**Movie Review Rating Predictor**

**Authors:** [Yuanzhe Marco Ma](https://canvas.ubc.ca/groups/259737/users/41176), Arash Shamseddini, [Kaicheng Tan](https://canvas.ubc.ca/groups/259737/users/502612), [Zhenrui Yu](https://canvas.ubc.ca/groups/259737/users/414443) (alphabetic order)

**Date:** Nov.-Dec. 2020

**Objectives**

In this project we tend to build a Regression model that uses the Support Vector Machines (SVMs) with RBF kernel algorithm to predict the *rating* of movies based on the *review* made. For this aim, we will be using a dataset called “Dataset\_IMDB” from SentimentDictionaries library compiled by Nicolas Pröllochs and Stefan Feuerriegel Copyright © under the MIT License. Here is a direct link to the dataset which is sourced open in github.com.

The main objective of this study is to answer the following predictive question using our model:

***“Given a review of a specific movie from one of the critics of the list below, what is the critic’s rating of the movie based on a number between 0 and 1?”***

|  |
| --- |
| **Critics** |
| Dennis Schwartz |
| James Berardinelli |
| Scott Renshaw |
| STEVE RHODES |

**Dataset**

The “Dataset\_IMDB” is a mid-sized dataset of 5006 movie reviews from the four critics. It has four columns as shown in below table.

Col. Table

**Analysis**

For the purpose of this project, we will use SVM with RBF kernel algorithm for our model. We have split the dataset into 80-20 % proportion for the training and test splits respectively. We have built our model using the training portion to come up with the best predictor. Next, we have examined the model using the test split to evaluate its performance. The results of the analysis is in this report.

**Exploratory Data Analysis**

The Exploratory Data Analysis (EDA) for this project can be found in EDA.md file which contains statistical information of the dataset presented in both tables and plots.

EDA.md 🡪 A histogram of rating distribution

EDA.md 🡪 The highest and lowest rated reviews

**Reproduction**

Refer to the Usage section Tiffany has provided. Should we just copy/paste it?!

**Dependencies**

Refer to the Usage section Tiffany has provided. Should we just copy/paste it?!

* Python 3.7.4 and Python packages:
  + docopt=0.6.2
  + requests=2.22.0
  + pandas=0.25.1R
  + feather-format=0.4.0
* R version 3.6.1 and R packages:
  + knitr=1.26
  + feather=0.3.5
  + tidyverse=1.3.0
  + caret=6.0-85
  + ggridges=0.5.2
  + ggthemes=4.2.0
* GNU make 4.2.1

**References**

Varada Kolhatkar. 2020, “MDS Machine Learning Repository.” University of British Columbia, Vancouver, <https://github.ubc.ca/MDS-2020-21/DSCI_571_sup-learn-1_students>

Dua, Dheeru, and Casey Graff. 2017. “UCI Machine Learning Repository.” University of California, Irvine, School of Information; Computer Sciences. <http://archive.ics.uci.edu/ml>

Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani, 2009, **“**[An Introduction to Statistical Learning](http://statlearning.com/)with Application in R”,Springer Publishing