Machine Learning Engineer Nanodegree

Using supervised learning to identify whether an employ will stay in company or not.

Udaykiran Dakavarapu

June 23rd, 2018

Proposal

Domain Background

Employee retention is of significant importance to companies. It should come as no surprise that companies that lead in employee engagement are often some of the most successful and profitable companies. Poor employee engagement means turnover, low morale, poor customer service, and a general blow to your bottom line. However, this issue become ultra-costly when high-performing employees start leaving because of low engagement. Supervised learning deals with Regression and Classification .Classification is being implemented in different fields Ex: Medical diagnosis and this model has gained a momentum to understand the change in behaviour of the consumer so this helps business works on the data.

Every company invest a lot on employee to train them and makes employee to work in any kind of situation and in any kind of project. And there are many reasons employees to quit job. Some of them are rude behaviour, work imbalance, Feeling undervalued, Employment misalignment etc. If a company Invest in skill enhancement of an employee he should use it for the growth of the business ,not for his personal profit. And I was searching for the dataset which has features that can be the reasons for the employee to quit job.

Link: https://www.kaggle.com/giripujar/hr-analytics/data#HR_comma_sep.csv

Problem Statement

HR managers have many responsibilities within an organisation as they protect the interests of both the employer and the employee. For employers, they manage employee relations and identify ways to cut labour costs. For employees ,they protect their rights ensuring that employers operate within the scope of employment and labour law.HR has a tough job to find the root cause of attrition. For suppose if HR knows that a particular employee is more probable of leaving and could take damage control action. So HR needs to find what features are making employees to leave .So I classified the data into sets ,the employees who stayed in the company and who left the company and I treated it as classification problem.

Features and Description:

- satisfaction_level : It is the level of satisfaction of the employee.
- last_evaluation : Its is the time since last performance is evaluation.
- number_project : It is the number of projects completed while at work.
- average_montly_hours : It is the average monthly hours at workplace
- time_spend_company : It is number of years spent in the company.
- Work_accident : whether the employee had a workplace accident or not.
- Left : Whether the employee left the work place or not.
- Promotion_last_5years : Whether employee promoted in last five years or not.
- Department : Department they work for.
- Salary : It is relative level of the salary.

By considering the above features,satisfaction_level is very important feature for the employee to stay in the company,because if the work what he is doing doesn't give him satisfaction means he will definitely leave the company.last_evalution is also an important feature which will give confidence to the employee.I think number_project is not that much important because if an employee engaged in big project means it will take time to complete the project.Average_monthly_hours also we have to consider because if he is putting his entire time in doing work he/she will not get enough time to relax so he may eventually quit the job.time_spend_company is high means he will be paid good.left feature is the decision column which we have to predict.And the main feature is salary of the employee.If he is satisfied with the salary

definitely he will continue in the company.promotion feature will give employee satisfaction about his work.Department is not that much important.

Datasets and Inputs

I have taken the Humar resource analytics dataset for the project and for the resources left the company and why are lefting the company and what features are helping them to left the company

The features in the data set are :

- satisfaction_level
- last_evaluation
- number_project
- average_montly_hours
- time_spend_company
- Work_accident
- left
- promotion_last_5years
- Department
- salary

The dataset consists of 14999 rows and 10 columns.And the number of employees left are 3571 .And the remaining employees are 11425 and the target set is left feature .In project we are predicting the employees who are going to left.And the data set is taken from the Kaggle

Link: https://www.kaggle.com/giripujar/hr-analytics/data#HR_comma_sep.csv

Solution statement

So the current situation of the company is talented and experienced employees are leaving the company because of various reasons and there may

be a lot of reasons for the employee to quit the job.So the job of the HR is to identify why the employees are leaving the job and causing damage to the company So the machine learning model will help HR to find who are going to quit the job and reasons for it based on the past historical data.So I am using classification model of supervised learning by passing the current employee data to the model and predicting whether the employee will leave the company or not.If null values are present I will replace it with mean for numerical column and mode for categorical column.And I will do hot encoding for categorical column and there are exactly 10 features in the dataset.I will use pca because all the features will not have the equal impact and will find the most affected features and will perform ensemble learning model to understand why the employees are moving on and will finalize only one model also will use SGDC and SVM to check f score and .If the model is performing better we will perform tuning and the main theme of the model is to find why employees are leaving the job.

Benchmark model

I will use logistic regression to predict the left feature(target feature).And accuracy , f-beta score and confusion matrix and will try to get better results in the ensemble learning models.

Evaluation Metrics

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

In the Numerator, are our correct predictions (True positives and True Negatives)(Marked as red in the fig above) and in the denominator, are the kind of all predictions made by the algorithm(Right as well as wrong ones).

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

$$\text{Precision} = TP/(TP+FP)$$

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).

$$\text{Recall} = TP/(TP+FN)$$

Fβ score – measures the effectiveness of retriveal with respect to a user who attaches beta times as much importance to recall as precision.

$$F_\beta = (1+\beta_2)\cdot \text{precision} \cdot \text{recall} / (\beta_2 \cdot \text{precision}) + \text{recall}$$

When beta=0.5 more emphasis is placed on precision .This is called f 0.5 score.

Project Design

First will import the csv file using pandas library and when the data is in the data frame I will use numpy library to calculate number of records,number of employes left and total employess choosed to stay and will use one hot encoder technique to convert text values into numerical values and will represent graphs of different columns with respective to left which is decision making column to understand the impact of features.And when the data is prepared I will split into two parts training and testing in the ration 80:20 and will drop the column of testing data set,so it is used to compare predicted value with actual value.And will divide the training data into 10 folds and will

perform kfoldcross validation to minimise overfitting and increasing the accuracy of the model. First will create model with logistic regression and to have a bench mark fscore and confusion matrix and will check accuracy. And will use ensemble learning to train the model and also will train model with SGDC and will check performance. And will improve performance of the model using XGboost for this I will use scikitlearn. And will apply gridsearch to tune the parameter and increase the performance. And once the model is trained I will use test set to predict the values and accuracy and confusion matrix is calculated based on the predicted values So this is my procedure for this project.

References

- https://www.techopedia.com/definition/28334/human-resources-analytics-hr-analytics
- https://www.analyticsinhr.com/blog/what-is-hr-analytics/
- https://www.forbes.com/sites/bernardmarr/2016/03/01/the-8-hr-analytics-every-manager-should-know-about/
- https://www.visier.com/clarity/reduce-employee-turnover-with-workforce-analytics/