Machine Learning for Employee Retention

MLND Capstone Proposal

Yanfei Wu

4/18/2017

Background

Recruiting excellent employees is one thing, but keeping them is another. While losing employees who are poor performers can have positive effects, high employee turnover rate is generally regarded as bad for the business. It increases expenses since the process of identifying, hiring and training employees is very expensive. Studies have shown that cost related to directly replacing an employee can be as high as 50–60% of the employee's annual salary, and the total cost of turnover can reach as high as 90–200% of the employee's annual salary.[1] Even worse, frenquent employee turnover can destroy the company morale, resulting in decreased performance in the workplace. Therefore, retaining its valuable and talented employees is vital to a company's success.

A novel approach to implementing an effective retention program and preventing key workers from leaving prematurely is to use machine learning techniques. For example, a supervised classification model can be trained on a dataset containing features related to the employees and whether they left (a dataset that could be available in many companies). Building such a model provides insights to key factors that result in employee turnover. It also allows the management and the human resource team to predict which employee is going to leave so that they could intervene immediately.

Problem Statement

The goal of this project is to build a supervised binary classification model using a simulated dataset containing a series of employee-related features and a binary class label of whether the employee left or not. The expected outcome of this project is to help the management and the human resource team in the company
1). to predict which current employee is going to leave (class label) so that they can intervene immediately;
2). to identify which are the most important factors (features) that lead to employee turnover so that changes can be implemented to ensure employees remain in place while maintaining high work performance and productivity.

Datasets and Inputs

The dataset was obtained from **Kaggle**. It is a simulated dataset containing 14,999 rows and 10 columns. The features include:
- `'satisfaction_level'`: employee satisfaction level (numerical, float 0-1)
- `'last_evaluation'`: the score from the employee's last evaluation (numerical, float 0-1)
- `'number_project'`: the number of projects the employee worked on (numeric, integer)
- `'average_monthly_hours'`: the average monthly hours the employee spent on work (numeric, integer)
- `'time_spend_company'`: number of year the employee spent at the company (numeric, integer)
- `'work_accident'`: whether the employee had a work accident (categorical, '0' - no, '1' - yes)
- `'promotion_last_5year'`: whether the employee had a promotion in the last 5 years (categorical, '0' - no, '1' - yes)
- `'sales'`: the Department the employee works (categorical, 'sales', 'accounting', 'hr', 'technical', 'support', 'management', 'IT', 'product_mng', 'marketing', 'RandD')
- `'salary'`: the salary of the employee (categorical, 'low', 'medium', 'high')

The target variable in the dataset is:
- `'left'`: whether the employee has left or not (categorical, '0' - no, '1' - yes)

Solution Statement

Based on the dataset, binary classification models using algorithms such as Logistic Regression, K-Nearest Neighbors, Decision Trees, and ensemble methods such as Random Forest can be trained and evaluated to find the best model for the problem. An optimized model can then be used to predict whether a current employee with given features will leave or not. Furthermore, feature importance analysis can be carried out to understand the most important factors that lead to employee turnover.

Benchmark Model

A naive benchmark model for this problem would be to predict that all the employees are going to leave. In other word, we convince that the management and HR team treat every single employee as if he/she is leaving. This will probably lower the turnover rate but most of the time, it is not a realistic retention program for most companies.

Evaluation Metrics

One way to evaluate the performance of the benchmark model and the solution model is to measure accuracy using cross-validation. Accuracy is defined as the number of correct predictions from all the predictions made. But it is not the best metric in this case due to the unbalanced classes (~20% class 'leave' vs ~80 class 'not leave'). Instead, a confusion matrix would give an unambiguous way to show the prediction results of the classifiers.

We can also obtain precision and recall from the classification results. Precision is the number of True Positives (TP) divided by the number of True Positives (TP) and False Positives (FP). A low precision indicate a large number of FP. Recall is the number of True Positives (TP) divided by the number of True Positives (TP) and the number of False Negatives (FN). A low recall indicates a large number of FN.

In this specific problem, we are better off to make false positive classifications than predicting valuable employees who are actually leaving as staying and taking no actions. Therefore, we can use the F2 measure which weights recall higher than precision.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, (\beta = 2)$$

Project Design

A workflow for this project is outlined below.

0. Environment

The project will be carried in **Python 3.5** with the following libraries:
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Scikit-Learn

1. Exploratory Analysis and Data Preprocessing

The goal of this part is to understand the characteristics of the dataset and to prepare it for the models.
The steps include:
- Visualize each individual feature
- Visualize the scatter matrix of features
- Normalize skewed numerical features
- Encode categorical variables

3. Model Building and Performance Evaluation

The goal of this part is to build classifiers with different models and select the best model for this problem.
The steps include:
- Implement classification models: Logistic Regression, K-Nearest Neighbors, Decision Trees, Ensemble Methods
- Select a best model by evaluating the performance metrics
- Optimize the model by tuning the hyperparameters

4. Feature Importance and Insights

The goal of this part is to extract the imformation on feature importance so that actionable insights can be provided to the management and the HR team.
The two steps are:
- Evaluate the feature importance
- Recommend actions companies can take to retain valuable employees

---

[1] Cascio, W.F. 2006. Managing Human Resources: Productivity, Quality of Work Life, Profits (7th ed.). Burr Ridge, IL: Irwin/McGraw-Hill. Mitchell, T.R., Holtom, B.C., & Lee, T.W. 2001. How to keep your best employees