**Report**

**XXX**

1. **data set**

The data we use is given by the exercise, it has four attributes: gender, age, ograde and igrade. Gender is recognized as M and F, age is range from 18 to 60, and ograde and igrade is range from 1 to 5. The main characteristics is described in the chart below:

|  |  |
| --- | --- |
| Size | 3500 |
| Number of NAs | 2977 |
| Number of attributes | 4 |
| Numbers of male | 1573 |
| Numbers of female | 1927 |
| Male without missing value | 270 |
| Female without missing value | 253 |

As we can see, though this data set have thousands of records, there are many missing value in it, most of them missing the “OGradethe” attributes, only 523 record is intact. Also, the gender distribution is not very balanced.

1. **Preprocessing**

The preprocessing I applied mainly has three steps. First, we wipe out all the records which have missing value, and 523 record remained.

Second, I applied dummy variables theory, for “gender” column, I converted the data into 1 and 0. For each record, 1 for male and 0 for female.

Third, as the assignment required, I randomly choose 70% of records become my training data set, the remain records become testing data set. This split is random and reproducible.

1. **histogram chart**

To draw the histogram, first I divide the subjects into two part according to their gender, and then designed six groups, each group have 10 years old arrange, after the counting jobs, the chart is draw as below:



As we can see, most subject is ages 20 -40, the distribution is not very balanced.

1. **scatter chart**

Before drawing this chart, I sorted all the data based on their ages. Also add a random scattered on the x and y, make all points can’t obscured by other points. The difference here is calculate from the difference between the third column and the fourth column. The chart is draw as below:



From this chart, we can find that young people is more likely to have igrade bigger than ograde.

1. **logistics regression**

The logistic regression results show that I only have 50% – 60% precision with the random choose training data set.

To describe the result, here I use one random record. The confusion matrix is:

|  |  |  |
| --- | --- | --- |
| Predicts  Actual | Female | Male |
| Female | 42 | 35 |
| Male | 32 | 47 |

1. **random forest**

The results from random forest is the same as the logistic regression, also 50% - 60% precision. And the confusion matrix is very familiar to the chart above.

1. **model recommendation**

From my experiment, both of the two algorithms didn’t have nice precision. But in compare with these two algorithms, I recommend the random forest, it is better for high dimension data, and also in its training process, it can detect the inner influence between features.