STA 523 Final Project Proposal

Group members:

Mengrui Li, Xin Liu, Wenli Shi, Hanyu Song, Azeem Zaman

Title: Estimation of Opening Weekend Box Office

Abstract:

We plan to estimate the domestic box office of opening weekend (or the first month) of a film's theatrical run using a multiple linear regression model. Proposed explanatory variables include:

1. User reviews: quantitative, 0 - 10
2. Critic reviews: quantitative, 0 -10
3. Director: categorical, 1: famous directors, 0: otherwise
4. Popular actor/actress: categorical,
   1: if including actor/actress in the top 100 "Most Popular Females/Males" has a role in the film (Source:http://www.imdb.com/search/name?gender=male,female&ref_=nv_tp_cel_1), 0: otherwise
5. Production budget: quantitative (Inflation will be taken into account)
6. Country of origin: categorical, 1: Domestic Movies, 0: Otherwise
7. Number of theaters in which film was shown during opening weekend: quantitative
8. Genres: categorical (one variable for each genre)

Procedures:

1. Data Collection and Preparation
   a. Relevant information will be extracted from http://www.imdb.com and http://www.boxofficemojo.com/alltime/weekends/ using the web scraping techniques covered in class.
   b. Explanatory variables will be standardized to balance variable contribution to data variability.

2. Data Analysis
   a. A multiple regression model will be fitted to have a general idea of goodness of fit and collinearity among explanatory variables. Interactions terms or higher order terms will also be added to explore the model structures. Possible data transformations will be considered based on histograms.
   b. A two-stage LASSO will be performed for variable selection and model fitting. If time permits, Bayesian Model Averaging will also be adopted for model selection.