The Data Open

Improving *MPOWER* Effectiveness through Clustering and Regression Analysis

Izzy Cheng, Nicholas Clark, Yoon Tae Park, Amy Tai

November 21, 2021

1 Problem Statement

MPOWER is a framework created by the World Health Organization (WHO), consisting of six measures designed to aid countries in implementing policies to reduce the demand for tobacco and mitigate the negative health effects of tobacco usage. For each measure (Monitor, Protect, Offer, Warn, Enforce, and Raise), WHO has defined four categories to classify each country's progress. Since its inception in 2008, the policy package has seen widespread adoption, with 146 countries enacting at least one of the six primary measures at the highest level of achievement. While the program has been met with positive reception, the efficacy of *MPOWER* may be sub-optimal. With that in mind, we pose the following primary question, along with sub-questions intended to guide the analysis process.

Primary Question: What suggestions can be made to improve the effectiveness of *MPOWER*?

1. Which countries have seen the best results in controlling tobacco usage through the use of *MPOWER*?
2. Have countries seen success through other means, or have countries seen minimal improvement in controlling tobacco usage despite utilizing *MPOWER*?
3. What is the predictive power of *MPOWER* in determining tobacco usage and are there other attributes with similar or higher correlations to tobacco usage?

> "*MPOWER* is the only document of a somewhat strategic nature that is a source of information on the spread of the tobacco epidemic, as well as of suggestions concerning specific actions for supporting the fight against this epidemic." - MPOWER – Strategy for Fighting the Global Tobacco Epidemic (2009)

As the primary strategic framework for tackling the tobacco epidemic, even a minor improvement to *MPOWER* has the potential to increase its adoption on a global scale, and as a result, save lives.

2 Executive Summary

Higher levels of implementation of *MPOWER* levels corresponded to a lower level of tobacco on a country basis. As expected, *MPOWER* did appear to provide positive benefits in reducing tobacco usage in adults over fifteen years of age.

There are several countries that have implemented *MPOWER* measures to a high extent and seen a great reduction in tobacco; those include Brazil, Iran, Panama, Turkey, United Arab Emirates, and Uruguay. These countries may serve as possible case studies for the effective implementation of *MPOWER* measures, and provide evidence to suggest that *MPOWER* is indeed an effective means of reducing tobacco usage.

As for possible counterexamples, there are two cases to consider: countries that have utilized *MPOWER* but seen minimal improvement, and countries that have not used *MPOWER* but still seen progress in controlling tobacco. An interesting pattern emerged in that of those countries with minimal *MPOWER* measures who still had low tobacco usage values, the top ten were all located in Africa. Of those countries with high utilization of *MPOWER* but subpar results, the most notable were North Macedonia and Russia. These countries may suggest that *MPOWER* may not be the sole option or comprehensive solution in limiting tobacco usage.

In order to find the predictive power of *MPOWER* in determining tobacco usage, as well as other attributes with similar or higher correlations to tobacco usage, the objective was to find the optimal linear regression model. *MPOWER* alone is not a key factor in decreasing tobacco usage. In addition, categorization by region and government expenditure data, when paired with the *MPOWER* summary indicators provided a significant improvement in model performance.

The final model correctly predicted at least 56% of the variation in tobacco usage, and a substantial improvement in comparison to the initial model which only utilized *MPOWER* measures. This supports the conclusion that there are means other than *MPOWER* measures that offer a complementary approach to decreasing tobacco usage: chief among these being regions and total government expenditure. Therefore, it is recommended when devising a strategy for combating the tobacco epidemic on a country-level, to comprehensively consider regional and cultural differences which may significantly dictate the relative effectiveness of *MPOWER* . While *MPOWER* as a set of guiding principles is valuable in its simplicity, in order to provide optimal resources for countries, WHO may consider modifying their recommended measures and evaluative indicators on a continent or sub-region basis.

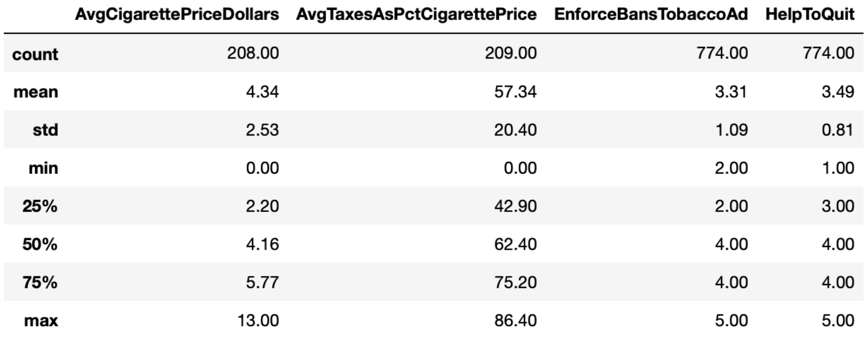
3 Technical Exposition

The technical exposition is compromised of three primary components: our exploratory data analysis with hypothesis testing, the clustering of countries by *MPOWER* engagement using K-Means, and the utilization of linear regression models to determine the predictor power of each *MPOWER* summary indicator.

3.1 Exploratory Data Analysis

The primary data sets utilized were **stop_smoking** and **tobacco_use_ww**. As the object of analysis was *MPOWER*, the **stop_smoking**, provided essential data relating to the implementation of *MPOWER* on a country-level basis. The **tobacco_use_ww** dataset, when taken in conjunction with **stop_smoking** yielded an essential metric for evaluating the effect of implementing *MPOWER* measures. The **death_rates_smoking_age** was also considered for this purpose but was ultimately discarded. Drawing explicit connections between death rate and *MPOWER* would be difficult, as the causal relationship, if present, would likely occur over an extended time frame, and therefore conclusions drawn from year-to-year fluctuations would likely not be representative or persuasive.

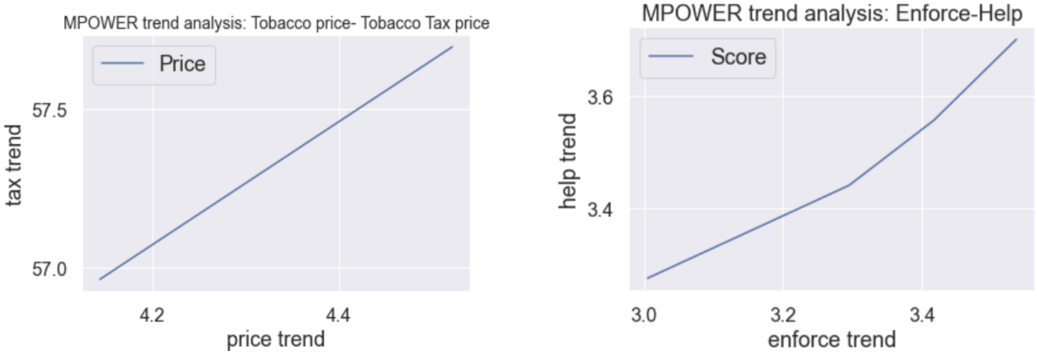
Relevant data was available for 2007, 2010, 2012, and 2014. The following summary statistics were found using **stop_smoking**, with the notable observations being the average of ‘Taxes as a percentage of Cigarette Price’, ‘Enforce’, and ‘Offer’ being 57.3%, 3.3 and 3.5 respectively (Figure 1).



	AvgCigarettePriceDollars	AvgTaxesAsPctCigarettePrice	EnforceBansTobaccoAd	HelpToQuit
count	208.00	209.00	774.00	774.00
mean	4.34	57.34	3.31	3.49
std	2.53	20.40	1.09	0.81
min	0.00	0.00	2.00	1.00
25%	2.20	42.90	2.00	3.00
50%	4.16	62.40	4.00	4.00
75%	5.77	75.20	4.00	4.00
max	13.00	86.40	5.00	5.00

Figure 1: Summary Statistics of **stop_smoking**

Trend on *MPOWER* is increasing for all kind of means, which are enforcing bans to tobacco advertising, helping people to quit smoking, average cigarette price, and average taxes (Figure 2).

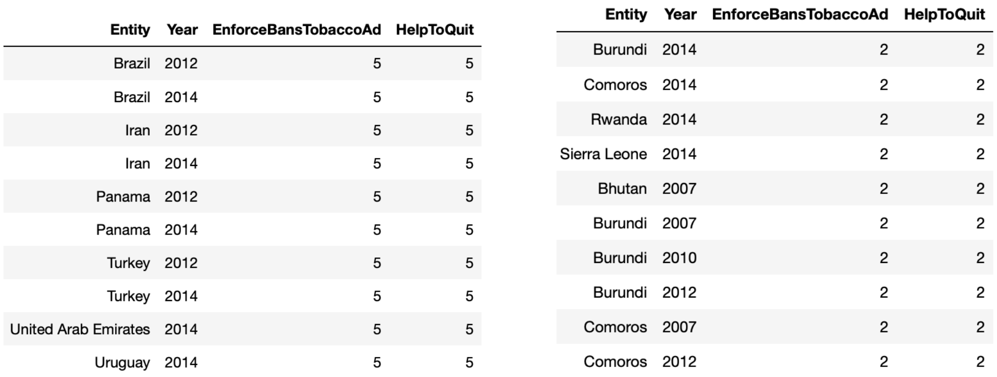


(a) Tobacco Price - Tobacco Tax

(b) Enforce-Offer

Figure 2: *MPOWER* Trend Analysis

Countries that have implemented the Enforce, and Offer measures to a high extent include Brazil, Iran, Panama, Turkey, United Arab Emirates, and Uruguay, while countries that have not include were Burundi, Comoros, Rwanda, Sierra Leone, Bhutan (Figure 3).



Entity	Year	EnforceBansTobaccoAd	HelpToQuit
Brazil	2012	5	5
Brazil	2014	5	5
Iran	2012	5	5
Iran	2014	5	5
Panama	2012	5	5
Panama	2014	5	5
Turkey	2012	5	5
Turkey	2014	5	5
United Arab Emirates	2014	5	5
Uruguay	2014	5	5

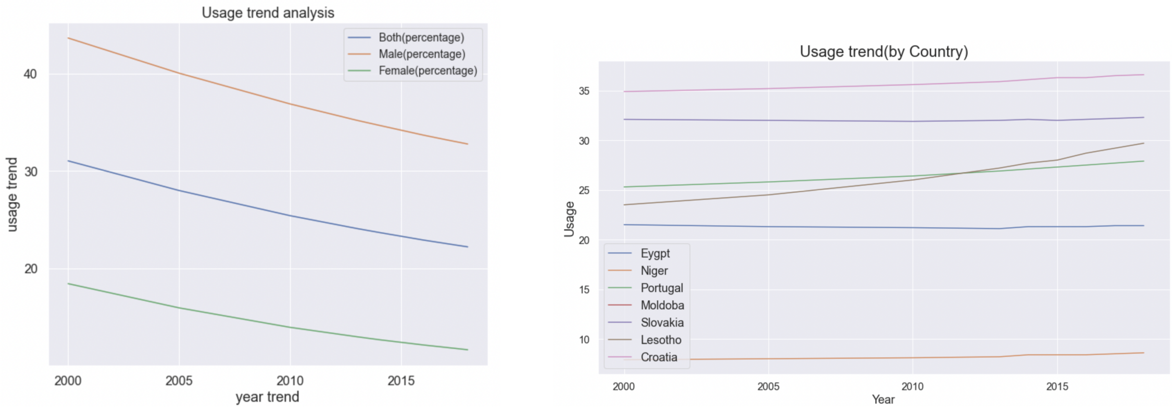
(a) High Level

Entity	Year	EnforceBansTobaccoAd	HelpToQuit
Burundi	2014	2	2
Comoros	2014	2	2
Rwanda	2014	2	2
Sierra Leone	2014	2	2
Bhutan	2007	2	2
Burundi	2007	2	2
Burundi	2010	2	2
Burundi	2012	2	2
Comoros	2007	2	2
Comoros	2012	2	2

(b) Low Level

Figure 3: Enforce, Offer Implementation Level

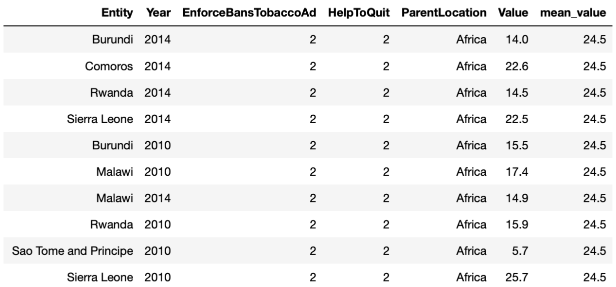
The overall proportion of individuals using tobacco products has been steadily decreasing across both males and females (Figure 4). Countries where this trend is not present include Egypt, Niger, Congo, Portugal, Moldova, Slovakia, Lesotho, and Croatia (Figure 4).



(a) Sexes Level

(b) Country Level

Figure 4: Usage Trend Analysis



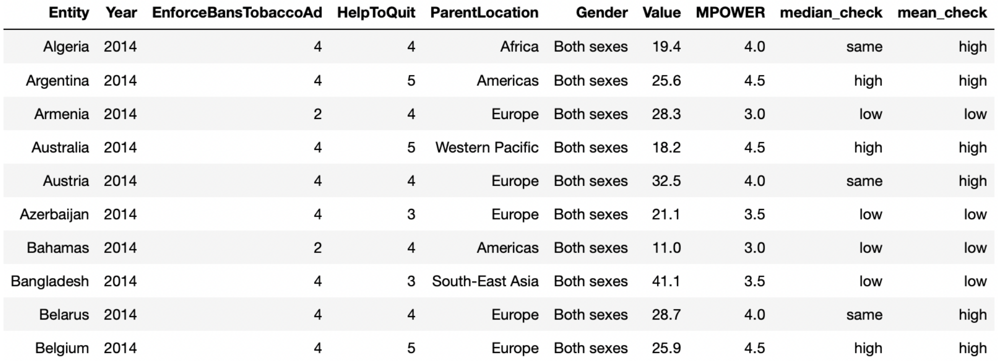
Entity	Year	EnforceBansTobaccoAd	HelpToQuit	ParentLocation	Value	mean_value
Burundi	2014	2	2	Africa	14.0	24.5
Comoros	2014	2	2	Africa	22.6	24.5
Rwanda	2014	2	2	Africa	14.5	24.5
Sierra Leone	2014	2	2	Africa	22.5	24.5
Burundi	2010	2	2	Africa	15.5	24.5
Malawi	2010	2	2	Africa	17.4	24.5
Malawi	2014	2	2	Africa	14.9	24.5
Rwanda	2010	2	2	Africa	15.9	24.5
Sao Tome and Principe	2010	2	2	Africa	5.7	24.5
Sierra Leone	2010	2	2	Africa	25.7	24.5

Figure 5: Summary of the countries with low *MPOWER* implementation and low tobacco usage

Countries were divided into two categories, high *MPOWER* and low *MPOWER*, in order to conduct a preliminary hypothesis test. and tried to see if they were significantly different or not.

- Hypothesis: High *MPOWER* countries have less absolute tobacco usage than low *MPOWER* countries.
- Null Hypothesis: Tobacco usage is constant across both categories.

A new field "MPOWER" was created by taking the mean of "EnforceBansTobaccoAd" and "HelpToQuit". The mean and median where then used to split the countries into two categories: High *MPOWER* and Low *MPOWER* (Figure 6). Both a T-test and Mann-Whitney U-test were conducted to compare the mean and median tobacco usage percentage across the two categories (Table 1). Both p-values were essentially 0. As a result, we rejected the null hypothesis at the $\alpha = .005$ significance level and concluded that High *MPOWER* countries tend to have less tobacco usage than Low *MPOWER* countries.



Entity	Year	EnforceBansTobaccoAd	HelpToQuit	ParentLocation	Gender	Value	MPOWER	median_check	mean_check
Algeria	2014	4	4	Africa	Both sexes	19.4	4.0	same	high
Argentina	2014	4	5	Americas	Both sexes	25.6	4.5	high	high
Armenia	2014	2	4	Europe	Both sexes	28.3	3.0	low	low
Australia	2014	4	5	Western Pacific	Both sexes	18.2	4.5	high	high
Austria	2014	4	4	Europe	Both sexes	32.5	4.0	same	high
Azerbaijan	2014	4	3	Europe	Both sexes	21.1	3.5	low	low
Bahamas	2014	2	4	Americas	Both sexes	11.0	3.0	low	low
Bangladesh	2014	4	3	South-East Asia	Both sexes	41.1	3.5	low	low
Belarus	2014	4	4	Europe	Both sexes	28.7	4.0	same	high
Belgium	2014	4	5	Europe	Both sexes	25.9	4.5	high	high

Figure 6: *MPOWER* Field Inclusion and Categorization

	Statistic	P-value
T-test	24.397	1.9E-70 $\approx$ 0
U-test	1.914	3.1E-22 $\approx$ 0

Table 1: Test Results

3.2 K-Means Clustering

The primary data sources used for this portion are **tobacco_use_ww** and **stop_smoking**. Both data sets were included in the competition materials, with the former coming from WHO and the latter from GHO. K-Means clustering was used to group countries based on their respective values. K-Means clustering is an iterative solution clustering analysis algorithm, with the goal being to divide N points into K clusters, so that each point belongs to the cluster corresponding to the nearest mean value, known as the cluster center. The equation used for fitting being either one of the following equations.

$$\sum_{i=0}^{n} \min_{\mu_j \epsilon C}(||x_i - \mu_j||^2) \tag{Clustering}$$

$$arg_S \min \sum_{i=1}^{k} \sum_{x \epsilon S_i} ||x - \mu_i||^2 \tag{Fatal Errors}$$

In order to determine the optimal number of clusters, two complementary evaluation techniques, the elbow method and the silhouette coefficient were employed. The elbow method records the sum of the squared error (SSE) for various cluster sizes and tries to find the best trade-off point between the error and the number of clusters. The silhouette coefficient ranges between -1 and 1 where a greater number corresponds to a better fit of data points to their clusters compared to other clusters (K-means clustering in Python: A practical guide).

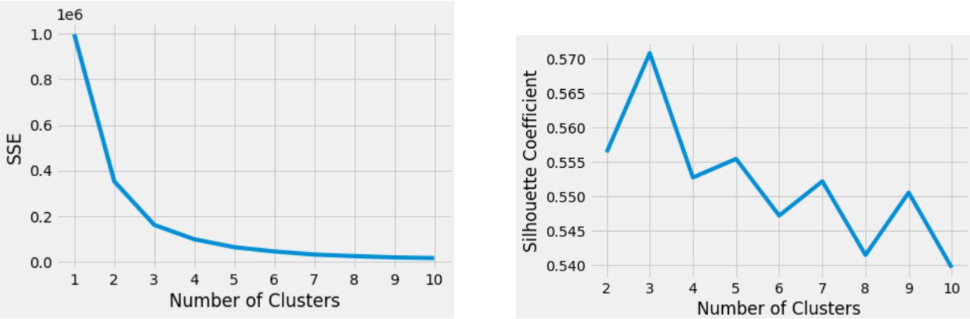
The first model clustered the countries based on their usage. Since there were no issues with duplication or NA values in the **tobacco_use_ww** data set, all of the data was kept for model development (4023 rows). Only one feature was used, the usage value, corresponding to the percentage of the population over 15 years of age that currently uses any tobacco product.

The data was transformed into a 2d array which was then fed to the KMeans sklearn algorithm to determine the appropriate clusters. The results from the elbow method and the silhouette coefficient are shown below (Figure 7). Using kneed's KneeLocator and visual assessment of these graphs, three clusters was determined to be the optimal number.

As such, KMeans was fit again using three clusters with the associated division (Figure 8). Based on the chart, it was evident that label 0 corresponded to low tobacco usage, 1 with medium usage, and 2 with high tobacco usage. Figure 9 presents the average number of countries in each parent location for all years of the data set (2000, 2005, 2010, 2013, 2014, 2015, 2016, 2017, and 2018) compared with only the most recent year (2018). As shown in the table, Africa and America have the most countries move from high tobacco usage to low tobacco usage in 2018. There were two country-gender combinations that also shifted from label 2 to label 0, i.e., from high tobacco usage to low tobacco usage and 90 country-gender combinations that moved down one label, i.e., from label 1 to label 0 or from label 2 to label 1. There was also one country-gender combination whose label group shifted upwards, with the remaining 354 country gender combinations remaining constant. Additionally, some of the specific countries that have seen the highest improvement in tobacco usage from before *MPOWER* was introduced (2005) to the most recent year (2018) are shown below (Figure 24).

Notably:

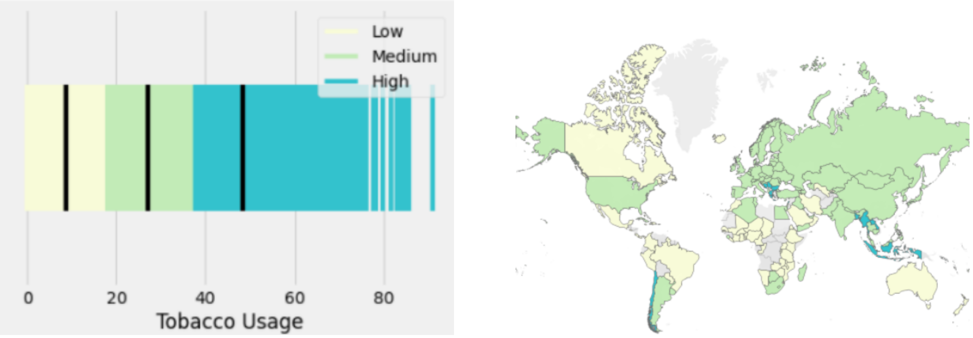
- Nepal-Female and Peru-Male both improved from high usage to low usage
- Portugal-Female regressed from low usage to medium usage
- All other country-gender combinations with changes in the usage category can be found in the csv file named **country-gender-usage-1.csv** located within Team_14_data.zip.



(a) Elbow Curve Plot.

(b) Silhouette Coefficient Plot.

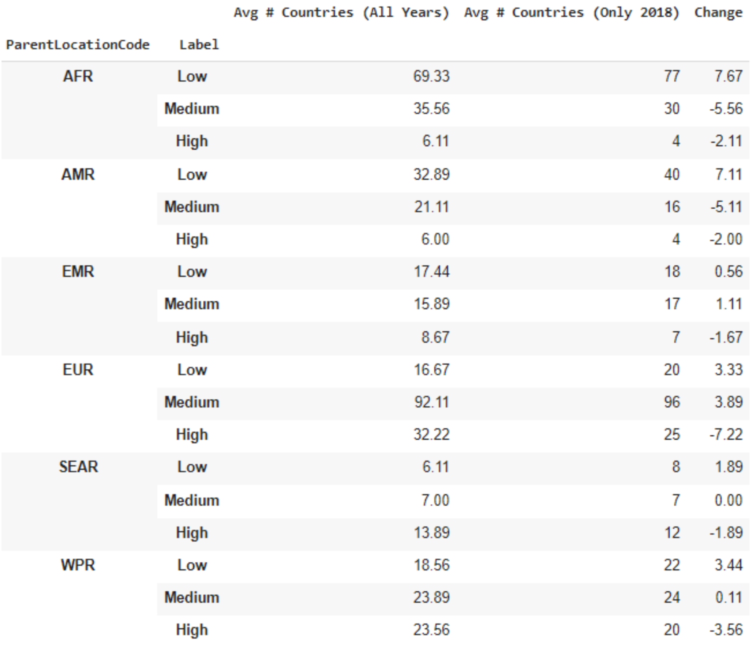
Figure 7: Evaluation Results for Various Cluster Sizes.



(a) Cluster Usage Division.

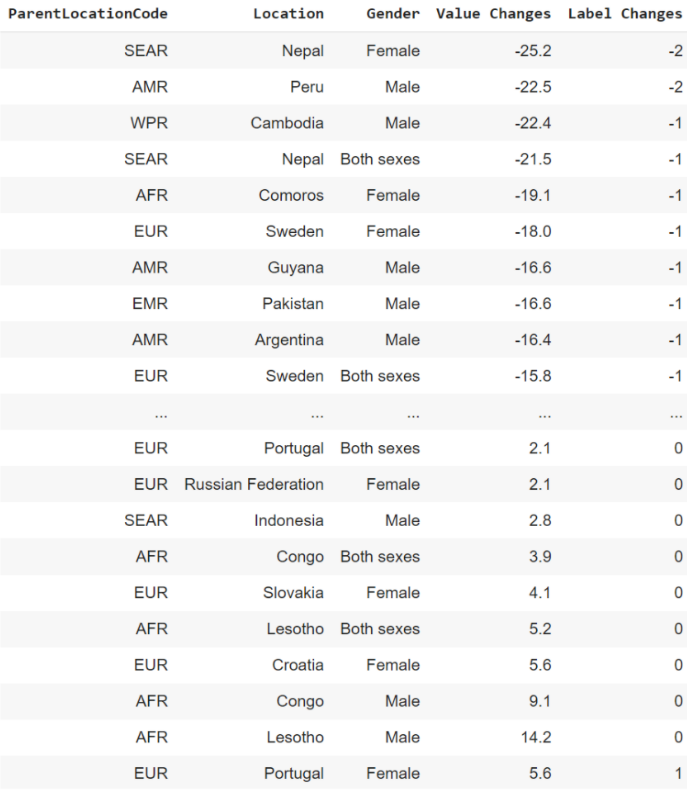
(b) Clusters On World Map (Tableau).

Figure 8: Attributes of Clusters



ParentLocationCode	Label	Avg # Countries (All Years)	Avg # Countries (Only 2018)	Change
AFR	Low	69.33	77	7.67
	Medium	35.56	30	-5.56
	High	6.11	4	-2.11
AMR	Low	32.89	40	7.11
	Medium	21.11	16	-5.11
	High	6.00	4	-2.00
EMR	Low	17.44	18	0.56
	Medium	15.89	17	1.11
	High	8.67	7	-1.67
EUR	Low	16.67	20	3.33
	Medium	92.11	96	3.89
	High	32.22	25	-7.22
SEAR	Low	6.11	8	1.89
	Medium	7.00	7	0.00
	High	13.89	12	-1.89
WPR	Low	18.56	22	3.44
	Medium	23.89	24	0.11
	High	23.56	20	-3.56

Figure 9: Tabular comparison of the average number of country gender entries in each cluster for all years and based on the most recent year.

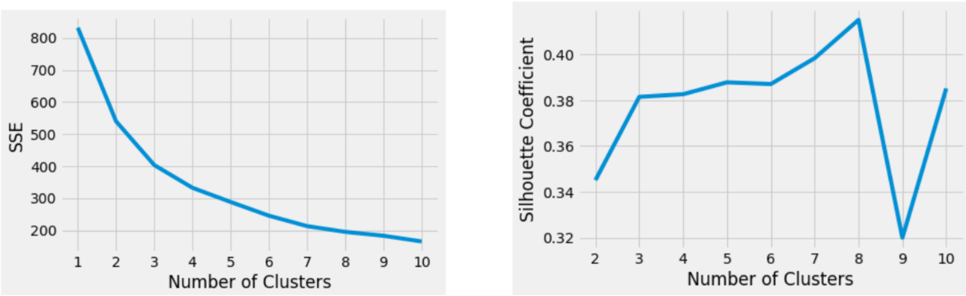


ParentLocationCode	Location	Gender	Value Changes	Label Changes
SEAR	Nepal	Female	-25.2	-2
AMR	Peru	Male	-22.5	-2
WPR	Cambodia	Male	-22.4	-1
SEAR	Nepal	Both sexes	-21.5	-1
AFR	Comoros	Female	-19.1	-1
EUR	Sweden	Female	-18.0	-1
AMR	Guyana	Male	-16.6	-1
EMR	Pakistan	Male	-16.6	-1
AMR	Argentina	Male	-16.4	-1
EUR	Sweden	Both sexes	-15.8	-1
...	...	...	...	...
EUR	Portugal	Both sexes	2.1	0
EUR	Russian Federation	Female	2.1	0
SEAR	Indonesia	Male	2.8	0
AFR	Congo	Both sexes	3.9	0
EUR	Slovakia	Female	4.1	0
AFR	Lesotho	Both sexes	5.2	0
EUR	Croatia	Female	5.6	0
AFR	Congo	Male	9.1	0
AFR	Lesotho	Male	14.2	0
EUR	Portugal	Female	5.6	1

Figure 10: Tabular summary of the top 10 and bottom 10 country gender combinations sorted by the magnitude of their changes in label and usage.

The second model clustered the countries based upon *MPOWER* indicator values. There was a data quality issue with the **stop_smoking** table as there were 566 rows with NA values in the "AvgCigarettePriceDollars" and "AvgTaxesAsPctCigarettePrice" fields. As such, these rows were dropped from the data for this K-Means model leaving a total of 208 rows for fitting. Afterwards, the values were scaled using the StandardScaler package in sklearn. The StandardScaler package standardizes features by subtracting the mean and then scaling to unit variance (K-means clustering in Python: A practical guide). The standard score of a sample x is $z = \frac{(x-\mu)}{s}$ where μ is the mean of training samples and s is the variance.

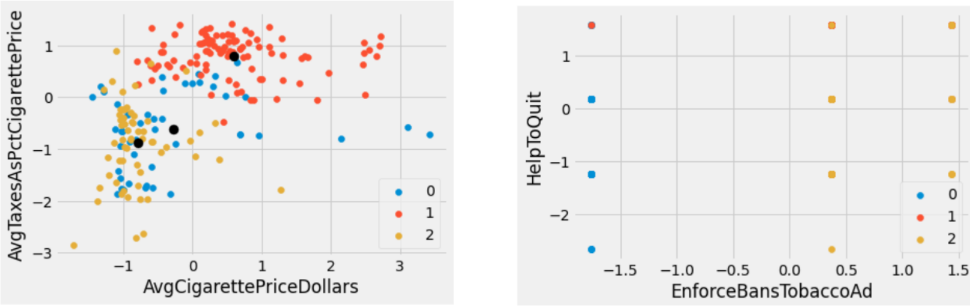
The "AvgCigarettePriceDollars", "AvgTaxesAsPctCigarettePrice", "EnforceBansTobaccoAd", and "HelpToQuit" fields were used as features for this model. The results from the elbow method and the silhouette coefficient are shown below (Figure 11). Using kneed's KneeLocator and visual assessment of these graphs, the best number of clusters was determined to be either three or eight. Due to ease of interpretability, three clusters were used instead of eight, with the graphical results from three clusters shown below (Figure 12. Note that the values in this figure are scaled rather than original values. Since a higher value for all four features was deemed more favourable, it was determined that label 0 corresponded to mediocre implementation, label 1 was great implementation except for banning tobacco ads, and label 2 was for poor implementation except for banning tobacco ads. Between 2012 and 2014, most countries stayed within the same category except for India, Morocco, North Macedonia, South Africa, Pakistan, Russia, and Senegal (Figure 13).



(a) Elbow Curve Plot.

(b) Silhouette Coefficient Plot.

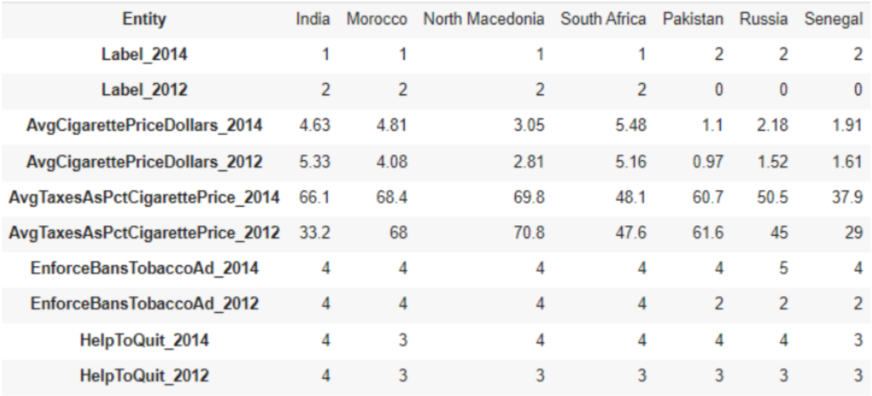
Figure 11: Plot of the evaluation results for various cluster sizes.



(a) "AvgCigarettePriceDollars" and "AvgTaxesAsPctCigarettePrice".

(b) "EnforceBansTobaccoAd" and "HelpToQuit". Plot.

Figure 12: Plot of the feature relationships for the three clusters



Entity	India	Morocco	North Macedonia	South Africa	Pakistan	Russia	Senegal
Label_2014	1	1	1	1	2	2	2
Label_2012	2	2	2	2	0	0	0
AvgCigarettePriceDollars_2014	4.63	4.81	3.05	5.48	1.1	2.18	1.91
AvgCigarettePriceDollars_2012	5.33	4.08	2.81	5.16	0.97	1.52	1.61
AvgTaxesAsPctCigarettePrice_2014	66.1	68.4	69.8	48.1	60.7	50.5	37.9
AvgTaxesAsPctCigarettePrice_2012	33.2	68	70.8	47.6	61.6	45	29
EnforceBansTobaccoAd_2014	4	4	4	4	4	5	4
EnforceBansTobaccoAd_2012	4	4	4	4	2	2	2
HelpToQuit_2014	4	3	4	4	4	4	3
HelpToQuit_2012	4	3	3	3	3	3	3

Figure 13: Summary of the countries with deviations in their *MPOWER* cluster labels.

Based on the two K-Means models, it appears there were several who have seen improvement in controlling tobacco through the usage of MPOWER: India, Morocco, South Africa, Pakistan, and Senegal. On the other hand, countries that have seen success through other means include Nepal and Peru. Countries that have seen minimal improvement despite implementing *MPOWER* measures include North Macedonia, and Russia.

3.3 Linear Regression

Several iterations of linear regression models were conducted to determine the predictive power of *MPOWER* and potential avenues of improvement for this framework. The first few iterations leveraged the provided **tobacco_use_ww** and **stop_smoking** data sets. However, the predictive power of the models derived from these two data sets were limited and yielded low coefficients of determination (r-squared). As such, other data from WHO was combined with the provided data sources to improve model performance. Yet, the results still showed low predictive power of tobacco usage from these detailed and more comprehensive *MPOWER* features. Based on these charts and low r-squared values, the team also explored related metrics to *MPOWER* such as taxes as share of cigarette price and share of tobacco retail price that is tax. These features showed promise with taxes as share of cigarette price having an r-squared value of 0.1172 on its own and share of tobacco retail price possessing an r-squared value of 0.0651 by itself. The team also looked at the predictive power of features relating to national health such as the current health expenditure, central government expenditure, gross domestic product, and population of the country. Only general government expenditure as % of gross domestic product had a high r-squared value on its own (around 0.1011). Please refer to the Appendix and code files for more details on these model iterations.

Initially, the team wanted to leverage the features which had more than 0.05 r-squared value on their own and the parent of the country (through one-hot encoded features). However, due to the discrepancies between the years of the data, some of these features were omitted as merging them with the current data would significantly reduce the data size.

The optimal model included all kinds of means of *MPOWER*, regional differences, and government expenditure ratio with a coefficient of determination of 0.5619 (Section 3.3). There are three main features in this model: regional difference, *MPOWER*, and government expenditures. Below are findings when optimizing the model to have the greatest impact on tobacco usage decrease (Section 3.3). Please refer to the Appendix and code files for more details on the intermediate model versions.

- **Regional difference:** Region was a key indicator of this model, which means that even with the positive effect of *MPOWER*, thinking of cultural difference would be significant.
- **MPOWER:** *MPOWER* was not the greatest factor, but still had a positive impact on decreasing usage. Raising taxes has relatively more impact on the model.
- **Government expenditure:** Increase in government expenditure is highly recommended to help people not to use tobacco. Surprisingly, expenditure on medical expenses did not make a significant difference. But, an overall level of government expense is needed.

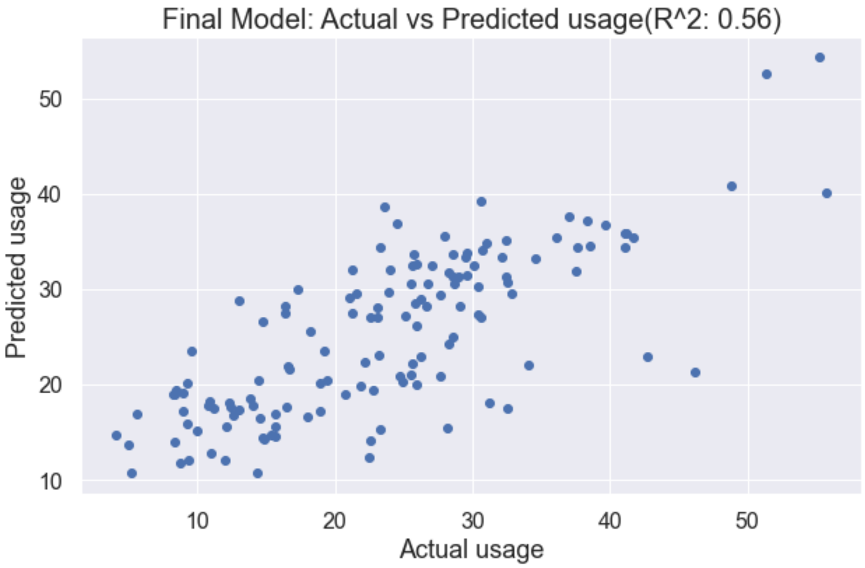


Figure 14: Visual comparison of the actual vs predicted usage from the final model.

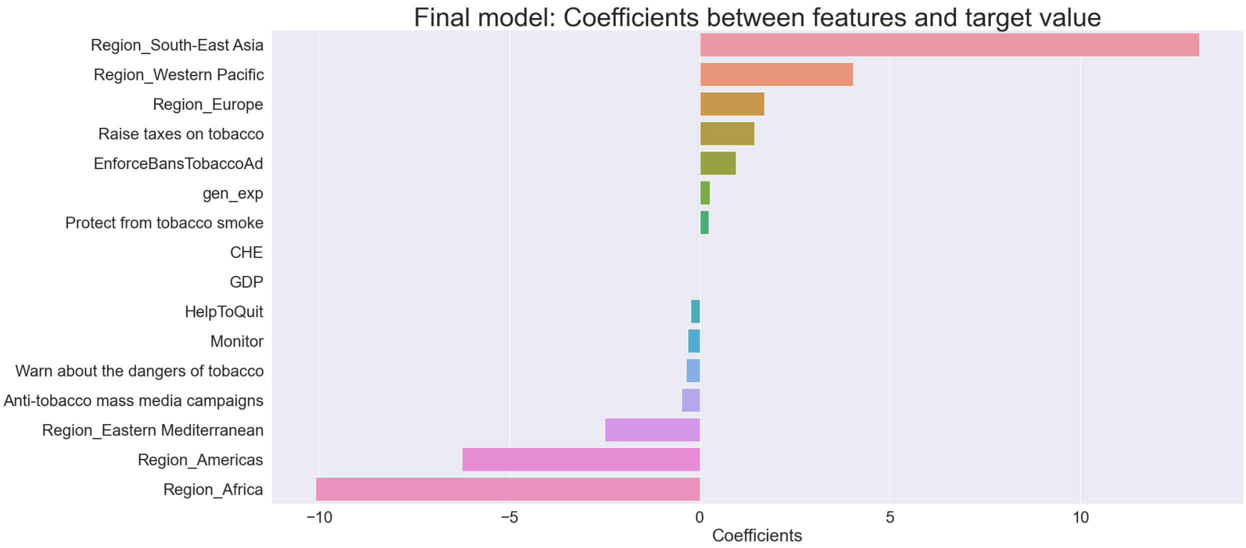


Figure 15: Visual comparison of the coefficient weights in the final model.

The preprocessing of the provided data sets used for the final model was as follows.

- **Stop_smoking.csv** had 774 rows with 7 columns, having 4 years(2007, 2010, 2012, and 2014) of 194 countries' data. There were no duplicates, but 2 columns(average price on tobacco, average tax price on tobacco) had null values that are more than 70% of total values. Since those columns didn't have sufficient data, the team decided to move the columns.

- **Tobacco_use_ww.csv** had 4023 rows with 7 columns, having 9 years(2000, 2005, 2010, and 2013 2018) of 447 data each. There were no duplicates, and no null values as well. It contained 149 countries' data for 3 different measures which were 'Both sexes', 'Male', 'Female'. In terms of getting an average usage, the team decided to just use 'Both sexes' as a measurement.

- Combining above those two data set on ‘Entity’ and ‘Year’, the team got intersection of only two years(2010, 2014). Merged data set contained 274 rows with 10 columns, having 2 years(2010, 2014) of 137 countries’ data. There were no duplicates, and no null values as well.

The preprocessing of additional data sets is described below.

- **M_group,P_group,O_group,W_group,W-MM_group,E_group,R_group.csv** had 1365 rows with 9 columns, having 7 years(2007, 2008, 2010, 2012, 2014, 2016, 2018) of 195 countries’ data. There were no duplicates, and no null values as well. It had 7 columns of *MPOWERs*, which were ‘Monitor’, ‘Protect from tobacco smoke’, ‘Offer help to quit tobacco use’, ‘Warn about the dangers of tobacco’, ‘Enforce bans on tobacco advertising’, ‘Raise taxes on tobacco’, ‘Anti-tobacco mass media campaigns’. Since ‘Enforce bans on tobacco advertising’ and ‘Offer help to quit tobacco use’ were duplicated in data set **Stop_smoking.csv**, the team removed those two columns and merged them to the combined dataset. So, in this data set the team used 5 columns as additional features for the model.
- **NHA_indicators.xlsx** had 1152 rows with 21 columns. There were no duplicates, and no null values as well. This data set was intended to be used for merging to previous data set, so the team filtered this data set by year 2010 and 2014, and used indicator of ‘Current Health Expenditure (CHE) per Capita in US$’, ‘General Government Expenditure (GGE) as % Gross Domestic Product (GDP)’, and ‘Gross Domestic Product (GDP) per Capita in US$’. Data set from each indicator contained 384 rows with 3 columns(192 countries for two years (2010, 2014)), with no duplicates or null values.
- By merging two additional data set to the previous combined data set(provided by Datathon), the team got 272 rows with 17 columns, having 2 years(2010, 2014) of 136 countries’ data set. There were no duplicates, and no null values as well.

Specifics on the modelling procedure used for the final model are listed below.

- The team used linear regression, and got a coefficient of determination: 0.5619(R-Squared). Also used the train data set as year 2010(136 rows), and test data set as year 2014(136 rows), and set 16 features and 1 target value.
- There were 16 features:
  - **Datathon(8, from stop_smoking.csv)**: *MPOWER* (2, “EnforceBansTobaccoAd”, “HelpToQuit”), Region(6, one-hot encoding))
  - **WHO data set(8, Global Health Expenditure Database)**: Additional *MPOWER* (5, Monitor, Protect, Offer, Warn, Enforce, and Raise tax), Other sources (3, GDP per capita, Current Health Expenditure(by government), General Government Expenditure)
- There was one target value:
  - **Datathon(tobacco_use_ww.csv)**:Tobacco usage percentage of both sexes
- For splitting the train and test data set, the team assigned the train data set as year 2010((136 rows), and test data set as year 2014((136 rows). Since splitting randomly might result in some situations like predicting past year by training future year, and this doesn’t make sense.

4 Conclusion

Through the use of both unsupervised learning (K-Means clustering) and supervised learning (Linear Regression), the predictor power of *MPOWER* in determining the tobacco usage of a country was assessed. Leveraging additional features about regional differences, and government expenditures, we present suggestions on ways to improve *MPOWER* and subsequently, decrease smoking usage across the world. We hope that considering the importance of these additional attributes along with *MPOWER* will diminish this global epidemic and lessen the public health threat that tobacco usage poses on the world.

5 Work Cited

"Clustering." Scikit-Learn, https://scikit-learn.org/stable/modules/clustering.htmlk-means.

Fatal Errors, https://www.fatalerrors.org/a/0dh01j8.html.

Global Health Expenditure Database. World Health Organization, https://apps.who.int/nha/database/Select/Indicators/en.

Kaleta, Dorota et al. "MPOWER-strategia na rzecz walki ze światowa epidema uzywania tytoniu" [MPOWER–strategy for fighting the global tobacco epidemic]. Medycyna pracyvol. 60,2 (2009): 145-9.

Real Python. "K-Means Clustering in Python: A Practical Guide." Real Python, 8 Jan. 2021, https://realpython.com/k-means-clustering-python/.

Roser, Max, and Esteban Ortiz-Ospina. "Tertiary Education." Our World in Data, https://ourworldindata.org/tertiary-education.

6 Appendix

Using **tobacco_use_ww** and **stop_smoking**, linear regression modeling was applied to determine the optimal level of indicator levels. However, the predictive power of models derived from only the two data sets was limited, therefore other data from WHO was combined with the provided data sources to improve model performance.

Features:

- *MPOWER* (Monitor, Protect, Offer, Warn, Enforce, and Raise)
- GDP per capita
- Region
- Medical expenditure ratio (by government)

Target Value

- Tobacco Usage Percentage of Both Sexes

After merging **tobacco_use_ww** and **stop_smoking**, the only data remaining came for 2010 and 2014. "EnforceBansTobaccoAd" and "HelpToQuit" were taken as features and 'Value' as the target. In addition, the data from 2010 was used for training while the data from 2014 was set aside for evaluative purposes. If the data was split randomly, the possible scenario of 2014 data being used to predict 2010 values is inherently illogical. The first model did not provide minimal insight due to the lack of features (Figure 16).

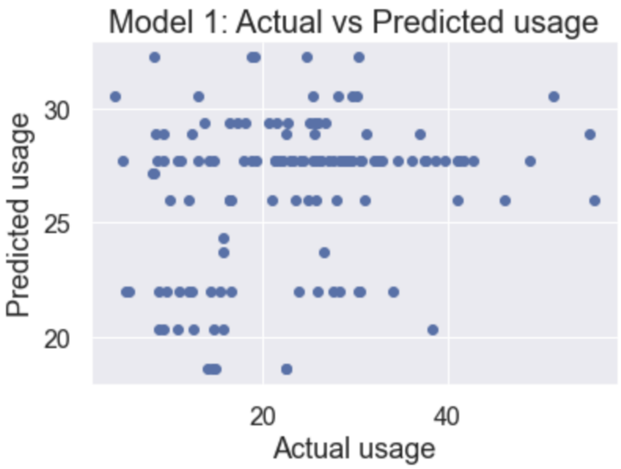


Figure 16: Model 1 with “EnforceBansTobaccoAd” and “HelpToQuit”

The second model was an attempt to improve performance by converting the target value from a percentage to number of people. In order to do so, population data from WHO was multiplied by the percentage value given. The population was then log-scaled to ensure all countries were comparable. However, the resulting model offered no advantages to Model 1 (Figure 17). Even for low *MPOWER* values, there were some countries with relatively low tobacco usage values (Figure 18). These countries were nearly all located in Africa, therefore the field “ParentLocation” was added (Figure 19).

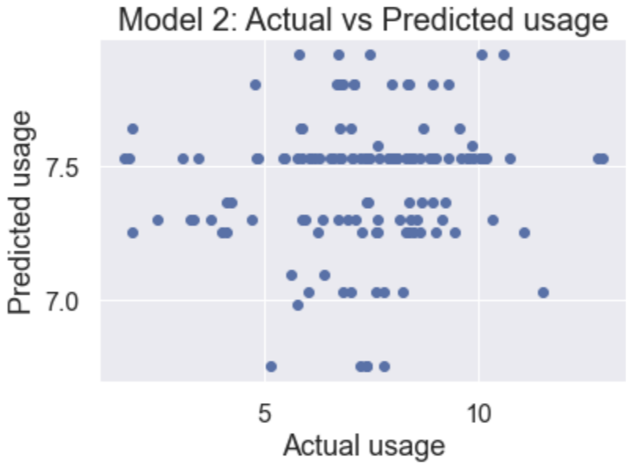


Figure 17: Model 2 with Number of People as Target Unit

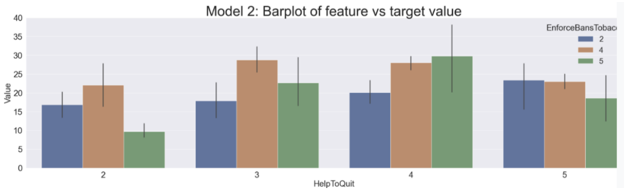
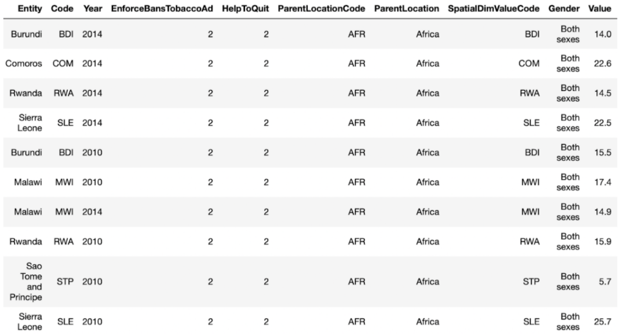


Figure 18: Distribution of “HelpToQuit”-Value



Entity	Code	Year	EnforceBansTobaccoAd	HelpToQuit	ParentLocationCode	ParentLocation	SpatialDimValueCode	Gender	Value
Burundi	BDI	2014	2	2	AFR	Africa	BDI	Both sexes	14.0
Comoros	COM	2014	2	2	AFR	Africa	COM	Both sexes	22.6
Rwanda	RWA	2014	2	2	AFR	Africa	RWA	Both sexes	14.5
Sierra Leone	SLE	2014	2	2	AFR	Africa	SLE	Both sexes	22.5
Burundi	BDI	2010	2	2	AFR	Africa	BDI	Both sexes	15.5
Malawi	MWI	2010	2	2	AFR	Africa	MWI	Both sexes	17.4
Malawi	MWI	2014	2	2	AFR	Africa	MWI	Both sexes	14.9
Rwanda	RWA	2010	2	2	AFR	Africa	RWA	Both sexes	15.9
Sao Tome and Principe	STP	2010	2	2	AFR	Africa	STP	Both sexes	5.7
Sierra Leone	SLE	2010	2	2	AFR	Africa	SLE	Both sexes	25.7

Figure 19: Tabular summary of countries with low *MPOWER* implementation but low tobacco usage.

Region categories were then converted into numerical values using one hot encoding (Figure 20). The resulting model had a significantly higher coefficient of determination: 0.404

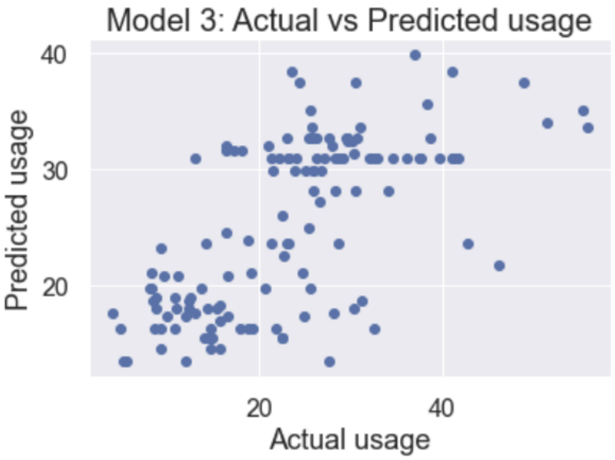
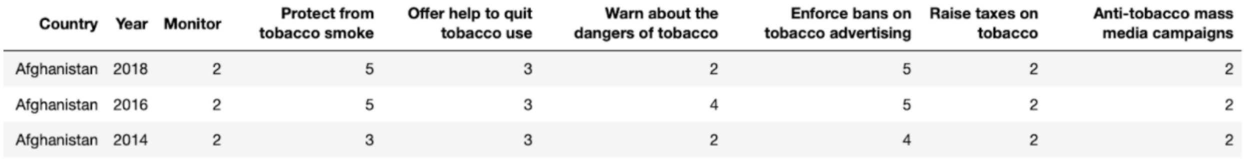


Figure 20: Model 3 with Region categorical features

The fourth model added *MPOWER* summary indicators that were not included in **stop_smoking**. The data was sourced from WHO (Figure 21). Fields include all *MPOWER* measures, and so those columns were included as additional features. The resulting model had a coefficient of determination of 0.465 (Figure 22).



Country	Year	Monitor	Protect from tobacco smoke	Offer help to quit tobacco use	Warn about the dangers of tobacco	Enforce bans on tobacco advertising	Raise taxes on tobacco	Anti-tobacco mass media campaigns
Afghanistan	2018	2	5	3	2	5	2	2
Afghanistan	2016	2	5	3	4	5	2	2
Afghanistan	2014	2	3	3	2	4	2	2

Figure 21: Tabular summary of *MPOWER* summary indicator levels

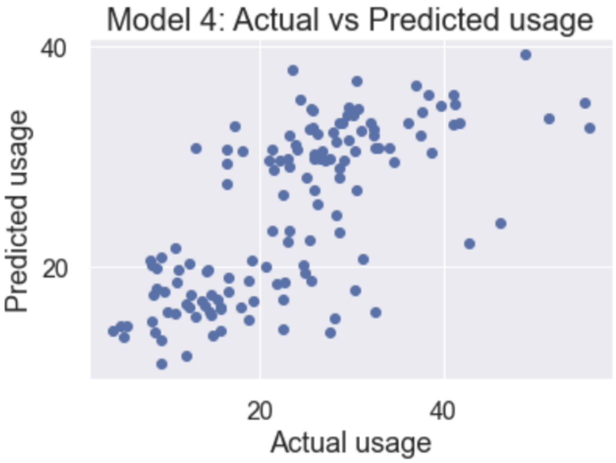


Figure 22: Model 4 with Additional *MPOWER* attributes

The fifth model added additional features in addition to *MPOWER* summary indicators. Data relating to medical infrastructure, government expenditure, and GDP per capita were sourced from WHO. The assumption was made that data of this type would influence the model. Three features were added, Current Health Expenditure (CHE) per Capita in US$, Gross Domestic Product (GDP) per Capita in US$, and General Government Expenditure (GGE) as % Gross Domestic Product (GDP). The resulting model had a coefficient of determination of 0.562 (Figure 23). In order to understand the model, coefficients relating features and target values were checked. Regional difference was a key indicator, and some *MPOWER* measures such as increasing tax rates also had an impact on the model. Surprisingly, Current Health Expenditure and GDP per capita did not have an impact on the model (Figure 24).

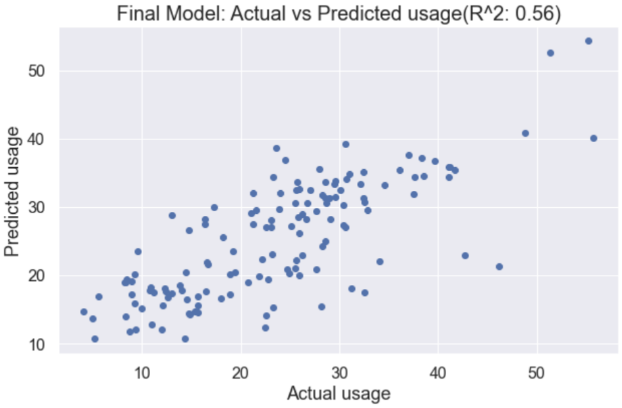


Figure 23: Model 5 with Government Expenditures Added As Features

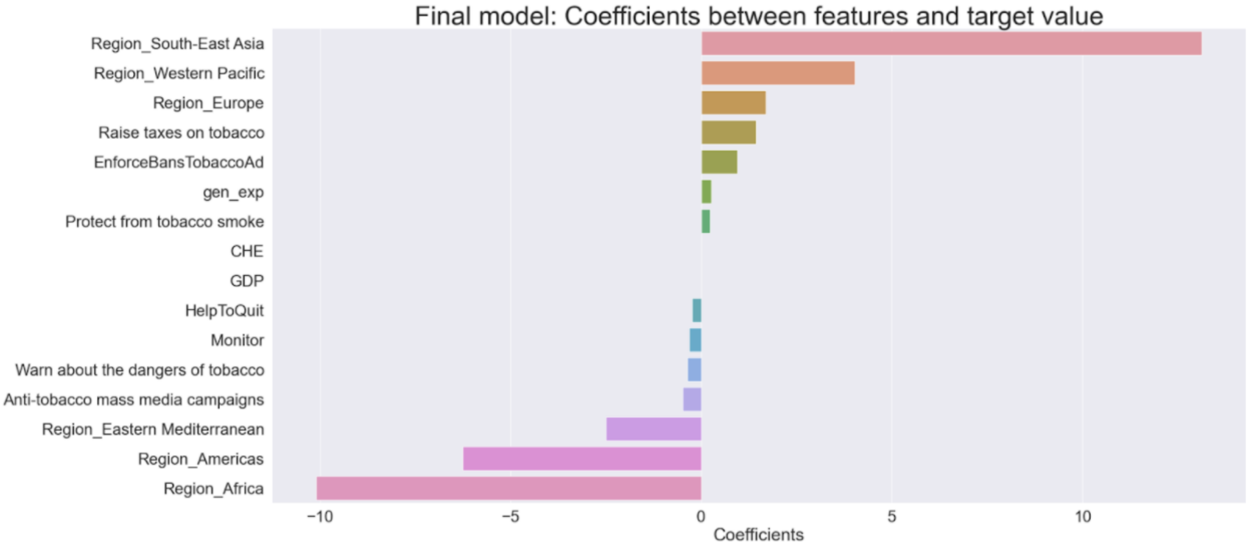


Figure 24: Feature Coefficients in Predicting Target Value

Additional models were created which incorporated features such as educational maturity. Based on the source that is originally provided by the World Bank (Tertiary Education), we were able to build a model. The model's predictive power saw no improvement in comparison to the previous model, with a coefficient of determination of 0.420.

The optimal model included all summary indicators of *MPOWER*, regional differences, and government expenditure ratio, resulting in a coefficient of determination of 0.562. The important features were:

- Regional difference: Region was a key indicator of this model, this may suggest regional and cultural qualities may impact tobacco usage regardless of *MPOWER* implementation.
- *MPOWER*: *MPOWER* had a positive impact on decreasing usage. Of the individual measures, raising taxes has the greatest impact.
- Government expenditure: Increasing government expenditure is highly recommended to reduce tobacco usage. Surprisingly, medical expenditure did not make a tangible difference.