Department of City and Regional Planning — Fall 2019
University of Pennsylvania — Dr. Ken Steif

FINAL PROJECT
(Due Various dates, 2019)

The goal of your final project is to impress your peers and your instructors by utilizing some of the tools you have learned this semester. The assignment is to communicate an engaging public policy use case of predictive modeling by showing off both your analytical skills and your ability to convert those skills into a relevant policy use case.

1. You are required to work in pairs (signup here). Both team members should work on the model, but each member will present a different part of the deliverable. Each pair will receive a grade that comprises both parts of the project.

While the emphasis for past assignments has been on goodness of fit for a single random test set, here I want to see the emphasis on cross-validation. We've spent some time discussing spatially-dependent test sets, but here, when possible, I also want to see time dependent test sets where you train your data on some period of time and test it on another. This is a great measure of generalizability.

Schedule

11/15 – Project introduced
11/22 – Prepare to discuss **your chosen project** in lab.
12/4 - Present app wireframes
12/20 – Final markdown due

1. **Deliverable 1 (Due 11/22):** Be prepared in lab to talk for a few minutes about the questions at the bottom of this document.

2. **Deliverable 2 (Due 12/6):** Team member 1 will be responsible for a 3 minute 'PechaKucha' presentation that 'sells' us on the idea of this fancy new planning app that you've designed to solve an important problem. At this point of the project, you are expected to have some well thought out exploratory analysis, but **not** a robust machine learning model. You should still have a good sense for the purpose of the predictions in the context of the use case. What is the use case? Who is the user? How does the app put the model into the hands of a non-technical decision maker? Who is creating the app? Have you created something that is usable by the client? This is a presentation where the slides are set to change automatically, **every 20 seconds**. **This is a requirement**. The presentation should follow this format:

Slide 1	Problem motivation: What is the business as usual approach to this planning problem?	30 seconds
Slide 2	What is the proposed data driven approach?	20 seconds
Slide 3	What is the data?	20 seconds
Slide 4	Exploratory analysis (spatial process)	20 seconds
Slide 5	Exploratory analysis	20 seconds
Slide 6	How does the model work in theory – features, spatial process etc	20 seconds
Slide 7	App wireframe 1	20 seconds
Slide 8	App wireframe 2	20 seconds
Slide 9	App wireframe 3 and conclude	30 seconds
		3 minutes and 10 seconds

3. **Deliverable 3 (12/20):**
Team member 1 will have the pechakucha uploaded on youtube with a recorded narration.

Team member 2 will be responsible for a **markdown** write up that would allow someone to replicate your analysis. If possible, post this markdown on your Github. At minimum, please hit on the below components:

a. Motivate the analysis – "What is the use case; why would someone want to replicate your analysis and why would they use this approach?"
b. Describe the data you used.
c. Describe your exploratory approach using maps and plots.
d. What is the spatial or space/time process?
d. Describe your modeling approach and show how you arrived at your final model.
e. Validate your model with cross-validation and describe how your predictions are robust (accuracy vs. generalizability).
f. Provide additional maps and data visualizations to show that your model is robust.
g. Talk about how your analysis meets the use case you set out to address.
h. What could you do to make the analysis better?

I expect to see data visualizations that are of high quality. I want to see codeblocks in your markdown.

Project options

Project option 1 – Predict heroin overdose events in Cincinnati, Ohio to better allocate prevention resources

The City of Cincinnati very recently released a dataset of heroin overdose locations. Using these data extracted from the Cincinnati EMS Open Data, your job will be to estimate a geospatial risk prediction model, predicting overdoses as a function of environmental factors like crime, 311 and inspections. You should validate your model against a kernel density as we have did in class. Also, you should try to train your model from one time period (long enough to have enough data) and test it on an out of out of sample test set time period (the following year, for instance).

Think critically about how you might offer these predictions to a public health official in your app. What do they want to know? Also remember that while your predictions are about overdose, it may be safe to assume that these are also places where people are just using heroin.

Project option 2 – Predict food inspection failures in Chicago to better allocate inspectors The Chicago Health Department wants to come up with a better way to allocate their limited health inspectors across the many food establishments in the City. How well can you predict if a food establishment will fail a health inspection? Can you figure out an interesting way to use the model to help the Health Department prioritize their inspections? This will use a logistic regression.

Specifically, your goal is to estimate a model using inspection data from one year to predict for the next. Does your model work better for certain kinds of establishments? Certain types of neighborhoods? Find your data on the Chicago Open Data Site.

Project option 3 – Predict EMS call to better allocate ambulances: Virginia Beach has shared data on emergency management responses to 911 calls. Can you predict where *and when* these calls will be made and test on the next few weeks? If so, perhaps it makes sense to put ambulances at certain places and times to reduce response times.

The first set of questions you have to wrestle with is, where are calls coming from and what are the response times to these places? You would likely aggregate these calls to a larger geography which means that you would be predicting a continuous outcome like number of calls per hour. Your app would probably be aimed at ambulance drivers to figure out where they should hang out to reduce response times. I wonder if it makes sense to know where they are currently dispatched from?

Project option 4 - Forecast traffic counts using built environment data (NEW): Austin has installed 54 sensors across the City to detect traffic in space and time. Can you forecast traffic counts in space (aggregated across time) as a function of land use, street characteristics and other built environment features? This is a tricky model to setup. It's not quite like the ride share approach because the spatial fixed effect won't work here. What is the appropriate unit of analysis? How can you model the spatial aspects of this process?

How could you use these predictions to inform a development-oriented app?

Project option 5 – Forecasting wildfire risk for a region in California (HARD):

With climate change, the State of California is exhibiting increased threat of wildfire. No doubt fire risk is a function of climate and weather, but also a host of time-invariant, spatial variables such as vegetation, elevation, land cover and more. Your challenge is to integrate California's Fire Perimeter data for 2-3 or years with other fire data, vegetation, land cover data, elevation data and other, to estimate fire risk. Can you use spatial cross-validation to validate this model?

There are multiple ways of doing this and if you choose this project, you should meet with me to discuss some possibilities. For an app, granted none of us are forestry experts, but can you design a fire management app that prioritizes where naturalist should clear brush, do burns, etc. Maybe, this is an app aimed at insurance companies or homeowners…

Project option 6 – Forecasting space/time scooter demand in Louisville (NEW)

This spring we're planning a big scooter project. Get a jump start by forecasting scooter demand in space time using data from Louisville. Where is demand in space/time? Where are underserved areas and are there demographic differences in service delivery that may give rise to equity issues? Can you build a scooter rebalancing app? Maybe an app for scooter riders?

Project option 7: The 'Michael Fichman Project' – Forecasting parking demand

What drives parking demand? If you knew, could you create a tool that would predict parking demand when development changes? Using Seattle open parking data, forecast parking demand as a function of built environment and road characteristics.

Project Option 8: Forecast new construction permits in Philadelphia in space and time

What are the economic and social features of neighborhoods that predict where new construction permits are filed in Philadelphia? Can you model the process of these permits proliferating across space and time for the say, the last 10 years? Would the neighborhood (maybe census tract) trend help you understand the speed of gentrification in certain places?

Project Option 9: Forecast train occupancy levels (Hard)
Can you forecast train occupancy for various OD pairs? There is a really great Kaggle dataset (here is a training and test csv and there is a station location csv). Note that there are three occupancy outcomes, low, medium and high, so you are going to have a three-way confusion matrix. Can you create an app that would help transportation planners do a better job planning the system?

Questions to prepare for Friday Nov 22nd.

What is the policy question?

How could data make a difference in answering this question?

What datasets have you identified to help you answer this question?

What kind of model would you build and what is the dependent variable?

How will you validate this model?

How do you think that stakeholders would want to consume this data?

What are the use cases for your app?

What should the app do?