Predict the population of high earnings with correlated features

Team: Error 404 Sleep Not Found

Zihan Ding
Meiru Zhang
Lingjun Liu
Mingrui Zhang
Ran Yan
Jiahong Wang

Outline

1. Introduction & Motivation
2. Methodology
3. Dimensionality Reduction
4. Results
5. Conclusion
6. Reference

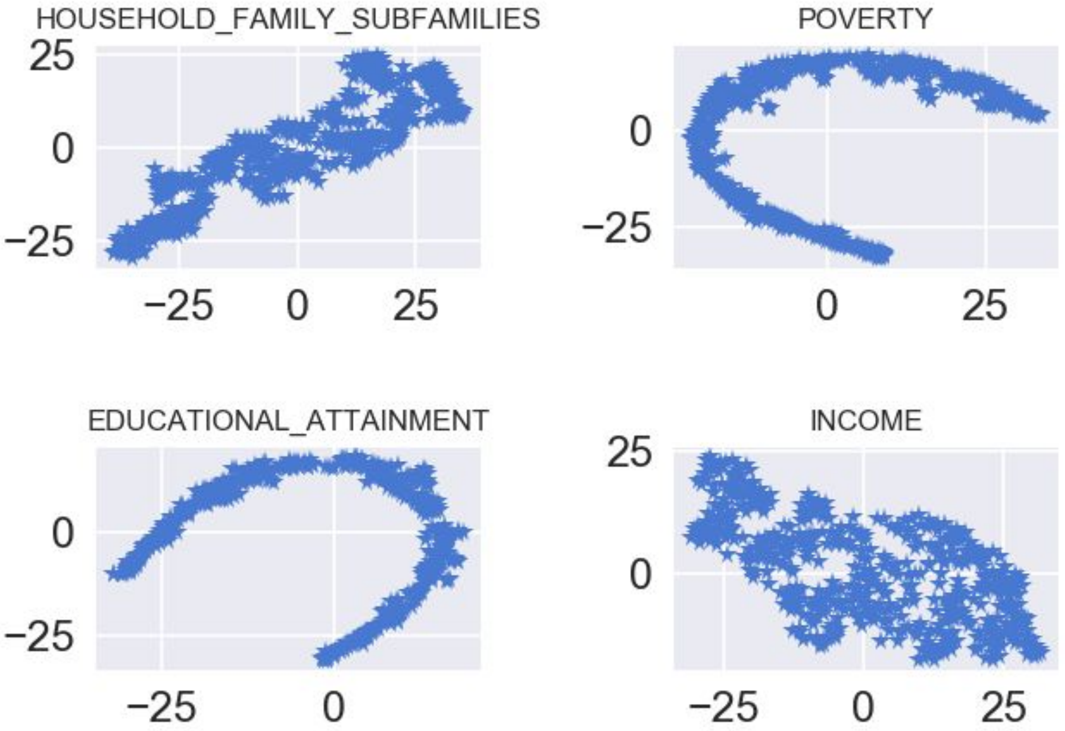
Introduction & Motivation

1. Socioeconomic status increasingly plays important roles in analyzing the behaviour of society, e.g. health, incomes and educations.
2. The gap of earnings could be a potential issue.
3. We could provide some remarkable suggestions in a quantitative way to reduce the gap between the rich and poor.

Methodology

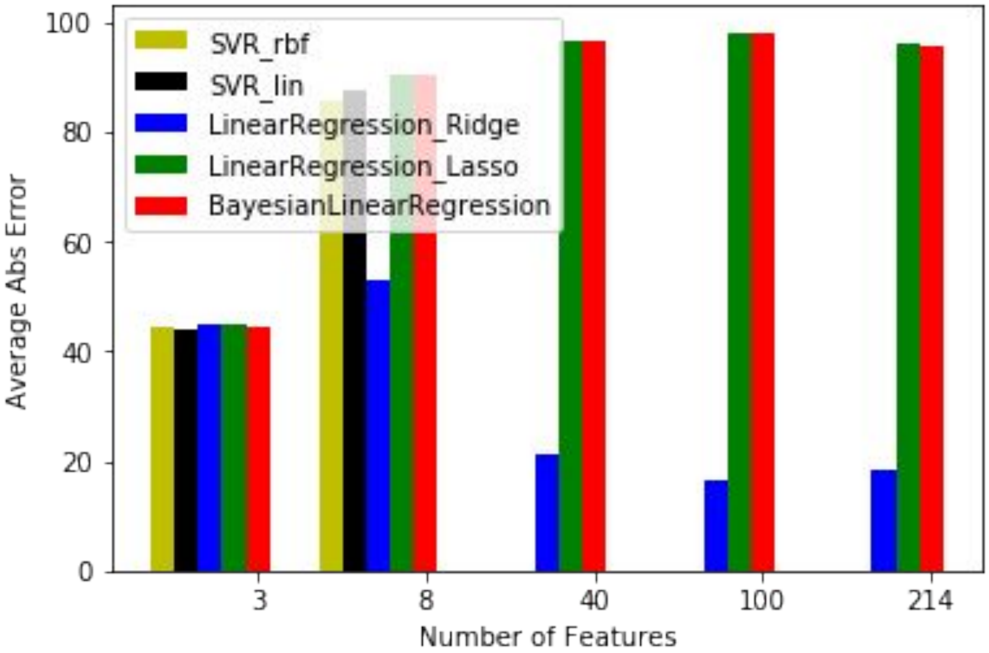
- Feature Selection
  - Calculate the Pearson correlation coefficient between the feature of the size of population of high earnings and other features
  - $\rho_{X,Y} = \dfrac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$ ∈ [-1, +1]
  -
  - Threshold of correlation coefficient 0.5, 0.6 and 0.7
  - Separate the original dataset to training(80%) and validation data(20%)
  - Remove “margin to error” features and columns with null values
- Prediction
  - Support vector regression
  - Linear regression
  - Bayesian linear regression
  - etc.

Dimensionality Reduction



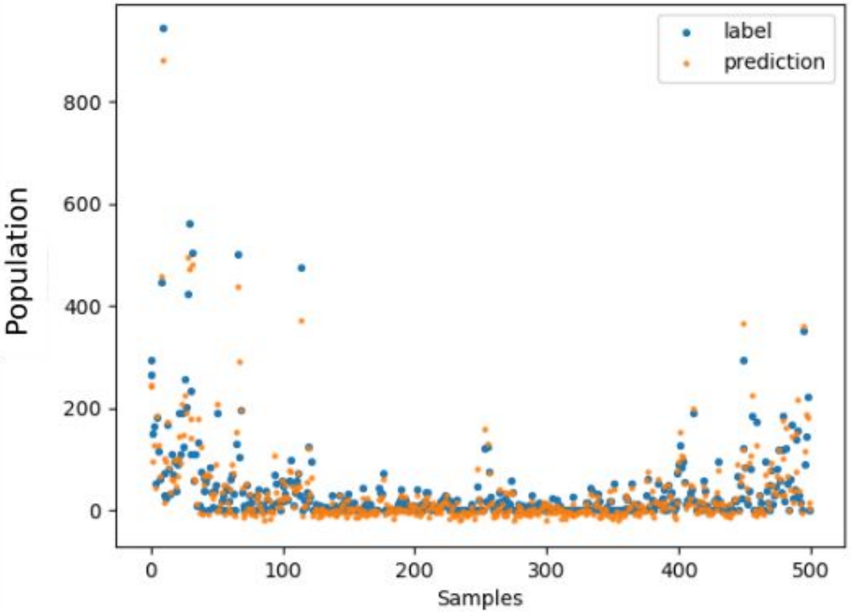
Methodology: T-distributed Stochastic Neighbor Embedding

Results



Comparisons of different regression models with different number of features. Bayesian Linear Regression, Linear Regression with Ridge or Lasso regularization, Support Vector Regression with Linear or RBF kernel.

Results



Prediction results are close to the true label values. (~ average absolute error of 16.5 for each sample, range of label value is [0,1600])

Conclusion

- High-earning population in certain location is highly correlated to education attainment and household family relationship

- The other key features also include: poverty, income and earnings(of females in the family)

- Decent prediction results on validation set.

- Future investigation on predictions on other feature (e.g. health condition) or more robust model structure.

Reference

[1] Samuel Bowles, Herbert Gintis, and Melissa Osborne. Incentive-enhancing preferences: Personality, behavior, and earnings.American Economic Review, 91(2):155–158, 2001.

[2] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In Advances in neural information processing systems, pp. 155–161, 1997.

[3] David L Featherman and Robert M Hauser. Sexual inequalities and socioeconomic achievement in the us, 1962-1973.American Sociological Review, pp. 462–483, 1976.

[4] John Neter, William Wasserman, and Michael H Kutner. Applied linear regression models. 1989.

[5] Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models.Journal of the American Statistical Association, 92(437):179–191, 1997.

[6] Gary Solon. Cross-country differences in intergenerational earnings mobility.Journal of Economic Perspectives, 16(3):59–66, 2002

Question?