Machine Learning Engineer Nanodegree

Capstone Proposal

Topic: Attrition prediction with machine learning approach in human resource management

Ivan Chen
December 28, 2017

Proposal

Domain Background

In human resource management, HR professionals take the responsibility in designing suitable systems such as work culture, compensation or promotion system to help the company or organization to retain top employees. Attrition in human resources refers to the gradual loss in labor over time.

In general, average attrition rate compared with other companies in same business is acceptable but relatively high attrition may cause trouble for companies. A major problem in high attrition rate is high cost to an organization; for example, human resources team may need to start hiring processes, providing new hire training. Besides, to the organization, high attrition rate may lead to lose knowledge and experience; it may impact daily or business operation. Usually human resources teams take attrition rates as factors into their department budgets to account for potential losses in productivity and the costs with replacing employees.

Problem Statement

The attrition rate is determined by dividing the number of employees who left the company or organization during a specific period by the number of employees during the same period. Attrition rate can be computed by monthly, quarterly or annual periods. Consistent rate of attrition is treat as the norm for a company but high attrition rate may induce trouble to a company. To predict and classify high attrition risk of employees are critical for human resources members. With the classified result human resources teams may evaluate the causes of attrition and find solutions.

The goal of the project is to generate a relatively accurate model to predict attrition or stay of employees based on collected employees data. In the study, a solution was proposed and to be implemented. Benchmark model, evaluation metrics, key feature importance observation will be considered and discussed.

Datasets and Inputs

The dataset used in this project was downloaded from kaggle, (https://www.kaggle.com/ludobenistant/hr-analytics-1/data) which was provided by Ludovic benistant for HR Analytic study and practice. The dataset consists total 14999 rows with 10 columns. The last one “attrition” column renamed by me from “left” for this study was used as target label of the prediction. Other columns provide related information of the employee such as average monthly working hours, number of projects. Although the number of columns is not many but they provide important work related information of employees, I think the dataset is suitable for this attrition prediction problem.

#	Column name	Meaning	Type	Possible value
1	satisfaction_level	Satisfaction Level	numerical	float
2	last_evaluation	Last evaluation	numerical	float
3	number_project	Number of projects	numerical	integer
4	average_monthly_hours	Average monthly hours	numerical	float
5	time_spend_company	Time spent at the company	numerical	float
6	Work_accident	Whether they have had a work accident	numerical	Binary, 1 denotes 'Yes', 0 denotes 'No'.
7	promotion_last_5years	Whether they have had a promotion in the last 5 years	numerical	Binary, 1 denotes 'Yes', 0 denotes 'No'.
8	Departments (original column name: sales)	Departments	categorical	sales, accounting, hr, technical, support, management, IT, product_mgn, marketing, RandD.
9	salary	Salary	categorical	low, medium, high.
10	Attrition (original column name: left)	Whether the employee has left	numerical	Binary, 1 denotes 'Yes', 0 denotes 'No'.

Solution Statement

For the attrition prediction of human resource management, I propose to adopt supervised machine learning approaches to build a training and prediction pipeline, such as Support Vector Machine, Decision Tree or AdaBoost ensemble method to build models with input dataset of employees’ information. Using accuracy and F-score measurement metrics can help to evaluate the model quality. According to high accuracy requirement of attrition prediction, smaller beta value of F-score method can achieve the quality measurement need.

Benchmark Model

I propose to generate a naive predictor, which always predicts attrition (value 1) or always non-attrition (value 0) as base model without any intelligence for comparison. Accuracy and F-score measurement can be adopted to measure the result. With this generated cost-effective naïve model, we may have base model to compare performance once the proposed solution model is ready.

Evaluation Metrics

The goal of the model is predict whether an employee will leave or stay in the company, accuracy and F-beta score are appropriate metrics for this kind of binary classification problem.

Accuracy measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of test data points. That is, accuracy = (True Positives + True Negatives) / Total number of data points.

"F-beta score" is a metric that considers both precision and recall. The formula was described below:

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

In the formula of F-beta score, the precision is the ability of the classifier not to label as positive a sample that is negative. For example, in a spam email classification application, it can tell what proportion of messages was classified as spam, actually was spam. That is,

Precision = True Positives/(True Positives + False Positives)

The recall is the ability of the classifier to find all the positive samples. In a spam email classification application, it can tell what proportion of messages that actually was spam that was classified by us as spam. That is,

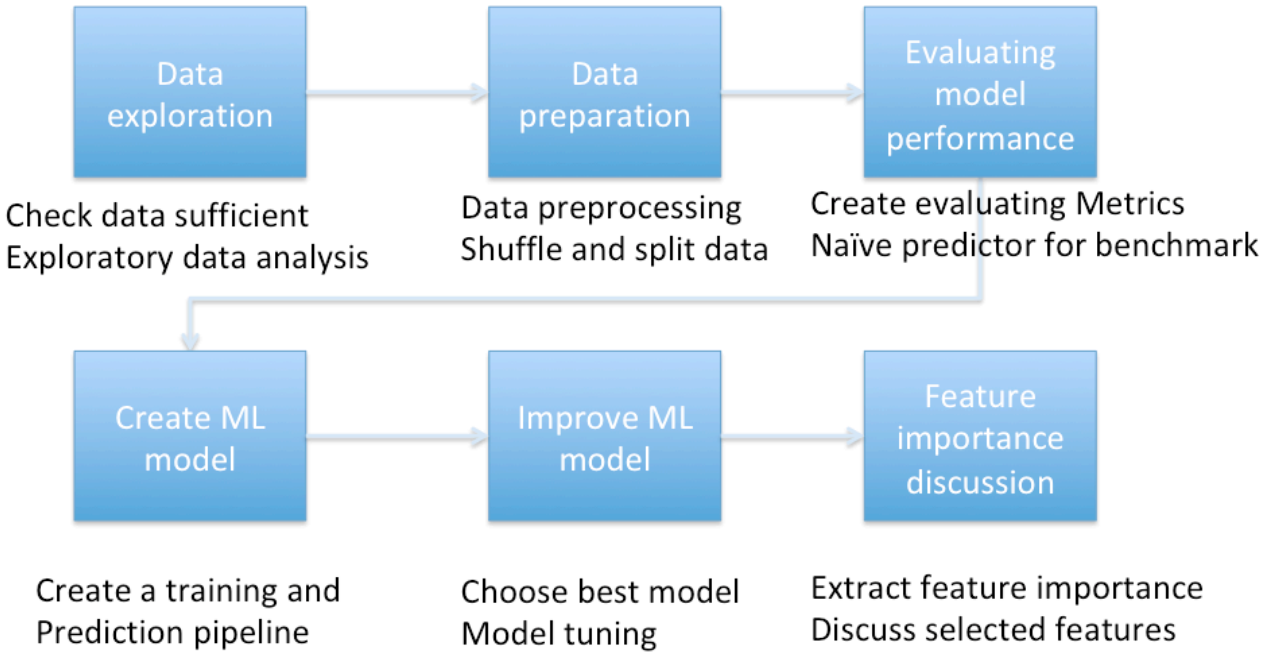
Recall = True Positives/(True Positives + False Negatives)

The F-beta score is a weighted harmonic mean of the precision and recall. User for different interest of application can adjust the weight beta. If the value of beta equals 1 means recall and precision are equally important.

For the attrition prediction problem, I thought precision is more important than recall. If an employee were classified as high risk of attrition candidate, it's better to be a real attrition because HR professionals may take some action based on the prediction result, for example, providing compensation program if the employees are valuable or preventing the employee's improper operations that hurt the company before leaving. So I propose to adjust the beta to smaller than 1 because the model requires high precision.

Project Design

The workflow of the solution for this attrition prediction consists the following main steps and described below:



1. Data exploration

First, I will check the dataset whether it is sufficient or suitable for the attrition study or not, then perform exploratory data analysis, such as exploring how may rows in the dataset, what is the attrition ration in the raw data, which I can judge is there large skew of the target label (attrition)? After feature exploration, I may know the types of features, what types are numerical (or binary), what types are categorical they are for later data preprocessing.

2. Data preparation

After data exploration, data preprocessing for different kind of types is needed, such as transferring skewed continuous features, normalization numerical features, using one-hot encoding scheme to convert categorical features to numerical dummy variables. After data was preprocessed properly, then I may shuffle and split data for later training and prediction model creation.

3. Evaluating model performance

In this attrition prediction project, accuracy and F-beta score metrics were chosen to measure the performance of model. Meanwhile, human resources member may care about more accuracy than recall, a smaller than 1 beta may suitable to get high accuracy model. Because there is no previous version of model for benchmark, I propose to generate a naive predictor, which always predicts attrition or always non-attrition as base model without any intelligence for comparison.

4. Create machine-learning model

In this attrition perdition, I will adopt three supervised learning models creating a training and prediction pipeline, which are Support Vector Machine (SVM), Decision Trees and AdaBoost ensemble learning. They both have advantages and disadvantages and I think intuitionally the AdaBoost ensemble leaning or Decision Trees may have better performance in the study, but the measurement metrics may tell us the result finally.

5. Improve machine-learning model

Based on previous evaluation of performance, I may choose the best model and perform model tuning for optimization. If the training time is not a big issue, then I may choose the model with better accuracy and F-beta score as best model. By using grid search approach with various values of hyper parameters to fine tune the chosen model can get final optimal model.

6. Feature importance discussion

In the final stage, I will explore the feature importance and their relevance with target label (attrition). With the feature importance observation, if the training time is a critical concern, then new model with these top features (say top five) may have a little lower score than the model with completed features but training time may significantly reduced. Besides, with these top important features observation, human resources teams may consider to provide compensation programs to valuable employees or take prevention actions to avoid employees' improper operation to the company before leaving.