Machine-learning-assisted creation of optimally located STEM programs

Clément Sicard, Kajetan Pyszkowski, Alexis Tabin
D-INFK Faculty, ETH Zurich

1. Introduction

- Nowadays, large cities collect **many data about socio-economic criteria**, such as unemployment rate, crime and education level
- In this project, using the public data, we aim to assist **the City of Chicago's Education Department** in finding **optimal** locations for new Science, Technology, Engineering et Mathematics (**STEM**) programs in order to **help fight crime, unemployment and poverty**

2. Method Description

- We created an **interactive dashboard** to help the Education Department to better visualize socio-economic criteria
- In the **back-end**, we serve **public data** (that we had previously pre-processed for performance reasons) and the inference result of our machine-learning model

The dashboard can be used in **two modes:**

1. **Interactive data visualization mode:** Explore the datasets first
    - Visualization consisting of an **interactive map** and a **collection of interactive graphs** related to this dataset are displayed
2. **Interactive prediction mode**: Make data-driven predictions
    - Visualize model predictions on a map
    - The user can **weigh interactively the contribution of each feature** to inform the prediction and **choose the number of predictions**

- In both modes, the user can **change/choose interactively** the dataset that is being visualized in the front-end

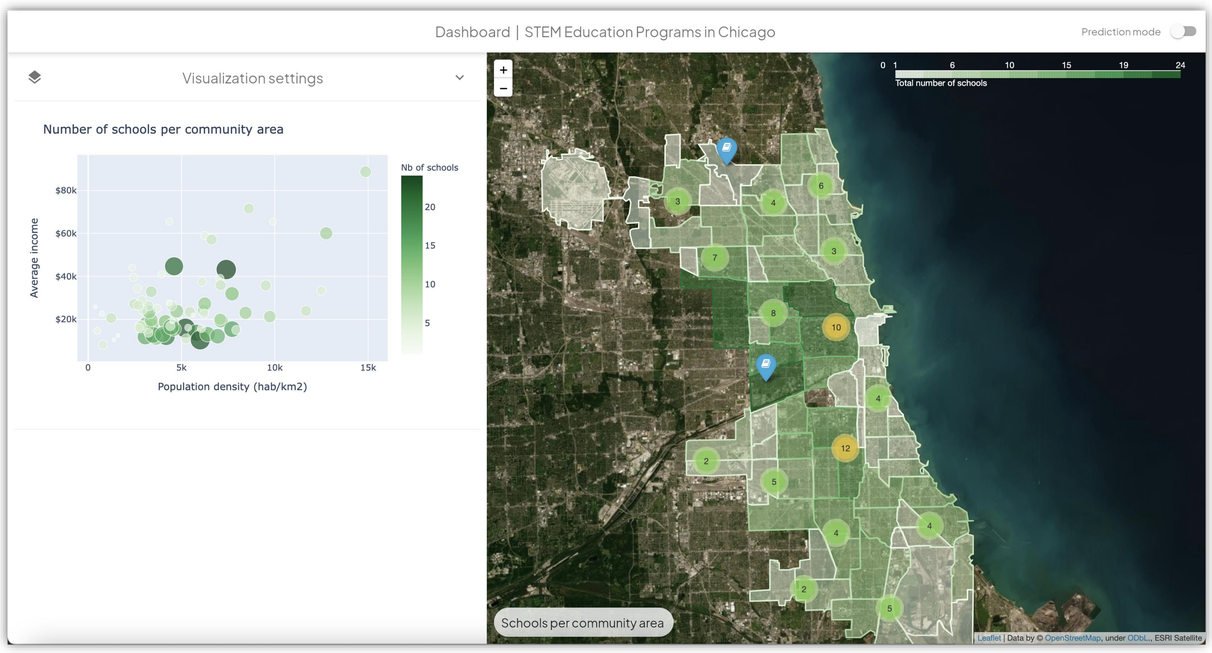

Figure 1: Screenshot of the dashboard in **interactive data visualization mode**

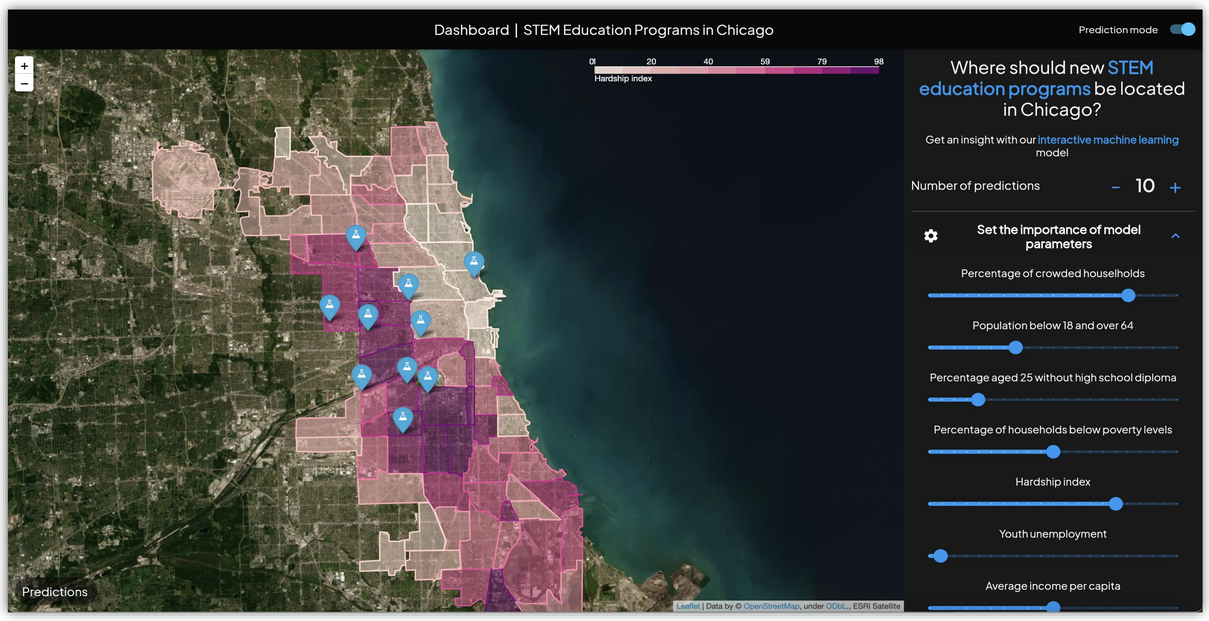

Figure 2: Screenshot of the dashboard in **interactive prediction mode**, with the feature weights sliders

3. Data Sources & Pre-Processing

- We used **15+ public datasets** from the **City of Chicago Open Data Portal** (which includes unemployment, crime, hardship index, average/median household income, school and library locations)
- Our reference geographical data unit is the **community area**
- We **grouped** every socio-economic criteria **by community area**, aggregating on the most coherent statistic to each feature

Figure 3: Division of the City of Chicago by community areas

4. Tech Stack

Back-end API & Data Manipulation	Front-end Interactive Dashboard, Graphs & Maps
python™ Flask (web development, one drop at a time) GeoPandas	Flutter Leaflet plotly

Back-end: Python and Flask for the API, GeoPandas for geospatial data pre-processing & interactive map serving
Front-end: Flutter (built on top of Dart language) for the dashboard UI, Leaflet for the interactive maps and plotly for the interactive graphs

5. Interactive Machine-Learning Model

- **Our goal:** to allow experts to choose which importance they give to each feature, closely following **human-in-the-loop** model
- To do so, our model is largely inspired by the **Geographically Weighted Regression (GWR)²** (Fotheringham, et al., 1998) model
- Weight distribution is **interactively set** in the front-end **by the user** (Figure 4)
- We first decide **in which** community area we will create a new program, then determine **where** in this area it will fit the best, according to the selected weighted criteria

Figure 4: Screenshot of the interactive feature importance sliders in the dashboard

6. Conclusion & Discussion

- Our model predictions **can be largely improved.** This project was mostly a proof-of-concept more than a directly usable model (for instance, it doesn't take protected spaces into account when inferring a location – eg. a STEM inside O'Hare Airport)
- We have no qualification in public governance, so the **criteria we chose might not be pertinent**, even though they appear to be. However, the model is largely adaptable to any other dataset that a qualified individual would find pertinent
- Visualization mode **helps** the qualified user **quantify** the selected socio-economic criteria – then prediction mode helps them **make informed decisions** based on the model predictions
- We can conclude that even though the model is not perfect, it fulfills its role of a decision-making tool to assist public policies

References

1. City of Chicago Open Data Portal, 2022, https://data.cityofchicago.org/
2. Fotheringham, A. S., et al. "Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis." Environment and Planning A: Economy and Space, vol. 30, no. 11, Nov. 1998, pp. 1905–1927, doi:10.1068/a301905