**CPLN 675 – Land Use And Environmental Modeling**

**HW3 – Forecasting flood inundation**

**Spring, 2022**

**Instructor: Michael Fichman**

**Deliverables and Due Date**: ~3 page report (in pdf form) *or* knitted R Markdown document (in html form) due in the dropbox linked below by the beginning of class on March 30th.

[Homework dropbox](https://upenn.app.box.com/f/8e484163b88f424781b004637fad31af)

The purpose of this project is to estimate a predictive model that yields predictions in space interpreted as ‘the probability that an area will be inundated with flood water”. You will train and validate your model on Calgary and then predict for a **comparable city**.

You can work in teams of two. Sign up for a team in the teammate signup sheet below. If you have a person you’d prefer to work with, sign up as a team, otherwise just put your name on the sheet – either as Team Member 1 on a new team, or as Member 2 next to another solo student. If you can’t find a team (we might have an odd number of students), let me know and we will accommodate you in a team of three. If you would prefer to work alone, let me know.

[Teammate Signup Sheet](https://docs.google.com/spreadsheets/d/1Qg07cZj5JNdGa9Er8eEyFkTqI5Qi4vnLIA57iQvbW5Y/edit?usp=sharing)

One team member should gather available features from Calgary, the other from your comparable city. You must stick to glm only. You may wrangle data in either ArcGIS or R, but you will be doing your modeling in R.

**Basic procedure**

1. Gather open data from both Calgary’s open data site and your comparable city’s open data site as well as other internet sources.
2. Using what we’ve learned about feature engineering over the first part of the semester, build as many useful variables to explain flood inundation:
   1. natural, hydrological and built environment features
   2. **You must include at least one feature from the watershed analysis.**
3. Join these features to the vector Fishnet.
   1. *Remember that ‘distance or density to Feature A’ might describe the spatial relationship better than simply, ‘Feature A’.*
4. Move your Fishnet dataset into R and run some logistic regressions with **both a test set and a training set**. Experiment until you find a model with enough statistically significant variables.
5. Run goodness of fit metrics; More advanced groups (e.g. if you have taken MUSA 508) should experiment with **spatial cross-validation.** Visualize your results in chart and map form.

**Deliverables**

**Your markdown or pdf report should contain the following:**

1. The Planning motivation for your algorithm and how you would deploy such an algorithm (borrow from your memo).
2. One page showing four of your more original (as you deem it), yet statistically significant features. Annotate as you see fit. *4 maps on 1 page.* **This must include at least one watershed feature (even if it’s not significant).**
   * Clarify why you chose certain features
   1. Flood Plains/Watershed based on DEM
   2. Sand & Gravel / Soil Type
   3. Non-Permeable Surfaces from Land Use Classification
   4. Steepness from DEM
   5. Distance to Waterbodies
3. One page with your final logistic regression model summary including your ROC curve, confusion metrics and associated goodness of fit/cross-validation. Annotate briefly.
4. One page showing 3 maps – The first shows true positives, true negatives, false negatives and false positives for the **training set in** Calgary. Second, your inundation predictions for Calgary (entire dataset); Third, predictions for your comparable city.
   1. Ability to flood (not on the coast)
   2. Climate (4 seasons, no hurricanes, freezable)

**Your narrated powerpoint youtube video should be as follows:**

No longer than 4 minutes.

Make sure you state up front why planners might be interested in this sort of analysis; how this differs from other approaches; what your goal is; what you modeling strategy is; what features are important; how did you create them; what were your results; how do you defend your predictions; what are the next steps.

**Show lots of interesting data visualizations. You will be graded on the quality of data visualizations and your ability to explain to a non-technical decision maker. *Include the link to this video in your markdown or pdf submission.***

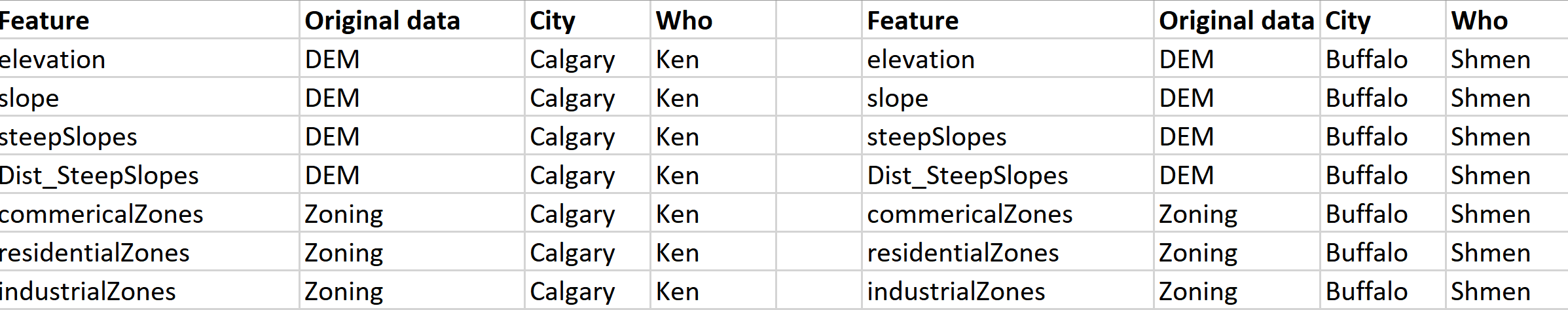
A picture containing text

Description automatically generatedA picture containing text

Description automatically generated

There are three files as illustrated above. There is **the satellite image with flood inundation** in light blue and cloud cover in red. You will have to reclassify into a raster of 0=no inundation and 1=inundation. **The pink polygon is the study area** which you will note has a greater area than the raster. You can

1) create a fishnet at the study area extent in ArcGIS with [this tool](https://desktop.arcgis.com/en/arcmap/10.3/tools/data-management-toolbox/create-fishnet.htm), and in R with [sf::st\_make\_grid.](https://www.rdocumentation.org/packages/sf/versions/0.8-0/topics/st_make_grid) You will then have to 2) create an unique id field in the fishnet shapefile and 3) use zonal statistics as table to get the raster values into the fishnet. Finally 4) remove the grid cells to the north for which we have no raster data from the fishnet. **You also have a DEM raster** as above. For your comparison city, look to their open data site for Digital Elevation Map (DEM). Otherwise, go to [EarthExplorer](https://earthexplorer.usgs.gov/) and download Shuttle Radar Topography Mission data ([SRTM](https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1-arc?qt-science_center_objects=0#qt-science_center_objects)) data.



Recall, you need to pick a comparable city (Miami, Houston and Phoenix are all examples of cities *not* comparable with Calgary). You and your partner need to collect **the same features** (with the same variable names) for both cities. I suggest you split your cities by teammate and create a shared tabled on google docs like below. This 'data matrix' will help you keep track of what data is being collected by whom. Keep in mind that if you create a field called ‘distanceToSteepSlopes’ where ‘steep’ is defined as ‘slopes > *x*’ and your partner in the other city defines it as ‘slopes > *y*’, your resulting predictions will be nonsensical.

Finally**, to be clear, to do this analysis, you will copy the code from the land conservation markdown and create new independent variables for a model that predicts inundation probability**. You can wrangle data in ArcGIS or R, it’s your choice, but your modeling will take place in R. Once you’ve trained your model on Calgary and are satisfied, you can use the ‘predict’ command to predict for your comparable city.