Methodology

We selected three widely used prediction models to obtain predictions, including naive Bayes, association rules, and decision trees. These models will be trained and tested using the same data set and the accuracy will be determined.

Naïve Bayes

Naïve Bayes is a Classification Method Based on Bayes' Theorem and Characteristic Condition Independent Hypothesis. This technique is mainly used to classify the possibilities of each situation. Using this approach is useful for predicting what we want to do, because it can easily help us to predict employee attrition and can be used against other technologies to see differences. In the learning process, we found that the algorithm has the ability to produce high precision, and the logic behind this is P(B|A) = P(A|B)P(B)/P(A). (A stand for feature and B stands for category).

Decision Tree

Decision tree is a decision analysis method that evaluates the risk and judges its feasibility by constructing a decision tree to obtain the probability that the expected value of the net present value is greater than or equal to zero, based on the known probability of occurrence of various situations. A graphic method. Since this decision branch is drawn like a tree, it is called a decision tree. In machine learning, the decision tree is a predictive model, which represents a mapping relationship between object attributes and object values. Entropy = system messiness, using algorithm ID3, C4.5 and C5.0 spanning tree algorithm Use entropy.

A decision tree is a tree structure in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category.

Decision trees are a very common classification method. It is a kind of supervisory learning. The so-called supervisory learning is to give a bunch of samples. Each sample has a set of attributes and a category. These categories are determined in advance. Then learn a classifier that can be used for new appearances. The object gives the correct classification. Such machine learning is called supervised learning.

Association rule

The association rule is an implication of the form X→Y, where X and Y are respectively called the leading and succeeding of the association rule. Among them, the association rule XY has the support degree and the trust degree.

Association rules are often used to find relationships between each attribute. It is primarily used to determine hidden rules in transaction data by applying purchases based on strong rule concepts and measurement of product frequency. By applying association rules to our data sets, we can clearly determine the extent of employee turnover and what factors are related to employee attrition.

Data Exploration

We download the data set from Kaggle named “Company”. There’re 1470 rows with 35 variables contains employees’ details, working condition, job details, attrition condition and so on. Each column with 1 being low and 5 being very high. In this data set, our goal is to understand “Attrition” to explore and solve this problem.

Here, we will use the data set for simple analysis, the relationship between several major categories of factors and attrition. For example, gender, age, working hours, working years, salary levels, etc. From the data we can see that there are more men in the company than women, so the number of attritions is more, but not much, so there is no strong correlation with gender. Then looking at the age, most people who leave the company are around 30 years old. People who are probably around 30 years old have more ideas and will choose their own lives. Regarding the time of work, the average monthly working hours of the attrition employees is higher than the average monthly working hours of the unemployed employees and the average monthly working hours of the attrition employees is higher than the monthly average working hours of the unemployed employees. For the salary issue, there is no doubt that the higher the salary, the lower the attrition rate. The prediction of the relationship between other factors and attrition will be analyzed and explored in more detail below.