Human Resource Turnover Prediction

Gautam Harinarayanan, Mayank Gulati, Matthew A. Lanham
Purdue University Krannert School of Management
gharinar@purdue.edu; mgulati@purdue.edu; lanhamm@purdue.edu

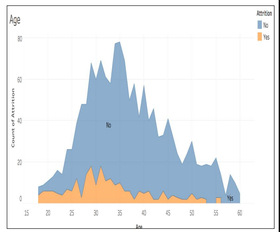
Abstract

In this study, we will be predicting employee turnover for an organization using linear and non-linear classification models, and find out the most important variables that affect turnover. This study is important because employee turnover is a major cost to an organization, and predicting turnover is at the forefront of needs of Human Resources (HR) in many organizations.

However, studies have shown that analytics and statistics are rarely used in the HR space and this opens up the possibilities for our project to be implemented. In order to effectively predict turnover, we used 4 different models such as Decision Trees, Support Vector Machine, and XGBoost. Modeling was performed using R language.

Introduction

Measuring employee turnover can be helpful to employers that want to examine reasons for turnover or estimate the cost-to-hire for budget purposes as well as understand how to run their HR department and how to retain their employees. Turnover should not ideally be eliminated, as a healthy turnover rate shows the willingness of a company to stay fresh and hire new minds. A high turnover rate however, shows employee dissatisfaction and decreases productivity.



If a company can not only understand its turnover rate, but also have a capability to predict it and find the underlying factors that affect it, there will be much more energy and time the company will have to focus on its customers and its products. While there is an increasing demand for analysts and statisticians, the use of analytics in the HR space is minimal. Due to this reason, HR analytics is still a domain yet to be explored entirely, which makes this project a vital tool and a necessity in the HR space.

Shown on the right is an employee satisfaction index for various job roles. From the index we expect that job roles with lowest satisfaction would have highest attrition rate.

Answering this question would help us understand drivers of attrition, giving the organization necessary intelligence.

Satisfaction

Job Role	
Healthcare Representative	2.7863
Human Resources	2.5577
Laboratory Technician	2.6911
Manager	2.7059
Manufacturing Director	2.6828
Research Director	2.7000
Research Scientist	2.7740
Sales Executive	2.7546
Sales Representative	2.7349

Methodology

Data Sources

The dataset used for modelling is IBM Employee dataset available on Kaggle and provided by IBM. The data consisted of around 1,500 employees throughout the United states, each observation had 35 features.

Exploratory Data Analysis

To understand the drivers for employee turnover, we analyzed the data by plotting turnover against key drivers such as Job Satisfaction, Education, and Performance rating. The data was not temporal, as date of employment or resignation date was Unknown. Hence, we could not deduce if the company was going through a rough patch or if there was a period where resignations were high due to economic reasons. In the process, we discovered that most employees were satisfied with the work environment and most got promoted in last 4 years. All performance ratings were either 3 or 4, which indicated that either the data was improperly recorded or a sub-set of data was provided for analysis.

Figure 2 outlines our study design, starting from data collection, data cleaning, data pre-processing, feature creation and selection, model/approach selection, cross-validation design, and model assessment/performance measures.

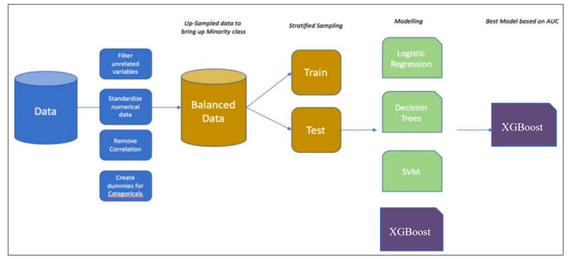


Figure 2. Study Design

Data Preparation

Data transformation was done to prepare data for modeling. Some of the independent variables, such as “Department”, “Education”, “Job Role” were dummy transformed for model building. We removed multicollinearity among independent variables and removed colinear features, such as ‘Performance Rating and Percent Salary Hike’ and ‘Monthly Income and Job Level’. We dropped certain features such as Employee ID, Over18, and Standard Hours, which were unrelated or had zero variance.

The dataset was low on the minority class of employee with Attrition as ‘No’, hence up-sampling was used to create near equal distribution of both classes in the data set.

Data Partition

The data was divided into two groups, by performing stratified sampling (80/20) to ensure that both response classes are proportionality represented in train and test data sets. Cross validation approach was incorporated for certain models (SVM: 10-fold; XGB: 3-fold) and results were validated against the held out test sets. The test set was used to assess, fine-tune, and compare the models.

Model Building and Comparison/Selection

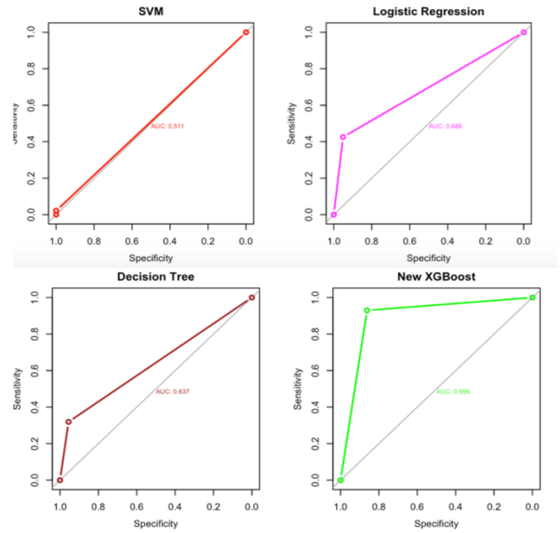
We used classification models to predict the likelihood of an employee to quit, in order to increase HR’s ability to intervene on time and take control of the situation. We used linear model for this classification type problem, i.e. Logistic regression and also used other non-linear models such as Decision trees and Support Vector Machines. Finally, we utilized a popular machine learning algorithm XGBoost to determine the probability of an employee satisfying the condition of Attrition and thus at risk of quitting.

Model Evaluation

The predictive models were tuned and evaluated on accuracy, Area under the curve (AUC), and Sensitivity to determine the best model to predict attrition. The business performance measure we considered is ‘Area under the Curve’, as the stakeholders wanted to understand the likelihood of employee attrition, so that department-wide or job-role-wide initiatives can be planned for retention.

Results

The results from different models were compared on the metric ‘Area under the Curve’ and not Accuracy as the dataset was small and our response class was unbalanced.



We used SMOTE function to balance our data and can clearly see that XGBoost had the best results with the greatest ‘Area under the curve’ and high accuracy for prediction.
HR Managers can get a probabilistic estimate of how likely an employee is to quit the company, so they can focus on methods for retention and/or succession planning. The result set from model can also to utilized to understand variables of importance, so that HR managers can focus on specific areas to work on.

Conclusions

The use of classification models to predict turnover could increase HR managers ability to intervene and plan resources accordingly. The model can offer business benefits in the following ways:

1. An increase in cost-savings
2. Maintain strong relationships between customers and clients.
3. Strategic HR decisions can be made to achieve the most ideal work environment for employees, while maintaining a proper turnover rate

There is room for further improvement and tuning of the models. Moreover, there can be several other variables of importance not currently present in the data such as peer evaluation of employee state in the company, which we believe will play a big role in understanding the position of an employee in a better way.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project.