***Introduction***

*By Yin Bian*

This paper is to describe our contributions to the final project of analyzing the profitability of movies based on various factors which was conducted jointly by Wenjun Zeng, Guanghong Fu, Rongzhao Huang and Yin Bian from June 30 to August 9. The paper highlights the bottom up approach to determine the most important driving factors, some basic statistical tests to understand the mechanism underlying the effect of some variables like the popularity of actors, the art of directors and the net budget from the beginning, on the profitability of movies.

1. **Problem identification**

More and more people are willing to spend their free time watching a movie outside with their families or friends. However, consumers have different preference based on their incomes, geographies and their own favor for certain celebrities. To exploits this preference sensitivity, =investors of those movies wish to know what factors make their product widely accepted; To achieve an appropriate estimate of revenue for a new movie, analysts wish to have a clear map of customer specific factors (favorite stars, spending behavior) and other factors like budgets as inputs on which they can build price predicting software products. Therefore, in this project, we attempt to identify how the gross revenue of a movie is influenced by some crucial factors like the date of release, the popularity of the leading actor and actress, the art of the director and the budgets the investors wish to put at the beginning.

Our data source, IMDb , is a website that can incorporate almost all the potential insights that we want to take into considerations. In this project, we only focus on a specific task of scraping, wrangling and analyzing the data form IMDb to draw a tentative conclusion about how we can predict the popularity of a movie based on those potential parameters although the ultimate broader goal is to possibly develop stress testing models for movies on the existing economic scenario. These stress test models are crucial as they enable both investors and consumers to answer the question like-What would happen to movies if the director invite James Cameron to be a leading actor in their new movie? And design their operations accordingly.

1. **Workflow and task breakdown**

To efficiently tackle profitability forecasting, the work is partitioned into three subparts: preliminary understanding and literature survey-to understand previous work done in this field and to check if anything can be reapplied or some useful insights be drawn, Variable Modelling and Analysis – to narrow down the large number of indicators to a set of five to ten most important ones , data management- to import , clean and pre-process huge amount of available data to bring data to usable form, statistical testing – to run some basic statistical analysis like unsupervised learning to solidify and concretize insights from Literature review.

1. **Methodology**

In order to analyze the popularity of movies in the US, we mainly tap into the very large data published by US Website. The internet movie database(IMDb) is an online database of information related to films, television programs and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia and review, which deliver the world’s best film data in terms of both abundance and quality. Those data are being carefully measured, computed and revised on a predefined way. For example, we scape the gross data online including the consumers review for each film after release.

Chronologically, this project had an initial focus on the public data, we have built a basic model analyzing the rating of a movie in the US on a macro level. What is much more accurate and actionable, is a model based on more influential factors. Our following work has been therefore based on data from other source, “on top” of this micro-level modeling, we leverage our macro-economic models to gain new insights and improve the accuracy of our models.

1. **Tools**

Different tasks involve specific tools that are well adapted. The software R has been our primary choice in this project, given its focus on data wrangling and cleaning. In terms of capability when it comes to very large dataset. we experienced some of its limitations, so we combine SQL to help us overcome this difficulty.