PROPENSITY MODELING FOR EMPLOYEE RE-SKILLING

Moninder Singh, Karthikeyan Natesan Ramamurthy

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA.

Shrihari Vasudevan

IBM Research
Bangalore, KA 560045, India

ABSTRACT

Due to the rapidly changing, dynamic nature of today's economic landscape, organizations are often engaged in a continuous exercise of matching their workforce with the changing needs of the marketplace. Re-skilling offers these enterprises the ability to effectively manage and retain talent, while also satisfying business requirements. We describe an analytics-based propensity scoring model for re-skilling by combining historical employee job-role/skill records, relationships between different job-roles/skills, employee resumes, and job postings. This is used to determine the source features that are the closest to a required target skill and hence identify employees that can be easily trained for the target skill. We evaluate this approach for a representative set of target skills at a multinational with a large services/consulting arm. We show that the propensity model learnt from the combined data sources has a high accuracy that is also substantially better than that achieved by using features from job-roles or resumes alone. The performance is improved further by using an ensemble model to evaluate the propensity scores.

***Index Terms*—** HCM, re-skilling, skill adjacency, workforce analytics, resume parsing

1. INTRODUCTION

With rapidly changing products, technologies, and platforms in the market, organizations today face a perpetual challenge in maintaining a workforce that can not only meet current needs but also satisfy future demands. However, the skills needed in the future may be vastly different from those needed currently. This is especially true for companies with large consulting and/or service arms. For example, as clients shift more of their computing infrastructure from the desktop to the cloud, organizations that serve them may find that they have a shortage of skills needed in the future (e.g. those related to cloud technologies) while facing a looming glut of legacy skills that are currently needed but for which there is declining demand (e.g. desktop application developers).

One way for a company to address this skill gap is by hiring people with the new skills[1], while letting go employees with legacy skills. However, this may not be the best way of handling it. Not only will it likely be expensive and inefficient, with market demand for hot skills probably outstripping their availability, the organization also risks losing otherwise valuable talent. Re-skilling employees with expertise in skills with declining demand to hot skills expected to be in growing demand offers a much better alternative.

The problem then is to identify the employees to be retrained. Apart from the logistics involved, such as balancing current need of diminishing-demand skills against expected demand of newer skills, number of people available and willing to be retrained, as well as business constraints, an important consideration is the "ease" with which an employee with one skill can be re-skilled to another.

Analytics-based solutions are commonly used in the industry to identify people with a certain skill-set, both within an enterprise as well as from amongst a pool of potential employees. Resume parsing and skill-extraction solutions are commonly deployed within HR systems. For example, in order to determine if someone is suitable for the position of an *iOS Application Developer*, such systems will extract skills from resumes and match against the skill-set required for such a position, such as 'iOS', 'Swift', 'mobile applications', etc. Other solutions use similar techniques to optimally match people to positions while incorporating business constraints [1], or identify such people based on statistically profiling and scoring needed skill-sets [2]. However, such systems cannot be used to identify people who do not possess the required basic skills, but could potentially do the job, with little or no training. In the above example, could the job be done by an A*ndroid Application Developer* or a *Java Application Architect*, or even a *Windows Database Developer*, and is one of these more suited than the other?

Ramamurthy *et al.* [3] addressed this issue by considering historical employee job-roles, and developing an adjacency model for skills as a weighted-combination of other skills. Using our example, the underlying hypothesis of this

[1] Note that when looking to re-skill or hire an employee, an organization typically thinks in terms of the 'job' or 'job role' that the individual is needed to perform (e.g. *Big Data Architect*) with a set of skills/experiences, rather than a single basic skill (e.g. *Java*) she should have. In this paper, we use the terms 'skill(s)', 'skill set', 'job', and 'job role' interchangeably to refer to an integrated set of skills and experiences necessary to perform a certain job role. A single basic skill, such as *Java*, will be explicitly referred to as such.

approach is that if many employees have *iOS Application Developer* and *Java Application Architect* in their job roles history, then it is likely that a *Java Application Architect* could be re-skilled to an *iOS Application Developer*. While this approach works well in practice[3], it has several weaknesses (as discussed in Section 2.1), such as the inability to infer a signal from, or apply the model to, an employee with no prior job-role history, such as a recently-hired employee.

In this paper, we propose to combine job-role based features with features extracted from resumes (using knowledge from job-role descriptions and job postings) to estimate the propensity of an employee to be re-trained in a new skill. We show that including the additional text-features provides a measurable improvement in the accuracy of the propensity model over just using skills from the jobs taxonomy. We use sparse logistic regression as our base propensity model, and demonstrate additional improvements in performance using ensemble learning.

2. DATA AND METHODOLOGY

We first describe the various data sources used along with our rationale for using them. Then, we describe how we use features extracted from such data sources to model an employee's propensity for re-skilling to a given target skill.

2.1. Data Sources and Feature Creation

The first data source we use is the **job-role taxonomy** into which organizations typically arrange the various job roles that are needed around the enterprise. Such a taxonomy describes job categories and job roles at the top, while gradually breaking each such job category/role into the skill sets and individual skills needed to satisfactorily perform such jobs at lower levels of the taxonomy. For example, job categories such as *IT Architect* and *Consultant* may be used to describe broad segments of work an employee is expected to perform, and are typically consistent with job categories recognized externally. Each job category may then cover several different job roles, each describing an integrated cluster of work responsibilities and activities that must be performed by an employee. For example, *Application Architect* and *Infrastructure Architect* can be two different job roles that belong to an *IT Architect* job category. While each job role shares the trait of the job category, different skill sets are needed to perform each job role, and each job role may be further broken down into more specific jobs (e.g. *Big Data Architect* and *Mobile Architect*, both being specific types of Application Architects). As mentioned earlier, this is the level where hiring/re-skilling decisions are typically made, and the level that is the focus of this paper. The lowest level of such a taxonomy will describe the basic skills and responsibilities. Each item in the taxonomy is typically accompanied by a concise description, which may or may not be very informative. However, taken together, descriptions for the various levels in the taxonomy for a particular job role can provide substantial information about the skills and experiences an employee needs to possess in order to perform that job. For example, in the taxonomy for the organization we consider in this paper, the basic skills needed for a *Big Data Architect* are simply listed as "skills to perform Big Data Architect role". However, the job-role itself is described as someone who "designs business intelligence solutions...understands data lakes and Hadoop framework,.. has one or more of these skills: HDFS, HBase, ...".

The second source of data available typically is the human-resource data that can provide **job-role history** of an employee at the organization during her career. This data, combined with the skills taxonomy, can help identify the evolution of skills amongst the employees of an organization, and, as discussed in Section 1, can help identify the sets of skills/job-roles that can make it easier for an employee to transition to a new job-role of interest [3]. However, there are several potential challenges in using this data by itself. First, no signal can be inferred for a new employee, or an existing employee with very limited job-role history. Second, there is a lot of inherent noise in the data since these skills are often self-reported; not every role in the HR data will be a role the employee has actually performed, but will also include roles the employee feels she can perform. Third, an employee may have basic skills that are not reflected in her job role history but may still be strongly indicative of what other job roles that employee can perform, perhaps with little training. Continuing our example, it may well be that *iOS Application Developers* are typically proficient in Java programming and Windows application development. In that scenario, employees having these basic skills could be potential candidates for re-skilling to *iOS Application Developers*, regardless of their actual job-role histories.

The third source of data, **employee resumes**, may list such basic skills and experiences that are not reflected in the job-role history of an employee, but can help indicate what other jobs that employee could potentially due with little or no training. However, unlike an organization's skills/job-role taxonomy that is often carefully curated, resumes can be very noisy. A resume will typically contain extensive information about skills, experiences, projects, etc. that the employee wants to highlight, much of which may not be relevant to the employee's ability to perform the job-role of interest. More importantly, the same information may be expressed in many different ways, using different styles and concepts and may vary substantially from employee to employee. For example, consider the case of the organization where we applied the proposed approach. All resumes followed a common template, with sections for employment history, previous jobs, and skills, etc. We specifically focused on the section pertaining to 'skills'. Nevertheless, the specific content in this section varied significantly across the employee base. While some employees provided a succinct list of basic skills (e.g.

”Java/J2EE, JSP, ATG Commerce, ...”), others provided more general write-ups of experience and projects undertaken (e.g. ”...Strong team building skills and management of cultural change process generated by experience in different cultures...”). Moreover, the amount of detail varied significantly from resume to resume. In order to effectively use this data, it is imperative to accurately identify and extract concepts that are relevant to the job roles of interest.

The fourth data source consists of an organization’s **external job listings**. Unlike a taxonomy description that typically used organization-specific terminology, job listings usually describe the job in externally accepted language, and also provide more detail about the job than is typically provided by a job role description. For a given job-role of interest, it is thus possible to use these three data sources to identify basic skills and experiences that employees with that job-role have. Then, as in the case of job-role histories, resumes of other employees can be scoured for such basic skills and experiences and used to evaluate their suitability for re-skilling to the job in question.

As such, we propose the following methodology to construct text-based features:

First, concepts are extracted from the three text-data sources to build a common library of concepts.

Second, this common library of concepts is ‘normalized’ so that different ways of referring to the same concept across the various sources are all consolidated into one concept.

Third, for each data source separately, concepts are extracted using this normalized concept library as follows:

(i) For each job-role, we look at the corresponding taxonomy descriptions for the job-role and corresponding specific skills, and extract all available concepts for that job-role

(ii) Similarly, for each job-role, we look at all available, distinct job-postings and similarly extract concepts used for that job-role

(iii) Finally, for each job-role, we combine resumes of all employees with that role, and similarly extract concepts

Fourth, for each data source separately, we use tf-idf wighting [4] to identify important concepts for each job-role for the corresponding data source.

Fifth, the union of all these concepts is now used to create the set of binary, text-based features.

Essentially, for each job-role, we look for concepts that are important to descriptions or job postings, or are important to people while writing their resumes. Everything else is redundant and is discarded.

2.2. Propensity to Acquire the Target Skill

Our goal in this work is to identify the employees best suited for re-skilling by combining information from the multiple data sources listed above. We refer to a new skill under consideration as the *target skill*. Our hypothesis is that for a given target skill, we can score the employees based on their *source features*. This score indicates the relative ease with which anyone else who has the same source features can transition to that target skill. As described previously, the source features for an employee consist of her historical skills acquired in the organization, as well as *text-based features* extracted from her resume. Given these *source features*, we model the propensity of the employee to acquire the target skill.

Let us assume that the number of employees is N. The source features for the employees are represented in the matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$. $\mathbf{x}_i$ is the row i of $\mathbf{X}$ and x_{ij} is the element (i, j). The number of historical skills is P_s, the number of text-based features is P_t, and hence $P = P_s + P_t$. We also create a binary target vector $\mathbf{y} \in \mathbb{R}^N$, which denotes the presence and absence of the target skill for the employee. The propensity to acquire the target skill given the source features for each employee is given by the conditional probability $P(Y_i = 1 | X_i = \mathbf{x}_i; \boldsymbol{\theta})$. Y_i is the random variable indicating presence or absence of the target, and X_i is the random vector for the source feature set. $\boldsymbol{\theta}$ is the parameter vector of the propensity model. We use the logistic function to model this conditional probability,

$$\pi(\mathbf{x}; \alpha, \boldsymbol{\beta}) = \frac{\exp\left(\alpha + \sum_{j=1}^{P} \beta_j x_{ij}\right)}{1 + \exp\left(\alpha + \sum_{j=1}^{P} \beta_j x_{ij}\right)}. \qquad (1)$$

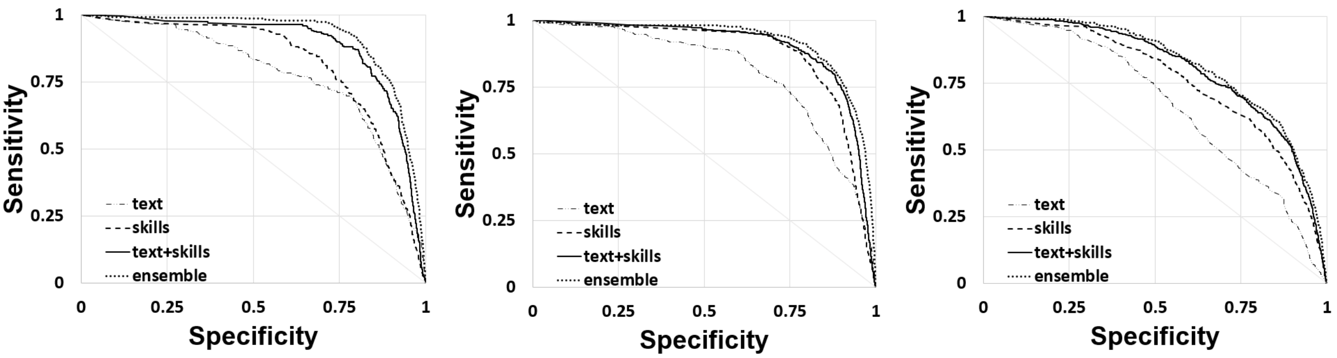
Estimation of the parameter vector $\boldsymbol{\beta}$ can be carried out using standard approaches [5]. Additionally, we impose sparsity regularization on $\boldsymbol{\beta}$ and hence our base propensity model is sparse logistic regression [6].

Since the number of source skills and the text-based features can be numerous, we perform several levels of filtering to remove irrelevant features from the source feature set before constructing the propensity model. One such screening approach [7] used for the source skills is described in [3].

In order to improve the accuracy of the propensity model, we propose to use ensemble learning approaches such as bootstrap aggregation [8], *Adaboost* [9], and a version of subsample aggregating (*subagging*) [10] that uses all positive examples and an equal number of random negative examples sampled without replacement in each iteration. The number of iterations in subagging is equal to the number of negative examples divided by the number of positives, so that each negative example is used in at most one iteration.

3. RESULTS AND DISCUSSION

We developed propensity models for re-skilling to three target skills: *Big Data Architect*, *Mobile Architect*, and *Cloud Java Developer*. The initial number of employees considered were over $120,000$ and the number of historical skills considered



(a) Big data architect

(b) Mobile architect

(c) Cloud Java developer

Fig. 1. ROC curves obtained from propensity models for three target skills with $5-$fold cross-validation.

were over 650. IBM SPSS Text Analytics was used to extract relevant concepts from the text data sources. Then, we performed a coarse normalization of the concept library manually, and used tf-df weighting to select the most relevant concepts for each job-role for the three data sources.The union of these concepts was then used to create text-features for each employee, based on whether a concept was mentioned in the resume or not. For example, concepts extracted for *Mobile Architect* included mobile, worklight, swift, xcode and objective C, while concepts extracted for *Cloud Java Developer* included Java, javascript, aws, bluemix, and web services. We also used the feature screening approach proposed in [3] to make the propensity models robust, and removed employees who did not have any source features turned on. This reduced the number of source features to less than 200 and the number of employees considered to about $47,000$ on an average. We note that the number of people who had each of the target skills was only a few hundred, resulting in a high class imbalance.

We used text-based features only, job-role features only, as well as their combination, to train the base learner (sparse logistic regression) as well as the ensemble learners. We used $5-$fold cross validation to evaluate the performance of the proposed models with the held-out data. The ROC curves and the area under ROC curves are provided in Figure 1 and Table 1 respectively. Note that we only present the results for the base learner and subagging ensemble learner. The text-based features when used alone have the worst performance, possibly because they can be more generic compared to historical skills. Moreover, this data is the noisiest, both due to the free text nature of resumes as well as the amount of irrelevant information that may be present in them. Using skills-based features resulted in better performance since the data is much less noisier. However, when combined with the text-based skills, they provided even better performance compared to using either of them separately. As discussed in Section 2.1, this is due to the complementary nature of the information typically provided by these two sources of information.

Table 1. Area under the ROC curve for propensity models with various target skills.

Target Skill	Text	Skills	Text +Skills	Text+Skills Ensemble
Big Data Architect	0.778	0.821	0.888	0.917
Mobile Architect	0.806	0.883	0.902	0.920
Cloud Java Developer	0.667	0.760	0.803	0.817

Finally, we observed that ensembling enhances the accuracy by a few percentage points. Since the employee data can only provide a noisy approximation to the underlying relationship between the source features and the target skill, using an ensemble of models is a natural way to reduce the prediction error. Results from only subagging ensemble is reported since it has the highest accuracy among the three ensemble learners tested. This could be because it mitigates the class imbalance in the data by using equal number of positives and negatives in each iteration.

4. CONCLUSION

We developed propensity models for identifying employees that can be trained to a new target skill. The features used in our learning algorithms included the historical skill records as well as text-based features extracted from employee resumes. We demonstrated that incorporating additional informative features as well as using ensembles of learners improved the accuracy of the propensity models. Future work involves investigating more sophisticated feature extraction approaches from the text data, as well as providing interpretable explanations to the propensity scores obtained.

5. REFERENCES

[1] Y. Richter, Y. Naveh, D. L. Gresh, and D. P. Connors, “Optimatch: applying constraint programming to workforce management of highly skilled employees,” *Int. J. Services Operations and Informatics*, vol. 3, no. 3/4, pp. 263–279, 2008.

[2] D. Wei and K. R. Varshney, “Optigrow: People analytics for job transfers,” in *Proc. IEEE Int. Congress on Big Data*, 2015, pp. 535–542.

[3] Karthikeyan Natesan Ramamurthy, Moninder Singh, Michael Davis, J Alex Kevern, Uri Klein, and Michael Peran, “Identifying employees for re-skilling using an analytics-based approach,” in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 345–354.

[4] G. Salton and C. Buckley, “Term weighting approaches in automatic text retrieval,” Tech. Rep. 87-881, Department of Computer Science, Cornell University, Ithaca, NY, 1987.

[5] Alan Agresti and Maria Kateri, *Categorical data analysis*, Springer, 2011.

[6] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[7] Jianqing Fan and Jinchi Lv, “Sure independence screening for ultrahigh dimensional feature space,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[8] Leo Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[9] Yoav Freund, Robert Schapire, and N Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, pp. 1612, 1999.

[10] Peter Bühlmann and Bin Yu, “Analyzing bagging,” *Annals of Statistics*, pp. 927–961, 2002.