Tobacco spending in Georgia: Machine learning approach

Maksym Obrizan[a*], Karine Torosyan[b], Norberto Pignatti[c]

[a]*Kyiv School of Economics, Kyiv, Ukraine*
[b]*International School of Economics at TSU, Tbilisi, Georgia and Global Labor Organization*
[c]*International School of Economics at TSU, Tbilisi, Georgia and IZA Institute of Labor Economics*

ABSTRACT

The purpose of this study is to analyze tobacco spending in Georgia using various machine learning methods applied to a sample of 10,757 households from Integrated Household Survey collected by GeoStat in 2016. Previous research has shown that smoking is the leading cause of death for 35-69 year olds. In addition, tobacco expenditures may constitute as much as 17% of the household budget. Five different algorithms (ordinary least squares, random forest, two gradient boosting methods and deep learning) were applied to 8,173 households (or 76.0%) in the train set. Out-of-sample predictions were then obtained for 2,584 remaining households in the test set. Under the default settings random forest algorithm showed the best performance with more than 10% improvement in terms of root-mean-square error (RMSE). Improved accuracy and availability of machine learning tools in R calls for active use of these methods by policy makers and scientists in health economics, public health and related fields.

Keywords: Tobacco Spending, Household Survey, Georgia, Machine Learning

1. Introduction

Earlier estimates of the very high premature mortality from smoking-related conditions indicated high smoking rates prevailing in the former Soviet Union. Peto et al. (1992) projected that during the 1990s tobacco will cause about 30% of all deaths at 35-69 years of age making it the largest cause of premature death, with about 5 million deaths occurring in the former USSR.

After the collapse of the Soviet Union smoking prevalence reduced in some of the former USSR republics. For example, a study of 8 countries of the Former Soviet Union reports a significantly lower smoking prevalence among men in 2010 compared with 2001 for 4 countries but not Georgia (Roberts et al., 2012).

Even nowadays smoking remains a serious health problem in Georgia. A national household survey of 1,163 adults in Georgia in 2014 reports smoking prevalence rates of 54.2% in men and of 6.5% in women (Berg et al. 2016). Secondhand smoke exposure (SHEs) is also quite high with 54.2% reporting SHEs at home.

Smoking is not only bad for health but also for household budget. Wang, Sindelar and Busch (2005) report a range of country estimates of total expenditures on tobacco. For example, US is on the lower bound in

[*] *Corresponding author.* E-mail address: mobrizan@kse.org.ua

terms of percent with nearly 4% of average expenditures being spent on tobacco product in smoking households (Busch et al. 2004). In monetary terms, however, this amount is more than $1,000 annually. China is on the upper bound of the range in terms of percent with current smokers spending 60% of personal income and 17% of household income (Gong et al., 1995).

In this paper we aim to provide the most recent evidence on household tobacco spending in Georgia using Integrated Household Survey collected by GeoStat in 2016. In addition to standard linear regression we employ a number of recent machine learning methods. From a policy and scientific perspectives it is important to select the estimation methods which give the best performance in terms of out-of-sample predictions of household spending on tobacco products.

The rest of the paper is organized as follows. First, we describe the GeoStat dataset used in this study. We provide descriptive statistics on key indicators of tobacco spending by Georgian households. Next, we develop a linear regression model which serves as a natural benchmark for future comparisons. Finally, we apply machine learning methods and compare their performance with OLS in terms of root-mean-square error (RMSE). The last section concludes.

2. Data and Methods

2.1 GeoStat data

Data on tobacco expenditures come from GeoStat file "tblShinda03_1" which includes daily data on consumption by household over week. The consumption data includes the amounts paid for different products defined by GeoStat version of COICOP. Total tobacco spending is defined by combining household expenditures on filter cigarettes (COICOP code of 22111), unfiltered cigarettes (22112), cigars (22121), tutuni (22131) and other tobacco products (22132). Share of tobacco expenditures in total household expenditures is computed by dividing expenditure on tobacco by total household expenditure in the same period.

The final dataset for 2016 includes information on 10,757 households which are divided for the purposes of this study into 8,173 households (or 76.0%) in the train set and 2,584 households in the test set. The train set is used for model building while the test set is employed for out-of-sample predictions for model comparison. 2,665 out of 5,508 (or 32.6%) households in the train set report non-zero expenditures on tobacco products. 876 out of 2,584 (or 33.9%) households in the test set report non-zero expenditures on tobacco products. The equality of means test cannot reject the null of equal means share of households reporting non-zero amounts on tobacco products. It is also reassuring that the share of households with non-zero tobacco expenditures is very close to estimated smoking prevalence rate reported in other studies (such as 36% provided in Bakhturidze, Peikrishvili, and Mittelmark 2016).

Fig. 1. in the Appendix shows tobacco spending patterns in Georgia for households with smokers based on IHS 2016. While the majority of such households spend less than 20 GEL (approximately 7.60 Euros) per week on tobacco products (top right graph), this relatively small amount may constitute as much as 40 or more percent of household income for sizable chunks of population (top left graph). Households with higher expenditures have lower share of tobacco spending (bottom left graph). Finally, higher tobacco expenditures in GEL imply (weakly) higher share spent on tobacco products by a household.

2.2 Benchmark OLS regressions

We begin with benchmark models estimated via Ordinary Least Squares (OLS). We assess spending on tobacco products along three different dimensions. First, we estimate a linear probability model for a discrete variable "Any Tobacco Spending" taking value of 100 if household reports any tobacco expenditures and 0 otherwise. Second, we estimate via OLS a model for "% of Tobacco in Spending" capturing household expenditures on tobacco products (including households with 0 expenditures on tobacco). Finally, we estimate a linear regression for "Tobacco Spending, GEL" capturing expenditure on tobacco products in local Georgian currency (also including households with 0 expenditures on tobacco).

Since expenditure data are only available at the household but not the individual level we have to be creative in defining potential correlates of tobacco expenditures. Specifically, for most variables we use average shares of household characteristics such as age, marital status, attained education level, profession. For example, variable "Secondary special school" is the average proportion of individuals in the household who have completed secondary special school. This variable will be equal to 2/3 for a households with 3 members 2 of whom completed secondary special school. In addition, we include number of males in the household, household size, measures of household income, characteristics of the local job market, region of Georgia and dummies for quarters 2, 3 and 4.

Results in Table 1 indicate that tobacco spending increases in average household age (but at a diminishing rate) and with a higher share of married members; for better educated households (somewhat surprisingly); for certain occupations; in Kvemo Kartli and Samtskhe-Javakheti regions. Spending on Tobacco is lower for Azerbaijani households and households with fewer members, households with internally displaced people (IDP) and recent movers (relative to those who al-ways live in the same place); for rural households and in Adjara a.r., Guria and Imereti and Racha-Lechkh regions. To provide one illustration, 1 additional male in the household is associated with 7.4 percentage points higher probability of "Any Tobacco Spending", 2.0 percentage points higher share of tobacco spending in total expenditures and 1.5 additional GEL spent on Tobacco. Other coefficients can be interpreted similarly.

Table 1. Descriptive statistics and OLS regressions for tobacco spending in Georgia

Variable	Mean (standard deviation)	Any Tobacco Spending (0/1)	% of Tobacco in Spending	Tobacco Spending, GEL
HH mean age	46.103	0.963***	0.312***	0.084**
	(17.203)	(0.175)	(0.050)	(0.035)
HH mean age squared	2421.387	-0.012***	-0.004***	-0.001***
	(1741.145)	(0.002)	(0.000)	(0.000)
Males in HH	1.709	7.360***	1.966***	1.429***
	(1.165)	(0.810)	(0.232)	(0.161)
Share of Azerbaijani in HH	0.072	-4.476*	-0.221	-1.485***
	(0.258)	(2.597)	(0.744)	(0.518)
Share of Married in HH	0.485	6.285***	0.690	1.391***
	(0.337)	(2.409)	(0.690)	(0.480)
HH size	3.516	-1.144**	-0.729***	-0.172
	(1.894)	(0.560)	(0.160)	(0.112)
Share of IDPs in HH	0.029	-4.583	-1.740*	-1.475**
	(0.155)	(3.464)	(0.992)	(0.691)
Share of Movers in HH	0.564	-6.852***	-1.604***	-1.011***
	(0.316)	(1.921)	(0.550)	(0.383)
Secondary special school	0.362	7.294***	2.690***	1.019***
	(0.350)	(1.907)	(0.546)	(0.380)
Hand craft school	0.075	7.620**	3.008***	0.568
	(0.182)	(2.991)	(0.857)	(0.596)
Higher education	0.032	17.932***	5.098***	1.861**
	(0.123)	(4.214)	(1.207)	(0.840)
Senior officials and managers	0.014	17.228**	3.324	8.566***
	(0.073)	(7.458)	(2.137)	(1.487)
Technicians and associate professionals	0.034	9.588**	1.510	3.405***
	(0.117)	(4.661)	(1.335)	(0.929)
Plant and machine operators	0.021	23.545***	4.876***	6.046***
	(0.089)	(5.613)	(1.608)	(1.119)
Elementary occupations	0.028	13.381***	3.497**	2.244**
	(0.109)	(4.908)	(1.406)	(0.978)
Share of public sector employees	0.071	-8.467**	-2.605**	-2.048***
	(0.172)	(3.721)	(1.066)	(0.742)
Labor inc., other HH adults	0.023	60.510***	-3.192	23.953***
	(0.043)	(14.336)	(4.107)	(2.858)
Non-lab income, HH	3.711	0.191**	0.013	0.111***

	(5.878)	(0.080)	(0.023)	(0.016)
Local LFP	0.700	80.201***	12.886***	12.160***
	(0.090)	(14.144)	(4.052)	(2.820)
Rural	0.614	-21.869***	-3.068***	-3.223***
	(0.487)	(3.102)	(0.889)	(0.618)
Shida Kartli	0.077	1.791	3.600***	-0.681
	(0.267)	(2.311)	(0.662)	(0.461)
Kvemo Kartli	0.115	7.960***	2.631***	1.623***
	(0.319)	(2.457)	(0.704)	(0.490)
Samtskhe-Javakheti	0.062	4.718	1.286	2.108***
	(0.242)	(3.097)	(0.887)	(0.617)
Adjara a.r.	0.074	-11.248***	-4.370***	-3.263***
	(0.262)	(2.538)	(0.727)	(0.506)
Guria	0.062	-7.040***	-2.550***	-2.424***
	(0.242)	(2.692)	(0.771)	(0.537)
Samegrelo-Zemo Svaneti	0.097	7.910***	0.444	-0.328
	(0.296)	(2.299)	(0.658)	(0.458)
Imereti, Racha-Lechkh	0.164	-6.612***	-1.961***	-1.780***
	(0.370)	(2.118)	(0.607)	(0.422)
		-33.569***	-6.357**	-5.923***
		(10.822)	(3.100)	(2.157)
Observations	10757	8173	8173	8173
Adjuster R-squared		0.113	0.083	0.112

*Notes: *p<0.1, **p<0.05, ***p<0.01. Models are estimated for 8,173 households in the train set. Independent variables represent share of individuals with certain marital status and education in the household. Only coefficients significant at 1% in at least one regression are shown to save space. All models also include share of Armenian, divorced, widowed, disabled individuals; share of individuals with no schooling or minimal education, 4 broad definitions of profession; local unemployment rate and households living in Tbilisi and Mtskheta-Mtianeti.*
Source: Authors' calculations based on GeoStat data.

While the OLS models are in general adequate their predictive power is quite limited as they tend to explain only 8.3 to 11.3% of variation in dependent variables. Hence, in the next section we turn to machine learning methods in order to improve predictive power of the models measured by RMSE in a test set of 2,584 households.

3. Machine learning approach to tobacco spending

Machine learning algorithm are widely used in business, scientific and policy applications nowadays. For the purposes of this study we will employ free R packages "h2o" (The H2O.ai team, 2017) and "xgboost" (Chen et al., 2018). The great advantage of "h2o" package is that it allows to run multiple machine learning algorithms from one R package. This makes machine learning tools accessible even to researchers with limited programming experience.

For example, just 9 lines of code will run random forest algorithm in R (provided train_data and test_data are already loaded):

```
library(h2o)
h2o.init()
trainHex <- as.h2o(train_data)
testHex <- as.h2o(test_data)
features<-colnames(train_data)[!(colnames(train_data)%in%c("Tobacco_","train_set"))]
RF_Hex <- h2o.randomForest(x=features, y="Tobacco_", train-ing_frame=trainHex)
pred.RF_ <- as.data.frame(h2o.predict(RF_Hex,testHex))
resid.RF_ = test_data$Tobacco_ - pred.RF_
```

```
RMSE.RF_ <- sqrt(mean(resid.RF_^2))
```

It is enough to replace line 6 with
GBM_Hex <- h2o.gbm(x=features, y="Tobacco_", training_frame=trainHex)
to run gradient boosting algorithm and so on.

Next we illustrate the application of machine learning algorithms to tobacco spending by Georgian households. Five different algorithms (ordinary least squares, random forest, two gradient boosting methods and deep learning) were applied to 8,173 households (or 76.0%) in the train set. In addition, we report the results of H2O's AutoML which is a simple wrapper function used for automatic training and parameter tuning within a pre-specified time. Out-of-sample predictions were then obtained for 2,584 remaining households in the test set.

Table 2. Relative performance of different machine learning algorithms in terms of RMSE for 2,584 households in the test set

Any Tobacco Spending (0/1)			
Model	RMSE	Δ RMSE vs OLS, %	Time, seconds
OLS	0.450	0.000	1.196
Random forest	0.402	-10.605	8.810
GBM	0.435	-3.162	4.704
Deep Learning	0.446	-0.763	12.298
XGBoost	0.436	-2.926	1.260
Automatic ML	0.421	-6.267	
% of Tobacco in Spending			
Model	RMSE	Δ RMSE vs OLS, %	Time, seconds
OLS	12.898	0.000	1.173
Random forest	11.901	-7.729	12.111
GBM	12.593	-2.366	4.560
Deep Learning	12.575	-2.502	13.760
XGBoost	12.896	-0.017	1.471
Automatic ML	12.219	-5.261	
Tobacco Spending, GEL			
Model	RMSE	Δ RMSE vs OLS, %	Time, seconds
OLS	8.791	0.000	1.095
Random forest	8.125	-7.577	10.895
GBM	8.547	-2.774	4.671
Deep Learning	8.680	-1.262	13.328
XGBoost	8.840	0.553	1.061
Automatic ML	8.265	-5.981	

Notes: Models are first estimated for 8,173 households in the train set (not shown to save space). RMSE is computed for 2,584 households in the test set to evaluate the quality of out-of-sample predictions. GBM stands for Gradient Boosting Machine. XGBoost stands for eXtreme Gradient Boosting. Automatic ML stands for Automatic machine learning command in "h2o" package. All models except for XGBoost are estimated using "h2o" package in R (The H2O.ai team, 2017). XGBoost is estimated using "xgboost" package in R (Chen et al., 2018). All models are estimated using the default parameters except for XGBoost which does not have default for "nrounds" (set to 2000).

Table 2 reports RMSE, improvement in RMSE over OLS in % as well as time spent on executing codes in R for three variables of interest – whether household has any tobacco spending, the share of tobacco spending in total expenditure and total expenditure on tobacco products in GEL.

The results in Table 2 clearly show the advantage of machine learning algorithms over OLS even in such a small sample and with default settings. Almost all algorithms perform better than OLS in all three models. Random forest is the best performing algorithm under default settings: out-of-sample RMSE is lower by 10.6% compared to OLS in a model for whether household has any tobacco spending; RMSE is lower by 7.7% compared to OLS in a model for share of tobacco spending; RMSE is lower by 7.6% compared to OLS in a model for total tobacco spending in GEL. Automatic machine learning (limited to 20 minutes of execution time in each of three models) is the second-best algorithm with default settings.

We would like to stress that these substantial improvements in predictive accuracy compared to OLS were achieved with default parameters in all algorithms. It is needless to say that fine-tuning of the parameters together with cross-validation may substantially improve the predictive accuracy of machine learning algorithms. In addition, the time spent on executing the algorithms is also very reasonable (but may increase in bigger samples). Hence, these results call for active use of modern machine learning tools in practical applications and scientific research even by experts with limited programming experience in R.

4. Conclusions

In this paper we provide the most recent estimates of tobacco spending by house-holds in Georgia in 2016. On the first stage, we describe the extent of tobacco spending by households and identify its important predictors using ordinary least square regressions. Given the limited predictive ability of OLS model we next turn to machine learning methods.

Random forest and automatic machine learning lead to a substantial reduction of RMSE of up to 10.6% even with default settings without any fine-tuning of hyperparameters. This improved accuracy together with the ease of use of machine learning tools which are readily available in R calls for active application of these methods by policy makers and researchers in health economics, public health and related fields.

REFERENCES

1. Bakhturidze, G., Peikrishvili, N. and Mittelmark, M.: The influence of public opinion on tobacco control policy-making in Georgia: Perspectives of governmental and non-governmental stakeholders. Public Participation in Tobacco Control Policy-making in Georgia. Tobacco Prevention & Cessation 2(January), 1 (2016).
2. Berg, C.J., Topuridze, M., Maglakelidze, N., Starua, L., Shishniashvili, M. and Kegler, M.C., 2016. Reactions to smoke-free public policies and smoke-free home policies in the Republic of Georgia: results from a 2014 national survey. International Journal of Public Health, 61(4), pp.409-416.
3. Busch, S.H., Jofre-Bonet, M., Falba, T.A. and Sindelar, J.L., 2004. Tobacco spending and its crowd-out of other goods (No. w10974). National Bureau of Economic Re-search.
4. Chen, T., He, T., Benesty, M., Khotilovich, V. and Tang, Y.: xgboost: Extreme Gradient Boosting. R package version 0.6.4.1. Accessed at https://CRAN.R-project.org/package=xgboost (2018).
5. Djibuti, M., Gotsadze, G., Mataradze, G. and Zoidze, A., 2007. Influence of household demographic and socio-economic factors on household expenditure on tobacco in six New Independent States. BMC Public Health, 7(1), p.222.
6. Gong, Y.L., Koplan, J.P., Feng, W., Chen, C.H., Zheng, P. and Harris, J.R., 1995. Ciga-rette smoking in China: Prevalence, characteristics, and attitudes in Minhang District. Jama, 274(15), pp.1232-1234.
7. The H2O.ai team (2017). h2o: R Interface for H2O. R package version 3.16.0.2. Ac-cessed at https://CRAN.R-project.org/package=h2o
8. Peto, R., Boreham, J., Lopez, A.D., Thun, M. and Heath, C., 1992. Mortality from to-bacco in developed countries: indirect estimation from national vital statistics. The Lan-cet, 339(8804), pp.1268-1278.
9. Roberts, B., Gilmore, A., Stickley, A., Rotman, D., Prohoda, V., Haerpfer, C. and McKee, M., 2012. Changes in smoking prevalence in 8 countries of the former Soviet Union between 2001 and 2010. American Journal of Public Health, 102(7), pp.1320-1328.
10. Torosyan, K., Pignatti, N. and Obrizan, M.: Job Market Outcomes of IDPs: The Case of Georgia. IZA DP No. 11301 (2018).
11. Wang, H., Sindelar, J.L. and Busch, S.H., 2006. The impact of tobacco expenditure on household consumption patterns in rural China. Social science & medicine, 62(6), pp.1414-1426.

APPENDIX

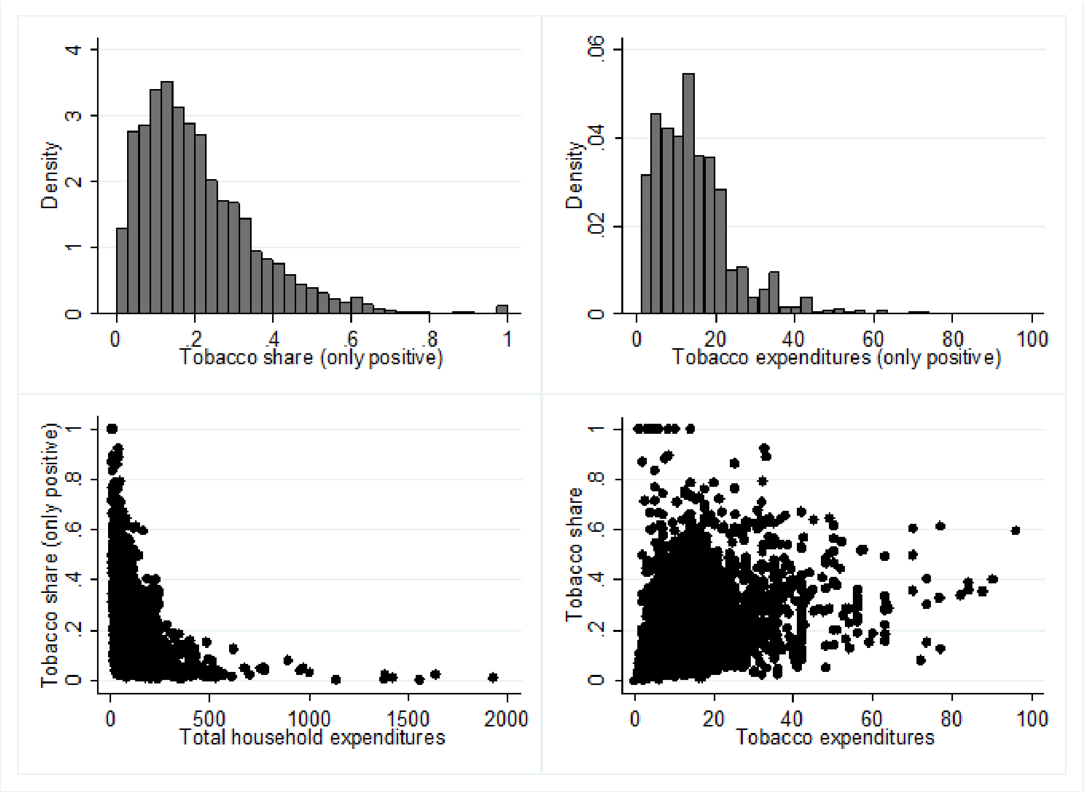

Fig. 1. Tobacco spending in Georgia based on IHS 2016. *Source: Authors' calculations based on GeoStat data.*