BADM Project Info

Websites to scrape from:

* <http://tw.dorama.info/drama/d_box_office.php?date=2015-10-25&ud=0>
* <https://tw.movies.yahoo.com/movieinfo_main.html/>
* http://tw.dorama.info/drama/d\_box\_office.php?date=2015-10-25&ud=0

Variables

* Movie Title
* Release Date
* Type of Movie
* Length of Movie
* Issuing Company
* Expected Ratings (before release)
* Actual received ratings (after release)
* Comment ratings (after release)
* Comment Date
* Number of Trailer
* Actors
* Directors
* Weekly Income

### Proposal Format

**2 pages**, 11-point font, 1-inch margins, 1.15 line space, PDF format. Please upload your file here.

### Content and Organization

The proposal should include the following components:

* **Suitable Title**: Should reflect the business and analytic goals
* **Business Goal and Humanistic Evaluation**
  + Who is the stakeholder or client?: Movie theaters in Taiwan.
  + Who will be affected by the solution?: Producing company of the movie, Theaters in Taiwan, Taiwanese customers of the theaters.
  + A description of the business goal. What are the business benefits of implementing this idea?

The business benefits relate to **(1) the theater companies maximizing their profits by meeting the demands of their customers more efficiently. i.e. showing the movies that the customers want to see. Customers will develop some form of loyalty to specific theaters when their demands are efficiently met.** (2) The producing company will be able to analyze how the relationship between the combination of actors, directors and type of movie will affect the income of the movie.

* + What opportunity is it creating? To meet the demands of Taiwanese customers of movie theaters. What shortcoming does it address? The following errors may be introduced when collecting data:

Sampling error, because there are many people who didn't give a rating nor comment (some people choose not to rate nor comment online)

* + What are possible dangers of this application? Illegal company will copy the movies which has high rating and sell it to customers before the movie is release.
  + Who might be harmed and how? Movie theaters, because customers will not go to theaters.
  + What would be considered a success? Predicting actual ranking accurately so movie theaters can better prepare the logistics of showing movies in advance.
* **Analytics/Data Mining Goal**
  + A description of the analytics objective.

(1)Whether there are strong relationship between actual rating and income

(2)To build the model using the expected rating to predict actual rating.

* + Is this a supervised or unsupervised task? Is it predictive or descriptive? Is it retrospective or forward-looking?

It is a supervised task, predictive and forward-looking.

* + What is the main outcome variable(s) of interest?

Actual rating is our outcome!

* **Data**
  + Brief description of the available data.
* Movie Title: The title of movie
* Release Date
* Type of Movie
* Length of Movie
* Issuing Company
* Expected Ratings (before release)
* Actual received ratings (after release)
* Comment ratings (after release)
* Comment Date
* Number of Trailer
* Actors
* Directors

Weekly Income

* + Some guidance on the data subset that will be used and the re-processing or preparation that might be needed based on your past experience.

(1)Categorical turn to numerical

(2)Handle missing data

* + Sample of ten rows (records) with ten columns (variables) that will be used.
* **Methods**
  + What are some data mining methods to consider?

(1) Visualization of actual rating and income

(2)Linear regression

(3)KNN

* + Which performance measures are appropriate? How do they map to the business goal?

To compare the error between predicted rating and actual rating.

Movie theaters will get more accurately predicted rating so they can get higher income by doing better preparation.

* **Implementation/Production**
  + Operational requirements and/or constraints (for example, will the solution be run in real-time? will it require collecting new data? will it be a one-time analysis or ongoing?)

We need to collect new data and it is a one-time analysis.