Disaster Relief Project: Part 1

DS-6030

Contents

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 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.

1 Project scope

In the project, you will build and evaluate several models using cross-validation and estimate their performance after threshold selection using a variety of performance metrics for the cross-validation results as well as for predictions on the hold-out set.

This course covers several model types:

- Models without tuning parameters
  - Logistic Regression
  - LDA (Linear Discriminant Analysis)
  - QDA (Quadratic Discriminant Analysis)
- Models with tuning parameters
  - KNN (K-nearest neighbor)
  - Penalized Logistic Regression (elastic net penalty)
  - Ensemble method: random forest (ranger) or boosting (XGBoost)
  - Support Vector Machines (SVM)
    - linear kernel
    - polynomial kernel
    - radial basis function kernel

Choose at least five of the above models to build, evaluate and compare. You will use the training data to build the models and the hold-out set to evaluate the performance of the models. The goal is to find a model that can accurately predict the location of displaced persons in the imagery data, and to do so in a timely manner.

This scope is vague on purpose. You will need to make decisions on how to preprocess the data, which models to use, how to evaluate the models, and how to select the final model. The project is designed to give you the freedom to explore different approaches and find the best solution for the problem.

2 Submission Format

The project is divided into three parts

1. Project plan (2-3 pages document, Rmd code file - Module 5): following EDA, outline the approach that will be taken to build and validate the model
2. Progress report (5-10 pages document, Rmd code file - Module 8): summarize preliminary results and define next steps
3. Presentation (video, slides and Rmd file - Module 11): present your results in a 15 minute video presentation

2.1 Project plan (2-3 pages document, Rmd code file - Module 5)

Following exploration and analysis of the data (training and holdout sets), summarize your key findings and outline the approach that will be taken to build and validate the model.

The project plan must cover the following topics:

- How did you preprocess the data and why did you choose this approach?
- What did you observe in the exploratory data analysis (EDA) of the training and holdout sets?

- How will your observations impact the model training and validation process?
- Which models do you plan to explore and why do you choose these models?
- How do you plan to train and validate the models?
  - you will need to train at least five models
  - the selection must include tuneable models
- How will you select the threshold for the classification models and why do you choose this approach?
- How will you evaluate the model performance and why do you choose this approach?
- How do you plan to select the final model and what is your rationale for this approach?

The project plan is an important component of the project, as it will guide your future work on it. It should be concise yet comprehensive, providing a clear roadmap for your project. However, it is not a final document, and you can adjust your approach as you progress through the project.

You will submit **two** deliverables:

1. **PDF document** which contains the results in a report format. You can use Word or any other text processing software to prepare this document. You are completely free in how you organize your project plan. However, the report must contain the minimum requirements listed above. **The PDF must not have more than 4 pages!**
2. **Rmarkdown (.Rmd)** file which contains your code. You will use *Tidyverse/ Tidymodels* to process your data and build models.

We will look at both documents.

2.2 Progress report (8-10 pages document, Rmd code file - Module 8)

By now you should have implemented a **substantial** part of the project. The progress report should summarize your preliminary results and define next steps.

The progress report must cover the following topics:

- What has changed from your initial project plan and why did you decide to make changes?
- What models have you implemented so far and what are your preliminary results?
  - Do you observe any issues with the models?
  - Are there differences between the models that are worth noting?
  - Which model do you favor so far and why?
- Considering the results so far, what are your next steps?
  - What models, if any, do you plan to implement next?
  - How will you address any issues you observed with the models?
  - How will you select the final model and what is your rationale for this approach?

You will submit **two** deliverables:

1. **PDF document** which contains the results in a report format. You can use Word or any other text processing software to prepare this document. You are completely free in how you organize your project plan. However, the report must contain the minimum requirements listed above. **The PDF must not have more than 12 pages!**
2. **Rmarkdown (.Rmd)** file which contains your code. You will use *Tidyverse/ Tidymodels* to process your data and build models.

We will look at both documents.

2.3 Presentation (video, slides and Rmd file - Module 11)

Prepare a video presentation of your project. The presentation should be about 15 minutes long and cover the following topics:

- Introduction
- Data
- Description of Methodology
- Results
- Conclusions

All graphs are suitably sized and clearly readable; axis labels and legends descriptive; captions describe the content of the graph.

You will submit **three** deliverables:

1. Recording of a **Presentation**. The presentation should be about 15 minutes.
2. **Presentation slides** (submit as PDF)
3. **Rmarkdown (.Rmd)** file which contains the code. You will use *Tidyverse/ Tidymodels* to process your data and build models.

We will look at all documents.

3 Coding

All code is well organized, and executes without errors. The R code should be easy to follow.

We will look at the code and include it in the grading. We check that the report and presentation reflects what the code does.

4 Collaboration and Help

- You are not permitted to copy code. You will no doubt come across examples on the internet. You can consult them to help understand the concept or process, but *code in your own words.*
- It is a scholarly responsibility to attribute all your work. This includes figures, code, ideas, etc. Think of it this way: will someone who reads your submission think that it is your original idea, figure, code, etc? Add a link and/or reference to all sources you used to solve a problem. It is really of no value to you when you just copy someone else's solutions (other then preserve a grade that you didn't earn).
- If you use generative AI, list it as a reference and describe what you used it for.

It is not always easy to tell what qualifies as an honor code violation, so do not be afraid to talk to me about it. Such discussions do not imply guilt of any kind.