Citadel Data Open Central Regional Datathon

The Effectiveness of Taxes and WHO's International EMPOWER Package on Tobacco Consumption

Daniel Park[1], Praneet Rathi[2], Suryasri Kalidas[2], and Aaron Miller[3]

[1]Purdue University
[2]University of Illinois at Urbana-Champaign
[3]University of North Carolina at Chapel Hill

November 21, 2021

Abstract

Since the 20th century, smoking tobacco has developed into a significant global epidemic, and countries around the world continue to impose legislations and apply preemptive measures to prevent the problem until this day. This is measured by the World Health Organization (WHO) through MPOWER, which is a policy package intended to assist in the country-level implementation of effective interventions to reduce the demand for tobacco. It is an internationally-recognized summary of the important elements of tobacco control.

MPOWER:

- **M**onitor tobacco use and prevention policies
- **P**rotect people from tobacco smoke
- **O**ffer help to quit tobacco use
- **W**arn about the dangers of tobacco
- **E**nforce bans on tobacco advertising, promotion and sponsorship
- **R**aise taxes on tobacco

As participants in the Fall 2021 Central Regional Datathon, we aim to support the efforts in tackling this worldwide health problem through data analysis. Our task was as follows:

...to use tobacco related data in order to discover and analyze patterns associated with tobacco usage. More broadly, you should aim to highlight and discover hidden patterns in the data that could be used to draw any meaningful conclusion on the subject.

Using data from the United Nations, we were able to create and optimize 8 multiple linear regression models that uses MPOWER metrics to predict tobacco consumption trends in countries with above R^2 accuracy. Discovering that putting countries into clusters and performing regression on each group significantly increased the prediction accuracy, we realized how significant environmental and economic factors can be in determining the effectiveness of anti-tobacco legislations.

1 Executive Summary

1.1 Formation of Topic Question

When we were provided the datasets, we explored ways to model an action and a response in order to pave the initial directions in our research. Therefore, we labeled the following CSV files as such:

Action: stop_smoking.csv. This global dataset "presents indicators that contribute to an individual to stop smoking. These contributions can be very direct, like offering government help, or less direct like increasing the tax and price of cigars and banning Tobacco advertisements." It has MPOWER measurements for 2012 and 2014 for each country.

Response:

- **tobacco_production.csv**
- **tobacco_use_ww.csv**
- **tobacco_use_us.csv**
- **sales_per_day.csv**
- **death_rates_smoking_age.csv**

Worldwide tobacco production, tobacco use, daily sales, and death rates for smoking age categories as well as United States tobacco use and prevalence of chronic respiratory diseases over the years could potentially be explained as the outcomes, given that the explanatory variable is anti-tobacco legislations and measures.

Knowing also that more than 10 years after the policy came into effect on February 7, 2008, the WHO can benefit from an assessment of MPOWER and the effectiveness of various tobacco control efforts, we decided to answer the following question:

According to MPOWER by the WHO, have legislations and control measures been effective in decreasing tobacco consumption worldwide? To what extent do tobacco legislations, such as taxes, drive down the prevalence of tobacco?

1.2 Areas of Examination

1.2.1 Naive Multiple Linear Regression

To answer the question above, we turned the task into a multiple regression analysis: explanatory variables are MPOWER measurements (monitor, protect from tobacco smoke, offer help to quit tobacco use, warn about dangers of tobacco, enforce bans on tobacco advertising, anti-tobacco mass media campaigns) and percent tax on the most sold tobacco brand, and the response variable is percent tobacco consumption per population.

This regression had a score of only $R^2 = 0.17$. This led us to reflect on the diversity of circumstances present in each separate nation that might affect their response to the UN program. Does location, specifically country type and their individual traits (potentially culture and economy) factor into tobacco consumption? Should we perform the regression analysis by clustering the countries into regions such as Europe and Africa (ParentLocation column) and determine if the regression score increases? Maybe assuming all countries will have the same tobacco consumption responses from MPOWER efforts is a naive proposition. This doesn't, however, necessarily mean that every country responds differently.

1.2.2 Country-Clustering Multiple Linear Regression

With this motivation, we set out to cluster countries by their response to the UN program, seeking to group together those that responded similarly. Through an optimization method, we clustered the countries into 8 groups each of around equal size. Performing a multiple linear regression on each of these, we were able to achieve an R^2 of above 0.92 for all of them. Notable countries from each of the 8 clusters include: Pakistan, China, Nigeria, India, Indonesia, USA, Uganda, and Yemen.

We then sought to understand what exactly is differentiating these clusters of countries, who based off the R^2 values seem to have natural partitions. In particular, we looked at international data on development indicators from the DataBank, and determined which features are most predictive of clustering groups. Various environmental factors and merchandise activity proved to be the two larger themes that tend to be predictive of clustering groups, and thereby predictive of countries' responses to the UN non-smoking initiatives.

1.3 Conclusion

We were able to devise regression models that, given the country and its MPOWER scores, predicts population tobacco consumption to at least $R^2 = 0.92$. Each model represents a cluster of countries. Region, or specifically country, seems to weigh into the changes in tobacco consumption with respect to MPOWER, as the score was $R^2 = 0.17$ without. When we cluster all the countries, we're able to increase this value to above 0.92 for each cluster.

2 Technical Exposition

2.1 Exploratory Data Analysis Data Cleaning

2.1.1 Provided Datasets

In Figure 1, we give a brief description of the tobacco_production_ww dataset.

Looking at Figure 2, we will drop the value footnotes as they are for citing sources and offer little value to our analysis. After doing

```
tobacco_production = pd.read_csv("drive/MyDrive/central-datathon-files/tobacco_production.csv")
print(tobacco_production.head())
print(tobacco_production.info())
print(tobacco_production.describe())
```

```
  Country or Area  Year         Unit        Value Value Footnotes
0         Albania  2006  Metric tons   546.600000             NaN
1         Albania  2006     Mil. USD     1.324113             NaN
2         Albania  2005  Metric tons  1878.500000             NaN
3         Albania  2005     Mil. USD     4.844285             NaN
4         Albania  2004  Metric tons   751.900000             NaN
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1208 entries, 0 to 1207
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Country or Area  1208 non-null   object
 1   Year             1208 non-null   int64
 2   Unit             1208 non-null   object
 3   Value            1208 non-null   float64
 4   Value Footnotes  187 non-null    object
dtypes: float64(1), int64(1), object(3)
memory usage: 47.3+ KB
None
              Year         Value
count  1208.000000  1.208000e+03
mean   2005.293874  1.403498e+04
std       6.199438  8.793434e+04
min    1995.000000  0.000000e+00
25%    2000.000000  6.078607e+00
50%    2005.000000  1.737481e+02
75%    2010.250000  7.696750e+03
max    2016.000000  2.056000e+06
```

Figure 1: tobacco_production_ww: Total tobacco production per year from 1995 to 2016 for 79 unique countries or areas.

```
print("# of Unique Countries:", len(np.unique(tobacco_production["Country or Area"])))
print("Unique Value Footnotes:", list(set(tobacco_production["Value Footnotes"])))
print("Unique Units:", np.unique(tobacco_production["Unit"]))

# of Unique Countries: 79
Unique Value Footnotes: [nan, '16', '4', '7', '12', '14', '6', '13,14', '17', '8', '5', '15', '3', '10', '2', '9', '1', '11']
Unique Units: ['Metric tons' 'Mil. USD']

tobacco_production = tobacco_production.drop("Value Footnotes", axis=1)
tobacco_production = tobacco_production[tobacco_production["Unit"] == "Metric tons"]
```

Figure 2: tobacco_production_ww data cleaning

so, there are no rows with null values. Additionally, there seems to be two rows for each country and year, where one measures tobacco production by metric tons and other measures in million USD. We decided to choose the universal metric tons, since the value of tobacco can change over the years due to the dynamic economy.

According to Figure 3, worldwide tobacco production has been relatively consistent ever since the drop from the spike in 1997, which makes it easy to raise questions on the effectiveness of tobacco preventive measures. However, considering the exponential growth in the world population and the tobacco supply remaining relatively constant, this supports the effectiveness of global prevention efforts. Additionally, more tobacco companies may have just migrated their production locations to countries that are relatively more tobacco-lenient, so this plot does not account for countries with varying levels of tobacco-prevention efforts and their relative production levels.

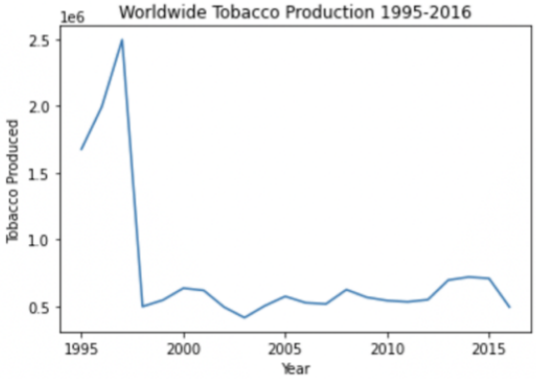


Figure 3: Worldwide tobacco production 1995-2016.

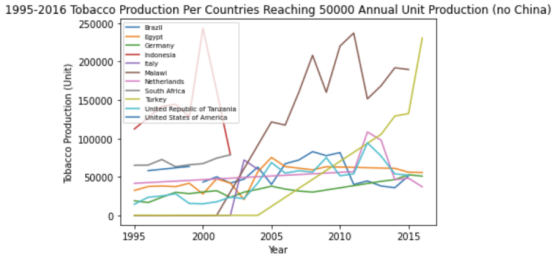


Figure 4: Worldwide tobacco production per countries reaching 50000 annual unit production (China excluded) 1995-2016.

```
print("Unique ParentLocations:", len(np.unique(tobacco_use_ww["ParentLocation"])))
print("Unique Countries:", len(np.unique(tobacco_use_ww["Location"])))
tobacco_use_ww = tobacco_use_ww.drop(["ParentLocationCode", "SpatialDimValueCode"], axis=1)
tobacco_use_ww = tobacco_use_ww[tobacco_use_ww["Gender"] == "Both Sexes"]

Unique ParentLocations: 6
Unique Countries: 149
```

Figure 5: tobacco_use_ww data cleaning

In Figure 4, we observe the history of tobacco production across some of the top producer countries. We decided to exclude China as they are the main producers of tobacco, making data visualization and making inferences difficult. Tobacco production value is an unspecified unit of measurement provided by the UNData. We can observe that tobacco production has increased exponentially for particular countries, including Turkey and Malawi. Since tobacco production is a less direct measure of usage prevalence for a particular country than consumption, we will look into changes in tobacco consumption worldwide.

Moving on to the tobacco_use_ww dataset, we do data cleaning and exploration to find out in Figure 5 that the dataset demonstrates total tobacco consumption per year from 2000 to 2016 for 149 unique countries or areas. Each row also contains the ParentLocation (Europe, Americas, Africa...), the respective ParentLocationCode (EUR, AFR...), SpatialDimValueCode (longitude, latitude), and Gender (Male, Female, Both Sexes). We will drop ParentLocationCode, which is redundant with ParentLocation, and SpatialDimValueCode, as relative geographical location is not the focus of our study. Additionally, we decided to only

Location		Location	
Nepal	-32.3	Egypt	-0.1
Norway	-24.6	France	0.0
Argentina	-24.4	Oman	0.1
Peru	-24.3	Slovakia	0.2
Comoros	-24.2	Niger	0.7
Sweden	-23.9	Republic of Moldova	0.8
Lao People's Democratic Republic	-23.3	Croatia	1.7
Cambodia	-22.6	Portugal	2.6
India	-21.9	Congo	4.9
Pakistan	-21.5	Lesotho	6.2

Figure 6: Percent population tobacco consumption 2018 - percent population tobacco consumption 2000 worldwide. 10 largest (left) and smallest (right) decrease countries.

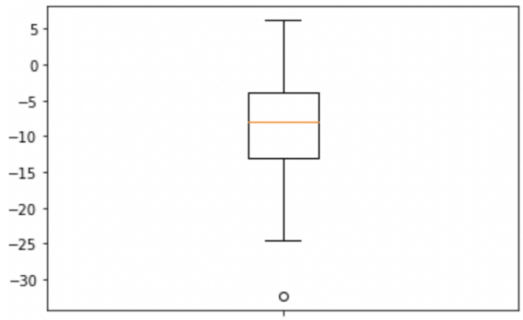


Figure 7: Percent population tobacco consumption 2018 - percent population tobacco consumption 2000 worldwide.

keep the Both Sexes row to simplify our tasks and narrow our focus to a more general one.

Similar to our analysis on tobacco production, we explored the changes in percent population tobacco consumption from 2000 and 2018 for countries worldwide. We found the top 10 and bottom 10 countries in terms of highest differences. Additionally, we created a box plot of all countries' data, finding out that the majority of the countries in the distribution have decreased tobacco consumption. The median percent decrease is about -10

As shown in Figure 8, some countries produce more than they consume. Therefore, we should just analyze consumption as a direct measure of tobacco usage and level of a health threat as production does not highly correlate with consumption.

Moving on, in Figure 9, the stop_smoking dataset provides the MPOWER measurements (categorical ratings between 1 to 5) for each of the 194 countries in 2007, 2010, 2012, and 2014. We excluded Code, which is redundant with Entity, AvgCigarettePriceDollars, since costs can significantly differ across countries, and AvgTaxesAsPctCigarettePrice, given too many null values exist. We searched and found a 2008 - 2018 MPOWER dataset by WHO. We also searched for a replacement for the current AvgTaxesAsPctCigarettePrice column.

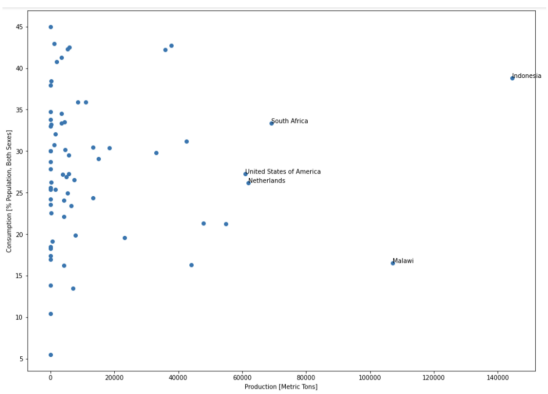


Figure 8: Percentage of population (both sexes) that consumes tobacco vs. annual tobacco production in metric tons. Note that China is excluded from this figure because their production is an extreme outlier. The countries with the five largest annual production values are labeled on the figure.

```
stop_smoking = pd.read_csv("drive/MyDrive/central-datathon-files/stop_smoking.csv")
print(stop_smoking.head())
print(stop_smoking.info())
print(stop_smoking.describe())
```

```
      Entity Code  ...  EnforceBansTobaccoAd  HelpToQuit
0    Algeria  DZA  ...                     4           3
1    Algeria  DZA  ...                     4           4
2  Argentina  ARG  ...                     4           4
3  Argentina  ARG  ...                     4           5
4    Armenia  ARM  ...                     2           4

[5 rows x 7 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 774 entries, 0 to 773
Data columns (total 7 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Entity                      774 non-null    object
 1   Code                        774 non-null    object
 2   Year                        774 non-null    int64
 3   AvgCigarettePriceDollars    208 non-null    float64
 4   AvgTaxesAsPctCigarettePrice  209 non-null    float64
 5   EnforceBansTobaccoAd        774 non-null    int64
 6   HelpToQuit                  774 non-null    int64
dtypes: float64(2), int64(3), object(2)
memory usage: 42.5+ KB
None
              Year  AvgCigarettePriceDollars  ...  EnforceBansTobaccoAd  HelpToQuit
count   774.000000                208.000000  ...            774.000000  774.000000
mean   2010.755814                  4.336394  ...              3.313953    3.493540
std       2.587373                  2.534659  ...              1.088995    0.807042
min    2007.000000                  0.000000  ...              2.000000    1.000000
25%    2010.000000                  2.195000  ...              2.000000    3.000000
50%    2012.000000                  4.155000  ...              4.000000    4.000000
75%    2013.500000                  5.767500  ...              4.000000    4.000000
max    2014.000000                 13.000000  ...              5.000000    5.000000
```

```
print("Unique Countries:", len(np.unique(stop_smoking["Entity"])))
print("Unique Years:", np.unique(stop_smoking["Year"]))
stop_smoking = stop_smoking.drop(["AvgCigarettePriceDollars", "AvgTaxesAsPctCigarettePrice"], axis=1)
```

```
Unique Countries: 194
Unique Years: [2007 2010 2012 2014]
```

Figure 9: stop_smoking data cleaning

2.1.2 External Datasets

Therefore, we downloaded these additional datasets:

- stop_smoking_extended.csv [1]
- The_Tax_Burden_on_Tobacco__1970-2019.csv [2]

The stop_smoking_extended.csv is the same as stop_smoking.csv with measurements in 2007, 2008, 2010, 2012, 2014, 2016, and 2018. It is missing the columns about taxes and contains the full MPOWER categories. The The_Tax_Burden_on_Tobacco__1970-2019.csv dataset needs cleaning and removal of unnecessary columns, but essentially we just want Location (country), Period (year), and Value (percent tax on the most sold brand of cigarette).

2.2 Preparing for Regression

We will merge international tobacco consumption, percent tax on the most sold brand of cigarette, and MPOWER measurements into one dataset for regression analysis.

Essentially, we make the variable names consistent ("Period", "Location") and merge on them. In stop_smoking_extend.csv's Anti-tobacco mass media campaigns and Price columns, there are entries called "Data not available" and "Not applicable". When the Anti-tobacco mass media campaigns score data is not available, I set it to the average of the score data for that location in all available years and floor it to be a categorical integer value. For the Price (percent tax on most sold brand of cigarette) column, I do the same, except in the case there is absolutely no tax data available for the given country, we set the value to 0.

2.3 General International Regression, Naive Attempt

2.3.1 Regression Equation and Coefficient of Determination

In Figure 10, the coefficient of determination is extremely low, which means our regression equation is not a good model to predict tobacco consumption based on MPOWER.

2.3.2 Individual MPOWER metrics vs Tobacco Consumption

From boxplots and scatterplot in Figure 11, we cannot determine there is a significant change

```
# MULTIPLE LINEAR REGRESSION
# dropped: location, period, parent location
# (Explanatory): MPOWER scores + price
# (Response): Usage as %

from sklearn.linear_model import LinearRegression
# from sklearn.model_selection import train_test_split

X = new_df.drop(["Usage", "Location", "Period", "ParentLocation"], axis=1)
y = new_df["Usage"]
# X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
regr = LinearRegression()
regr.fit(X, y)

print("Coefficients: ", regr.coef_)
print("Intercept: ", regr.intercept_)
print(regr.score(X, y))
```

```
Coefficients:  [ 2.11970079  0.36159386 -1.15686765 -1.25643293  0.98791536 -0.29572938
  0.15921392]
Intercept:  13.687304975792136
0.16688923565047875
```

Figure 10: y = 2.12 * (Monitor) + 0.36 * (Protect from tobacco smoke) - 1.16 * (Offer help to quit tobacco use) - 1.26 * (Warn about the dangers of tobacco) + 0.99 * (Enforce bans on tobacco advertising) - 0.30 * (Anti-tobacco mass media campaigns) + 0.16 * (Price) + 13.69 R2 = 0.17

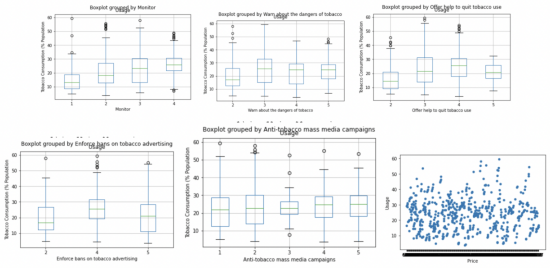


Figure 11: Individual MPOWER metrics vs Tobacco Consumption

in tobacco consumption due to the changes in MPOWER scores.

2.3.3 Comments

We can devise 2 implications:

1. Maybe anti-tobacco efforts, as measured by MPOWER, are not effective in alleviating tobacco consumption, which is counter-intuitive and not likely.
2. Countries and possibly years, which we normalized in our naive regression attempt do actually matter when predicting tobacco consumption. For example, cultural differences and living standards may weigh into the prevalence of tobacco.

We will investigate (2) in the next regression analysis, where we cluster countries into 8 sets.

2.4 Clustered International Regression

Our initial, undirected attempt at modeling the impact of UN's smoking initiatives simply runs a regression model on the dataset with

several entries for each country corresponding to different years. Taking these static data points, representing a single country at a single timestamp, is less intuitive and impactful than analyzing how each country has changed over time. Thus, it would be ideal to condense the several rows for each country into one row, where that row contains information regarding the UN initiatives and their impact on the country over the past decade.

We first look at the spread of the columns for each country over the course of the decade for the UN initiative metrics, price of tobacco, and usage of tobacco. These are as follows, with the number representing the average standard deviation of the metric for each country across 5 years of data:

Feature	St. Dev.
Monitor	0.371703
Protect	0.342358
Offer	0.293652
Warn	0.596275
Enforce	0.270773
Media	1.033693
Price	4.926539
Usage	1.438712

The relatively low deviations for each of the first six columns, all of which are UN initiatives, suggest that their changes in the past decade aren't particularly relevant, and we can simply average them. The same can't be said for price and usage, so we subtract the newest price and usage from the oldest price and usage for each country. Performing this subtraction also narrows the scope of modeling to fit the change in usage over time, or in other words the actual impact of the initiatives. We thus get the following dataset:

Location	Monitor	Protect	Offer	...
Albania	2.25	5	3.75	...
Algeria	2.5	3	3	...
Andorra	1.5	2.75	4	...
Argentina	3.75	4.25	4	...
Armenia	4	3	3.75	...
Australia	4	5	5	...
Austria	4	2	4	...
Azerbaijan	3.25	2.5	2.75	...

Now, the problem with the initial regression model was that it tried to fit one model for capturing the impact of UN initiatives and price of tobacco on its usage to all the countries, resulting in very poor accuracy. Instead, it's more likely that different countries respond differently to the initiatives, so a one-size-fits-all approach would not be effective. What we therefore set out to do is to determine which, if any, groups of countries respond similarly to UN initiatives and the price of tobacco. To do so, we present a method to generate optimized clusters.

In particular, we begin by randomly partitioning the data points, each corresponding to a different country, into similarly-sized clusters. This is the initial configuration from which we will optimize the clusters. For each of the clusters, we run a multiple linear regression model, the same format as was used for the initial universal attempt, that is best predictive of changes in usage of tobacco based on UN initiative metrics and the change in price of tobacco for the countries in that cluster. After these models are generated, we go through every country and quantify each model's performance in predicting that country's usage, regardless of whether that model is or isn't the one generated for the country's current cluster. The country then gets assigned to the cluster whose model was most predictive and therefore had the least error in predicting the country's change in usage of tobacco.

The motivation behind this is that even if a particular country was used to train one model, it may happen that the model generated by countries of another cluster is more predictive, meaning that this country behaves more similarly to countries of that cluster. We repeat this process until the configuration no longer changes, meaning that each country is now part of the cluster whose model best predicts that country's change in usage rate of tobacco.

This method is analogous to the popular k-means clustering method, with the exception that rather than assigning data points to the center they're closest to, we assign countries to the model that best predicts their tobacco usage. Like k-means clustering, however, this method does have one potential weakness- the final clustering configuration it outputs is prone to change depending on the initial random configuration that was generated. To overcome this, we run this entire method ten times, each with a new random partition of countries into initial clusters, and select the final configuration that is highest performing of all ten.

To quantify the performance of a general

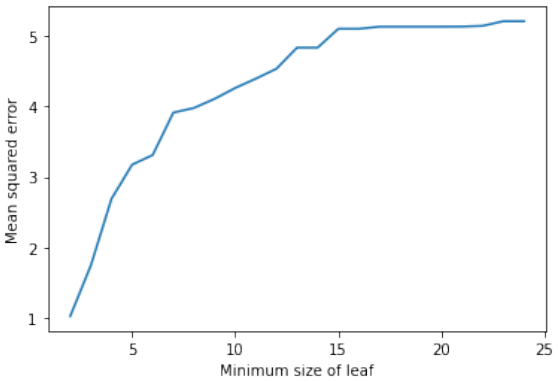


Figure 12: Mean squared error of decision trees tuned by minimum leaf size

clustering configuration, we used the following metric:

$$\sum_{n=1}^{k} R_n^2 \cdot c(n)$$

where the R_n^2 term represents the accuracy of the multiple regression model and the $c(n)$ term represents the number of countries in the nth cluster. This essentially gives us a weighted average of the clusters' model performances, where the weight is the size of the cluster.

The other remaining factor in this optimization method was the number of clusters present, as that would be set from the beginning. To determine this, we sought a proxy for the number of clusters that would best represent the natural distribution of countries with respect to their tobacco usage response to UN initiatives and tobacco price. We did this by fitting a decision tree regression model, in which each of the leaves of the tree represent countries that had similar values for the predictive variables and similar ultimate output of change in tobacco usage. To determine the optimal number of clusters, we generated decision trees for leaf sizes ranging from 2 to 25, resulting in Figure 12.

Intuitively, the mean squared error goes close to zero as the leaf size goes toward 1, as the resulting model would have a different leaf for each country and thus no error in predictions. What we want, however, is a balance between accuracy and minimum leaf size, as the cluster sizes and therefore the leaf sizes should be on the larger side to provide otherwise unknown insight regarding similarity of countries' tobacco usage responses. We thus selected a minimum leaf size of 14, seeing from the graph that this point corresponds to a divot in the graph, indicating this locally optimum with little improvement in MSE despite decreases in minimum leaf size. Using this minimum leaf size, we fit a decision tree with rules as seen in Figure 13.

We see that this decision tree produced 8 final leaves, suggesting that 8 is a natural number of clusters to use for our optimization method. The models for the 8 clusters had the following R^2 values:

0.9955447818713915
0.9857414513390416
0.99009790668601
0.9864343108521781
0.9203396224098466
0.9981156982496142
0.9871481551539502
0.9965356400620113

This optimization method produced a sharp increase in accuracy for regression models, with the results of one such cluster model shown in Figure 14.

From the Figure 14 we see that for the select cluster all of the features are highly significant in predicting usage rate, and this was the case for all of the clusters. The high R^2 value can be seen as well. This all suggests a sharp improvement over the initial 0.17 mark that was obtained by modeling all the countries together.

Having successfully determined an accurate partition of countries based on their response to tobacco usage, we now seek to investigate what factors are common between the countries within each cluster- and different between countries of different clusters. To do so, we turned to the World Bank's World Development Indicators dataset, containing 1443 different indicators with regards to world development, ranging in broad categories from energy consumption, to healthcare services, to demographics. These come at an annual level for each country, with many empty values, so we decided to only keep variables that had some value for each country in the past 5 years. We have provided a sample of the initial and final datasets to illustrate these transformations.

Initial Dataset

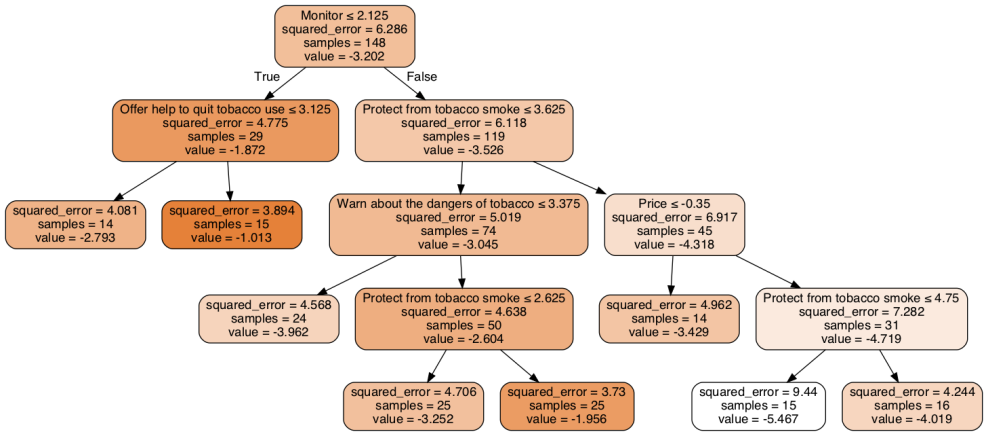
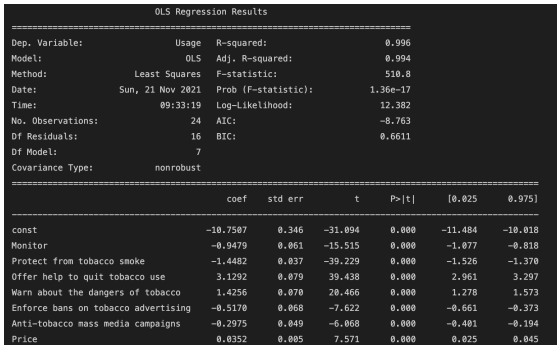


Figure 13: Decisions and statistics of trained decision tree with minimum leaf size of 14



OLS Regression Results

Dep. Variable:	Usage	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.994
Method:	Least Squares	F-statistic:	510.8
Date:	Sun, 21 Nov 2021	Prob (F-statistic):	1.36e-17
Time:	09:33:19	Log-Likelihood:	12.382
No. Observations:	24	AIC:	-8.763
Df Residuals:	16	BIC:	0.6611
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.7507	0.346	-31.094	0.000	-11.484	-10.018
Monitor	-0.9479	0.061	-15.515	0.000	-1.077	-0.818
Protect from tobacco smoke	-1.4482	0.037	-39.229	0.000	-1.526	-1.370
Offer help to quit tobacco use	3.1292	0.079	39.438	0.000	2.961	3.297
Warn about the dangers of tobacco	1.4256	0.070	20.466	0.000	1.278	1.573
Enforce bans on tobacco advertising	-0.5170	0.068	-7.622	0.000	-0.661	-0.373
Anti-tobacco mass media campaigns	-0.2975	0.049	-6.068	0.000	-0.401	-0.194
Price	0.0352	0.005	7.571	0.000	0.025	0.045

Figure 14: Multiple Linear Regression results from one of the 8 final clusters

Location	Series	2016	2017	2018	...
AFG	Fuels	32.44	NA	NA	...
AFG	Rural elec	97.10	97.09	98.31	...
AFG	Urban elec	99.5	99.5	99.9	...
AFG	Fin acct	NA	NA	NA	...

Final Dataset

Location	Fuels	Rural elec	Urban elec
AFG	32.44	96.9	100

We still were left with 61 variables, which could pose some correlation and predictive overlapping issues for modeling, so we filtered this further by using Anova tests for the columns, ultimately producing 30 variables to conduct modeling on. We then performed a Random Forest classifier on these 30 variables, and quantified which variables were most impactful in determining clusters. These came out to be **Methane emissions, Forest area, Fisheries production, Merchandise imports, and Merchandise exports**.

Following several analyses verifying the relative importance of these features with respect to other features, we can conclude that, among the wide variety of factors under the umbrella of world development factors, the most predictive ones fall under two realms: **environmental** and **economic** variables.

The results we obtained no doubt merited further analysis, perhaps into the dynamics between these two realms with regards to country's smoking responses. Nevertheless, with the contest constraints, we conclude having obtained insights into clusters of countries that respond similarly to UN's smoking initiatives and the characteristics of those countries that are best associated with and perhaps predictive of the cluster they're in. This insight can be utilized by the United Nations as well as any other movement or organization seeking to combat smoking and similar problems at an international scale. It paves the way for a possible balance between a naive, one-size-fits-all approach and an unrealistic, country-tailored approach.

References

[1] HealthData. (2019). The Tax Burden on Tobacco, 1970-2019. [Data File]. Retrieved from https://healthdata.gov/dataset/The-Tax-Burden-on-Tobacco-1970-2019/etts-u9ii/data

[2] MPOWER. (2018). World Health Organization. Retrieved November 21, 2021, from https://www.who.int/initiatives/mpower