Proposal

Lin liuyu 1701213060

1. Project description

This project is to establish a model to predict whether employees will quit from a given record of factors affecting employee turnover.

2. Project process

a) **Data analysis:** do some statistical analysis on the data as well as visualize the data to get the intuitively idea that which feature may be useful to clarify these two kinds of people. For example, we can see the distribution of different features in these two kinds of people.

b) **Data processing:** deal with the na data. Change some categorical data into dummies variable or continuous variable.

c) **Feature construction:** try to build some new features by the raw data to describe the behavior of this person more specific.

d) **Separate data:** spilt the data into training data and test data randomly.

e) **Model training:** use the in-sample data to train and validate the model, i may use perception, logistic

regression, svm, decision tree and combine different models for ensemble learning to improve the performance.

f) **Improvement:** try to use pca or lda and redo step e).

g) **Visualization:** do some visualization for the outcome.

h) **Evaluation:** evaluate the prediction model.

3. Dataset description

Dataset source:

http://www.dcjingsai.com/common/cmpt/%e5%91%98%e5%b7%a5%e7%a6%bb%e8%81%8c%e9%a2%84%e6%b5%8b%e8%ae%ad%e7%bb%83%e8%b5%9b_%e8%b5%9b%e4%bd%93%e4%b8%8e%e6%95%b0%e6%8d%ae.html

Dataset information: The data mainly include factors that affect employee turnover (salary, business travel, job satisfaction, work input, overtime, promotion, salary increase, etc.) And the corresponding record of employee turnover. The data is divided into training data and test data, which are stored in two files of pfm_train.csv and pfm_test.csv, respectively. The training data mainly consist of 1100 records and 31 fields. The main fields are as follows:

Data variables:

Age	Employee's age
Businesstravel	Business travel frequency, non-travel doesn't travel, travel_rarely says she doesn't travel much, travel_frequently says she travels frequently.
Department	The department of employees, sales means sales department, research & development indicates r & d department, human resources indicates human resources department.

Distancefromhome	The distance between a company and its home address. Range from 1 to 29, 1 is the nearest, and 29 represents the farthest.
Education	The education level of employees is from 1 to 5, and 5 indicates the highest level of education.
Educationfield	The field of professional learning by employees, life sciences for life science, medical for medical treatment, marketing for marketing, technical degree as a technology degree, human resources for human resources, and other for others.
Employeenumber	Employee number
Environmentsatisfaction	The satisfaction degree of employees to work environment. Range from 1 to 4, 1 is the lowest, and 4 is the highest.
Gender	The gender of employees, male means men, female means women.
Jobinvolvement	Employees' work engagement. Range from 1 to 4, 1 is the lowest investment level, and 4 is the highest investment level.
Joblevel	Occupation level. Range from 1 to 5, 1 for the lowest level, 5 for the highest level.
Jobrole	Sales executive is a sales director, research scientist is a science researcher, laboratory technician laboratory technician, manufacturing director is the manufacturing director, healthcare representative is a medical representative, manager is a sales representative. Ector is the research director and human resources is human resources.
Jobsatisfaction	Job satisfaction is from 1 to 4, 1 represents the lowest degree of satisfaction, 4 represents the highest degree of satisfaction.
Maritalstatus	Employee's marital status, single represents singleton, married represents married, divorced represents divorce.
Monthlyincome	Employees' monthly income ranges from 1009 to 19999.
Numcompaniesworked	The number of companies that employees once worked for.
Over18	Is the age more than 18 years old
Overtime	Overtime, yes means overtime, no doesn't work overtime.
Percentsalaryhike	The percentage of wage increase.
Performancerating	Performance evaluation;
Relationshipsatisfaction	Relationship satisfaction. Range from 1 to 4, 1 shows the lowest degree of satisfaction, 4 expresses the highest satisfaction.
Standardhours	Standard working hours;
Stockoptionlevel	The level of stock options;
Totalworkingyears	Total work age;
Trainingtimeslastyear:	Training time last year, from 0 to 6, 0 indicated no training, 6 showed the longest training time.
Worklifebalance	The balance between work and life. Range from 1 to 4, 1 indicates the lowest degree of balance, 4 indicates the highest degree of balance.
Yearsatcompany	The number of years worked in the company.

Yearsincurrentrole	The number of years of work at present.
Yearssincelastpromotion	The time from the last promotion.
Yearswithcurrmanager	The number of years worked with the current managers.

Whether or not turnover

Attrition	Whether employees have left, 1 said they have left, and 2 said they did not quit.