RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis

Chapter · October 2020

DOI: 10.1007/978-3-030-62008-0_35

6 authors, including:

Ziwei Jin
Eindhoven University of Technology

Jiaxing Shang
Chongqing University

Qianwen Zhu
City University of Hong Kong

RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis

Ziwei Jin[1,2], Jiaxing Shang[1,2(✉)], Qianwen Zhu[3], Chen Ling[1,2], Wu Xie[4], and Baohua Qiang[4]

[1] College of Computer Science, Chongiqng University, Chongqing, China
ziweikan@foxmail.com, shangjx@cqu.edu.cn, lingchen20638@163.com

[2] Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing, China

[3] School of Management and Engineering, Nanjing University, Nanjing, China
zhuqw@smail.nju.edu.cn

[4] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, China
xiesixchannels@126.com, qiangbh@guet.edu.cn

Abstract. In human resource management, employee turnover problem is heavily concerned by managers since the leave of key employees can bring great loss to the company. However, most existing researches are employee-centered, which ignored the historical events of turnover behaviors or the longitudinal data of job records. In this paper, from an event-centered perspective, we design a hybrid model based on survival analysis and machine learning, and propose a turnover prediction algorithm named RFRSF, which combines survival analysis for censored data processing and ensemble learning for turnover behavior prediction. In addition, we take strategies to handle employees with multiple turnover records so as to construct survival data with censored records. We compare RFRSF with several baseline methods on a real dataset crawled from one of the biggest online professional social platforms of China. The results show that the survival analysis model can significantly benefit the employee turnover prediction performance.

Keywords: Turnover prediction · Survival analysis · Random survival forests · Professional social networks · Machine learning

1 Introduction

The employee turnover problem has been widely concerned by companies since the demission of key staff members may cause great loss to the company. Even after ordinary employees quit their job, companies will have to re-invest time and money to find new alternatives and train them. Obviously, being able to foresee whether an employee is likely to leave would benefit the company in terms of retaining talented employees and reducing losses.

Z. Huang et al. (Eds.): WISE 2020, LNCS 12343, pp. 503–515, 2020.
https://doi.org/10.1007/978-3-030-62008-0_35

There have been tremendous research works tackling the employee turnover problem, some of them [4,7] fall within the scope of management, sociology, psychology, etc. In recent years, with the flourish of online social networks [8], data-driven approaches began to draw researchers' attention, examples include machine learning based methods [3,13,15,21], survival analysis based methods [14,16,19], social network analysis based methods [5,6,17], etc. However, existing machine learning based methods mainly focus on feature engineering for the binary prediction task, which ignored the historical events of turnover behaviors or the longitudinal data in job records. Traditional survival analysis models usually impose strict assumptions on data distribution, and they are mainly used for factor analysis rather than turnover prediction. In this paper, we propose a hybrid model that combines survival analysis with machine learning models, based on which we further propose a turnover prediction algorithm. In our framework, we focus on turnover events rather than employees through survival analysis from the perspective of events. Specifically, we use **R**andom **S**urvival **F**orests for survival analysis and **R**andom **F**orest form turnover prediction, leading the **RFRSF** algorithm. In addition, we take strategies to handle employees with multiple turnover records so as to construct survival data with censored records. We first calculate the probability of each turnover event occurrence at a certain time point. Then, we view the probability value along with other time-invariant information as features to learn whether an employee is likely to leave at a certain time. Finally, we use these features to train machine learning models for the employee turnover prediction task.

To evaluate our proposed model, we conduct experiments on a real-world dataset crawled from one of the largest professional social platforms in China [1]. The results show that compared with other baseline models, the proposed model can achieve higher prediction accuracy.

2 Related Work

In this paper, we mainly focus on data-driven studies about employee turnover problem, and these studies can be roughly divided into three groups.

The first group are machine learning based methods, which formulate employee turnover prediction as a binary classification problem. Liu et al. [15] evaluated the importance of job skills for departure prediction. Yang et al. [21] propose a causal structure learning based feature modification method (CSFM), which helps management to retain employees who are leaving. Studies [1,3] offered experimental comparisons of various common machine learning algorithms. In general, random forests [3] and XGBoost [1] perform relatively better. De Jesus et al. [13] trained different machine learning models based on different industry characteristics.

The second group are survival analysis based methods. Studies [14,16,19] analyzed the impacting factors which contribute to employee's turnover behavior and calculated turnover probabilities with a Cox proportional hazard model.

[1] Due to privacy consideration, we do not disclose the name of the platform.

Zhu et al. [22] combined Cox proportional hazard model with random forest for turnover prediction.

The third group are social network based methods. Bigsby et al. [5] extracted network features to improve prediction accuracy. Oentaryo et al. [17] modeled employee mobility as a directed network, where each edge represents a job hob behavior. Based on the network they performed connectivity analysis at job and organization levels to derive insights on talent flow. Cai et al. [6] modeled the connections between employees and companies as a bipartite graph and proposed a graph embedding approach to predict employee turnover behaviors.

Besides the above methods, there are also other studies, such as the semi-Markov based algorithms [9,10].

3 Survival Analysis

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems. It is straightforward to see that survival analysis can be used to analyze the employee turnover problem.

3.1 Basic Concepts

To begin, we first introduce some basic concepts about survival analysis, please refer to [2] for more details.

Event: Based on the specific problem, an event can refer to the death, failure, turnover, crime recommitment or other events of interest.

Survival Time: If an event happens, then survival time is the duration from the beginning of observation to the event occurrence. Otherwise it is the duration to the end of observation or the time when the object exits the experiment.

Censor: A record is censored if the event of interest is not observed due to the limited observation time or other reasons. Censoring can be divided into several subtypes [20], and in this paper, we only consider the right-censoring where the event is not observed due to the limited observation time.

Survival Function: It is the probability that an individual lives longer than a specific time t, which is defined as:

$$S(t) = P\{T \geq t\} \tag{1}$$

where T is event time. Survival function monotonically decreases with t, and given that all of the subjects are alive at the beginning, we have $S(0) = 1$.

Hazard Function: Consider the conditional probability of an event happening at time t, given that the event does not happen before t, i.e., $P\{X \in (t, t + dt)|X > t\}$, then the hazard function is defined as:

$$H(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt)}{dt \cdot S(t)} \tag{2}$$

Cumulative Hazard Function: the cumulative hazard function is defined as:

$$\Lambda(t) = \int_0^t H(u)du \tag{3}$$

3.2 Random Survival Forests Model

The RSF model, proposed by Ishwaran et al. [12], is an extension of traditional random forests (RF) model for survival analysis. Like traditional RF, RSF uses bootstrap to draw samples and grows multiple binary recursive survival trees. Compared to existing survival analysis models, RSF does not impose any assumption on data distribution.

In the RSF model, the trees are built according to the cumulative hazard function (CHF). It first calculates the CHF of the terminal nodes and then gets the CHF of each tree in the forest. Finally, to obtain an ensemble CHF, we average over all survival trees in the forest.

Harrell's Concordance Index (C-index): The RSF model can be evaluated by Harrell's Concordance Index [11]. The C-index measures whether an event occurred earlier is associated with a higher risk rate or lower survival rate. For each pair of samples $[i, j]$, to calculate C-index, we first define p_{ij} as:

$$p_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are comparable and the prediction is concordant} \\ 0.5 & i \text{ and } j \text{ are comparable and the prediction is inconsonant} \\ 0 & i \text{ and } j \text{ are not comparable} \end{cases} \tag{4}$$

Based on the above formula, the C-index is as follows:

$$C\text{-index} = \frac{\sum p_{ij}}{\sum l_{ij}} \tag{5}$$

where l_{ij} is an indicator of whether the sample pair $[i, j]$ is comparable and 1 means comparable while 0 means not comparable.

4 The Proposed Method

4.1 Problem Formulation

In this paper, we analyze the problem of turnover prediction from an event-centered perspective. A traditional way to solve this problem is to predict an employee's turnover behavior completely based on her previous turnover behaviors, which is an employee-centered manner. However, this kind of model assumption can be "biased", because an employee's turnover behavior only relates to her turnover history. Moreover, temporal factors also play an important role because an employee's decision changes over time. Therefore, a reasonable way to investigate this problem is to fully consider the three factors: time, current status and employee's past turnover behaviors.

Based on the above analysis, we incorporate the specific prediction time into the survival function, and formulate our problem as follows:

Definition 1. *(**Turnover Prediction Problem**): Given an employee u's past turnover events, the current job information, the social platform information, and a specific time t, the turnover prediction problem aims at predicting whether u will quit the current job at time t.*

4.2 Our Method

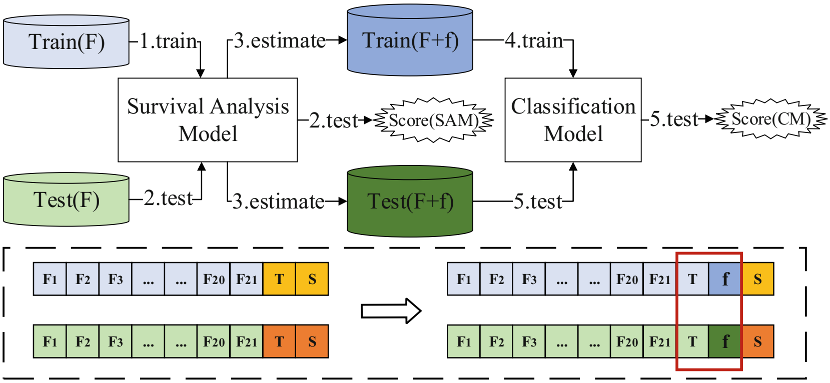

Fig. 1. The general framework of ExtTra

Framework. The framework of our method is shown in Fig. 1, which is divided into two parts: the survival model and the classification model. The survival analysis model is based on RSF, which is trained and tested to meet the preset requirements and generates survival rate (f) as the output before entering the classification model. In this part, we use 21 features and 2 tags (turnover status (S) and survival time (T)) to trained the model. Then, the classification model is trained and tested through the new testing data. In the second part, we use 23 features and 1 tag (turnover status (S)) to trained the model. It should be noted that these 23 features include the 21 features of the survival model, and the two new features of survival rate (f) generated in the survival analysis model and survival time (T) which used as a tag in the survival analysis model.

RFRSF Algorithm. According to the above framework, we propose our RFRSF employee turnover prediction algorithm, as shown in Algorithm 1. In Algorithm 1, OOBScore(RSF) and OOBScore(RF) are the out-of-bag estimation scores of RSF and RF respectively. These scores are used to find the best RSF models and RF models.

Algorithm 1. The RFRSF algorithm

```
Require:
    Train (F), Test (F)
Ensure:
    RSF, Score(RSF), OOBScore(RSF), RF, Score(RF), OOBScore(RF), Train(F+f),
    Test(F+f)
Initialize: Score(RSF)=0.5, OOBScore(RSF)=0.5, Score(RF)=[0,0,0,0], OOB-
Score(RF)=0
while OOBScore(RSF) does not meet the requirements do
    Train RSF with the algorithm Train(F)
    Out-of-package estimation of RSF to get OOBScore(RSF)
end while
Substituting RSF into the algorithm Test(F), calculate Score(RSF)
Substituting RSF into Train(F) and Test(F), calculate Train(F+f), Test(F+f)
while OOBScore(RF) does not meet the requirements do
    Train RF with the algorithm Train(F+f)
    Out-of-package estimation of RF to get OOBScore(RF)
end while
Test RF with the algorithm Test(F+f) to get Score(RF)
return RSF, Score(RSF), OOBScore(RSF), RF, Score(RF), OOBScore(RF),
Train(F+f), Test(F+f)
```

5 Evaluation

5.1 Dataset

The dataset used in this paper are crawled from one of China's largest online professional social platform, which contains employees' personal information, educational background, work experience, and platform activities. To analyze the data with survival model, we label each sample with its survival time and turnover status. Specifically, we inspect each job record to see if it has an end time (e.g., 2016-02), if so, the survival time is the duration from start time to end time and the sample is labeled 1 (indicating the occurrence of a turnover event), otherwise the survival time is the duration from start time to the time point when the user updates this job record, the sample is labeled 0 and the record is marked as right-censored. We remove records with no start time information. After data cleaning and preprocessing, we get 287,229 samples with 119,728 positive samples and 167,501 negative ones. We randomly divide the data set into training and testing sets according to a ratio of 7:3.

5.2 Feature Extraction

From the dataset, we extract 22 features for turnover prediction, which can be divided into six categories. Table 1 summarizes the features and their detailed descriptions.

Unlike the standard RSF survival model, in our dataset, an employee may have multiple job-hopping records. From the perspective of survival analysis, it means that the event occurs multiple times during the observation period for a single object, which is not supported by standard survival analysis models. To tackle this problem, we take the following two strategies:

i) First, we split an object with multiple events into several separate objects where each object is associated with exactly one event. For example, as shown in Fig. 2, during the observation period, employee A has two turnover events, we split the two events of A into two turnover events and marked as A1 and A2.
ii) Second, we align different events by converting absolute time into relative time. For example, as shown in Fig. 2, we align different events to make sure they have the same start time, in this way we can focus on the relative length of time before the event occurrence.

Table 1. Summary of selected features

Category	Feature	Description
Demographic	*gender*	Gender of employee
Job	*industry_type*	Employee industry type, such as education, IT, etc.
	cmp_scale	Company size, i.e., the number of employees
	position_level	Employee's position level at the company
Platform	*iteractions*	Number of online interactions with other user
	dongtai	Number of posts written by the employee
	guandian	Number of opinions expressed by the employee
	zhuanlan	Number of articles written by the employee
	dianping	Number of comments written by the employee
	likes	Number of likes received by the employee
	views	Number of views by other users
	recent_feeds	Number of recent feeds received by the employee
	influence	Online influence of the employee
	inf_defeat	Ratio of users defeated by the employee in terms of online influence ranking
	info_ratio	Online information integrity of the employee
	imp_tag_num	Number of impression tags given to the employee
	pro_tag_num	Number of professional tags given to the employee
Education	*degree*	The highest educational degree
	sch_type	The school type (in China) where the employee get the highest degree, such as 985, 211
Job change	*has_turnover_num*	Number of turnover records for the employee
	has_timelength	The time length the employee has worked for (including the current job)
	timelength	The time length of the current job
Survival	*survival_rate*	The survival rate generated through RSF model

After the aforementioned processing, we conduct a statistical analysis on the lengths of job records (measured in months), as shown in Fig. 3, where isexit=1 indicates a turnover event. The lengths of most of the job records are located between 1 to 24 months and the peaks of turnover behavior at year marks (after every 12 months). Also, we observe that the 13–24 month (1–2 year) interval has a significantly higher relative turnover rate. Therefore, we conclude that time length at the current job can be used as an important feature for turnover prediction.

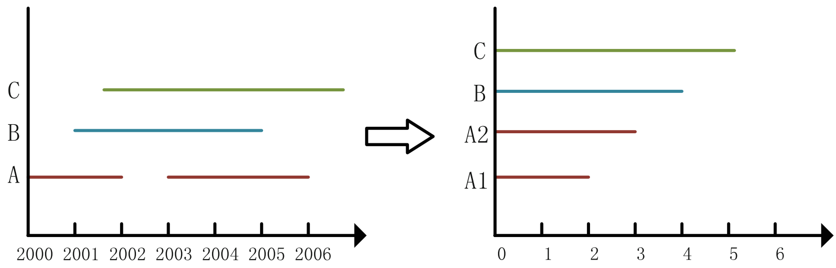

Fig. 2. The processing of objects with multiple events

5.3 Baselines

We include the following baselines for comparison:

Classic Machine Learning: The first group of baselines are classic machine learning methods without survival analysis feature (survival rate). We consider five representative machine learning models, which include Naive Bayesian (**NB**), Logistic Regression (**LR**), Decision Tree (**DT**), XGBoost (**XGB**) and Random Forest (**RF**). The model parameters are tuned to ensure a good overall performance.

Machine Learning with Cox: The second group of baselines combine classic machine learning methods with Cox proportional hazard model, which is proposed by Zhu et al. [22], we denote them as X+Cox, where X is the corresponding machine learning model.

Machine Learning with RSF: The third group of baselines combine classic machine learning methods with Random Survival Forests model, i.e., with the survival rate feature generated by RSF. We denote them as X+RSF, where X is the corresponding machine learning model.

5.4 Evaluation Metrics

We use *Accuracy*, *Precision*, *Recall*, *F1-score*, and *AUC*, which are widely used in classification tasks, as the evaluation metrics.

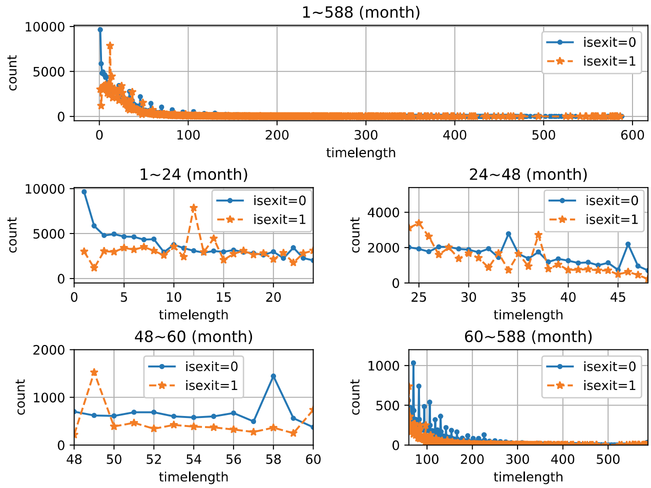

Fig. 3. The distribution of time

5.5 Results

Survival Analysis Results. We first compare the C-index scores of the RSF model with the Cox model. There are 21 covariates in the survival analysis and we perform hypothesis testing on these covariates. In the Cox model, only 7 covariates pass the testing, and it achieves the highest C-index score of 0.55 (which is unsatisfactory). We think this is mainly caused by the large amount of censored data in our dataset. On the contrary, the random survival forests model can use all the 21 covariates, and the C-index score is 0.68, which is higher than the Cox model.

To further analyze the differences between the two survival analysis models, we check their survival rate distribution, as shown in Fig. 4. It can be seen that the survival rate distribution of RSF is consistent with our common sense

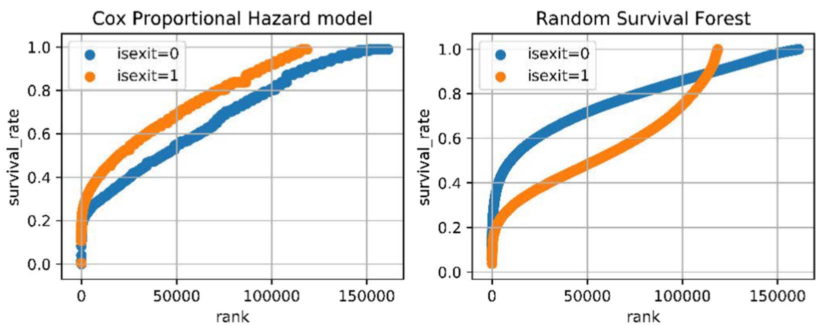

Fig. 4. The distribution of survival rate

that the survival rate of most turnover employees (isexit = 1) is significantly lower than that of the non-turnover employees. By contrast, the survival rate distribution of the Cox model obviously violates our intuition.

Classification Results. We compare the Accuracy, Precision, Recall, F1-score, and AUC index scores of different baselines, and the results are shown in Table 2, where the highest metrics are shown in bold. We see that the RFRSF model proposed in this paper achieves the best performance in terms of Accuracy, Precision, and F1-score and AUC, its Recall is also very close to the best value. Also, the RSF survival model can significantly benefit the machine learning methods than the Cox model, especially for the tree based models (DT, RF and XGBoost). We think this is mainly due to the tree-based structure of RSF.

Table 2. Classification results of different methods

Model	Accuracy	Precision	Recall	F1-score	AUC
NB	0.5181	0.4627	0.8706	0.6043	0.5653
LR	0.6400	0.6050	0.4266	0.5004	0.6114
DT	0.6246	0.5550	0.5641	0.5595	0.6165
XGBoost	0.7085	0.6845	0.5755	0.6253	0.6906
RF	0.7132	0.6945	0.5740	0.6285	0.6946
NB+Cox	0.6303	0.5616	0.5814	0.5714	0.6239
LR+Cox	0.6519	0.6210	0.4578	0.5270	0.6262
DT+Cox	0.6264	0.5585	0.5652	0.5618	0.6183
XGBoost+Cox	0.7046	0.6809	0.5700	0.6205	0.6868
RF+Cox	0.7106	0.6885	0.5787	0.6289	0.6931
NB+RSF	0.5183	0.4629	**0.8707**	0.6044	0.5655
LR+RSF	0.6497	0.6175	0.4493	0.5202	0.6228
DT+RSF	0.7915	0.7495	0.7608	0.7551	0.7873
XGBoost+RSF	0.8258	0.7941	0.7934	0.7938	0.8214
RFRSF	**0.8465**	**0.8217**	0.8315	**0.8174**	**0.8420**

Feature Importance Analysis. We further analyze the feature importance for the RFRSF and RF method, where the feature importance is measured by Gini Index [18] and the results are shown in Fig. 5. In the RFRSF model, the importance of survival rate is higher than 0.6, while the scores of other features are lower than 0.1. Compared to the RFRSF method, the importance differences between different features of RF method are less significant.

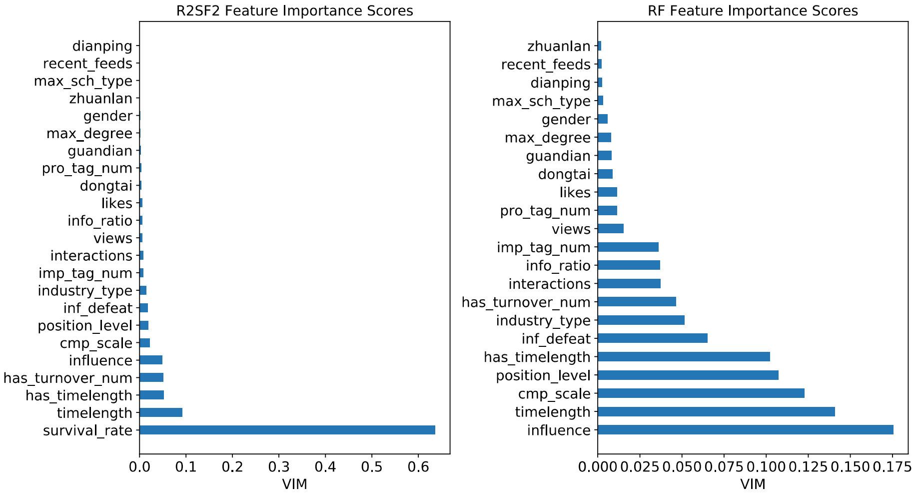

Fig. 5. Feature importance based on Gini Index

6 Conclusion

In this paper, we proposed an employee turnover prediction model by combining random survival forests with random forests. We constructed survival data based on employee's historical job records and then turned the employee turnover prediction problem into a traditional supervised binary classification problem. Experimental results on a real dataset verified the effectiveness of our model.

Acknowledgements. This work was supported in part by: National Natural Science Foundation of China (Nos. 61702059, 61966008), Fundamental Research Funds for the Central Universities (Nos. 2019CDXYJSJ0021, 2020CDCGJSJ041), Frontier and Application Foundation Research Program of Chongqing City (No. cstc2018jcyjAX0340), Guangxi Key Laboratory of Optoelectronic Information Processing (No. GD18202), Guangxi Key Laboratory of Trusted Software (No. kx201702).

References

1. Ajit, P.: Prediction of employee turnover in organizations using machine learning algorithms. Algorithms **4**(5), C5 (2016)
2. Allison, P.D.: Event History and Survival Analysis: Regression for Longitudinal Event Data, vol. 46. SAGE Publications, Thousand Oaks (2014)
3. Bao, L., Xing, Z., Xia, X., Lo, D., Li, S.: Who will leave the company?: A large-scale industry study of developer turnover by mining monthly work report. In: 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), pp. 170–181. IEEE (2017)
4. Baron, J.N., Hannan, M.T., Burton, M.D.: Labor pains: change in organizational models and employee turnover in young, high-tech firms. Am. J. Sociol. **106**(4), 960–1012 (2001)

5. Bigsby, K.G., Ohlmann, J.W., Zhao, K.: The turf is always greener: predicting decommitments in college football recruiting using Twitter data. Decis. Support Syst. **116**, 1–12 (2019)
6. Cai, X., et al.: DBGE: employee turnover prediction based on dynamic bipartite graph embedding. IEEE Access **8**, 10390–10402 (2020)
7. Carraher, S.M.: Turnover prediction using attitudes towards benefits, pay, and pay satisfaction among employees and entrepreneurs in Estonia, Latvia, and Lithuania. Baltic J. Manag. **6**, 25–52 (2011)
8. Chen, H., Jin, H., Wu, S.: Minimizing inter-server communications by exploiting self-similarity in online social networks. IEEE Trans. Parallel Distrib. Syst. **27**(4), 1116–1130 (2016)
9. Fang, M., Su, J.H., Wang, T., He, R.J.: Employee turnover prediction based on state-transition and semi-Markov-a case study of Chinese state-owned enterprise. In: ITM Web of Conferences, vol. 12, p. 04023. EDP Sciences (2017)
10. Fang, M., Su, J., Liu, J., Long, Y., He, R., Wang, T.: A model to predict employee turnover rate: observing a case study of Chinese enterprises. IEEE Syst. Man Cybern. Mag. **4**(4), 38–48 (2018)
11. Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests. JAMA **247**(18), 2543–2546 (1982)
12. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., et al.: Random survival forests. Ann. Appl. Stat. **2**(3), 841–860 (2008)
13. de Jesus, A.C.C., Júnior, M.E.G., Brandão, W.C.: Exploiting linkedin to predict employee resignation likelihood. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 1764–1771 (2018)
14. Li, H., Ge, Y., Zhu, H., Xiong, H., Zhao, H.: 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, pp. 917–925 (2017)
15. Liu, J., Long, Y., Fang, M., He, R., Wang, T., Chen, G.: Analyzing employee turnover based on job skills. In: Proceedings of the International Conference on Data Processing and Applications, pp. 16–21 (2018)
16. Mulla, Z.R., Kelkar, K., Agarwal, M., Singh, S., Sen, N.E.: Engineers' voluntary turnover: application of survival analysis. Indian J. Ind. Relat. **49**, 28–341 (2013)
17. Oentaryo, R.J., Lim, E.-P., Ashok, X.J.S., Prasetyo, P.K., Ong, K.H., Lau, Z.Q.: Talent flow analytics in online professional network. Data Sci. Eng. **3**(3), 199–220 (2018). https://doi.org/10.1007/s41019-018-0070-8
18. Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., Wang, Z.: A novel feature selection algorithm for text categorization. Expert Syst. Appl. **33**(1), 1–5 (2007)
19. Wang, J., Zhang, Y., Posse, C., Bhasin, A.: Is it time for a career switch? In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1377–1388 (2013)
20. Wang, P., Li, Y., Reddy, C.K.: Machine learning for survival analysis: a survey. ACM Comput. Surv. (CSUR) **51**(6), 1–36 (2019)

21. Yang, Y., Zhan, D.C., Jiang, Y.: Which one will be next? An analysis of talent demission. In: The 1st International Workshop on Organizational Behavior and Talent Analytics (Held in conjunction with KDD 2018) (2018)
22. Zhu, Q., Shang, J., Cai, X., Jiang, L., Liu, F., Qiang, B.: CoxRF: employee turnover prediction based on survival analysis. In: 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1123–1130. IEEE (2019)