Estimating Socioeconomic Attributes from Location Information

Shohei Doi (NII)

Takayuki Mizuno (NII/Univ. Of Tokyo)

Naoya Fujiwara (Univ. of Tohoku/Univ. Of Tokyo)

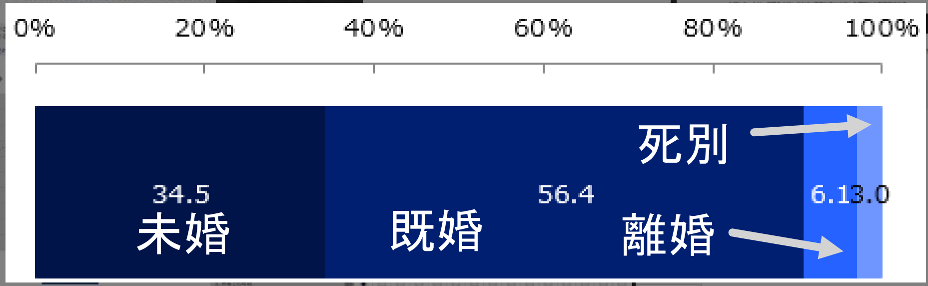
Motivation

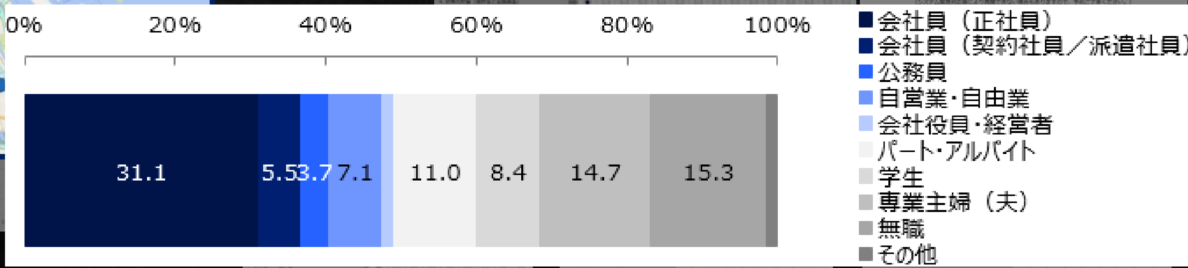
- Location-based service has been used for administrative and marketing purposes (Hammer et al., 2017; Huan et al., 2017).
- In particular, the distribution of socioeconomic attributes of individuals are crucial.
  - However, due to privacy security like GDPR, it is hard to obtain personal data associated with location information.
- To overcome this limitation, Lamanna et al. (2018) estimate office of a twitter user by geo-tagged tweets actively posted in the daytime while Lenormand et al. (2016) infer a user's house by tweets at night.
  - Personal attributes: mobile phone behavior (Ying et al., 2012; Al-Zuabi et al., 2019) , SNS (Cesare et al., 2018; Kosinski et al., 2013; Aletras and Chamberlain, 2018), photo (Lewenberg et al., 2016)
  - Distribution of attributes: content of talk via phone (Blumenstock et al., 2015), tweet (Montasser and Kifer, 2017), restaurant info (Dong, 2019)
- In this study, we collect survey data including location information and predict personal socioeconomic attributes.

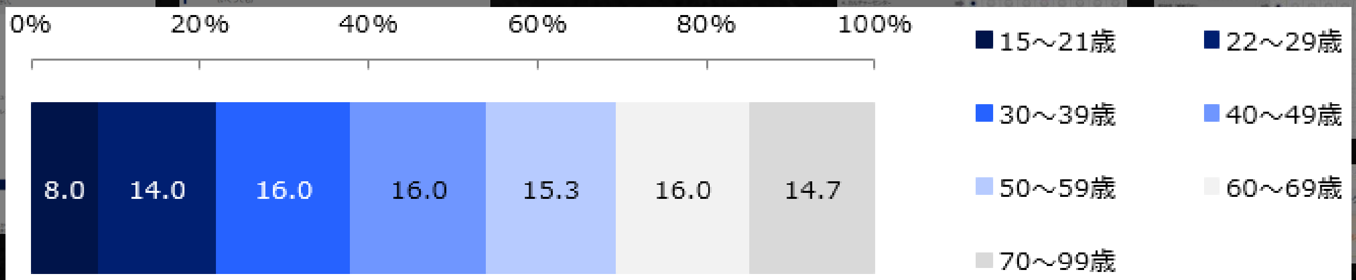
Contents

Rakuten Insight Survey Data

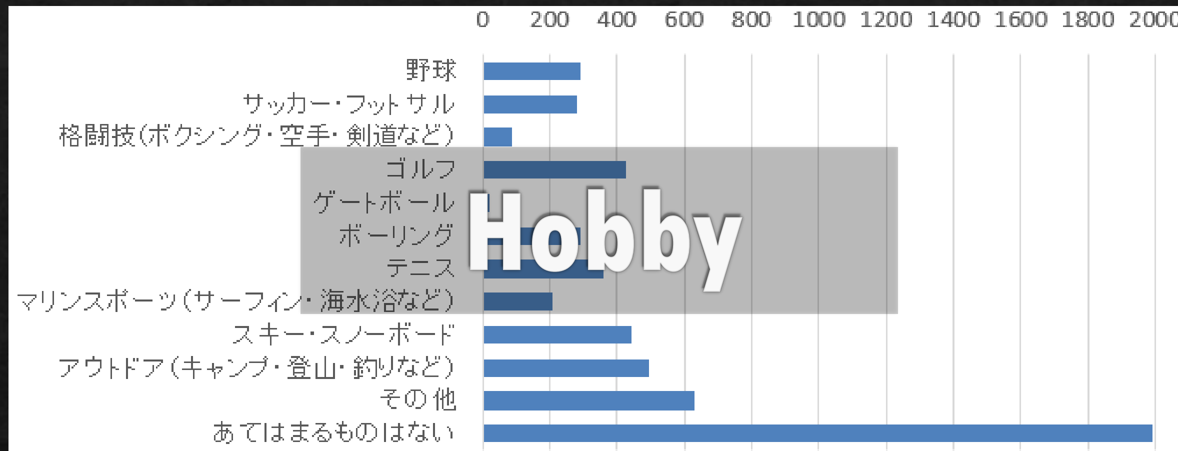
- Male:Female ＝ 1980:1980
- Tokyo:Miyagi:Hiroshima:Nagasaki ＝ 3000:400:400:160



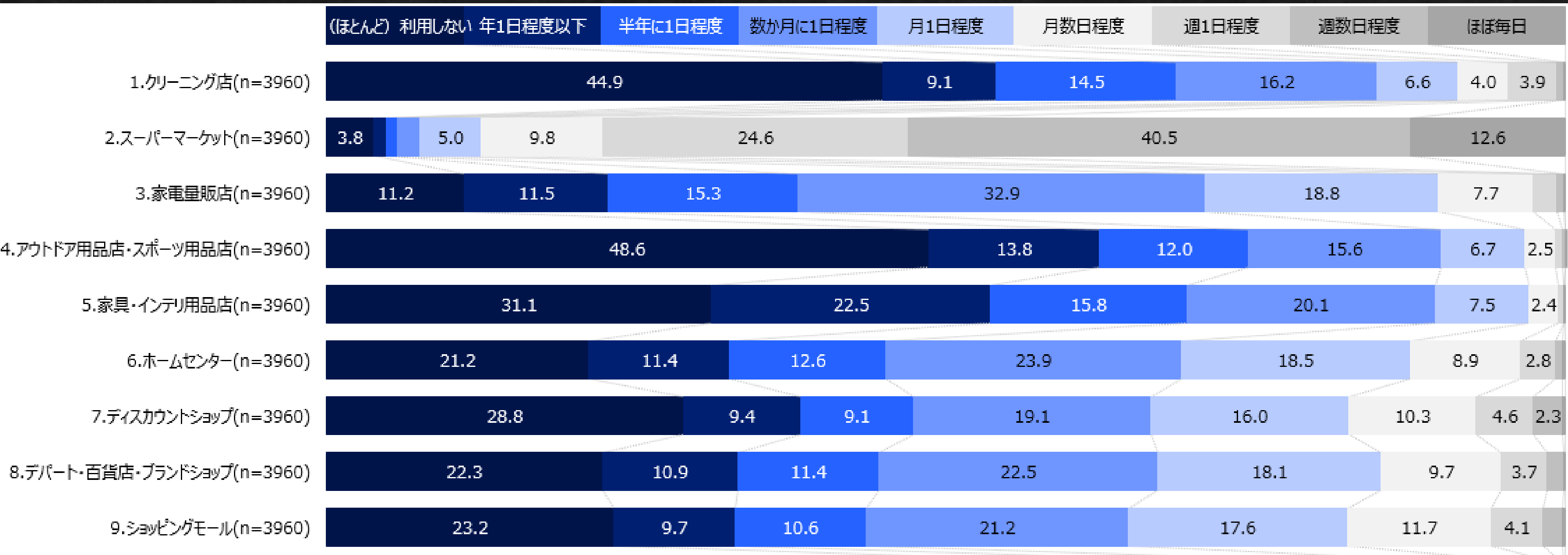




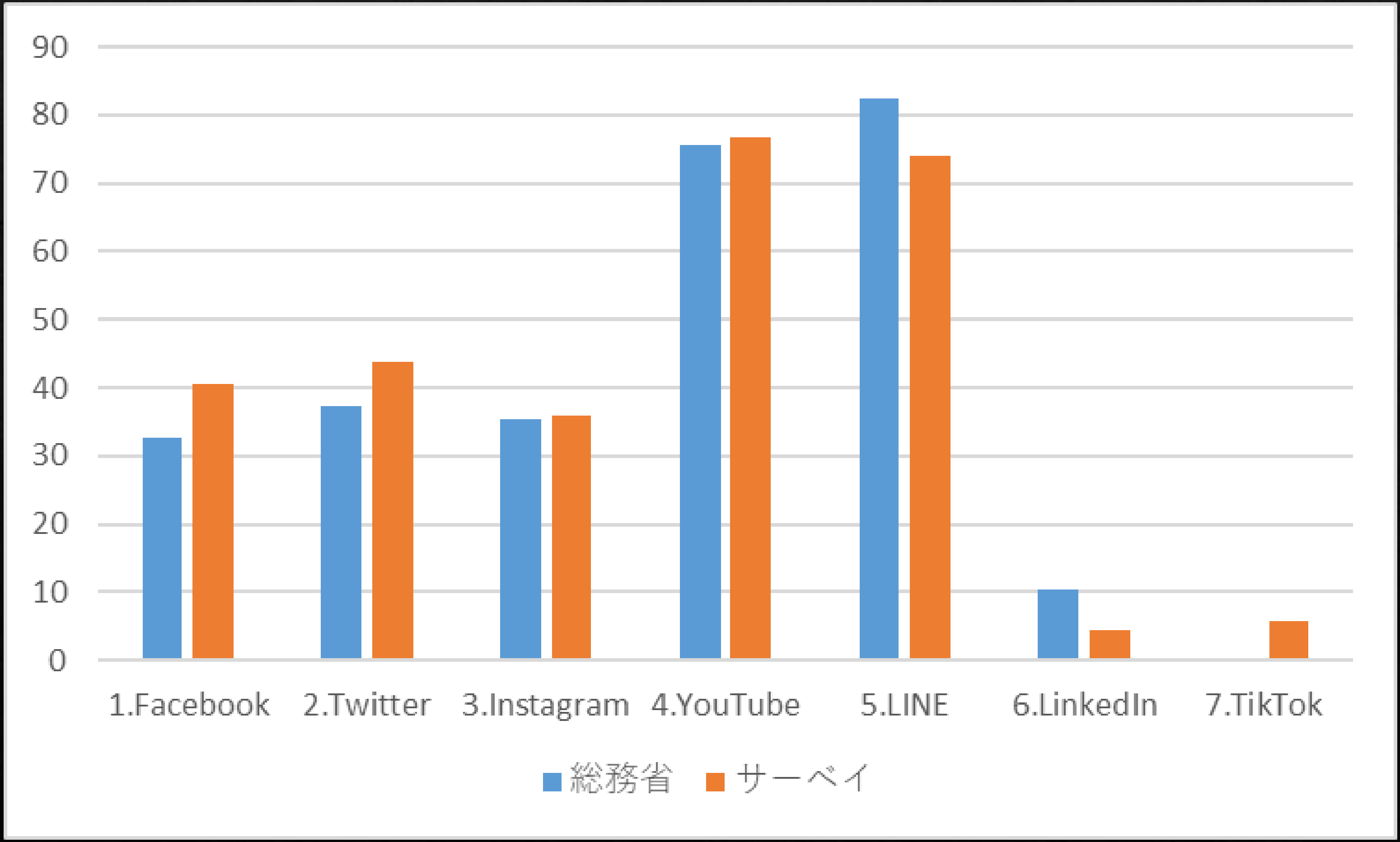
58 Personal Attributes



Frequency of Visiting 52 places



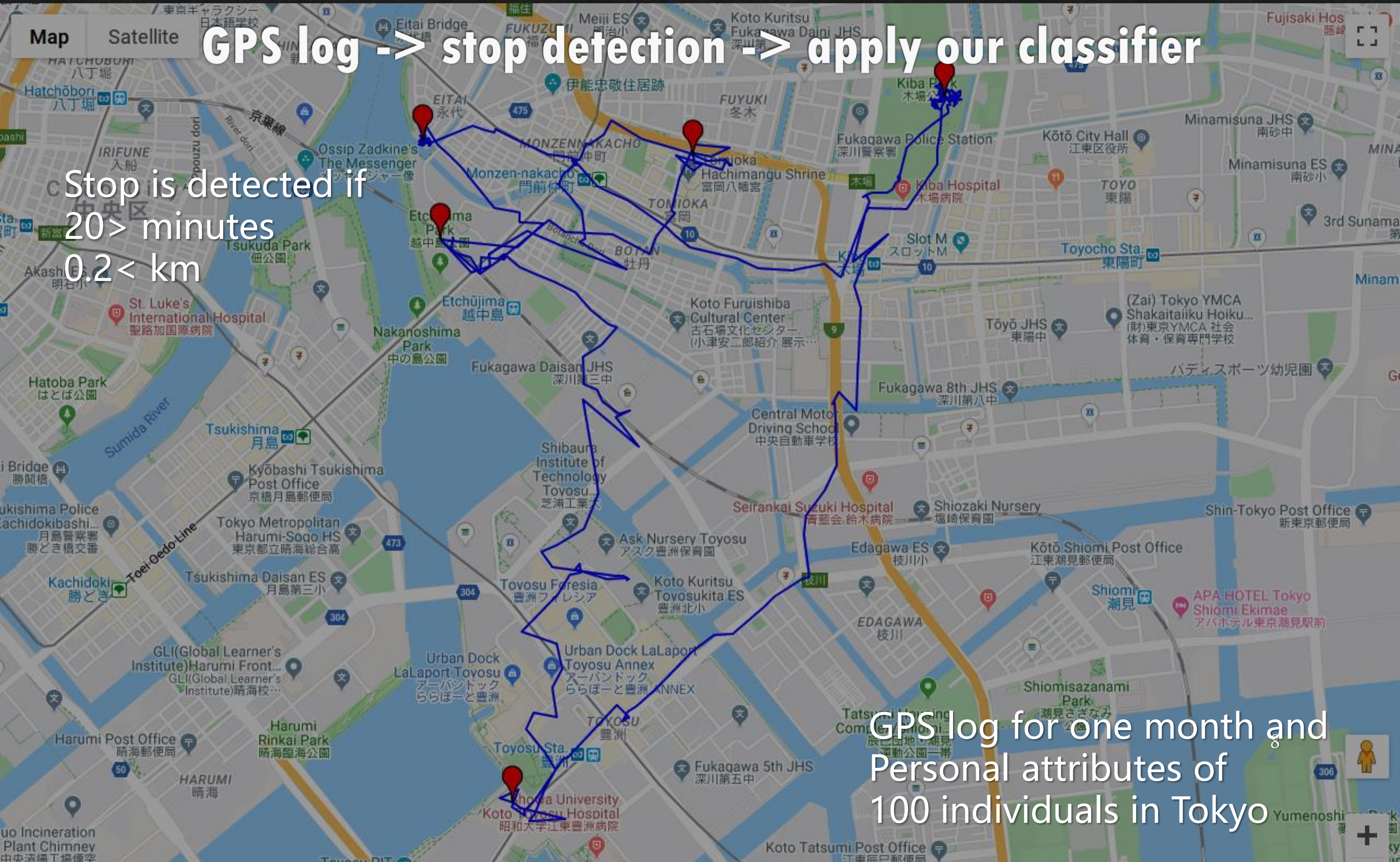
Representative Population?



Future Work: GPS Log

GPS log -> stop detection -> apply our classifier

Stop is detected if
20> minutes
0.2< km



GPS log for one month and
Personal attributes of
100 individuals in Tokyo

Contents

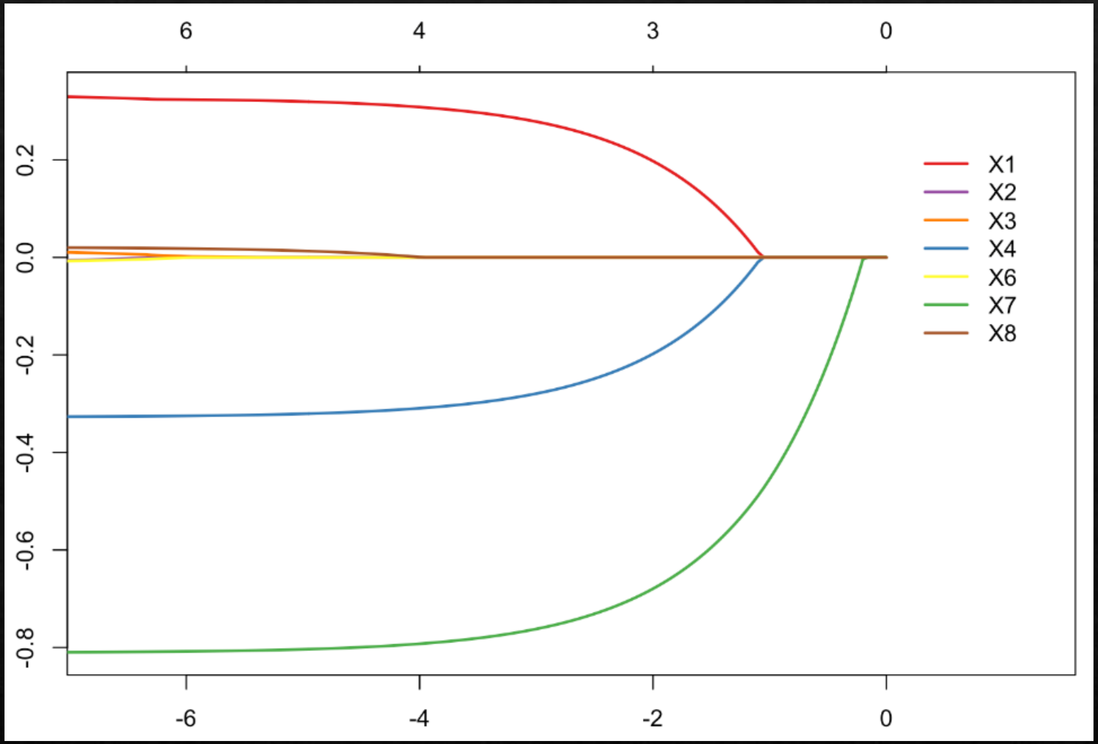
1. Motivation
2. Data Description
3. Methodology
    - Machine Learning
        - Lasso
        - Naive Bayes (Gaussian)
        - Random Forest
        - XG-boost
        - Light GBM
        - SVM
4. Results
5. Conclusion

Logistic Lasso

Logistic regression with L1 regularization

-> shrinking coefficients

-> avoiding over-fitting & selecting important features



Naive Bayes (Gaussian)

According to Bayes rule, the probability of personal attribute, y, conditional on location information, x:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

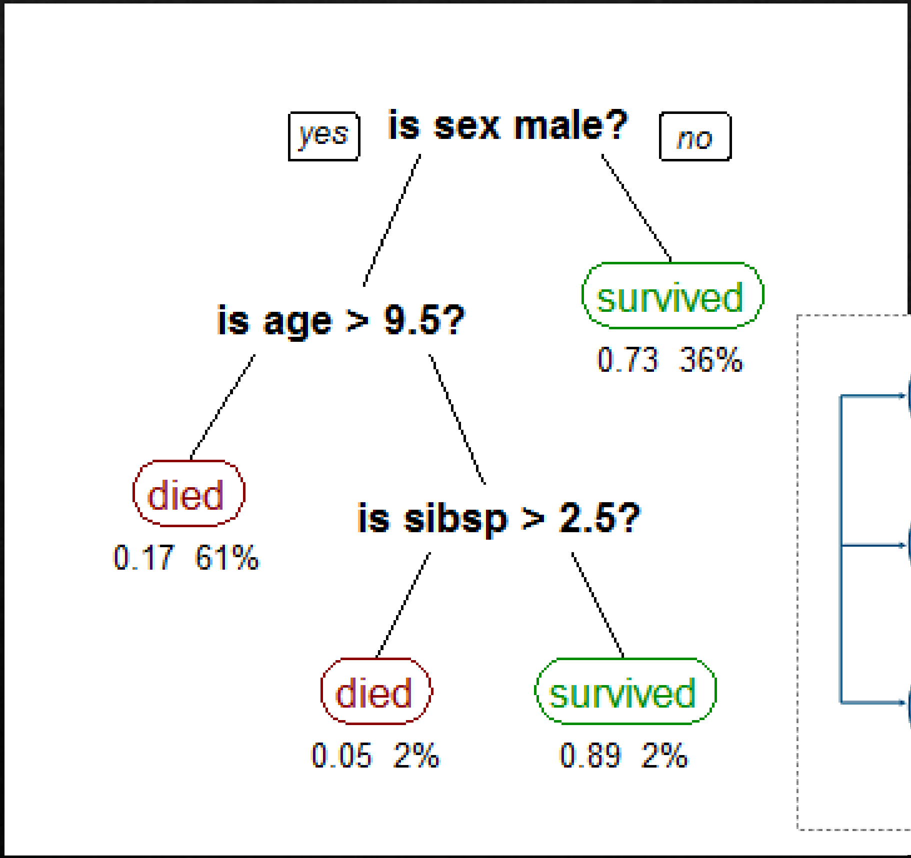
Elements of location history, $(x_1, \dots, x_n)$, are independent:

$$p(x|y) = \prod p(x_i|y)$$

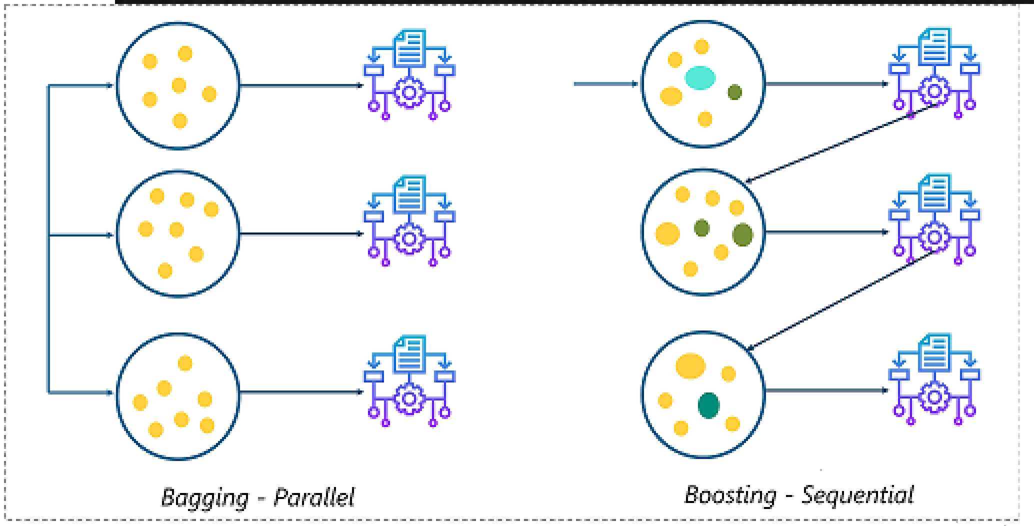
Each location, x_i, is normally distributed:

$$p(x_i|y) = \phi\left(x_i \middle| \mu_y, \sigma_y\right)$$

Random Forest, XGBoost, LightGBM

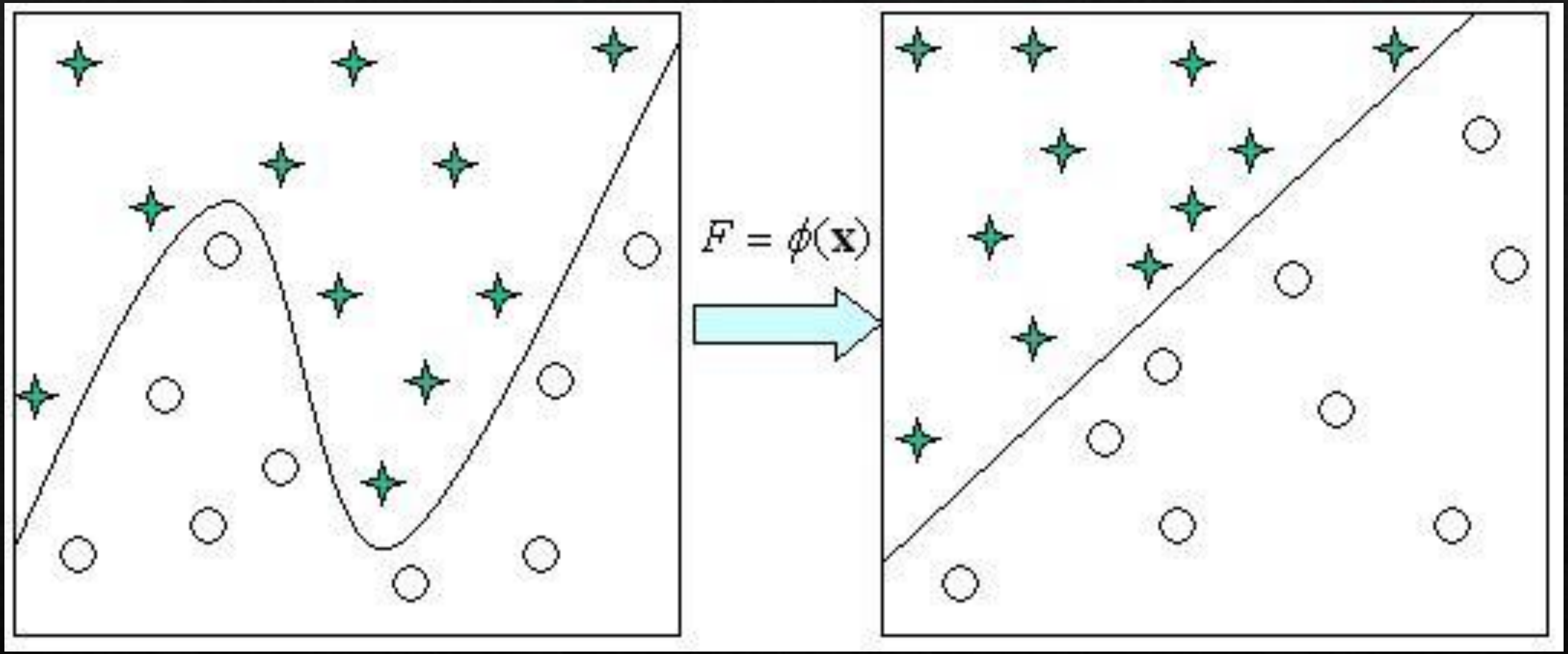


Prediction by many weak decision trees



In boosting, weak trees are generated based on previous failures.

Support Vector Machine (RBF)



SVM finds (hyper)plane separating sample into positive and negative ones.

Kernel trick expand the dimension of feature space to improve prediction.

目次

1. Motivation
2. Data Description
3. Methodology
4. Results
    1. Cross-validation and SMOTH
    2. Learning Evaluation Functions
    3. Gender
    4. Age
    5. Overall
5. Conclusion

ROC AUC, PR AUC, MCC

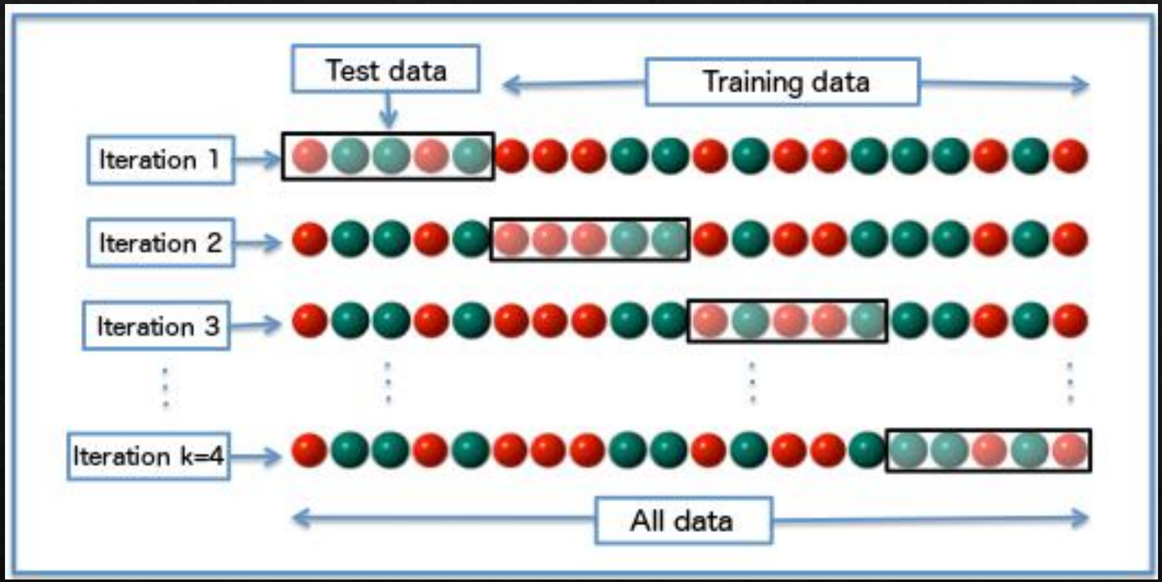
Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Imbalanced data -> Accuracy is unreliable

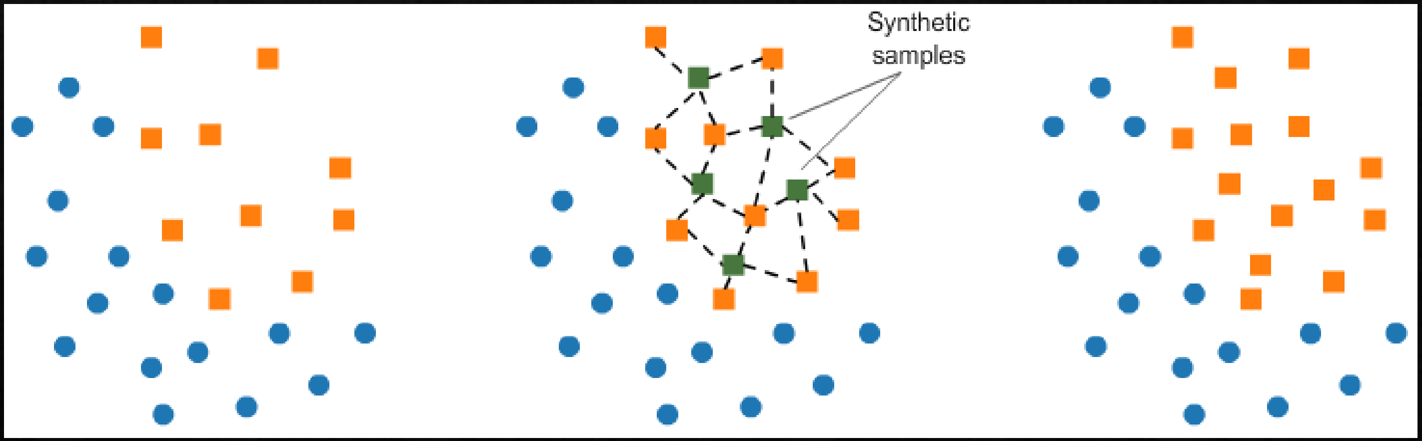
Matthews coefficient: correlation of confusion matrix

Cross-validation, SMOTE

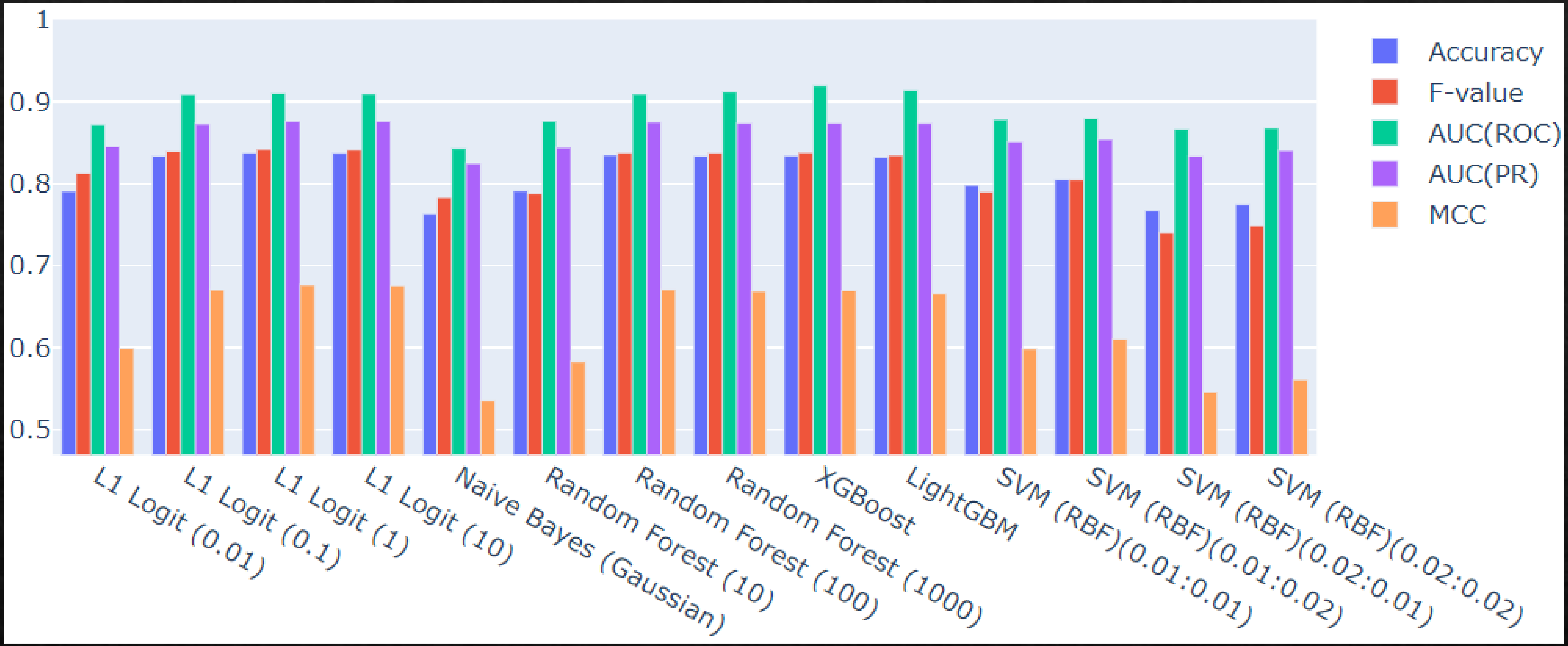


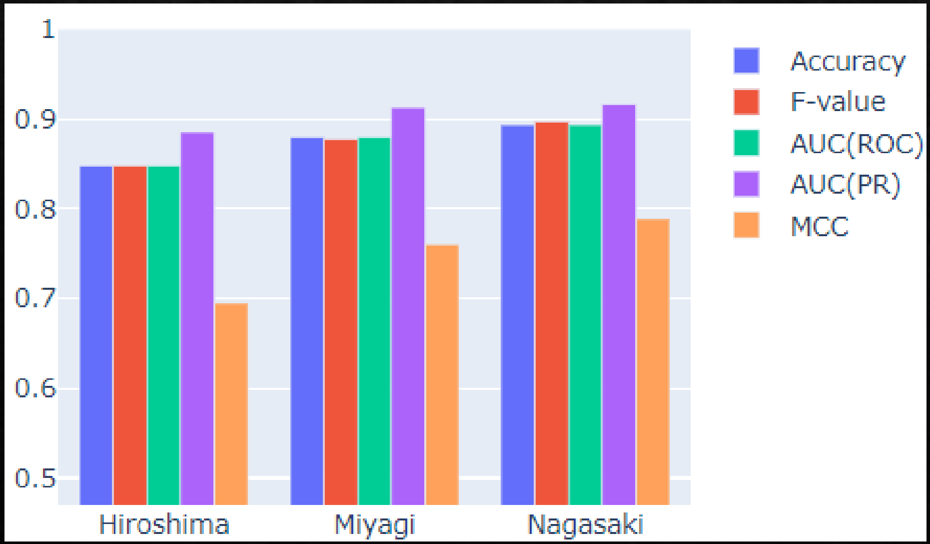
5-fold CV to evaluate performance

Up-sampling by SMOTE

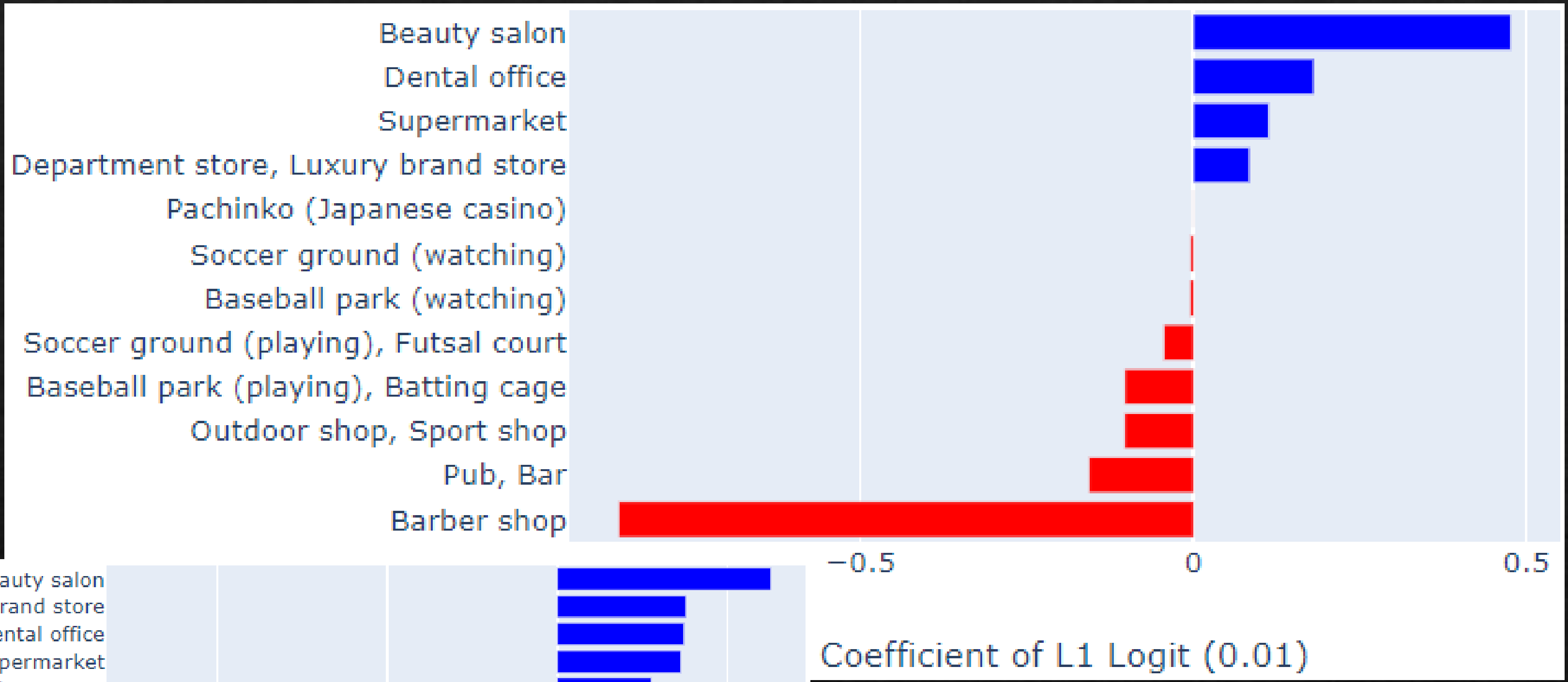


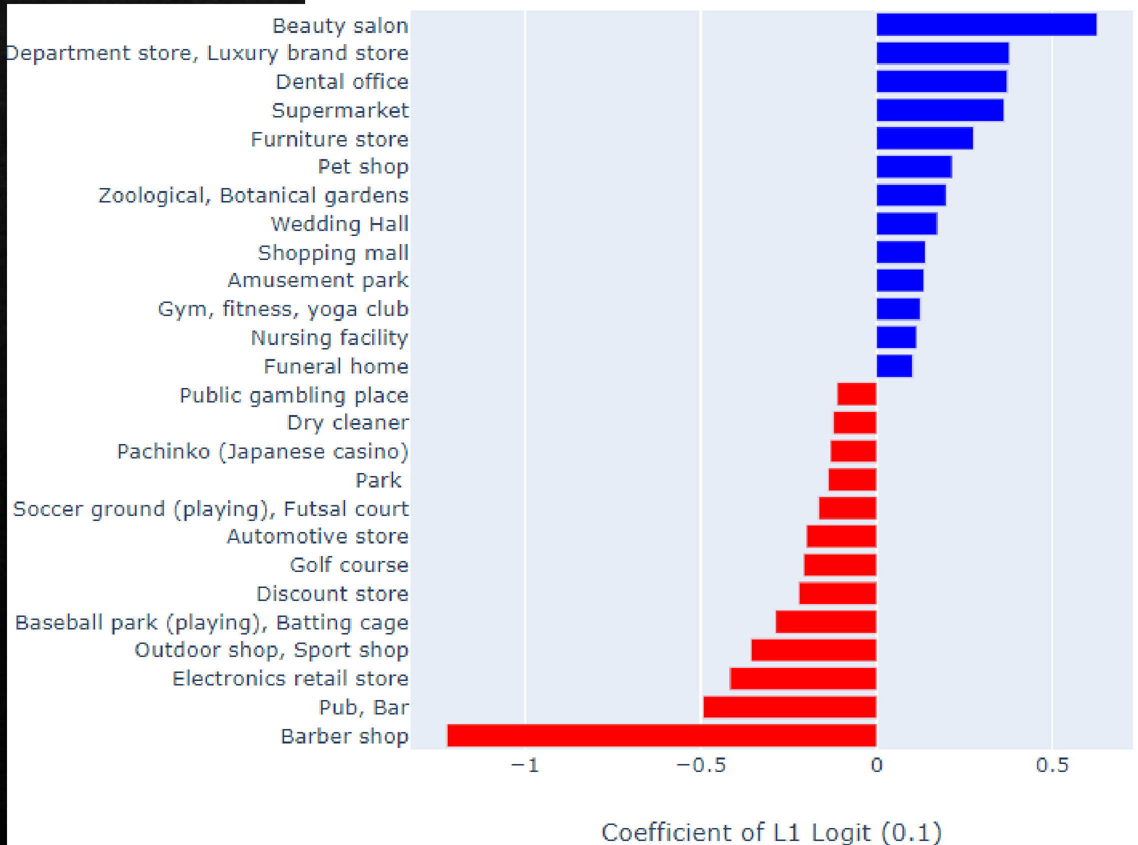
Gender





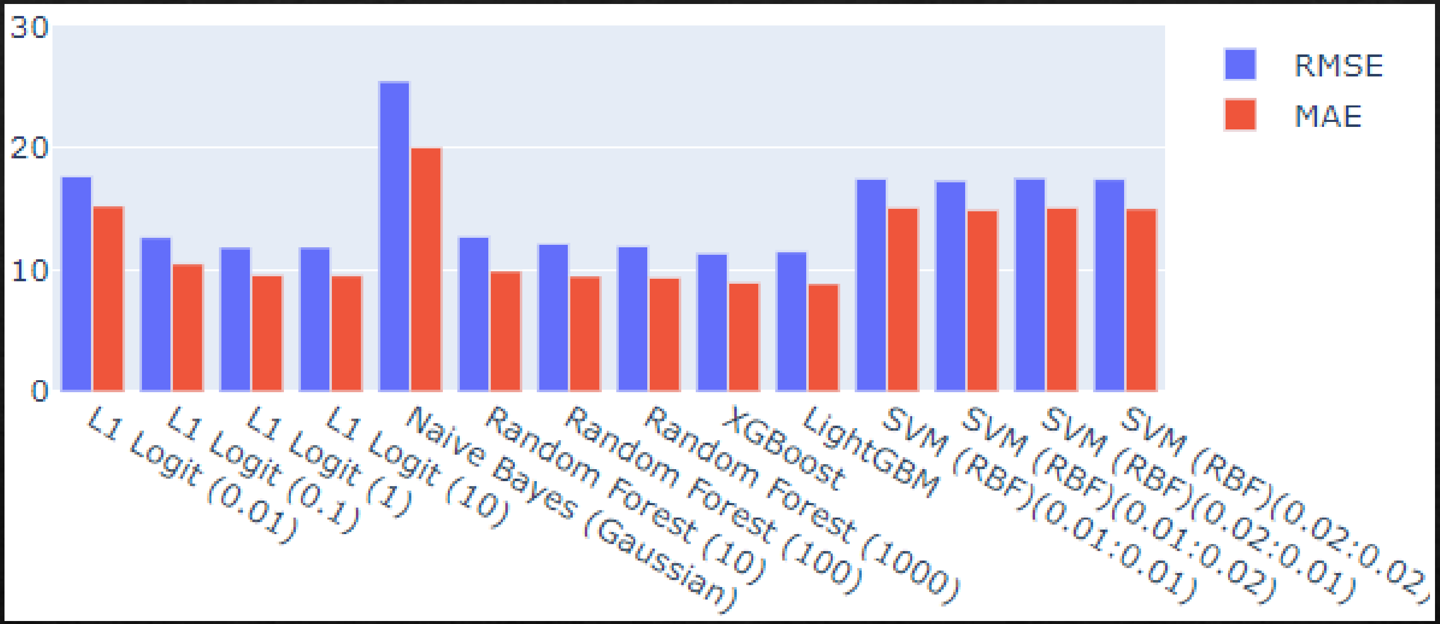
XG-Boost
Train: 3000 in Tokyo
Test: 400 in Miyagi
400 in Hiroshima
160 in Nagasaki

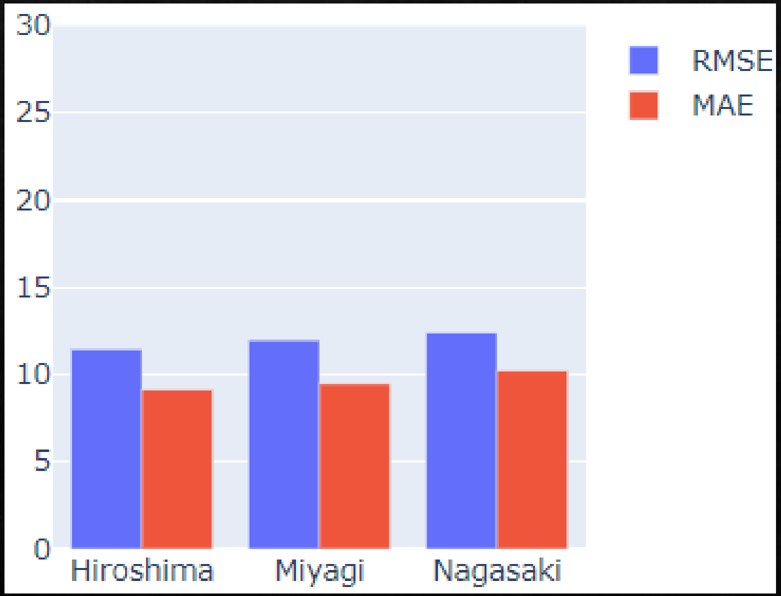




Important Features for predicting gender

Age





XG-Boost

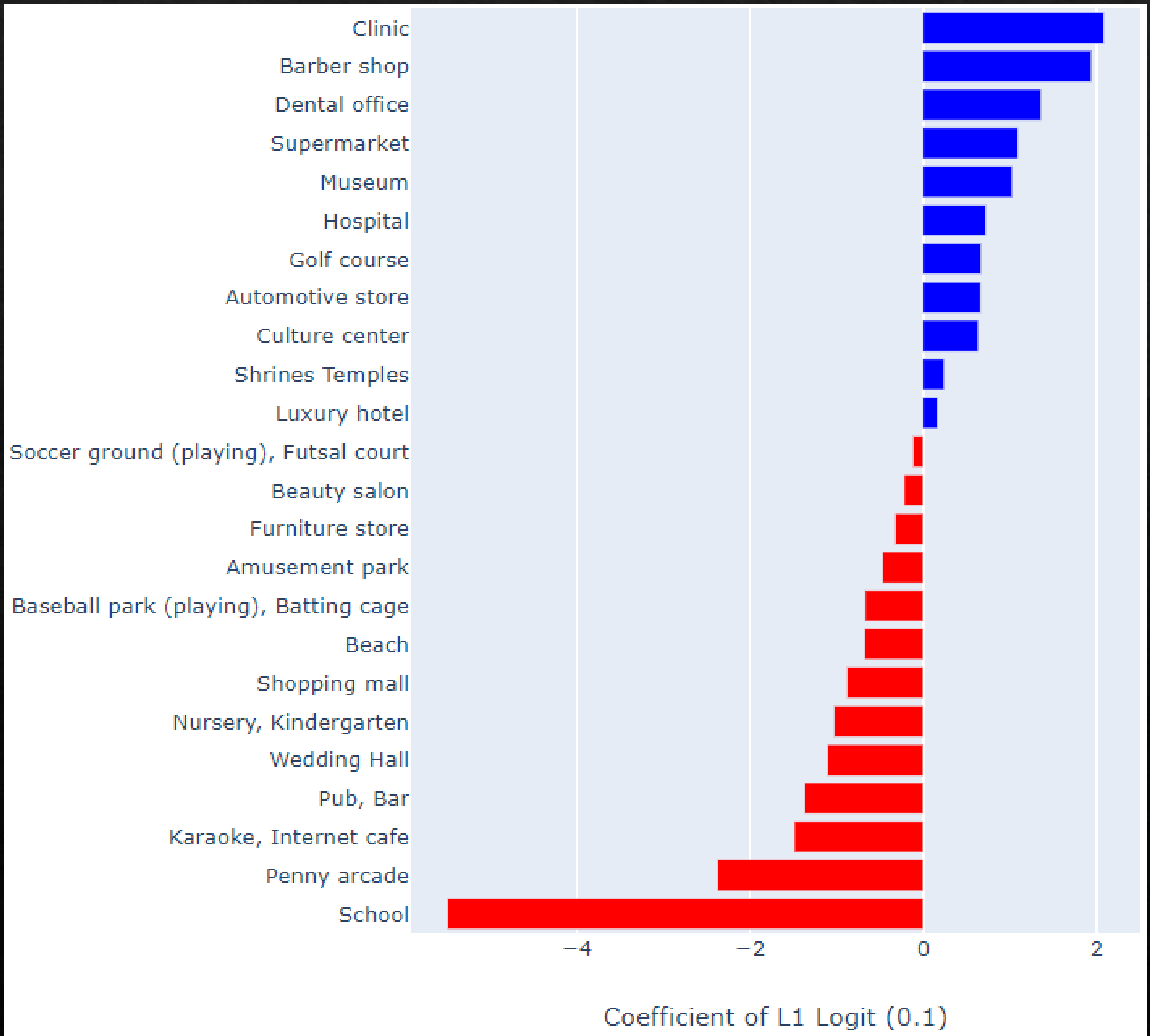
Train: 3000 in Tokyo

Test: 400 in Miyagi

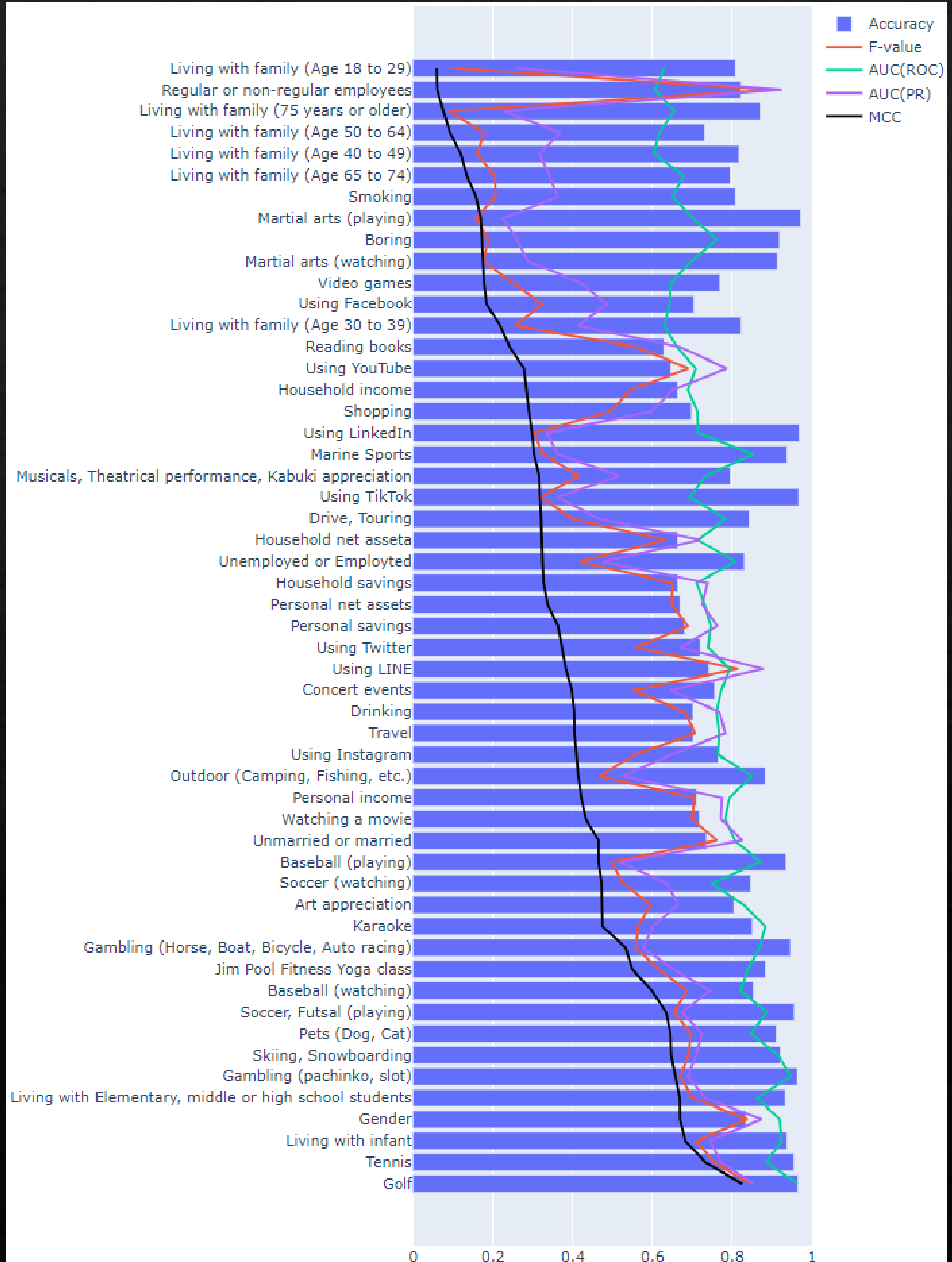
400 in Hiroshima

160 in Nagasaki

Important Features for Predicting Age



Overall Performance



Conclusion

- We estimate personal socioeconomic attributes from location information.
  - Gender is predicted as accurately as existing studies using other information (accuracy is around 85%).
  - Hobbies which requires specific facilities are well predicted.
  - Whether they have infants and children or not is predictable while whether they live with adults and elderly people is not.
  - Other attributes not explicitly related to location, like income , use of web apps and indoor activities are hard to estimate.
- In a future work, we apply the classifiers developed in this study to actual GPS log data to estimate personal attributes.