ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Predicting employee attrition with a more effective use of historical events

Abdel-Rahmen Korichi[1,2], Hamamache Kheddouci[1] and Daniel J West[2]

1- Université de Lyon - LIRIS UMR 5205
Bâtiment Nautibus 43, bd du 11 novembre 1918, 69622 Villeurbanne - France
2- Panalyt Pte. Ltd. - Singapore

Abstract. Attrition prediction research typically focuses on constructing models that involves one observation per employee over a limited time period, while the rest of the employees are discarded. Time-series attributes are transformed to non-time-series ones by applying statistical operations (e.g. sum, max, etc.). Such methods result in information loss and therefore less effective predictions. In this paper, we introduce a dynamic approach to employee attrition prediction, leveraging the longitudinal nature of the data, and allowing the models to generalize across behaviors and providing a closer estimate of the employee risk of leaving.

1 Introduction

Employee attrition is a reality that catches the attention of every business, no matter what they do. One of the most effective ways to tackle this issue is to build a predictive model - based on the wide variety of human resources (HR) data available [1] - to identify employees who are showing an intention to leave voluntarily. These models can vary in how they are built and operated but essentially follow the same principle - take a sample of employees that have already quit the company and use their data to build a model to highlight with an adequate advance who is likely to leave the company in the near future. By identifying the employees at risk and by understanding why employees choose to end their employment with the company, the latter can make improvements to reduce existing attrition problems.

To date, attrition prediction research based on HR data typically focuses on constructing classification models that are trained using only one observation per employee over a limited period of time, and do not consider its variation over time, discarding employees who have left earlier [2, 3, 4]. Historical records that change over time like the employee salary or the grade are transformed into non-time-series ones by applying statistical operations like the sum, the average, or looking at the latest values. Such approaches result in information loss and therefore less effective predictions.

In this paper, we introduce a dynamic method for generating training data from employee records and we use it to predict who is at risk of leaving the company within the next 6 months. By categorizing the employees in different 'buckets of risk' from *low risk* to *high risk*, our goal is to identify employees with high attrition risk to prevent them from leaving. A few studies have considered different approaches for attrition prediction. However, to the best of

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

our knowledge, none of them applies our framework to solve the employee attrition prediction problem. The obtained model for the prediction of employee's attrition is tested on an anonymized dataset from a real company provided by Panalyt Pte. Ltd. [5], which includes HR information from 1650 terminated employees.

2 Related work

Many researchers have made a lot of efforts to better understand which features (e.g. job satisfaction, compensation, engagement, etc.) are most influential in predicting employee attrition [6, 7], and a lot of studies have tried different approaches to predict voluntary attrition. Particularly, machine learning classification algorithms like decision Trees (DT), random forests (RF), gradient boosting trees (GBT), logistic regression (LR) or support vector machines (SVM) [8].

Yahia et al. (2021) [2] provide an extensive benchmark of different classification techniques. All these methods have in common that they use only one observation per employee selected during a defined period - usually a few months. As an example (Fig. 1), here employees 1, 3, and 4 are still active after the period considered ($T2$) and they would be labelled as active (0). Employees 2 and 5 are terminated during the period considered, and they joined the organization before the period considered, so the would be labelled as terminated (1). Employee 3 joins the organization during the period considered and is therefore discarded. Anyone else who left the organization before $T1$ or who joined after $T1$ is also discarded. For terminated employees, the observation used is the last one before they left the organization. For active employees, the observation used is as of the last month of the period considered (black dots).

The approach above does not consider the variation of the employees' features over time and is limited for a specific time period, which raises many issue: firstly, many employees are completely discarded of the training datasets. Secondly, for each employee, the information carried in only one observation can simply not contain the time-varying features (e.g. salary over time, grade over time, etc.). Thirdly, the trained models are not able to tell within which time period the employees are at risk to leave (will they leave within 3, 6 months?), which makes it unpractical for organizations to use. Finally, this approach does not consider the seasonality and how different periods of the year affect the risk of attrition.

More studies can be found in the area of customer churn, and as the approach to predict employee attrition is very similar to the case of customer churn [9], it helps us to deduce other methods. Ali et al. [10] show that using multiple training observations per customer from different time periods (Multiple Period Training Data) increases the predictive accuracy of churn models compared with the traditional approach of using only the most recent observation. However, Ali et al. acknowledge that their framework has a few issues. One of them is the potential lack of independence introduced by multiple observations with (almost) no change, resulting in many duplicates and an imbalanced training dataset. Another limitation is that instead of extracting multiple observations

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

from the whole tenure of the customers, they reduce it to a limited period, which implies that a lot of customers are again discarded.

This paper makes a contribution by proposing a framework that will exploit different observations per employee, making use of the longitudinal nature of the data, without the negative effects of the methods described above.

3 Method

We define *Employees at risk* a discrete variable as follows. Records where the value *number of months before the termination date* is smaller than 6 are labelled as 1 (at risk), and records where the value *number of months before the termination date* is greater than 6 are labelled as 0 (not at risk).

$$Employees\ at\ risk = \begin{cases} 1\ if\ \#\ months\ before\ the\ termination\ date <= 6 \\ 0\ if\ \#\ months\ before\ the\ termination\ date > 6 \end{cases}$$

One of the issues with the above labelling is that employees with a long tenure in the company can have many times more records compared to others. Another issue is that it usually takes at least a few month to see a noticeable change in the employee records, which makes the records very redundant and the dataset imbalanced. Thus, we use a sampling method to tackle those issues and we will compare the performance of the model using our sampling method and using only one observation per employee. For the records labelled as 0 (not at risk) and for each employee, we keep up to 4 equidistant records between the earliest record and T-6, with a minimum distance of 3 months between the records. For the records labelled as 1 (at risk), we keep up to 3 records : 2, 4, and 6 months before the termination date ($T-2$, $T-4$ and $T-6$). We need to keep different records for the model to be able to capture any change at different stages of the employees' employment in the company. Figure 2 helps us to visualize how the sampling works. The black dots represent the observations labelled as at risk, and the white dots represent the observation when the employees are not at risk.

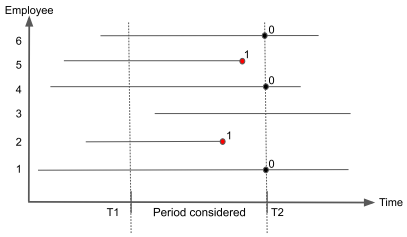

Fig. 1: Method using one observation per employee

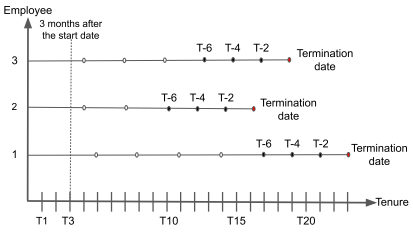

Fig. 2: Method using multiple observations per employee

Regarding the model, Ajit et al. [4] shows that to predict attrition, the XGBoost classifier performs better compared to Logistic Regression, Random Forest, KNN and others. XGBoost is capable of handling the noise and the null

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

values of the dataset as compared to the other classifiers, overcoming important challenges [11]. In light of these reasons, we will train our model using XGboost.

4 Experiments

4.1 Dataset

For our experiment, we use an anonymized dataset provided by Panalyt Pte. Ltd. [5]. The dataset contains monthly records of 1650 permanent employees, from their start date to their termination date. As the probation period is too unpredictable (many people leave during their onboarding because there is simply no good fit), we decide to exclude the first 3 records of each employee. The resulting dataset (that we name D1) contains data from 1650 employees that have already left the company, with a total of 56464 monthly records from 4 months after their start date to their termination date.

For the purpose of our study, the dataset is reduced to 15 attributes: *employee id* (unique identifier of the employee), *effective date* (date of the record), *start date* (date on which the employee started), *termination date* (date on which the emploee left), *tenure* (time since the employee joined the organization), *month of starting date* (month when the employee joined), *month of the observation* (month of the record), *salary* (current salary of the employee),*time since last salary* (number of months since no salary increase), *grade* (current grade of the employee), *time since last grade* (number of months since no grade increase), *median salary* (median salary of the employee's team), *gender* (gender of the employee), *age* (age of the employee), *months before the termination date* (number of months between the effective date and the termination date).

The features *number of months since no salary increase* and *number of months since no grade increase* are created using the historical records. The reduced dataset (named D2) contains 8755 records, from which 4948 are labelled 'at risk' and 3807 'not at risk'. 3338 out of 8755 records contains missing values, which is very common in HR dataset, and XGBoost is particularly convenient as it can handle missing values. To train the model, we exclude the variables employee id, effective date, start date, termination date and months before the termination date. The training dataset contains 3 types of variable:

- Static variables that never change: *gender*, *month of starting date*
- Dynamic variables that can change overtime: *age*, *tenure*, *salary*, etc.
- Lags: *time since last salary*, *time since last grade*

Dynamic variables and lags are particularly useful to organizations as it can tell them what to do to keep an employees at risk of leaving. For instance, if a manager knows that an employee is at risk of leaving mainly because he/she hasn't been promoted for a long time, the manager can take action.

4.2 Results

We apply a 10-fold cross-validation to provide a robust estimate of the performance of our model (Table 1). To avoid any data leakage, we make sure that

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

different records from the same employee are not used in the training datasets AND the testing datasets. Apart from the accuracy, all the other metrics are higher for the reduced dataset (D2). With a recall of 0.2, the model trained on D1 fails to identify the employees at risk compared to the model trained on D2.

Table 1: Model results

Dataset	Accuracy (std)	Precision (std)	Recall (std)	AUC (std)
D1	0.83 (0.00)	0.64 (0.03)	0.20 (0.03)	0.59 (0.1)
D2	0.72 (0.01)	0.74 (0.01)	0.79 (0.02)	0.71 (0.1)

If we pay attention to the model trained on D2 and the distribution of the probabilities (Fig. 3), we see that the model's predictions seem very accurate when the probabilities are smaller than 0.20 and greater than 0.80. When we take a closer look to the proportion of correct predictions for different probabilities (Fig. 4), we see that the higher the probability, the more confident the model is that an employee is at risk of attrition - it reaches 96.2% of accuracy when the probabilities are greater than 0.9. On the other hand, the lower the probability, the more confident the model is that an employee is not at risk. The model is less accurate when the probabilities are between 0.4 and 0.6, which makes sense.

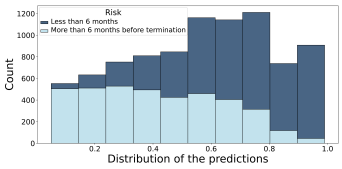

Fig. 3: Distribution of the predictions by class

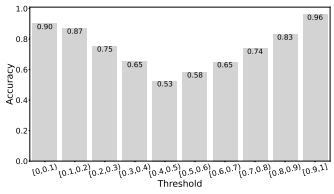

Fig. 4: Performance of the model for different probability buckets

We propose to have different 'buckets of risk' depending on the probabilities: very low Risk ([0,0.2)), low risk ([0.2,0.4)), intermediate risk ([0.4,0.6)), high risk ([0.6,0.8)) and very high risk ([0.8,1]). Using this classification can help organizations to highlight where to focus their attention.

Another question that we asked ourselves is how does the model perform for the records at T-2, T-4 and T-6? This time, we trained the model on 80% of the dataset and we kept track of which record were at T-6, at T-4, and at T-2 to see the distribution of the probabilities returned for these records (Fig. 5).

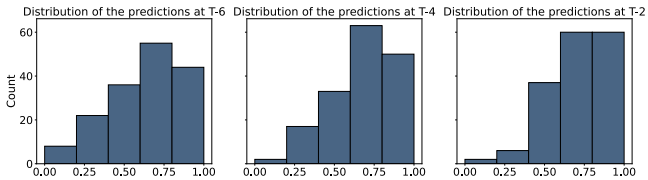

Fig. 5: Performance of the model at T-6, at T-4, and T-6

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

With an average score of 0.64 at T-6, 0.68 at T-4 and 0.72 at T-2, we notice that as expected, the closest we are to the termination date, the more confident the model is that the employee is at risk of leaving.

5 Conclusion

In this paper, we proposed an effective approach that takes advantage of the longitudinal nature of the data to solve the employee attrition prediction problem. Our sampling method reduced the dataset and we demonstrated that the model performs significantly better after sampling. We also showed that on average, the risk score gets higher and higher during the last 6 months before an employee leaves. Finally, using the probabilities returned by the model, we proposed a bucketization method to highlight different risk profiles for different employees. Further research would allow us to explore other methods of sampling, looking at the impact of training the data with more or less records. Another research could be to study more deeply the probabilities returned by the model and if, for each employee, a trend could be identified.

References

[1] J. Harris, E. Craig, and D. Light. Talent and analytics: new approaches, higher roi. *Journal of Business Strategy*, 32:4–13, 2011.

[2] Nesrine Ben Yahia, Jihen Hlel, and Ricardo Colomo-Palacios. From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access*, 9:60447–60458, 2021.

[3] Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. Predicting employee attrition using machine learning techniques. *Computers*, 9(4):86, 2020.

[4] Pankaj Ajit. Prediction of employee turnover in organizations using machine learning algorithms. *algorithms*, 4(5):C5, 2016.

[5] People analytics — panalyt. https://www.panalyt.com/.

[6] Thomas W Lee, Peter Hom, Marion Eberly, and Junchao Li. Managing employee retention and turnover with 21st century ideas. *Organizational dynamics*, 47(2):88–98, 2018.

[7] Alex Frye, Christopher Boomhower, Michael Smith, Lindsay Vitovsky, and Stacey Fabricant. Employee attrition: What makes an employee quit? *SMU Data Science Review*, 1(1):9, 2018.

[8] Yue Zhao, Maciej K Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. Employee turnover prediction with machine learning: A reliable approach. In *Proceedings of SAI intelligent systems conference*, pages 737–758. Springer, 2018.

[9] Sepideh Dolatabadi and Farshid Keynia. Designing of customer and employee churn prediction model based on data mining method and neural predictor. pages 74–77, 07 2017.

[10] Özden Gür Ali and Umut Aritürk. Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Syst. Appl.*, 41:7889–7903, 2014.

[11] Shital Kakad, Rucha Kadam, Pratiksha Deshpande, Shruti Karde, and Rushabh Lalwani. Employee attrition prediction system. *Int. J. Innov. Sci., Eng. Technol*, 7(9):7, 2020.