Birthrate prediction model

Introduction

- **This project aims to predict London boroughs GFR (general fertility rate) based on data from various sources, including Foursquare location data.**
- **This is a proof-of-concept, estimating GFR based on publicly available data.**
- **This can be of interest to social science researchers who investigate the impact of POIs availability on the fertility rate, or to city planners, etc..**

Data

- **Wikipedia - List of areas of London - Table of neighborhoods in London, the borough(s) they belong to.**
- **Wikipedia - List of London boroughs - Table of boroughs, their area and population and co-ordinates.**
- **Python geopy Yandex - to identify the co-ordinates of each London neighborhood.**
- **Foursquare API – location data - POI (Point Of Interest) categories rates for each neighborhood.**
- **London Datastore - Births and Fertility Rates, Borough – Live births by local authority of usual residence of mother, General Fertility Rates and Total Fertility Rates.**

Methodology

- **Data gathering and preparation:**
  - Data from different sources (wikipedia pages, London datastore) was gathered, cleaned and merged
  - Yandex (using geopy python package) and Foursquare APIs were used for location data retrieval
- **Regression models building:**
  - Data were split to training and test sets, 80%/20%, for accuracy evluation
  - PCA was used to decrease dimensionality of the data: an important step, since number of originally gathered features is too large compared to the dataset volume
  - A linear regression with ridge normalization, and support vector regression (non-linear, RBF kernel) were deployed
  - K-fold validation was used to choose models hyper parameters: grid search (linear regression) and random search (SVR)

Results

Using k-fold validation, following hyper parameters were selected:

- **Linear model:**
  - alpha ~ 41.41
- **SVR model:**
  - C ~ 160.3
  - gamma ~ 0.1
  - Epsilon ~ 0.87

Accuracy scores of the models:

- Linear regression RMSE: 7.73 live births per 1,000 women aged 15-44

  Linear regression R2 score: 0.66

  London has GFR of 62.9, hence RMSE of 7.73 is approximately 12.29% error.
- SVR RMSE: 6.19 live births per 1,000 women aged 15-44

  SVR R2 score: 0.8

  London has GFR of 62.9, hence RMSE of 6.19 is approximately 9.84% error.

Discussion

- **London has GFR of 62.9**
- **RMSE of the Linear regression model (7.73) is approximately 12.29% error**
- **RMSE of the SVR model (6.19) is approximately 9.84% error**
- **If a larger dataset is used (e.g., by using historical data, or if GFR per location is given), accuracy is expected to increase**

Conclusion

- **This project illustrates an adequate approach for gathering, pre-processing and merging of publicly available data from different sources for the construction of a regression model that can predict General Fertility Rate (GFR)**
- **A median error of 12.29% and 9.84% for linear and SVR models, respectively, was obtained**
- **The error is expected to decrease if a larger dataset is provided**
- **This can be of interested for social scientists, as a proof of POIs influence on birth rates, city planners, etc..**