Department of City and Regional Planning Fall 2021
University of Pennsylvania Profs Michael Fichman & Matthew Harris

FINAL PROJECT

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.

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.

Please focus on cross-validation (random, spatial and time, where appropriate) and on goodness of fit indicators (accuracy and generalizability) **that relate directly to the business process**. You may have to make up the business process, but please consider social, economic etc. costs/benefits.

Schedule

11/19 – Project introduced
12/3 – Prepare to discuss **your chosen project, data, methods** and some wireframe ideas in lab breakout.
12/10 - Presentations in class
12/17 – Final markdown and video due at noon.

1. **Deliverable 1 (Due 12/3):** Have a partner chosen, pick and project, and be prepared in lab to talk for a few minutes about the questions at the bottom of this document.

2. **Deliverable 2 (Due 12/10):** Team member 1 will be responsible for a 4 minute 'PechaKucha' presentation that 'sells' us on the idea of this fancy new planning app that you've designed to solve an important problem. Spend ~50% of your time on exploratory analysis and model results/validation. The expectation is that you will have preliminary model at this stage, which you will sharpen by the time the assignment is due. The other 50% should focus on questions like, 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**. Remember – sell it to us. What should come first - the model or the app? Don't forget to constantly remind the audience about the use case to keep your solution relevant.

3. **Deliverable 3 (12/17 - noon):**
Team member 1 will have the pechakucha uploaded on youtube with a recorded narration. Link to the video in your markdown.

Team member 2 will be responsible for a **markdown** write up that would allow someone to replicate your analysis (**show your code blocks**). Post this markdown on **your Github** not in a google folder. 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 analysis 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 useful (accuracy vs. generalizability).
f. Provide additional maps and data visualizations to show that your model is useful.
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. Please include codeblocks.

Project options

Project option 1 – Predict heroin overdose events to better allocate prevention resources

The City of Mesa has a dataset of heroin overdose locations. Using these data extracted from the city’s Open Data portal, 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). Note the fact that the data have some accuracy diminished to make them more anonymous – think about how this plays into your prediction and use case.

You can also undertake this project using similar data from Cincinnati, Ohio.

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 (NEW): The City of Santa Monica has shared data on emergency management responses to 911 calls.

How do EMS calls vary with demographics? Can you predict where *and when* these calls (Call Type == 'EMS') 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. If multiple teams want to take this on, you can also do this analysis in Virginia Beach.

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 either a count or binary outcome about 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 Metro train delays in and around NYC: An amazing new dataset has popped up on Kaggle recently that list origin/destinations delays for Amtrak and NJ Transit trains. Can you predict train delays? Consider the time frame that it would be useful to have such predictions. Predicting 5 minutes out is not going to be as useful as 2-3 hours out. Consider training on a month and predicting for the next week or two. Consider time/space (train line, county etc.) cross validation. Many app use cases here.

Project option 5 – Forecasting wildfire risk for a region in California:
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 possible model approaches here. 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 – Forecast Airbnb Prices in Amsterdam
You can predict home prices – how about apartment rents? Using a scraped Airbnb dataset, predict property prices based on property characteristics and other neighborhood level data. For your app, do not reinvent the Airbnb app, but think about a consumer facing or public-policy focused use case. Consider the fact that regulation of Airbnb is a contentious issue in tight real estate markets and the service has been outlawed in some cities.

Project option 7: Forecasting parking demand What drives parking demand (revenues)? If you knew, could you create a tool that would predict parking demand over time and space. Using San Francisco open parking data (locations), forecast parking demand as a function of built environment, neighborhood and road characteristics. Lots of app use cases possible depending if the user is public or private sector. There are *105m rows here* – so you'll have to use the API to download a small slice of the data.

Project Option 8:
Forecast train occupancy levels

Can you forecast train occupancy for various OD pairs? There is a really great Kaggle dataset (the data are here as 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? This table shows which stations are on which lines.

Questions to prepare for Friday December 3rd.

What is the use case?

How could data make a difference in answering this question? Do you have a sense for the business as usual decision making?

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 (cross-validation & goodness of fit metrics that relate to the business process)?

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

What are the use cases for your app and what should the app do?