Udacity Machine Learning Nanodegree
Capstone Proposal

Hoang Duy Trinh

Human Resource Analytics

Proposal

Domain Background

In any company, Human Resources (HR) have the task of regulating, monitoring and analyzing the complex interactions that occur between the employees and the company itself. Employee retention is of significant importance to companies. However, due to several reasons, the company may need to replace some of the employees. This would make the company to incur in high costs due to many parts, including hiring new people, and also in side effects, for example, lowering the morale of the other employees. Employee retention is therefore a delicate matter that needs to be addressed considering a wide range of factors: studies show an immediate reduction in the productivity, and major efforts are usually required to keep pace with economic constraints (for example demand/supply management etc.).

Due to these reasons, a data-driven approach is the most significant in order to understand the complex relationships that take place in a work environment and it can help to alleviate the surplus of management costs. Similar to this, multiple efforts have been done in machine learning to cope with churn problems. In [1], the report provides a case study for customer churn linked to a well-known online payment platform.

Datasets and Inputs

The dataset used in this project is the Human Resource Analytics from Kaggle [2] and consists of nearly 15000 samples, each containing 9 features corresponding to one employee. The dataset is labelled under the column **left** (=1 if the employee left, =0 otherwise), the other columns are the following:

1. **satisifcation_level:** employee satisfaction level, num in [0,1]
2. **last_evaluation:** last employee evaluation, num in [0,1]
3. **number_project:** number of projects, num
4. **average_monthly_hours:** avg number of working hours, num
5. **time_spend_company:** years spent at the company, num
6. **work_accident:** whether the employee was involved in a work accident, boolean (0,1)

7. **promotion_last_5years:** whether the employee was promoted in the last 5 years, boolean (0,1)
8. **department:** which department the employee worked in, text
9. **salary: l**abel of low, medium or high, text

The class distribution is unbalanced, meaning that one class is much more represented in the dataset. In particular, the employees with left = 0 are more present than the employees who left. Approximately, the distribution is (75-25%), therefore, insightful metrics need to be chosen to take care of this fact (e.g. accuracy only may not be sufficient alone and may lead to the accuracy paradox).

Problem & Solution

We want to analyze the characteristics of the employee turnover, and understand how a company could minimize the negative effects. To this end, the solution should include:

1. A detailed characterization with deep insights in the data, studying first the importance of each features, their correlations and their statistics

2. A possible classification of the employees, including clustering methods

3. An anticipatory model that could predict a possible employee leave

Benchmark Model

A comprehensive analysis can be found from Kaggle kernel in [3], which includes investigation of the variables, their correlations. Also good insights and visualizations are given to determine which are the main factors that correlate the satisfaction level & employee retention.

A set of models is considered in a logistic regression problem, to predict the employee retention. By using accuracy as one of evaluation metrics, we are able to directly compare the proposed predictive model against this one.

Evaluation Metrics

In this problem, accuracy would be a good metrics to evaluate and compare the presented models. However, due to the unbalance class distribution it is better to consider multiple metrics, including recall, precision and F1 score. In particular, the recall of leavers, will be important to evaluate the performance of the algorithms.

Project Design

Our project will include the basics and fundamental part of a data-driven problem solving:

- we will start from a data preprocessing step, where missing features and categorical components of the dataset will be treated;
- then, we will perform some EDA, where we will adopt useful visualizations to understand the features and their correlations;

- we will start our modeling and analysis: first a clustering method, can help to individuate sub-groups of employees and make us understand how to separate the employee classes;

- finally, we will study which type of model will fit best our goal: to do this, a parameters validation and search will be implemented to maximize the evaluation metrics. Also, new models will be explored: to this end, we can rely on different implementation of the gradient boosting algorithm, including Xgboost or LightGBM.

[1] https://www.h2o.ai/wp-content/uploads/2019/02/Case-Studies_PayPal.pdf

[2] https://www.kaggle.com/jacksonchou/hr-data-for-analytics

[3] https://www.kaggle.com/jacksonchou/hr-analytics