HOLLYWOOD MOVIE DATASET

DATA DESCRIPTION

The data set was obtained from several movie databases using both automated as well as manual means. It is more than likely that some of the values are captured/entered incorrectly. Hence, the accuracy of the data set cannot be guaranteed and provided to you as is, for educational purposes.

In the context of predictive analytics, the dependent variable can be the box-office gross revenues (i.e., *GrossBoxOffice*) for forecasting/regression type prediction models. For a classification type prediction modeling exercises, the dependent variable can be the success category (i.e., *BoxOfficeClass*). In the representation of *BoxOfficeClass* variable, a movie is assigned to one of nine success categories based on its gross box-office receipts, ranging from "1 - Flop" to "9 - Blockbuster." The following table shows the breakpoints/bins used to convert the gross box-office revenues to 1 of 9 success categories:

Class No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Range (in Millions)	**< 1 (Flop)**	**> 1 < 10**	**> 10 < 20**	**> 20 < 40**	**> 40 < 65**	**> 65 < 100**	**> 100 < 150**	**> 150 < 200**	**> 200 (Blockbuster)**

This data set can be used for descriptive as well as predictive modeling purposes. It also has a textual field (i.e., *ShortStoryLine*), which can be used for text mining (i.e., Word Cloud in SAS VA) exercises. A brief data dictionary is provided in the next page.

Here are a few refernces/publications that used some version of this dataset:

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast guru: a web-based DSS for Hollywood managers. *Decision Support Systems*, 43(4), 1151-1170.

Sharda, R., & Delen, D. (2006). How to Predict a Movie's Success at the Box Office. *Foresight: The International Journal of Applied Forecasting*, (5), 32-36.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243-254.

DATA DICTIONARY

Field Name	Short Definition	Possible Values
MovieID	A Unique Identifier for the movie.	An integer number
YEAR	The year movie is shown in theatres	Year 2000 to 2010
StarValue **(… -Director; -Producer, -Cast)**	Signifies the star value of the director, producer and the cast (as per the recent past box-office success).	An ordinal category from 1 to 5 (1: lowest, 5: highest)
OriginalScreenPlay	The movie is based on an original screen play	Yes/No
Genre	Specifies the content category the movie belongs to. A movie can be classified in more than one content category at the same time (e.g., action as well as comedy). Therefore, each content category is represented with a separate binary variable.	Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Mystery, etc.
Competition	Indicates the level at which each movie competes for the same pool of entertainment dollars against movies released at the same time.	High, Medium, Low
MPAA Rating	The rating assigned by the Motion Picture Association of America (MPAA).	G, PG, PG13, R, NR
Number of screens	Indicates the number of screens the movie is expected to be shown at its debut.	An integer number
BoxOfficeClass	Box-office success category	An integer from 1 to 9
GrossBoxOffice	Box-office gross revenue on theatres	An integer number
ShortStoryLine	A short textual description of the script/story	A few sentences
EstimatedBudget	Estimated movie budget – this field has values only for a subset of the movies	An integer number
MovieLength	The number of minutes the movie runs – this field has values only for a subset of the movies	An integer number