Classification of Disadoption of Water Efficient Technologies in Rural Costa Rica

Marisa L. Henry[1], Tianshi Liao[4], Shashank Agarwal[2,3], Xinhan Luo[2]
[1]Department of Environmental Health and Engineering, [2]Department of Applied Mathematics and Statistics,
[3]Center for Leadership Education, [4]Department of Computer Science

10 May 2018

INTRODUCTION

The data in this analysis comes from an economic field experiment conducted in 2015-2016 in Guanacaste, Costa Rica. Households in the area were randomly assigned to one of three groups: (1) *control,* which received no new technologies or money; (2) *treatment with bonus,* which received water efficient technologies (faucet aerators and efficient shower heads) for their houses and a monetary bonus of ~20 USD to keep the technologies installed; or (3) *treatment without bonus,* which received only the water efficient technologies. An audit and survey were conducted at baseline and 16 months after installation.

Our goal is to use information about the households collected during the audit/survey to classify households into three categories: those that totally disadopt all technologies; those that disadopt *some* of the technologies; and those that keep all the technologies.

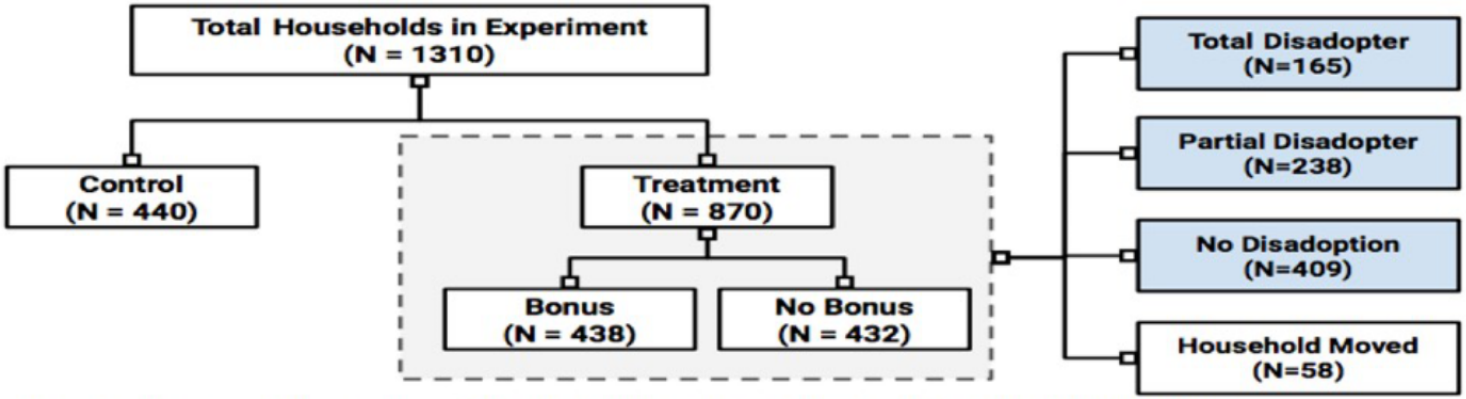


Figure 1. Schematic of the experimental design with the outcome classes of interest highlighted in blue

Our team is thankful to Dr. María Bernado Del Carpio (Assistant Professor at University of Maryland Baltimore County, Department of Economics) **and Dr. Paul Ferraro** (Bloomberg Distinguished Professor of Business and Engineering at JHU) **for providing access to this dataset.**

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING
JOHNS HOPKINS CAREY BUSINESS SCHOOL
UNIVERSITY OF MARYLAND Department of Economics

FEATURE ENGINEERING

Data Cleaning

The data was received in a very large, unorganized file, and all of the variables were labeled in Spanish. Many of the columns in the file were not useful for our analysis, e.g. linear combinations of several columns, large amounts (>75% missing data), or identification information, so they were removed. Using documents on how the data was collected, each feature was identified and renamed in English.

Missing Data

Many of the features in the dataset had missing data. To use *sklearn,* we had to fill in the missing values. Several techniques were used depending on the amount of missingness:

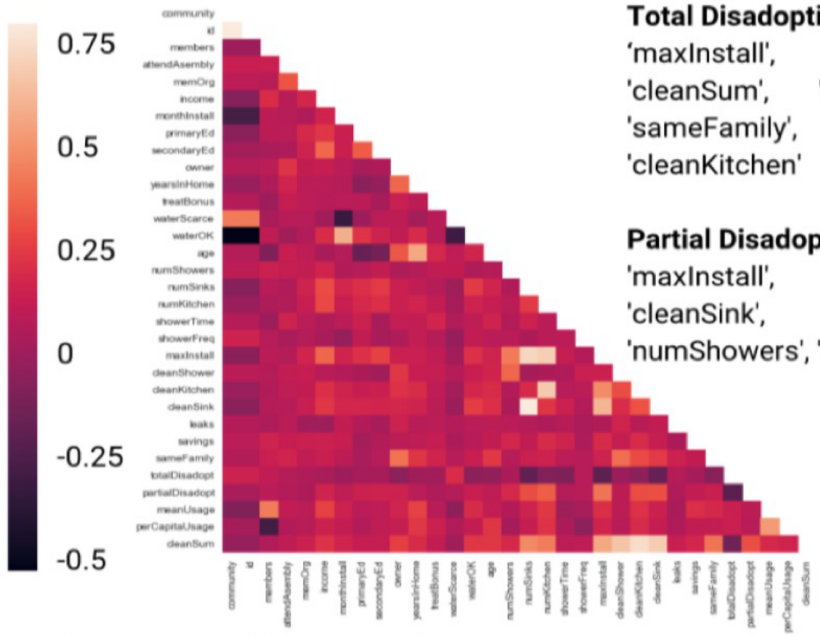
Table 1. Methods used for filling in missing feature values.

Amount of Missingness	Fill Value
>15 %	Change to binary, s.t. 1 = missing, 0 = not missing
between 1% and 15%	Random Forest Regressor
<1 %	Mean of feature

Random Forest Regressor(RFR) is used for its ability to capture complex relationships within data and good generalization on prediction due to bagging. We exploited top 5 correlated features with regard to each feature with moderate amount of missing values and run RFR to fit the portion with value and then predict the portion with missing value as fill-ins.

Feature Selection

To visualize correlations between features, we used a heatmap. We selected the top 10 features correlated with each outcome of interest to run our classification methods.



Total Disadoption

'maxInstall', 'numSinks', 'cleanSink', 'cleanSum', 'showerTime', 'numKitchen', 'sameFamily', 'waterOK', 'owner', 'cleanKitchen'

Partial Disadopt

'maxInstall', 'numKitchen', 'cleanSum', 'cleanSink', 'numSinks', 'cleanKitchen', 'numShowers', 'income', 'secondaryEd', 'owner'

Figure 2. A heatmap of the correlations between feature variables, after missing data is filled in.

CLASSIFICATION

We tested using 5 different classification techniques: k-nearest neighbors, decision tree, random forest, Gaussian naive bayes, and support vector machine. For each, we ran a one-v-all classification.

KNN

We tested prediction using from 3 to 9 nearest neighbors. Results were quite similar, with the best results achieved using 4 nearest neighbors for total disadopt and 9 for partial disadopt.

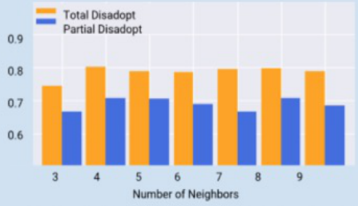


Figure 3. Comparison of results of k-nearest neighbors using different number of neighbors.

GNB

Gaussian Naive Bayes performs the worst among all the models because it assumes independence between the features , which is not the case with our data set.

SVM

Support vector machine performed really well on the data set. The fact that it uses regularization, SVM tends to resist over-fitting.

SVM

Support vector machine performed really well on the data set. The fact that it uses regularization, SVM tends to resist over-fitting.

Decision Tree

We selected the best depth of the decision trees based on the maximum cross validation score. The best depth was 3 for the total disadoption prediction and 5 to predict partial disadoption. Below is the tree for total disadoption.

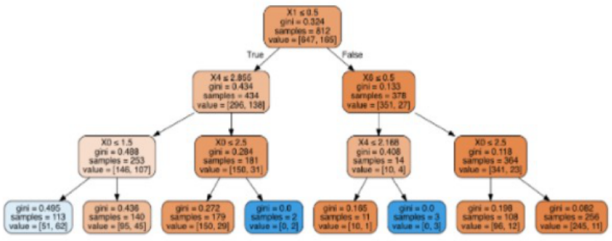


Figure 4. A decision tree used for classification of total disadoption.

Random Forest

Random forest with 20 trees in the random forest performed the best in terms of accuracy for total disadopt and that with 80 trees performed the best for partial disadopt.One possible reason for such a good fit for the random forest model can be the fact that most of the features in our model are categorical.

CONCLUSIONS & FUTURE WORK

After cleaning the data and filling in missing values we used the dataset to successfully classify households into total and partial disadopters. KNN and Random Forest performed particularly well in classification. Future research will be done on how data mining can be used in heterogeneity analysis in economic field experiments.

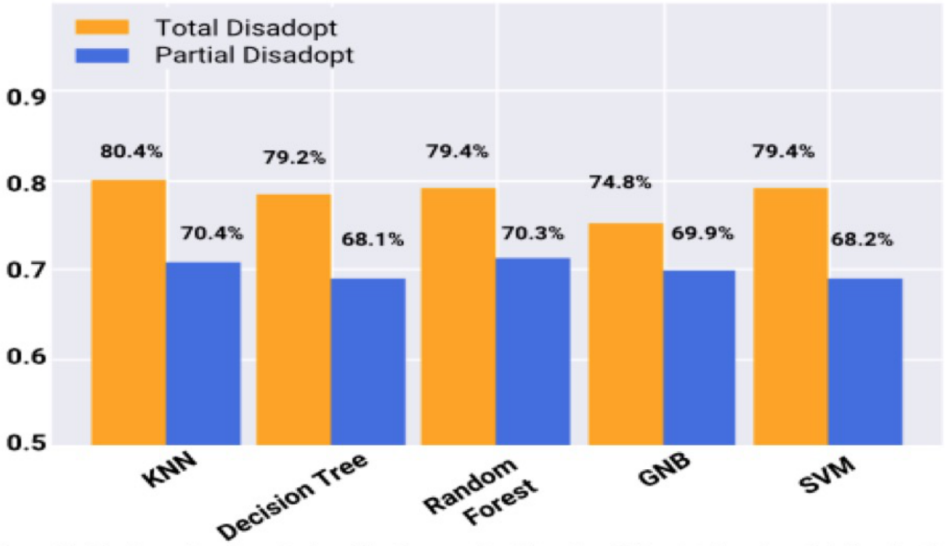


Figure 5. Final results of each classification method for classifying total and partial disadoption.