Machine Learning Engineer Nanodegree

Predict Employee Retention

Mamidi Karthik Kumar
February 16, 2019

Proposal

Domain Background

Every year a lot of companies hire several employees. These companies invest time and money for training those employees, not only this but there are training programs within for their existing employees too. The main aim of these programs is to increase the effectiveness of their employees [1]

HR Analytics:

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

Retention in HR:

Employee Retention refers to the techniques employed by the management to help the employees stay with the organization for a longer period. Employee retention strategies go a long way in motivating the employees so that they stick to the organization for the maximum time and contribute effectively. Sincere efforts must be taken to ensure growth and learning for the employees in their current assignments and for them to enjoy their work.

Employee retention has become a major concern for corporates in the current scenario. Individuals once being trained tend to move to other organizations for better prospects. Lucrative salary, comfortable timings, better ambience, growth prospects are some of the factors which prompt an employee to look for a change. Whenever a talented employee expresses his willingness to move on, it is the responsibility of the management and the human resource team to intervene immediately and find out the exact reasons leading to the decision.

I personally motivated for this project, as an organization invests time and money in grooming an individual and make him ready for work and to understand the corporate

culture and when an individual resigns from his present organization, it is most likely that he would join the competitors. So, the management must understand the difference between a valuable employee and an employee who doesn’t contribute much to an organization. Hence efforts must be made to encourage the employees so that they stay happy in the current organization and do not look for change.

Problem Statement

The main problem I want to address here is: How can a company better act to minimise the negative of employee turnover?

Every Organization is trying to gain maximum results and especially employees looking for better opportunities to fulfil their demands. So, the retention of an individual in an organization is not for long and without the employees the organization cannot function well. So, It is important to understand the various problems an organization faces in order to maintain the employees and use methods to overcome these problems and retain employees.

It is a classification problem and it uses the data of previous employees which have worked for the company and by finding pattern it predicts the retention in the form of yes or no. It uses various parameters of employees such as salary, number of years spent in the company, promotions, number of hours, work accident etc.

Datasets and Inputs

For this project, I will be using the HR Employee Retention [2] dataset from Kaggle. This was uploaded for examining employee retention and will be suited for this analysis. The dataset consists of 15,000 samples with 9 features plus 1 label for retention. They contain a mixture of numeric, Boolean, and label values. From the dataset it is possible to see that satisfaction_level could be correlated with other variables such as work_accident.

This dataset contains numerical features, Boolean features and categorical features as well. The total number of data points are 14999. As for splitting the dataset, we split our data into features and labels in separate training, validation and test sets. We will use validation set to help measure our model's performance as we develop it and hold out the test set until final evaluation of the model.

- **satisfaction_level:** It is a numerical measure of employee satisfaction from 0 which is low, to 1, which is high.
- **last_evaluation:** It is a numerical measure of how well the employee was performing at their last evaluation, from 0 which is low, to 1, which is high.
- **number_project:** The number of projects the employee was assigned to.
- **average_monthly_hours:** The average numbers of hours the employee worked each month, rounded to an integer value.
- **time_spend_company:** The time the employee has spent at the company, the number denoting the number of years.

- **work_accident:** It shows whether the employee was involved in a work accident, 0 which is no, and 1, which is yes.
- **performance_last_5years:** It shows whether the employee was promoted in the last 5 years, 0 for no and 1 for yes.
- **sales:** It shows which department the employee worked in.
- **salary:** A text label for low, medium or high.

And finally, the label representing the retention is:

- **left:** It shows whether the employee left or not, 0 for no and 1 for yes.

Note: It must be noted that this is a simulated dataset. Unfortunately, real company data suitable for this problem is hard to available, the reason may be due to a combination of sensitive corporate value and personal privacy the dataset would contain. The only other similar dataset I found was also simulated, which is also from Kaggle IBM HR Analytics Employee Attrition & Performance [3].

As a result, this will limit the direct application of our results. Although we should be able to apply the same principles & model to a real-world dataset.

Solution Statement

To solve this problem, I plan to:

1. Provide insight into the underlying pattern in the data.
2. Look for any significant clusters within the dataset.
3. Provide a model that can predict retention of specific employees.

For this problem I would like to try supervised learning techniques such as Logistic Regression, Support Vector Machines, Decision Trees and Random Forest Classifier.

I believe with these insights, it would be possible to make more informed decisions about how to improve employee retention and to better target efforts on specific groups or individuals. Furthermore, a predictive model could be used to more quickly evaluate the effect of company on employee retention.

Benchmark Model

Apart from the dataset on Kaggle there is an analysis authored by the [4] using the same dataset.

This analysis includes investigation into the variables and their correlations and looks for variables contributing to whether employees leave, using statistics and visualisations to demonstrate findings along the way. As this analysis already shows a significant correlation between satisfaction_level and employee retention. It also highlights groups of employees who are favourable to retain, based on their evaluation score.

Then the analysis proceeds to the main modelling stage, using cross validation on the dataset with three learners: decision tree, logistic regression, Random Forest and Support Vector Machine.

Of them the final chosen model was Random Forest Classifier as it could predict retention with about 97% accuracy on the test set.

By also using accuracy as one of our evaluation metrics, we will be able to directly compare our predictive model against this one. We will also be able to compare it to the other features of the model through:

1. understandability and simplicity.
2. Ability to provide actionable insight.

Evaluation Metrics

This dataset is an example of a class imbalance problem because of the skewed distribution of employees who did and did not leave. More skewed the class means that accuracy breaks down. As a result, evaluating our model based on accuracy is wrong thing to do.

I prefer to use recall rate:

Recall = correctly_predicted_leavers / total_leavers

and I also prefer to use precision:

Precision = correctly_predicted_leavers / predicted_total_leavers

and, F1-Score:

F1-score = 2 (precision * recall) / (precision + recall)

This is motivated by wanting to minimise the effect of employees leaving, so it makes sense to place extra importance on minimising misclassifying leavers that is false negatives.

Project Design

To solve this problem, I intend to follow the below mentioned workflow:

1. Data Pre-processing: Look over the data to understand the features. Perform any appropriate feature removal or scaling. In this step we will perform operations like replacing some of the feature labels like ‘sales’ to ‘department’ and ‘Work_accident’ to ‘work_accident’ etc. We can visually inspect all our numeric features to make sure we don't have any outliers. We use a boxplot to show the variability of the data, including outliers. For ease of visualisation, we can apply simple rescaling ((xv-min) / (max - min)) so that they are displayed on the same axis scale in our visualisation. For Boolean features as these are provided as integer values of 0 or 1. We'll first that ALL values are 0 or 1, and then convert to explicit Boolean data types and the we will inspect the categorical data to ensure labelling is consistent, and then encode the categories appropriately.

2. Feature Investigation: Measures central tendency, variation within features. In this step we will find central tendency by using mean, standard deviation. For feature correlations we plot all numerical features against each other, then just continuous numerical features against each other.
3. Principle component analysis: Are there any underlying features? and what are the most and least significant features?
4. Cluster Analysis: Are there any significantly different groupings of employees, with regards to retention rates? I would like to consider using Gaussian Clustering or K-Means Clustering for this dataset.
5. Modelling: Identify and evaluate multiple appropriate algorithms to model the data to make classification predictions. Choice of model based on predictive power and suitability for use. I plan on using supervised learning techniques such as Logistic Regression, Decision Trees, Random Forest Classifier and Support Vector Machines.

References

[1] Dubey, A. K., Maheshwari, I., & Mishra, A. Predict Employee Retention Using Data Science.

[2] https://www.kaggle.com/gummulasrikanth/hr-employee-retention

[3] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[4] https://www.kaggle.com/randylaosat/predicting-employee-kernelover