CASA0006 Week 1 Quiz

Huanfa Chen

15 January 2025

Overview

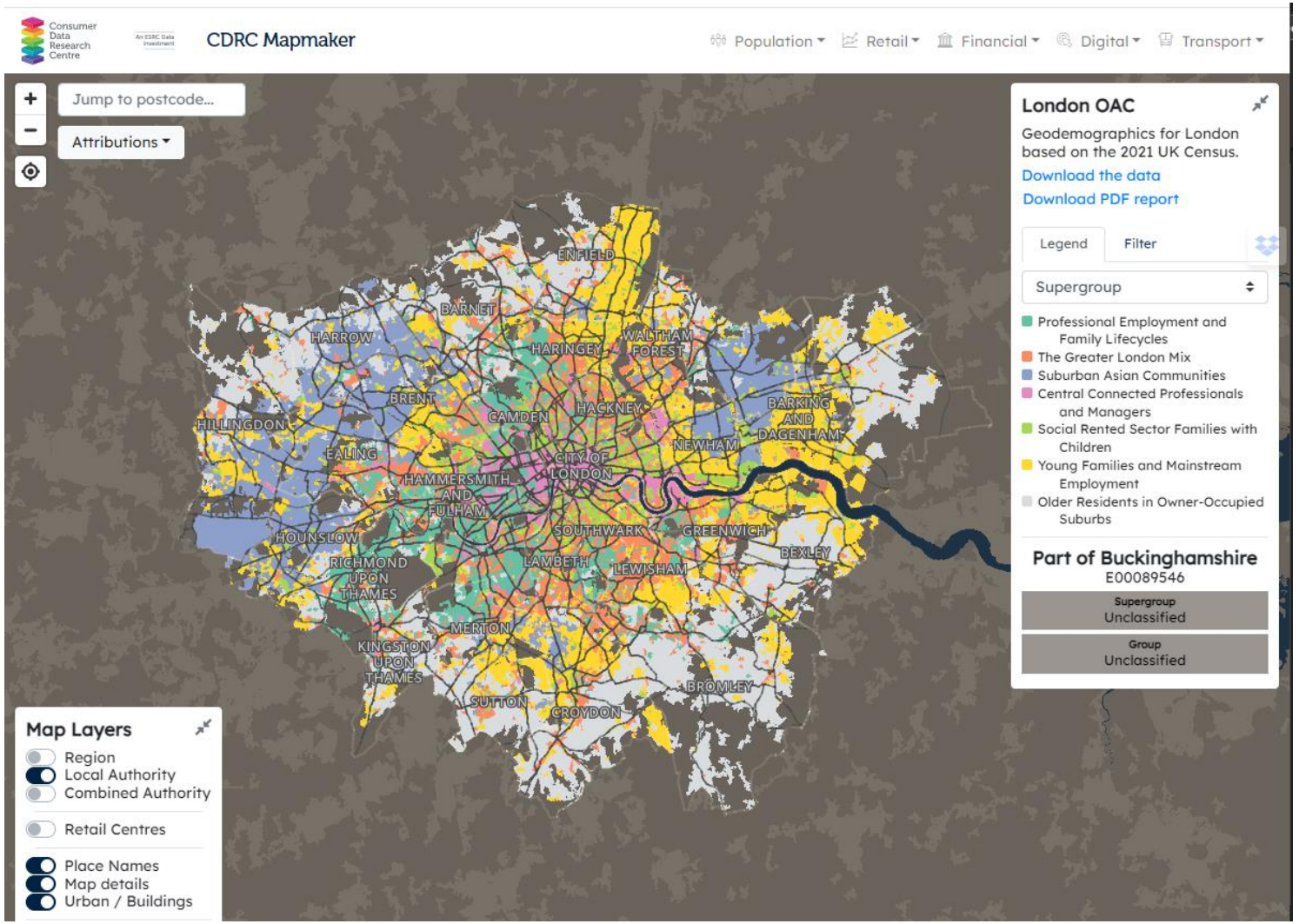
?? Participants

Q1: Which of the following are applications of supervised machine learning?

- London Output Area Classification: use Census data and kmeans to cluster output areas into several classes. (https://data.london.gov.uk/dataset/london-area-classification)
- Sentiment analysis of twitter - predict whether a twitter is happy or not based manually-labelled twitter data
- Predict tomorrow's ridership of Santander bikes based on the data of the past 5 years
- Build a model to predict the champion of Qatar world cup based on historical match results

Q1: Which of the following are applications of supervised machine learning?

- London Output Area Classification used clustering techniques, which is unsupervised machine learning.
- Don't judge a book by its cover, or its name



https://mapmaker.cdrc.ac.uk/#/london-oac/

Q2: Which of the following are NOT true about machine learning?

- Only neural network models can be called machine learning. *(Counter example: random forest)*
- All machine learning algorithms require labelled data. *(Counter example: unsupervised learning, clustering)*
- For any problem, the predictive accuracy of neural networks is always higher than linear regression.
- Data size and quality is very important for machine learning models.

Neural Network with negative R2

$$R^2 = 1 - \frac{\text{sum of square of residuals}}{\text{variation of actual y (or total sum of square)}}$$

- If you fit an OLS linear regression on a dataset, the R^2 is always non-negative.
- If you carelessly fit an artificial neural network on a dataset, you might get a negative R^2, which means the model is fitting the data very badly.
- Potential reason is that hyperparameters are not properly selected.

Q3: Which are true about 'No free lunch theorem'?

- It implies that no model is better than the others for all problems.
- It implies that we can't find the best machine learning algorithm for a given problem and dataset. *(Given a dataset and a set of ML algorithms, we can empirically find the best algorithm)*
- It indicates that it is not worthy to train large machine learning models. *(Generally, the larger model, the better fitting performance)*
- It implies that the validation of models should be conducted empirically for a specific task.

Q4: Which of the following are assumptions of linear multiple regression?

- Linear relationship between x and y
- Independent errors
- Normally distributed errors
- Equal variance of errors
- The x variable is normally distributed
- The y variable is normally distributed

Q5: Which of the following are true regarding linear regression?

- Linear regression models can be used for clustering analysis. *(No, linear regression is supervised learning)*
- Linear regression models can be used to predict y value given x values
- Linear regression models has good interpretation
- Linear regression has no assumptions about the dataset. *(Linear regression has four assumptions, LINE)*