# Who are Job Hoppers? Analysis and Prediction Using LinkedIn Data

**AAAI Press**
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

**Abstract**

In this study, we use data from LinkedIn to predict job hopping at scale. The attributes under investigation mainly consists of location, industry, skill, age, highest degree attained, last school attended, and length of working experience. We employ three different prediction models, including Support Vector Regression, Neural Networks, and Random Forest, which are evaluated by 10-fold cross-validation. More importantly, we analyze the attribute distributions and the correlation between job changes and the attributes to reveal valuable insights.

## Introduction

Employee's turnover is common in our society, meanwhile the information-based prediction of job shift is concerned with multiple agents in job market. For employers, training and adaption of employees are time and money consuming, so they tend to recruit people with better loyalty and pay more attention to current employees who are likely to leave. For employees, they would like to find a job where they can stay for a relatively long time. For headhunters, understanding career path of potential workers is what their works focus on.

Currently, there are works done related with this issue. However, they emphasized the importance of prediction that whether workers would leave their current position within a predefined time (Sikaroudi, Ghousi, and Sikaroudi 2015; Sexton et al. 2005). None of them considered to predict when a person would leave, which is significantly more difficult but also more valuable.

In this paper, we build models to measure the loyalty of workers via predicting duration they stay in the current job, compare the effectiveness of these models, and analyze the correlation between job changes and attributes. To achieve our goal of prediction, we get dataset from LinkedIn to model for the prediction when employees with specific background would leave their current position. Using this dataset, we show that Random Forest is relatively effective in comparison with other two.

Our work makes the following contributions:

- We identify the leading factors for employees to change jobs frequently;
- We evaluate a number of models in order to build a reliable predictor for job changes;
- We extract valuable insights that determine job changes in the hope to help both employers and employees achieve their goals.

## Related Work

Generally, job hoping is treated as a harmful behavior. Adler (2015) thinks job hoping is devastating to anyone who wants to develop a long-term career. Suster (2010) believes job hoppers make terrible employees so for start-ups, they should not be hired.

In contrast, some hold an opposite view that there is no harm by job hoppers for both employees and employers (Sullivan 2015, Lowman 2014, Zimmerman 2016, Trunk 2010). From their perspectives, job hoppers are regarded as easy-going employees who learn fast, bring new information, and are likely to be top performers in any company. As for job hoppers themselves, being a serial job hopper improves their careers while loyalty does not count much.

Regardless of the pros and cons, it is valuable to identity the characteristics of a job hopper. There are significant literatures on predicting jobs changes and career paths.

Kapur et al. (2016) introduced a method for ranking universities based on career outcomes of their graduates by leveraging career paths' data on LinkedIn. The algorithm consists of company ranker determining desire to a career and school ranker based on career outcomes.

Liu et al. (2016) used multi-source multi-task training data from Twitter, Facebook and LinkedIn to predict future career path. The method is accomplished via a multi-source learning framework and jointly penalizes disagree-

ments among sources, which also learns the task-sharing and task-specific features.

Xu et al. (2014) built a novel Similar Profiles algorithm which treats each member's profile as a sequence of nodes in a large social network, using a variety of features such as job titles, companies, skills, schools from LinkedIn to represent each member's identity. They use this method to help recruiters find similar profiles.

## Dataset

### Data Acquisition

From Linkedin.com, we search and download html files of employees' profile, and parse each file to extract data.

The final data attributes we organize include **Location**, **Industry**, **Skills**, **Age**, **Degree**, **School**, **Company**, **Title**, **Start Time**, **Length**, and **y**, where y is the target job stability value representing the length of one's *last job* (which we try to predict).

### Data Cleaning

To get a proper dataset, we apply the following restrictions:

- **Location** – United States only
- **Years of Experience** – more than 3 years.

### Data Selection

We further parse the data to obtain selected information:

- **Experience** – up to 6 most recent jobs (except the current one) per person will be stored, each of which contains **company name**, **title**, **start time**, and **length** staying in the position.
- **Education** – the first and the last education will be used to compute **age/seniority** and **the highest degree** with **school name**, respectively.

### Data Transformation

- **Skills** –Management, Finance, Language, Communication, Sales, Software, and Programming. Detail about the encoding rules is shown in Table 1 (on page 4).
- **Age/Seniority** – difference between the current year (2016) and the starting year of the first education.
- **Start Time –** difference between the start time of the job and 01/1950 in month. 0 stands for missing value.
- **Degree** – 0 stands for missing value; 1 for others; 2 for Associate; 3 for Bachelor; 4 for Master; 5 for Doctor.
- **School ranking** – binned by the range of USNEWS ranking of universities (e.g. any universities in rank 30-50 are binned into 40). Detail about the encoding rules is shown in Table 2 (on page 4).

### Other Data Issues

Profile pages of regular LinkedIn accounts (excluding premium accounts) are only accessible to recruiter users or users within their network; in addition, many LinkedIn users choose to keep their data private, including some premium accounts. As a result, the data we can reach is limited. Currently, we collect data about 5000 users.

Nearly all profiles accessible are premium users with good quality of data, which means their experiences and education background are relatively rich and accurate.

## Methodologies

### Support Vector Regression

Support Vector Regression (SVR) performs regression on continuous variables. To train an SVR model for our data, we choose the most widely used Radial Basis Function (RBF) as kernel, and set gamma as 0.01, cost as 1, and degree as 3 (Dai, Chuang, and Lu 2015).

### Back-Propagation Neural Network

Back-Propagation Neural Network (BPNN) was adopted in turnover rate prediction (Fan et al. 2012). In this study, we use 3-layer net, where the hidden layer has 64 units using sigmoid as the kernel, a rigid penalty of 0.01, and a loss function of squared error.

### Random Forest

Random Forest (RF) is an ensemble learning method for classification and regression. We set the Random Forest with 300 iterations. Note that its training takes a significantly shorter time compared with the other two models.

## Analysis

### Data Analysis

To analyze the correlation between each dimension and the target value, we visualize data for every attribute. Some of the plots are interesting and worth a discussion.

**Age/Seniority and stability Y.** Figure 1 shows the age/seniority distribution with the 75-quantile in each age. Note that we only show the 75-quantile for clarity because the full scatter plot is too crowded to discern. From the plot,

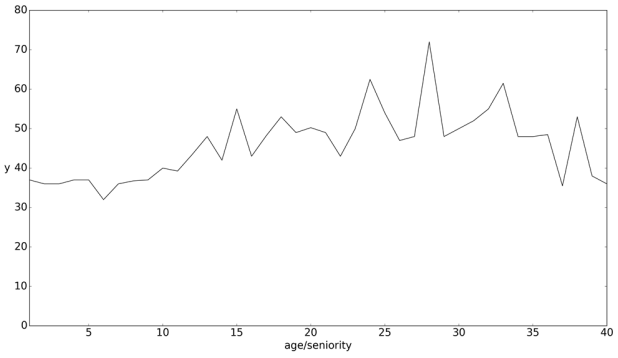


*Figure 1. Age/Seniority (months) vs. 75-Quantile Stability Value.*

we found that new graduates are more likely to leave the current position (lower *y*), while people after 10 years tend to be more stable in the position. Note that due to the size of our data, senior professionals are obviously fewer than youth professionals, which results in the appearance that older people (after 35, for example) are likely to change jobs. That is *not* the case.

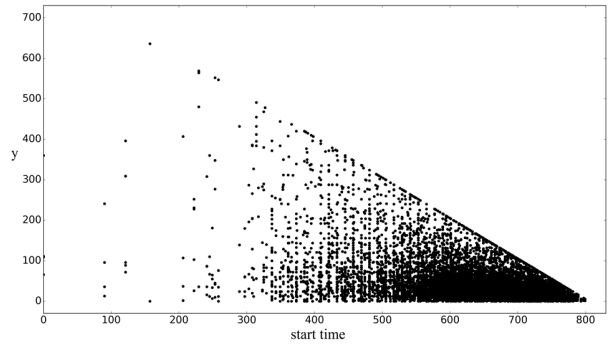


*Figure 2. Start Time (months since 01/1950) vs. Stability Value.*

**Start Time and stability Y.** Figure 2 shows the job change activities over the years. Most people started working in the last 30 years (thus are distributed between 440-800 months since 1950). People tend to change jobs often.

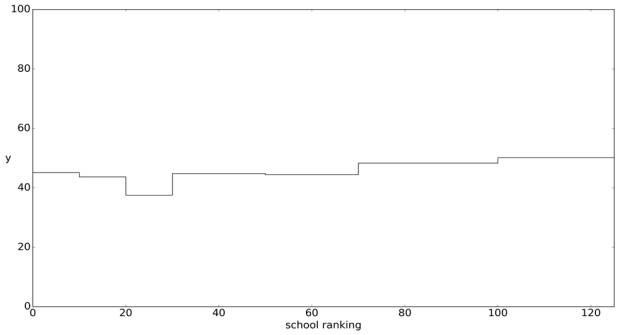


*Figure 3. School Ranking vs. 75-Quantile Stability Value.*

**School ranking and stability Y.** We can derive from Figure 3 that graduates from the universities ranked between 20 and 30 tend to stay in their current job for a bit shorter time. This is quite interesting in some ways.

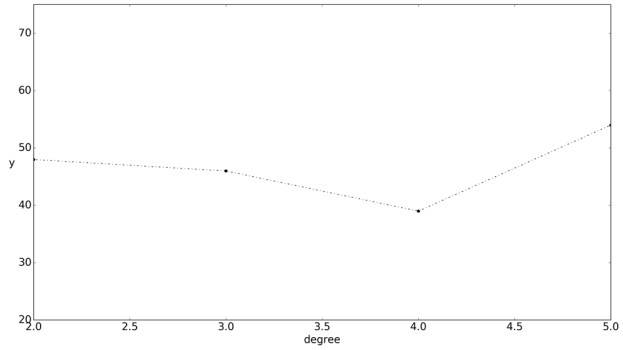


*Figure 4. Degree (2 for Associate; 3 for Bachelor; 4 for Master; 5 for Doctor) vs. 75-Quantile Stability Value.*

**Degree and stability Y.** Figure 4 shows the relationship between degree and the target value. Degree 2 stands for Associate Degree, 3 for Bachelor Degree, 4 for Master Degree, and 5 for Doctor Degree. It shows that people with a Doctor degree are more stable in their positions. People with a Master Degree, on the contrary, tend to stay in the same position for a relatively shorter time.

## Model Analysis

We use Mean Absolute Error (MAE) as one metric to evaluate the effectiveness of our prediction models:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$$

where $y$ stands for the true $y$, and $\hat{y}$ stands for the estimated $y$. From this metric, we can intuitively tell the range of true $y$ when the evaluated model returns an estimated $y$.

Another metric we use is Relative Absolute Error (RAE), which is similar with MAE but takes the scale of real $y$ into account:

$$RAE = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{\sum_{i=1}^{n}\left|\overline{y}_i - y_i\right|}$$

where $\overline{y}$ stands for the mean of real y. This metric reveals the relative error when the scale of real y is considered.

Using 10-fold cross-validation, we compare three models in MAE and RAE. The result is shown in Table 3.

| Model | Time to Build (sec) | MAE | RAE (%) |
|---|---|---|---|
| **SVR** | 148.7 | 28.8 | 88.5 |
| **BPNN** | 4905.1 | 19.0 | 58.6 |
| **RF** | **18.7** | **18.5** | **57.1** |

*Table 3. Evaluation of each model.*

Random Forest costs significantly shorter time to build and produces the best results. It means we can predict the time a person staying in its current position with an average error of 18 months, i.e. one year and a half. It is interesting to note that in another work about turnover prediction (Sikaroudi, Ghousi, and Sikaroudi 2015), even though they focused on predicting whether or not an employee would change job in a predefined time, it appears that Random Forest performs the best as well.

## Conclusion and Future Work

We use rich data from LinkedIn to predict job hoppers at scale. Among different prediction models, Random Forest performs the best. Moreover, we analyze the attribute distributions and the correlation between job changes and the major attributes to reveal valuable insights.

We plan to investigate more features, such as the number of job changes and the number of connections. Additional data sources to enlarge the dataset are also useful.

## References


Adler, L. 2015. http://www.inc.com/lou-adler/why-job-hopping-syndrome-is-a-career-killer.html

Suster, M. 2010. https://bothsidesofthetable.com/never-hire-job-hoppers-never-they-make-terrible-employees-e30cd5ff7322

Sullivan, J. 2015. https://www.eremedia.com/tlnt/hiring-job-hoppers-10-reasons-why-they-are-so-very-valuable

Lowman, E. 2016. https://www.themuse.com/advice/career-lessons-from-a-serial-job-hopper

Zimmerman, K. 2016. http://www.forbes.com/sites/kaytiezimmerman/2016/06/07/millennials-stop-apologizing-for-job-hopping/#1b6d57d8697d

Trunk, P. 2010. http://www.cbsnews.com/news/why-job-hoppers-make-the-best-employees

Sikaroudi, A. M. E.; Ghousi, R.; and Sikaroudi, A. E. 2015. A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). *Journal of Indus*

*trial and Systems Engineering* 8(4): 106-121.

Sexton, R. S.; McMurtrey, S.; Michalopoulos, J. O.; and Smith, A. M. 2005. Employee turnover: a neural network solution. *Computers & Operations Research* 32: 2635-2651.

Kapur, N.; Lytkin, N.; Chen, B.; Agarwal, D.; and Perisic, I. 2016. Ranking Universities Based on Career Outcomes of Graduates. In *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 137-144. San Francisco, Calif: ACM Press.

Liu, Y.; Zhang, L.; Nie, L.; Yan, Y.; and Rosenblum, D. S. 2016. Fortune Teller: Predicting Your Career Path. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 201-207. Phoenix, Arizona: AAAI Press.

Xu, Y.; Li, Z.; Gupta, A.; Bugdayci, A.; and Bhasin, A. 2014. Modeling professional similarity by mining professional career trajectories. In *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1945-1954. New York, New York: ACM Press.

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, 144-152. Pittsburgh, Penn: ACM Press.

Dai, W.; Chuang, Y.; and Lu, C. 2015. A clustering-based sales forecasting scheme using support vector regression for computer server. *Procedia Manufacturing* 2: 82-86.

Fan, C.; Fan, P.; Chan, T.; and Chang, S. 2012. Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications* 39(10): 8844-8851.


| Value | Superclass | Subclass |
|---|---|---|
| 0 | Management | management, leadership, project management, strategy, business development, training, team leadership, business strategy, event planning, program management, crm, start-ups, event management, strategic partnerships, strategic planning, time management, content strategy, business analysis, process improvement, entrepreneurship, change management, operations management, product development, coaching, competitive analysis, business planning, project planning, forecasting, product management, leadership development, portfolio management, brand management, vendor management, management consulting, executive management, requirements analysis, user experience, inventory management, content development, hiring, employee relations, creative strategy |
| 1 | Finance | budgets, financial analysis, investments, fundraising, financial modeling, mergers & acquisitions, financial services, capital markets, budgeting, corporate finance, small business |
| 2 | Language | English, Chinese, Spanish, Japanese |
| 3 | Communication | public speaking, social networking, marketing communications, copywriting, press releases, editing, writing, consulting, media relations, wordpress, teamwork, strategic communications, recruiting, teaching, video, copy editing, creative writing, entertainment, video editing, problem solving, community outreach, editorial, scrum, corporate communications, ppc, troubleshooting, storytelling, publishing |
| 4 | Sales | marketing, marketing strategy, sales, social media marketing, sales management, market research, e-commerce, product marketing, sales operations, integrated marketing, selling, retail, marketing management |
| 5 | Internet | online marketing, online advertising, email marketing, facebook, blogging, powerpoint, word, excel, seo, web analytics, google analytics, interactive marketing, b2b, mobile devices, salesforce.com, photoshop, photography, sem |
| 6 | Software | javascript, web development, enterprise software, c++, xml, html5, web applications, unix, css, cloud computing, jquery, saas, git, testing |

*Table 1. Skills Classification*

| Rank | Universities |
|---|---|
| 5 | Princeton University, Harvard University, University of Chicago, Yale University, Columbia University, Stanford University, Massachusetts Institute of Technology, Duke University, University of Pennsylvania, Johns Hopkins University |
| 15 | Dartmouth College, California Institute of Technology, Northwestern University, Brown University, Cornell University, Rice University, University of Notre Dame, Vanderbilt University, Washington University in St. Louis, Emory University, University of California—Berkeley, Georgetown University |
| 25 | University of California—Los Angeles, University of Virginia, Carnegie Mellon University, University of Southern California, Tufts University, Wake Forest University, University of Michigan—Ann Arbor, University of North Carolina—Chapel Hill |
| … | … |

*Table 2. US News College Ranking (only the Top 30 schools are shown due to the page limit).*