ISQA: 4890

Name: Yu Guo

Professor: Yong Shi

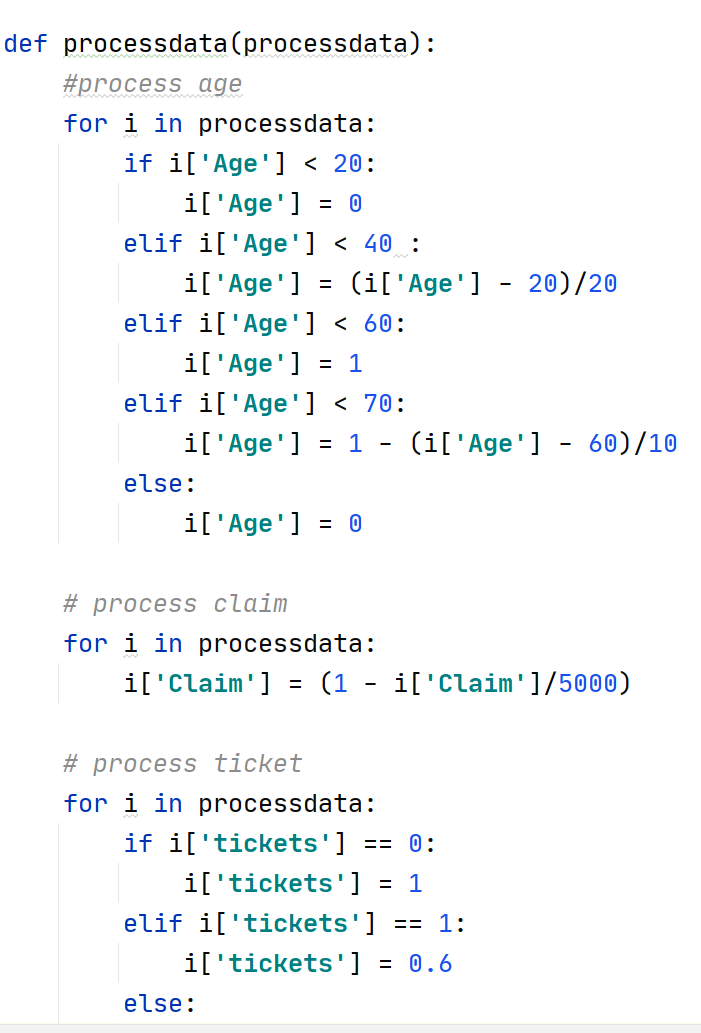
Task1: For the task1, I used python to solve both questions.

The code has 6 parts include:

1. Read data



2. Process data



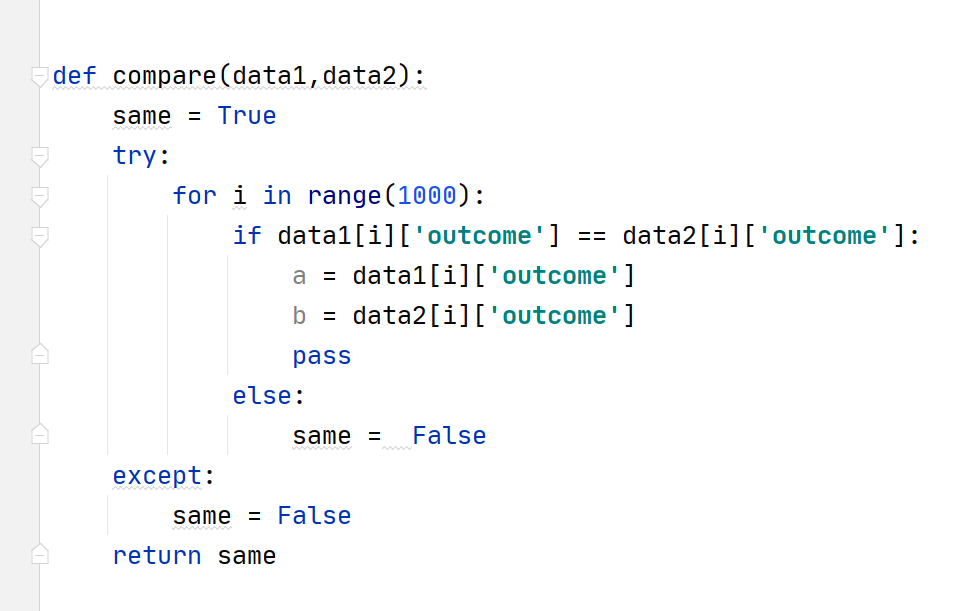
3.Trainning data



4.Getting new cluster



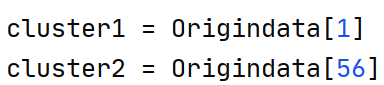
5.Compare two data result



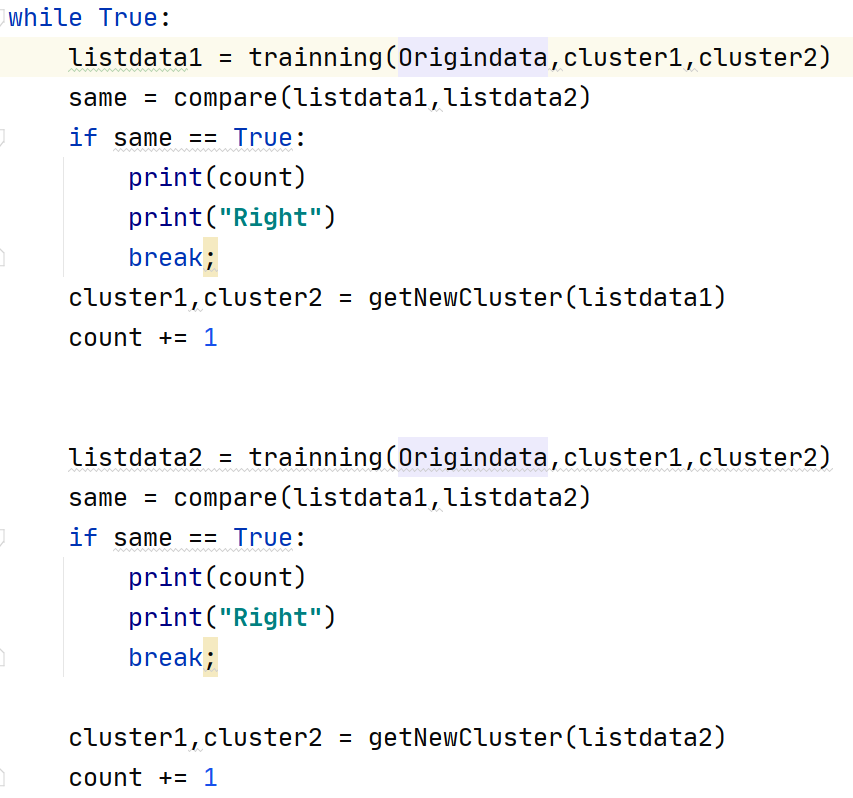
6. Main executive code



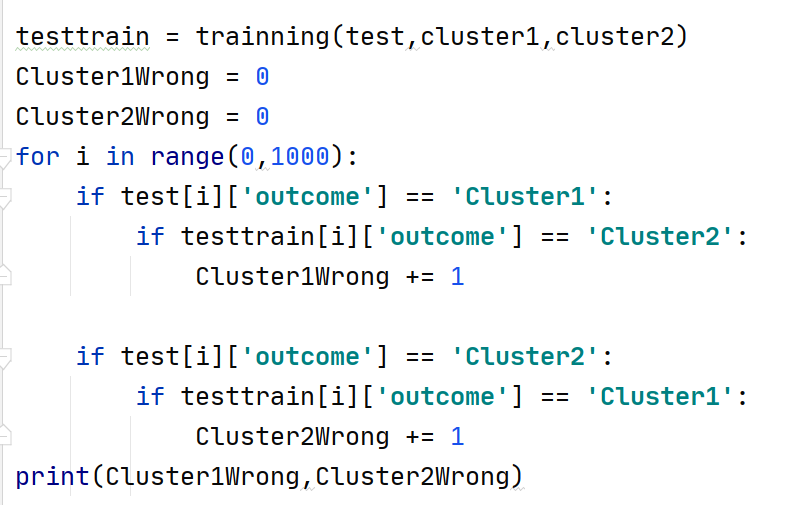
Question1:

For the Q1, I chose data in the first row, and data in the 56th row as the cluster1 and 2. 

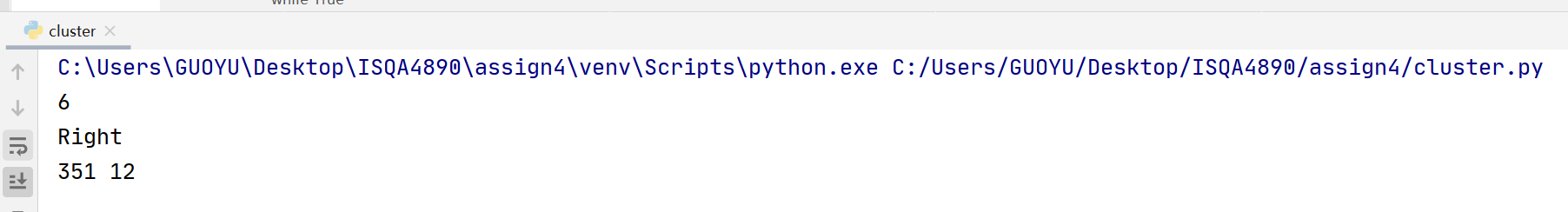
Then I used loop structure to training the data and generate the new cluster util two training data ge the same result of the cluster.



Then I used the latest clusters to continue training the other 1000 data and determine the difference between the real outcome and the mining outcome.

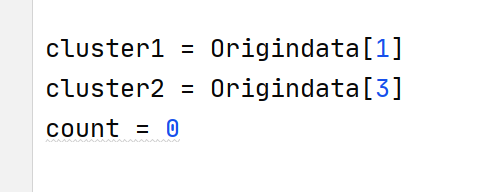


The result indicate that six loops have been executed to get the same result. And 351 + 12 errors. For the 351 errors, the model determine the real cluster1 as cluster2, and 12 errors the model determine the real cluster2 as the cluster1.



Question4:

The basic parts are the same also include six parts: Get data, process data, training data, get new cluster and determine result and main executive part. And the difference is the data is less dimensional than the question1. Since the data set itself has no tendency to compare, the main result of my validation process is to get the CLUSTER result of the test data and the center of the two CLUSTERs

.

Because the data do not provide the two clusters at the beginning, So I chose data in the first row and the data in the third row as the two clusters.

The result shows that it used 3 loops to finish the training to get the same cluster result. And in the test data, 22 data are cluster1, and 28 data are cluster2.

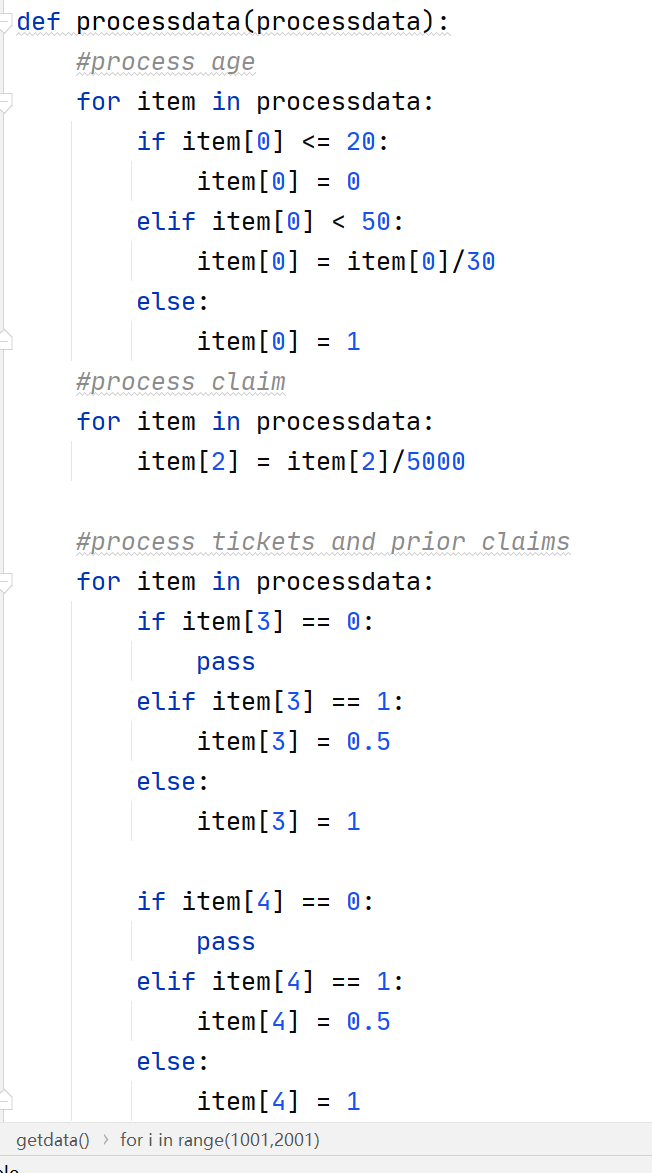
The latest clusters are:

{'Age': 0.07556905864197534, 'Income': 0.23762843366165942, 'Credit Rating': 0.2829861111111111, 'Risk': 0.9947916666666666, 'On-time': 0.8298611111111112, 'cluster': ''}

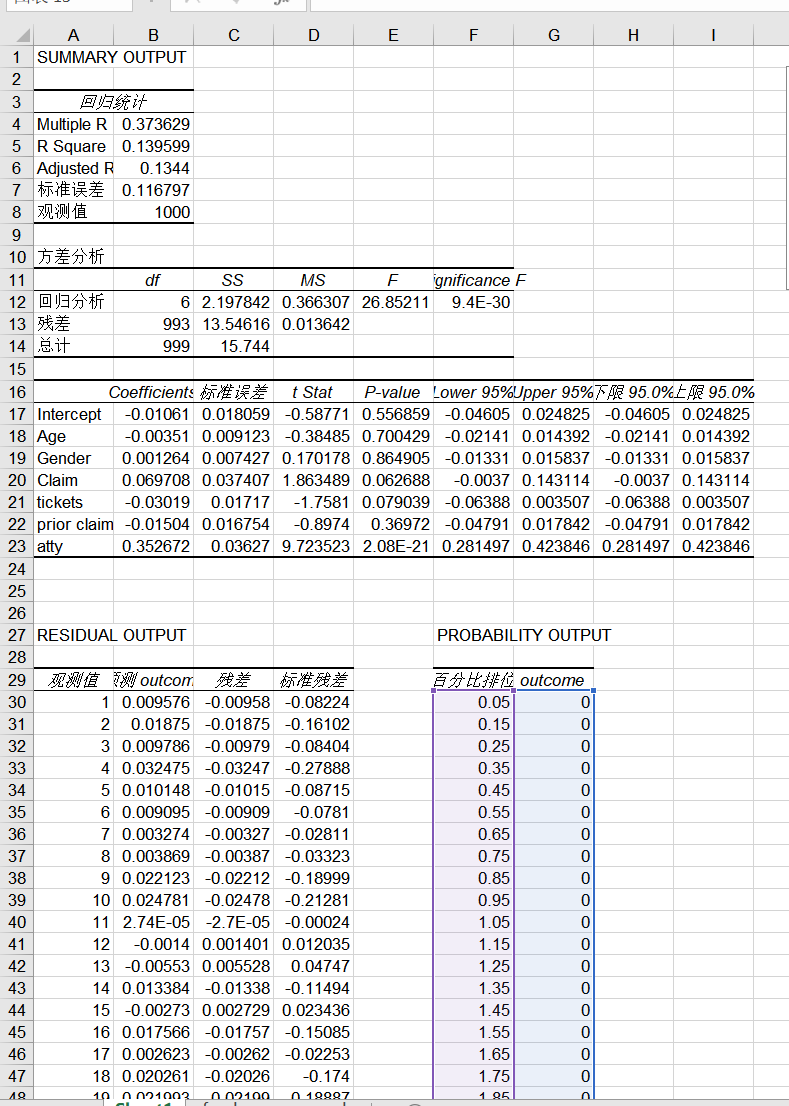
{'Age': 0.2768429487179485, 'Income': 0.2825627924958214, 'Credit Rating': 0.26121794871794873, 'Risk': 0.03365384615384615, 'On-time': 0.9647435897435898, 'cluster': ''}

Task2:

At the first, I used python to do the data transfer and store it into another xls file.



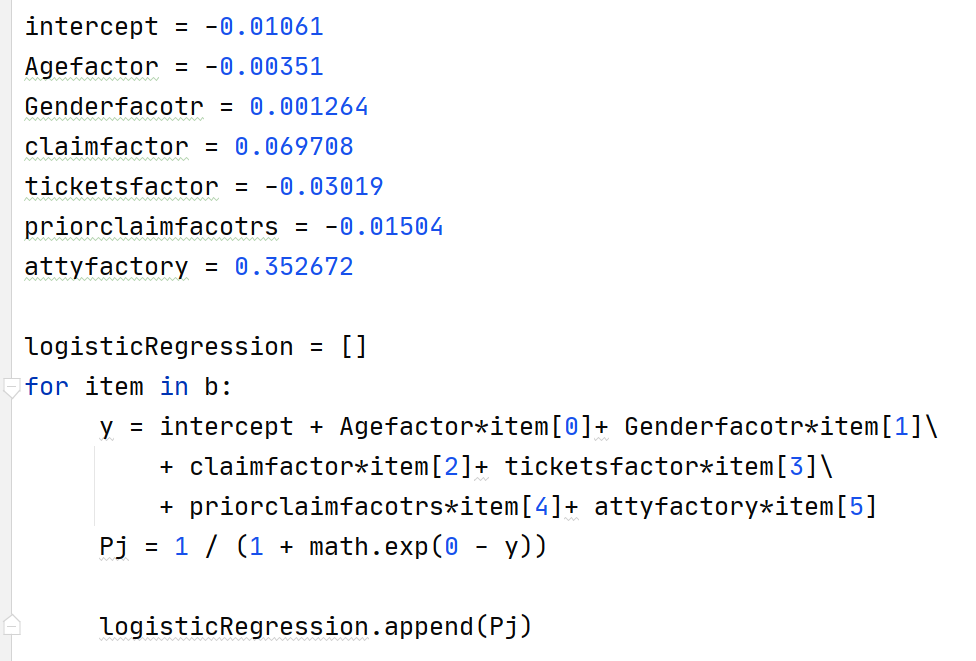


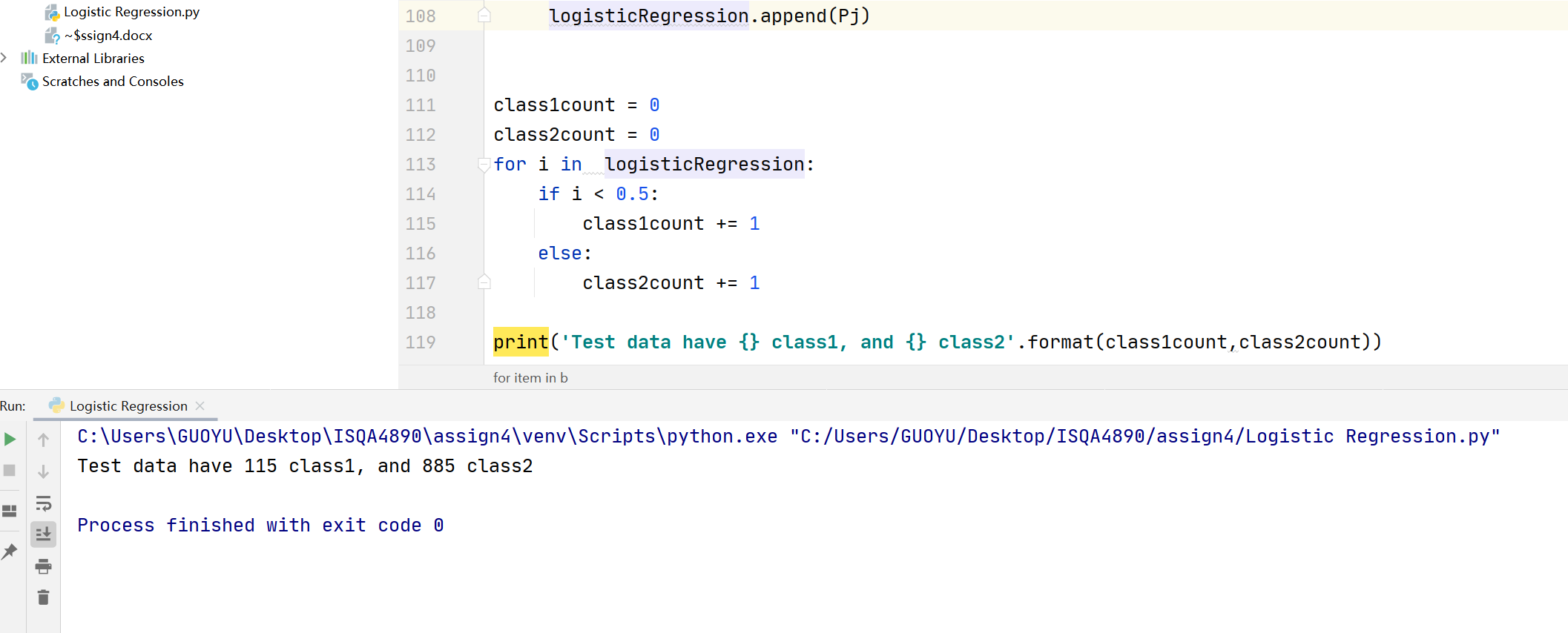
And then I used excel to do the multiple linear regression for the first 1000 data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *回归统计* | |  |  |  |  |  |  |  |
| Multiple R | 0.373629 |  |  |  |  |  |  |  |
| R Square | 0.139599 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.1344 |  |  |  |  |  |  |  |
| 标准误差 | 0.116797 |  |  |  |  |  |  |  |
| 观测值 | 1000 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| 方差分析 |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| 回归分析 | 6 | 2.197842 | 0.366307 | 26.85211 | 9.4E-30 |  |  |  |
| 残差 | 993 | 13.54616 | 0.013642 |  |  |  |  |  |
| 总计 | 999 | 15.744 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *标准误差* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *下限 95.0%* | *上限 95.0%* |
| Intercept | -0.01061 | 0.018059 | -0.58771 | 0.556859 | -0.04605 | 0.024825 | -0.04605 | 0.024825 |
| Age | -0.00351 | 0.009123 | -0.38485 | 0.700429 | -0.02141 | 0.014392 | -0.02141 | 0.014392 |
| Gender | 0.001264 | 0.007427 | 0.170178 | 0.864905 | -0.01331 | 0.015837 | -0.01331 | 0.015837 |
| Claim | 0.069708 | 0.037407 | 1.863489 | 0.062688 | -0.0037 | 0.143114 | -0.0037 | 0.143114 |
| tickets | -0.03019 | 0.01717 | -1.7581 | 0.079039 | -0.06388 | 0.003507 | -0.06388 | 0.003507 |
| prior claims | -0.01504 | 0.016754 | -0.8974 | 0.36972 | -0.04791 | 0.017842 | -0.04791 | 0.017842 |
| atty | 0.352672 | 0.03627 | 9.723523 | 2.08E-21 | 0.281497 | 0.423846 | 0.281497 | 0.423846 |

So the multivariate linear return formula:

Y = -0.01061 + age \* -0.00351 + gender\*0.001264 + claim \* 0.069708 + tickets\*-0.03019+ prior claims\*-0.01504 + atty\*0.352672

And then I used this data to verify the last 1000 test data.

And finally get the result:

In the 1000 test data, 115 are class2, and 885 are class2.