# Developer Guide for Predicting the Impact of Broadway Extension in Vancouver City Using Double Machine Learning in Panel Data Framework

## Introduction

This project investigates the impact of the Broadway Subway extension on crime rates in Vancouver using a double machine learning approach within a panel data framework. The research combines environmental criminology theories with machine learning and econometric methods to analyze historical data and forecast crime occurrences.

## Data Sources

**Crime Data:**
- Source: Vancouver Police Department's GeoDASH (2003 to May 2024)
- Description: Includes instances of various crime types, with a focus on 'Break and Enter Commercial' for this study.

**Transit Stations Data**
- Source: City of Vancouver Open Data Portal
- Description: Locations and operational dates of transit stations, including the Expo Line, Millennium Line, Canada Line, and R5 RapidBus.

**Additional Features:**
- Source: City of Vancouver Open Data Portal
- Description: Includes business locations, street lighting poles, graffiti locations, and more.

**Group-Specific Data:**
- Source: Abundant Housing Vancouver
- Description: Neighborhood-specific data such as population, income, and unemployment rates. Census data from 1996, 2001, 2006, 2011, and 2016 were linearly interpolated to estimate values for 2008 to 2023.

**Street and Junction Data**:
- Source: ArcGIS, GeoDASH, and Translink Open API
- Description: Street segment lengths, neighboring junctions, and the structure of the streets and junctions.

## Data Preparation

Data preparation involves multiple steps, primarily carried out in the following notebooks:

**1. census_data_extraction.ipynb:**
- Extracts neighborhood-specific data from census records.
- Interpolates data to fill gaps between census years.

**2. merged_dataprep.ipynb:**
- Merges various data sources to create a comprehensive dataset.
- Prepares panel data with observations of 6,179 street junctions from 2008 to 2023.

- Counts features by finding the closest junction to each feature instance using longitude and latitude coordinates.

**3. visualizations.ipynb**:

- Visualizes the initial data distributions and correlations.
- Helps in understanding the spatial distribution of crime and other features.

**4. counts_prep.ipynb:**

- Prepares count features for various attributes (e.g., number of crimes, businesses, transit stations).

## Reach Calculation

The concept of reach, as developed by Stephen Borgatti, was used instead of simple counts to better capture the influence of neighboring junctions. The reach calculations were performed in the Data_wrangles notebooks, as detailed in the paper. Reach measures provide a more accurate representation of the proximity effect by normalizing results and managing high-density areas effectively.

## Double Machine Learning Implementation

The DML approach was implemented in the DML_FINAL_5_Models.ipynb notebook. This notebook includes:

- Implementation of five models: K-Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Long Short-Term Memory (LSTM), and Spatial-Temporal Graph Convolutional Networks (STGCN).
- Evaluation of models using metrics like RMSE and R-squared.
- Visualization of model performance and results.

## Predictions and Visualization

The predictions.ipynb notebook contains various methods for visualizing spatial data, showing the differences in crime reach before and after the introduction of new transit stations. This notebook includes:

- Visualization of the estimated effect of new transit stations on crime reach.
- Comparison of old and new data values.

## Requirements

To run the notebooks and models, all necessary packages are listed in the requirements.txt file. Install these packages using the following command:

```
pip install -r requirements.txt
```

## Essential Notebooks

Only the essential notebooks for data preparation, reach calculation, DML implementation, and visualization are included in this guide. Other notebooks used for testing different approaches and models are omitted as they do not interfere with the final results.

### List of Essential Notebooks:

1. census_data_extraction.ipynb - Extracts and processes census data.
2. merged_dataprep.ipynb - Merges various data sources and prepares panel data.
3. visualizations.ipynb - Visualizes initial data distributions.
4. counts_prep.ipynb - Prepares count features.
5. Data_wrangles notebooks - Calculates reach measures.
6. DML_FINAL_5_Models.ipynb - Implements and evaluates DML models.
7. predictions.ipynb - Visualizes predictions and spatial data changes.