**Employee Retention with R Based Data Science Accelerators**

Le Zhang, Data Scientist, Microsoft, Graham Williams, Director of Data Science, Microsoft



Employee retention has been and will continue to be one of the biggest challenges of a company. While classical tactics such as promotion, competitive perks, etc. are practiced as ways to retain employees, it is now a heated trend to rely on machine learning technology for exploiting patterns with which company can understand their employees better.

Employee demographic data have been studied and primarily used for analyzing employees’ inclination of leaving a company [1]. Nowadays, as the proliferation of internet, employees’ behavior records in company IT system such as social media post can be leveraged for such kind of analysis [2]. In addition to that, novel cognitive computing technology based on artificial intelligence tool empower the HR department to predict staff churn before it actually happens.

This blog post introduces an R based data science accelerator that can be quickly adopted by data scientist to prototype a solution for employee attrition prediction scenario. The prediction is based on two types of employee data that can be collected by companies without much effort:

1. Static data which do not involve over time. This type of data may refer to demographic and organizational data such as age, gender, title, etc. The characteristics of this type of data are that, within a certain period of time, they do not change or solely change in a deterministic way. For instance, years of service of an employee is static as the number increments every year.
2. Second type of data is the dynamically involving information about an employee. Recent studies revealed that sentiment is playing a critical role in employee attrition prediction. Classical measures of sentiment require employee survey about their satisfaction with work. Social media posts become useful for sentiment analysis as employees may express their feelings about work on internet. Non-structural data such as text can be collected for mining patterns for segmenting employees with different inclination of churn.

Attrition prediction is a scenario that takes the above two types of data as input, and then identifies individuals that implicitly present their inclination of leaving. The basic procedure is to extract features from the available data, and build predictive models based on a training set with labels of employment status. Normally it can be formalized as a supervised classification problem, while the uniqueness is that population of employees with different employment status may not be equal. Training such an imbalanced data set requires resampling or cost-sensitive learning techniques. For sentiment analysis on unstructured data such as text, pre-processing techniques that extract analysis-friendly quantitative features should be applied. Commonly used feature extract methods for text analysis include word-to-vector, term frequency, or term frequency and inverse document frequency, etc. Algorithms for building the model depends on data characteristics. In case a specific algorithm does not yield desired result, ensemble techniques can be taken to further boost model performance.

R is a convenient tool for performing HR churn prediction analysis. A lightweight data science accelerator that demonstrates the process of predicting employee attrition, given both static and dynamic information of employees, is introduced in the GitHub repository. Codes of the analytics are embedded in an R markdown, which can be interactively executed step by step. Especially, sentiment analysis on employee review comments are performed for employment status segmentation. Text analytical APIs of Microsoft Cognitive Services are applied for such analysis and comparatively studied with popular R packages.

More about Microsoft Cognitive Services can be found at <https://www.microsoft.com/cognitive-services/en-us/>. R packages that wraps up text related APIs of Microsoft Cognitive Services can be found at <https://github.com/yueguoguo/Azure-R-Interface/tree/master/utils/cognitiveR>.

References:

1. Terence R. Mitchell, et al, “Why People Stay: Using Job Embeddedness to Predict Voluntary Turnover”, Academy of Management Journal, vol. 44, no. 6, pp 1102- 1121, December 1, 2001.
2. <https://hbr.org/2016/01/sentiment-analysis-can-do-more-than-prevent-fraud-and-turnover>