Predicting why employees leave prematurely

Udacity Machine Learning Nanodegree

-Siddharth Pratap Singh

October 5 2017

Domain Background:

It is very important for an establishment to understand the mindset of their employees. To know the reasons behind the premature leaving of the employees is a major objective of the companies. If understood well, this can help in increasing the employee productivity and overall growth of the company.

Problem Statement:

The employees have been leaving from the company and the problem is to understand and predict if an employee will probably leave or not.

The reasons behind leaving any company can be many. Some of these can be less number of projects that a person was involved in, if the people have promotions in a very long time, if they are frequently asked to work overtime in a company where they are not satisfied with their work quality, a number of bad evaluations from the managers etc. These factors along with many others are involved in contributing towards what kind of a relationship an employee has with his/her company.

If we have information about such factors like mentioned above, we can get a hint as to what might have been an underlying relationship between different factors, or which ones of them contribute the most. This dataset presents us with some of such factors. We have values ranging from time spent at the company, to average monthly hours that an employee spends at the company, whether they had a work accident that turned them off from the company or what kind of an evaluation they had the last time. The dataset also presents us salary and the department of the employee too.

The problem statement is simple. We want to understand and predict the reasons to know why the employees of any establishment leave. My goal is to understand which among these factors contribute towards this the most and predict if the employee will leave or not given this data about him.

Datasets and Inputs:

The data is taken from a Kaggle competition. A brief description of the features in the dataset :

- Satisfaction Level (ranges from .01 to 1 with 1 being the highest).
- Last evaluation (ranges from .36 to 1, presents us with a measurable quality of the employee as rated by the managers, 1 represents the best evaluation).
- Number of projects
- Average monthly hours
- Time spent at the company (number of years that the employee has been at the company).
- Whether they have had a work accident (1 if yes 0 if not)

- Whether they have had a promotion in the last 5 years (1 if yes 0 if not)
- Departments (column sales)
- Salary
- Whether the employee has left (1 for yes/0 for no)

The dataset contains these features' information about 14999 employees of which nearly 24% have left. We have information about the employees who have left (nearly ¼ of the dataset contains information about such employees) and we can use this information to make a prediction on other employees, i.e if they will leave or not by comparing these features against each other, comparing which features are independent and which are dependent etc.

Benchmark model:

The benchmark model is the base rate model. A simple heuristic to understand the minimum accuracy in predicting if an employee will leave or not will be predicting the majority label of the dataset. Here 76% of the employees fall in the not left label (0), hence the lowest accuracy would be to say that 76% would not leave. This will be the minimal accuracy on this dataset. I will be using this reference as my benchmark model.

Solution Statement:

After cleaning and necessary preprocessing of the dataset, I will try to find correlations between the features and possibly work with PCA. After this, a set of classifiers like RF, SVM will be used to predict the final value and the best model will be put forward.

Evaluation metrics:

False Positives: People we predicted as leaving but did not leave.

False Negatives: People we predicted as not leaving and left.

Since both the metrics are useful we will ube comparing the models on precision_score, recall_score, accuracy_score.

Project Design:

1- Downloading the dataset.

2- Exploratory analysis. This constitutes a major part of this project. We will try to find correlations between the features etc.

3- Data Preprocessing and distribution : There aren't any missing values so preprocessing wont be much of a factor here. Segmenting data into training, validation and testing data.

4- Training predictive models: Here I would try to evaluate performance of different classification models on the data. Some of the examples are logistic regression, SVM, Random Forests, knn etc.

5- Selecting the best model and final summary: The best performing model will be chosen and a few summarized points to check regarding if the person will leave. This can also help in changing present conditions of an employee to try and make them stay.