Career Path Prediction

JEREMIAH BILL, University of Rochester
RAAYAN PILLAI, University of Rochester

CCS Concepts: •**Information systems** →**Data mining;** •**Applied computing** →**Forecasting;**

Additional Key Words and Phrases: Career prediction, professional analysis, neural network, decision tree, data science, data mining, web scraping

ACM Reference format:
Jeremiah Bill and Raayan Pillai. 2017. Career Path Prediction. *ACM Trans. Web* 1, 1, Article 1 (December 2017), 10 pages.
DOI: 0000001.0000001

1 INTRODUCTION

With the excess of career opportunities in our society, it has become increasingly difficult to choose the right career path. This is an example of the paradox of choice where having more options causes individual to struggle when making decisions, due to fear of making the wrong decision. Further, this indecisiveness causes a gap in the job market, filling jobs with individuals who are not passionate about their work, causing decreased levels of productivity. Our mission, is to limit the range of choices for prospective professionals, giving them the opportunity to better understand career paths based on their work experience and the current education and job market.

2 RELATED WORK

As of late companies are searching for ways to attract and retain top talent in order to further business initiatives. In turn, research in the field of talent acquisition, retention and mobility is gaining a substantial amount of traction. Two studies specifically [7, 8], applied data mining strategies to model talent career paths and career outcomes based on education respectively.

The research team in [8] sought to model talent career paths by focusing on turnover and career progression. Their study is focused on a dataset containing anonymized employee career data. For each employee, there is temporal data reflective of start/end dates of the employee joining the company or holding a position at a given occupational level. Further, their research was supplemented by both static and dynamic information for each employee. They describe static information as unchanged data such as gender and age. On the other hand, dynamic information included numeric performance ratings and hierarchical report chains with respective time stamps. The use of such dynamic factors provided a qualitative look at what governs an employee leaving a company or being promoted within a company which was unique in comparison to the datasets we were able to gather for analysis. The researchers found that without the numeric performance rating attribute the model's performance would decrease substantially, meaning that this attribute is

a viable metric for predicting career path and status within a company. All in all, the team proposed "a novel survival analysis approach for modeling the career paths of employees, which is based on multitask learning with ranking constraint formulation."[8] which significantly outperformed other multi-task learning and survival analysis methods.

The research team from LinkedIn [7] leveraged valuable proprietary data in order to rank universities based on industry performance of their graduates. The team created a system with two main ranking components: a company ranker and school ranker. The company ranker used LinkedIn member data to generate desirable companies for a given profession. The school ranker was then used to rank universities based on the number of graduates attaining a job at a desirable company for a given profession. Their company ranker is a graphical model, where companies represent the nodes of the graph and employee movement from company to company is represented by the edges of the graph. Finally, in order to generate the desirability score of each company, the team applied PageRank on the graph. Using this company ranker they could then begin to generate university rankings. The research proposed by this group provided a novel approach to understanding the importance universities play in attaining a job for a given profession. They were able to analyze career= trajectories of graduates and use that information to generate a ranking for a university as opposed to other metrics based on "reputational assessments". Ultimately, the team's work provided substantial insight for universities and students; adding to the resources provided by LinkedIn Higher Education while gaining media coverage from major publishers.

Within our analysis, we sought to accomplish the following: gain insight on the current job and education market, mine user data relevant to career trajectories(LinkedIn profiles and resumes), use mined data to predict potential career paths. All in all, our research culminates a variety of data sources and data mining approaches to help prospective professionals make confident decisions in the job industry.

3 METHODOLOGY

3.1 Data Acquisition

Gathering data began with searching for openly available datasets relevant to our problem statement. In turn, we gathered data relevant to the current education and job market, leading company data and user data.

(1) Job Market data
  (a) US Jobs on Monster.com [2]
    (i) Includes job listings of various US based positions. Provides, title, description, location, sector and organization of each listing
    (ii) Provided us with a means for analyzing how job descriptions play a role in predicting job industry through the use of text analysis and classification methods
    (iii) Allowed us to mine popular locations amongst job listings
(2) Company Data
  (a) Fortune 500 Companies [1]
    (i) List of top companies ranked by revenue
    (ii) Provided a means for tiering companies
  (b) The Open Data 500 [10]
    (i) General company metadata including: size, category, location, business model attributes
    (ii) Provided data to analyze industry, company size and location popularity
(3) Education Data
  (a) College Scorecard Data [4]

(i) Contains relevant statistics about US universities including: admission rates, unemployment rates, post graduation earnings and degree percentages.
(ii) Provided data for modeling salary expectations based on university type
(iii) Allowed us to analyze correlation of admission rates and unemployment rates.

(b) University and Degree Data [3]
(i) Included 3 datasets regarding US universities and degree programs
(ii) Salary by degree: earnings statistics for various degree programs
(iii) Salary by college type: earnings statistics for universities of a given type (Public, Ivy League, Party, Liberal Arts)
(iv) Salary by region: earnings statistics for universities within given regions. (This dataset was joined with the college type dataset as they both contained the same universities.)

(c) The Times Higher Education World University Ranking 2018 [5]
(i) List of World Universities ranked
(ii) Allowed for a means of tiering universities for predictive analysis

(4) User Data
(a) Linkedin.com Scrapper [9]
(i) Extraced previous work and education experience
(ii) Used to model career paths

(b) Indeed.com Scrapper [6]
(i) Extracted job title, education and 3 most recent work experiences based on job query
(ii) Collected a total of records to use in conjunction with linkedin data to predict (add more)

3.2 File Structure

- root
  - analysis
    - college-scorecard-analysis.ipynb
      - code for analyzing College Scorecard Dataset [4]
    - education-analysis.ipynb
      - code for analyzing Wall Street Journal University datasets [3]
    - job-market-analysis-refined.ipynb
      - code for analyzing job listings from Monster.com dataset [2]
    - trajectory-analysis-rev-ntr.py
    - trajectory-analysis.py
    - us-companies-analysis.ipynb
  - data
    - images: analysis generated images
    - companies: company data [1, 10]
    - education: data for education analysis [5]
    - jobs: job analysis data [2]
    - users: scraped user data [6, 9]
    - scraping: various scrapers

4 EXPERIMENT

4.1 Job Listing Analysis

To begin analysis of the job listing data we first sought to gain an understanding of the dataset by constructing visualizations of the data. Throughout this process we were able to gain insight into popular job locations, distribution of organization and sector values, and frequently occurring words within job titles. As seen in Figure 1, healthcare and retail listings dominate this dataset which validates the important words found in the job title field due to the fact that ”manager”,

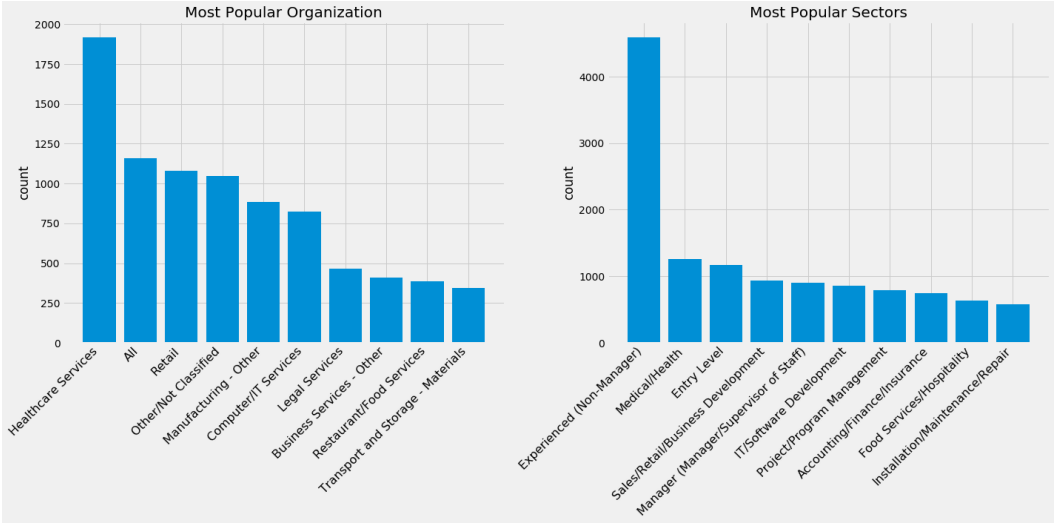

Fig. 1

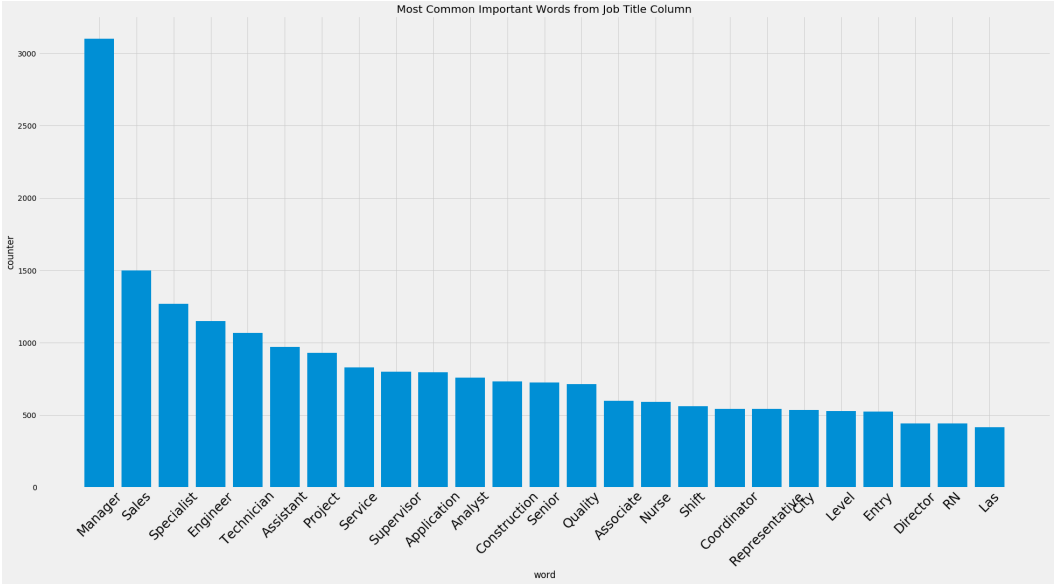

Fig. 2

”sales”, ”specialist”, ”technician” , ”supervisor” are among the top ten most frequently occuring words.

Next, we sought to analyze the job description field to see if we could use this attribute to determine what organization the listing corresponded to. The motivation here was that each organization would have a unique set of words within the description field that are exclusive to

that organization and using this assumption we can create a consistent feature set for each job listing to train a classifier with. The following steps were taken in our text classification process:

(1) Feature Creation
  (a) Create of the bag-of-words
    (i) Iterated through all job listings
    (ii) Removed irrelevant words(stop words) and punctuation
    (iii) Lemmatized each non-stop word in order to reduce derived words down to one common lemma. i.e.(walked, walks, walking = walk)
    (iv) Our bag-of-words contains the 1000 most frequent words
  (b) Use our bag-of-words to create a consistent feature set for each listing's description field
    (i) Given our bag-of-words of length 1000 we create a length 1000 feature vector
    (ii) Each index i in the feature vector corresponds to a binary value(1 if word is in description else 0) based on whether or not the word in index i of our bag-of-words is in the description of the current listing
(2) Label Creation
  (a) One hot encode organization values
  (b) Given 5 organization values create a label vector of length 5
  (c) Each index of the label vector corresponds to a unique organization value
  (d) For each listing, set the listing's organization value index to 1 in the label vector, every other index will remain 0

4.2 Classification Process and Results

After creation of our feature and label sets we were prepared for classification. The algorithms we used for classification were Naive Bayes and a Multilayer Neural Network. Our training and testing sets where split 75%/25 (total = 6091 , training = 4568 , test = 1523) and the top 5 most common organization values were used as labels.

4.2.1 Naive Bayes. We used two forms of Naive Bayes namely, Multinomial and Bernoulli Bayes. The core difference between these two algorithms is that Bernoulli Bayes can only operate with binary feature values but Multinomial Bayes can operate with both binary and absolute count feature values. The accuracy of our Naive Bayes classifiers were just over chance, attaining 64% accuracy for Multinomial Bayes and 63% for Bernoulli Bayes.

4.2.2 Multilayer Neural Network. Our Multilayer Neural Network was designed using TensorFlow[11]. In order to create a neural network using TensorFlow, we had to create a computation graph to represent our neural network model which would then be run by a TensorFlow session with our input data. In order to create the computation graph, each layer's weights and biases were defined, then the computation between each layer was defined. The computation between each layer is simply (input data * weights) + biases applied to an activation function (reLu for classification purposes) where each layer is then passed as input data to the next layer. Next, we defined an optimization function to reduce the cost after each iteration of training. Finally, to ensure our network was operating appropriately, we ran our TensorFlow session in batches, outputting the loss to ensure that as the network was training the loss was decreasing. Ultimately, our neural network model attained an accuracy of 75%.

4.2.3 Job Listing analysis results. Ultimately, organization prediction based on job description could be improved with a larger data set and cleaner data. The data set proved to be very sparse leaving us with a small fraction of the dataset to analyze(only 6091 listings out of 22,000). Therefore,

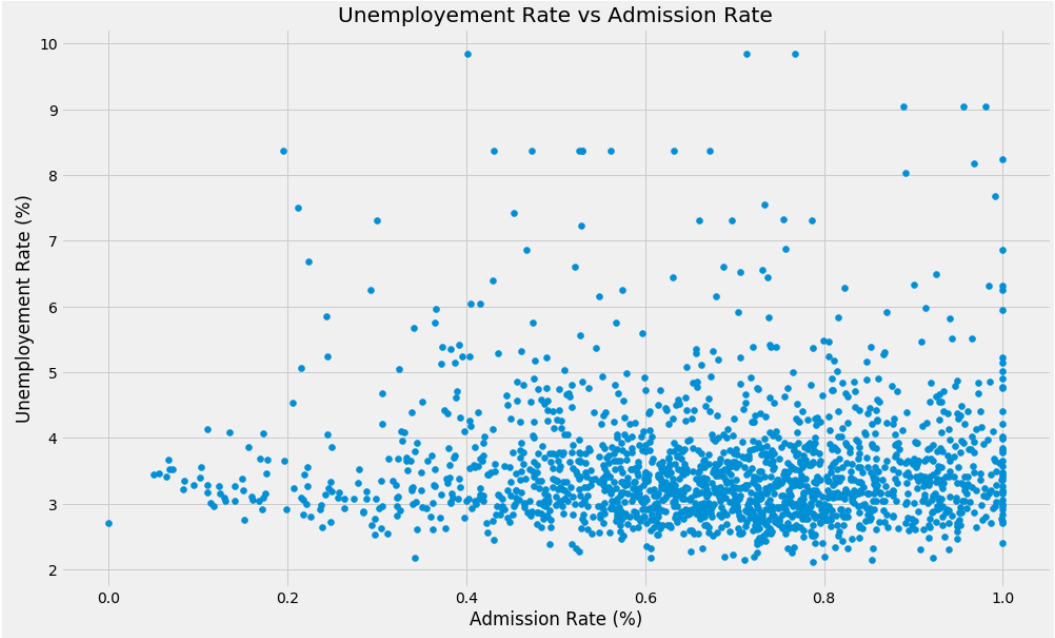

Fig. 3

given more data from a variety of job listings our text based classification approach could yield promising results.

4.3 Education Analysis

4.3.1 College Scorecard data. The motivation for analyzing the college scorecard data [4] was to understand how various university metrics play a role in determining employment and earnings of prospective graduates. The main attributes of focus were admission rates, unemployment rates, and degree percentages(percentage of students awarded a given degree) for each university. One of the first metrics we analyzed was the correlation between admission and unemployment rates. Looking at the scatterplot below, it is evident that there is no correlation between admission and unemployment rates. This is an interesting finding because this contrary to the popular belief that universities with lower admission rates yield increased job placement. But after careful consideration, this lack of correlation can possibly be explained by universities with higher admission rates offering degree programs where the job market is less competitive and there are more readily available jobs. As opposed to universities with lower admission rates placing students in highly competitive fields. Another possible explanation could be that universities with lower admission rates are on average smaller than universities with higher admission rates. For example, two universities(one large, one small) with the same number of unemployed graduates, will lead to the smaller university having a higher unemployment rate than the larger university.

To further investigate the effects of university statistics on career outcomes we created a multivariate linear regression model. In doing so, we used degree percentages, admission rates and unemployment rates as independent variables and median 10 year salary as our dependent variable. The accuracy of the regression model is characterized by a standard error of 25% and R^2 value of .5. This tells us that the dependent variables play a moderately strong role in determining the

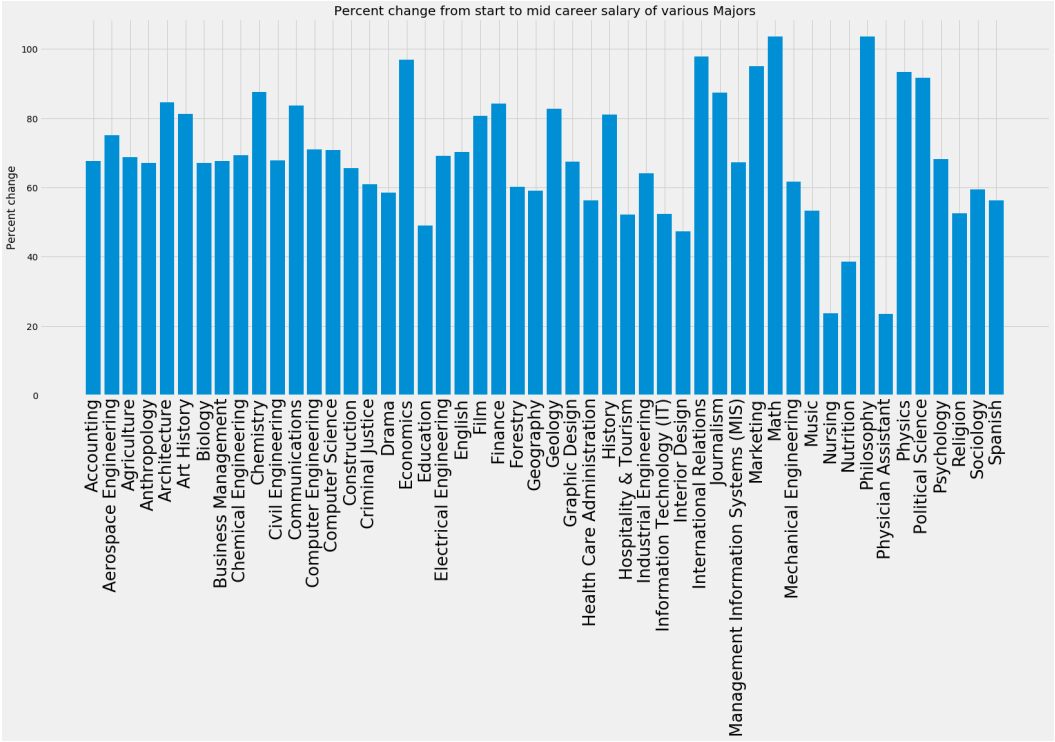

Fig. 4

salary of a graduate of a given university with relatively high accuracy. However, we noticed a potential source of bias within the data. We recognized that degrees for health professions were the most common degree type awarded, nearly doubling the next highest category. This imbalance could potentially skew our regression to favor salaries of health professions as opposed to the other degree programs. Therefore, we pruned universities which had health profession degree percentage attribute larger than 50%. This yielded a .3 increase in our R^2 value and a 5% increase in accuracy.

Taking an even granular look at degree programs we were able to visualize percent change from start to mid career for various degree programs (Figure 4). The two lowest percent changes in salary were for Nursing and Physician Assistant degree programs. Our hypothesis for this finding is that these two degree programs do not adopt the typical corporate hierarchy. In order for nurses or physician assistants to advance to higher levels of the medical hierarchy, more schooling would be required. Overall, this is quite common in the medical field, in order for employees to advance more certification is inevitable. On the other hand, the majority of degree programs under study do adopt a corporate hierarchy where it is common for an employee to "climb the corporate ladder". For example, a business student moving from an analyst, to an associate, to vice president, etc. Ultimately, the corporate hierarchy trend causes the mean percent change to be 70% while Nursing and Physician Assistant percentages are substantially lower 50%.

4.4 Company Analysis

To further supplement our predictive modelling of career outcomes we sought to extract various characteristics of industry leading companies. For example, when analyzing the The Open Data

Fig. 5

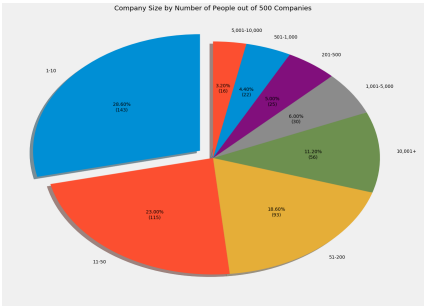

Fig. 6

500 dataset [10], we found that the most popular company categories were data/technology and Finance/Investment validating the high frequency of company locations being in Silicon Valley and New York. Further, startups dominate this data set, Figure 6 shows over 50% of companies having less than 50 employees.

4.5 Career Trajectory Prediction

One aspect of our project was to be able to predict the type of company a person would work at given their schooling and first couple jobs. To do this we decided to use our data from LinkedIn and Indeed to create tuples of users. This data, as aforementioned, was scraped and converted into the following tuples: *(School, Company1, Company2, Company3).* Around 600 user tuples were generated from our web scraping. The first step of preprocessing was to index the schools against the Times Higher Education, World University Rankings 2018 [5]. Schools of rank 0-50 were bucketed as T1, 50-200 as T2, 200-500 as T3, 500-1000 as T4. Schools not on the list were given T5, and users without schooling were given T6 known as 'NONE'. For companies the same process was performed, except the indexes were matched from Fortune 500, Top 1000 Companies [1]. After this was completed the buckets were transformed back into integers (0 to 5) for use in various classifier softwares. Now having tuples of type (int, int, int, int) we could use many of *sklearn*'s classifiers. The classifiers would be trained on the tuples, considers *School*, *Company1*, and *Company2* as the features and *Company3* as the class. For each classifier 550 tuples would be used for training and 85 would be used for testing. The first classifier used was a *DecisionTree*, it resulted in a 66% accuracy. The next classifier sued was an *ExtraDecisionTree*, it resulted in a similar accuracy of around 66%. Decision trees were chosen mainly because they are easy to visualize for this type of problem. The next classifier used was a *MultinomialNB*, it resulted in 67% accuracy. The final classifier used was a *MLPClassifier*, it resulted in 68% accuracy. From these results it's easy to see that the average accuracy was around 67%. The testing was done using an 7 fold cross validation. An interesting finding from the visualization of the decision tree was that the only feature used to classify was *Company2*. This leads us to believe that the most recent job you had is most likely to predict your next job, which falls in line with common sense.

5 CONCLUSION

In conclusion, our analysis provides a holistic view of the current education and job market. The creation of this project forced us to consider many of the aspects of *knowledge discovery*. Choosing classifiers, proper preprocessing and finding apporpiate sources were among many of the import decisions we made. We also provide a unique approach to predicting career paths with self mined

data using a tier based approach with various classification methods. Further, we executed text classification to predict industries aligned with certain job descriptions. Ultimately, our analysis yielded promising results which applied to more data could yield even more prominent results.

ACKNOWLEDGMENTS

The authors would like to thank Professor Jeibo Luo for a great semester and the oppurtunity to work on an interesting project.

REFERENCES

[1] Fortune 500. 2017. Fortune 500. (2017). Retrieved December 14, 2017 from http://fortune.com/fortune500/
[2] Aleksey Bilogur. 2017. US Jobs on Monster.com. (2017). Retrieved December 14, 2017 from https://www.kaggle.com/PromptCloudHQ/us-jobs-on-monstercom/version/1
[3] Chris. 2017. Paysacale Inc. Where it Pays to Attend College. (2017). Retrieved December 14, 2017 from https://www.kaggle.com/wsj/college-salaries
[4] U.S. DOE. 2017. College Scorecard Data. (2017). Retrieved December 14, 2017 from https://collegescorecard.ed.gov/data/
[5] Time Higher Education. 2017. World University Rankings 2018. (2017). Retrieved December 14, 2017 from https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats
[6] Indeed. 2017. Indeed Resumes. (2017). Retrieved December 14, 2017 from https://www.indeed.com/resumes
[7] Navneet Kapur, Nikita I Lytkin, Bee-Chung Chen, Deepak Agarwal, and Igor Perisic. 2016. Ranking Universities Based on Career Outcomes of Graduates.. In *KDD*. 137–144.
[8] Huayu Li, Yong Ge, Hengshu Zhu, Hui Xiong, and Hongke Zhao. 2017. Prospecting the Career Development of Talents: A Survival Analysis Perspective. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 917–925.
[9] LinkedIn. 2017. LinkedIn. (2017). Retrieved December 14, 2017 from https://www.linkedin.com
[10] Sharan Pai. 2017. The GovLab, Open Data 500. (2017). Retrieved December 14, 2017 from https://www.kaggle.com/govlab/open-data-500-companies
[11] TensorFLow. 2017. TensorFlow Basic Usage. (2017). Retrieved December 14, 2017 from https://www.tensorflow.org/versions/r0.12/get_started/basic_usage

Received November 2017; revised Decemeber 2017