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 | 0.479 | 0.446 | 0.255 | 0.324 |
| Test set | 0.506 | 0.469 | 0.285 | 0.354 |

As such, given the current dataset, logistic regression model is useless for solving this problem. It requires more complicated features to better generate the model.