Statistics Poland

Estimation of tourist expenditures

Sebastian Wójcik,

Statistical Office in Rzeszów, Statistics Poland

Tourist expenditures

Statistics Poland conducts a quarterly-based households sample survey on participation of Polish residents in trips.

Due to level of detail of a questionnaire and low subsample size for some destination countries, the survey suffers from

- a high variability of expenditures in a sample
- item non-response problem pertaining to expenditures by expenditure category.

Tourist expenditures

With Big Data sources we are able to estimate the number of trips on the low level of aggregation. To use these results in regular production we need to produce whole records to put it into database.

The records consist of

- descriptive variables of trips such as destination country, purpose of trip, mean of transportation, type of accommodation
- expenditures by expenditure category

Tourist expenditures

As explanatory variables, eleven categorical variables were used:

- destination country
- purpose of trip
- mean of transportation
- type of accommodation
- if the accommodation is booked via Internet
- nights spent
- if the trip is organised or not
- year and quarter, etc.

All of these variables had no missing values.

Tourist expenditures

Aim of this study is to compare Machine Learning methods and statistical methods when dealing with missing values. To this end, five expenditure variables were selected:

- expenditure on accommodation
- expenditure on restaurants & café
- expenditure on transport
- expenditure on commodities
- other expenditures

Models tried

Non-Machine Learning models:

- Linear Model (OLS)
- General Linear Model (GLS)
- Robust Linear Model
- LARS

Machine Learning models:

- CART
- Random Forest
- Optimal Weighted Nearest Neighbour
- Support Vector Machine (linear and radial kernel)

Model selection

Mixture of Bootstrapping and K-fold Cross Validation was used to find the optimal set of hyperparameters with respect to RMSE for each tested model.

Tuning was carried out in the following way:

- Draw B samples without replacement with size amounting 90% of size of the dataset.
- Train model with a given set of hyperparameters
- Make predictions on the 10% remaining cases.
- Calculate error statistics
- Average error statistics over all B draws.

Results

Method (R function)	MAE	MAPE	RMSE	R2
Linear Model OLS (lm)	92.62	1.4003	183.05	0.225
General Linear Model GLS (glm)	92.62	1.4003	183.05	0.225
CART (rpart)	94.15	1.4636	185.79	0.203
Robust Linear Model (rlm)	85.64	1.0406	188.56	0.216
LARS (lar)	93.26	1.4766	184.35	0.217
Random Forest (randomForest)	92.01	1.3507	185.13	0.212
Optimal Weighted Nearest Neighbour (kknn)	86.15	1.2400	171.13	0.329
Support Vector Machine (svm) radial kernel	91.25	0.9569	202.84	0.172
Support Vector Machine (svm) linear kernel	86.03	0.9381	192.45	0.203

Distributional accuracy

- In a case of selected model, predictions were slightly biased - mean prediction was 2% higher than true mean. Other methods did not produced significantly biased predictions (bias up to ±0.5%).
- Based on Kolmogorov-Smirnoff test, the distribution of predictions significantly differed from the true one for every model (p-value < 10-16).
- Other descriptive statistics were not calculated

Other conclusions

- All tested machine learning methods produced plausible predictions.
- Traditional regression models produced negative values except LARS which gives several sets of prediction and the final set can be selected with respect to non-negativity condition.
- The machine learning methods can deal with “singular” problems since they do not take into account any correlations.
- Machine learning methods are much more powerful than traditional models and they can easily overfit to the dataset. Therefore, estimating the out-of-bag error is one of the relevant way to compare various methods by bootstrapping or cross validation.

Statistics Poland

Thank you for your attention!