Disaster Relief Project

SYS 6018: Data Mining

September 1, 2020

1 Introduction

In this project, you will use classification methods covered in this course to solve a real historical data-mining problem: locating displaced persons living in makeshift shelters following the destruction of the earthquake in Haiti in 2010.

Following that earthquake, rescue workers, mostly from the United States military, needed to get food and water to the displaced persons. But with destroyed communications, impassable roads, and thousands of square miles, actually locating the people who needed help was challenging.

As part of the rescue effort, a team from the Rochester Institute of Technology were flying an aircraft to collect high resolution geo-referenced imagery. It was known that the people whose homes had been destroyed by the earthquake were creating temporary shelters using blue tarps, and these blue tarps would be good indicators of where the displaced persons were – if only they could be located in time, out of the thousands of images that would be collected every day. The problem was that there was no way for aid workers to search the thousands of images in time to find the tarps and communicate the locations back to the rescue workers on the ground in time. The solution would be provided by data-mining algorithms, which could search the images far faster and more thoroughly (and accurately?) then humanly possible. The goal was to find an algorithm that could effectively search the images in order to locate displaced persons and communicate those locations rescue workers so they could help those who needed it in time.

This disaster relief project is the subject matter for your project in this course, which you will submit in two parts. You will use data from the actual data collection process was carried out over Haiti. Your goal is to test each of the algorithms you learn in this course on the imagery data collected during the relief efforts made Haiti in response to the 2010 earthquake, and determine which method you will use to as accurately as possible, and in as timely a manner as possible, locate as many of the displaced persons identified in the imagery data so that they can be provided food and water before their situations become unsurvivable.

You will document the accuracy and runtime for classification for each algorithm in the `.Rmd`-based presentation files that you create (as variants of the template that is provided for you). In Module 6, you will submit your `.Rmd` with some of the results of your analyses filled out. In Module 12, at the end of the course, you will submit your completed `.Rmd` file, which will include the conclusions of your work as well as your recommendations regarding algorithm preferability for this application.

2 Collaboration, teamwork, and help

1. This is *not* a group project; however...

2. You *are* welcome to discuss what you're doing with your classmates
3. The intent of any peer-to-peer conversations about this project may be "supportive" in nature
4. But on a spectrum from "empowering" (in the positive sense of this world) to "enabling" (in the negative sense of this word), your exchanges must stay *well* to the "empowering" end of the equation

This is a "give a friend a fish, feed them for a day; teach a friend to fish, feed them for a lifetime" kind of situation. Do not feel obligated to ensure the success of your peers in this project – that is their responsibility – but help them succeed when you are able. This is not a zero-sum game and your personal expertise and knowledge will benefit tremendously from opportunities to make this sort of contribution to your fellow man and woman.

3 Project Part 1 (60 points) – *DUE MODULE 6*

The `.Rmd` file you submit will address and provide the entries in Table 1. If it is not clear from your `.Rmd` file how the values in your table were produced you will not receive credit for the entry.

3.1 K-Folds Out of Sampling Performance (30 points)

There are 33 blank spaces in Table 1, each of which contributes 1 point – to a maximum of 30 points – to your score. To acquire points for the **ROC** column, you must provide the actual curve in a plot. To acquire points for the **threshold** column, you must both (a) provide a threshold, and (b) explain and justify your rationale for the threshold you chose. Acquiring points for all rows below the **threshold** row will of course require a **threshold** value for the column in question. All entries are numeric, with the aforementioned exceptions for ROC curves and part (b) – the justification and explanation of your chosen values – for the **threshold** row.

Method	KNN ($k =$)	LDA	QDA	Logistic Regression
Accuracy				
AUC				
ROC				
Threshold				
Sensitivity=Recall=Power				
Specificity=1-FPR				
FDR				
Precision=PPV				

Table 1: Performance Metrics: **10-Fold** Cross-Validation Metrics

3.2 Conclusions (30 points)

Report on at least THREE conclusions.

1. One must be your determination of which algorithm works best (10 points).

Additional conclusions should be observations you've made based on your work on this project, such as:

- Are there additional recommend actions that might be taken to improve results?
- Were there multiple adequately performing methods, or just one clear best method?

- What is it about this data formulation that allows us to address it with predictive modeling tools?
- How effective do you think your work here could actually be in terms of helping to save human life?
- Do these data seem particularly well-suited to one class of prediction methods, and if so, why?

Your best two additional conclusions will be graded and will each be worth 10 points.

4 Project Part 2 (132 points) – *DUE MODULE 12*

The `.Rmd` file you submit will address and provide the entries in Table 2. If it is not clear from your `.Rmd` file how the values in your table were produced you will not receive credit for the entry.

4.1 K-Folds Out of Sampling Performance (36 points)

Table 2 is filled out the in same manner as in Table 1, but the new added columns are worth more points than the original columns. There are 18 blank spaces in the rightmost two columns of Table 2 – each of these spaces is worth 1 point for a total of 18 points. The remaining (33) blank spaces in columns shared with Table 1 – which you should have already completed in part 1 of this project – are each worth $\frac{1}{4}$ point up to a maximum of 8 points. The final 10 points will be given for providing an (a) appropriate interpretation of your choses tuning parameter values, and (b) explanation of how they were chosen.

Method	KNN ($k =$)	LDA	QDA	Logistic Regression	(tuning parameters =) Random Forest	SVM
Accuracy				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
AUC				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
ROC				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
Threshold				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
Sensitivity=Recall=Power				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
Specificity=1-FPR				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
FDR				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	
Precision=PPV				$\longleftarrow \frac{1}{2}$	$1 \longrightarrow$	

Table 2: Performance Metrics: **10-Fold** Cross-Validation Metrics

4.2 Hold-Out Test Sample Performance (36 points)

Table 3 is filled out in the same manner as in Table 1. If it is not clear from your `.Rmd` file how the values in your table were produced you will not receive credit for the entry.

Each of the six columns is worth 6 points, with 1 point lost for each entry *not* provided in the table. *E.g., if you have all the row values for the SVM column, but fail to note the values of the tuning parameters you used, you will get only* 5 *of the* 6 *available points.*

Method	KNN ($k =$)	LDA	QDA	Logistic Regression	(tuning parameters =) Random Forest	SVM
Accuracy						
AUC						
ROC						
Threshold						
Sensitivity=Recall=Power						
Specificity=1-FPR						
FDR						
Precision=PPV						

Table 3: Performance Metrics: **Hold-Out** Test Data Set Scores

4.3 Conclusions (60 points)

Report on at least SIX conclusions.

Four of your conclusions must be the following (each worth 10 points):

1. A discussion of the best performing algorithm(s) in the cross-validation and hold-out data
2. A discussion or analysis justifying why your findings above are compatible or reconcilable
3. A recommendation and rationale regarding which algorithm to use for detection of blue tarps
4. A discussion of the relevance of the metrics calculated in the tables to this application context

Your best two additional conclusions will each be worth 10 points.