COMP9321 Assignment 3 Report

Wenxuan Shi
z5182291
23 APR 2021

Part-I: Regression

Performance on CSE Vlab machine:

```
zid,MSE,correlation
z5182291,6869109729388241.0,0.68
```

Evaluation and Improvement:

At the beginning, I only chose to select top 6 casts/crews/production_companies, and added them to the final comparison model.

However, I found that the results were getting greater while the number of top_num increased, but also getting worse when the top_num was very large (e.g. 25 or more). After detailed testing for the range of 6 to 22, I found top_num with 16 produced the best results.

In addition, I didn't take column **budget** and **runtime** into consideration, and the result was not satisfied before adding these two columns. After involving **budget** and **runtime** column, the result was getting higher as well.

For column **crew**, I initially considered all crews such as Director, Director of photographer etc. However, I found for a movie, the position of Director has a huge effect on movie performance, compared with other crews, so I decided to only consider Director, and ignored other crews in a movie.

For model using, I chose RandomForestRegressor. One of the main reason is that it reduces overfitting problem which happens in decision trees, and so it helps to improve the accuracy.

I also changed my code for program efficiency concern. Initially, I extracted the name of each item I chose, such as the name of an actor, the name of a production_company etc. But I found it is a little bit more efficient if I just extracted the ids of them.

Problems faced:

The data type of some columns are different, so I need to use different techniques to pre-process the data. Also, I found it more achievable and efficient to store items into dictionaries and then count the number of occurrences.

Part-II: Classification

Performance on CSE Vlab machine:

```
zid,average_precision,average_recall,accuracy
z5182291,0.70,0.54,0.71
```

Evaluation and Improvement:

I used the same technique as Part-I to prepare the data. I also paid attention to the performance of Classification while I changed my strategy which described in Part-I Evaluation and Improvement.

Besides, for training model selection, I chose KNeighbors Classifier. I found that n_neighbors ranged from 5 to 14 produced similar results. With the consideration of efficiency, I decided to chose a lower n_neighbors. Finally, I set it to 5.

Problems faced:

I faced the same problems mentioned in Part-I. Also, in order to improve my model, I researched the parameters that I could changed, and finally tested that the number of n_neighbors would affect on my model.