Recommending industry jobs with machine learning

Pedro Seguel [*1] **Shih-Tsai Wei** [*1]

1. Introduction

The internet has shaped the way people search for jobs nowadays (Jansen et al., 2005). The popularity of job market sites such as Indeed, Monster or LinkedIn provides an organized set of job offerings where employees can find their dream job and apply present or prepare for it in the future. Although these platforms help the potential candidate to evaluate specific jobs positions to apply, they don't provide an updated overview of the market trends that can help them to see the big picture.

We plan to apply machine learning (ML) techniques using job advertisement and recruitment data to enhance the search strategy of candidates and help them to identify the gap between the job post requirement and their current ability. We expand current research by focusing on the role of the predictive character that hard and soft skills have over specific industry jobs among the same job domain. We extend prior research on job market analysis by using automated techniques of ML, using the soft skill in our dictionary, and by proposing a user interface for job-seekers.

We plan to leverage a labeled dataset based on 760 jobs posting information for marketing internships extracted from Indeed to train our multi-label classification models. Our dataset includes labels for 31 job skills, which are also grouped between soft and hard skills. Using a supervised approach, we will be able to explore how these diverse skills are predictors of getting a job in a desirable industry. Moreover, this project explores a user interface that can use this information to recommend people how to map their careers.

2. Related Work

Job market analysis using job advertisements Prior research has provided an in-depth analysis of specific job market trends using online job postings and advertisements for domain-specific professions (Ritzhaupt et al., 2010; Jansen et al., 2005; Sugar et al., 2012). For instance, Ritzhaupt et al. (2010) paper examines the multimedia competencies of educational technologists via combining insights from job posting analysis and a survey of professionals within the field. This in-deep study highlights changes in job trends, and how the general curriculum match with those job trends, showing the opportunities of using job posting data to understand variations on the demand side of the market. However, as insightful these studies are, limitations arise from the use of standard statistical techniques for analysis and coding of the data on these studies. Our approach is distinctive from this research because we plan to take an automated approach to use the text-data that is available related to the job search process, and we provide broader suggestions to job seekers about their suitable industries instead of specific job postings. This work will be beneficial especially to people who don't have a clear preference regarding the industry. Also, we will present a UI prototype for users to visually give our career suggestions and provide a better user experience.

Matching jobs and applicants Based on published papers on applied machine learning to the online job seeking process (Guo et al., 2016; Fazel-Zarandi & Fox, 2009; Poch et al., 2014), for instance, Poch et al. (2014) explored the use of supervised classifiers to learn implicit relations between resumes and jobs offers, which cannot be found with similarity or pattern based search skills, names of professions and degrees. While this type of research tackles the broader challenges of matching resumes with job descriptions and desirable characteristics for an employee, our project focus on a more specific area of the job searching process. Our interest is on the job-seeker that is planning its training for a future application. Consequently, we focus on automating the process of labeling a type of job using job skills features, rather than semantically matching an existing resume with a job offer. Also, by focusing on more specific distinctions between job skills (including hard and soft skills), we plan to visualize hidden patterns present on the job market under the same job title, or for our case marketing interns.

Job skills relatedness Finally, another interesting -and less frequent- line of work focus on the relatedness among job skills using text mining techniques (Hughes, 2015; Van-Duyet et al., 2017; Lau & Sure, 2002). Van-Duyet et al. (2017) develop a method called Skill2vec, which applies machine learning techniques in recruitment to enhance the

[1]The University of Texas at Austin. Correspondence to: Pedro Seguel <ppseguel@utexas.edu>, Shih-Tsai Weid <stwei@utexas.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

search strategy regarding candidates possessing the appropriate skills. While Van-Duyet et al. (2017) creates a useful architecture to automate this process, its based primary on a general dictionary of technical skills. On the other hand, our work focuses on predictive classification techniques using labeled job skill data to explore the relationships of skills and industries. Also, we expand current research on job skill by adding soft, and hard skills to the analysis.

3. Methods

We decided to use a domain-specific dataset that could provide rich text information from the real job posting information, as well as a set of labeled skill data coded from those text descriptions. We will use Woods (2017) dataset on marketing internships job postings from a diverse set of enterprises in the U.S.A., which includes 760 job postings collected at the beginning of February 2017 from Indeed.com. We plan to use the list of 31 hard- and soft-skills as features which had been coded in the original dataset.

Since the targeted labels for the industry were not available on the dataset, we coded them by hand using the Linkedin information from each company named on the dataset. The company's key agent who created the company's profile on Linkedin had to select one industry category among total 149 different options created by Linkedin. Since the options are fixed and the industry is well self-defined, the resources are unified and reliable. After that, we recode these industry labels to a group of 17 main groups used by LinkedIn (Table 1). This process can provided us with the targeted classes for our study.

Table 1. Industry code based on LinkedIn (2019).

Group	Industry	Company
good, manufacture	Retail	Dierbergs
organizational, services	Individual and Family Services	Children's Bureau
corporate, media	Marketing and Advertising	A2 Advertising
...	...	...

According to this information, we address the job industry classification as a Multi-Label Classification problem (Jain, 2017), which means each instance can be assigned with multiple industry group categories. This process works well for our dataset since the predicting properties of a data-point are not mutually exclusive. We coded our set of target labels using the steps described above, and we plan to use the in-built adapted algorithm available on the sklearn.multiclass library to support multi-label classifications (Sklearn, n.d). This approach allows us to directly perform multi-label classification, rather than transforming the problem into different subsets of problems (Jain, 2017). We will use the adapted versions for the Decision Tree (sklearn.tree.DecisionTreeClassifier) and K-Nearest Neighbours (sklearn.neighbors.KNeighborsClassifier) classifiers. Moreover, since multi-label classification differs from a single-label setting we will use distinctive performance measures proposed by the literature (Wu & Zhou, 2017) that will be addressed with more detailed in our experiment sections.

The Figure 1 is a flowchart illustrating the steps of our machine learning system.

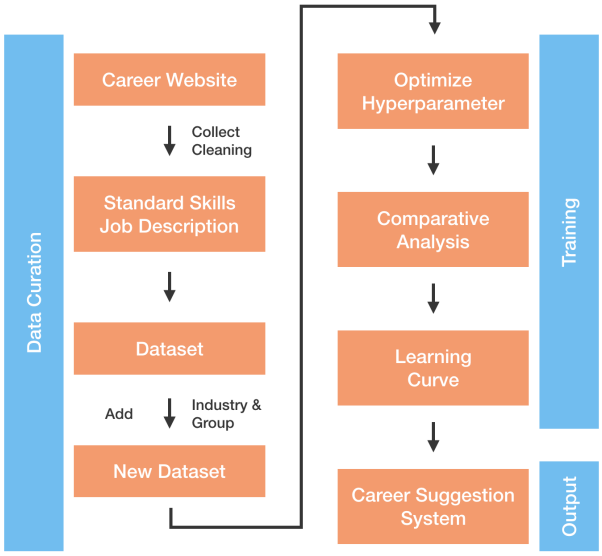

Figure 1. Flowchart

4. Experimental Design

4.1. Experiment 1

We will compare the performance of a Decision Tree, and K-NN adapted classifiers to analyze the data. There are two parts to this experiment. First, we will optimize hyperparameter(s) for each classification model. Second, based on the first result, we will do a comparative analysis of optimized classification models using overall evaluation metrics, and comparing the results on an adapted multi-label confusion matrix to analyze the targeted group industry labels.

Datasets. We will use Woods (2017) dataset on marketing internships job postings, which includes 760 job postings collected at the beginning of February 2017 from Indeed.com. However, we add one category of assigning the group of each industry as multi-labeled to do analysis.

Baselines. We will follow the two steps to conduct this part. First is to optimize hyperparameter(s). About the Decision Tree, we will find the optimal hyperparameters for the split criterion by testing gini and entropy and for

the tree depth by testing at least 6 different values when training a decision tree. For K-NN, we will find the optimal hyperparameters for the distance metric by testing Euclidean and Manhattan and for the number of nearest neighbors by testing at least 6 different values when using k-NN.

Second, we will conduct comparative analysis by using hyperparameters for regularized models to retrain each of the two models on all the training data using the optimal hyperparameters found in part 1. We will report the predictive performance on the test dataset for each of the two models.

Evaluation Metrics. We will use adapted measurements of Accuracy, Precision, Recall to compare the multi-label classifiers. To be able to apply these measures into multi-label classification, it is necessary to binarize the output (per target label). Also, the Sklearn library provides implementations that allow the analysis of the overall score of the classifier, by enabling each sample to give a weighted contribution to the total score with a *sample weight* parameter.

4.2. Experiment 2

On this experiment we will test how the size of training data affects the performance of the model.

Datasets. We will use the same dataset that we used in the Experiment 1.

Baselines. We will plot the learning curves for two multi-label adapted models including the Decision Tree, and K-NN classifiers. To create each curve, we will vary the amount of training data (at least ten different training sizes) in relation to the performance measured in coverage scores.

Evaluation Metrics. We will report our results using coverage scores as our performance metric. As Wu & Zhou (2017) show coverage score provides the number of more labels on average that should be included to cover all relevant labels, and it is considered as a distinctive performance metric for multi-label models.

5. Experimental Results

5.1. Experiment 1

Here we will report the overall evaluation metrics for each model compare the performance of each model regarding the targeted labels. Moreover, Figure 2 shows an example of a multi-label confusion matrix that we will use to provide a more in-depth analysis. This type of matrix -according to Sklearn documentation- is a format used to represent multi-label data, where each row of a 2d array or sparse matrix corresponds to a sample, each column corresponds to a class, and each element is 1 if the sample is labeled with

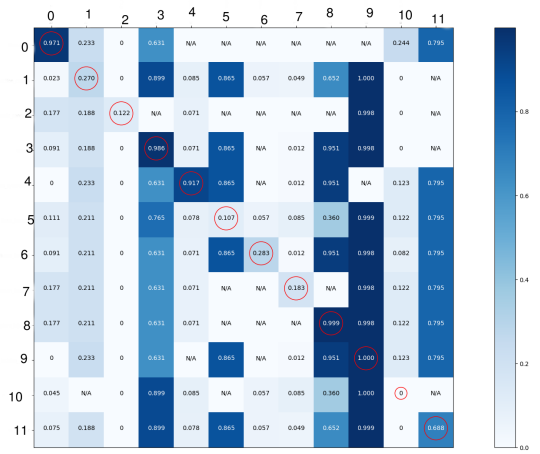

Figure 2. Multi-label confusion matrix

the class and 0 if not. Therefore, when the value is close to 1 in the diagonal (top-left to bottom-right) cells of the figure, it can be can interpreted as the model predicted the label for the right target. Also, the darker the color blue, the closet the number of the cell is to value 1. These results will give us a better idea of the relative performance between the chosen algorithms on average. However, we won't be able to compare their performance on a trade-off perspective regarding the analyzed metrics. This type of insight could be explored by using adapted ROC curves or other similar visualizations, which could be done in future research.

5.2. Experiment 2

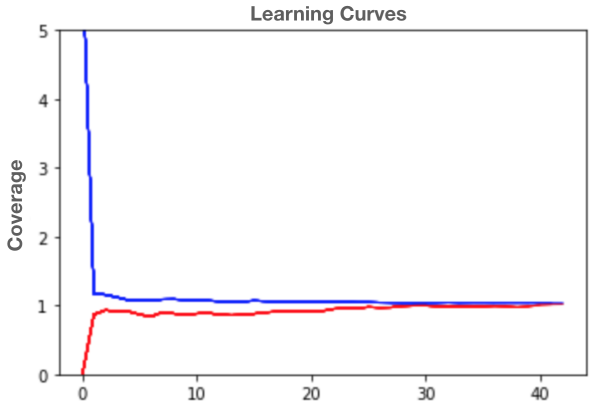

Figure 3. Learning Curve

Figure 3 shows an example of a learning curve that will be plotted for each algorithm. For instance, if the learning

curve of Decision Tree classifier performs better than the k-NN classifier, we expect that the coverage for both training and testing data set should be higher. By comparing the curves for the training and testing data when sample size increases, we can also check if the models are overfitted.

6. Conclusions

According to our experiments and the training of machine learning models, we show that our models have high levels of prediction of targeted industry labels. Furthermore, we can see that the Decision Tree performs better than the k-NN on the Multi-label confusion matrix. Besides, it takes a smaller amount of training and test size to get the best results.

We foresee that steps for future work might include the analysis of the relationships between the skills and other information contained in job posting data across different industries. Moreover, the study can be applied to other datasets not limited to marketing internships, including other domains and full-time jobs or contractors. Also, it would be interesting to compare our models with additional ones using non-labeled text data from the job posting. This comparison could help us understand if there is an emerging pattern within the text (style, use of words, word counts) across industry jobs.

Our research expands current efforts leveraging ML to provide suggestions on career development to people are given specific skills sets, including soft skill data. We believe that our preliminary results can be used for the design of prototypes of interfaces to assist job-seekers. Although the limitations of our evidence and our data, we believe that this is a significant step to develop AI solutions that can help people who are struggling in deciding or changing their career path.

References

Fazel-Zarandi, M. and Fox, M. S. Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In *Proceedings of the 8th International Semantic Web Conference*, volume 525, 2009.

Guo, S., Alamudun, F., and Hammond, T. Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60:169–182, 2016.

Hughes, S. How We Data-Mine Related Tech Skills kernel description. https://insights.dice.com/2015/03/16/how-we-data-mine-related-tech-skills/, 2015. Accessed: 2019-04-15.

Jain, S. Solving multi-label classification problems (case studies included). https://www.analyticsvidhya.com/blog/2017/08/introduction-to-multi-label-classification/, 2017. Accessed: 2019-04-23.

Jansen, B. J., Jansen, K. J., and Spink, A. Using the web to look for work: Implications for online job seeking and recruiting. *Internet research*, 15(1):49–66, 2005.

Lau, T. and Sure, Y. Introducing ontology-based skills management at a large insurance company. In *Proceedings of the Modellierung*, pp. 123–134. Citeseer, 2002.

LinkedIn. Industry Codes. https://developer.linkedin.com/docs/reference/industry-codes, 2019. Accessed: 2019-04-23.

Poch, M., Bel, N., Espeja, S., and Navio, F. Ranking job offers for candidates: learning hidden knowledge from big data. In *LREC*, pp. 2076–2082, 2014.

Ritzhaupt, A., Martin, F., and Daniels, K. Multimedia competencies for an educational technologist: A survey of professionals and job announcement analysis. *Journal of Educational Multimedia and Hypermedia*, 19(4):421–449, 2010.

Sklearn. 1.12. multiclass and multilabel algorithms. https://scikit-learn.org/stable/modules/multiclass.html/, n.d. Accessed: 2019-04-23.

Sugar, W., Hoard, B., Brown, A., and Daniels, L. Identifying multimedia production competencies and skills of instructional design and technology professionals: An analysis of recent job postings. *Journal of Educational Technology Systems*, 40(3):227–249, 2012.

Van-Duyet, L., Quan, V. M., and An, D. Q. Skill2vec: Machine learning approach for determining the relevant skills from job description. *arXiv preprint arXiv:1707.09751*, 2017.

Woods, R. Marketing internship postings. https://data.world/rdowns26/marketing-internship-postings, 2017. Accessed: 2019-04-23.

Wu, X.-Z. and Zhou, Z.-H. A unified view of multi-label performance measures. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3780–3788. JMLR. org, 2017.