Project 3: Movie Review Sentiment Analysis

**Fall 2022 STAT 542**

Authored By:

|  |  |
| --- | --- |
| Mohammad Hassanpour | MH57 |
| Jason Young | JasonY6 |
| Unnati Narang | Unnati |

# Abstract

Lasso logistic regression and data pre-processing were used to predict sentiment on five folds of test data resulting in the accuracy on test data in Table I.



Table I reports the overall accuracy and run time of the model on test data in each of five folds.

# Introduction

The objective was to use a classification model to predict sentiment of movie reviews on five folds of test data. The first step was to create a vocabulary from training data to use for all folds. Then a model was built from training data to predict sentiment from test data reviews.

Predictions for test data are stored in a two-column data frame containing a test id in the first column and predicted probabilities in the second column.

Model performance is evaluated using Area Under the Curve (AUC). This refers to the area under the ROC curve. Receiver-operating characteristic (ROC) analysis was originally developed during World War II to analyze classification accuracy in differentiating signal from noise in radar detection.1 The curve simply plots true positive rate against false positive rate. A large AUC is desirable.

# Methods

**Build Vocabulary**

Prior to model building and prediction, a suitable set of terms is obtained from training data to use for classification.

Please see buildVocab.html for details on the process to create the vocabulary.

**Data Pre-processing**

The processing of training and test data is identical.

Reviews are loaded, periods followed by any number of spaces are replaced with a single space, certain words are removed, and a Document Term Matrix (DTM) is generated. A DTM is a matrix with a document (in this case, review) represented in each row and a term represented in each column with counts in each entry.

**Model Training**

Once training data is processed to produce a DTM a lasso logistic regression model is trained on this data using the glmnet function in R to produce a glmnet object.

Finally, sentiment is predicted from a DTM based on test data using the smallest lambda stored in the glmnet object.

# Results



Table I reports the overall accuracy and run time of the model on test data in each of five folds.

**The computer system used for prediction:**

MacBook Pro (13-inch, 2020, Four Thunderbolt 3 ports)

Processor: 2.3 GHz Quad-Core Intel Core i7

Memory 16 GB 3733 MHz LPDDR4X

# Discussion

R packages had several tools to help with sentiment analysis. It was important to learn to clean html tags and to use R package text2vec to construct DTMs of limited size.

# References

Liang, Feng. “Project 3: What we have tried (I).”, *Campuswire.com,* 9 November 2022, <https://campuswire.com/c/G3D46BBBA/feed/1183>*.*

1Lusted LB. Signal detectability and medical decision making. **Science***.*1971; *171*: 1217–1219.