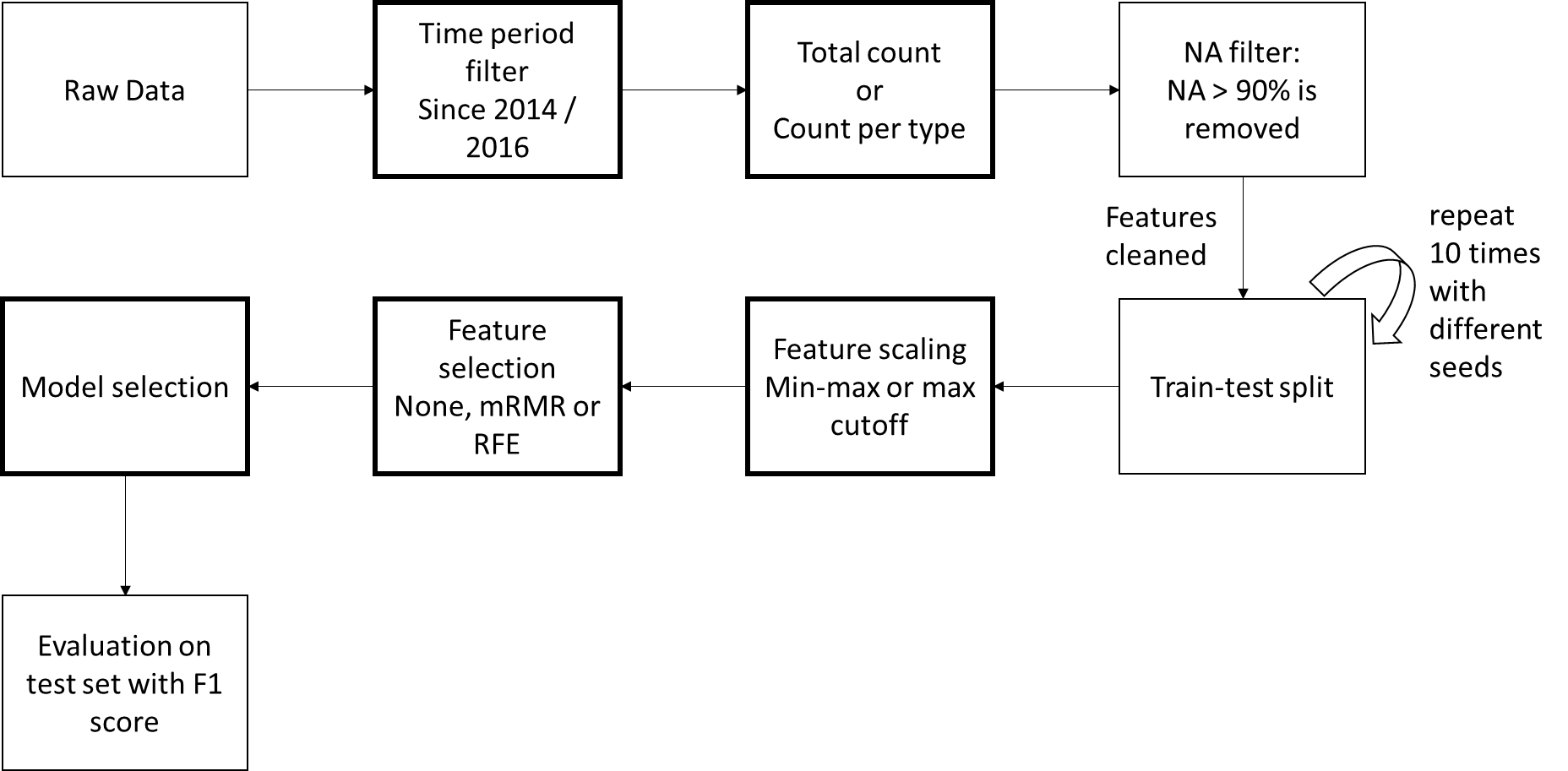
Experiment pipeline



The scores from rating platform is float. But the pipeline I used in LTS includes classification models. To use the existing code with the least modification, the following results is done by rounding the float score to integer labels (e.g. 2.4 -> 2, 4.51 -> 5).

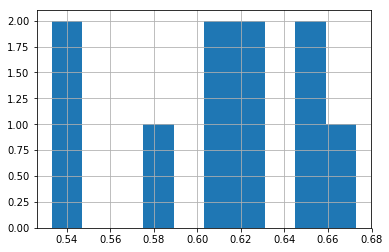
These experiments use all ratings (409 ratings by 41 users on 314 segments).

|  |  |
| --- | --- |
| For LTS   1. Using recent (since 2016) car related and neighborhood related features has a slightly better chance to perform better than longer period (since 2014). 2. For count-based feature, count per type is always better than total count. 3. Max cutoff has a slightly better chance to perform better than min-max. 4. In most cases, no feature selection is better. 5. Gradient boosting classification is the best among all models. Train errors of tuned gradient boosting in all 10 runs are 0. 6. Best test f1 score range between 0.71 and 0.73. | For rating platform   1. For car related and neighborhood related features, using longer period (since 2014) is slightly better than recent (since 2016). 2. For count-based feature, total count has a better chance to perform better than count per type. 3. Max cutoff is about the same as min-max. 4. Whether or not to use feature selection is inconclusive. 5. Which machine learning model is the best is inconclusive. 6. Best test f1 scores range from 0.53 to 0.673. |

The rest are the results of predicting rating platform score.

### Parameters of the best test f1 score of each run:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **seed** | **time period** | **total or by type** | **data scaling** | **feature selection** | **# features** | **model** | **test\_f1** |
| 0 | since 2016 | TOTAL | max\_cut | None | 65 | GDBreg | 0.672936 |
| 100 | since 2014 | TOTAL | max\_cut | rfecv\_linsvc | 40 | SVM | 0.611029 |
| 168352 | since 2016 | TOTAL | min-max | rfecv\_linsvc | 18 | logistics | 0.655987 |
| 291592 | since 2016 | TOTAL | min-max | rfecv\_linsvc | 25 | SVM | 0.533157 |
| 40918 | since 2014 | TOTAL | max\_cut | None | 65 | GDBcls | 0.62626 |
| 5258 | since 2014 | BY\_TYPE | min-max | None | 278 | linearSVM | 0.588693 |
| 57852 | since 2014 | BY\_TYPE | max\_cut | rfecv\_linsvc | 27 | RFcls | 0.53351 |
| 7821 | since 2014 | TOTAL | min-max | None | 65 | RFcls | 0.646994 |
| 789729423 | since 2014 | BY\_TYPE | max\_cut | None | 278 | MLPcls | 0.615597 |
| 972 | since 2014 | TOTAL | min-max | None | 65 | DTcls | 0.627967 |

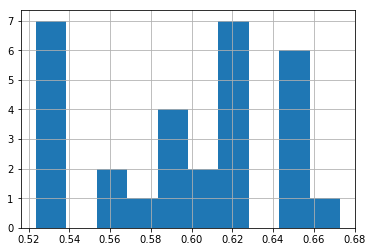


Distribution of test f1 score of 10 runs.

The distribution of 4 parameters are shown below

### Parameters of the best 3 test f1 scores of each run:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **seed** | **time period** | **total or by type** | **data scaling** | **feature selection** | **keep** | **model** | **test\_f1** |
| 0 | since 2016 | TOTAL | max\_cut | None | 65 | GDBreg | 0.672936 |
| 0 | since 2014 | BY\_TYPE | max\_cut | mrmr | 161 | GDBcls | 0.65779 |
| 0 | since 2016 | BY\_TYPE | max\_cut | rfecv\_linsvc | 33 | BAGcls | 0.657189 |
| 100 | since 2014 | TOTAL | max\_cut | rfecv\_linsvc | 40 | SVM | 0.611029 |
| 100 | since 2014 | TOTAL | min-max | rfecv\_linsvc | 40 | MLPcls | 0.555556 |
| 100 | since 2016 | BY\_TYPE | min-max | rfecv\_linsvc | 38 | linearSVM | 0.533865 |
| 168352 | since 2016 | TOTAL | min-max | rfecv\_linsvc | 18 | logistics | 0.655987 |
| 168352 | since 2014 | TOTAL | max\_cut | mrmr | 52 | BAGcls | 0.626114 |
| 168352 | since 2016 | BY\_TYPE | max\_cut | mrmr | 169 | GDBcls | 0.625587 |
| 291592 | since 2016 | TOTAL | min-max | rfecv\_linsvc | 25 | SVM | 0.533157 |
| 291592 | since 2016 | TOTAL | min-max | rfecv\_linsvc | 25 | ols | 0.532562 |
| 291592 | since 2016 | TOTAL | max\_cut | rfecv\_linsvc | 28 | RFcls | 0.523567 |
| 40918 | since 2014 | TOTAL | max\_cut | None | 65 | GDBcls | 0.62626 |
| 40918 | since 2016 | TOTAL | min-max | None | 65 | MLPreg | 0.604417 |
| 40918 | since 2014 | TOTAL | min-max | mrmr | 50 | ADAreg | 0.592915 |
| 5258 | since 2014 | BY\_TYPE | min-max | None | 278 | linearSVM | 0.588693 |
| 5258 | since 2014 | TOTAL | max\_cut | rfecv\_linsvc | 27 | ols | 0.585516 |
| 5258 | since 2014 | TOTAL | min-max | rfecv\_linsvc | 26 | SVM | 0.583354 |
| 57852 | since 2014 | BY\_TYPE | max\_cut | rfecv\_linsvc | 27 | RFcls | 0.53351 |
| 57852 | since 2016 | BY\_TYPE | min-max | None | 235 | GDBcls | 0.531836 |
| 57852 | since 2014 | TOTAL | min-max | None | 65 | BAGcls | 0.530121 |
| 7821 | since 2014 | TOTAL | min-max | None | 65 | RFcls | 0.646994 |
| 7821 | since 2014 | TOTAL | min-max | mrmr | 51 | RFcls | 0.646834 |
| 7821 | since 2016 | BY\_TYPE | max\_cut | rfecv\_linsvc | 33 | linearSVR | 0.644425 |
| 789729423 | since 2014 | BY\_TYPE | max\_cut | None | 278 | MLPcls | 0.615597 |
| 789729423 | since 2014 | TOTAL | max\_cut | rfecv\_linsvc | 31 | GDBcls | 0.57242 |
| 789729423 | since 2016 | BY\_TYPE | max\_cut | rfecv\_linsvc | 35 | MLPcls | 0.557559 |
| 972 | since 2014 | TOTAL | min-max | None | 65 | DTcls | 0.627967 |
| 972 | since 2016 | BY\_TYPE | max\_cut | mrmr | 165 | GDBcls | 0.620269 |
| 972 | since 2016 | TOTAL | max\_cut | mrmr | 51 | linearSVM | 0.619289 |



Distribution of top 3 test f1 score of 10 runs.

The distribution of 4 parameters are shown below: