Machin Learning for the Big Data

Project 2

M. Zaher Abd Ulmoula

B00761188

Project 2

In this project we read data about user daily trajectories which content those data ("t\_user\_id", "transportation\_mode", "collected\_time", "latitude", "longitude") and from that data we calcualte those features for each user by each day also for each transportation mode:

● Distance (e.g. Haversine)

● Speed (m/s)

● Acceleration (m/s2)

● Bearing (0 to 360 degrees)

in other python file Project\_2\_Hierarchical\_vs\_Flat we used the result to predict the transportation mode just from feature we have (Distance, speed, Acceleration, Bearing).

For the prediction we will use two structures first one is Hierarchical Structure and second one is Flat Structure.

# Method

## Calculate the t-test for the class:

To compere classes ["bus", "car", "walk", "taxi", "subway", "train"] the means of the (“Speed”, “Distance”, “Acceleration”) for each one feature so, I did the t-test between each pair of the classes this is the result:

Speed means:

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Statistic | P-Value | Null hypotheses |
| ["bus", "car"] | 4.76642940887e-108 | -23.53 | Reject |
| ["bus", "walk"] | 0.0 | 65.268 | Reject |
| ["bus", "taxi"] | 7.01341063315e-84 | -20.413 | Reject |
| ["bus", "subway"] | 2.29375450136e-118 | -24.786 | Reject |
| ["bus", "train"] | 0.0 | -62.742 | Reject |
| ["car", "walk"] | 0.0 | 76.0856 | Reject |
| ["car", "taxi"] | 0.0903408975559 | 1.69541 | Failed to Reject |
| ["car", "subway"] | 0.000773304055671 | -3.37262 | Reject |
| ["car", "train"] | 2.63370779644e-131 | -31.8343 | Reject |
| ["walk", "taxi"] | 0.0 | -71.7553 | Reject |
| ["walk", "subway"] | 0.0 | -68.0994 | Reject |
| ["walk", "train"] | 0.0 | -111.884 | Reject |
| ["subway", "train"] | 2.97240787069e-93 | -24.5277 | Reject |

Acceleration means:

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Statistic | P-Value | Null hypotheses |
| ["bus", "car"] | 2.65013729765e-09 | 5.97955728192 | Reject |
| ["bus", "walk"] | 2.33082044281e-114 | 23.4298513773 | Reject |
| ["bus", "taxi"] | 2.15983010078e-59 | -16.8378363986 | Reject |
| ["bus", "subway"] | 2.24252010761e-84 | -20.4343609455 | Reject |
| ["bus", "train"] | 7.51849577447e-309 | -47.6737075798 | Reject |
| ["car", "walk"] | 1.59691418259e-18 | 8.83464373085 | Reject |
| ["car", "taxi"] | 4.67317586158e-45 | -14.8963791172 | Reject |
| ["car", "subway"] | 2.46057811966e-45 | -14.8741220891 | Reject |
| ["car", "train"] | 5.35378337449e-129 | -31.3885130371 | Reject |
| ["walk", "taxi"] | 4.71850223657e-193 | -31.768844906 | Reject |
| ["walk", "subway"] | 6.73772145824e-239 | -35.9032888743 | Reject |
| ["walk", "train"] | 0.0 | -64.1216751826 | Reject |
| ["subway", "train"] | 1.52949027291e-52 | -16.8208022949 | Reject |

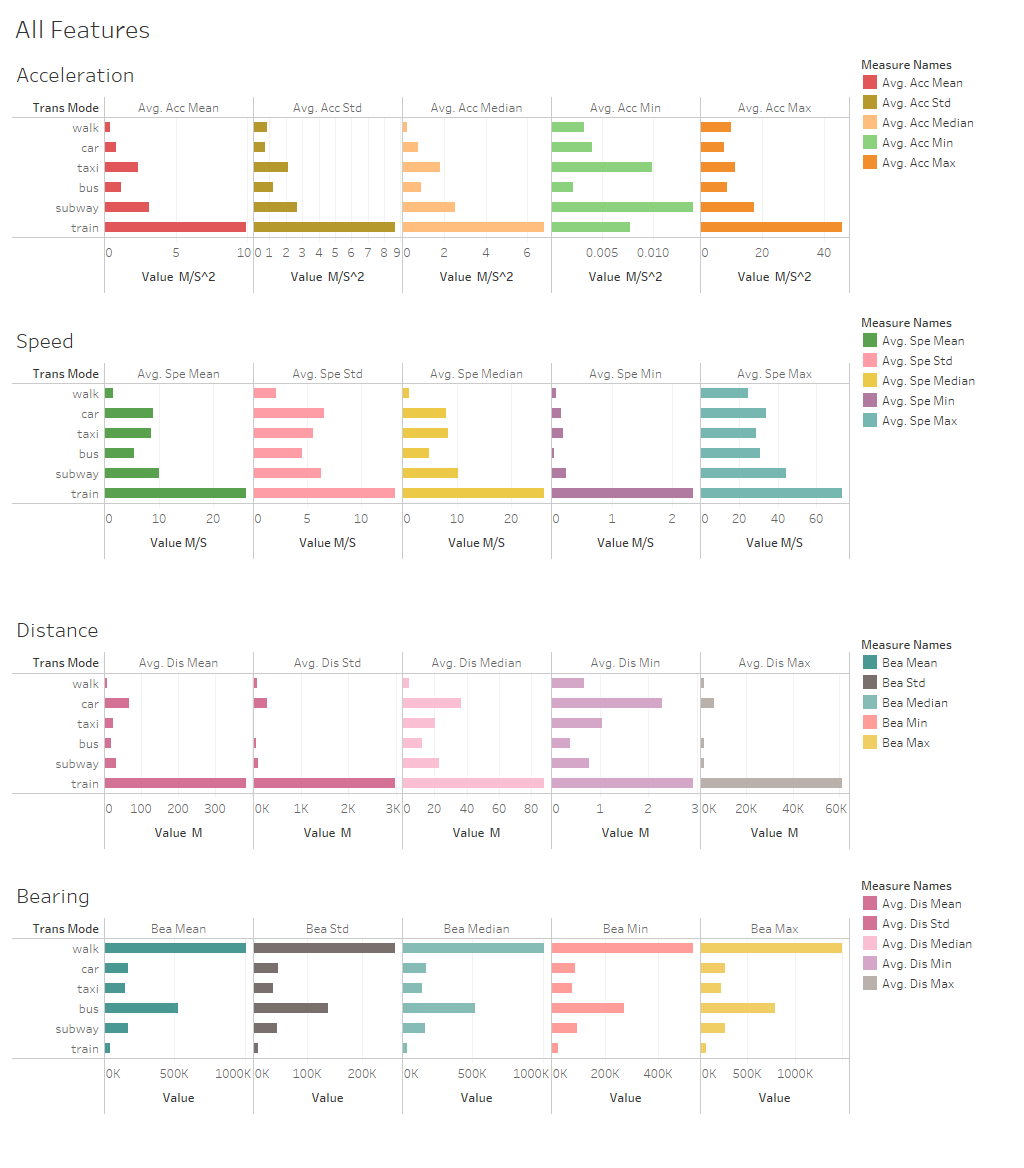
Distance means:

|  |  |  |  |
| --- | --- | --- | --- |
| Classes | Statistic | P-Value | Null hypotheses |
| ["bus", "car"] | 9.64441720598e-38 | -13.1164604133 | Reject |
| ["bus", "walk"] | 1.59988013392e-10 | 6.41113941713 | Reject |
| ["bus", "taxi"] | 4.42945690977e-05 | -4.09340457352 | Reject |
| ["bus", "subway"] | 5.33351772609e-17 | -8.4549341426 | Reject |
| ["bus", "train"] | 1.02190838427e-29 | -11.5557377554 | Reject |
| ["car", "walk"] | 3.35926312994e-68 | 17.8613902328 | Reject |
| ["car", "taxi"] | 4.33666553031e-10 | 6.31079195249 | Reject |
| ["car", "subway"] | 5.0138353073e-09 | 5.89879395809 | Reject |
| ["car", "train"] | 2.47807673746e-08 | -5.64967979534 | Reject |
| ["walk", "taxi"] | 4.56834005989e-12 | -6.9445500626 | Reject |
| ["walk", "subway"] | 2.88011772919e-26 | -10.6941797309 | Reject |
| ["walk", "train"] | 2.86210008427e-56 | -16.1487936106 | Reject |
| ["subway", "train"] | 1.50922133088e-10 | -6.51508036529 | Reject |

So, all the null hypotheses are rejected except Acceleration means of the Car class and Taxi which is excepted because the behaver of both is same and they are driving in same traffic.

## Plot the Features of each class:

Because the t-test didn’t help to find the similarity between the class I plot the features to see it.



## Analyses the class

### Bus, Taxi, and Car

From the plot of all feature we can see there is some similarity between the (bus, taxi, and car) because the acceleration and the speed is in same range but there are different in distance maybe because the number of users have car more from other or who use the public transportation each time use different one depends on his trip. Also, is idea of that class being similar is from there are using the same road, so they deal with traffic and they need to stop and drive again therefore the acceleration is close, and they have same limitation of speed which make the max and min speed is same.

### Train and Subway

Also, those class train and subway are similar and that because they are driving on there won rails there for they don’t deal with traffic or stop for passengers likes (bus, taxi, and car). Therefore, they drive in higher speed and acceleration is close too. They have some differentness like even the mean acceleration of the train the higher, but the min acceleration of subway is higher than the train which is true because the train need more time to increase the speed and train is haver than the subway. Also, the users site in train for longer distance like to travel between the cities therefor the distance is longer for the train.

### Walk

That leave us with the walking is different from any other class, which is obvious because the user walking is less from any other transportation, except for distance because there is outliers in the data. However, will not change in the classifications.

So, the hierarchical Structure in this project will be like this

The first model will fit by data have those labels (walk, ctb, sub\_train) which mean it just will learn the different between those different class. But the second model we lean to find the different between those (car, bus, train). The third one is for (subway, train).

### Classifiers

To do this project I need for two classifiers from SciKetLean library, so I chose the DecisionTreeClassifier and RandomForestClassifier because the implementation in the library is great with multi classification problems like what I have here and will good to find how they do if there different structure for classification.

## Code Files

* Project\_2\_cal\_features.ipynb
* Project\_2\_Hierarchical\_vs\_Flat.ipynb

# Result

After doing the code for this project to compeer between Hierarchical and Flat structure by using the DecisionTreeClassifier (DTC) and RandomForestClassifier (RFC) I got this result for Hierarchical:

A screenshot of a cell phone

Description generated with very high confidence

From the preives plot I found the RFC is doing better for classify the data in Hierarchical structure.

I got this result for Flat:

A close up of a map

Description generated with very high confidence

Now for the flat structure the DTC is doing better from RFC which give us idea how the classification structure could change the result of the algorithm because both of the algorithm got better result for the Hierarchical structure even those RFC had bad result for flat it got better in the Hierarchical. So, I did the t-test to find if there any similarity between there result I found

Hierarchical

p\_val 0.00097988283406 statistic -3.93076624398

the mean of model DTC is not identical to mean of model Random\_Forest

Flat

p\_val 0.000528543268696 statistic -4.20816659712

the mean of model DTC is not identical to mean of model Random\_Forest

there isn’t any similarity between them.