**Introduction**

The Hong Kong film awards ( HKFA) was established in 1982. Since 2019, it has been successfully hold for 38 years. The purposes of the Hong Kong Film Awards are to promote Hong Kong films locally and abroad, recognize outstanding achievements, encourage professional development and promote film culture. Through this competition, more excellent films could emerge in Hong Kong film industry. There are 21 awards in HKFA, including, for example, Best Film, Best Director, Best Actor, Best Actress, Best New Performer and so on.

Aim at having a better understanding of HKFA, this paper tries to explore the competition by following research questions:

1. Which factors affect most to win the Best Film in HKFA ?

2. Who is the runner-ups ,which in Chinese “陪跑王” , means the actors who had been nominated most while never won the prize, in HKFA?

**Data Aquisition**

Mtime is China's famous movie portal with 170 million unique visitors per month and is the largest online ticketing and movie merchandising company in the country. Mtime has one of the largest movie entertainment community in China.On Mtime, we found the complete information of hkfa, including films, actors and film reviews. we choose to scrap the data from Mtime as our data source.

**1. Best Runner-ups**

In viewing of the best runner-ups, we target at Best Actor nominees and winners, movie reviews, movie ratings. On a single webpage of the HKFA’s best actor or actress section, all the data related to the awards including time, actor’s name, movie’s name were wrapped in a <class=‘event\_awards event\_list’ , with the information of winners in tag < class=‘yellowbox’ >, nominees in tag <class=‘bluebox’>.

Based on this understanding of the webpage structure, we first found all the <class\_='event\_awards event\_list> tags in which we subsequently found all related tags by using the “requests” and “beautifulsoup4” packages, and appended them to their respective lists. Before starting harvesting, three empty lists named “years”, “wins”, “nominations” had been created to store the harvested data. Since the data we needed were store on multiple webpages, we wrote a for-loop to scrape all pages automatically.



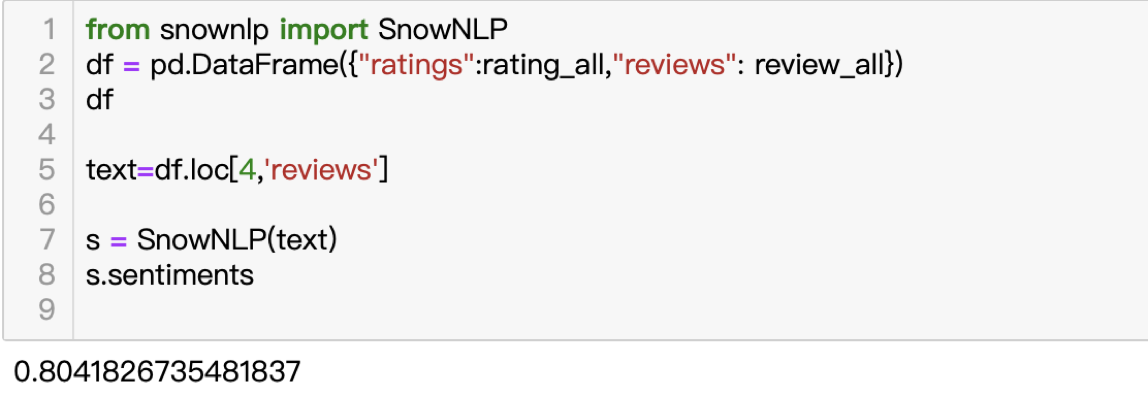
**2. Best Film Awards**

In viewing of the best film, we target the movie ratings of the nominated and the best film. On a single webpage of the HKFA’s best film section, the score of each movie related to four aspects, which are music, frame, director and story. And the full mark of each aspect is 10. However, the score of each aspect did not appear directly on the webpage. We found that the score of each aspect present to the 'elements-check' in HTML. All the data related to the score of a certain aspect were wrapped in a<div pan="M14\_Movie\_Overview\_Rating\_Final" class=gradebox \_\_r\_c\_"> html tag, with the title in a <dd class="hm"> contains the score of movie frame, the title in a <dd class="yy"> contains the score of the music, the title in a <dd class="dy"> contains the score of the director, the title in a <dd class="gs"> contains the score of the story. Four empyt lists named "画面","音乐","导演", and "故事" had been created to store the harvested data.



In addition, we also consider the comments of each movie with some importance. So we use Natural Language Processing (NLP) to calculate the audience sentiment score through their short reviews of the best films and those nominated films for over the years.





At last, we combined the five factors "director", "frame", "music", "story" and "NLP" as our input data to explore the best film awards. The picture below shows our dataset and we define the best movie is "1" while the nominated movie is "0". This dataset finally exported as a CSV file. Our total dataset contains 165 movies with their five feature data.



**Data exploration-*Best Film***

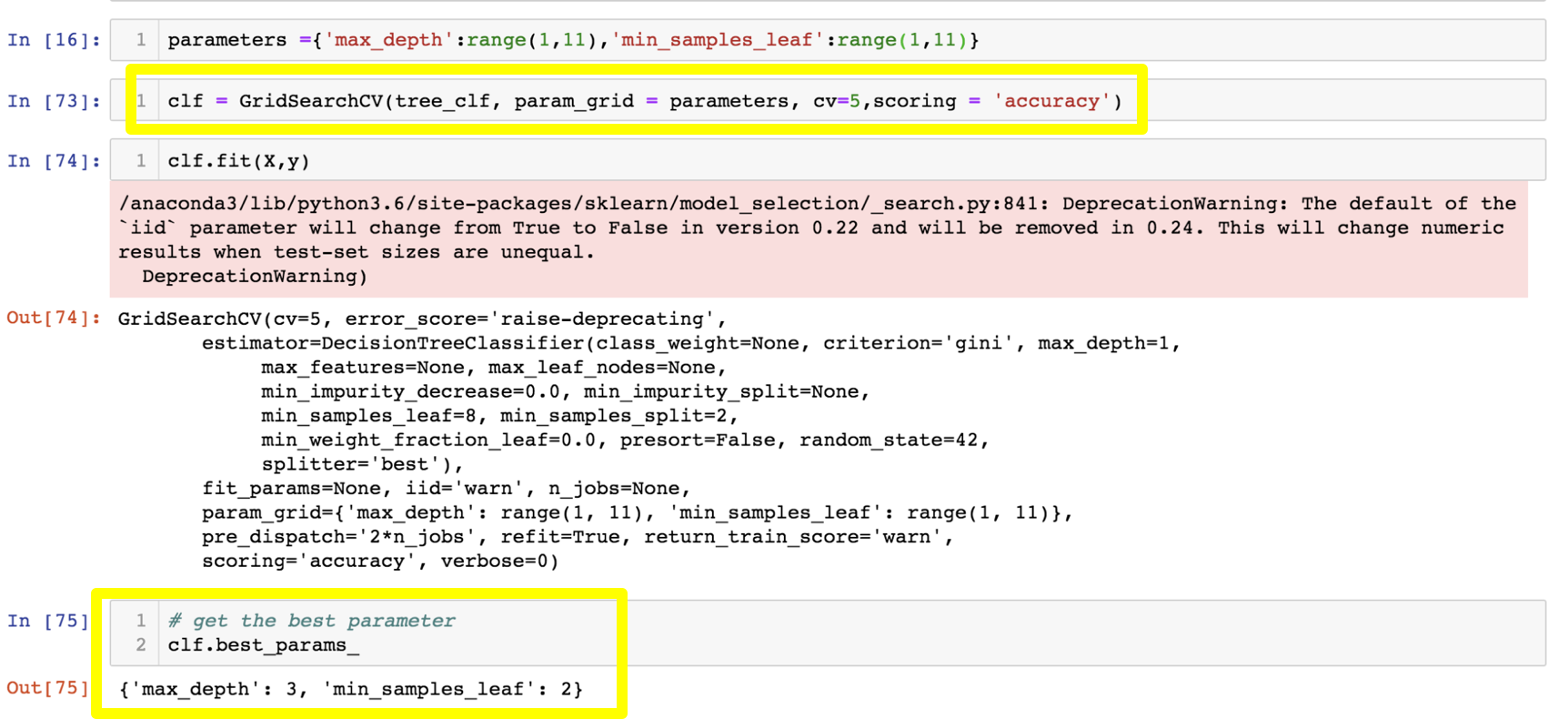
Aiming to predict and test which factors affect the best film awards most and answer our first research question. We decided to use the decision tree to train the data and found out which factor matters the most. Firstly, we computed the percentage of winning movie in the whole dataset was almost 0.21. It means that the percentage of winning the best film is 20%.



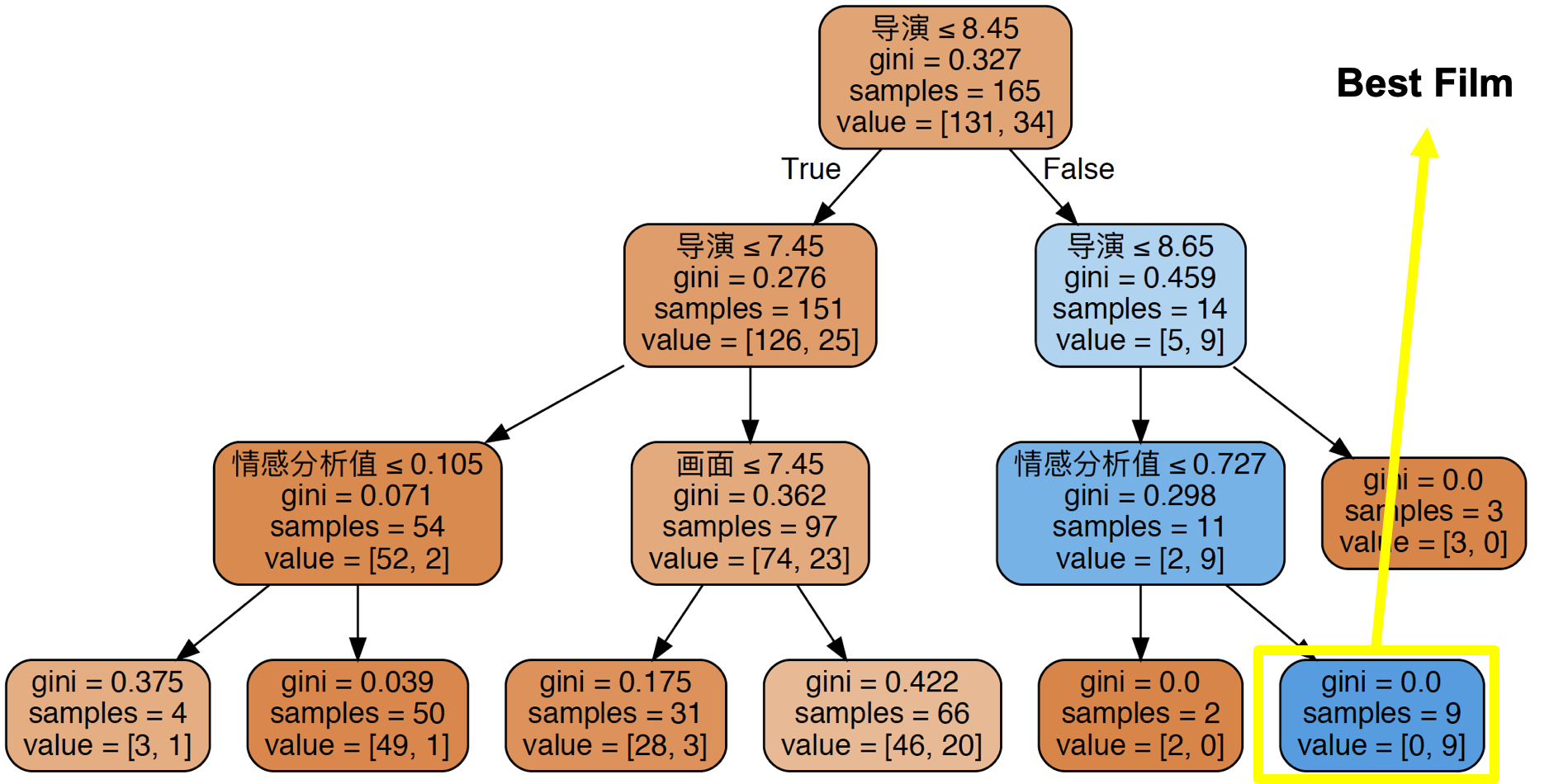
Secondly, we split the datasets into trainset and test set. The test size is the proportion of the test set in the whole dataset.



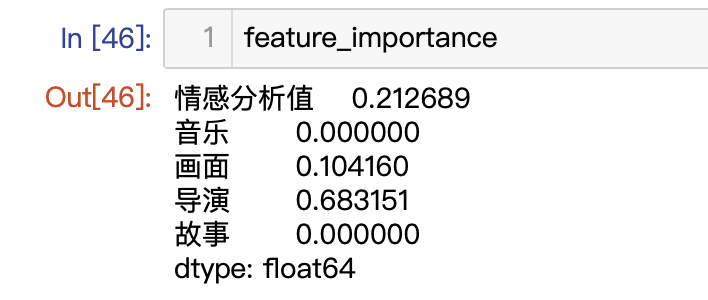
After that, we computed that our accuracy score of the test set is 0.76 while our accuracy score of the train set is 1. Our model does not perform well on the test set. So we choose the parameter for the decision tree. We used GridSearch to find out the best suitable parameters. Then it will classify the train set until different labels are separated in certain parameters to avoid overfitting.



Meanwhile, our dataset is small (165 movies). If we use train test split, then there is not enough data point to train our model. So to use all the data. We apply cross-validation to tune the parameters. Through trial-and-error, we append 5 to k value and find our best parameters. As a result, our best parameters in the screenshot above contained the max depth as 3 and the minimum sample leaf as 2.

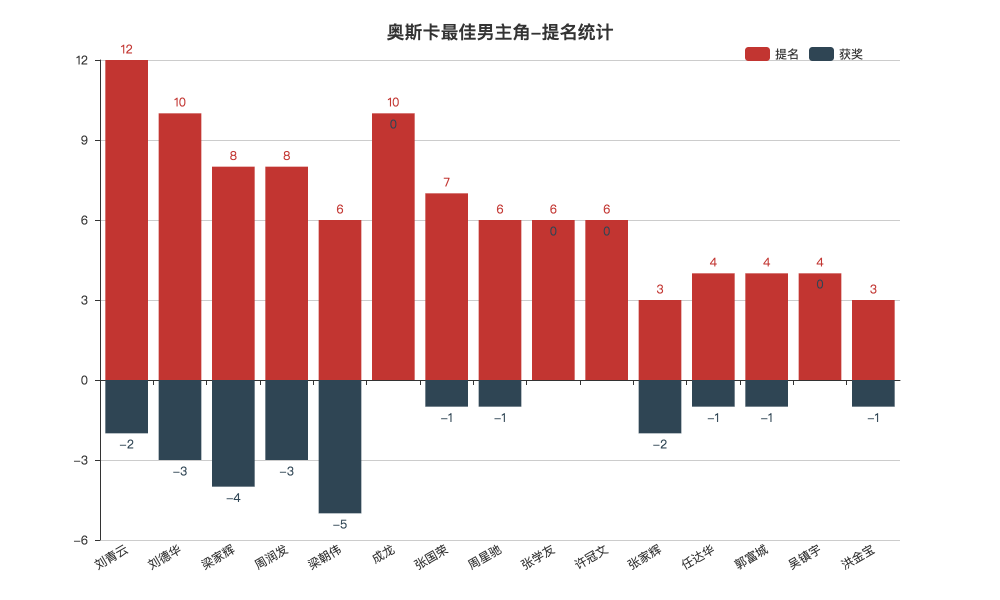


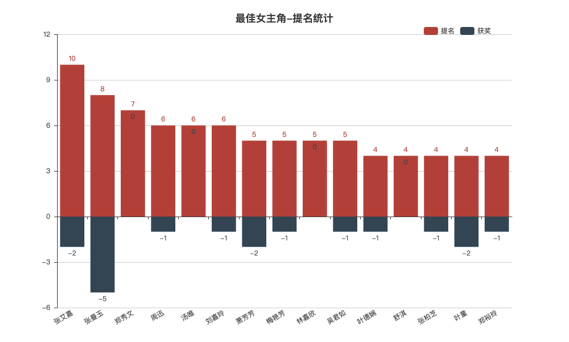
We made full use of the best parameters and graphviz to visualize the decision tree. The blue box which has been pointed indicated the films that won the award. As we can see, the gini equals to 0. The total sample is 9 while the first parameter of the value means the number of films which won the best film. Besides, the feature importance of those five factors as following:



Therefore, we figured out that the grading to the director of a movie matters the most while the NLP score also with some importance. The decision tree tells us that if the score of the director of a movie is higher than 8.45, lowerthan or equal to 8.65 with NLP score higher than 0.727, it will hasa higher chance to become the best film. Finally, our decision tree gets 0.84 of accuracy score.

**Data exploration*-Best Runner-ups***





In order to find out the answer of our research question, Who is the best runner-up in HKFA, we collect the data of Best Actor's winners and nominees. Using Pyecharts package to draw the bar above. We define best runner-up as the actors who had been nominated most while never won the prize. From the result, we can found that Jackie Chan and Sammy Cheng are the Best Runner-ups. Jackie has been nominated for 10 times and Sammy for seven times, while both of them never win the prize in the past 38 years.



Word cloud of Jackie Chan's Best Actor nominated movies review



Word cloud of Best Actor winners movies review

Moreover, we use jieba and WordCloud package to do the text mining. We scrap the movie reviews text and choose jieba as our text processing tool. Then we apply WordCloud to visualize the result and explore the key focus point audiences have toward best runner-ups and winners. Through comparing the differences and similarities among these word clouds, we tend to suggest that when considering who would win the best actor or actress, not only the performance of actors matter, but directors also play an essential role.

**Limitations and improvements**

Firstly, our data sample size is not big enough. We just focusing on the hk

film awards and it only last about 38 years until now.

Secondly, the Mtmime website we explore is not so well-known.

Thirdly, the marking criteria of the movies are changing due to the development

of high-quality movie frames. So actually it is not so precise that we put all the movie into the same criterion.

As for further improvement, we consider using TF-IDF (term frequency–inverse document frequency) to have a better analysis of movie reviews. Applying feature engineering process to improve our model's performance.