Topic

Predict what you use when

backing home / travel

1. What we predict

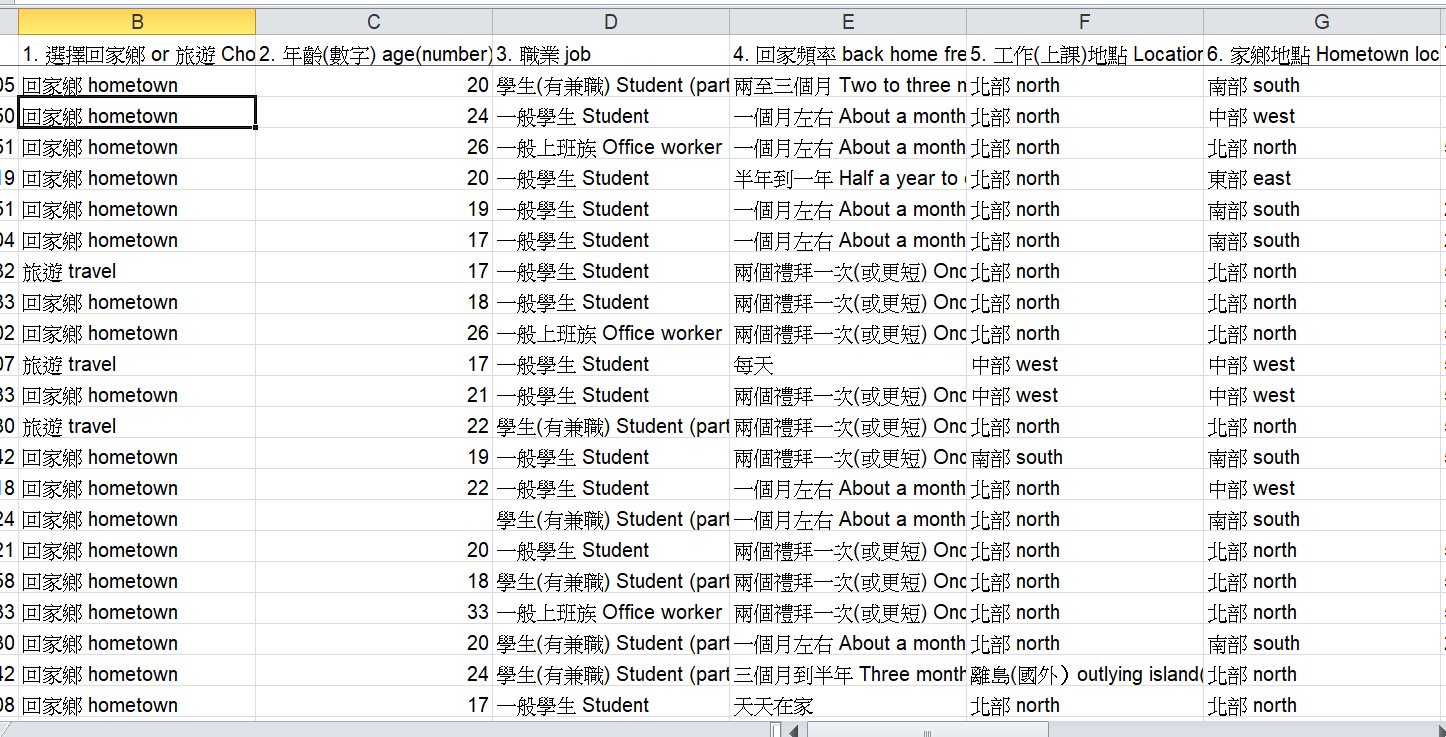
We collect a lot of data which contains one’s age, back home frequency, job, and other which may affect what transportation one person choose in backing home.

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.

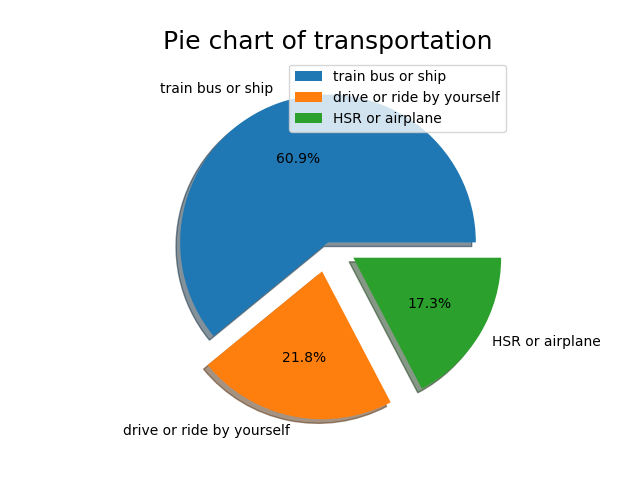
Note: someone’s hometown and location of work/school is at the same place, so they choose the data od traveling

1. 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.

1. The data distribution

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.

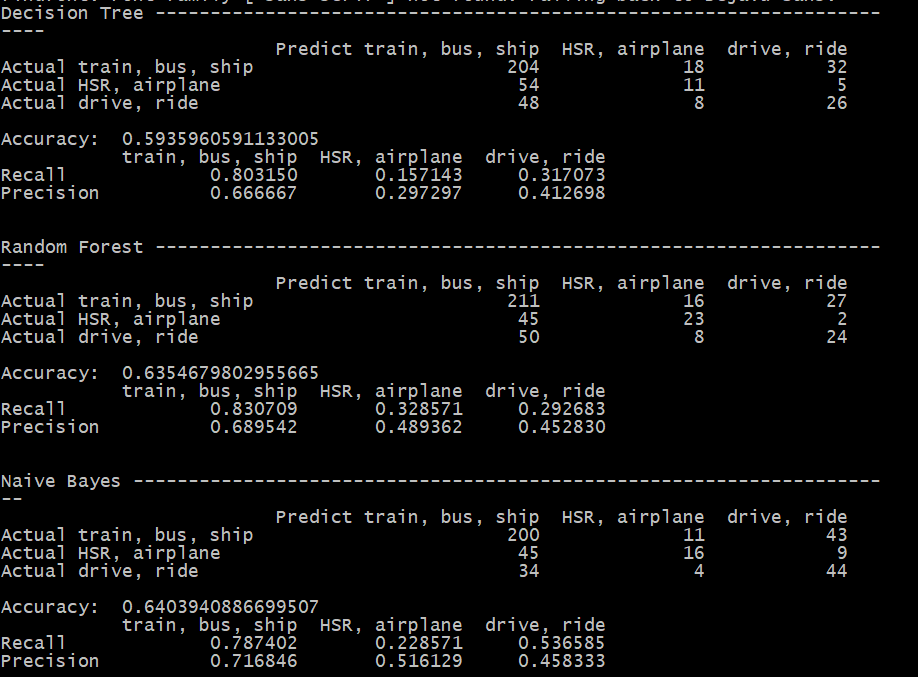
1. Data processing

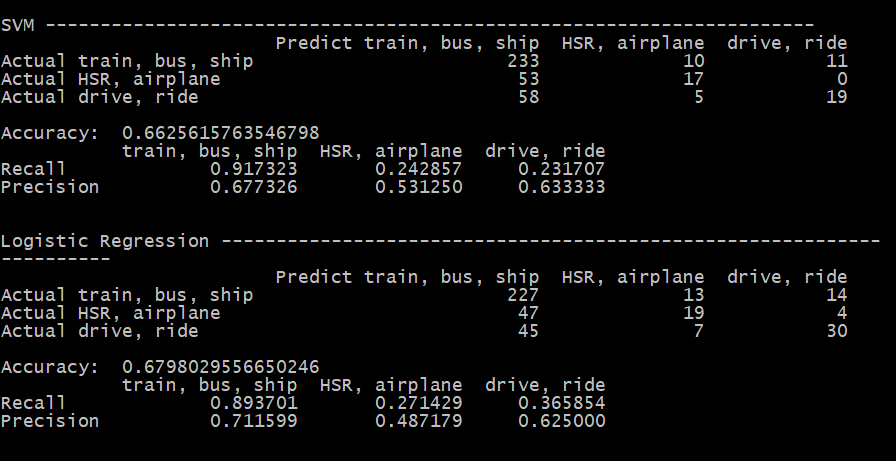
We fill the missing with mode. Then transform them to numeric data to fit in sklearn.

1. Model

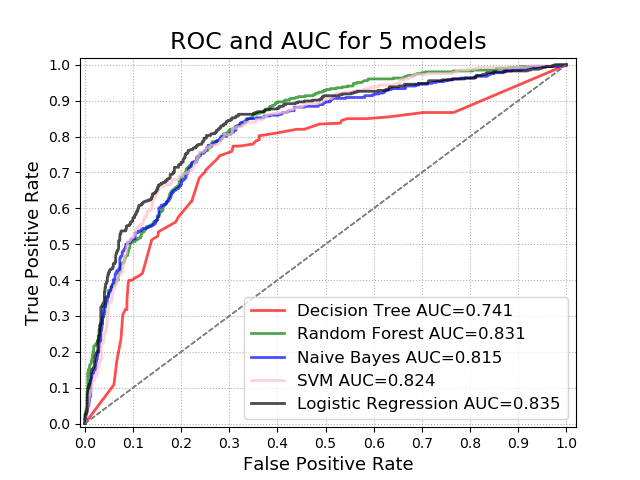
Five models be created, which are Decision tree, Random forest, Naïve bayes, SVM, Logistic. We implement our model first, but the result is worse than sklearn, so choose the sklearn.

1. Predict result





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.