logistic regression [12], more advanced classification techniques such as incremental dictionary learning [14]. The survival and Hawkes point process model predicts the likelihoods of attrition (or success) in next period with a given duration for a particular client. Different from survival model, we can use Hawkes point process model to incorporate the influences of external activities on the outcome. For example, Yan et al. used two Hawkes point process model to capture the activities of sell person to predict the success of sale [11], and the impacts of disastrous weather events of oil pipes [13].

While, for time horizontal predictive model, it predicts the attrition (or success) in next period at a given time. Therefore, survival analysis model considers the whole life cycle of a

client (or employee). A prediction model considers the likelihood of attrition at a given time snapshot.

In addition to analyzing the attrition using classification technique, Cung et al. utilized a clustering technique called spectral clustering for voluntary attrition problem [8]. Different from *k-means* or *PAM* based on the object distances, spectral clustering utilizes the spectrum of the data similarity matrix to perform this separation. Their study focuses on identifying a set of employees those who are likely voluntarily to leave the company from those who are not.

However, in this paper, we are facing a different problem. Partially due to the confidential consideration, we do not have the attrition instance at the individual level as other attrition analytics problem. Therefore, we cannot apply the classification approaches as there is a lack of prediction features (either a job role or a skill level). Instead of focusing on attrition on an individual level, we have to leverage the unsupervised clustering approach at job category level, i.e. to extract those job categories with the tendency of higher attrition rate or requiring specific retention consideration due to internal business needs. Considering this, we have chosen attrition rate and three business need dimensions as the features for clustering. More important, we have extended the research into impact analysis that provides the meaningful business answer to a retention program.

Finally, our effort aligns with the general trends of applying business intelligence to solve business management problem, and more specific the human resource related problems. For example, researchers are applying more advanced supervised machine learning techniques [10] to business attrition and other business problems.

VII. CONCLUSION

With the help of machine learning and statistical techniques, in this paper, we study a very general business problem that can expand this into a more complex and profound problem. The developed framework makes targeted, variable incentive investments in hot skill employees to drive retention / attract talent and grow critical skills. Subsequently, we can continuously develop this advanced analytical method and expand its focus to the job market in both growth and mature markets and executive capabilities in different service industries.

Different from majority approaches using predictive analysis of identifying key factors in term of voluntary attrition based on the dataset with individual employee information, this study only utilizes the aggregated group (job category) information. It has the advantage of better applicability in term of privacy protection and the fairness of allocating variable payment. We achieve the fairness by separating the decision of providing allowance into two steps. We only involve the managers in the last step for the allowance allocation. Therefore, it might be more applicable in term of applying to real business problems. The weakness of this approach is that it cannot reveal the detailed causality of attrition. Therefore, it cannot provide insights how to avoid high attrition if the primary reason for voluntary leave is not just monetary payment.

The contribution of the paper comes from a complete end-to-end solution to identify candidate categories of studied job roles and skills, the assessment of the effective of an incentive mechanism of targeted categories, and finally the business outcome measured with financial metrics. From a system perspective, we developed an integrated solution allowing dynamically analyzing and monitoring the business performance. As there are well-defined job roles, and the structured hierarchy of seniority in retail and IT industries, and government agencies, we can apply this method to those institutes where the major assets are their top performance employees.

REFERENCES

[1] M. Subramony and B. C. Holtom, “The Long-Term Influence of Service Employee Attrition on Customer Outcomes and Profits”, Journal of Service Research, 15(4) 460-473, 2012.

[2] W. F. Cascio, “The High Cost of Low Wages”, Harvard Business Review December 2006 Issue

[3] WorldatWork and Deloitte Consulting LLP, “Incentive Pay Practices Survey: Publicly Traded Companies”, February 2014.

[4] S. Daniels, “Retaining a Workforce That Wants to Quit”, Harvard Business Review, July 07, 2010.

[5] D. Poel, and B. Larivière, “Customer attrition analysis for financial services using proportional hazard models”, European Journal of Operational Research, Volume 157, Issue 1, 16 August 2004, Pages 196–217.

[6] A. Adhikari, “Factors Affecting Employee Attrition: A Multiple Regression Approach”, The IUP Journal of Management Research, May 2009.

[7] M. J. Somers, “Thinking differently: Assessing nonlinearities in the relationship between work attitudes and job performance using a Bayesian neural network”, Journal of Occupational and Organizational Psychology, Volume 74, Issue 1, pages 47–61, March 2001.

[8] Cung, B., Jin, T., Ramirez, J., Thompson, A., Boutsidis, C., Needell, D., "Spectral Clustering: An empirical study of Approximation Algorithms and its Application to the Attrition Problem", SIAM, Research Journal, vol. 5, pp. 283-303, 2012.

[9] J. Bhatnagar, “Talent management strategy of employee engagement in Indian ITES employees: key to retention”, Employee relations 29 (6), 640-663.

[10] Jatinder N. D. Gupta (Author, Editor), Kate A. Smith (Editor), “Neural Networks in Business: Techniques and Applications” Hardcover, July 1, 2001.

[11] J. Yan, C. Zhang, H. Zha, M. Gong, C. Sun, J. Huang, S. Chu, X. Yang, "On Machine Learning towards Predictive Sales Pipeline Analytics", AAAI 2015

[12] J. Yan, M. Gong, C. Sun, J. Huang, S. Chu, "Sales pipeline win propensity prediction: a regression approach", Integrated IFIP/IEEE Network Management (IM), 2015

[13] J. Yan, Y. Wang, K. Zhou, J. Huang, C. Tian, H. Zha, W. Dong, "Towards effective prioritizing water pipe replacement and rehabilitation", IJCAI 2013

[14] J. Yan, C. Tian, J. Huang, F. Albertao, "Incremental dictionary learning for fault detection with applications to oil pipeline leakage detection", Electronics letters 47(21), 1198--1199, IET, 2011