Who's Next: Evaluating Attrition with Machine Learning Algorithms and Survival Analysis

Jessica Frierson(✉) and Dong Si(✉)

University of Washington, Bothell, WA 98011, USA
{jess2018, dongsi}@UW.edu

Abstract. Every business deals with employees who voluntarily resign, retire, or are let go. In other words, they have employee turnover. Employee turnover, also known as attrition can be detrimental if highly valued employees decide to leave at an unexpected time. This paper aims to find the employee(s) that are most at risk of attrition by first identifying them as someone who will leave. Second, identify if their department increases the probability of them leaving. And third, identify the individual probability of the employee leaving at a given time. This paper found Logistic regression to consistently perform well in attrition classification compared to other Machine Learning models. Kaplan-Meier survival function is applied to identify the department with the highest risk. An attempt is also made to identify the individual risk of an employee leaving using Cox proportional hazard. Using these methods, we were able to achieve two of the three goals identified.

Keywords: Machine learning · Attrition · Logistic regression
Survival analysis

1 Introduction

All companies experience a degree of attrition and high attrition rates are concerning. However, what is more concerning is having experienced, knowledgeable, high performing employees leave. This leads to a company having to hire and train someone new in the domain of the previous employee, at possibly a higher market rate. This could lead to a period of low productivity while the new employee ramps up with company policies and trainings [1]. Therefore, it is vital that any business, big or small know what contributes to their employees leaving. In addition, being able to identify those red flags can lead to identifying the potentially at-risk employees. Remedial steps can then be taken to balance out the contributing factors leading to attrition and lessen the probability of the employee leaving. This project will aim to identify those contributing factors based on sample data provided by IBM which was found on Kaggle, and to identify employees at risk of attrition. In addition, through survival analysis, attempt to answer: Who's next?

This paper aims to use the best model based on experimental runs to identify employees that are at risk of attrition. Principal component analysis (PCA) will be

F. Y. L. Chin et al. (Eds.): BIGDATA 2018, LNCS 10968, pp. 251–259, 2018.
https://doi.org/10.1007/978-3-319-94301-5_19

performed to choose the top N feature components that contribute the most to attrition. In addition, survival analysis is used on each department to identify which one has the highest probability of its members leaving the company at a given time. Lastly, we will use a different survival analysis algorithm to find the employee most likely to leave from the department with the highest attrition rates.

1.1 Related Work

Previous literature, primarily in management, social, and other fields, has explored the issue of attrition. Limited research has been done with identifying employees that have higher risks of attrition with machine learning. For instance, literature that identifies a better subset of attributes to give a better predictor of attrition has been explored by Chang [2]. He found that it can be difficult to choose and limit the features to use specifically when pulling data straight from HR databases since there is a variety of information available. Identifying and selecting a good set of attributes from databases is a start.

A problem with HR datasets is that it can become difficult to not overfit models [1]. One method to reduce overfitting is to forgo the accuracy and make the model more generic to accommodate a more varied dataset. Extreme Gradient Boosting is found to be more robust than other models due to its implementation of regularization [3]. A previous study from Singh et al. [1] went one step further than just identifying attrition risks. They not only identified employees at a higher risk of attrition, but also explored if proactive steps, such as a salary increase, lowered the attrition rates.

In addition, there are papers that explored the aspect of employee retainability. In the case of Ramamurthy et al. [4], interest was in creating models to help employers reduce attrition by identifying employees that are good candidates to be retrained in a new skill based on their current skill set. In another case, Saradhi and Palshikar [5], drew parallelisms between customer turnover and employee turnover. They evaluated customer models and applied them to employees. They found that some of those models can be used for employee attrition prediction.

There is also a lot of literature available for survival analysis, but those literatures do not cross over to the machine learning fields too frequently. Goli and Soltanian sought to compare the performance of support vector regression (SVR) with Cox model on both a simulated and real survival problem. They found that with certain parameters, the SVR outperformed the Cox model and with others– the performance increase was diminished [6].

2 Methods

Previous papers that use human resource data to explore attrition have mainly concentrated on comparing the performance of various machine learning models and not so much on applying the model. In other words, we have not found papers that apply their models to predict if there are specific departments, or employees that are at a higher risk of turnovers. This is what sets this work apart from other work. In addition, other work does not combine different statistical models to the issue of attrition in combination with machine learning models. Palshikar [5] came close by recommending as part of future work to use survival analysis models in combination with their learning models but did not implement it. As of this time, we have not come across another paper who has.

The approach for this paper, is to use various learning models to find the best performant model for classification. Figure 1 shows the process pipeline. The process begins with pre-processing. After preprocessing, the data is split 70/30 for training and testing sets. With this split, we use and compare six different classification models. These models are, Decision tree, Logistic regression, Support vector machines, Naïve Bayes, K Nearest Neighbors, and Neural Networks. All models will be evaluated based on their accuracy, precision, and recall. Once the models have been implemented, we apply two different survival analysis algorithms.

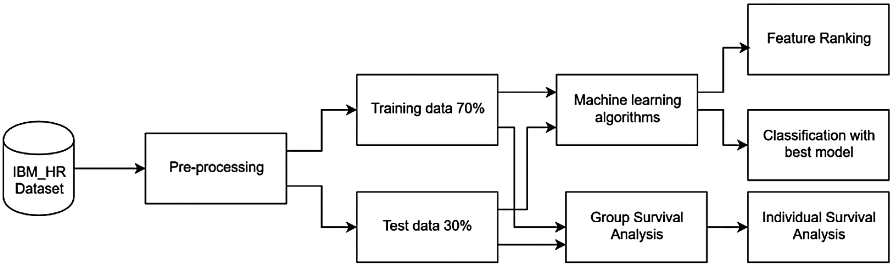

Fig. 1. This shows the ML pipeline for this term project. Preprocessing is needed to remove attributes that provide no value. The dataset is split into a training set and a test set. The training data is used for data familiarity, machine learning, and survival analysis. Test data is only used for machine learning and survival analysis.

Survival analysis is defined as: "A set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest" [7]. In our case the event of interest is attrition. Two survival analysis models are explored. The first model, Kaplan-Meier estimator, is to evaluate each department group. The Kaplan-Meier estimate involves computing the probabilities of an event occurring at a specific point in time [8]. The second algorithm is the Cox proportional hazards. This model will help identify the probability that an employee will leave, ultimately answering "Who's next."

2.1 Data Preprocessing

The dataset used is the IBM_HR simulated dataset found on Kaggle. This dataset contains 1470 data rows and 34 features. Data preprocessing is done by removing features from the dataset as seen in Table 1. These attributes contained the same value for all the data rows or contained non-relevant identifying information such as Employee number. In addition, there are some attribute fields that where re-mapped from string text to enumeration values. Such fields can be seen in Table 2. This was done in consideration of the machine learning algorithms that will be applied since it's easier to process numbers than text.

Table 1. Attributes removed from the dataset.

Attribute	Reasons
Employee count	Employee count
All records had the same value of (1), no valuable information could	All records had the same value of (1), no valuable information could
Be gained from this attribute.	Be gained from this attribute.
Employee number	Employee number

Table 2. Mapping of attributes that need to be enumerated.

Attribute	Source values	Target values
BusinessTravel	Non_Travel, Travel_Rarely, Travel_Frequently	0, 1, 2
Department	Human resources, research & development, sales	1, 2, 3
EducationField	Human resources, life sciences, marketing, medical, other, technical degree	1, 2, 3, 4, 5, 6
Gender	Female, male	1, 2
JobRole	Healthcare representative, human resources, laboratory technician, manager, manufacturing director, research director, research science, sales executive, sales director	1, 2, 3, 4, 5, 6, 7, 8, 9
Marital Status	Divorced, married, single	1, 2, 3
Overtime	No, yes	0, 1

2.2 Factors Leading to Attrition

One of the research goals was to identify top contributing attributes for the simulated data. To get those top features, we used an ensemble tree classifier to extract the feature importance rankings. The algorithm for the ensemble is based on a perturb-and-combine technique, which is designed for trees [9]. The rankings can be seen in Table 3. The feature that holds the highest ranking is Overtime with a ranking of .79.

Table 3. Top ten features ranked based on the results from the tree classifier. The name of the attribute is followed by the ranking.

No.	Feature	Ranking
1	Overtime	0.791443
2	Age	0.512488
3	YearsWithCurrentManager	0.427158
4	MonthlyIncome	0.416188
5	TotalWorkingYears	0.414962
6	DistanceFormHome	0.403594
7	YearsAtCompany	0.368960
8	WorkLifeBalance	0.360206
9	DailyRate	0.347088
10	JobRole	0.346335

The following feature only holds a ranking of .51. If we consider features ranked .5 or higher as significant, we would only have those two features. All other features ranked below that arbitrary threshold value. However, these results are not consistent. The attribute Overtime is always the highest ranking and the last few features remain in their ranked position with each subsequent run. However, the features in between vary by one or two positions with multiple runs. After exploring the data and finding the rankings, we then focused on selecting a model for the classification objective.

2.3 Machine Learning Algorithm Comparison

The models we compare in this section are: Decision Tree, Logistic regression, Support vector machine, Gaussian Naïve Bayes, K-Nearest Neighbors, and Neural Networks. These models are selected because they are appropriate for supervised binary classification. The label to identify is Attrition, which contains 0 for data rows that remain in the company, and 1 for data rows that have left the company.

Python's scikit learn library contains implementations of the models we target to explore. We consume the library models by using the standard pattern by first fitting the training data and the training label. Then the fitted model is used to predict using the test data. From the predicted model, metric scores can be calculated. Table 4 contains each model's respective metrics. The model that best performed is Logistic regression. Logistic regression contains the highest accuracy with (.8752), precision (.8603), and recall scores (.8752). The least performant model is Neural Networks. It is assumed that this is because our data is not ideal for neural networks, which works best with sequential data such as image and audio files.

Table 4. Model comparison table.

Model	Accuracy	Precision	Recall
Decision tree	.7823	.7950	.7823
Log regression	.8752	.8603	.8752
SVM	.8458	.7153	.8458
G. Naïve Bayes	.8140	.8318	.8141
KNN	.8344	.7867	.8344
NN	.6938	.7504	.6939

The receiver operating characteristic curve (ROC) of all models can be seen in Fig. 2 along with the area under the curve (AUC) value. From this plot, Gaussian Naïve Bayes algorithm performs almost as well as Logistic regression. After conducting these runs, the best classifier model for the IBM_HR dataset is consistently Logistic regression.

After identifying the Logistic regression as the best model, we wanted to see if it can still be improved by including feature dimension reduction since the data contained the full feature set after the initial pre-processing of the data. We combined Principal Component Analysis with Logistic regression using sklearn pipeline make_pipeline library. The results are seen in a combined plot with standalone Logistic regression in

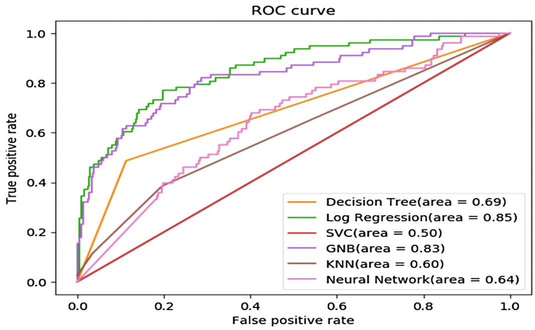

Fig. 2. Receiver operating characteristic curve of all six ML models with their area under the curve score.

Fig. 3. There is a slight almost negligible improvement using PCA with Logistic regression. Figure 3 shows the output from the best performing run of the piped model. When modifying the number of components for PCA, the model performance decreased if the number of components is reduced from the full feature set.

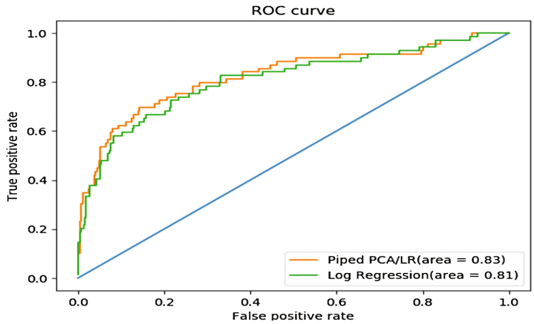

Fig. 3. Receiver operating characteristic curve of piped and not piped Logistic regression with their area under the curve score.

2.4 Survival Analysis

This survival analysis section aims to identify the department with the highest probability that its personnel will leave the company at a given time. To do this we used two different Python survival packages: 1. Sksurv and 2. Lifelines. Each package requires the data to be in a specific format. Therefore, we had to do more pre-processing to get the data in the right formats. Learning models typically contain a minimum of two parameters that need to be provided, 1. the data matrix/array and 2. the label array. With the survival packages, it is still expected to provide a data matrix/array, but instead of a label file, a censored event and time needs to be provided. We therefore created a censored event and time file using the Attrition label as the event (either 0 or 1) and added the YearsAtCompany attribute as the time parameter. With this modification we were able to get the overall probability of survival for the dataset as well as the probabilities of survival for each department using Kaplan-Meier model. From the plots provided in Fig. 4, we can estimate the probability or survival (no attrition) by choosing a time. For example, the probability of someone who just joins (t = 0) has a survival rate of or near 100. As time goes by the probability of survival decreases. Figure 4(b) contains the survival step functions for each of the three departments present in the data (refer to Table 2 to identify the departments). It's a bit tough to clearly see if department 1 or department 3 has more people leaving. However, it looks like department 3 has steeper and more frequent steps than department 1. Each step indicates that the event (attrition) was observed.

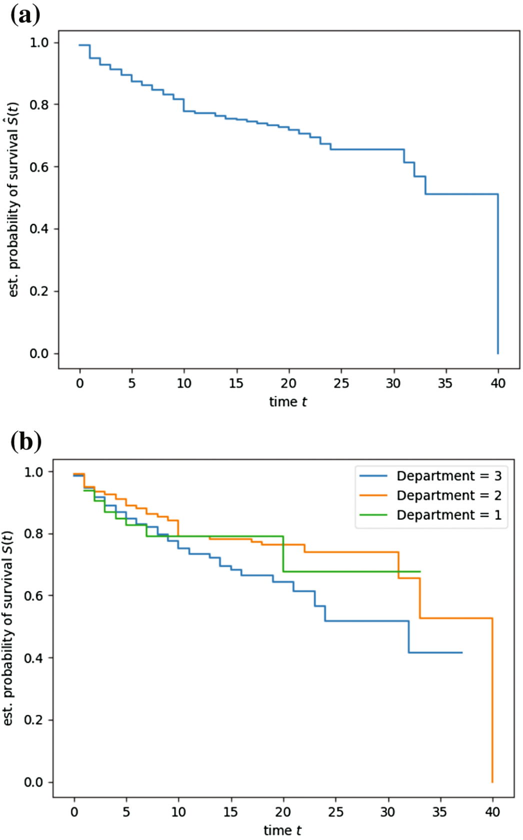

Fig. 4. Kaplan-Meier plotted as step functions. (a) Is the function for the overall dataset, whereas (b) is the plot for each department. The probability can be deduced by evaluating a specific time (t).

Figure 5 helps to distinguish between department 1 and department 3 as the department with the highest probability of attrition. In combination with the above Kaplan-Meier plots and the bar graph, it can be determined that Department 3 is the department that has the highest probability of employees leaving.

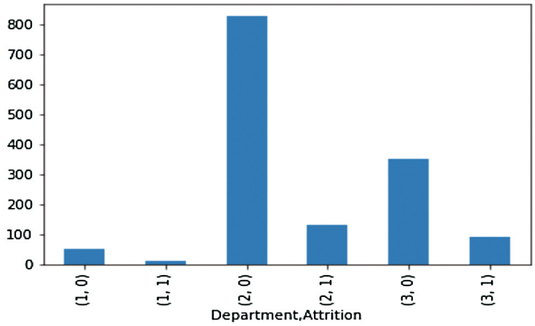

Fig. 5. Bar graph sets representing the number of people who stay in the company vs those that leave. Labels (1, 0) represent the department (1, 2, or 3) and Attrition (1, 0).

The last objective of this paper is to identify the employees from department 3 that have the highest probability of leaving next. To accomplish this, we decided to use the Cox proportional hazard model. The documentation for using the scikit-survival package mentioned that this function can be used to get the individual probability based on the "patient" or in our case, the employee. However, the algorithm kept failing with our dataset. The failing point was on an issue where a calculation received a NaN or infinite value. We decided to look for another survival package and came across Lifelines. After we installed the second package, we were still unable to get results using the Cox model. The Lifelines package also reported the same error of an expected value being NaN or infinite. This lead us to believe the issue was with our data. The data was either not in the proper format or missing additional information for the functions to yield results.

3 Conclusion and Discussion

This paper explored the possibility of predicting who within a company can be next to leave. This was done by first identifying an employee as someone who will leave or stay using binary classification models. The best performant model for this data is Logistic regression. Logistic regression consistently outperformed the other models and retained an AUC score of 85%. Next, Kaplan-Meier function was used to get the survival function for the overall company and each department. Through the Kaplan-Meier step function plot and bar graph, department 3 was shown to contain the least survival rate. Lastly, the Cox's proportional hazard model was attempted with two different libraries ultimately both failing to complete and reported the same issue. Since the same issue was reported, this lead us to conclude that the data was not in the right state for these functions.

The primary challenge we faced in this work, was the lack of in depth knowledge around survival functions. With a stronger background knowledge of the functions used, we may not have reached a blocking stop. Another challenge was getting the feature rankings to produce a consistent ranking. Each run to produce a ranking gave slightly different results.

Future work will require processing the data into the right expected format to complete the last objective of this paper. An alternative to using the library functions could be to implement our own survival functions. In addition, these methods can be applied to different datasets and measure the fragility of the approach taken in predicting who will be next to leave the company.

References

1. Singh, M. et al.: An analytics approach for proactively combating voluntary attrition of employees, In: IEEE ICDM, pp. 317–323 (2012)
2. Chang, H.: Employee turnover: a novel prediction solution with effective feature selection. In: Information Science and Applications, pp. 417–426 (2009)
3. Ajit, P., Punnoose, E.: Prediction of employee turnover in organizations using machine learning algorithms. Int. J. Adv. Res. Artif. Intell. **5**(9) (2016). http://dx.doi.org/10.14569/IJARAI.201.050904
4. Ramamurthy, K., et al.: Identifying employees for re-skilling using an analytics-based approach. In: IEEE ICDM, pp. 345–354 (2015)
5. Saradhi, V., Palshikar, G.K.: Employee churn prediction. Expert Syst. Appl. **38**, 1999–2006 (2011)
6. Goli, S., et al.: Performance evaluation of support vector regression models for survival analysis: a simulation study. Int. J. Adv. Comput. Sci. Appl. (IJACSA), **7**(6), 381–389 (2016)
7. Cornell University. Cornell Statistical Consulting Unit
8. Khanna, P., Kishore, J., Goel, M.K.: Understanding survival analysis: Kaplan- Meier estimate. Int. J. Ayurveda Res. **1**(4), 274–278 (2010)
9. Breiman, L.: Arcing Classifiers. Annals Stat. **26**(3), 801–849 (1998)