Appendix

Table 1: Variable descriptions of the original data set

Variable Name	Data Type	Description
Budget	Integer	The cost of creating the moving in dollars
Genres	List	The different genres that a movie belongs to
Homepage	String	The official movie website
Movie_id	Integer	Unique ID for the movie
Original_language	String	Original language that the movie was released in
Original_title	String	Original title of the movie
Overiew	String	The general premise of the movie
Popularity	Float	Popularity of the movie is based on viewer ratings
Production_companies	List	The production companies that worked on the movie
Production_countries	List	The countries where the movie was filmed
Release_date	Date	The date the movie was released
Runtime	Float	Length of the movie in minutes
Spoken_languages	List	The different languages that is available for the movie
Status	String	Release status of the movie
Tagline	String	The tagline used in the advertising campaign for the movie
Title	String	The final title of the movie on release
Keywords	List	Descriptive words used for the movie
Cast	List	The cast members for the movie
Crew	List	The crew members for the movie
Revenue	Integer	The total revenue accrued by the movie
Vote_average	Float	The average rating of the movie based on viewer votes
Vote_count	Float	The total number of ratings a movie was given

Table 2: Features engineered

Feature Name	**Description**
Release_dayofweek	The day of the week that the movie was released. Taken from release date
Release_quarter	The annual quarter that the movie was released. Taken from release date
inflationBudget	the budget of movie after considering inflation factors
Num_keywords	number of keywords of the movie
Original_title_word_count	number of words in the movie's original title (before translated into English)
Title_word_count	number of words in the movie's title
Overview_word_count	number of words in the overview
Tagline_word_count	number of words in tagline
Cast_count	number of casts
Crew_count	number of production crew
Production_countries_count	how many countries have been involved in the production of the movie
Production_companies_count	how many production companies have collaborated in making the movie
Has_homepage	whether the movie has a homepage
Has_tagline	whether the movie has tagline
isOriginalLanguageEng	whether the original language of this movie is English

Table 3: Performance of regression models

	MLR	**Random Forest**	**XGBoosting**
Test MSLE before param tuning	2.23	2.79	2.34
Test MSLE after param tuning	-	2.15	2.25
5-fold C.V. before param tuning	2.06	2.61	2.01
5-fold C.V. after param tuning	-	2.04	1.79