Philip Morris International: PMI IMPACT Project

IT for Illicit Trade Risk Management (IT^{2}RM)

Data Science Delivery Report 4:

Modeling and Model Selection Exercise

Prepared by: Prof. Anton Gerunov, Ph.D., PMP

Delivery Date: 28. February 2020

Contents

List of Figures

Executive Summary

The current delivery report focuses on the fourth project quarter and reflects **two main objectives**:

1. Outline the drivers of illicit cigarette contraband and trade;
2. Propose an optimal model structure and modeling algorithm that can successfully forecast non-domestic cigarette incidence.

Building upon the results of Delivery Report 3, and taking into account the correlational structure of data, we focus our investigation of a smaller subset of relevant independent variables that have the highest correlations with the target dependent variable (domestic cigarette incidence). Those variables are as follows:

- Country
- Year
- Quarter
- Time Index
- Age Dependency Ratio
- Compulsory Education in years
- Agricultural Employment
- Total Employment
- Real Growth
- Gross Capital Formation
- Inflation
- Remittances
- Population Growth
- Vulnerable Employment
- Non-domestic Cigarette Incidence (named Total Contraband)

The preliminary analysis leads to a **few key conclusions**:

- Non-domestic cigarette incidence is extremely highly and negatively correlated with objective economic conditions. The higher the economic growth, the less contraband will be observed.
- Likewise, employment (and to a smaller extent – vulnerable employment) shows a negative correlation with non-domestic cigarette incidence.
- Similarly, remittances are negatively correlated with contraband. This variable likely proxies disposable incomes for vulnerable households, and thus, the higher the remittances (thus higher disposable income), the lower the illicit cigarette trade and contraband.
- Population growth is also negatively associated with contraband. This likely reflects a generational shift in cigarette consumption as younger cohorts consume less and the increase in the no-consumption group leads to decreased demands and thus decreased contraband.
- Delivery Report 3 results clearly show that non-domestics cigarette incidence correlates highly with crimes such as sexual assault, trafficking, and burglary. This is partly reflected in the sign and size of the correlations of vulnerable employment.

To assess the quantitative effects of those focus variables we estimate two different multiple regression models. Model I is based purely on the effect of each variable on the explanatory power of the model. In short, each independent variable is entered into a simple linear regression and regressed on the target variable. Due to the **unsatisfactory amount of explained variance in Model I, we also fit another model with alternative specifications**. Model II leverages both the results from the 13 preliminary regressions, the insights from Model I, as well as key associations, reflected in the correlation matrix in Figure 1. The new specification thus retains non-domestic cigarette incidence as dependent variables and leverages the following independent variables: Real.Growth, Age.Dep, Agri.Emp, Employment, Remittances, Population.Grow. The relative importance of all those variables is graphically presented in Figure 2.

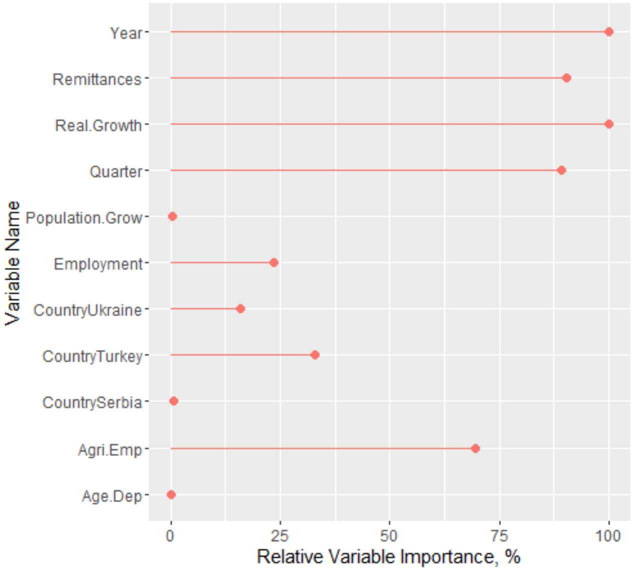

Figure 1: Variable Importance Plot for Multiple Linear Regression Model

The figure shows that the **strongest effects on contraband** is concentrated in the following variables:

- Time Trend – captured by the year and the quarter variables
- Economic growth
- Remittances from abroad

Given the analyzed set of predictors, we endeavor to find the best modeling algorithm that can provide accurate forecasting results. To this end we investigate a wide selection of regression algorithms and estimate them on a training set (comprising about 80% of total data). Then, we find the forecast generated on new observations (the test set, which is about 20% of data) and see what the forecast error is when compared with actual realizations. **Given the calculated accuracy metrics, the best performing forecasting algorithm is bagEarthGCV, which is an implementation of the method Bagged MARS (Multivariate Adaptive Regression Splines) using gCV Pruning**. Essentially, this is a regression method with splines. It can divide the given sub-sample into groups and then fit individual level linear models on each of those groups. Since we leveraged the complete sample for training the model, we can safely assume that the different clusters of data correspond to the four target countries. This shows that essentially linear models are appropriate for the modeling task at hand, and, if needed, alternative linear models can be fruitfully leveraged to analyze and predict the non-domestic cigarette incidence.

The **project has thus moved according to schedule and no delays have been registered** in its data science components. The data and modeling foundations are to be expanded in the coming 4 quarters and the knowledge accrued will be used to further refine the forecasting model, and also develop the risk management model.

More specifically**, the upcoming tasks are as follows**:

1. Ensure proper data pipeline for forecasting model, so that forecasts can be reliably generated on a country level across all target countries;
2. Provide alternative sources of data for the model;
3. Procure data for Ukraine at the regional level;
4. Start developing detailed risk management model at the regional level for all target countries, ensuring sustainable forecast generation and risk scoring;
5. Expand work on results visualization and overall UX when using the analytics module.

Background

This PMI Impact project – IT for Illicit Trade Risk Management (IT^{2}RM) aims at utilizing publicly and privately available data, link them in a unified data warehouse and develop sophisticated analytic capabilities on top of it. Leveraging data on crime, socio-economic development, consumer sentiment, legitimate trade, consumer behavior, illicit cigarette and tobacco market and intercepted illegal imports the project will create a unified database that can be used to visualize and analyze key trends in illicit trade and outline the main drivers at a regional level. This will be used to gain insight into the connection between illicit trade in cigarettes and other criminal activities at a detailed level of granularity. Furthermore, a sophisticated forecasting and risk management system is to be built on top of that, dynamically showing increases in the risk of illicit cigarette trade in different regions that can guide both producers and law enforcement authorities.

The current delivery report focuses on the fourth project quarter and reflects two main objectives:

- **Outline the drivers** of illicit cigarette contraband and trade;
- Propose an **optimal model structure and modeling algorithm** that can successfully forecast non-domestic cigarette incidence.

While those two tasks are connected, they do not necessarily overlap. The former is an exploratory task that focuses on the structural drivers of non-domestic cigarette incidence. In this sense, the insights generated from this can be used not only for field-work and tactical decisions, but also to inform strategy and guide policymaking.

The latter problem aims to create a useful model for forecasting. While it may be a fruitful approach to include process drivers in a structural model it may not always be possible or necessary to do so. Due to issues of missing data for target countries, high multi-collinearity (see Delivery Report 3), or inadequate proportion between independent variables and sample size, it is clear that only a sub-sample of the process drivers will be included even in a purely structural model. This report thus makes an overview of process divers, estimates their relative effect and contribution within a multiple linear regression framework, and then makes an extensive search of the best forecasting algorithm among a large space of possible candidates. The most accurate model is then presented and further commented.

Process Drivers

Building upon the results of Delivery Report 3, and taking into account the correlational structure of data, we focus our investigation of a smaller subset of relevant independent variables that have the highest correlations with the target dependent variable (domestic cigarette incidence). Those variables are as follows:

- Country
- Year
- Quarter
- Time Index
- Age Dependency Ratio
- Compulsory Education in years
- Agricultural Employment
- Total Employment
- Real Growth
- Gross Capital Formation
- Inflation
- Remittances
- Population Growth
- Vulnerable Employment
- Non-domestic Cigarette Incidence (named Total Contraband)

The descriptive statics of those focus features are shortly presented in the following Table 1.

Table 1: Descriptive Statistics of Key Explanatory Variables

Variable Name	mean	sd	median	mad	min	max	range	skew	kurtosis	se
Country	2.30	1.08	3.00	1.48	1.00	4.00	3.00	0.01	-1.44	0.16
Year	2017	1.75	2017	1.48	2014	2019	5.00	-0.22	-1.31	0.25
Quarter	2.81	0.99	3.00	1.48	1.00	4.00	3.00	-0.01	-1.41	0.14
Time.Index	10.81	5.83	11.00	7.41	1.00	20.00	19.00	-0.04	-1.33	0.85
Age.Dep	51.08	2.82	50.36	1.72	45.41	56.60	11.19	0.58	-0.49	0.41
Comp.Educ	10.94	1.33	11.00	1.48	8.00	12.00	4.00	-1.42	0.75	0.19
Agri.Emp	14.84	6.03	17.15	4.54	6.75	22.92	16.18	-0.42	-1.61	0.88
Employment	47.83	3.04	46.88	2.47	41.03	52.56	11.53	0.10	-0.60	0.44
Real.Growth	3.56	2.00	3.40	1.71	-0.75	8.49	9.24	-0.05	0.03	0.29
GrossCap.Form	23.83	4.62	21.52	3.63	16.74	30.64	13.90	0.29	-1.60	0.67
Inflation	6.30	5.36	6.29	5.62	-1.09	22.61	23.70	0.79	0.23	0.78
Remittances	3.92	4.01	3.09	4.36	0.13	12.46	12.34	0.78	-0.70	0.59
Population.Grow	0.24	1.08	-0.49	0.34	-0.73	1.70	2.43	0.46	-1.79	0.16
Vulnerable.Emp	48.29	28.07	35.83	17.40	12.90	85.65	72.76	0.41	-1.54	4.09
Total.Contraband	0.08	0.06	0.07	0.06	0.01	0.22	0.21	0.80	-0.30	0.01

The associations between those variables are presented graphically in the focus correlation matrix in Figure 1.

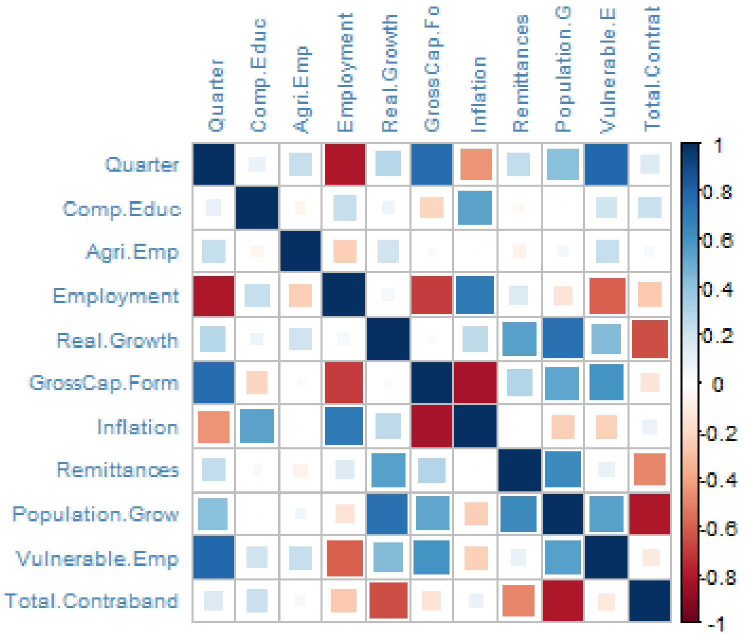

Figure 2: Correlation Matrix of Focused Sub-set of Explanatory Variables

Table 1: Overviews of assessed indicators and their likelihood of being a major determinant of illicit cigarette trade and consumption

The matrix reflects a few key conclusions:

- Non-domestic cigarette incidence is extremely highly and negatively correlated with objective economic conditions. The **higher the economic growth, the less contraband** will be observed. In a similar vein, gross capital formation (a measure of economy-wide investment) also shows a negative association with illicit cigarette incidence. Due to the large overarching effects on both the demand and the supply side of the economy, we expected the overall coefficient sign to reflect net effects given other predictors. Thus, its sign in the forecasting equation is unclear despite the negative correlation.

- Likewise, **employment (and to a smaller extent – vulnerable employment) shows a negative correlation with non-domestic cigarette incidence**. Employment means both higher disposable incomes (thus decreasing demand for cheaper tobacco products such as contraband ones), as well as provides an alternative occupation for potential perpetrators, thus putting a negative pressure on the supply side.
- Similarly, **remittances are negatively correlated** with contraband. This variable likely proxies disposable incomes for vulnerable households, and thus, the higher the remittances (thus higher disposable income), the lower the illicit cigarette trade and contraband.
- **Population growth is also negatively associated with contraband**. This likely reflects a generational shift in cigarette consumption as younger cohorts consume less and the increase in the no-consumption group leads to decreased demands and thus decreased contraband.
- Apart from that we observe expected cross-correlations. Objective economic conditions tend to group together, as do demographic ones.
- Finally, this correlation matrix does not include other types of crimes due to a number of incomplete time series. However, Delivery Report 3 results clearly **show that non-domestics cigarette incidence correlates highly with crimes such as sexual assault, trafficking, and burglary**. This is partly reflected in the sign and size of the correlations of vulnerable employment.

Quantitative Effects of Key Process Drivers

To assess the quantitative effects of those focus variables we estimate two different multiple regression models. Model I is based purely on the effect of each variable on the explanatory power of the model. In short, each independent variable is entered into a simple linear regression and regressed on the target variable. We then use the metric adjusted R^2 as a measure of explained variance to prioritize which features will be included in the final Model I. The relative contribution of each predictor is summarized in Table 2.

Table 2: Summarized Results of 13 Simple Linear Regressions for Feature Selection

	Explanatory Variable	Coefficient	P-Value	Adjusted R^2
1	Country	-0.068	0.004	0.321
2	Year	-0.020	0.000	0.376
3	Quarter	-0.001	0.877	0.001
4	Age.Dep	0.000	0.925	0.000
5	Comp.Educ	0.022	0.000	0.265
6	Agri.Emp	0.001	0.683	0.004
7	Employment	-0.003	0.276	0.026
8	Real.Growth	0.014	0.000	0.250
9	GrossCap.Form	0.005	0.002	0.191
10	Inflation	-0.001	0.362	0.018
11	Remittances	-0.007	0.000	0.263
12	Population.Grow	0.022	0.003	0.176
13	Vulnerable.Emp	0.000	0.124	0.052

The strongest effects observed are for the following variables:

- **Country** – showing strong specific for each of the targeted four countries (Bulgaria, Serbia, Turkey, Ukraine);
- **Year** – the model shows a clear and slightly downward trend over time;
- **Years of compulsory education** – a variable that traditionally has a wide effect on socio-economic processes. Due to its low variance, however, we hypothesize that it reflects country-specific effect and can be meaningfully replaced by the Country variable;
- **Real growth rate of the economy** – robust and sustainable growth affects both supply and demand side of illicit cigarette trade and contraband. On the supply side, it lowers incentives and provides alternative occupation and revenue streams for potential participants. On the demand side, it leads to more disposable income, thus increasing overall consumption, including that of cigarettes;
- **Gross capital formation** – exhibits similar effects and transmission channels like real growth but through the lens of investment processes;
- **Remittances** – we observe a negative relationship of relatively small magnitude, with the effects likely coming across the channel of increasing disposable income, as discussed above;
- **Population growth** – the relatively small but significant coefficient shows that while generational differences are to be discerned, their effects is not large in size'

Based on the estimated simple linear regression, we formulate Model I, which is a multiple regression with non-domestic cigarette incidence as dependent variable and the following independent variables: GrossCap.Form, Real.Growth, Employment, Remittances, Population.Grow. The estimated coefficients and results are presented in the following Table 3.

Table 3: Multiple Linear Regression Model I Results

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.355	0.162	2.189	0.034
GrossCap.Form	0.001	0.008	0.073	0.942
Real.Growth	0.014	0.005	2.831	0.007
Employment	-0.006	0.005	-1.339	0.188
Remittances	-0.008	0.003	-2.488	0.017
Population.Grow	-0.024	0.037	-0.639	0.527
	R^2	**0.396**	**Adj. R^2**	**0.323**

In this model only two of the explanatory variables reach significance – real growth is highly significant below the 1% level and has a robust positive effect on non-domestic cigarette incidence. Remittances also reach significance at the 5% level and they exert a negative effect upon cigarette contraband. The model itself is not particularly good, as it **explains just above 32% of the variance of the target variable**. This means that this model is able to capture less than a third of its dynamics. Such results are not satisfactory for either field-work or policy-making and thus call for the inclusion of more variables and further calibration of the independent variable set.

Due to the **unsatisfactory amount of explained variance in Model I, we also fit another model with alternative specifications**. Model II leverages both the results from the 13 preliminary regressions, the insights from Model I, as well as key associations, reflected in the correlation matrix in Figure 1. The new specification thus retains non-domestic cigarette incidence as dependent variables and leverages the following independent variables: Real.Growth, Age.Dep, Agri.Emp, Employment, Remittances, Population.Grow.

Table 4: Multiple Linear Regression Model II Results

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.518	0.231	6.584	0.000
Real.Growth	0.014	0.004	3.447	0.001
Age.Dep	0.001	0.005	0.142	0.887
Agri.Emp	-0.024	0.005	-4.879	0.000
Employment	-0.026	0.006	-4.674	0.000
Remittances	0.013	0.007	1.902	0.064
Population.Grow	0.106	0.037	2.861	0.007
	R^2	**0.688**	**Adj. R^2**	**0.642**

This model shows **much better results with over 64% of the observed variance explained**. Given the complexity of the phenomenon under modeling and the short time series, this result is satisfactory. Among the new predictors all but the age dependency ratio reach statistical significance, and most of them – below the 1% level. Traditional economic variables such as real growth and employment are highly significant with p-values of 0.001 and $p < 0.0005$, respectively. It seems that growth mainly affects the demand side, spurring consumption of all sorts, while employment affects the supply side, providing alternative sources of income rather than contraband. This holds true for both traditional as well as agricultural employment. Remittances also seem to work mainly on the demand side. Population growth that reflects generational shifts is also statistically significant in this framework, and again its sign reflects a net effect, given the set of predictors under observation.

The relative importance of all those variables is graphically presented in Figure 2.

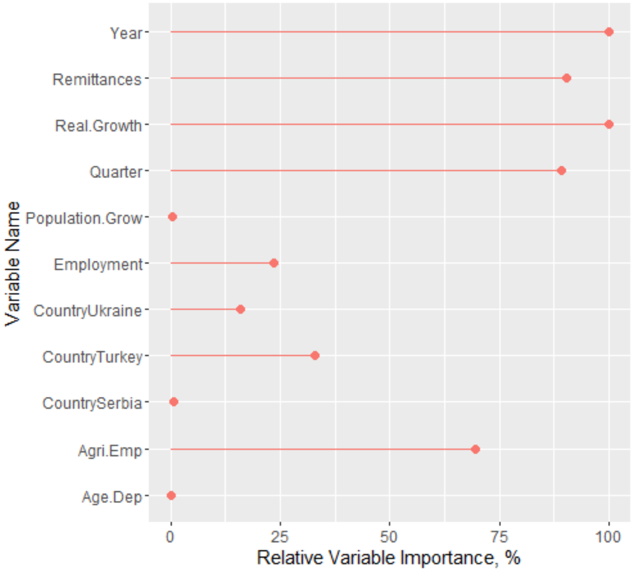

Figure 3: Variable Importance Plot for Multiple Linear Regression Model

The figure shows that the strongest effects on contraband is concentrated in the following variables:

- Time Trend – captured by the year and the quarter variables
- Economic growth

- Remittances from abroad

After we have reached a satisfactory set of predictors that have complete time series and piloted a simple forecasting model, the next step is to find the optimal regression algorithm that can model the data at hand. This is done in the next section.

Exhaustive Search of Optimal Model

Given the analyzed set of predictors, we endeavor to find the best modeling algorithm that can provide accurate forecasting results. To this end **we investigate a wide selection of regression algorithms and estimate them on a training set** (comprising about 80% of total data). Then, we find the forecast generated on new observations (the test set, which is about 20% of data) and see what the forecast error is when compared with actual realizations.

The 105 algorithms investigated are as follows:

- Model Averaged Neural Network
- Bagged MARS
- Bagged MARS using gCV Pruning
- Bayesian Additive Regression Trees
- Bayesian Generalized Linear Model
- Boosted Tree
- The Bayesian lasso
- Bayesian Ridge Regression (Model Averaged)
- Bayesian Ridge Regression
- Bayesian Regularized Neural Networks
- Boosted Linear Model
- Boosted Tree
- Conditional Inference Random Forest
- Conditional Inference Tree 1
- Conditional Inference Tree 2
- Cubist
- Stacked AutoEncoder Deep Neural Network
- Multivariate Adaptive Regression Spline
- Elasticnet
- Tree Models from Genetic Algorithms
- Random Forest by Randomization
- Ridge Regression with Variable Selection
- Generalized Additive Model using LOESS
- Generalized Additive Model using Splines
- Gaussian Process
- Gaussian Process with Polynomial Kernel
- Gaussian Process with Radial Basis Function Kernel
- Multivariate Adaptive Regression Splines
- Fuzzy Rules via MOGUL
- Generalized Linear Model
- Boosted Generalized Linear Model
- glmnet
- Generalized Linear Model with Stepwise Feature Selection
- Hybrid Neural Fuzzy Inference System

- Independent Component Regression
- Partial Least Squares
- k-Nearest Neighbors 1
- k-Nearest Neighbors 2
- Polynomial Kernel Regularized Least Squares
- Radial Basis Function Kernel Regularized Least Squares
- Least Angle Regression
- Least Angle Regression
- The lasso
- Linear Regression with Backwards Selection
- Linear Regression with Forward Selection
- Linear Regression with Stepwise Selection
- Linear Regression
- Linear Regression with Stepwise Selection
- Model Tree
- Model Rules
- Multi-Layer Perceptron 1
- Multi-Layer Perceptron, with multiple layers
- Multi-Layer Perceptron 2
- Multi-Layer Perceptron, multiple layers
- Monotone Multi-Layer Perceptron Neural Network
- Neural Network
- Neural Network
- Non-Negative Least Squares
- Tree-Based Ensembles
- Non-Informative Model
- Parallel Random Forest
- Neural Networks with Feature Extraction
- Principal Component Analysis
- Penalized Linear Regression
- Partial Least Squares
- Partial Least Squares Generalized Linear Models
- Projection Pursuit Regression
- Quantile Random Forest
- Quantile Regression Neural Network
- Random Forest
- Radial Basis Function Network
- Relaxed Lasso
- Random Forest
- Random Forest Rule-Based Model
- Ridge Regression
- Robust Linear Model
- CART 1
- CART 2
- CART 3
- Quantile Regression with LASSO penalty
- Non-Convex Penalized Quantile Regression
- Regularized Random Forest
- Regularized Random Forest

- Relevance Vector Machines with Polynomial Kernel
- Relevance Vector Machines with Radial Basis Function Kernel
- Subtractive Clustering and Fuzzy c-Means Rules
- Partial Least Squares
- Spike and Slab Regression
- Sparse Partial Least Squares
- Supervised Principal Component Analysis
- Support Vector Machines with Linear Kernel 1
- Support Vector Machines with Linear Kernel 2
- L Regularized Support Vector Machine (dual) with Linear Kernel
- Support Vector Machines with Polynomial Kernel
- Support Vector Machines with Radial Basis Function Kernel 1
- Support Vector Machines with Radial Basis Function Kernel 2
- Support Vector Machines with Radial Basis Function Kernel 3
- Bagged CART
- Partial Least Squares
- Wang and Mendel Fuzzy Rules
- eXtreme Gradient Boosting 1
- eXtreme Gradient Boosting 2
- eXtreme Gradient Boosting 3
- Self-Organizing Maps
- Ensembles of Generalized Linear Models

To compare the different models, we use **the following forecast accuracy metrics** (equations use standard notation):

- Mean Error: $ME = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$
- Root Mean Squared Error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i)^2}$
- Mean Absolute Error: $MAE = \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|$
- Mean Percentage Error: $MPE = \frac{1}{n}\sum_{i=1}^{n} 100 * \frac{(y_i - f_i^m)}{y_i}$
- Mean Absolute Percentage Error: $MAPE = \frac{1}{n}\sum_{i=1}^{n} 100 * \left|\frac{(y_i - f_i^m)}{y_i}\right|$

In terms of the optimal model selection, we mostly rely on the RMSE as this is a consensus metric, easy to implement and communicate, and provides comparability to other similar exercises. The summary table of forecast accuracy metrics for all the estimated models is presented in Table 5.

Method	Time	ME	RMSE	MAE	MPE	MAPE
avNNet	8.393	-0.007	0.059	0.050	-69.398	93.699
bagEarth	9.693	-0.009	0.018	0.016	-29.088	35.171
bagEarthGCV	3.748	0.001	0.010	0.009	1.097	16.183
bartMachine	663.993	-0.002	0.023	0.018	-21.250	36.106
bayesglm	2.692	-0.004	0.030	0.023	-20.981	38.186
blackboost	32.084	-0.015	0.054	0.041	-80.764	96.133
blasso	5.040	-0.002	0.040	0.032	-37.396	59.261
blassoAveraged	4.760	-0.002	0.040	0.032	-37.396	59.261
bridge	3.765	-0.003	0.038	0.030	-35.162	54.413
brnn	2.502	-0.001	0.030	0.023	-13.370	33.037

BstLm	5.583	0.028	0.055	0.031	24.953	32.194
bstTree	50.069	0.000	0.020	0.014	-8.875	16.165
cforest	4.665	-0.004	0.045	0.033	-49.012	68.548
ctree	1.898	-0.026	0.071	0.055	-111.518	136.248
ctree2	3.569	-0.026	0.071	0.055	-111.518	136.248
cubist	4.674	0.000	0.015	0.012	-6.613	17.671
dnn	25.854	-0.001	0.059	0.048	-57.590	84.609
earth	1.836	-0.023	0.040	0.028	-74.871	79.104
enet	1.678	-0.004	0.029	0.024	-20.639	39.809
evtree	21.729	-0.007	0.059	0.043	-44.244	86.699
extraTrees	6.151	-0.005	0.017	0.014	-23.223	30.718
foba	1.695	-0.005	0.031	0.023	-26.784	38.484
gamLoess	3.690	-0.002	0.024	0.019	-11.880	34.192
gamSpline	2.959	-0.003	0.029	0.024	-17.994	40.661
gaussprLinear	5.261	-0.005	0.039	0.028	-36.182	51.607
gaussprPoly	2.741	0.004	0.023	0.018	9.047	22.683
gaussprRadial	1.249	0.000	0.036	0.024	-23.951	36.218
gcvEarth	1.386	-0.018	0.030	0.022	-55.743	60.140
GFS.FR.MOGUL	1035.820	-0.028	0.065	0.059	-107.552	124.456
glm	0.771	-0.003	0.029	0.024	-17.996	40.662
glmboost	1.210	-0.002	0.041	0.032	-37.759	58.711
glmnet	1.640	-0.006	0.032	0.023	-28.180	38.682
glmStepAIC	1.296	-0.003	0.029	0.024	-17.241	40.771
HYFIS	115.540	0.012	0.051	0.038	27.541	56.925
icr	1.104	0.027	0.055	0.035	12.405	41.182
kernelpls	0.894	-0.011	0.048	0.042	-55.044	92.085
kknn	1.305	0.007	0.042	0.031	12.449	35.255
knn	0.676	0.016	0.032	0.022	22.045	27.753
krlsPoly	1.009	0.004	0.059	0.047	-48.601	78.711
krlsRadial	1.353	0.003	0.027	0.019	2.378	21.891
lars	0.784	-0.005	0.033	0.023	-28.546	38.276
lars2	0.827	-0.001	0.040	0.031	-35.426	56.362
lasso	0.686	-0.005	0.033	0.024	-28.702	38.841
leapBackward	1.136	-0.002	0.048	0.043	-36.951	89.961
leapForward	0.684	-0.007	0.038	0.032	-41.310	58.940
leapSeq	0.827	0.000	0.039	0.029	-27.867	48.011
lm	0.712	-0.003	0.029	0.024	-17.996	40.662
lmStepAIC	1.073	-0.003	0.029	0.024	-17.241	40.771
M5	14.361	0.009	0.022	0.020	20.155	35.776
M5Rules	3.568	0.009	0.022	0.020	20.155	35.776
mlp	2.038	0.028	0.065	0.047	-5.665	57.062
mlpML	1.868	0.028	0.065	0.047	-5.665	57.062
mlpWeightDecay	3.999	0.040	0.071	0.049	17.719	47.596
mlpWeightDecayML	3.921	0.040	0.071	0.049	17.719	47.596
monmlp	14.993	-0.033	0.060	0.042	-77.523	102.665
neuralnet	0.982	0.004	0.059	0.047	-48.465	78.883
nnet	1.427	-0.007	0.059	0.050	-69.407	93.706
nnls	0.607	0.020	0.062	0.037	0.169	41.564
nodeHarvest	71.069	-0.002	0.029	0.018	-7.289	16.728
null	0.593	0.004	0.059	0.047	-48.617	78.978

parRF	1.963	-0.002	0.032	0.024	-33.964	49.105
pcaNNet	1.455	-0.018	0.054	0.044	-80.823	93.907
pcr	0.816	0.028	0.056	0.039	12.353	49.516
penalized	1.855	0.004	0.050	0.039	-39.503	75.638
pls	0.738	-0.011	0.048	0.042	-55.044	92.085
plsRglm	8.719	0.002	0.039	0.034	-22.693	58.269
ppr	0.724	-0.006	0.020	0.016	-20.814	28.918
qrf	3.569	0.011	0.022	0.016	11.407	24.394
qrnn	725.935	0.007	0.016	0.014	14.782	20.700
ranger	3.701	-0.007	0.029	0.023	-39.751	51.044
rbfDDA	2.018	0.085	0.104	0.085	99.930	99.930
relaxo	1.182	0.085	0.104	0.085	100.000	100.000
rf	1.630	-0.006	0.035	0.026	-42.495	56.026
rfRules	137.390	-0.026	0.064	0.059	-104.711	122.166
ridge	1.773	-0.004	0.029	0.024	-20.639	39.809
rlm	2.744	-0.001	0.030	0.024	-16.519	38.493
rpart	1.717	-0.021	0.071	0.055	-104.699	136.136
rpart1SE	1.387	-0.028	0.070	0.053	-115.488	135.466
rpart2	1.803	-0.028	0.070	0.053	-115.488	135.466
rqlasso	3.127	-0.005	0.038	0.028	-36.927	47.945
rqnc	2.440	-0.004	0.037	0.027	-35.053	47.821
RRF	28.850	-0.003	0.033	0.025	-36.274	50.383
RRFglobal	5.725	-0.007	0.036	0.028	-46.829	60.296
rvmPoly	6.993	0.003	0.042	0.033	-16.211	51.916
rvmRadial	2.458	0.021	0.033	0.028	39.723	45.168
SBC	6.901	0.003	0.045	0.032	6.691	35.757
simpls	1.743	-0.011	0.048	0.042	-55.044	92.085
spikeslab	20.952	-0.005	0.040	0.030	-39.268	55.967
spls	8.265	-0.004	0.029	0.023	-20.651	36.272
superpc	2.153	0.098	0.112	0.098	137.994	137.994
svmLinear	5.854	-0.002	0.039	0.027	-31.829	44.138
svmLinear2	1.563	0.007	0.040	0.027	-14.944	40.563
svmLinear3	2.792	-0.026	0.064	0.058	-103.298	120.993
svmPoly	5.427	0.006	0.025	0.018	10.554	22.039
svmRadial	1.775	0.005	0.036	0.022	-8.976	23.567
svmRadialCost	1.786	0.005	0.035	0.021	-7.489	21.686
svmRadialSigma	2.558	0.005	0.036	0.022	-8.976	23.567
treebag	3.133	-0.003	0.042	0.032	-43.358	62.424
widekernelpls	1.476	-0.011	0.048	0.042	-55.044	92.085
WM	21.652	0.010	0.050	0.039	25.335	56.980
xgbDART	517.148	-0.017	0.038	0.024	-66.751	71.167
xgbLinear	50.018	-0.031	0.055	0.035	-102.382	105.442
xgbTree	92.528	-0.001	0.037	0.026	-34.411	59.389
xyf	5.377	-0.001	0.021	0.018	8.002	27.404
randomGLM	250.600	-0.002	0.038	0.030	-31.016	53.171
xgbDART	380.361	-0.017	0.038	0.024	-66.751	71.167
xgbLinear	41.509	-0.031	0.055	0.035	-102.382	105.442
xgbTree	70.967	-0.001	0.037	0.026	-34.411	59.389
xyf	3.828	-0.001	0.021	0.018	8.002	27.404

Optimal Model Estimation and Characteristics

Given the calculated accuracy metrics, the **best performing forecasting algorithm is bagEarthGCV, which is an implementation of the method Bagged MARS (Multivariate Adaptive Regression Splines) using gCV Pruning**. Essentially, this is a regression method with splines. It can divide the given sub-sample into groups and then fit individual level linear models on each of those groups. Since we leveraged the complete sample for training the model, we can safely assume that the different clusters of data correspond to the four target countries. This shows that **essentially linear models are appropriate for the modeling task at hand**, and, if needed, alternative linear models can be fruitfully leveraged to analyze and predict the non-domestic cigarette incidence.

The regression spline model producing the highest forecast accuracy may not necessarily have the highest R squared (and indeed, its adjusted R^2 is at 0.573) but is able to generate the best predictions among alternative models. The relative contribution of the tested variables to this models (their relative importance in %) is presented visually in Figure 3, as follows.

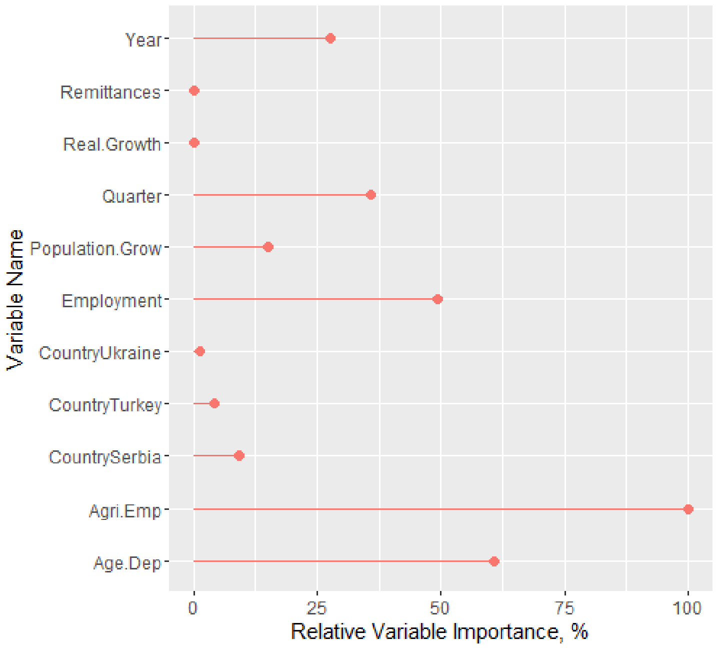

Figure 4: Variable Importance Plot for Bagged Multivariate Adaptive Regression Splines (MARS) Model

Next Steps

The modeling exercise of the first four quarters of the **project has achieved the following**:

- Collated, procured and organized a large database of socio-economic, demographic, business, and crime statistics for the purposes of the project
- Performed initial data analysis and visualization of the data
- Outlined possible drivers of non-domestic cigarette incidence and interpreted their associations within the framework of correlational analysis
- Selected data series of non-domestic cigarette incidence that meet minimum standard for inclusion in forecasting model
- Tested focus predictors within simple and multiple linear regression
- Selected optimal forecasting model and estimated it using data at hand

The project has thus moved according to schedule and no delays have been registered in its data science components. The data and modeling foundations are to be expanded in the coming 4 quarters and the knowledge accrued will be used to further refine the forecasting model, and also develop the risk management model.

More specifically**, the upcoming tasks are as follows**:

- Ensure proper data pipeline for forecasting model, so that forecasts can be reliably generated on a country level across all target countries
- Provide alternative sources of data for the model
- Procure data for Ukraine at the regional level
- Start developing detailed risk management model at the regional level for all target countries, ensuring sustainable forecast generation and risk scoring
- Expand work on results visualization and overall UX when using the analytics module