## Part 1 : Regression

To get the performance of my model, I need to preprocess the data based on the dataset. First, I used a dictionary to calculate the times of each person in cast column and then I sort the dictionary to select the ten people who are most frequent in the column. I also did this for crew and production companies. For the rest column, I set 1 to the row whose movie has a homepage and set 0 to which does not have a homepage. I also did this for the original language. Set 1 to English language and set 0 to other languages. For production countries, I calculate the length of each row, which means the number of the production countries for this movie. I also did this for spoken language. For the release date column, I divided into four seasons. From January to March, I set to 1. From April to June, I set to 2. From July to September, I set to 3 and I set 0 to the rest month. After the preprocess the data, I use random forest method to train the data and found the correlation performances not good. So for the cast column, I calculate the number of the top ten people in each row and did the same action to crew and production companies. I retrain the model and finally got the satisfying result.

## Part 2 Classification

In order to do the classification, I did the same preprocess data method for part 2. And I used gradient boosting classifier method to train the data. For the first run, the result performance well, so I did not do any changes to improve the model.