Data Mining Project Report on

Employee Turnover

by :
Faik Marwa
Azdad Mohamed
Alla Mina
Wiaam Sahraoui
Supervised by : Yasser El MADANI EL ALAMI

INPT
المعهد الوطني للبريد والمواصلات
Institut National des Postes et Télécommunications

March 2020

Contents

1. Summary

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. Understanding why and when employees are most likely to leave can lead to actions to improve employee retention as well as plan for new hires in advance.

This project presents a relatively simple data mining approach, using R, to harnessing workforce data to understand a company's employee turnover, and predict future employee attrition before it happens so that actions can be taken now, before it's too late. In this challenge, we have a dataset with employee information and need to predict when employees will leave and understand the main factors of employee churn.

2. Problem definition and background

You may be asking what is Employee turnover? In one word, turnover, its when employees leave the organization. In another word, terminates, whether it be voluntary or involuntary. In the widest sense churn or turnover is concerned with both the calculation of rates of people leaving the organization and the individual terminates themselves.Employees turnover is a very costly problem for companies. The cost of replacing an employee if it is often more than 100,000 USD, taking into account the time spent interviewing and finding a replacement, placement costs, connection bonuses and loss of productivity for several months.

The Business Question

"What are the characteristics of employees that can help managers predict the possibility of a particular employee leaving the company?"

In many businesses such a cell phone companies and others, it is far harder to generate and attract new customers than it is to keep old ones. So businesses want to do what they can to keep existing customers. When they leave, that is 'customer churn' for that particular company.

This kind of thinking and mindset applies to Human Resources as well. It is far less expensive to 'keep' good employees once you have them, then the cost of attracting and training new ones. So we have a marketing principle that applies to the management of human resources, and a data science set of algorithms that can help determine whether there are patterns of churn in our data that could help predict future churn.

Goal

Predict when employees will leave by understanding the main factors behind employee churn.

Dataset

We have the data of employees in some companies. It contains all the employees who joined from 24/01/2011 to 13/12/2015. For each employee, we also know if he's still in the company on 12/13/2015 or if he has resigned. Also, we have general information about the employee, such as the average salary during her mandate, her department and her years of experience.

Columns:

`employee_id`: employee ID. Unique per employee per company

`company_id`: company ID.

`dept`: employees department

`seniority`: number of years of work experience at the time of hiring

`salary`: average annual salary of the employee during his mandate within the company

`join_date`: when the employee has joined the company, this can only be done between 24/01/2011 and 13/12/2015.

`quit_date`: date on which the employee left his job (if he is still employed on 12/13/2015, this field is NA)

`churn`: yes if the employee has left his job, No if he is still in the company.

`duration_days`: duration in days between `join_date` and `quit_date`

3. Design of project

Process

The Project consists of :

Part 1 - Data pre-processing : We cleaned the data:

-Standardize column titles into lowercase names

-Check column data types

-Transform join_date and quit_date into dates

-Eliminate the missing values for the attribute "salary"

-Change the type of attribute

Part 2 - Description:

-Generate a description of the dataset.

-Identify the main statistical characteristics of the data set.

-Suppose, for each company, that the workforce starts at zero on 01/23/2011. What is the estimate of the workforce, for each company, each day, from 24/01/2011 to 13/12/2015.

-Get the number of new employees each day for each company

-Get the number of employees who left each company each day

-Plot the churn / salary distribution in a boxplot

Part 3 - Classification:

-Create a decision tree to classify contributors according to churn (Yes or No) with seniority, salary entry attributes

NB: we followed the classification steps using a decision tree (Training / test, classification, evaluation, and performance ["Accuracy score" and "Confusion Matrix"])

-Classify employees according to churn (Yes or No) using KNN using cross validation

-Report KNN performance

Part 4 -Prediction:

- We performed a linear regression to predict the duration in days based on the input attributes seniority, salary

-Predict duration in day for employee x salary = 9800 and seniority = 25

Project Management

This project is a result of , Given that the project management involves coordinating various aspects in order to bring forth a positive result. After the decision of our project's subject, each member has perfectly finished his task: Wiaam was charged on the first part of the project (cleaning the dataset). Mohamed and me (Marwa) worked on the second and third parts and wrote this report. At last, Mina focused on the last part which is the prediction.

Present Results and Document

All the files (code, csv data file) for this project can be found accompanied with this report.