Can Machine Learning Predict Consumers' Environmental Attitudes and Beliefs?

Kiana Yektansani[1], SeyedSoroosh Azizi[2]

Abstract

Individuals' beliefs and attitudes towards climate change have a significant impact on their support for environmental policy regulations and willingness to take voluntary actions. In this paper, we aim to predict public beliefs about global warming, anthropogenic climate change, support for an environmental tax policy, and likelihood of buying an energy efficient home appliance. We use the data from the European Social Survey and employ four machine learning techniques to predict individuals' environmental beliefs and attitudes. We can correctly identify more than 70% of the green respondents in different settings. For policymakers, being able to predict these preferences is crucial to successfully implement an environmental policy. These results help politicians avoid substantial resistance costs that can arise from an environmental policy without broad public support. This information is also helpful for green producers to predict consumers' willingness to pay and environmental preferences to deliver targeted marketing strategies and product features.

JEL classification: Q54, Q58, C53

Keywords: environmental attitudes, consumer behavior, machine learning

---

[1] University of Illinois at Chicago, Tel: 5093306063. Email: k.yektansani@gmail.com

[2] Purdue Northwest, Tel: 8156085308. Email: SeyedSoroosh.Azizi@gmail.com

1. Introduction

The observed increase in the globally averaged temperature since the mid-20th century is very likely due to the increase in anthropogenic greenhouse gas concentrations (IPCC Working Group II Fourth Assessment Report). Anthropogenic climate change has a significant negative impact on physical and biological systems globally (Rosenzweig et al., 2008) and marine ecosystems (Riou et al., 2011), it intensifies the potential for western US forest fire activity (Abatzoglou and Williams, 2016), and creates challenges and costs for societies worldwide (Hoegh-Guldberg and Bruno, 2010).

The significant and worldwide impacts of human activities on the climate and environment call for actions from policymakers. On the other hand, climate change policy entails significant economic and lifestyle changes for residents of a country and demands substantial sacrifice from the public (e.g., see Söderholm, 2012; Sutherland, 2000). Thus, the direction and strength of public opinion is a critical factor in developing an appropriate policy response.

Individuals' beliefs about climate change and its underlying causes affect their support for environmental policy regulations and willingness to take voluntary actions. For instance, attitudes towards climate change have a strong influence on the levels of support for an emissions trading scheme (Pietsch and McAllister, 2010; Kotchen et al., 2013), a carbon tax, or a GHG regulation (Kotchen et al., 2013).

Several papers have studied the factors influencing individuals' attitudes towards climate change. For example, Ziegler (2017) finds that while environmental values are the major factors for climate change beliefs in USA, Germany, and China, conservative identification in the USA still has a significant negative effect on beliefs in general climate change as well as anthropogenic climate change. McCright and Dunlap (2011) also find that conservative

white males in the US are significantly more likely to deny climate change and the effect is even stronger among those who self-report great understanding of global warming. Other research finds that outdoor temperature, being exposed to heat-related primes, high anchor (Joireman et al., 2010), environmental beliefs about earth's limited resources, and humans' interaction with the nature are associated with one's approach to climate change urgency (Gadenne et al., 2011).

However, none of these papers have focused on predicting the values of individuals' attitudes towards climate change on a new dataset (or part of the dataset that is not used for building the model). Identifying the underlying factors shaping individuals' attitudes and beliefs is useful when the officials have the means and intensions to influence them; e.g., for developing public education programs to educate the public about causes of climate change and gain public support for various policies (O'Connor et al. 2002). In reality, a lot of times, what matters most is identifying the final value of consumer's orientations rather than the forces affecting them. For instance, for policymakers, the information about consumers environmental preferences is important as such preferences shape the optimal policy response (see Espinola-Arredondo and Zhao, 2012; Bansal and Gangopadhyay, 2003). Our study aims to fill this gap by focusing on predicting individuals' orientations.

In addition to designing the optimal environmental policy, predicting such attitudes and beliefs has other important benefits for policymakers. The lack of public support can be an obstacle in implementation of effective environmental policies and achieving environmental goals (Kallbekken et al., 2011). By knowing people's standpoint (which is our focus in this study), politicians can avoid numerous resistance costs that can arise from an environmental policy that is not backed by broad public support. For instance, during the Yellow Vest

movement that started in 2018, France was forced to cancel the fossil fuel tax increase in the aftermath of increasingly violent protests (Rubin and Sengupta, 2018). The damage claims associated with this movement was estimated to be €170 million by the French Insurance Federation ("French 'yellow vest' demos", 2019). Canada is in a similar position because of the backlash against increasing carbon tax to fight climate change: "Trudeau now needs to figure out how much, and how quickly, Canadians actually want climate action" (Forrest, 2020, para. 5).

Therefore, it is critical for the authorities to know about consumers' potential reaction to a certain policy before incurring the developing and implementation costs. And if the prospects of consumers reaction are not bright, they can undertake programs to change them before incurring such costs.

For green producers, it is crucial to be able to identify consumers' beliefs and environmental preferences so that they can deliver customized products and targeted advertising depending on the specific market they face. For example, Gadenne et al. (2011) find that individuals who are concerned about global warming are more likely to have favorable attitudes toward environmental behaviors and purchases and there is a strong association between environmental attitudes and energy saving behaviors. According to O'Connor et al. (1999), individuals' beliefs about likelihood of climate change is a strong predictor of their willingness to take voluntary actions (e.g., carpooling, installing more insulation, or using more energy efficient appliances) and support government policies (e.g., higher taxes to reduce CO2 emissions or rainforest preservation). Research findings by Carlsson et al. (2012) suggest that disbelief in global warming has a significant and negative impact on the probability of stating a positive willingness to pay (WTP) for reducing CO2 emissions. In

addition, the belief that climate change is caused by humans is effectively what separates those with a positive WTP for CO2 emission reductions from others. In conclusion, environmental attitudes and beliefs are common explanatory variables in surveys to estimate WTP for climate policy (Nemet and Johnson, 2010).

For green producers, targeting specific marketing strategies to potential green consumers with higher environmental awareness and WTP is more efficient than delivering the same strategy to the entire population (Mostafa, 2009). Producers of environmentally friendly products need to profile and segment the population and target each segment based on their WTP and climate change beliefs. Poorly designed marketing methods and product features arising from ignoring the differences in consumers' environmental awareness and concerns can be costly for firms. This was the case for Whirlpool, a home appliance company, when they realized consumers would not pay a price premium for a CFC-free refrigerator because they did not know what CFCs were (Singh and Pandey, 2012).

While econometrics aims to discover causality and inference (e.g., what are the major determinants of individuals' environmental preferences, what is the effect of education on environmental beliefs, etc.), machine learning is a useful tool for prediction. With a set of machine learning techniques, this study aims to predict individuals' beliefs about climate change, the share of anthropogenic activities on climate change, individuals' attitudes towards an environmental tax policy, and likelihood of buying energy efficient home appliances using their self-reported characteristics. The results will help policymakers and green producers identify the public's and/or consumers' viewpoints and potential reactions to a new policy, green product, or green marketing strategy.

The organization of the paper is as follows. Section 2 presents the data, section 3 describes the model and methodology, section 4 outlines the prediction results, and section 5 contains some concluding remarks.

2. Data

The data for this study is taken from round 8 of the European Social Survey (ESS) that was conducted in 2016 and released in 2018. In face-to-face interviews, the ESS measures attitudes towards a wide range of areas including media use, politics, climate change, welfare, and health in more than thirty European nations. The original dataset has 44387 observations and 534 variables. We remove the observations with missing values, meaning the number of used observations in each model varies. Table 1 includes a list of the dependent variables (group 1) and selected independent variables (group 2). The variables of the dataset we use are about: time spent on news and internet, political and civic engagement, social interactions, gender, age, socio-demographics, environment, energy, and climate change[3]. The dependent variables we analyze are respondents' opinion about climate change, the role of human activity on climate change, increasing taxes on fossil fuels, and purchasing an energy efficient home appliance. The respondents can choose between a set of options (e.g., strongly in favor to strongly against). For simplicity, we code all the dependent variables into binary variables.

[3]In selection of these variables, we have followed the literature about their potential impact on individuals' environmental beliefs and attitudes. For example, knowledge, beliefs about human responsibility, volume of news media coverage (Krosnick et al., 2006), demographic characteristics (O'Connor et al., 1999), income, education, political views (Nemet and Johnson, 2010), and altruistic values (Mostafa, 2009) are some of the variables mentioned in the literature. We have also used variable importance values reported by random forest for the selection of our variables. In addition, the variables with minimal missing values were added because they do not hurt the model and can potentially improve the predictions.

Group	Variable	Definition	Mean	Min	Max
1	ccnthum[4]	Climate change caused by natural processes, human activity, or both	3.404	0	5
	clmchng[5]	Do you think world's climate is changing	3.496	1	4
	eneffap[6]	How likely to buy most energy efficient home appliance	7.874	0	10
	inctxff [7]	Favor increase taxes on fossil fuels to reduce climate change	2.833	1	5
2	agea	Age of respondent	50.76	15	98
	Clmthgt3	How much thought about climate change before today	3.308	1	5
	edulvlb	Highest level of education	13.61	1	28
	eduyrs	Years of full-time education completed	13.3	0	54
	eisced	Highest level of education, ES - ISCED	4.123	1	8
	eiscedf	Father's highest level of education, ES – ISCED	3.053	1	8
	gvslvol	Standard of living for the old, governments' responsibility	8.17	1	11
	happy	How happy are you	7.556	1	11
	hinctnta	Household's total net income, all sources	5.619	1	10
	impfun	Important to seek fun and things that give pleasure	3.27	1	6
	imptrad	Important to follow traditions and customs	3.449	1	6
	inprdsc	How many people with whom you can discuss intimate and personal matters	3.781	1	7
	iphlppl	Important to help people and care for others well-being	3.663	1	6
	nwspol	News about politics and current affairs, watching, reading or listening, in minutes	84.69	0	1410
	pplhlp	Most of the time people helpful or mostly looking out for themselves	5.603	1	11
	rdcenr	How often do things to reduce energy use	5.022	1	7
	sclmeet	How often socially meet with friends, relatives or colleagues	4.863	1	7
	wrenexp	How worried, energy too expensive for many people	3.668	1	5

Table 1. List of dependent and independent variables

[4] In our coding 0 indicates “I don't think climate change is happening”, 1 indicates “entirely by natural processes” and 5 indicates “entirely by human activity”.
[5] In our coding 1 indicates “definitely not changing” and 4 indicates “definitely changing”.
[6] In our coding indicates 0 “not at all likely” and 10 indicates “extremely likely”.
[7] In our coding 1 indicates “strongly against” and 5 indicates “strongly in favor”.

3. Model and Methodology

Suppose that $X = (X_1, \ldots, X_p)$ are the independent variables capturing respondents' characteristics and Y is the response variable which measures respondents' environmental beliefs and preferences. Assume the number of observation is represented by n and x_{ij} represents the value of the jth predictor for the ith observation where $i = 1, 2, \ldots, n$ and $j = 1,2, \ldots, p$.

Also assume that $Y = f(X) + \varepsilon$ where f is an unknown but fixed function of X and ε is a random error term with mean 0. Note that in machine learning the goal is prediction and we are not concerned about the correlation between ε and X. Our goal is to predict Y using X:

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ is our estimate of f and $\hat{Y}$ is our prediction of Y. In machine learning, $\hat{f}$ is treated as a black box, meaning one is not concerned about the particular functional form of $\hat{f}$ (James et al., 2013). All that matters is that $\hat{f}$ provides an accurate prediction of Y, how individual independent variables are associated with the dependent variable is not relevant in the prediction problem. The response variable is:

$$Y = \begin{cases} 0, & if\ the\ respondent\ is\ green \\ 1, & if\ the\ respondent\ is\ brown. \end{cases}$$

We divide our dataset into a training and test datasets. The observations in the training data are used to train the model while the test data is used to examine the performance of the model. Since our response variables are qualitative or categorical, we use classification techniques. The classifiers we use in this paper are logistic regression, random forest, boosting, and neural networks. We briefly describe each method below.

Since our response variable is qualitative and binary, we use logistic regression rather than linear regression.[8] Logistic regression models the probability that Y belongs to a particular category, e.g.,

$$\Pr(Y = 1|X) \equiv p(X).$$

One might predict $Y = 1$ for any individual for whom $p(X) > 0.5$. The 0.5 threshold is not fixed and can change depending on the circumstances. For example, if the cost of incorrectly classifying someone as green is high (e.g., due to high cost of targeted product feature design), we can take on a more conservative approach in predicting individuals who are green and choose 0.7 as our threshold. If, however, the cost of incorrectly classifying someone as brown is high (e.g., due to high cost of carried educational programs), we can choose 0.3 as the threshold for our analysis. In addition, when a dataset is imbalanced (i.e., the majority of data are from one class), the learning algorithm tends to label everything as the majority class. In this setting, varying the decision threshold helps with imbalanced data (Maloof, 2003).

Next, we move to random forest which is a tree-based method. Random forest constructs multiple decision trees and averages over them using a bootstrapped dataset. Boosting is another model that uses decision trees to make predictions. In boosting, the trees are grown sequentially and each tree is influenced by the errors made by the preceding tree. The last model that we use is neural networks (NNs). In NNs, each neuron is a node connected to other nodes via links. Most NNs have three types of layers made of neurons: input, hidden, and output layers. Every NN has one input layer and one output layer. The number of neurons in the input layer is equal to the number of

[8] This is because we want to avoid the problem of predicting below zero or above one probabilities that is possible when we fit a straight line to a binary response variable coded as 0 and 1.

predictors in the model. The number of neurons in the output layer of a classifier NN depends on the number of classes. For the hidden layers, we follow the suggestion by Heaton (2008, p. 158) and use one hidden layer for the NN predictions of this paper.

There are several methods to evaluate the performance of a classifier. A classifier's accuracy is defined as the percent of correct classifications and the error rate is the percent of incorrect classifications (accuracy $= 1 -$ error rate). One shortcoming of the accuracy is that it assumes equal costs for all types of misclassification of respondents.

A binary classifier can make two types of classification errors for a given threshold value: type I error happens when the brown respondents are incorrectly classified as green. Related to type I error is specificity (also called true negative rate) is the fraction of brown individuals that are correctly identified, and it equals to 1−Type I error. Type II error happens when green individuals are misclassified as brown. Related to type II error is sensitivity (also called the true positive rate) is the portion of green respondents that are correctly classified as such, and it equals to 1−Type II error. A confusion matrix is used to illustrates this information (Hlaváč, 2016). Based on the confusion matrix (Table 2Table 2), accuracy is $\frac{a+d}{a+b+c+d}$, sensitivity is $\frac{d}{c+d}$, and specificity equals $\frac{a}{a+b}$.

		Predicted	
		0	1
Actual observation	0	a: TN: True Negative	b: FP: False Positive
	1	c: FN: False Negative	d: TP: True Positive

Table 2. Confusion matrix

Another tool used to assess the performance of a binary classifier is a receiver operating characteristic (ROC) curve. The ROC curve is created by plotting sensitivity on the y-axis against the 1-specificity on the x-axis as we vary the discrimination threshold of the classifier. The area under the ROC curve (AUC) illustrates the overall performance of a classifier summarized over all possible thresholds. Higher AUC values signal a better classifier.

4. Prediction Results

The prediction results are summarized in the tables below. Each table has a different dependent variable and different set of predictors. The cutoff for classification is varied in each table in order to get the best combinations of sensitivity, specificity, and accuracy. In addition, the response variables are coded into binary variables.

4.1. Beliefs about climate change

The first response variable we analyze is clmchng, i.e., whether the respondent thinks that the world's climate is changing. The respondents choose between definitely changing, probably changing, probably not changing, and definitely not changing. After removing the observations with missing values, we are left with 25093 observations for our classification problem. The table below shows the share of respondents in each category:

Definitely changing	14434
Probably changing	9109
Probably not changing	1060
Definitely not changing	490

Table 3. Observations' distribution for clmchng

The predictors are the features in group 2 of Table 1. To keep our response variable binary, climate change skeptics, i.e., the respondents who choose any of the last three options are grouped together and coded as zero:

$$Y = \begin{cases} 0, & \text{climate change skeptics} \\ 1, & \text{climate change believers.} \end{cases}$$ [9]

Table 4 summarizes the prediction results. Random forest has a higher sensitivity, specificity, accuracy, and AUC values than logistic regression, meaning logistic regression is outperformed by random forest. Boosting correctly identifies 76% of the climate change skeptics while correctly classifying more than half of climate change believers. Random forest correctly identifies 74% of climate change believers while correctly classifying more than half of climate change skeptics. Depending on the costs of misclassification of each class, we can use either of these algorithms. For example, if the costs of incorrectly classifying climate change skeptics as believers (i.e., the cost of type I error) is high,[10] we should use boosting to have a high value of specificity. If, on the

[9] In creating dummy dependent variable, we tried to reduce the imbalance in the data while keeping the economic intuition. In the appendix, we redo these predictions with a different threshold for the dependent variables.

[10] E.g., a politician in a country with frequent protests should not underestimate the number of climate change skeptics when enacting a tough climate change policy since the potential resistance costs may be very high. For instance, France has the highest number of average annual protests among Western European countries (Nam, 2007), signaling French authorities to pay close attention to climate change skeptics.

other hand, the costs of misclassifying climate change believers (i.e., the cost of type II error) is high,[11] we should use random forest instead of boosting to have a high value of sensitivity.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.729	0.738	0.573	0.703
Specificity	0.524	0.584	0.764	0.553
Accuracy	0.641	0.672	0.655	0.639
AUC	0.626	0.661	0.669	0.628

Table 4. Prediction results; dependent variable: clmchng, independent variables: group 2 of Table 1

Figure 1 shows the importance of each predictor in the random forest algorithm of Table 4.[12] According to this figure, how much the respondents had thought about climate change before the day of the interview, their age, years of education, followed by how much time they spend following news about politics and current affairs contribute most to the model. With the help of Figure 1, we develop a new model with only the top three features as predictors. The results are shown in Table 5 where the response variable is still clmchng and the predictors are clmthgt3, agea, and eduyrs. With boosting for example, we can correctly identify 80 % of climate change skeptics and half of the climate change believers (i.e., random guessing for climate change believers). Random forest correctly classifies 70% of climate change believers while correctly classifying 58% of the skeptics.

[11]E.g., in a country with low frequency of annual protests, e.g., Iceland or Luxembourg (Nam, 2007), the resistance cost of misclassifying climate change skeptics is low which increases the relative costs of ignoring climate change. As another example, a green producer incurs a lot of unnecessary costs if they develop and implement an expensive advertising program to educate the already-aware customers of a market (who have been misclassified as skeptics) about the consequences of climate change. In these settings, underestimating the number of climate change believers has high costs and thus a higher value of sensitivity is more desirable.

[12] The measure used here is the MeanDecreaseGini which is an indicator of variable importance and is the total amount that the Gini index decreases by splits over a given predictor, averaged over all trees (James et al., 2013, p. 319). The Gini index is a measure of node purity; a small value of a Gini index indicates that a node contains mostly observations from a single class.

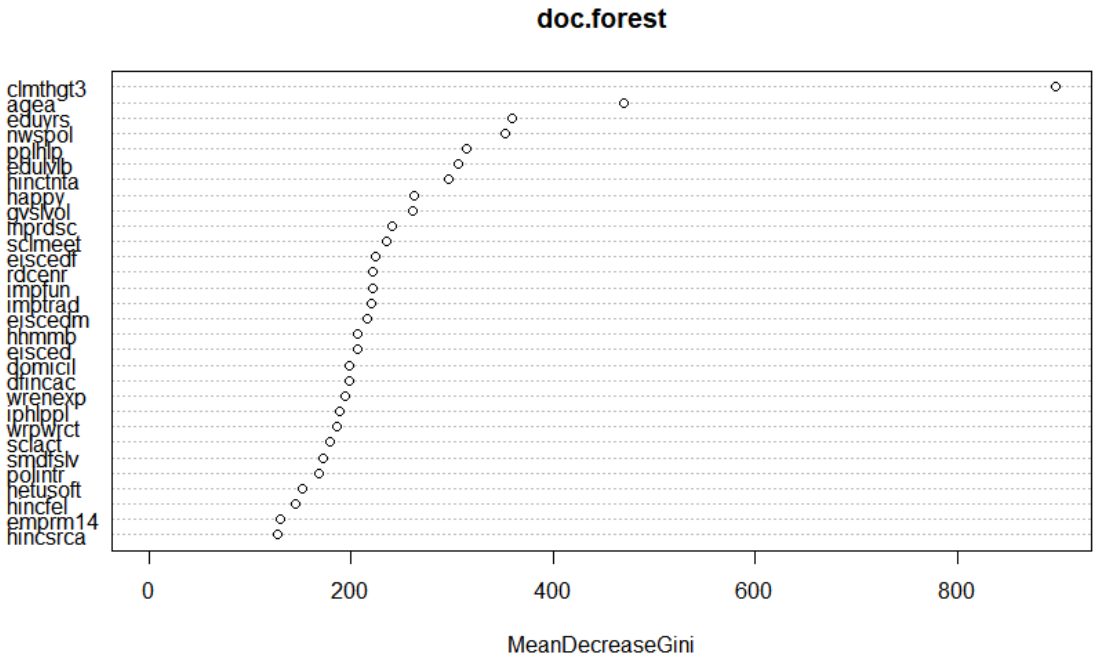

Figure 1. Variable importance plot for Table 4

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.682	0.701	0.504	0.454
Specificity	0.521	0.578	0.798	0.844
Accuracy	0.612	0.647	0.632	0.623
AUC	0.602	0.639	0.651	0.649

Table 5. Prediction results; dependent variable: clmchng, independent variables: top three predictors shown in Figure 1

4.2. Beliefs about the share of anthropogenic climate change

Next, we attempt to predict ccnthum, i.e., whether the respondent thinks that climate change is caused by natural processes, human activity, or both. The respondents choose among these options: entirely by human activity, mainly by human activity, about equally by natural processes and human activity, mainly by natural processes, entirely by natural processes, and I don't think

climate change is happening.[13] For our classification problem, after removing the observations with missing data, we are left with 24559 observations . Table 6 shows the share of respondents in each category:

entirely by human activity	1580
mainly by human activity	9540
about equally by natural processes and human activity	11243
mainly by natural processes	1684
entirely by natural processes	393
I don't think climate change is happening	119

Table 6. Observations' distribution for ccnthum

To transform the response variable into a binary variable, we group the respondents into a group of those who acknowledge significant anthropogenic climate change (i.e., they choose one of the first three options) and the deniers (i.e., they choose one of the last three options):

$$Y = \begin{cases} 0, & \text{deniers of anthropogenic climate change} \\ 1, & \text{supporters of anthropogenic climate change.} \end{cases}$$

The results are demonstrated in Table 7. The predictors in this table are the features in group 2 of Table 1. For example, logistic regression is able to accurately classify approximately 74% of the supporters of anthropogenic climate change while it correctly classifies 54% of deniers of anthropogenic climate change and has an overall accuracy rate of 71%.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.744	0.771	0.868	0.657
Specificity	0.543	0.530	0.395	0.629
Accuracy	0.711	0.731	0.789	0.653
AUC	0.644	0.651	0.631	0.643

[13] Even though this question was not asked from respondents who believe climate is definitely not changing (based on their answer to clmchng), still some respondents chose the last option mentioned above.

Table 7. Prediction results; dependent variable: ccnthum, independent variables: group 2 of Table 1

Figure 2 shows the importance of each predictor in the random forest algorithm of Table 7. According to this figure, the age of the respondent, how much they had thought about climate change before the day of the interview, how much time they spend following news about politics and current affairs, followed by years of education are the biggest contributors of the model. With the help of Figure 2, we develop a new model with only the top three features as predictors. The results are shown in Table 8 where the response variable is still ccnthum and the predictors are agea, Clmthgt3, and nwspol. Even though removing most of the features hurts the predictive ability of the classifiers, they are still informative in classifying the respondents using only three features. For example, logistic regression classifies 68% of the supporters of anthropogenic climate change while it correctly classifies 51% of the deniers of anthropogenic climate change. With neural networks, we can correctly identify 76% of climate change skeptics and almost half of the climate change believers. Depending on the misclassification costs, either of these approaches can be applied to predict individuals' beliefs about anthropogenic climate change.

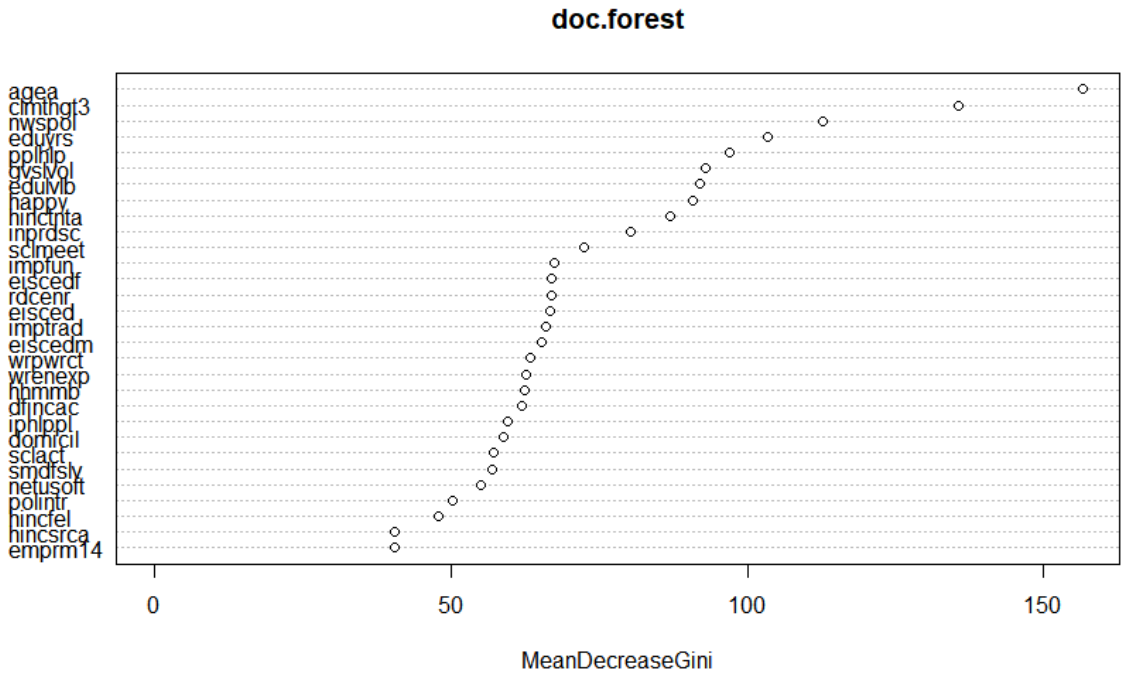

Figure 2. Variable importance plot for Table 7

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.684	0.855	0.857	0.497
Specificity	0.512	0.349	0.407	0.758
Accuracy	0.655	0.764	0.779	0.543
AUC	0.598	0.602	0.632	0.628

Table 8. Prediction results; dependent variable: ccnthum, independent variables: top three predictors shown in Figure 2

4.3. Attitudes towards an environmental tax policy

Next, we aim to predict inctxff, i.e., respondents' position on increasing taxes on fossil fuels such as oil, gas, and coal. The individuals choose among strongly in favor, somewhat in favor, neither in favor nor against, somewhat against, and strongly against. We remove the missing data and the remaining 24866 observations' distribution for inctxff is as follows in Table 9:

strongly in favor	2131
somewhat in favor	6555
neither in favor nor against	5410
somewhat against	6436
strongly against	4334

Table 9. Observations' distribution for inctxff

The respondents who chose the first three options are grouped together and coded as 1. Thus, the binary response variable is codes as below:

$$Y = \begin{cases} 0, & \text{against policy} \\ 1, & \text{in favor of or indifferent to policy.} \end{cases}$$

The independent variables selected for this model are the features in group 2 of Table 1. The prediction results are summarized in Table 10. With the given predictors, boosting and random forest outperform the other classifiers. Boosting for example, correctly classifies approximately 51% of tax supporters and 72% of antitax respondents.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.644	0.702	0.509	0.410
Specificity	0.523	0.515	0.719	0.752
Accuracy	0.592	0.620	0.602	0.561
AUC	0.585	0.609	0.614	0.581

Table 10. Prediction results; dependent variable: clmchng, independent variables: group 2 of Table 1

Figure 3 shows the importance of each predictor in random forest of Table 10. According to this figure, the age of the respondent, how much time they spend following news about politics and current affairs, years of education, followed by their opinion on whether people are mostly helpful or selfish are the most important features in predicting the respondents' position on an environmental tax policy. With this information, we construct a new model with only the top three features as predictors. The results are shown in Table 11 where the response variable is still inctxff and the predictors are agea, nwspol, and eduyrs. Removing the majority of the predictors reduces the predictive ability of the models, but they are still useful in providing some guidance to make predictions.

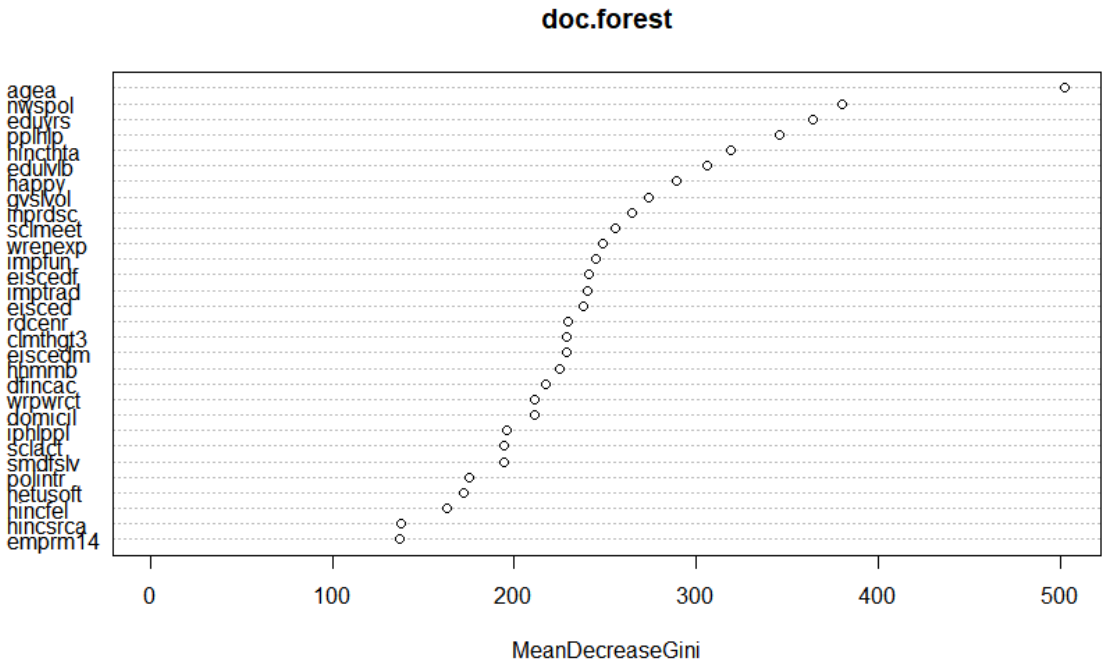

Figure 3. Variable importance plot for Table 10

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.617	0.732	0.492	0.414
Specificity	0.465	0.345	0.625	0.706
Accuracy	0.549	0.559	0.551	0.543
AUC	0.541	0.538	0.558	0.560

Table 11. Prediction results; dependent variable: inctxff, independent variables: top three predictors shown in Figure 3

4.4. Buying energy efficient home appliances

Next, we aim to predict eneffap, i.e., how likely an individual is to buy one of the most energy efficient models of a large electrical appliance for their home. This variable is the closest variable to measure WTP for an environmentally friendly product in the ESS questionnaire. So the prediction results of this variable is one of the more useful measures for green producers to predict WTP for green goods.

The individuals choose among not at all likely to extremely likely. We remove the missing data and the remaining 25113 observations' distribution for eneffap is as summarized in Table 12:

Not at all likely										Extremely likely
0	1	2	3	4	5	6	7	8	9	10
337	129	316	525	543	1852	1565	3004	5262	4349	7231

Table 12. Observations' distribution for eneffap

The respondents who chose the first six options are grouped together and coded as 0. The remaining respondents are grouped together and coded as 1. Thus, the binary response variable is codes as below:

$$Y = \begin{cases} 0, & \text{unlikely to buy green goods} \\ 1, & \text{likely to buy green goods.} \end{cases}$$

The independent variables selected for this model are the features in group 2 of Table 1. The prediction results are summarized in Table 13. With the given predictors, logistic regression, boosting, and neural networks have similar AUC (indicating that they have similar performance). Random forest has the highest AUC value and it correctly classifies 75% of the respondents who are likely to buy the green good while correctly classifying 56% of those who are unlikely to buy green goods. This has useful information for green producers to do market segmentation. After identifying green consumers, green firms can offer their green products to them without the need for substantial marketing expenditures.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.702	0.547	0.905	0.724
Specificity	0.537	0.754	0.331	0.512
Accuracy	0.678	0.579	0.819	0.692
AUC	0.619	0.650	0.618	0.618

Table 13. Prediction results; dependent variable: clmchng, independent variables: group 2 of Table 1

Figure 4 shows the importance of each predictor in random forest of Table 13. The age of the respondent, how much time they spend following news about politics and current affairs, followed by how often do they do things to reduce their energy use (e.g., switching off appliances that are not being used, walking for short journeys, or only using the heating or air conditioning when really needed) are the top three predictors of eneffap. Overall, the most important predictors are similar in all models, showing that we can get good predictions for environmental beliefs and attitudes by knowing the same set of limited predictors. In Table 14, we use the top three predictors as the only features to predict eneffap. Even though the models' prediction abilities drop, they are still informative in predicting the values of eneffap.

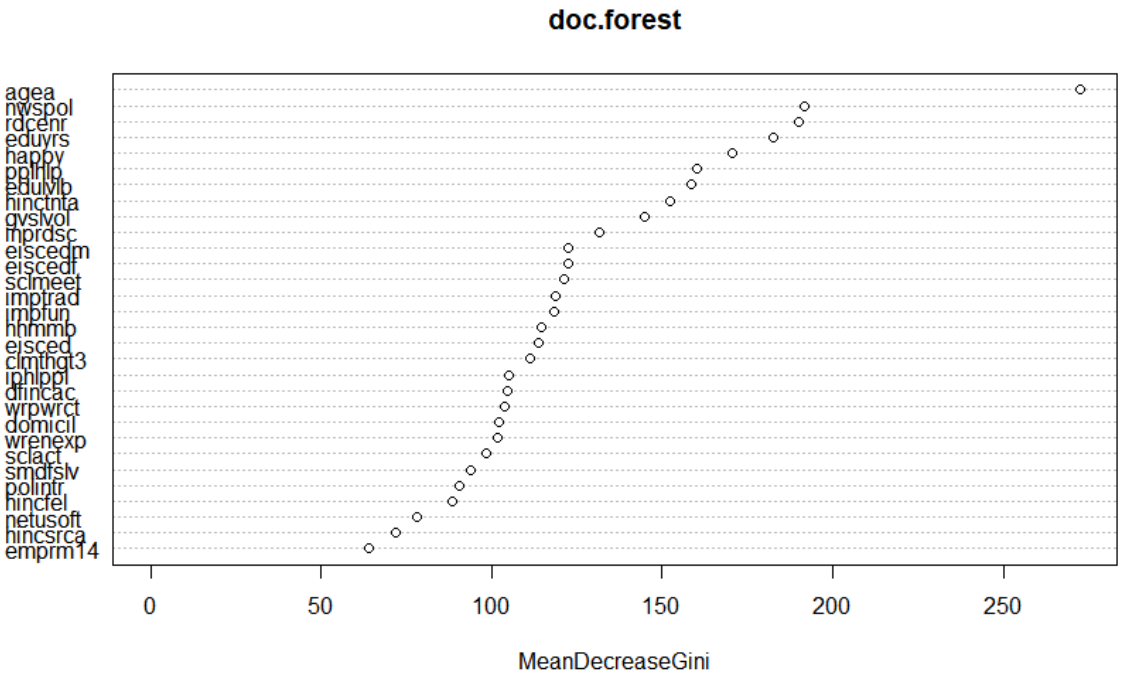

Figure 4. Variable importance plot for Table 13

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.696	0.904	0.899	0.629
Specificity	0.404	0.278	0.314	0.505
Accuracy	0.649	0.804	0.801	0.609
AUC	0.550	0.591	0.608	0.567

Table 14. Prediction results; dependent variable: eneffap, independent variables: top three predictors shown in Figure 4

5. Conclusion and Policy Implications

Human activities have had considerable and global impacts on the environment. There is scientific consensus about the reality of anthropogenic climate change (Oreskes, 2004). Therefore, anthropogenic climate change is a serious issue that needs to be addressed. Depending on the social preferences, environmental problems within any nation are perceived, interpreted, and prioritized

differently and environmental policy needs to take these differences into account (Walter and Ugelow, 1979). Thus, it is crucial for policymakers to be able to estimate these preferences to make informed policy decisions. While numerous papers have investigated the underlying factors that affect environmental preferences, there are not any papers, to the best of our knowledge, that have focused on predicting the values of such preferences. In this paper, we aim to fill this gap. We use the European Social Survey (ESS) data, an academically driven cross-national survey conducted across thirty European nations.

Based on the prediction results, we are able to correctly identify 74% of climate change skeptics while still correctly classifying 58% of climate change believers. In the case of respondents' opinion about the role of anthropogenic climate change, we can correctly identify 77% of supporters and 53% of deniers and achieve a 73% rate of accuracy. When it comes to predicting respondents' position on increasing taxes on fossil fuels, we are able to correctly identify 72% of those opposed to the tax policy and more than half of supportive or neutral respondents. Our model can also correctly identify 75% of individuals who are likely to buy a very energy efficient home appliance while correctly identifying 56% of the other group. In addition, in each model, when we limit the predictors to the top three predictors, we are able to correctly identify 80% of climate change skeptics, 76% of deniers of anthropogenic climate change, and 71% of environmental tax opposers. These features can also correctly classify approximately 63% of individuals who are more likely to buy energy efficient appliances for their home.

The design and implementation of environmental policy are expensive. Failure to incorporate the public's standpoint may result in their resistance, protests, violence, and potentially revoking the policy, e.g., violent protests by the Yellow Vest movement in France (Rubin and Sengupta, 2018).

To avoid these costs, the officials can use the information and methodology provided in this paper to identify the citizens that are more likely to oppose the policy. Depending on the portion of the population that falls into this group, the government may decide to overturn the policy or initiate educational programs to enlighten the potential opposers about the detailed consequences and benefits of the policy. Additionally, green firms can use these prediction results for market segmentation to determine opportunities, deliver tailored marketing strategies, and product characteristics to increase profits and better serve customers' needs and wants. Ignoring individuals' environmental awareness and concerns can be costly for firms, e.g. CFC-free refrigerators produced by Whirlpool that resulted in loss for the company (Singh and Pandey, 2012).

Despite the richness of the data in this survey, it is important to note that all the values for all features and responses are self-reported values and individuals' self-reported intensions may not necessarily carry to actions. In addition, the survey is for European nations which are generally more environmentally aware than US citizens (e.g., see Ziegler, 2017). For instance, more than 90% of the respondents in this survey believe that the role of anthropogenic climate change is greater than or equal to the role of natural processes in causing climate change (Table 6). This imbalance in the data, while it does not make our results incorrect, hinders the predictive ability of the classifiers. Future research can combine this survey with similar surveys from other continents. By making the sample observations more diverse, we can expect higher values of sensitivity, specificity, accuracy, and AUC and improve the predictions for individuals' environmental values and attitudes.

6. Appendix

In this appendix, we do some sensitivity analysis by redoing the predictions of section 4 but with different thresholds for the dependent variables.

6.1. Beliefs about climate change

Here, we group the respondents who chose probably not changing and definitely not changing together and code them as 0 and group the respondents who chose definitely changing and probably changing together and code them as 1. Note that this method of categorizing would imply that less than 7% of the data are in group 0 (see Table 3). This remarkable imbalance reduces the predicting abilities of the models. The results are shown in Table 15. The specificity values are small, meaning the models tend to predict almost everyone as the more populated group which is 1 here.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.886	0.938	0.981	0.795
Specificity	0.193	0.291	0.135	0.328
Accuracy	0.843	0.898	0.929	0.765
AUC	0.539	0.615	0.558	0.561

Table 15

6.2. Beliefs about the share of anthropogenic climate change

In our sensitivity analysis for predicting ccnthum, we group those who chose entirely and mainly by human activity together and code them as 1. The remaining respondents are grouped together and coded as 0. The results are demonstrated in Table 16. The models have good predictive powers.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.754	0.788	0.868	0.507
Specificity	0.515	0.558	0.382	0.732
Accuracy	0.714	0.750	0.786	0.545
AUC	0.635	0.673	0.625	0.620

Table 16

6.3. Attitudes towards an environmental tax policy

In this part, we categorize the respondents who are neither in favor nor against an environmental tax policy as brown consumers. Hence, in Table 9, the first two options are grouped together and coded as 1 and the remaining three options are grouped together and coded as 0. Table 17 shows the results. For instance, neural networks can correctly classify 72% of the brown consumers while correctly classifying more than half of green consumers.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.586	0.564	0.266	0.505
Specificity	0.642	0.684	0.924	0.721
Accuracy	0.623	0.643	0.698	0.647
AUC	0.614	0.624	0.595	0.613

Table 17

6.4. Buying energy efficient home appliances

One could argue that only those who report being extremely likely to buy an energy efficient home appliance would buy one. Therefore, in this part of our sensitivity analysis, we group those who reported values of 9 and 10 in Table 10 as green consumers and code them as 1. The remaining respondents are considered brown and coded as 0. The results are shown in Table 18. For example, boosting can predict 78% of brown consumers correctly while correctly classifying half of green consumers.

	Logistic	Random Forest	Boosting	Neural Networks
Sensitivity	0.523	0.578	0.519	0.586
Specificity	0.696	0.733	0.782	0.691
Accuracy	0.615	0.661	0.659	0.642
AUC	0.609	0.656	0.650	0.639

Table 18

7. References

Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. Proceedings of the National Academy of Sciences, 113(42), 11770-11775.

Bansal, S., & Gangopadhyay, S. (2003). Tax/subsidy policies in the presence of environmentally aware consumers. Journal of Environmental Economics and Management, 45(2), 333-355.

Carlsson, F., Kataria, M., Krupnick, A., Lampi, E., Löfgren, Å., Qin, P., Chung, S., & Sterner, T. (2012). Paying for mitigation: A multiple country study. Land Economics, 88(2), 326-340.

ESS Round 8: European Social Survey Round 8 Data (2016). Data file edition 2.1. NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. doi:10.21338/NSD-ESS8-2016

Espinola-Arredondo, A., & Zhao, H. (2012). Environmental policy in a linear city model of product differentiation. Environment and Development Economics, 17(4), 461-477.

Forrest, M. (2020, February 3). Trudeau faces defining challenge of his second term. Politico. https://www.politico.com/news/2020/02/03/justin-trudeau-climate-dilemma-canada-110388

Gadenne, D., Sharma, B., Kerr, D., & Smith, T. (2011). The influence of consumers' environmental beliefs and attitudes on energy saving behaviours. Energy policy, 39(12), 7684-7694.

Hlaváč, V. (2016). Classifier performance evaluation. Czech Technical University.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, pp. 3-7). New York: springer.

Joireman, J., Truelove, H. B., & Duell, B. (2010). Effect of outdoor temperature, heat primes and anchoring on belief in global warming. Journal of Environmental Psychology, 30(4), 358-367.

Heaton, J. (2008). Introduction to neural networks with Java. Heaton Research, Inc.

Hoegh-Guldberg, O., & Bruno, J. F. (2010). The impact of climate change on the world's marine ecosystems. Science, 328(5985), 1523-1528.

IPCC, 2007: Summary for Policymakers. In: Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press, Cambridge, UK, 7-22.

Kallbekken, S., Kroll, S., & Cherry, T. L. (2011). Do you not like Pigou, or do you not understand him? Tax aversion and revenue recycling in the lab. Journal of Environmental Economics and Management, 62(1), 53-64.

Kotchen, M. J., Boyle, K. J., & Leiserowitz, A. A. (2013). Willingness-to-pay and policy-instrument choice for climate-change policy in the United States. Energy Policy, 55, 617-625.

Krosnick, J. A., Holbrook, A. L., Lowe, L., & Visser, P. S. (2006). The origins and consequences of democratic citizens' policy agendas: A study of popular concern about global warming. Climatic change, 77(1-2), 7-43.

Maloof, M. A. (2003, August). Learning when data sets are imbalanced and when costs are unequal and unknown. In ICML-2003 workshop on learning from imbalanced data sets II (Vol. 2, pp. 2-1).

McCright, A. M., & Dunlap, R. E. (2011). Cool dudes: The denial of climate change among conservative white males in the United States. Global environmental change, 21(4), 1163-1172.

Mostafa, M. M. (2009). Shades of green: A psychographic segmentation of the green consumer in Kuwait using self-organizing maps. Expert Systems with Applications, 36(8), 11030-11038.

Nam, T. (2007). Rough days in democracies: Comparing protests in democracies. European Journal of Political Research, 46(1), 97-120.

Nemet, G. F., & Johnson, E. (2010). Willingness to pay for climate policy: a review of estimates.

O'Connor, R. E., Bord, R. J., Yarnal, B., & Wiefek, N. (2002). Who wants to reduce greenhouse gas emissions?. Social Science Quarterly, 83(1), 1-17.

O'Connor, R. E., Bard, R. J., & Fisher, A. (1999). Risk perceptions, general environmental beliefs, and willingness to address climate change. Risk analysis, 19(3), 461-471.

Oreskes, N. (2004). The scientific consensus on climate change. Science, 306(5702), 1686-1686.

Pietsch, J., & McAllister, I. (2010). ‘A diabolical challenge’: public opinion and climate change policy in Australia. Environmental Politics, 19(2), 217-236.

Riou, S., Gray, C. M., Brooke, M. D. L., Quillfeldt, P., Masello, J. F., Perrins, C., & Hamer, K. C. (2011). Recent impacts of anthropogenic climate change on a higher marine predator in western Britain. Marine Ecology Progress Series, 422, 105-112.

Rosenzweig, C., Karoly, D., Vicarelli, M., Neofotis, P., Wu, Q., Casassa, G., ... & Tryjanowski, P. (2008). Attributing physical and biological impacts to anthropogenic climate change. Nature, 453(7193), 353-357.

Rubin, A. J., & Sengupta, S. (2018, December 6).'Yellow Vest' Protests Shake France. Here's the Lesson for Climate Change. New York Times. https://nyti.ms/2zICkcc

Singh, P. B., & Pandey, K. K. (2012). Green marketing: policies and practices for sustainable development. Integral Review, 5(1), 22-30.

Söderholm, P. (2012). Modeling the economic costs of climate policy: An overview. American Journal of Climate Change, 1(1), 14-32.

Sutherland, R. J. (2000). " No cost" efforts to reduce carbon emissions in the US: An economic perspective. ENERGY JOURNAL-CAMBRIDGE MA THEN CLEVELAND OH-, 21(3), 89-112.

Walter, I., & Ugelow, J. L. (1979). Environmental policies in developing countries. Ambio, 102-109.

Yahoo! News. (2019, March 18). French 'yellow vest' demos caused 170 mln euros damage: insurers.https://news.yahoo.com/french-yellow-vest-demos-caused-170-mln-euros-151403596.html

Ziegler, A. (2017). Political orientation, environmental values, and climate change beliefs and attitudes: An empirical cross country analysis. Energy Economics, 63, 144-153.