Limjeerajarat, J., & Naparat, D. (2022). Preserving talent: Employee churn prediction in higher education. In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 22* (pp. 661-667). ICEB'22, Bangkok, Thailand, October 13-17, 2022

Preserving talent: Employee churn prediction in higher education

(Work-in-Progress)

Jariya Limjeerajarat [1]
Damrongsak Naparat [2,*]

*Corresponding author
[1] Computer Technical Officer, Chiang Mai University, Chiang Mai, Thailand, jariya.lim@cmu.ac.th
[2] Lecturer, Chiang Mai University, Chiang Mai, Thailand, damrongsak.naparat@cmu.ac.th

ABSTRACT

Retaining employees in a knowledge-based organisation, such as a university, is a significant challenge, especially as the need to keep knowledgeable workers is key to sustaining their competitive advantage. Knowledge is the organisations' and employees' most valuable and productive asset, but this intrinsic character leads to a high employee turnover. Often, universities learn about employees' imminent departure too late. To prevent the loss of high-performing employees and to detect the warning signs early, business firms have been using advanced data mining techniques to predict "customer churn". Recently these techniques have been used with "employee churn" in various industries, but not in higher education. This research bridges this gap by applying data mining techniques to predict employee churns in a university. The contributions of this research will be: 1) to identify critical factors that lead to talent losses; 2) to help universities devise appropriate strategies to retain their employees' talents.

Keywords: Employee attrition, churn prediction, higher education, data mining.

INTRODUCTION

Retaining employees in a knowledge-based organisation, like a university, is a significant challenge. For universities, retaining knowledge workers is key to sustaining their competitive advantage. Knowledge is the employees' most valuable and productive part, but this intrinsic character leads to turnover. Often, universities learn about imminent departure too late. Disruptions in higher education signify an increased need for a university to retain high-performing employees to remain competitive. For example, the growth of postsecondary alternatives to further education, including Massive Open Online Courses (MOOCs), industry-driven certification programs and coding boot camps, offers a broad range of options for learning. While online learning platforms, such as Coursera, provide additional channels for universities to reach students, they also invite competition to universities as companies are now offering similar content to those offered by universities.

Moreover, there has been a significant change in how people view their work and the actual purpose of work during and after the COVID-19 pandemic. Recently, 41% of employees globally are considering leaving their jobs, and 36% of those leaving their jobs do so without having their next job available (Frick et al., 2021). Therefore, organisations are going to need to update their 'retention of talent' strategy to retain valuable employees, and universities are no exception. As an organisation strives to keep its valuable employees or "top performers", human resource management becomes critical as it helps an organisation to preserve talents to sustain its competitive advantage (cf. Sooraksa, 2021).

Universities are competing for talent. In the age of a knowledge-based economy, the need for 'creative knowledge' workers has skyrocketed. Employees' skills have now become a critical determining factor for the survival of an organisation (Holdford, 2019). Realising the value of talented employees, companies have started to leverage 'people analytics' and digital technologies in different ways. For example, Uniliver, a European-based consumer goods company, uses digital solutions to increase administration and employee engagement. Saudi Aramco, Saudi Arabia's national petroleum and natural gas company, has adopted virtual reality (VR) and gamification for training their employees (Patwardhan et al., 2019). Microsoft uses "Workplace Analytics" to measure how work patterns across teams change to improve the employee experience. Google is among the pioneers who use people analytics intensely to help to profile its employees. Google has an array of initiatives, such as devising predictive models to forecast upcoming people management problems and opportunities, analysing people data to improve diversity and developing a mathematical algorithm, which proactively and successfully predicts which employees are most likely to become a retention problem (Grace, 2022).

While a growing number of examples and papers demonstrate how private companies are moving towards people analytics, less is known about how the public sector, particularly universities, utilises people analytics to retain top-performing employees. In addition, research on predictive analytics for employee churn is growing and is still attractive for academia and practitioners. An analysis of existing predictive employee churn research points towards an opportunity for future research since the nature of datasets and the context of an organisation significantly affect the prediction. However, it is unlikely to be a comprehensive predictive employee churn model that can be used across different organisations. Therefore, this research is designed to bridge this knowledge gap and create a predictive employee churn model for a public research university. The

outcomes from this research can: 1) identify critical factors that lead to losses of talent; 2) help universities devise appropriate strategies to retain their talents.

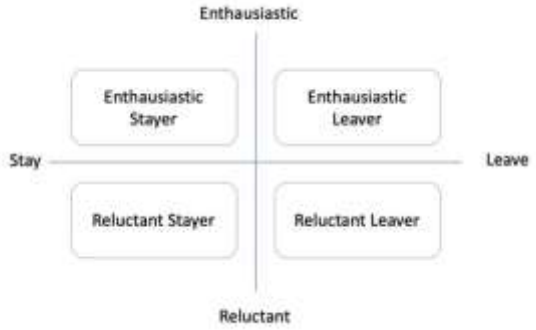
LITERATURE REVIEW

Employee Attrition

Retaining talented employees is a constant challenge for every organisation. Research studies seek to understand factors causing employees to leave an organisation. A wide range of factors has been suggested as predictors for an employee to quit. These factors include individual attributes such as personal characteristics and demographics, perceptions, and performance, as well as the attributes of the institutional environment, such as the organisation's structure, reward systems, and competition in a labour market – creating constraints and opportunities for individuals moving to a new job.

Theories on voluntary employee turnover have advanced consistently. In 1958, James March and Herbert Simon proposed *'perceived desirability'* and *'ease of movement'* as two primary constructs to suggest why employees stay with an organisation. These two constructs, later, are construed as *'job (dis-) satisfaction* and *'actual job alternatives'* consecutively. Later, Mobley (1997) suggests how job dissatisfaction culminates in a high turnover and explains how employees depart (Lee et al., 2017). Intermediate linkages such as subjective expected utility (SEU) analysis of the benefits and costs of seeking new jobs and search intentions have been proposed (Lee et al., 2017). Prior to job dissatisfaction, such as pay, centralisation, and alternative available employment, were identified as factors driving job (dis -) satisfaction. Based on cost/benefit analysis alone, although employees might be dissatisfied with their current jobs, they will not switch to new jobs if they anticipate future promotions in their existing job, or if there are worse payoffs from the new job.

Another strand of research examines *'why'* (rather than how) employees leave (e.g., Price, 1989; Beach & Connolly, 2005; Lee & Mitchell, 1994). According to Lee and Mitchell (1994), there are four paths an employee might take in terms of deciding whether to leave a job. The first path is based on the classic dissatisfaction-induced leaving. The remaining three alternative paths are based on *'shock'* – the jarring event(s) evoking thoughts of leaving. In Path 1, employees leave because a shock activates a pre-existing plan. Path 2 occurs when (an) undesirable event(s) prompt immediate quitting. And in Path 3, outside job offers cause employees to question their current job commitment. Hom et al. (2012) propose Proximal Withdrawal States Theory (PWST) to classify various mindsets about remaining or leaving an organisation. Simply put, there are four archetypal mindsets: 1) enthusiastic stayers (i.e., "I want to stay, and I can stay"), 2) enthusiastic leavers (i.e., "I want to leave, and I can leave"), reluctant stayers (i.e., "I want to leave, but I have to stay", and reluctant leavers (i.e., "I want to stay, but I have to leave"), which can be depicted as Figure 1.



Source: This study
Figure 1: PWST matrix.

Drawing from the above two strands of literature, a meta-analysis of research in turnover research during 1995 – 2008 (Holtom et al., 2008) shows six categories of antecedents of job turnover (see Table 1).

Table 1: Factors that lead to employee churn.

No.	Categories	Factors
1	Individual Difference	Ability, biodata/attribute, personality
2	Nature of the job	Routinisation, job scope, autonomy, role states
3	Traditional Attitude	Job satisfaction, met expectations, organisation commitment, job involvement,
4	Newer attitude	stress & strain, exhaustion & well-being, psychological uncertainty, change acceptance/perceptions, challenge/hindrance stressors
5	Organisational/Macro	Organisation size, group cohesion, demography, reward system, organisation culture, organisation prestige, climate, unit-level attitudes, normative/institutional pressures
6	Person-context interface	Justice, leadership, attachment/ties, person fit, realistic job preview, interpersonal relations, position history, socialisation

Source: Holtom et al., 2008

Lee et al. (2017) contend that previous research has proposed many factors that cause people to leave their work. Future research should embrace the varied research designs and analytical tools available to push the knowledge forward. The availability of big data and advanced data analytic tools provide greater opportunities for researchers to be innovative about gaining valuable insights into employee attrition.

Data Mining Employee Churn

Predictive models are now widely used in many business domains to predict occurrences of events, including predicting employee churn. The techniques for predicting employee churn have been ported from those used to predict *"customer churn"*. Customer churn is a key problem in highly competitive service markets (Saradhi & Palshika, 2011). A voluntary churn happens when customers stop using a company's current services and switch to a competitor company's services. It is critical to solve this problem because acquiring new customers is more complex and expensive. Losing customers leads to loss of revenue, which in turn negatively affects the 'bottom line'. In the same vein, *"employee churn"* is when an organisation loses its *"internal customers"* – i.e., the employees. Losing employees incur additional costs for an organisation, such as recruitment costs, onboarding costs, lost productivity, lost engagement and employee morale, training costs, and lost institutional knowledge.

Using data mining and machine learning allows an organisation to identify high-performing employees who are suddenly about to leave. Machine learning algorithms can be categorised into two groups: supervised machine learning and unsupervised machine learning. Supervised machine learning requires labelled input and output data, or the *"training dataset"*, during the training phase. The algorithm will learn about the relationship between input and output and construct a model, as this type of learning is called supervised learning, because it requires human oversight. Conversely, unsupervised machine learning does not require labelled input and output data. Algorithms of this type will recognise patterns in a raw dataset and construct a model based on the identified patterns. This approach is often used in the early exploratory phase to better understand the datasets (Provost & Fawcett, 2013). Employee churn prediction falls into the supervised learning category. The machine learning algorithms are fed with the training dataset containing multiple attributes and the label indicating which employee is already left and which are still with an organisation (Ekawati, 2019).

Existing literature shows that the commonly used data mining techniques for predicting employee churn are: Naïve Bayes, decision tree and random forests, logistic regression, and Support Vector Machine (SVM).

Naïve Bayes

In the literature on machine learning, naive Bayes is a common classification technique that has gained popularity for its effectiveness and simplicity (Mitchell, 1997). *A posteriori* probabilities are calculated using this procedure for each class. In predicting employee churn, these are the chances of seeing churn and non-churn, given an employee record. These *a posteriori* probabilities are calculated using the Naive Bayesian assumption and the Bayes rule for each class given a specific employee record. As a result, the learned function *'f'* is nothing more than a probability table. The conditional independence of the attributes used to describe employees is a crucial presumption in the Naive Bayes classifier. This method has been used in numerous case studies, particularly in the wireless telephony sector. However, it has only moderate success in predicting employee attrition (Fallucchi et al., 2020).

Decision Tree and Random Forests

Decision trees are popular in the literature, because of their simplicity in interpreting the discovered rules. Given a training data set, this learning technique produces a tree where each node represents an *n* attribute, and each branch represents a value for that property. The decision tree nodes are determined by the explanation power of the attributes (as evaluated by an information gain number) (Duda et al., 2001). The primary issue with the decision tree learning technique is its instability. Little changes in the training dataset typically result in substantial variances in classification performance. To solve this problem, Breimen proposes
"random forests" (Breimen, 2001). The objective of random forests is to generate numerous decision trees using sampled data (via bootstrap resampling) using only a subset of attributes. The final model is derived from the combination of the decisions of each of the constructed trees (using a voting-based approach).
Despite their instability, decision trees have been used in many case studies. Taiwan cellular telecommunication uses decision trees to estimate client turnover (Hung et al., 2006). The data was collected from 160,000 customers with an 8.75% customer attrition rate (14,000). Attributes for constructing the model include demographic information (age, tuner, and gender), billing payment history, call details, and customer care service characteristics. In this study, the performance of random forests is very high, at 98%. Random forests are also effective in other cases, such as predicting newspaper subscriptions (Lariviere & den Poel, 2005) and customer churn in the banking industry (Courssement & den Poel, 2008).

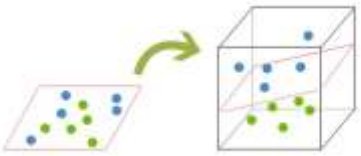
Logistic Regression

Logistic regression is a multivariate statistical technique that predicts group membership in a dichotomous dependent variable, given a collection of independent variables. Logistic regression is superior to other statistical methods when classifying data into a binary dependent variable. Multiple linear regression, for instance, assumes the normal distribution of the dependent variable and the linear relationship between the independent and dependent variables. These assumptions are not established during logistic regression model development. When the data are not normally distributed, and the dependent variable is binary

(i.e., predicting employee churn), logistic regression could be an effective technique, as the dependent variable is dichotomous. Quinn et al. (2008) use logistic regression to predict case worker and supervisor turnover in human services agencies. To develop the model, they used the first 429 cases as the training dataset and the remaining 109 cases to validate and test the model. The results show that the logistic regression model could predict employees who left the agency with 79% accuracy. However, the prediction was more accurate when considering the number of stayers only, at 92%. Attributes included in the analysis were demographic data, such as age, education, gender, and country, and several job-related attributes, such as job type, length of tenure, and current rank in the agency.

Support Vector Machine (SVM)

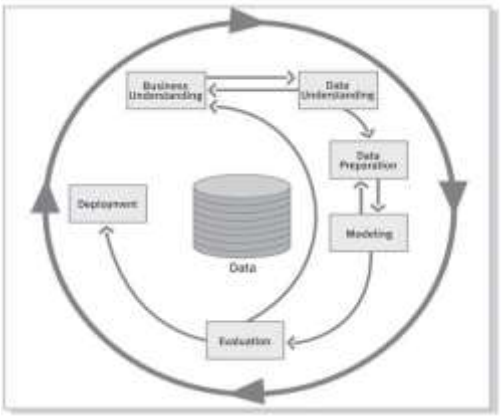
A Support Vector Machine (SVM) is a supervised learning algorithm that can be used for classification and regression, but it is commonly used in classification problems. The key concept of SVM is that it will find a "hyperplane" that best divides data into two classes. Support vectors are the nearest two data points (one from each side of the hyperplane) next to the hyperplane. Removing the support vectors would alter the position of the hyperplane. In essence, a hyperplane is a line that linearly separates and classifies a set of data (in a two-dimensional view), or it is a plane that classifies a collection of data (in a three-dimensional view) (see Figure 2). A distance between support vectors and a hyperplane (or a line) is called "margin". The primary goal of SVM is to maximise the margins of the support vectors so that it can separate two datasets as much as possible. Saradhi and Palshika's (2011) study shows that SVM, when used to predict employee churn, could outperform other algorithms, such as random forests and Naïve Bayes.



Source: www.towardsdatasicne.com
Figure 2: SVM hyperplane.

RESEARCH METHOD

This research follows the CRISP-DM reference model to create an employee attrition prediction in the context of higher education. CRISP-DM stands for Cross-Industry Standard Process for Data Mining. CRISP-DM, also funded by the European Commission, was intended to be industry-tool and application-neutral. Currently, a CRISP-DM consortium puts the CRISP-DM methodology forward, and this methodology is well accepted among practitioners and academia (Chapman et al., 2000). CRISPDM views a data mining process as a life cycle consisting of six stages (See Figure 3): 1) business understanding; 2) data understanding; 3) data preparation; 4) modelling; 5) evaluation; and 6) deployment. In reality, CRISPR-DM is proven to be an effective process. It at least provides general guidelines that are useful for planning, documentation, and communication (Wirth & Hipp, 2000). We will follow CRISP-DM up to step five since model deployment at the university level would require greater collaboration from different units and demand more significant resources.



Source: Cross-Industry Standard Process for Data Mining
Figure 3: CRISP-DM Life Cycle.

The Business Understanding Stage

In this stage, the business objective is clarified, and the criteria for success are clearly defined. For this research, the main goal is to use data mining techniques to find an appropriate model to predict employee attrition in a public research university. We collected data from one of the top research universities in Thailand. In 2022, the university has approximately 12,000 employees, classified into three main groups: faculties, supporting staff, and professional staff.

The Data Preparation Stage

We gathered data from the human resource database of the university mentioned previously. The database contains approximately 15,904 employee records, including 2,750 churners and 13,154 non-churners, accumulated from 2015 to 2022. The data include 18 attributes for each employee record, as in Table 2 below.

Table 2: Human Resource Dataset Attributes

No	Features/Characteristics	Description	Data Type
1	Record ID	Unique record ID	Categorical
2	Sex	Gender	Categorical
3	Nationality	Nationality	Categorical
4	WorkLineID	Employee's line of work (e.g., academic, supporting)	Categorical
5	MariageStatus	Marital status (e.g., single, married)	Categorical
6	EmployeeTypeID	Employee types (e.g., permanent, temporary)	Categorical
7	Education	Education level	Categorical
8	Age Range	Age range (range of 10)	Categorical
9	Work time	Years of stay with the university	Numeric
10	Salary Range	Salary range (range of 10,000)	Categorical
11	SubEmployeeType	Employee types (e.g., permanent, temporary, retired)	Categorical
12	Academic Main Group	Fields of study (e.g., health science, technology, social science and humanity)	Categorical
13	Welfare	Amount of financial benefit received each year	Numeric
14	LeaveDay	Number of annual leave days	Numeric
15	Performance	Performance evaluation results (e.g., excellence, good, fair, poor)	Categorical
16	Salary Raised (percentage)	Percentage of salary raised	Numeric
17	Tenure	Tenure position or not	Categorical
18	Stay/Left	Stay or left	Categorical

Source: This study.

One-hot Encoding

One-hot encoding is a method to convert categorical data variables to improve a model's prediction and classification accuracy. Many machine learning models do not take categorical values as their inputs, so they must be converted into numbers. However, the conversion could hamper the prediction model's performance, because the model might see these values as having an ordinal relationship while they do not. One-hot encoding is designed to solve this problem. The one-hot encoding creates a new binary feature for each possible category and assigns a value 1 to the feature of each sample that corresponds to its original category. One-hot encoding is a crucial feature engineering procedure. For example, a feature containing values for three colours, "red", "green", and "blue", can be encoded into a three-element binary vector as Red: [1, 0, 0], Green: [0, 1,0], Blue: [0,0,1]. It should be noted that one-hot encoding might not be appropriate for a feature that has too many categorical values. When the cardinality of the categorical features is large, a dictionary that maps categorical features will be large and can significantly strain a computer's memory resources (Weinberger et al., 2009). Having acknowledged this limitation, we will convert all features with categorical attributes using the one-hot encoding technique, since the cardinality of these features is not considered as high, and our computing resources can handle this amount and complexity of data.

Feature Selection

Feature selection will be used to eliminate redundant features to improve model accuracy. This process could involve many statistical techniques. The process involves building various models with different subsets of training features to build a more accurate classification model. For example, Trivedi (2020) performs a study on the credit scoring model with a different feature selection approach. Using freely available data - the German Credit dataset, the researcher uses three feature selection techniques (i.e., Information Gain, Gain Ratio, and Chi-Square) to choose the best subset of features from the dataset. After comparing the three techniques, she found that the Chi-Square feature selection is the most suitable for the German Credit dataset. This result should be interpreted cautiously, since it has not been generalised to other datasets. We will perform the feature selection process to firstly improve the model's accuracy and, secondly, to identify key factors that lead to voluntary leave.

The Modelling Stage

In this phase, we will use four machine learning algorithms, including Naïve Bayes, decision trees and random forests, logistic regression, and SVM, commonly adopted techniques for predicting employee churn. At this stage, we will need to calibrate our models and optimise the parameters of each model. Also, we will have to move back and forth between this stage and the data preparation stage, since each model can handle certain data types. Data conversion and data transformation are expected.

The Evaluation Stage

At this stage, we will compare the results from the four machine-learning algorithms and examine which algorithm yields the best outcome for our employee churn prediction problem. We will use the *"area under the receiver operating characteristics curve"* (ROC-AUC) to measure the performance of the models. AUC is a general measure of 'predictiveness'. It is preferable to other metrics (such as error rate), because it measures the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one (Punnoose & Ajit, 2016). We will use the *"lift chart"* to visualise the improvement of a particular model when compared with a random guess.

CONCLUSION

This research establishes that it is necessary for knowledge-based organisations, particularly universities, to retain their highperforming employees. These employees are essential for universities to sustain their competitive advantages in the era of highly competitive and disruptive environments. Universities need to re-consider their people management strategies and prevent unnecessary losses of employees. Machine learning and data mining techniques are helpful tools that allow universities to predict the voluntary leaving of high-performing employees. Nonetheless, research in predictive analytics in employee churn is still scarce. There is still a need to improve the accuracy of employee churn prediction, and there should be more cases from a wide range of industries (Ekawati, 2019). To advance this field of study, we will create a predictive model for employee churn in the context of a public research university. The results of our research will be beneficial in at least two ways. First, from the theoretical perspective, it will allow us to identify the key factors that lead to voluntary leave. Second, from a practical standpoint, the university can use the results as input to initiate appropriate people strategies to retain valuable employees and their talents.

REFERENCES

Beach, L. R., & Connolly, T. (2005). The psychology of decision making: People in organizations. Sage Publications.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide (). The CRISP-DM consortium.

Coussement, K., & Poel, D. V. D. (2008). Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers. In Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium (No. 08/527; Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium). Ghent University, Faculty of Economics and Business Administration. https://ideas.repec.org/p/rug/rugwps/08-527.html

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern Classification (2nd ed). Wiley. http://public.eblib.com/choice/publicfullrecord.aspx?p=699526

Ekawati, A. D. (2019). Predictive Analytics in Employee Churn: A Systematic Literature Review. Journal of Management Information & Decision Sciences, 22(4), 387–397.

Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. Computers, 9(4), 86.

Frick, J., George, K. C., & Coffman, J. (2021). How to Attract Top Tech Talent. Harvard Business Review. https://hbr.org/2021/11/how-to-attract-top-tech-talent

Grace, E. (2022). How Google is using people analytics to completely reinvent HR. Peoplehum. https://www.peoplehum.com/blog/how-google-is-using-people-analytics-to-completely-reinvent-hr (accessed 18 February 2022).

Holford, W. D. (2019). The future of human creative knowledge work within the digital economy. Futures, 105, 143–154. https://doi.org/10.1016/j.futures.2018.10.002

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Eberly, M. B. (2008). 5 Turnover and Retention Research: A Glance at the Past, a Closer Review of the Present, and a Venture into the Future. Academy of Management Annals, 2(1), 231–274. https://doi.org/10.5465/19416520802211552

Hom, P. W., Mitchell, T. R., Lee, T. W., & Griffeth, R. W. (2012). Reviewing employee turnover: Focusing on proximal withdrawal states and an expanded criterion. Psychological Bulletin, 138(5), 831.

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. Expert Systems with Applications, 31(3), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080

Lee, T., Hom, P., Eberly, M., Li, J., & Mitchell, T. (2017). On The Next Decade of Research in Voluntary Employee Turnover. The Academy of Management Perspectives, 31, amp.2016.0123. https://doi.org/10.5465/amp.2016.0123

Lee, T. W., & Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary employee turnover. Academy of Management Review, 19(1), 51–89.

Mitchell, T. M. (1997). Machine learning (Vol. 1). McGraw-hill New York.

Patwardhan, R. S., Hamadah, H. A., Patel, K. M., Hafiz, R. H., & Al-Gwaiz, M. M. (2019). Applications of Advanced Analytics at Saudi Aramco: A Practitioners' Perspective. Industrial & Engineering Chemistry Research, 58(26), 11338–11351. https://doi.org/10.1021/acs.iecr.8b06205

Price, J. L. (1989). The impact of turnover on the organization. Work and Occupations, 16(4), 461–473.

Provost, F., & Fawcett, T. (2013). Data Science for Business. 409.

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 5(9). https://doi.org/10.14569/IJARAI.2016.050904

Quinn, A., Rycraft, J. R., & Schoech, D. (2008). Building a Model to Predict Caseworker and Supervisor Turnover Using a Neural Network and Logistic Regression. Journal of Technology in Human Services, 19(4), 65–85. https://doi.org/10.1300/J017v19v04_05

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. Expert Systems with Applications, 38(3), 1999–2006. https://doi.org/10.1016/j.eswa.2010.07.134

Sooraksa, N. (2021). A Survey of using Computational Intelligence (CI) and Artificial Intelligence (AI) in Human Resource (HR) Analytics. 2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), 129–132. https://doi.org/10.1109/ICEAST52143.2021.9426269

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in Society, 63, 101413.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., & Attenberg, J. (2009). Feature hashing for large scale multitask learning. Proceedings of the 26th Annual International Conference on Machine Learning, 1113–1120. https://doi.org/10.1145/1553374.1553516

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 1, 29–39.