SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA AND DATA ANALYTICS
LAB PROJECT 5

This lab project is based on a dataset about movie success in 2014 and 2015 by Ahmad et al. (2015) which is available on the online platform by Lichman et al (2013). The dataset file movidata.csv can be downloaded from Blackboard.

<u>Important:</u> Before you start with the exercises, run the following code to <u>remove</u> all observations from the dataset where budget, screens, or aggregate followers have NA values and create a few more variables.

```
moviedata <- subset(moviedata, !is.na(budget) & !is.na(screens)
   & !is.na(aggregate_followers))
moviedata$profit <- moviedata$gross - moviedata$budget
moviedata$netlikes = moviedata$likes - moviedata$dislikes
```

EXERCISE 1 (2 MARKS) *[R-CODE]*

Use R to create a variable called “posrating” in the dataframe. The variable takes on the value 1 if ratings >= 6.8. For ratings < 6.8 it takes on the value 0. Use R to perform a logistic regression that regresses the newly created variable “posrating” on aggregate_followers, dummy_sequel, netlikes, and sentiment in order to predict the probability that a movie has a positive rating. Interpret the coefficients and report the results of the logistic regression in APA style (including a logistic regression table and reporting of AIC-values).

EXERCISE 2 (1 MARK) *[R-CODE]*

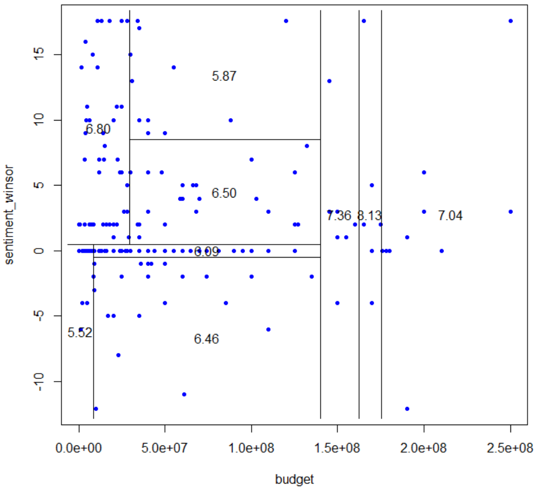
Use R to create a confusion matrix. Report the confusion matrix. Calculate and interpret sensitivity, specificity, and accuracy.

EXERCISE 3 (2 MARKS) *[R-CODE]*

Use R to create two logistic regression models. In the first model, regress posrating on sentiment. In the second model, regress posrating on budget, dummy_sequel, and sentiment. Plot an “Area under the ROC Curve” plot for each of the two models and explain the plots. Use the AUC values to compare the two models and interpret the results.

Exercise 4 (2 marks) ***[R-Code]***

Use the winsor function discussed in Week 3 to create a variable “sentiment_winsor” with a multiplier of 2.2. Use R to create a regression tree that uses budget and sentiment to predict ratings. Then, create a scatterplot of budget and sentiment, and add the partitions of the regression tree. Interpret at least 2 partions of the partitioned scatterplot.



Exercise 5 (1 mark) ***[R-Code]***

Based on the regression tree created in Exercise 4, use cross-validation to determine the optimal tree size and prune the tree. Plot and interpret the tree.

Exercise 6 (2 marks) ***[R-Code]***

Use R to create a classification tree to predict posrating. As predictors, take into account the variables aggregate_followers, comments, likes, dislikes, and sentiment. Use cross-validation to determine the optimal tree size and prune the tree. Plot and interpret the tree.

References

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. In Smart City/ SocialCom/S ustainCom (SmartCity), 2015 IEEE International Conference on 2015 Dec 19 (pp. 273-278). IEEE. https://ieeexplore.ieee.org/document/7463737

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

DATASET

moviedata *Conventional and Social Media Movies 2014 and 2015*

Description

A dataset about the success of movies in 2014 and 2015.

Usage

moviedata

Format

A data frame with 231 observations on the following 14 variables.

`movie`	Name of the movie
`year`	Year of movie release
`ratings`	Rating of the movie (0 – 10)
`genre`	Identifier for the genre of the movie (e.g., action, adventure, drama)
`gross`	Gross world-wide income from the movie (in US$)
`budget`	Budget for the movie
`screens`	Number of screens that the movie was initially launched in on the opening weekend in the US
`sequel`	A number indicating whether the movie is sequel or original (individual) movie, where higher numbers indicate later sequels in a series. For instance, for Mission Impossible a sequel value of 5 indicates that this is the fifth movie in the series.
`dummy_sequel`	0 – Original movie 1 – Sequel movie
`sentiment`	A sentiment score assessed through an analysis of tweets about the movie on Twitter. 0 represents a neutral sentiment, a positive value represents a positive sentiment, and a negative value indicates a negative sentiment. The sentiment score for each movie was calculated by retrieving all tweets related to each movie, assigning the sentiment score to each of them and then aggregating the score.
`views`	Number of times the movie trailer was viewed on YouTube
`likes`	Number of likes the movie trailer received on YouTube
`dislikes`	Number of dislikes the movie trailer received on YouTube
`comments`	Number of times the movie trailer received a comment on YouTube
`aggregate_followers`	The aggregate number of actor followers: Equal to sum of followers of top 3 cast from Twitter

Source

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. In Smart City/ SocialCom/S ustainCom (SmartCity), 2015 IEEE International Conference on 2015 Dec 19 (pp. 273-278). IEEE. https://ieeexplore.ieee.org/document/7463737

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.