



Data Analysis of OpenMaps data

Paola Elefante

Paola Elefante, paolaelefante.com@gmail.com, tel. 040 0217654, paolaelefante.com

Description of data

Data frame with

- Customer's assigned id code
- 1153 i_geocode variables: number of points of interest (ex. fire station, café, ...) in 500 m, 2 km and 10 km area around the customer
- Target1: target value of 0 or 1

GOAL: predict Target1 entries where missing (about 30% of cases)

Questions

How do points of interest affect the Target1 value?

Which statistical model to choose?

How to cope with the size of data?



Assumptions & choices

I assume that the geocode variables are independent.

Since Target1 attains two values, I model it by binomial distribution.

I assume dependence is linear and use a probit prediction model.



Issue: overfitting

Too many geocode variables for Probit to work fine.

Solution 1

- Randomly pick N geocode variables.
- Compute the Probit model.
- Compute the AIC value to evaluate how well the model fits.
- Repeat M times and choose the model with lowest AIC value.

Solution 2

- Randomly pick N geocode variables.
- Compute the Probit model.
- Use the model to predict Target1 values.
- Repeat M times and take a majority vote to decide the final prediction values.

on average,
this works in:

67% of cases

65% of cases

Solution 1 works better and has better performance:
predicted data are stored in NAtarget_pred.csv

Description of the customer's area


Regardless of Target1, we can make a qualitative description of the customer's area.

- Exclude the points of interest which are 10km away: they are irrelevant.
- Give a smaller weight to Poi 2km away.
- Clean data (ex. remove columns with nonsense names).
- Description is not objective, the visual choice has to reflect that.

Random Customer Example



Customer #19488

Parking	157
Restaurant	41
Telephone	28,5
...	

(0.5 weight given to further locations)

Benefits of this model

- Flexible, easy, neutral visual representation (let the data talk).
- Automatic method.
- If the dataset is updated with new information, the method still works.