Topic
Predict what you use when
backing home / travel

1. What we predict

We collect a lot of data which contains one's age, back home frequency, job, and other which may affect what transportation one person choose in backing home.

The target has three items, cheap one, expensive one, and drive by oneself.

For instance: one is student(no part-time), and he has a high frequency of backing home, and his hometown is close to the location of the school, then he has a high probability to choose train. Or one has a high salary, and have a very low frequency of backing home, he might choose HSR.

Note: someone's hometown and location of work/school is at the same place, so they choose the data of traveling

## 2. Original data, is in the file 'original.xlsx' and 'final.csv'

The features include 12 items, which is hometown/travel, age, job, frequency, location of two, distance of two, relationship of friends and family, financial situation, gender, have married and so on.

We collect the data by ptt, baha, fb and so on, which contain 1300 samples.

3. The data distribution

Pie chart of transportation

- train bus or ship — 60.9%
- drive or ride by yourself — 21.8%
- HSR or airplane — 17.3%
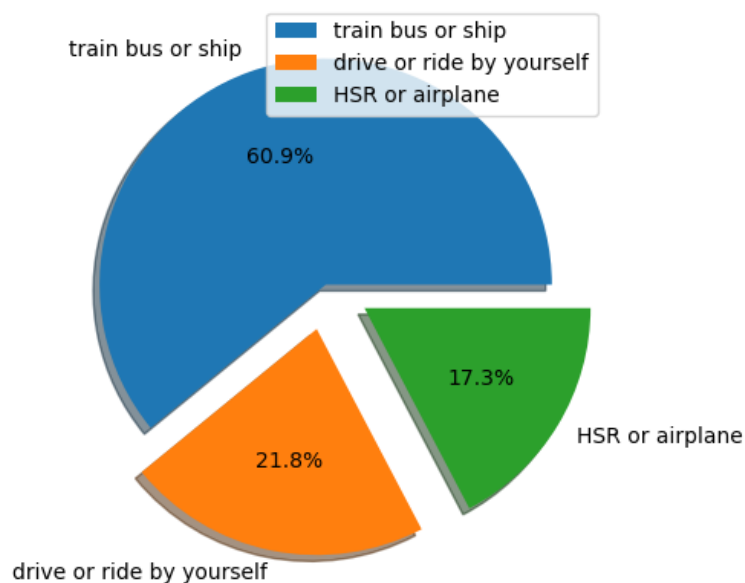
The above is our target distribution, many people choose train or bus, and this also makes the other two has less train. Others are in the 'picture' folder.

4. Data processing

We fill the missing with mode. Then transform them to numeric data to fit in sklearn.

5. Model

Five models are created, which are Decision tree, Random forest, Naïve bayes, SVM, Logistic. We implement our model first, but the result is worse than sklearn, so choose the sklearn.

6. Predict result

```
Decision Tree ---------------------------------------------------------------
----
                      Predict train, bus, ship  HSR, airplane  drive, ride
Actual train, bus, ship                    204             18           32
Actual HSR, airplane                        54             11            5
Actual drive, ride                          48              8           26

Accuracy:  0.5935960591133005
         train, bus, ship  HSR, airplane  drive, ride
Recall           0.803150       0.157143     0.317073
Precision        0.666667       0.297297     0.412698

Random Forest ---------------------------------------------------------------
----
                      Predict train, bus, ship  HSR, airplane  drive, ride
Actual train, bus, ship                    211             16           27
Actual HSR, airplane                        45             23            2
Actual drive, ride                          50              8           24

Accuracy:  0.6354679802955665
         train, bus, ship  HSR, airplane  drive, ride
Recall           0.830709       0.328571     0.292683
Precision        0.689542       0.489362     0.452830

Naïve Bayes -----------------------------------------------------------------
--
                      Predict train, bus, ship  HSR, airplane  drive, ride
Actual train, bus, ship                    200             11           43
Actual HSR, airplane                        45             16            9
Actual drive, ride                          34              4           44

Accuracy:  0.6403940886699507
         train, bus, ship  HSR, airplane  drive, ride
Recall           0.787402       0.228571     0.536585
Precision        0.716846       0.516129     0.458333
```

```
SVM -----------------------------------------------------------------
                     Predict train, bus, ship  HSR, airplane  drive, ride
Actual train, bus, ship                  233             10           11
Actual HSR, airplane                      53             17            0
Actual drive, ride                        58              5           19

Accuracy:   0.6625615763546798
          train, bus, ship  HSR, airplane  drive, ride
Recall            0.917323       0.242857     0.231707
Precision         0.677326       0.531250     0.633333

Logistic Regression ------------------------------------------------
----------
                     Predict train, bus, ship  HSR, airplane  drive, ride
Actual train, bus, ship                  227             13           14
Actual HSR, airplane                      47             19            4
Actual drive, ride                        45              7           30

Accuracy:   0.6798029556650246
          train, bus, ship  HSR, airplane  drive, ride
Recall            0.893701       0.271429     0.365854
Precision         0.711599       0.487179     0.625000
```
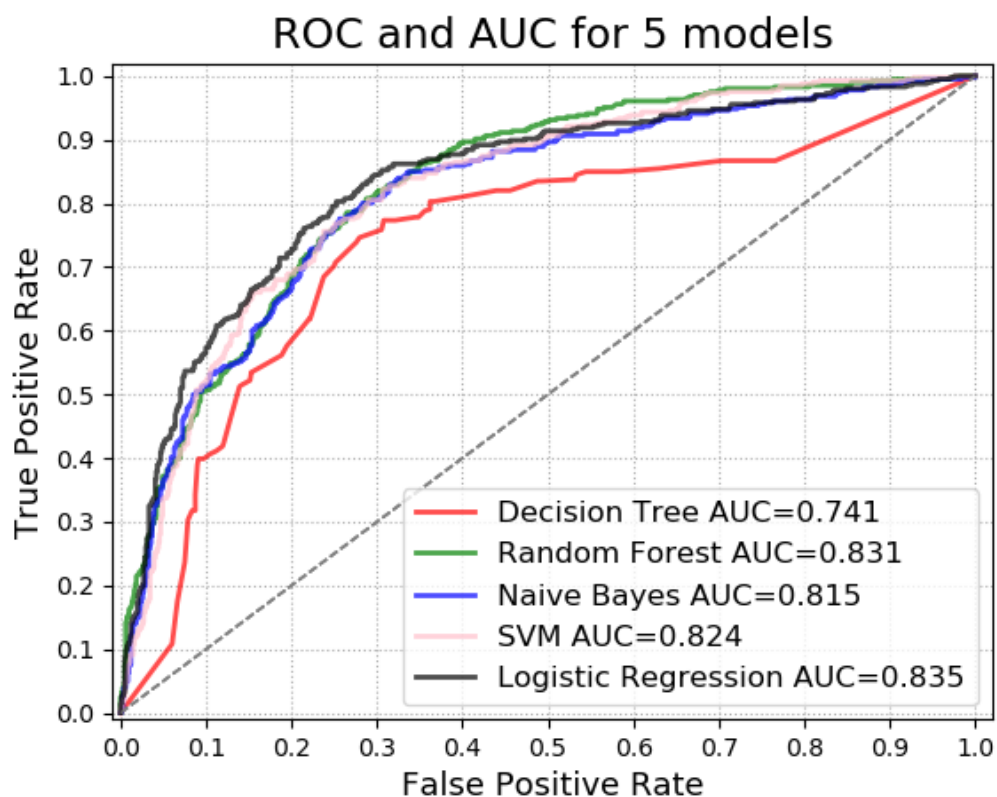
As the picture, Decision tree has the lowest accuracy, and logistic is the highest, but the prediction has too many on train and bus, this also decrease other two's recall.

The ROC curve shows that tree is the worst one, and

random forest and logistic are better.