Process Report - Group 19

Xiaoyu Yang [2640948, xyg230], Jiamian Liu [2632301, jlu510], and
Song Yang[2638728, syg340]

1 Project Schedule

From April 24 to April 30, we dived deep into the dataset and got a general understanding of the assignment and a feeling of the data. Besides, we also did detailed research on relevant literature. During this period, we learned common necessary steps for a recommendation system, including data-processing work to get suitable data format, feature engineering for model and several classifiers such as random forest, SVM and so on for this assignment.
From May 1 to May 7, we did some research on relevant materials to implement some simple and initial approaches for recommendation system. Then we did data preparation and pre-processing. Furthermore, we also tried some simple data processing work, such as filling in missing values and deal with classifiers without parameters.
From May 8 to May 12, we did business understanding work to understand some previous approaches and found some efficient approaches for our assignment. We thought we'd better balance data in our model, do some optimized approaches for missing values and make inner correlations between values.
From May 13 to May 16, we implemented several steps to train and establish the model: We trained our model using the processed dataset with random classifier and then SVM classifier.
From May 16 to May 20, we ran and got our prediction results for the recommendation system. We adjusted the trained model and then predicted the result of recommendation system along with changing some parameters for better results.
From May 20 to May 24, we evaluated our result and model, optimized our model for higher accuracy and wrote the report. We also tried to combine some other features and approaches further learned from some other materials.

2 Division of Work

Xiaoyu Yang was mainly responsible for feature engineering, which included extracting and finding the inner correlation within the dataset to get a better feeling about the dataset. Besides, he did the establishment of the model, which included using different classifiers to train the model and ensemble different models to get higher accuracy.
Jiamian Liu was mainly responsible for data pre-processing, which included handling the missing data with proper values, normalize data and so on. Also, some feature engineering work, which included extracting and finding the inner correlation within the dataset.
Song Yang was responsible for data pre-processing as well. Furthermore, he also did the work of training and evaluating the model to get the prediction results, which included adapting the trained model into testset to get different prediction results and optimizing the model.
The report is the teamwork of all of us.

3 Team Cooperation

For team cooperation, all our team members were very active. We hoped to improve our assignment through everyone's continuous efforts, which could hopefully lead to better results in the competition and the report. Moreover, we hoped that through this project, we could not only learn the knowledge of data mining and fight for the leaderboard but also exercised our team cooperation ability, which included communication during group meetings, discussion through our slack channel and using Git, agile methods for development.

3.1 Availability

All the team members tried their best to devote themselves to the project since the knowledge and the experience of the project was very interesting and useful. Moreover, we were eager to gain more data mining and machine learning experience, which helped us all to do research on the models and to combine different features and tricks for development. For example, every week all team members actively discussed the problems and shared their findings together.

3.2 Response Time

Since everyone was very active as mentioned above, we solved most of our problems quickly. We adopted the concept of agile development and improved our model through continuous iteration. In feature engineering, we brainstormed together to decided features to deal with. Anyone who had new ideas, all of us respond quickly to develop implementations and tested models. Through our rapid response and active cooperation, we constantly improved the work of feature engineering, and finally achieved decent models and results. For example, running on a single machine was time-consuming because of the large amount of data. We worked together on solutions and cut the dataset so that we could operate on different machines.

3.3 Knowledge Contribution

We handled different modules as is shown in Fig.1. For example, we every Sunday evening, there was a meeting to share and discuss the progress as well as relevant knowledge and findings of the project. At weekly meetings, we discussed the latest technology of different aspects of data mining and how to continuously improve our model. Moreover, we have also established a small knowledge sharing library in slack and Google drive to share state-of-the-art methods and papers. We believed that sharing the knowledge not only enable us to complete this project well, but also provided us with a good horizon for the future of data mining.

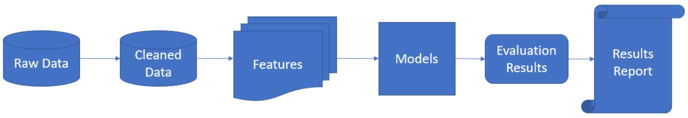

Fig. 1. The pipeline of assignment 2