Since the we are carrying out a binary prediction about whether the song will be a hit (labelled as 1) or not be a hit (labelled as 0), logistic regression is the first model that comes in mind. The dataset is first pruned to obtain an equal balance of the labels, then divided into the training set which contains 80% of the whole dataset and the test set which contain the left 20%. By applying stochastic gradient descent on the training set until the parameters converge, we are able to obtain the parameters for each feature and then apply it to the test set. Unfortunately, the model is not desirable to this problem, it is too simple which leads to underfitting, after 100,000 iterations, the result is as below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 |
| Training set | 0.512 | 0.511 | 0.398 | 0.447 |
| Test set | 0.508 | 0.531 | 0.402 | 0.458 |

In order to make the dataset more complicated, we added binary features regarding their genres and obtain the following result.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 |
| Training set | 0.558 | 0.584 | 0.503 | 0.541 |
| Test set | 0.570 | 0.569 | 0.554 | 0.561 |

In comparison, we can see that a more complicated dataset does make the model to fit better, but still, simply using logistic regression model is not effective enough to solve this problem.