**Crime Lab New York – Machine Learning Applied Task**

Yuqi Liao

May 8, 2019

# Summary/Instruction

This document provides a summary of machine learning applied task and answers the predefined questions. To access the code and the accompanying written analysis (that explains the code and documents rationales of certain analysis choices), please see the RNotebook file in [HTML](https://github.com/yuqiliao/CLNY/blob/master/RNoteBook.html) and the [RMD](https://github.com/yuqiliao/CLNY/blob/master/RNoteBook.Rmd) formats.

# Task 1 – build and evaluate a “freshness” predictor

I built an elastic net and a naive Bayes model to predict the outcome (“freshness”). The RNotebook file documents more details of the rationale of choosing these two models. In short, because (1) they tend to be faster (than the other models that I considered), (2) the logistical regression (and by extension, the elastic net model) is a foundational classification model, and (3) naive Bayes model is typically used in the context of text analysis modeling.

Accuracy is used as the main model evaluation metric because in this context, I believe the two levels of the outcome (“fresh” or “rotten”) are equally important. Also, in the final clean dataset, the cases of “fresh” and “rotten” comments (rows) are not unbalanced.

I then focused on investigating the result of the elastic net model further by looking at the variable importance of each input features. Among the top 20 variables with most predictive values, names of the producer/director/actor/actress (e.g. “lee”, “cage”, “ralph”, “william”), and words that indicate sentiment (e.g. “fortun” [for “fortunate” or “fortunately”], “exhilar” [for “exhilarated”] play important roles in predicting “freshness”.

# Task 2 – interrogate your predictor

To explore whether the performance of the model difference across subpopulations in the data, I focused on applying the trained elastic net model to a subset of review comments by each genre.

The results show that the accuracy metric varies across movie genre to be as low as 0.2857143 in “Television” and as high as 0.8333333 in “Gay and Lesbian”. Such results are most likely due to the fact that there are not too many review comments about movies of “Television” or “Gay and Lesbian” genre (with 84 and 12 observations in the test dataset, respectively) to begin with, therefore the model’s predictions on those rows would have a large variance. This is a reminder about the importance of looking at how the model affects different subpopulation differently and the importance of having each subpopulation adequately represented in the dataset (if analyzing subpopulation is something that the research cares about).

# Task 3 – summarize results

**Q1: Compare the performance between both models you created in task 1 – is there a significant difference between their performance?**

A1: The accuracy metric is 0.6297 for the elastic net model, and 0.608 for the naive Bayes model. For a benchmark model that predicts “fresh” for all observations in the test data, its accuracy would be 14923/(14923+9485) = 0.611 (not shown in the RNotebook file). Therefore, in comparison, the two selected models are not performing very well, and the difference between the two models is not big.

**Q2: Are features from the movie file (movies.tsv) useful in these predictions? Why or why not?**

A2: Because the features from movies.tsv are joined with the review level data (so each row is a review, and many reviews of the same movie will have the same movie-level information), the variation of the features from movies.tsv tends not to be very high, thus it is less likely that these features will be highly predictive of the outcome variable. Given that, however, if there’s enough variation, it will show up in the variable importance ranking.

The features used from movies.tsv are the categorical “rating” variable (e.g. G, PG, PG-13, NR, R, etc.) and the continuous movie “runtime” variable (in the unit of minutes). The result shows that (not reflected in the RNotebook), only the “ratingNR” and “ratingPG-13” (they are the dummy versions of the “rating” variable) made it as one of the top 30 most predictive variables in predicting the outcome.

**Q3: If you had additional time (again, please do not spend more than 5-6 hours on this task), what would your next steps be as you iterate on your predictor?**

A3: Given more time, the following steps should be considered and implemented to improve the performance of the model and to understand better of the implication of the model.

* Get more context of the data (e.g. be more familiar with movie genre. movie technical terms, etc.) in other words, it is important to connect the result back to the data with content expertise.
* Try out bi-gram and n-gram options
* Conduct topic modeling to generate new features for model building
* Normalize all variables before the model
* Try out more models (e.g. adaboost, random forest, etc)