Predicting the Propensity to Move Using Register Data in Flanders

Armen Abagyan
24/05/2022

Research Motivation

- Statistics Netherlands study “Replacing a survey question by predictive modeling using register data” (Joep Burger et al, 2018)
- Replaced the Dutch Housing Survey on desire to move houses within two years by applying predictive machine learning algorithms on Dutch register data.
    - Survey questionnaires expensive, time consuming, declining response rates, mismatch between responded behavior and actual behavior → general shift from primary observation with survey questionnaires to secondary observation from administrative registers and big data.
- Authors linked several registers from the Dutch System of Social Statistical Datasets (SSD) containing life history events from 1995-2016, and all features known up to reference data were used to predict moving behavior within two years of that reference date.
- What about Flanders?

Data sources and features of interest

- Data for 2010-2019 (2020-2021 excluded)
  - Stock data
  - Flow data (deaths, births, internal migration)
  - Financial data
  - Education data
  - Statistical Sector data
- Individual characteristics
  - Country of origin, nationality, civil status, position within household, income (classified low, middle, high), starter, educational level (status?), employment status/ type, home ownership
- Households characteristics
  - Household type, # of people in HH, # of children in HH, HH income (classified low, middle, high)
- Statistical Sector/neighborhood Characteristics
  - Percentage of people over 65, Percentage of persons with a parent not born in Belgium , percentage of low/middle/high -income households, percentage of family households, percentage of single parent households, percentage of HHs with home ownership
- Interaction terms
  - Household type with percentage of household type in SS, HH income classification with percentage of HHs income classification.

Features (cont) – Life Events

- Events/ Change in Household within year t-1 with respect to reference date for year t
    - Life Course Events
        - Change in Civil Status
        - Change in employment/educational level
        - Moved in previous year/ number of moves in previous year
    - Change in Household Composition
        - Death within HH
        - Birth within HH,
        - Someone else within HH moved
        - Change in HH type
- Time since last life course event or change in household composition with respect to reference date for year t

Sampling and Cross Validation strategies

- Sampling
  - Only internal migration within Flanders itself is considered.
  - Ages 16 -64.
  - Collective households (public institutions excluded)
  - 50,000 households in Flanders sampled in 2010, and individuals followed throughout years 2010-2019.
  - Stratified sampling to assure equal proportion of movers in full and sampled data set.
    - Within Flanders, less than 3% of the population move houses, want to assure same percentage in sample.
- Cross Validation

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
TRAIN	t-7	t-6	t-5	t-4	t-3	t-2	t-1	t		
VALIDATE		t-7	t-6	t-5	t-4	t-3	t-2	t-1	t	
TEST			t-7	t-6	t-5	t-4	t-3	t-2	t-1	t

Machine Learning Binary Classification Models:

- **1. Penalized Logistic Regression: Ridge, Lasso, and Elastic Net Regression**
  - Penalized logistic regression: imposes penalty to LR for high dimensionality, results in shrinking coefficients of less contributive variables towards zero (Regularization).
  - <u>Loss Function Ridge Regression:</u>
    - Optimization procedure keeps loss function minimal
    - Lambda controls how much emphasis is given to the penalty term.
    - Coefficients pushed to zero but never achieve zero, not ideal if we only want to select important features

$$L_{log} + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

  - <u>Loss Function: Lasso Regression</u>
    - Coefficients pushed all the way to zero.
    - Penalty tends to pick one variable when predictor variables are correlated

$$L_{log} + \lambda \sum_{j=1}^{p} |\beta_j|$$

  - <u>Loss Function: Elastic Net</u>
    - Combination of both Lasso and Ridge regression
    - Additional alpha parameter to give weight to Lasso or Ridge regression.

$$L_{log} + \lambda \sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$$

- **Class weights** implemented to deal with the highly **imbalanced data.**

$$log\ loss = \frac{1}{N} \sum_{i=1}^{N} [-(w_0(y_i * log(\widehat{y_i})) + w_1((1-y_i) * log(1-\widehat{y_i})))]$$

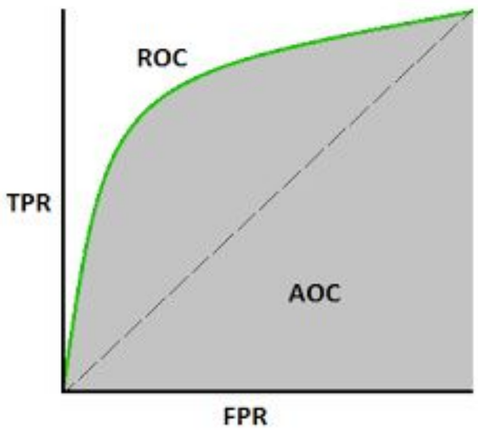
Machine Learning Binary Classification Models:

- 2. XGBoost: Extreme Gradient Boosting
    -
    - XGBoost vs. Random Forest:
        - 1. XGBoost prunes the tree immediately with “similarity score” before entering into the actual modeling purposes.
        - 2. XGBoost requires far less hyper parameters than Random Forest.
        - 3. XGBoost better handles unbalanced data sets and Random Forest is less reliable.

Evaluation of Machine Learning Methods

- Confusion Matrix
  - True Positive Rate/ Sensitivity: TP/(TP+FN)
  - Specificity : TN/(TN+FP)
  - False Positive Rate (1-Specificity)
  - Precision: TP/TP+FP
- ROC-AUC CURVE

Actual Predicted	Moved=1	Moved =0
Moved=1	TP	FP
Moved=0	FN	TN



- ROC- Alternative to large number of confusion matrices in case of change of threshold.
- Can compare the ROCs of different machine learning models with AUC- % chance that the model will be able to distinguish between positive and negative classes.

Thank you !