Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Predicting Emigration in the MENA region
Lab Report

MACIS
ETH Zurich

Core Seminar: Methods III
Spring Semester 2019
10. 06. 2019

Lecturer: Dominik Hangartner
Tutorial Instructor: Moritz Marbach
Graduate Teaching Assistant: Eroll Kuhn

Fride Sigurdsson
Matriculation number: 18 - 941 - 112

Jamila Issa
Matriculation number: 18 - 950 - 287

Mirjam Grünholz
Matriculation number: 13 - 213 - 350

The aim of this lab report is to predict emigration using survey data from the Wave IV of the Arab Barometer that includes responses from seven countries located in the MENA region.

Data Preprocessing

In order to do that, the preprocessing of the data is a vital step in ensuring high predictive performance and so a number of transformations were made before training any models. Initially, the types of variables were identified, and a subset of both types - ordinal categorical and non-ordinal categorical - was created. The ordinal non-categorical variables were altered from factor to numeric and standardized. This was to ensure better predictability by assuming an ordinal scale that moves linearly along responses such as "I strongly agree", "I agree", etc. The standardization for the numeric variables was made to ensure the same scale along different models and variables. Emig was kept as factor variable. Secondly, a correlation check between the outcome variable (emig) and the rest of the numeric variables was conducted. No correlation higher than 0.299 was detected, so no respective changes were made. Moreover, variables q1 (governorate) and q2 (district) were excluded from the subset as they might lead to strong overfitting of the models. Once all those transformations were done, the numeric and categorical subsets were merged into a full dataset. The remaining step to the preprocessing is the train and test split. The chosen split has 4620:1260:1260 observations in the train, test1, and test 2[1] splits respectively. The reason this split was chosen is because by making both test splits the same, low bias (high accuracy of the training model) and low variability (which affects how consistent the predictions are across the models) are ensured[2]. Additionally, this split reflects the size of the hold-out sample that our instructors will be testing our model on.

Models

The models that were used to predict emigration were logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic GAM, lasso, tree-based models (boosting, and bagging, boosted classification tree), and support vector machines (linear, polynomial, and radial). Cross-validation with 10 folds was applied to all the models. The performance of the models was evaluated by the area under the curve (AUC) for both the train split and the out-of-sample performance. The results can be found in Table 1.

Model	Parameters/ specification	AUC on train	AUC on test1	AUC on test2
Logistic reg.		0.7954	0.8370	
LDA		0.7966	0.8330	
QDA		0.7418	0.8494	
Logistic GAM	df = 1	0.7826	0.8278	
Lasso	α = 1 and λ= .003	0.8108	0.852	
Bagging	mtry = 12	0.9225	0.9569	0.9217
Boosting	mtry = 161	0.9188	0.9521	
Boosted Cl. Tree		0.9106	0.9532	
Linear SVM	cost = .01	0.7894	0.8299	
Radial SVM	sigma = .01, cost = 2	0.9174	0.9531	
Poly SVM	degree = 3, scale = .5, C = .01	0.9168	0.9532	

Table 1: Results for the train and test1 split for all the models and the result for the test2 split for the best model

Overall, the best model seems to be the bagging model with an AUC of 0.96 in the test1 split. However, an unbiased estimate of its predictive power can only be achieved once this model is tested on unseen data such as the test2, which resulted in an AUC of 0.9217498 or on the unseen evaluation data.

The most important variables in the bagging model to predict emigration are: age, time lived in the area (q1991c), marital status unmarried (q1010), non-use of internet (q409) and participation in a Facebook group (q4113).

[1] Test 1 refers to the subset that is used to calculate the out-of-sample performance, whereas the test 2 split refers to the subset that is used once the best performing mode is identified and ensures an unbiased estimate of the predictive performance.

[2] This was confirmed by trying other possible data splits as well. We are aware that bias and variance have a trade-off relationship, but the chosen split ensures an efficient level of both.