CSDA1050 Advanced Analytics Capstone Course

Project Proposal

Machine Learning in Employee Retention

Eugene (Yong Geun) Park

Introduction

I am a finance controller/HR manager at a small advertising agency. It is a creative agency where creative concepts are developed for clients' marketing purpose and developed ideas are delivered to public through various media means. It is a business that hugely relies on Human Resources. As much as it is important to have resources who can develop creative and effective marketing ideas, the business also requires resources who can execute the idea in market by developing partnership and trust with client.

While there definitely is a pool of resources with proven experience and well-known career milestones out in the market, recruitment from the pool is not always guaranteed solution for success. Availability, Suitability and Financial Feasibility are some of many factors that need to be considered for new hires and it is always very difficult to align them to our exact business situation.

As a result, it inevitably becomes very important to retain current Human Resources who are nurtured and trained to our current business model. Cost of losing existing resource incurs in many areas. Most immediately, the business gets exposed in handling client request which can be risk in losing current business. If a recruiting firm is engaged in finding a replacement, the fee is usually more than 20% of the salary of new hire. There also is transition period to have new hire trained, which consumes company resources and while then, the exposure to client relationship still extends.

Advertising is a very competitive field. Most ideally, the company wants to grow year-over-year to have scale to provide career advancement opportunities within the organization. However, what the company can offer to employees who seek further growth is very limited and we face challenges keeping valuable human resources.

Research Question

What are the primary factors that impact employee retention?

I want to be able to identify factors that have most impact employee retention. Based on findings, I wish to be able to identify what the company should focus on to reduce turn-over rate. In addition to identifying significant factors to employee retention, maybe I can identify employees who are at risk of leaving the company. And that can help build preventive measures, or plan new hires accordingly to least impact current operation.

Dataset

I have the record of employees since 2005. Age, gender, address, hire date, termination date, termination reason, starting salary, department, job title, job level, and education are the information I have in the dataset and I wish I can build a predictive model that can at least hint what needs to be focused for better employee retention.

There is a great advantage using a dataset I built. Although the dataset may not be complete, as I know the data by heart, I will be better at handling missing values and outliers. Also, by the end of the research, I might have better insight and comprehension.

There definitely is limitation to this dataset because it is a small operation and only has 127 employees record. I am unsure whether I can build a reliable model based on this. Aside from objective features from the existing dataset, I wish I had some subjective field as well, such as performance rating and satisfaction survey rating, which I do not have.

I would like to complement the research by using company's financial dataset. It is a much more comprehensive dataset than HR record. It may contribute in discovering new findings and help build employee retention strategy.

Methodology

Using R, I will start by applying some supervised machine learning techniques to my HR dataset to build predictive models. Decision Tree, Random Forest, Support Vector Machine and Logistic Regression are considered at the moment. Their accuracy will be compared and analysis will be done to decide what is most suitable.

In the process, my dataset will need cleaning. All fields containing personally identifiable information will be removed or replaced with suitable information for confidentiality. I foresee the need of adding, binning, deleting, reclassifying and filtering features.

I would also like to utilize financial data for the research purpose. Whereas my HR data has bigger emphasis on personal(internal) information, my financial data can provide external perspective. The dataset has company's sales, cost and profit information by month and client. Its changes in finance and client over time can potentially be linked to employee turn-over rate then maybe a model can be developed to tell how business performance may impact on employee retention.