图书馆画像标准更新的工作的主要难点在于参数和模型的选取。我们主要数据来源有3处：图书馆出入馆数据，讨论间预定数据和图书馆书籍借阅数据。经过讨论，我们抓取了“出入馆天数，在图书馆内的总时长，预订讨论间次数，使用讨论间总时长，借阅书籍总数，借阅各类书籍数”作为学生的特征。

在数据预处理部分，由于部分同学没有图书借阅或讨论间的数据，对应数据项为None，所以需要先将所有的None改为0。另外，由于借阅图书的人数占比非常小，直接聚类效果非常不好，为了适应大部分同学的情况，在数据预处理部分，还增加了对数据取log的操作。

该分类问题为无监督分类，我们首先使用了最经典的KMeans模型。调用sklearn库中的KMeans模型对数据进行聚类，根据需求，先后指定k等于2和k等于3来训练模型并聚类，并且指定了随机种子0，保证在不改数据和任何参数的情况下，模型多次聚类的结果是前后一致的。另外，还调用了GaussianMixture(GMM)模型进行了和KMeans相似的操作。

然后，通过轮廓分析，探究分类效果，并将其可视化。考虑到部分特征中可能包含重复信息，部分特征噪声较大，我们通过删减部分特征重新训练分类模型，并同样通过轮廓分析对比前后分类结果质量。后考虑到部分特征相关性不大，将特征分为两类，分别聚类，并且将聚类结果相应的提升至二维。（“怎么样”的“某某者”）

最后，对比了十余种分类方法，选择通过log后的使用讨论间累计时长，在图书馆累计时长与借阅书籍总数作为特征，使用KMeans进行分类，通过两轮聚类（图书馆，讨论间特征一轮， 借阅数据一轮），划分出2\*3一共6类画像。