Predicting health information avoidance using machine learning models

Zihan Hei

Table of contents

1 Introduction

Health information avoidance—defined as any behavior designed to prevent or delay access to available but potentially unwanted health information (Howell et al. (2020); Sweeny et al. (2010))—remains a significant barrier to realizing the public health benefits of personalized risk communication. Despite the rapid expansion of access to medical information in the Information Age, many people still choose not to learn about their personal health risks (Gigerenzer & Garcia-Retamero (2017); Ho et al. (2021); Kelly & Sharot (2021)). Health information avoidance is not a uniform behavior, but can take multiple forms, varying in its duration and degree of intentionality. Individuals may avoid health information by delaying the decision to learn about their screening results or they may avoid it completely, choosing to never know. Information avoidance may also manifest through both active and passive means. People may explicitly ask others not to disclose information, physically remove themselves from situations were information might be revealed, or passively refrain from questions that could reveal unwanted knowledge. These avoidance behaviors may manifest as refusing cancer screening, delaying medical care, or not requesting test results (Sweeny et al. (2010)).

Research indicates that health information avoidance is common. For example, approximately 15% of US adults avoid personalized health risk information across various contexts (Meese et al., 2022), and nearly 39% express a reluctance to learn about their cancer risk (Emanuel et al. (2015)). Similar avoidance rates (e.g., approximately 40%) have been observed for non-personalized health information (i.e., general health information not relevant to personal risk) (Chae et al. (2019); Orom et al. (2020)).

Such widespread avoidance underscores the need to understand how and why individuals avoid health information. Some individuals may avoid health information to protect themselves from unpleasant emotions, prevent exposure to information that conflicts with their worldview or creates an obligation to act. Even when the information may be critical to health,avoidance eliminates the discomfort of decision-making and the emotional burden of confronting potential illness (Sweeny et al. (2010)). Previous research has identified various psychological and cognitive factors underlying this avoidance phenomenon and explored potential explanations for this behavior. O'Brien et al. (2024) found that self-perceptions of health, such as low perceived risk, engagement in healthy behaviors, and demographic characteristics, often guide people's decisions to avoid learning about their health risks. Other studies have shown that information overload can increase anxiety and cognitive dissonance, leading to avoidance behavior (Dattilo et al. (2022); Song et al. (2021); Soroya & Faiola (2023)). Furthermore, heightened risk perceptions can exacerbate anxiety and sadness, which may further hinder people from seeking health information (Sultana et al. (2023); Zhao & Cai (2009)).

However, few studies have attempted to use machine learning methods to identify predictors of health information avoidance. This study applies predictive modeling and machine learning methods to examine patterns of cancer screening avoidance ("health avoiders") using sociodemographic and psychological data. By integrating behavioral and belief factors, this study aimed to better understand the complex dynamics that lead to health information avoidance. The key finding is that while many predictors showed statistical significance, none provided meaningful predictive power. This suggests that the key determinants leading people to avoid health information remain unidentified.

1.1 Review of Machine Learning Methods

Machine learning (ML) methods were applied to identify patterns associated with health information avoidance. This study focuses on three main models: **Logistic Regression**, **Random Forest**, and **Multivariate Adaptive Regression Splines (MARS)**. Each approach offers different advantages and trade-offs, summarized in Table 1 below.

1.1.1 Table 1. Comparison of Machine Learning Models

Model	**Description**	**Advantages**	**Limitations**	**Interpretability**
Linear Regression	Models the relationship between predictors and a continuous outcome using a straight-line equation.	Simple, interpretable, and efficient to train; effective for linearly related data; allows understanding of variable relationships.	Assumes linearity and independence among variables; sensitive to outliers; limited for complex or nonlinear data.	High
Logistic Regression	Estimates the probability of a binary outcome based on a linear combination of predictors.	Simple, interpretable, good baseline; easy to assess predictor importance.	Assumes linearity; struggles with nonlinear or high-dimensional data.	High
Random Forest	Ensemble of decision trees using bootstrapped samples and random feature selection.	Handles nonlinear and complex data; reduces overfitting; provides feature importance.	Less interpretable; slower with large datasets; harder to explain model logic.	Moderate

Model	Description	Advantages	Limitations	Interpretability
MARS (Multivariate Adaptive Regression Splines)	Builds flexible regression models using piecewise linear splines.	Captures nonlinear relationships; performs automatic feature selection; minimal preprocessing needed.	Computationally intensive; interpretation can be challenging with correlated predictors.	Moderate - High

Linear / Logistic Regression served as a baseline interpretable model, Random Forest captured complex nonlinear interactions, and MARS modeled adaptive spline-based relationships between predictors and cancer avoidance. By comparing these models, this study aimed to identify the most effective approach to predict health information avoidance while maintaining interpretability.

2 Method

2.1 Data Source

This study used the de-identified Health Avoiders dataset provided through Cloud Research in collaboration with Dr. Heather Orom (University at Buffalo). The dataset includes sociodemographic, psychological, and behavioral variables collected by Cloud Research. All analyses were conducted in R, and reproducibility was ensured through a README file and a Quarto documentation workflow.

2.2 Outcome Variable

The outcome variable, `Cancer_Avoidance_Mean`, represents the average score across 8 items (`Avoid_Cancer_`)measuring participants' avoidance of cancer-related health information. Because this variable exhibited non-normality, both logarithmic and square-root transformations were tested; however, these transformations did not improve model performance or interpretability, so the untransformed variable was retained for analysis.

2.3 Predictor Models

Heather Orom (University at Buffalo) developed a theoretical framework to group predictors as a health psychologist studying health information avoidance.

1. Demographic Model (`demo_data`) — `Ethnicity`, `Political_Party`, `Gender4`, `Job_Classification`, `Education_Level`, `Age`, `Income`, `Race`, and `MacArthur_Numeric`.

2. Media Use Model (media_data) — Social_Media_Usage, AI_Use, Video_Games_Hours, Listening_Podcasts, Facebook_Usage_cat, TikTok_Use, X_Twitter_Usage, Social_Media_type, and Influencer_Following.

3. Health Condition Model (health_condition_data) — Stressful_Events_Recent, Current_Depression, Anxiety_Severity_num, PTSD5_Score, Health_Depression_Severity_num, and Stress_TotalScore.

4. Health Behavior Model (health_behavior_data) — Fast_Food_Consumption, Meditation_group, Physical_Activity_Guidelines, Cigarette_Smoking_num, Supplement_Consumption_Reason_num, Diet_Type, and Supplement_Consumption.

5. Other Factors Model (other_data) — Home_Ownership, Voter_Registration, Climate_Change_Belief, and Mental_Health_of_Partner.

2.4 Variable Scaling and Scoring

Several psychological and health condition variables were standardized using validated scales:

- PC-PTSD-5: a 5-item yes/no screen for post-traumatic stress.
- GAD-7: a 7-item measure of anxiety severity (0–3 scale).
- PHQ-9: a 9-item measure of depression severity (0–3 scale).
- Life Events Checklist: summed to represent cumulative stress exposure.

Composite averages were computed for each domain, resulting in 4 continuous variables: PTSD5_Score, Anxiety_Severity_num, Health_Depression_Severity_num, and Stress_TotalScore.

2.5 Analysis Plan

Predictive modeling was conducted in R using the tidymodels package. Due to the non-normal outcome variable, two complimentary approaches were applied:

1. Regression Models: predicted the continuous outcome Cancer_Avoidance_Mean were fitted using Linear Regression (baseline), Random Forest, and Multivariate Adaptive Regression Splines (MARS). Model performance was assessed using root mean squared error (RMSE), mean absolute error (MAE), and the correlation between predicted and observed values.

2. Classification Models: predicted the binary outcome Cancer_Avoiders01 (0 = non-avoider, 1 = avoider) were fitted using Logistic Regression (baseline) and Random Forest. Model performance was assessed using area under the curve (AUC) and accuracy with basic calibration applied when class imbalance occurred.

Some models used cross-validation, and Linear / Logistic Regression served as the baseline for comparison.

3 Results

All analyses were conducted in Posit Cloud, a collaborative online platform for writing R scripts and developing Quarto markdown documents.

3.1 Overview of Model Performance

Table 2 summarizes the predictive performance across all modeling approaches. Consistent with the study's central finding, all models demonstrated **statistically significant associations but no meaningful predictive power**.

3.1.1 Table 2. Summary of Model Performance Across All Approaches

Domain	Model Type	Key Metric	Value	Interpretation
Demographics	Regression (RF)	Correlation	0.109	No predictive relationship
Demographics	Regression (MARS)	Cross-val R^2	0.017	Poor generalization
Demographics	Classification (RF)	ROC AUC	0.449	Below chance performance
Media Usage	Regression (RF)	Correlation	0.021	No predictive relationship
Media Usage	Classification (RF)	ROC AUC	0.574	Marginally above chance
Health Condition	Regression (RF)	Correlation	0.103	No predictive relationship
Health Condition	Regression (MARS)	Cross-val R^2	-0.003	Negative (no pattern)
Health Condition	Classification (RF)	ROC AUC	0.446	Below chance performance
Health Behavior	Regression (RF)	Correlation	0.060	No predictive relationship
Health Behavior	Classification (RF)	ROC AUC	0.472	Below chance performance
Other Factors	Regression (RF)	Correlation	0.145	Weak relationship
Other Factors	Classification (RF)	ROC AUC	0.420	Below chance performance
Full Model	**Regression (RF)**	**Correlation**	**0.300**	**Weak relationship**
Full Model	**Classification (RF)**	**ROC AUC**	**0.230**	**Worse than individual models**

Note: RF = Random Forest; MARS = Multivariate Adaptive Regression Splines. Classification models showed high accuracy (94-96%) due to class imbalance but uniformly poor discrimination (AUC < 0.60). All R^2 values were < 0.03.

The following sections detail findings for each predictor domain, focusing on the strongest associations identified despite their limited predictive utility.

3.2 Overview of Findings

This analysis examined five predictor domains: demographic, health status, health behaviors, media usage, and other attitudinal variables. In all modeling approaches, predictors were statistically significant with cancer avoidance behavior, but their predictive power was limited. The following paragraphs detail the findings for each domain.

3.2.1 Demographic Predictors

Linear regression analysis showed a significant negative association between socioeconomic status (MacArthur numerical score) and cancer avoidance behavior (= -0.025, p = 1.89e-10), but the model explained only 0.5% of the variance ($R^2 = 0.005$). Age analysis identified a threshold of 35 years, with older individuals exhibiting significantly higher cancer avoidance scores (= 0.063, p = 6.45e-06).

In both age groups (under 35 and over 35), Democrats had lower cancer avoidance scores than Republicans: under 35 (= -0.165, p = 8.97e-08) and over 35 (= -0.225, p = 2.67e-15). MARS analysis confirmed that party affiliation was the most important factor (importance = 100). Although the classification model showed high accuracy (95.5%), its discrimination was poor (ROC AUC = 0.449), indicating statistically significant but not practically predictive effects.

3.2.2 Health Condition Predictors

Predictive relationships among health condition variables were weak. The total stress score showed no significant trend (= 0.005, p = 0.073, $R^2 = 0.0005$). While the MARS model identified anxiety severity as the most important factor, cross-validation performance was negative (CVRSq = -0.003), suggesting no generalizable pattern. The classification model also failed to meaningfully differentiate cancer avoiders.

3.2.3 Health Behavior Predictors

Cigarette smoking demonstrated a significant positive relationship with cancer avoidance. Smokers had higher cancer avoidance scores (= 0.186, p < 2e-16, $R^2 = 0.011$) and 1.87 times higher odds of being cancer avoiders (= 0.627, p = 1.05e-06). Approximately 8–9% of smokers were cancer avoiders compared to 4–5% of non-smokers. Despite statistical significance, the overall predictive power remained limited.

3.2.4 Media Usage Predictors

Media usage variables showed minimal predictive value (overall correlation = 0.021). Influencer following was not significantly associated with cancer avoidance (= -0.011, p = 0.613, R^2 = 0.00008). Facebook usage showed a statistically significant but weak effect (= 0.236, p = 0.039), with nearly equal proportions of cancer avoiders among users (5%) and non-users (4%). These effects are statistically significant but not practically meaningful.

3.2.5 Other Predictors: Climate Change Beliefs

Climate change beliefs were statistically associated with cancer avoidance (R^2 = 0.024). Individuals who strongly believe in human-caused climate change had lower cancer avoidance scores (= -0.271, p = 3.32e-13) and 0.29 times the odds of being cancer avoiders compared to climate deniers (= -1.236, p = 1.87e-10). Climate deniers showed 12–13% cancer avoiders compared to 3–4% among strong believers. Despite statistical significance, model differentiation remained poor (ROC AUC = 0.42).

3.2.6 Comprehensive Model: Combining Major Predictors

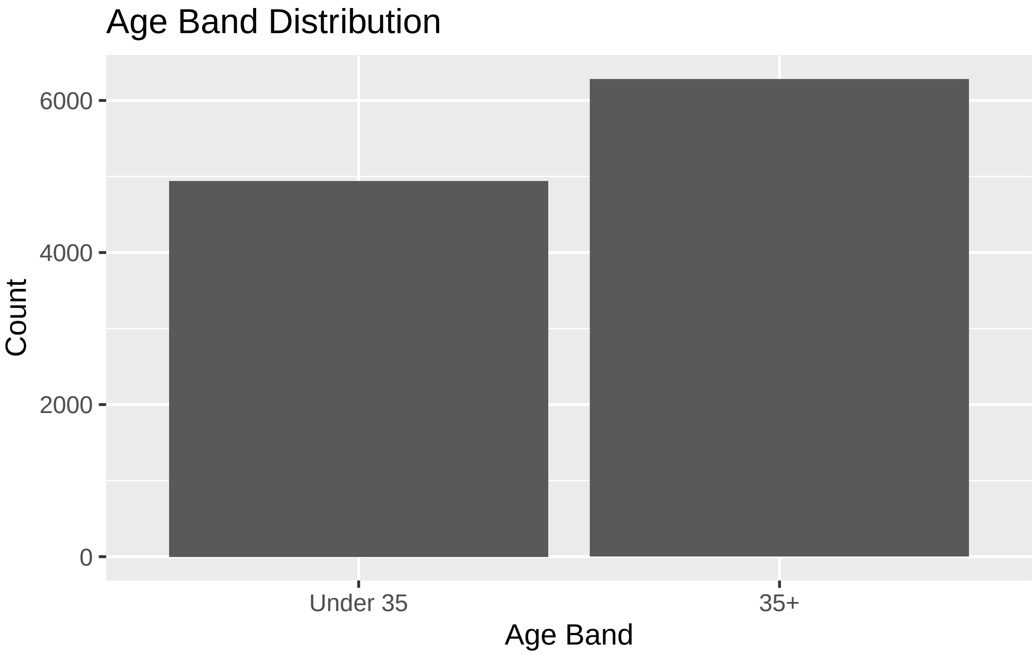
Random forest models incorporating all main predictors showed modest improvement. The regression model achieved a correlation of 0.30 (RMSE = 0.62, MAE = 0.52). The classification model achieved 95% accuracy but failed to meaningfully differentiate cancer avoiders (ROC AUC = 0.23).

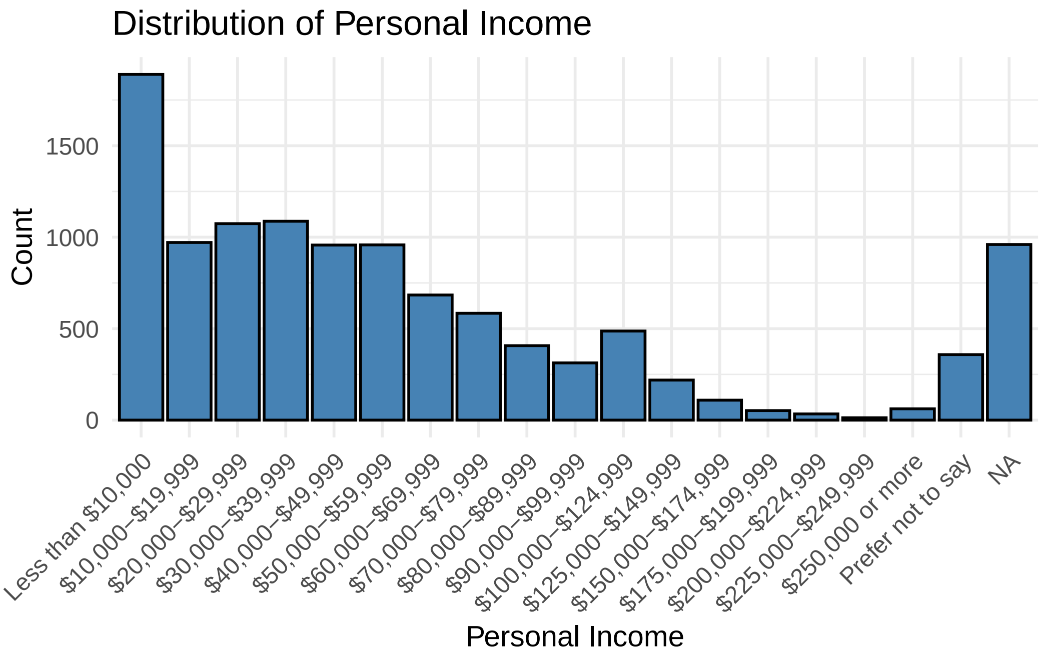
Across all analyses, predictors were statistically significant but not practically predictive ($R^2 < 0.03$, ROC AUC < 0.50), indicating that the measured variables have limited explanatory power for cancer avoidance behavior.

3.3 Data Preprocessing and Descriptive Statistics

3.4 Predictor Distributions

3.4.1 Predictors



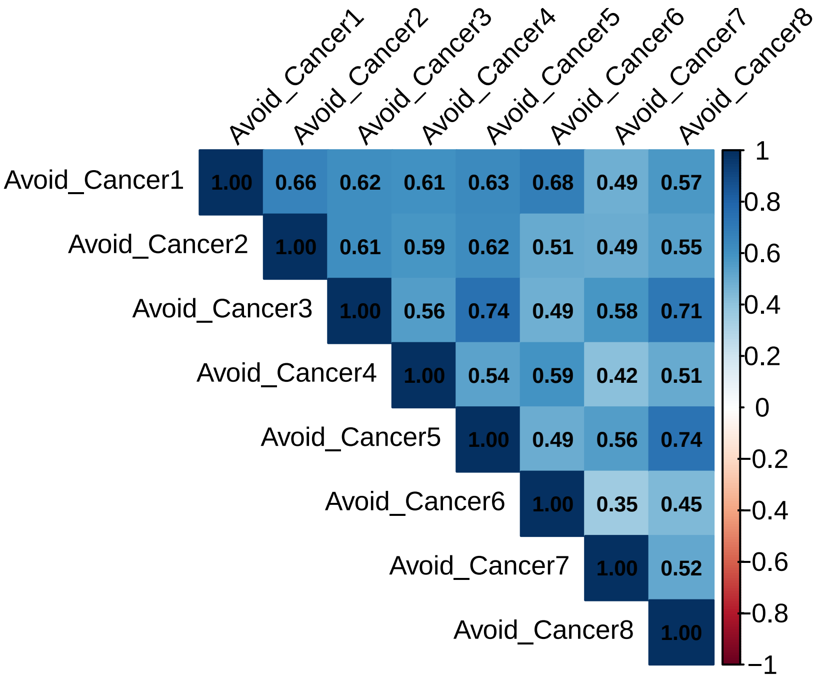


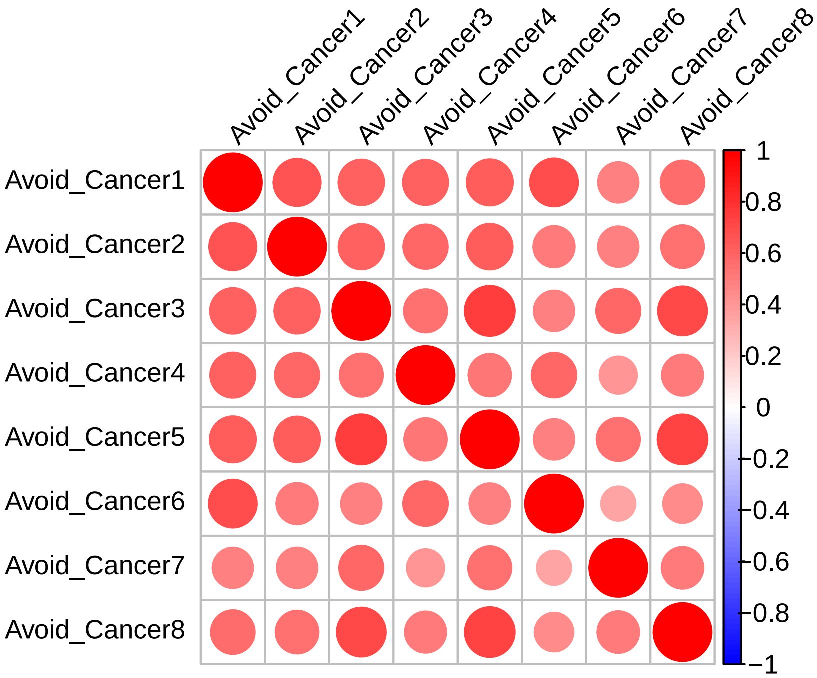
3.5 Outcome Distribution and Item Correlations

- The outcome variable, `Cancer_Avoidance_Mean`, was computed as the average of 8 items measuring participants' avoidance of cancer-related health information. Four of these items `Avoid_Cancer3`, `Avoid_Cancer5`, `Avoid_Cancer7`, `Avoid_Cancer8` were reverse-coded to ensure higher scores consistently reflected greater avoidance. Reverse scoring was calculated using the formula:
  - Reversed score=(Max+Min)−Original score

3.5.0.1 Correlation

```
              Avoid_Cancer1 Avoid_Cancer2 Avoid_Cancer3 Avoid_Cancer4
Avoid_Cancer1     1.0000000     0.6630162     0.6192841     0.6072410
Avoid_Cancer2     0.6630162     1.0000000     0.6141968     0.5869138
Avoid_Cancer3     0.6192841     0.6141968     1.0000000     0.5556186
Avoid_Cancer4     0.6072410     0.5869138     0.5556186     1.0000000
Avoid_Cancer5     0.6338078     0.6208662     0.7444480     0.5385535
Avoid_Cancer6     0.6840374     0.5089422     0.4899384     0.5936301
Avoid_Cancer7     0.4878094     0.4908217     0.5812488     0.4152808
Avoid_Cancer8     0.5700417     0.5459774     0.7130865     0.5062153
              Avoid_Cancer5 Avoid_Cancer6 Avoid_Cancer7 Avoid_Cancer8
Avoid_Cancer1     0.6338078     0.6840374     0.4878094     0.5700417
Avoid_Cancer2     0.6208662     0.5089422     0.4908217     0.5459774
Avoid_Cancer3     0.7444480     0.4899384     0.5812488     0.7130865
Avoid_Cancer4     0.5385535     0.5936301     0.4152808     0.5062153
Avoid_Cancer5     1.0000000     0.4917697     0.5572648     0.7388645
Avoid_Cancer6     0.4917697     1.0000000     0.3526875     0.4478447
Avoid_Cancer7     0.5572648     0.3526875     1.0000000     0.5192653
Avoid_Cancer8     0.7388645     0.4478447     0.5192653     1.0000000
```





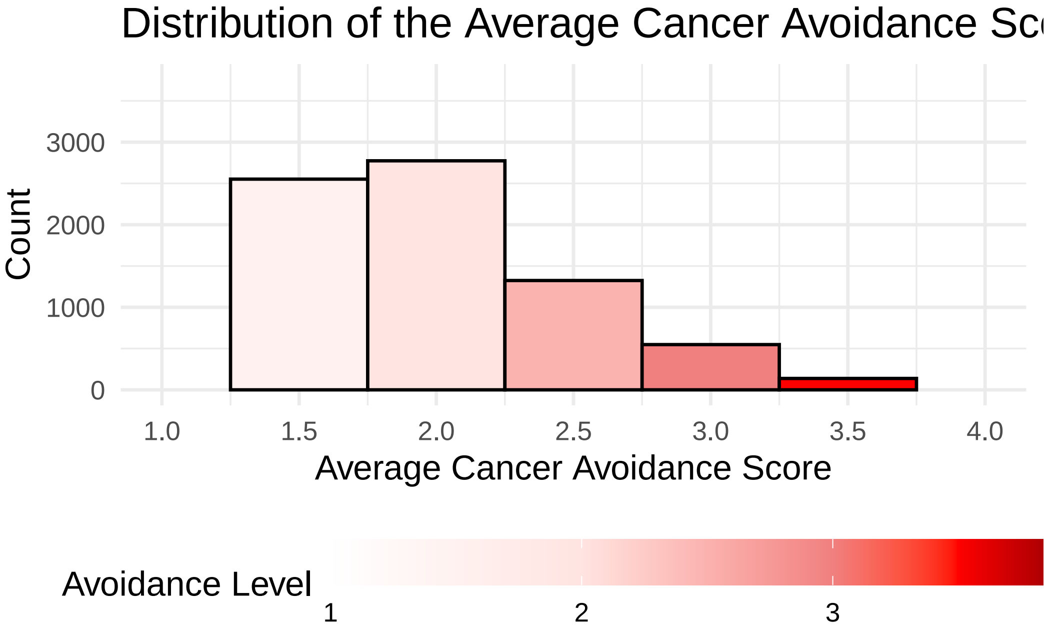
- All the 8 of the cancer items are moderate to high correlate with each other, so we can add them up and get the average score.

3.5.0.2 Alpha

```
Reliability analysis
Call: psych::alpha(x = selectdata[, c("Avoid_Cancer1", "Avoid_Cancer2",
    "Avoid_Cancer3", "Avoid_Cancer4", "Avoid_Cancer5", "Avoid_Cancer6", 
    "Avoid_Cancer7", "Avoid_Cancer8")])

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean   sd median_r
      0.91      0.91    0.91      0.57  10 0.0013  1.7 0.65     0.56

    95% confidence boundaries 
         lower alpha upper
Feldt     0.91  0.91  0.91
Duhachek  0.91  0.91  0.91

 Reliability if an item is dropped:
              raw_alpha std.alpha G6(smc) average_r  S/N alpha se  var.r med.r
Avoid_Cancer1      0.89      0.90    0.89      0.55  8.7   0.0016 0.0098  0.55
Avoid_Cancer2      0.90      0.90    0.90      0.56  9.1   0.0015 0.0107  0.56
Avoid_Cancer3      0.89      0.90    0.89      0.55  8.6   0.0015 0.0084  0.55
Avoid_Cancer4      0.90      0.90    0.90      0.58  9.5   0.0014 0.0103  0.57
Avoid_Cancer5      0.89      0.90    0.89      0.55  8.6   0.0015 0.0079  0.56
Avoid_Cancer6      0.91      0.91    0.90      0.59  9.9   0.0013 0.0070  0.58
Avoid_Cancer7      0.91      0.91    0.91      0.59 10.2   0.0014 0.0070  0.59
Avoid_Cancer8      0.90      0.90    0.90      0.56  9.0   0.0015 0.0083  0.58

 Item statistics 
                  n raw.r std.r r.cor r.drop mean   sd
Avoid_Cancer1 11155  0.85  0.84  0.82   0.78  1.8 0.90
Avoid_Cancer2 11155  0.79  0.80  0.76   0.73  1.6 0.77
Avoid_Cancer3 11155  0.83  0.84  0.83   0.78  1.6 0.75
Avoid_Cancer4 11155  0.78  0.76  0.71   0.69  1.9 0.93
Avoid_Cancer5 11155  0.83  0.84  0.83   0.78  1.6 0.76
Avoid_Cancer6 11155  0.75  0.72  0.68   0.65  2.3 1.02
Avoid_Cancer7 11155  0.67  0.70  0.63   0.59  1.4 0.59
Avoid_Cancer8 11155  0.79  0.80  0.77   0.72  1.7 0.83

Non missing response frequency for each item
              0    1    2    3    4 5 miss
Avoid_Cancer1 0 0.45 0.32 0.18 0.05 0    0
Avoid_Cancer2 0 0.56 0.32 0.10 0.03 0    0
Avoid_Cancer3 0 0.54 0.35 0.09 0.02 0    0
Avoid_Cancer4 0 0.43 0.27 0.26 0.05 0    0
Avoid_Cancer5 0 0.50 0.38 0.10 0.03 0    0
Avoid_Cancer6 0 0.29 0.23 0.36 0.12 0    0
Avoid_Cancer7 0 0.69 0.27 0.03 0.01 0    0
Avoid_Cancer8 0 0.53 0.31 0.12 0.04 0    0
```

- Based on the analysis, the alpha does not improve for the measure if any of the items are dropped. All final alphas will be equal or lower than the 0.89 raw alpha and 0.9 standardized alpha.
- The internal consistency of the `Information Avoidance - Cancer` items was high (std.alpha

= 0.895), suggesting that the items measured a coherent construct. The average inter-item correlation was (average_r = 0.55), indicating moderately strong associations among the items without redundancy. Reliability estimates (G6 = 0.894) further supported this consistency.



Percentage of participants with Cancer_Avoidance_Mean >= 2.5

```
[1] 14.4061
```

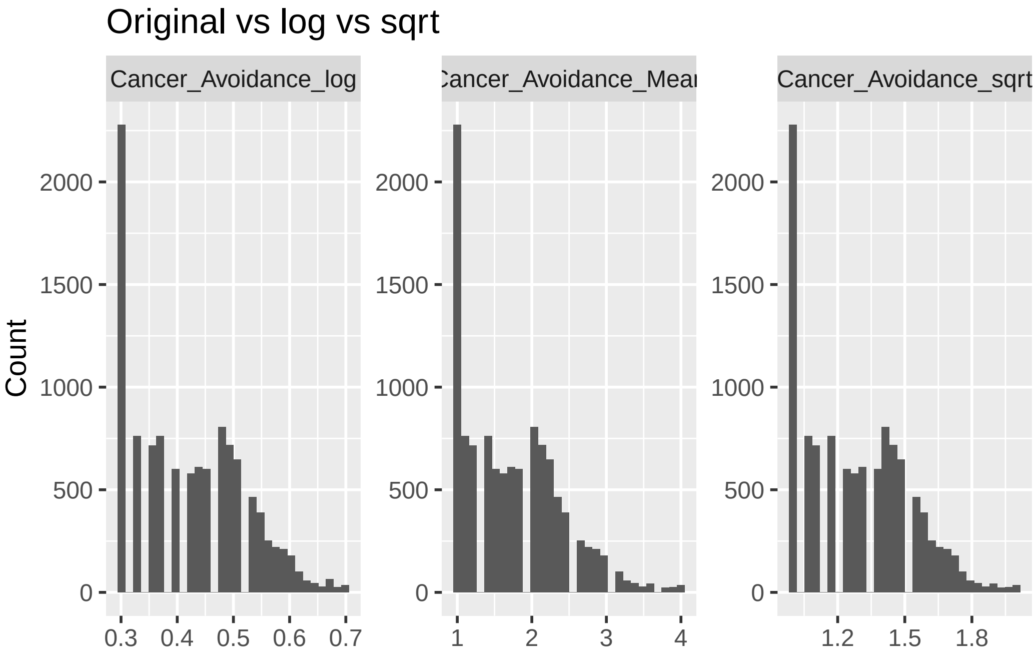
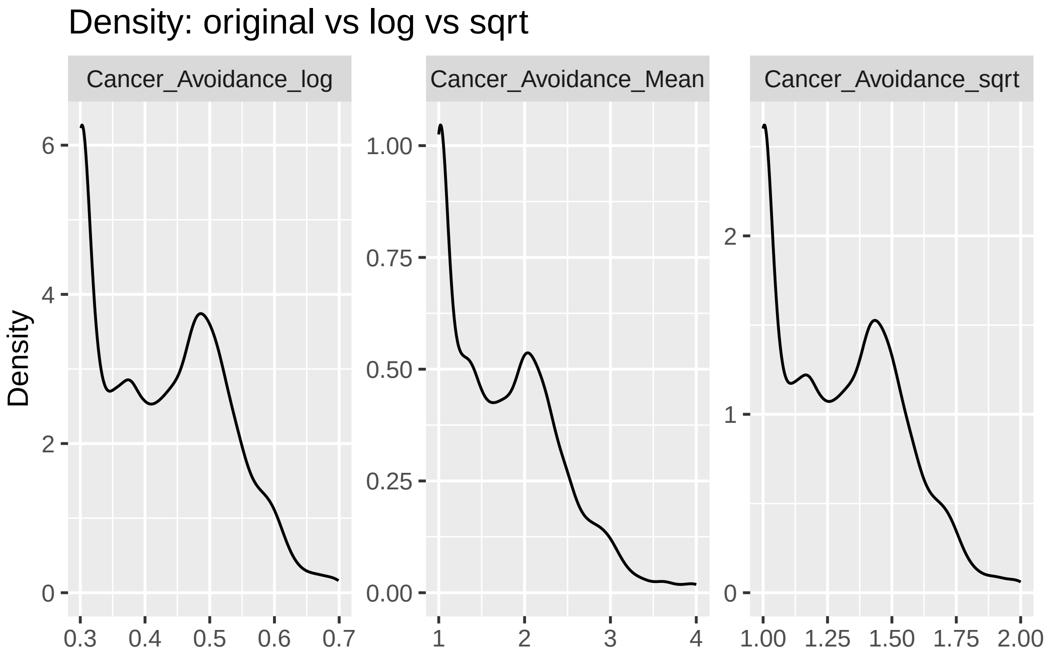
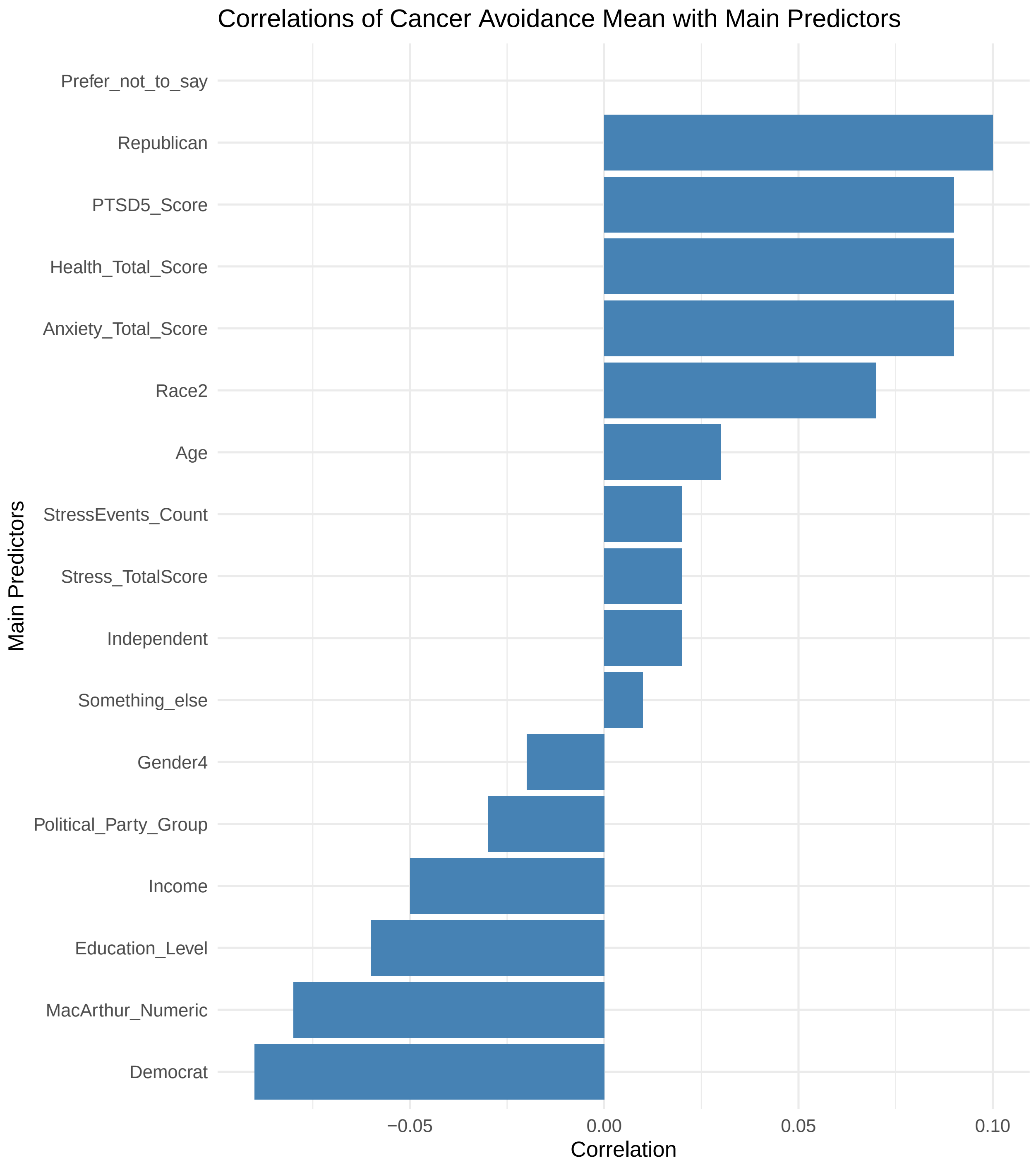


Table 3: Correlations between Cancer Avoidance Mean and main predictors

Predictor	Correlation
Age	0.03
Republican	0.10
Democrat	-0.09
Independent	0.02
Something_else	0.01
Prefer_not_to_say	NA
Gender4	-0.02
Education_Level	-0.06
Income	-0.05
MacArthur_Numeric	-0.08
Political_Party_Group	-0.03
Race2	0.07
Anxiety_Total_Score	0.09
PTSD5_Score	0.09
Health_Total_Score	0.09
StressEvents_Count	0.02
Stress_TotalScore	0.02



3.5.0.3 Matrix



3.6 Descriptive tables for categorical variables

3.7 REGRESSION MODELS

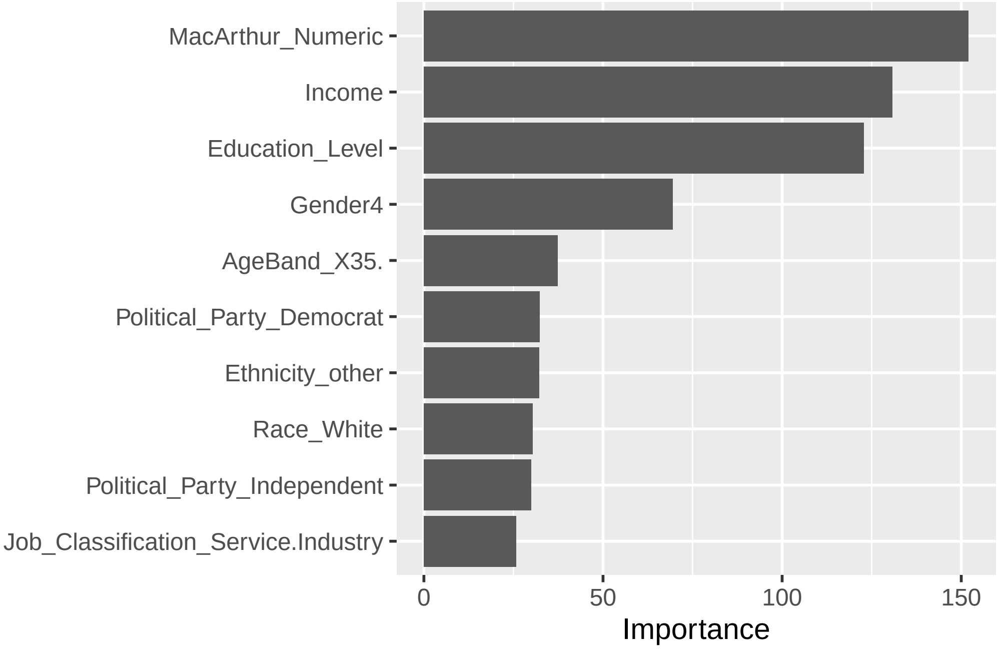
(continuous outcome `Cancer_Avoidance_Mean`)

Table 4: Sample of Categorical Variable Counts and Percentages

Variable	Value	n	percent
AI_Use	Missing	41	0.4%
AI_Use	No	2262	20.3%
AI_Use	Yes	8852	79.4%
AgeBand	35+	6275	56.3%
AgeBand	Under 35	4880	43.7%
AgeGroup	18–24	1318	11.8%
AgeGroup	25–34	3562	31.9%
AgeGroup	35–44	3089	27.7%
AgeGroup	45–54	1738	15.6%
AgeGroup	55–64	962	8.6%
AgeGroup	65+	486	4.4%
Anxiety_Feeling_Afraid	Missing	42	0.4%
Anxiety_Feeling_Afraid	More than half the days	974	8.7%
Anxiety_Feeling_Afraid	Nearly every day	678	6.1%
Anxiety_Feeling_Afraid	Not at all	5839	52.3%

3.7.1 Random Forest Analysis (tidymodel)

3.7.1.1 Demographic Model



- RMSE (~0.63) and MAE (~0.53) are pretty close, meaning errors aren’t heavily dominated by outliers.
- Correlation = 0.109 is extremely low. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.

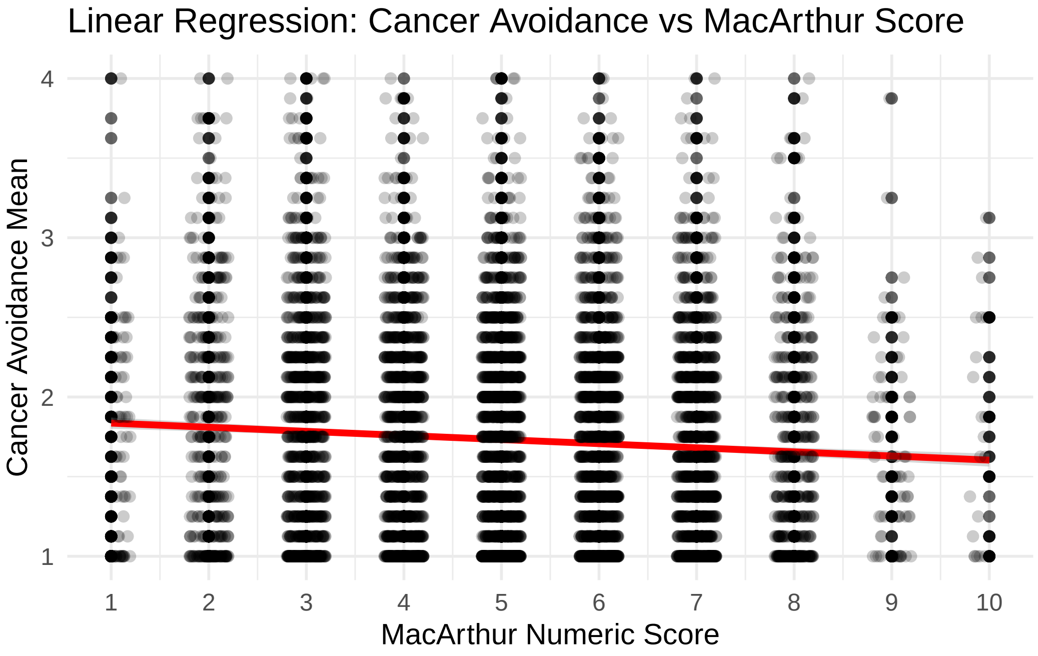
Table 5: Random Forest Model Performance (Demographic Predictors)

Metric	Value
RMSE	0.643
MAE	0.529
R-squared	0.012
Correlation	0.107

- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that `MacArthur_Numeric` is the most predict variable.

Term	Estimate	Std Error	Statistic	p value
Intercept	1.862	0.022	86.363	0e+00
MacArthur Numeric	-0.026	0.004	-6.451	1e-10

- The linear regression analysis shows a statistically significant negative relationship between MacArthur Scale scores and cancer avoidance mean (β = -0.025, p = 1.89e-10). For each one-unit increase in the MacArthur score (indicating higher subjective socioeconomic status), cancer avoidance mean scores decreased by 0.025 units.



- The scatter plot displays a slight negative trend (indicated by the red regression line), but the data points are widely dispersed across all MacArthur score levels.
- This visual pattern confirms the weak correlation, showing that while a statistical relationship exists, MacArthur score (subjective socioeconomic status) is not a strong predictor of cancer avoidance behaviors on its own.

Table 6: Linear Regression: Predicting Cancer Avoidance Mean from Age Group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.699	0.019	87.746	0.000
AgeGroup25–34	-0.003	0.023	-0.153	0.879
AgeGroup35–44	0.062	0.024	2.623	0.009
AgeGroup45–54	0.078	0.027	2.909	0.004
AgeGroup55–64	0.065	0.032	2.040	0.041
AgeGroup65+	-0.037	0.047	-0.796	0.426

Table 7: Linear Regression: Predicting Cancer Avoidance Mean from Age Band

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.69667	0.01024	165.71770	0e+00
AgeBand35+	0.06408	0.01411	4.54045	1e-05

- The initial analysis examined cancer avoidance scores across six age categories, revealing that middle-aged adults (35-64) showed significantly higher cancer avoidance behaviors compared to the youngest group (18-24). However, the model explains only 0.33% of the variance (R^2= 0.003), indicating weak predictive power despite statistical significance.
- So we break the `AgeGroup` into `AgeBand` with only 2 categories "above 35" and "under 35" to see if the result will be better.

- Individuals aged 35 and older had cancer avoidance scores 0.063 units higher than those under 35 (p = 6.45e-06), a highly significant difference. The model remains weak in explanatory power (R^2 = 0.0025), but the clear age threshold at 35 years provides a meaningful distinction for further investigation.
- Now, we wanted to see if other predictors would influence the `AgeBand` for under 35 and over 35, so we created two multivariable linear regression models (Age_under35 and Age_over35) predicting Cancer Avoidance Mean.

- For individuals under 35, political party affiliation emerged as the strongest predictor of cancer avoidance. Democrats showed 0.165 lower scores than Republicans (p = 8.97e-08),

Table 8: Linear Regression: Under 35 Predicting Cancer Avoidance Mean

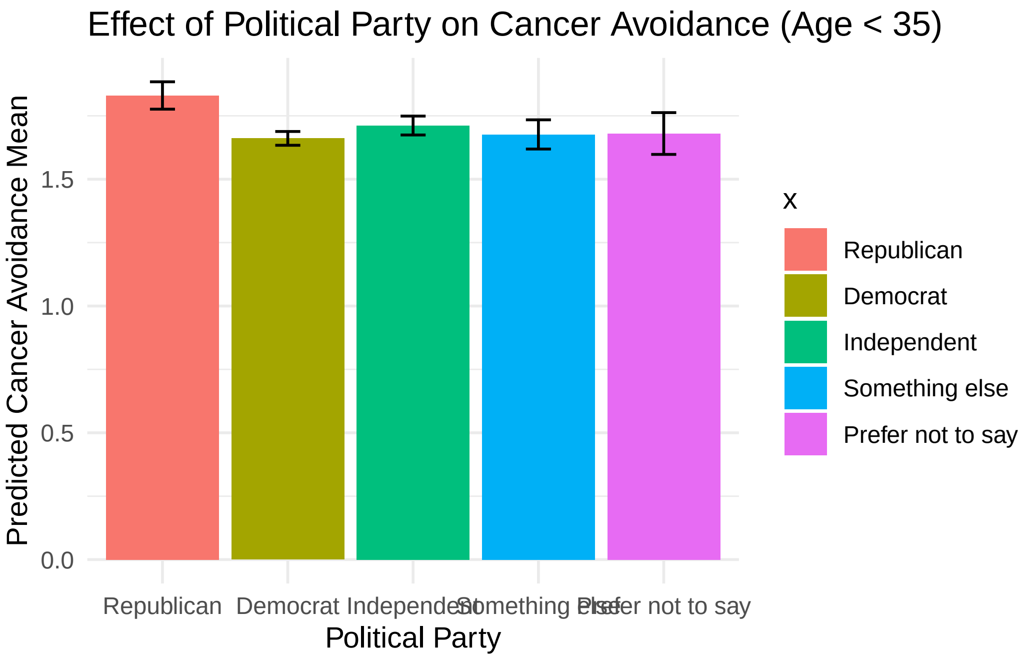
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.927	0.046	42.108	0.000
Education_Level	-0.009	0.008	-1.190	0.234
Income	0.003	0.006	0.484	0.628
MacArthur_Numeric	-0.013	0.006	-2.295	0.022
Political_PartyDemocrat	-0.169	0.031	-5.477	0.000
Political_PartyIndependent	-0.119	0.033	-3.542	0.000
Political_PartySomething else	-0.154	0.040	-3.805	0.000
Political_PartyPrefer not to say	-0.150	0.050	-2.990	0.003

Table 9: Linear Regression: Over 35 Predicting Cancer Avoidance Mean

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.160	0.044	48.839	0.000
Education_Level	-0.011	0.008	-1.349	0.178
Income	-0.016	0.007	-2.286	0.022
MacArthur_Numeric	-0.029	0.007	-4.499	0.000
Political_PartyDemocrat	-0.225	0.028	-7.935	0.000
Political_PartyIndependent	-0.131	0.030	-4.368	0.000
Political_PartySomething else	-0.138	0.041	-3.397	0.001
Political_PartyPrefer not to say	-0.093	0.057	-1.632	0.103

Independents showed 0.120 lower scores (p = 0.0003), and those selecting “Something else” showed 0.155 lower scores (p = 0.0001).

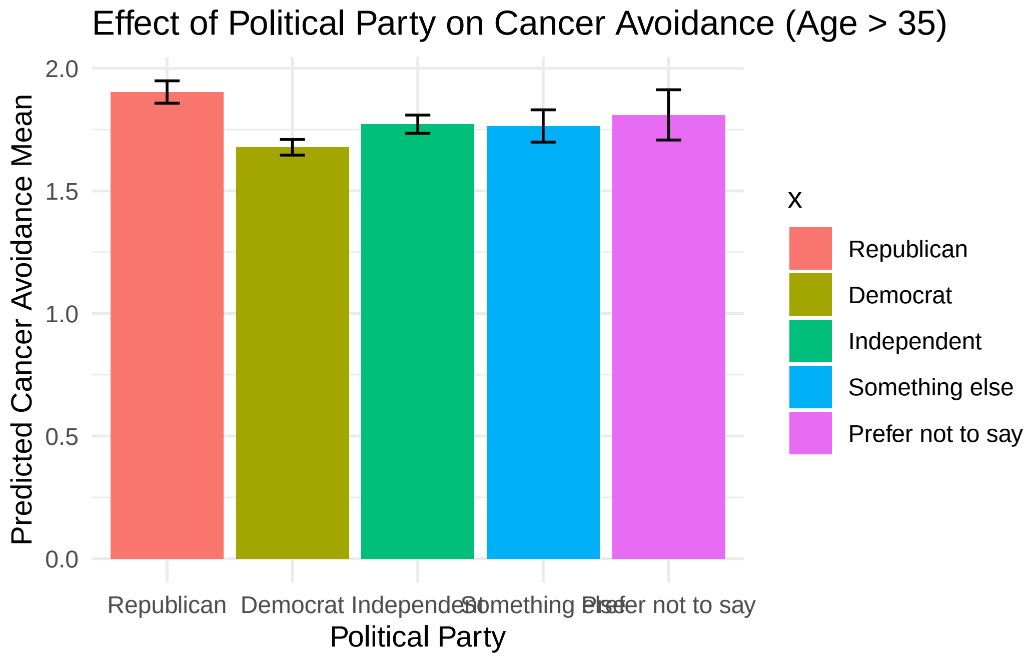
- But the overall model explaining less than 1% of variance ($R^2 = 0.0097$).



- The bar plots shows predicted cancer avoidance means across political affiliations, stratified by age. For those under 35, Republicans show the highest cancer avoidance behaviors (~1.80), while all other political groups cluster around 1.65, with error bars indicating moderate variability.

- For individuals aged 35 and older, political party affiliation showed even stronger effects than in the younger cohort. Democrats had 0.225 lower cancer avoidance scores than Republicans (p = 2.67e-15), Independents had 0.131 lower scores (p = 1.28e-05), and “Something else” respondents had 0.138 lower scores (p = 0.0007).

- Additionally, both income (β = -0.016, p = 0.022) and MacArthur score (β = -0.029, p = 7.00e-06) showed significant negative associations, with this model explaining 2.8% of the variance ($R^2 = 0.028$), the highest among all models tested.



- The bar plots shows predicted cancer avoidance means across political affiliations, stratified by age. For those 35 and older, Republicans maintain the highest scores (~1.93), while Democrats show notably lower scores (~1.68), and other groups fall in between.
- There has been a consistent gap between Republicans and Democrats in the two age groups. Republicans exhibited higher levels of cancer avoidance scores, suggesting that political party is strongly associated with these avoidance behaviors regardless of age, although this effect is more pronounced in older adults.

3.7.1.2 Media Use Model

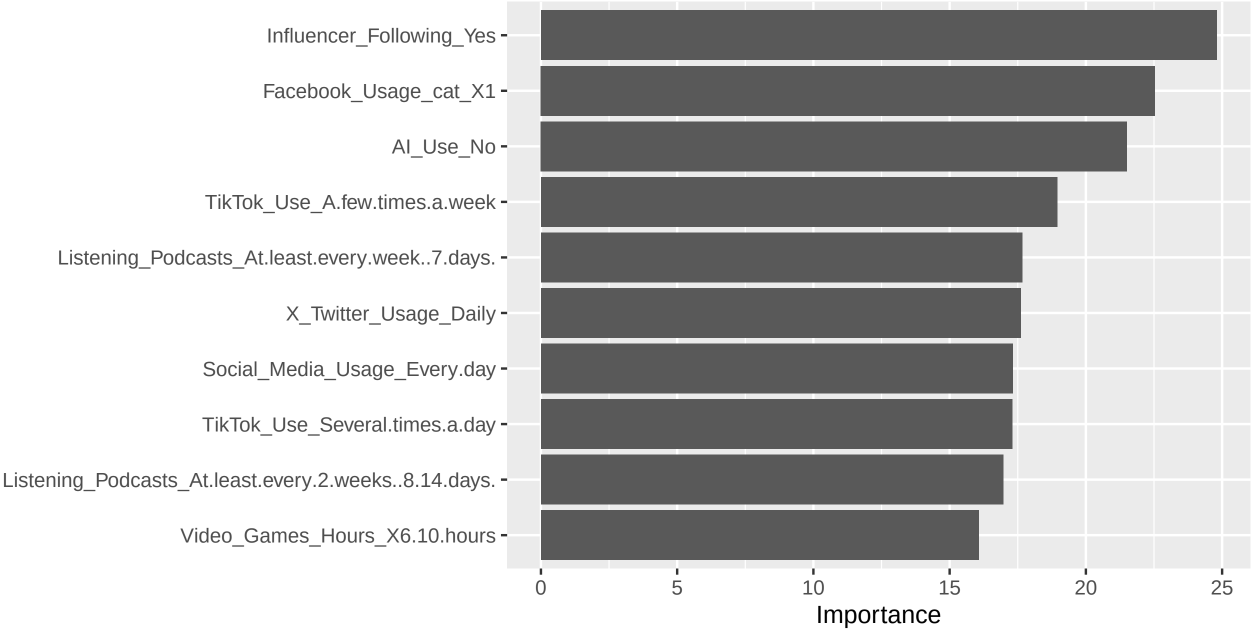


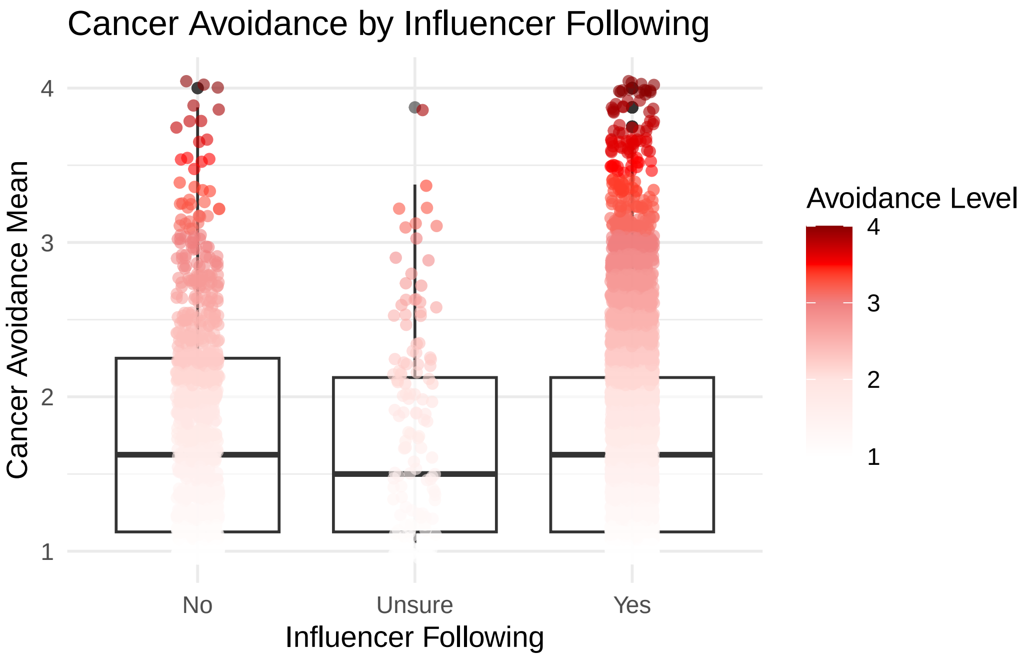
Table 10: Random Forest Model Performance (Media Usage Predictors)

Metric	Value
RMSE	0.664
MAE	0.551
R-squared	0.000
Correlation	0.021

Table 11: Linear Regression: Influencer Following Predicting Cancer Avoidance Mean

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.752	0.020	85.664	0.000
Influencer_FollowingUnsure	-0.038	0.058	-0.656	0.512
Influencer_FollowingYes	-0.011	0.022	-0.506	0.613

- RMSE (~0.66) and MAE (~0.55) are pretty close, meaning errors aren't heavily dominated by outliers.
- Correlation = 0.021 is extremely low. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.
- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that `Influencer_Following` is the most predictive variable.
- The linear regression analysis shows a none meaningful relationship between Influencer_Following scores and cancer avoidance mean (p > 0.05).



- The boxplot examining cancer avoidance scores across influencer following categories (No, Unsure, Yes) reveals remarkably similar distributions across all three groups. This visual pattern strongly supports the conclusion that influencer following status has no substantial relationship with cancer avoidance score.

3.7.1.3 Health Condition Model

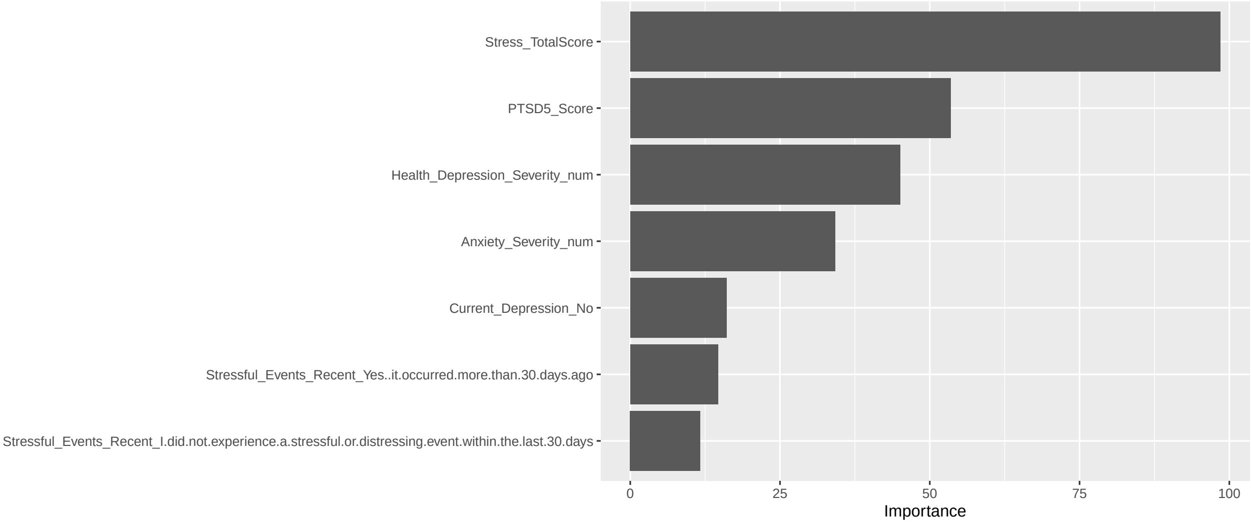


Table 12: Random Forest Model Performance (Health Condition Predictors)

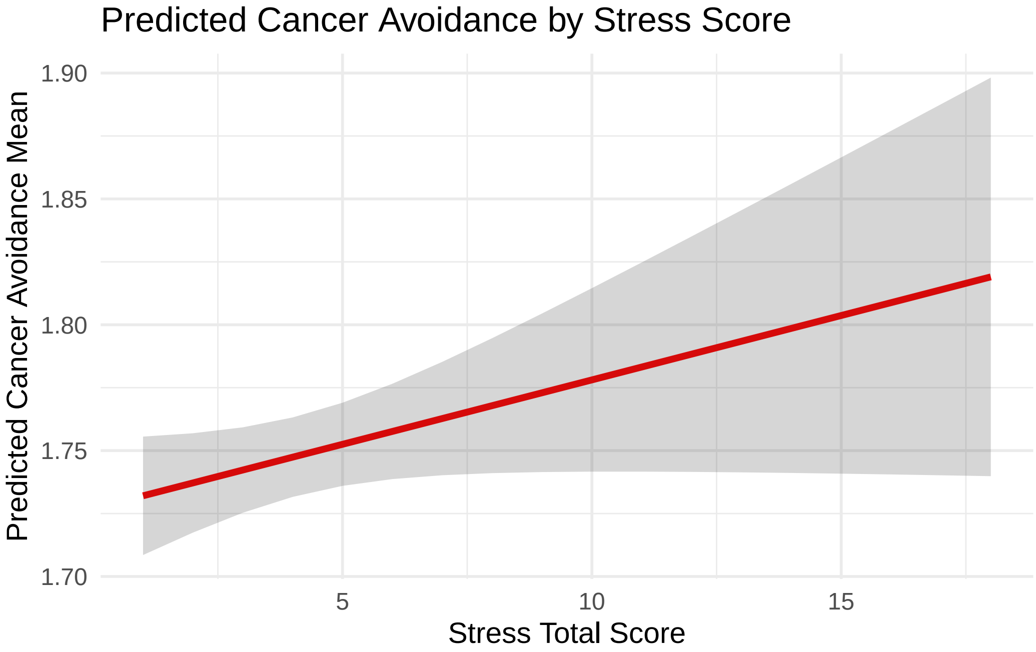
Metric	Value
RMSE	0.653
MAE	0.548
R-squared	0.011
Correlation	0.103

- RMSE (~0.65) and MAE (~0.55) are pretty close, meaning errors aren't heavily dominated by outliers.
- Correlation = 0.103 is extremely low. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.
- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that `Stress_TotalScore` is the most predict variable.

Table 13: Linear Regression: Stress Score Predicting Cancer Avoidance Mean

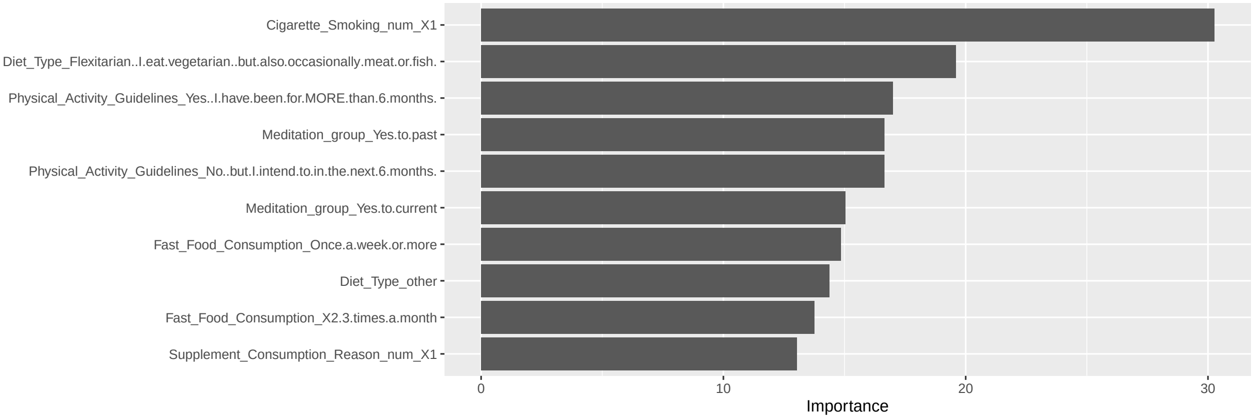
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.727	0.014	121.214	0.000
Stress_TotalScore	0.005	0.003	1.794	0.073

- The linear regression analysis shows a none meaningful relationship between stress total scores and cancer avoidance mean (p > 0.05).



- The plot displays predicted cancer avoidance scores across the range of stress total scores (0-18), showing a slight positive trend indicated by the red regression line. The broad gray confidence band around the regression line reflects the considerable uncertainty in these predictions. While higher stress levels may be weakly associated with slightly higher cancer avoidance behaviors, but this relationship is not statistically significant.

3.7.1.4 Health Behavior Model



- RMSE (~0.67) and MAE (~0.56) are pretty close, meaning errors aren't heavily dominated by outliers.

- Correlation = 0.06 is extremely low. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.

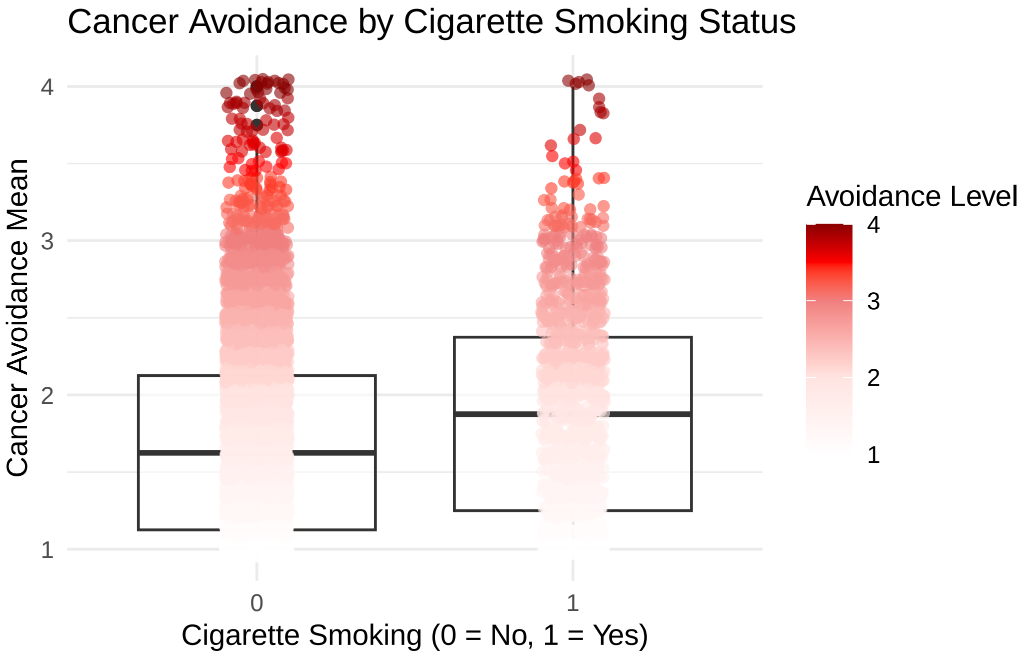
Table 14: Random Forest Model Performance (Health Behavior Predictors)

Metric	Value
RMSE	0.671
MAE	0.556
R-squared	0.004
Correlation	0.060

Table 15: Linear Regression: Smoking Predicting Cancer Avoidance Mean

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.708	0.008	201.753	0
Cigarette_Smoking_num1	0.186	0.021	8.860	0

- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that number of `Cigarette_Smoking` is the most predict variable.
- The linear regression analysis shows a statistically significant positive relationship between cigarette smoking and cancer avoidance behaviors ($\beta = 0.186$, p < 2e-16). For every one-unit increase in cigarette smoking status (from non-smoker to smoker), the predicted cancer avoidance score increases by 0.186 units, meaning individuals who smoke tend to have slightly higher cancer avoidance scores.
- While this relationship is conventionally significant at the 0.05 level, the model explains only 1.1% of the variance ($R^2 = 0.011$), indicating that smoking status alone provides minimal predictive power for cancer avoidance scores.



- Look at the box plot, smokers show higher cancer avoidance scores on average compared to non-smokers. This could mean that smokers are more aware of or concerned about cancer risks, possibly leading them to engage in other cancer-preventive behaviors. However, the low variance reminds us that while this pattern is statistically significant, smoking status explains only a small fraction of the variation in cancer avoidance scores.

3.7.1.5 Other Model

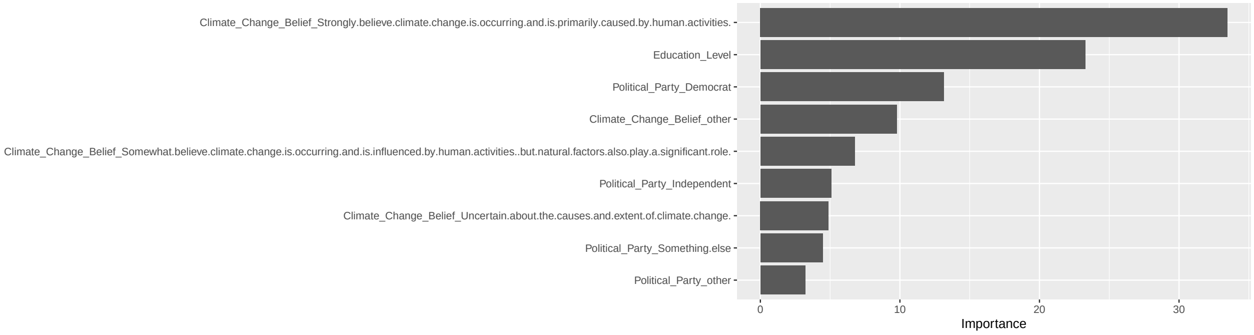


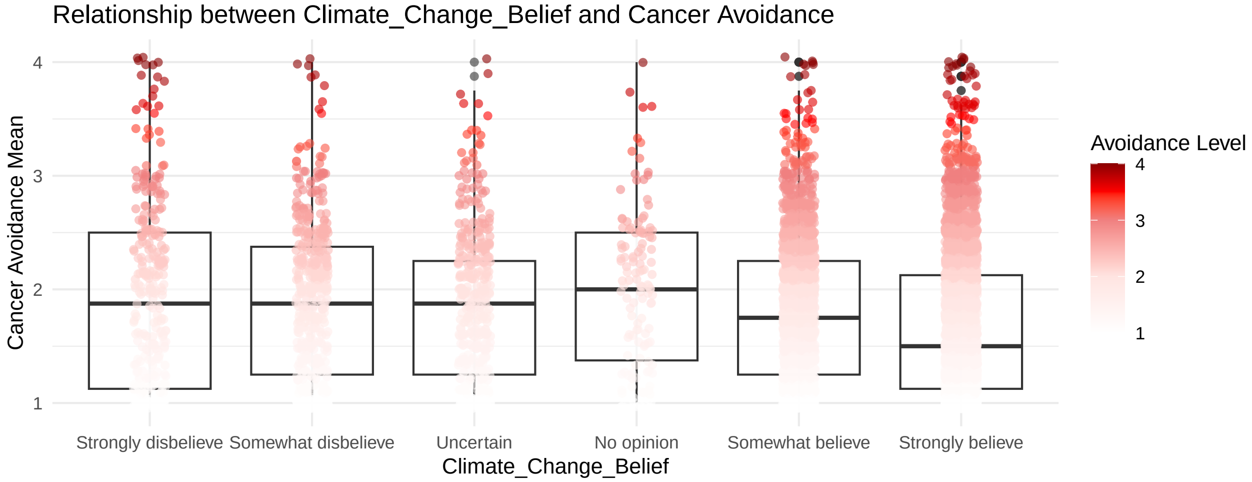
Table 16: Random Forest Model Performance (Other Predictors)

Metric	Value
RMSE	0.631
MAE	0.525
R-squared	0.021
Correlation	0.145

- RMSE (~0.63) and MAE (~0.52) are pretty close, meaning errors aren't heavily dominated by outliers.
- Correlation = 0.145 is extremely low. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.
- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that `Climate_Change_Belief` (Strongly believe climate change is occurring and is primarily caused by human activities) is the most predict variable.
- The linear regression shows significant associations between climate change beliefs and cancer avoidance scores. Individuals who strongly believe in human-caused climate change showed significantly lower cancer avoidance scores (β = -0.271, p = 3.32e-13), meaning their scores were 0.271 units lower on average. The model explains 2.4% of the variance (R^2 = 0.024), indicating a weak but statistically significant relationship.

Table 17: Linear Regression: Cliamte Change Belief Predicting Cancer Avoidance Mean

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change, believing
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human activities, bu
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by human activitie



- The boxplot displays cancer avoidance behaviors across six climate change belief categories, revealing a clear gradient pattern. Climate change deniers and no opinion show higher median cancer avoidance scores around 1.9 - 2, with relatively compact distributions. Which means Individuals who strongly believe in human-caused climate change showed significantly lower cancer avoidance scores. However, the low variance reminds us that while this pattern is statistically significant, undefined explains only a small fraction of the variation in cancer avoidance scores.

3.7.2 Multivariate Adaptive Regression Splines (MARS)

3.7.2.1 Demographic

- `Political_PartyDemocrat` is the most influential predictor (baseline = 100).
- `MacArthur_Numeric` is about 75% as important.
- `RaceWhite` is about 63%.

Interactions appear as products (such as):

- Ethnicity Prefer not to say * Political_Party Democrat = an interaction term between ethnicity and political party.
- Selected model at 17 terms

Table 18: MARS Model Coefficients (Demographic Predictors)

Term	Coefficient
(Intercept)	1.748
Political_PartyDemocrat	-0.113
h(MacArthur_Numeric-9)	0.275
h(9-MacArthur_Numeric)	0.019
h(Age-56)	-0.014
h(56-Age)	-0.003
RaceWhite	0.080
h(Income-2)	-0.017
Job_ClassificationProfessional	-0.093
Job_ClassificationBlue Collar	0.127
Job_ClassificationUnemployed/Student/Parent	-0.074

Table 19: MARS Variable Importance (Demographic Predictors)

Variable	N Subsets	GCV	RSS
Political_PartyDemocrat	6	1	10
MacArthur_Numeric	37	1	9
RaceWhite	33	1	8
Job_ClassificationProfessional	17	1	7
Job_ClassificationUnemployed/Student/Parent	18	1	6
Job_ClassificationBlue Collar	16	1	5
Age	21	1	4
Income	22	1	2

Table 20: MARS Model Summary Statistics

Metric	Value
Selected Terms	11.000
R-squared	0.028
GRSq	0.021

- GRSq/RSq converging around 0.03-0.04
- Mean out-of-fold RSq near zero (poor predictive performance)
- High cross-validation error (CVRSq = 0.017, sd = 0.019)
- This suggests demographic variables don’t explain Cancer_Avoidance_Mean very well.

Table 21: MARS Model Coefficients with Cross-Validation (Top 10 Terms)

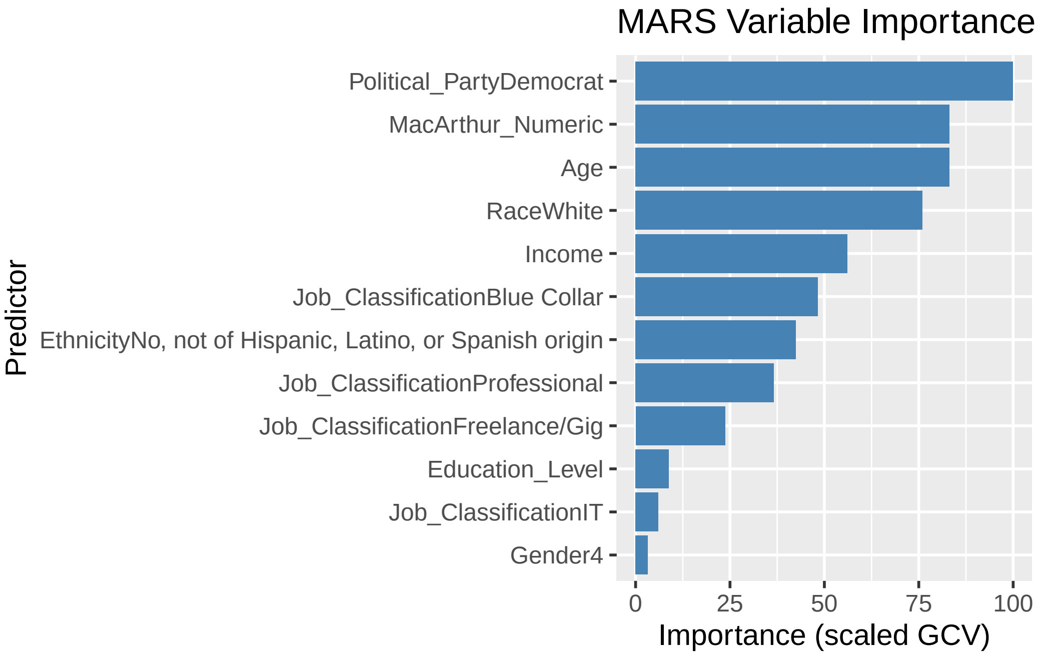
Term	Coefficient
(Intercept)	1.519
Political_PartyDemocrat	-0.124
h(MacArthur_Numeric-9)	0.917
h(56-Age)	0.009
h(Age-26)*h(9-MacArthur_Numeric)	0.001
RaceWhite	0.184
Job_ClassificationProfessional*h(9-MacArthur_Numeric)	-0.024
h(Income-2)*RaceWhite	-0.039
EthnicityNo, not of Hispanic, Latino, or Spanish origin*h(MacArthur_Numeric-9)	-0.957
Job_ClassificationBlue Collar*h(9-MacArthur_Numeric)	0.038

Table 22: MARS Model Summary Statistics

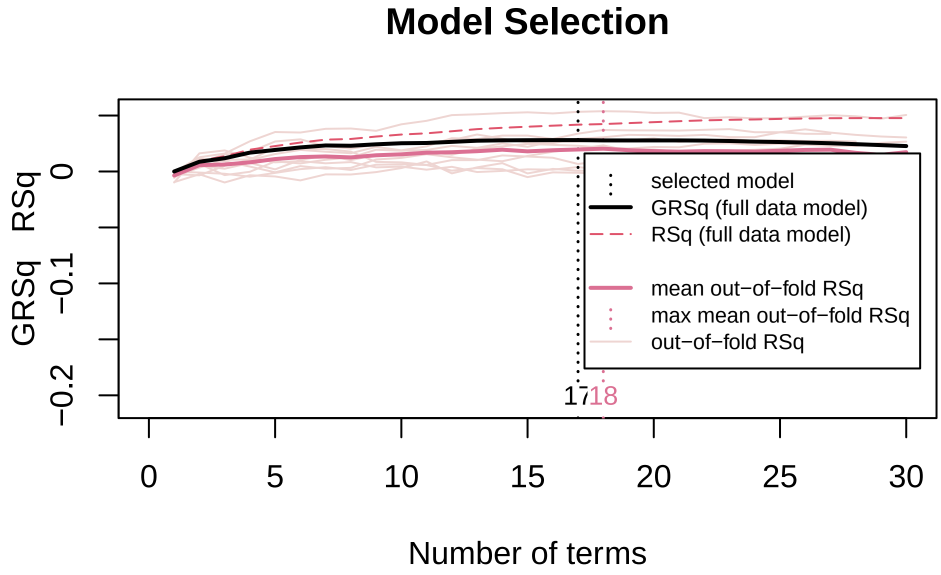
Metric	Value
Selected Terms	17.0000
R-squared	0.0418
GRSq	0.0281

Table 23: Cross-Validation Results by Number of Terms

Terms	Cancer_Avoidance_Mean	mean
fold7	0.0062	0.0062
fold8	0.0258	0.0258
fold9	0.0267	0.0267
fold10	-0.0009	-0.0009
mean	0.0175	0.0175



Visualizing MARS models



- The model selection plot shows that both RSq (training fit) and GRSq (generalized R-square) values remain low. This indicates that demographic characteristics provide limited explanatory power for cancer avoidance mean, and adding nonlinear terms does not meaningfully improve predictive accuracy.
- Demographics status explain very little of the variation in cancer avoidance. Where effects appear, they are narrow interactions, such as between political affiliation and ethnicity, or between Job_Classification Blue Collar and MacArthur_Numeric > 3.

3.7.2.2 Health condition

Table 24: MARS Model Coefficients (Health Condition Predictors)

Term	Coefficient
(Intercept)	1.760
h(Anxiety_Severity_num-3)	0.144
h(1-PTSD5_Score)	-0.088

Table 25: MARS Variable Importance (Health Condition Predictors)

Variable	N Subsets	GCV	RSS
Anxiety_Severity_num	4	1	2
PTSD5_Score	5	1	1

- `Anxiety_Severity_num` is the most influential predictor (baseline = 100).
- `PTSD5_Score` is about 61%

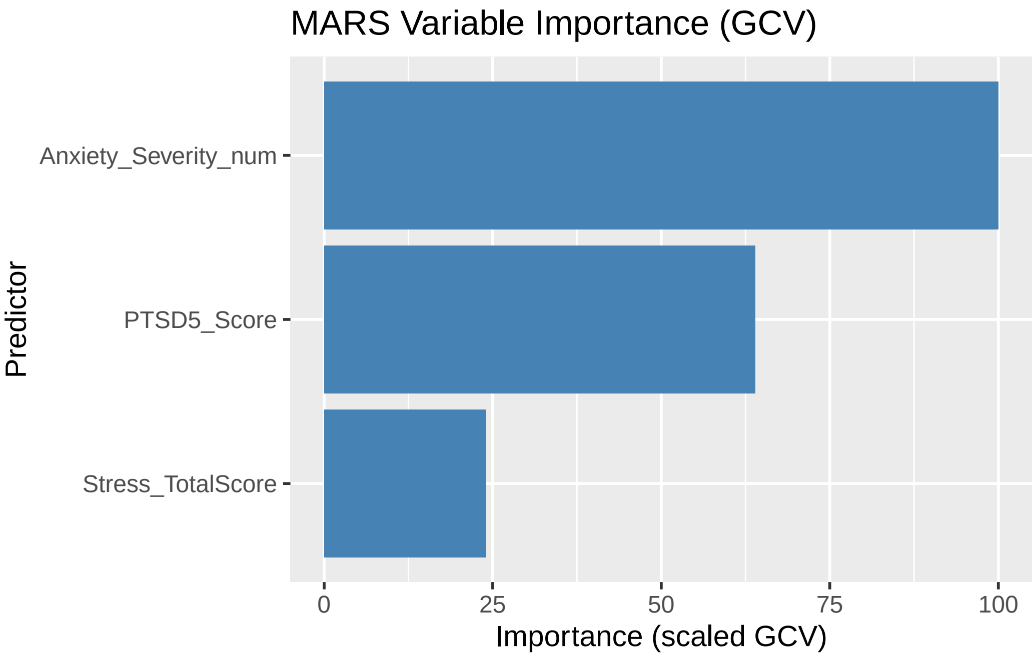
Table 26: MARS Model Summary Statistics (Health Condition)

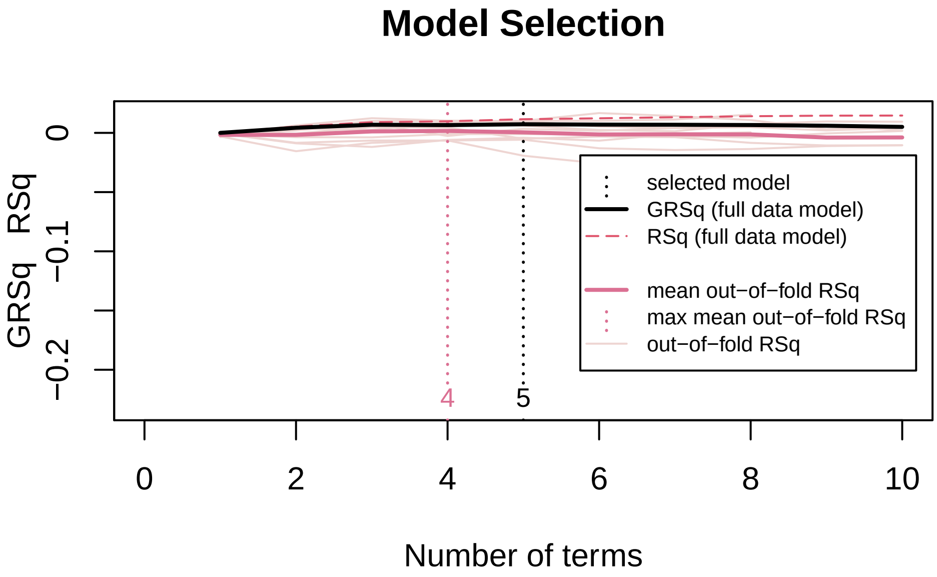
Metric	Value
Selected Terms	3.000
R-squared	0.009
GRSq	0.007

```
                           [,1]
Cancer_Avoidance_Mean 0.1452411
```

Interactions appear as products (such as):

- h(Anxiety_Severity_num-3) = 0.191: When anxiety severity score > 3, cancer avoidance increases by 0.191 units.
- h(1-PTSD5_Score) = -0.088: When PTSD score < 1, cancer avoidance decreases by 0.088 units, meaning higher PTSD scores are associated with slightly higher avoidance.
- Selected model at 5 terms
- GRSq/RSq barely above zero (GRSq = 0.0073, RSq = 0.012)
- Mean out-of-fold RSq is negative - confirms no predictive power
- CVRSq = -0.003, MaxErr = 2.35, This suggests health condition variables don't explain Cancer_Avoidance_Mean very well.
- Predictive power is weak: training R-square ~ 0.065, The model explains about 6.5% of the variance in Cancer_Avoidance_Mean on the training data and is a weak fit.
- The correlation between Cancer_Avoidance_Mean and health condition is also very week (0.1452411)





- The model selection plot shows that both RSq (training fit) and GRSq (generalized R-square) initially increase, indicating that adding the first few hinge functions improves model performance. However, after approximately 4 terms, GRSq reaches its maximum and then slightly declines, suggesting that additional terms provide minimal benefit and may cause overfitting.
- The selected model at 4 terms uses 4 predictors and achieves a modest GRSq (0.01), indicating that although weak nonlinear relationships exist, the overall predictive power of the model remains low.

3.8 CLASSIFICATION MODEL

(binary outcome `Cancer_Avoiders01`)

3.8.1 Random Forest

3.8.1.1 Demographic Model

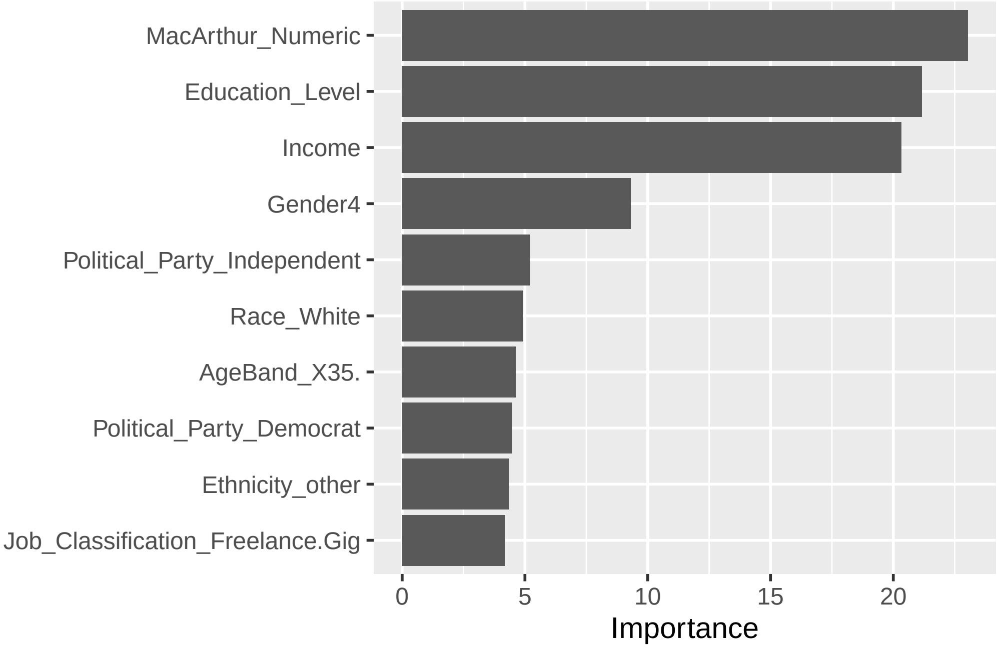


Table 27: Random Forest Classification Model Performance (Demographic Predictors)

Metric	Value
accuracy	0.955
kap	0.000

Table 28: Confusion Matrix

Truth	0	1
0	2306	108
1	0	0

Table 29: ROC AUC Score

Metric	Value
ROC AUC	0.449

- High accuracy (0.955) simply reflects predicting majority class.

Table 30: Logistic Regression: MacArthur Numeric Predicting Cancer Avoiders

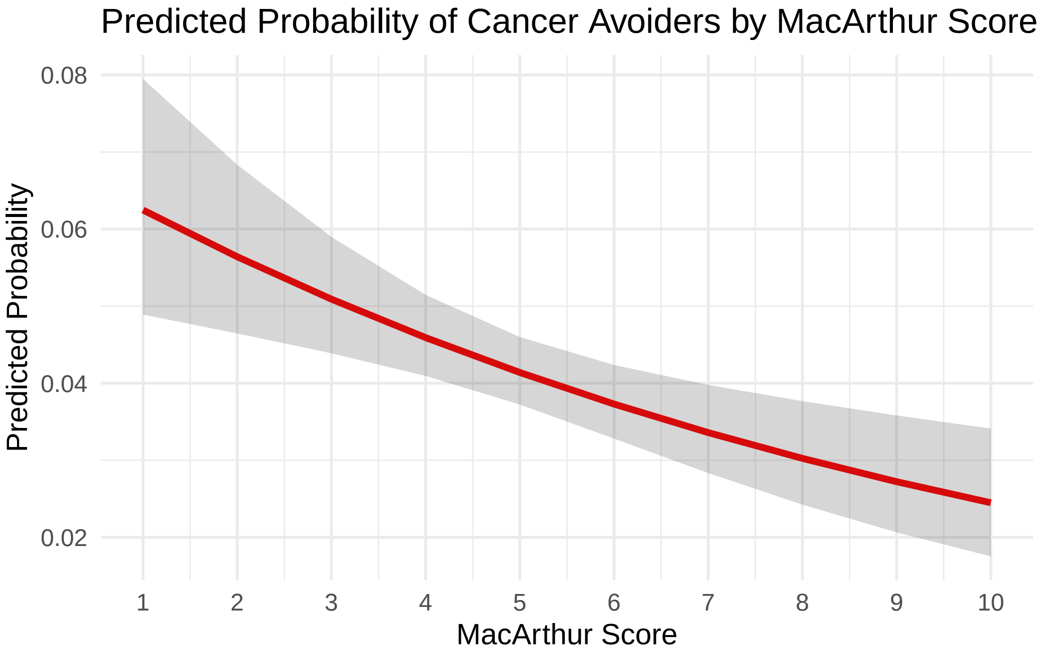
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.600	0.162	-16.100	0.000
MacArthur_Numeric	-0.108	0.032	-3.427	0.001

- ROC AUC = 0.449 , which means no better than random
- The random forest model shows that `MacArthur_Numeric` is the most predict variable.

- The logistic regression shows a statistically significant negative relationship between MacArthur scores and the probability of being a cancer avoider (β = -0.108, p = 0.0006). For each one-unit increase in the MacArthur score, the log-odds of being a cancer avoider decrease by 0.108, meaning individuals with higher subjective socioeconomic status are less likely to be classified as cancer avoiders.
- However, the model shows minimal improvement with only 11.8 point reduction in deviance over the null model (null deviance = 2785.6, residual deviance = 2773.8).

```
Area under the curve: 0.5511
```

- ROC AUC = 0.55, which means slightly better than random



- The plot displays the predicted probability of being a cancer avoider across MacArthur scores ranging from 1 to 10. The red line shows a clear downward trend. The gray confidence band widens slightly at the extremes but remains relatively narrow, indicating consistent uncertainty across the score range. This visualization confirms the negative relationship identified in the logistic regression, higher socioeconomic status is associated with lower probability of being a cancer avoider.

- However, the poor ROC AUC value (0.449) indicates that although the relationship is statistically significant, the model has little practical use in identifying cancer avoiders.

3.8.1.2 Media Usage Model

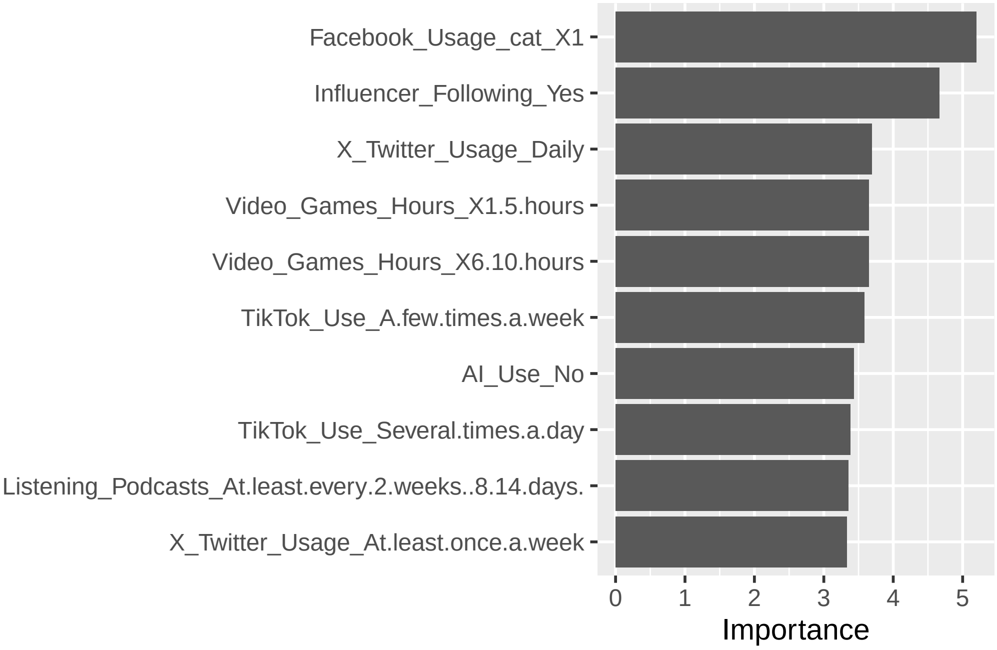


Table 31: Random Forest Classification Model Performance (Media Usage Predictors)

Metric	Value
accuracy	0.947
kap	0.000

Table 32: Confusion Matrix

Truth	0	1
0	1939	109
1	0	0

Table 33: ROC AUC Score

Metric	Value
ROC AUC	0.574

- High accuracy (0.947) simply reflects predicting majority class.
- ROC AUC = 0.574 , which means slightly better than random
- The random forest model shows that `Facebook_Usage` is the most predict variable.

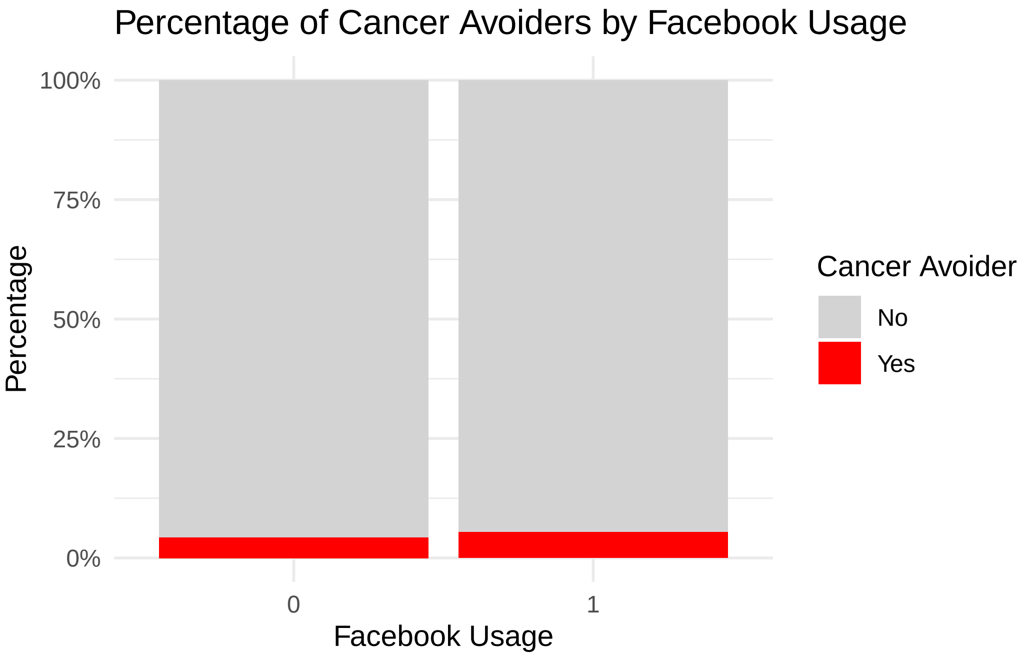
Table 34: Logistic Regression: Facebook Usage Predicting Cancer Avoiders

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.104	0.087	-35.680	0.000
Facebook_Usage_cat1	0.236	0.114	2.067	0.039

- The logistic regression shows a statistically significant positive relationship between Facebook usage and the probability of being a cancer avoider ($\beta = 0.2358$, p = 0.0388). Facebook users have odds of being a cancer avoider that are 1.27 times higher ($e^{0}.2358 = 1.266$) than non-users, representing a 26.6% increase in odds.
- However, the model shows minimal improvement over the null model, with only a 4.3 point reduction in deviance (null deviance = 2654.9, residual deviance = 2650.6), indicating that Facebook usage explains none of the variation in cancer avoider.

```
Area under the curve: 0.5291
```

- ROC AUC = 0.53, which means slightly better than random



- The stacked bar chart displays the percentage distribution of cancer avoiders and non-avoiders across Facebook usage groups (0 = No, 1 = Yes). Both groups show similar distributions, with cancer avoiders (red) representing a very small percentage at the bottom of each bar. While the logistic regression identified a statistically significant difference (p = 0.039), the visual similarity between the two bars reinforces that this difference has minimal value for predicting cancer avoider status.

3.8.1.3 Health Condition Model

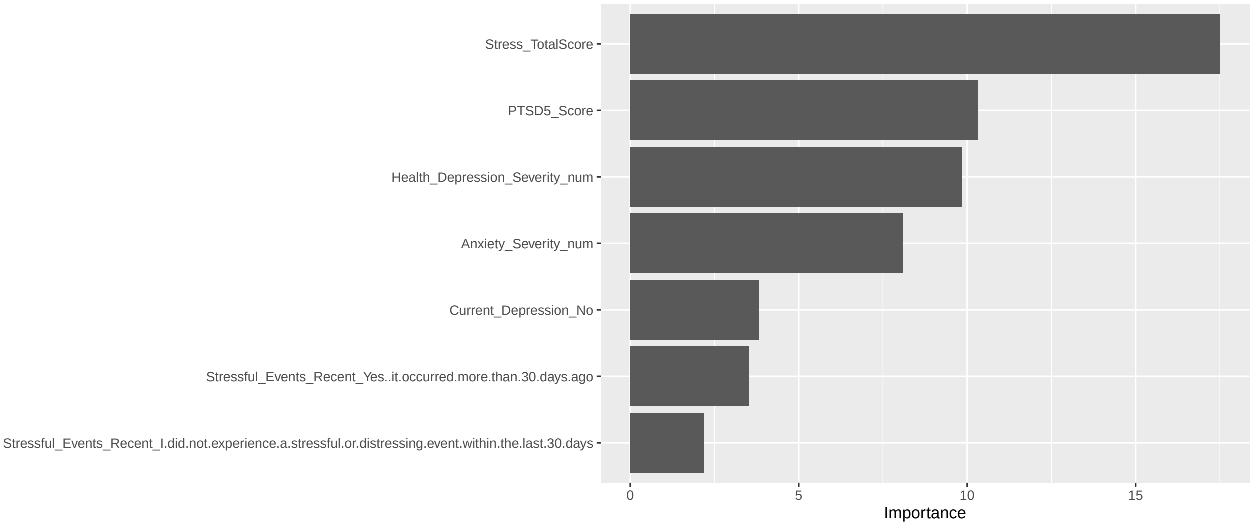


Table 35: Random Forest Classification Model Performance (Health COndition Predictors)

Metric	Value
accuracy	0.948
kap	0.000

Table 36: Confusion Matrix

Truth	0	1
0	1895	103
1	0	0

Table 37: ROC AUC Score

Metric	Value
ROC AUC	0.446

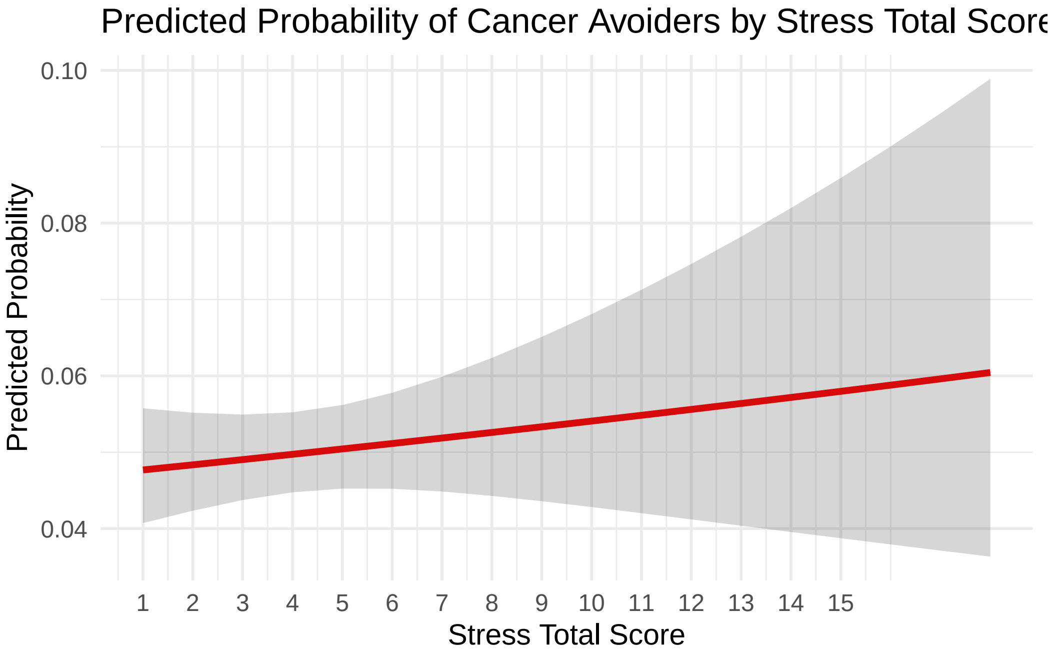
- High accuracy (0.948) simply reflects predicting majority class.
- ROC AUC = 0.446, which means no better than random
- The random forest model shows that `Stress_TotalScore` is the most predict variable.
- The logistic regression shows a none significant relationship between stress total score and the probability of being a cancer avoider ($\beta = 0.01475$, p = 0.447).

```
Area under the curve: 0.5127
```

- ROC AUC = 0.51, which means slightly better than random

Table 38: Logistic Regression: Stress Total Score Predicting Cancer Avoiders

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.009	0.099	-30.260	0.000
Stress_TotalScore	0.015	0.019	0.761	0.447



- The plot displays the predicted probability of being a cancer avoider across stress total scores ranging from 1 to 18. The red line shows a very slight upward trend. The gray confidence band widens at higher stress scores, indicating increasing uncertainty in predictions for individuals with very high stress levels. Which demonstrates that stress total score provides essentially no predictive information for identifying cancer avoiders, consistent with the lack of statistical significance in the model.

3.8.1.4 Health Behavior Model

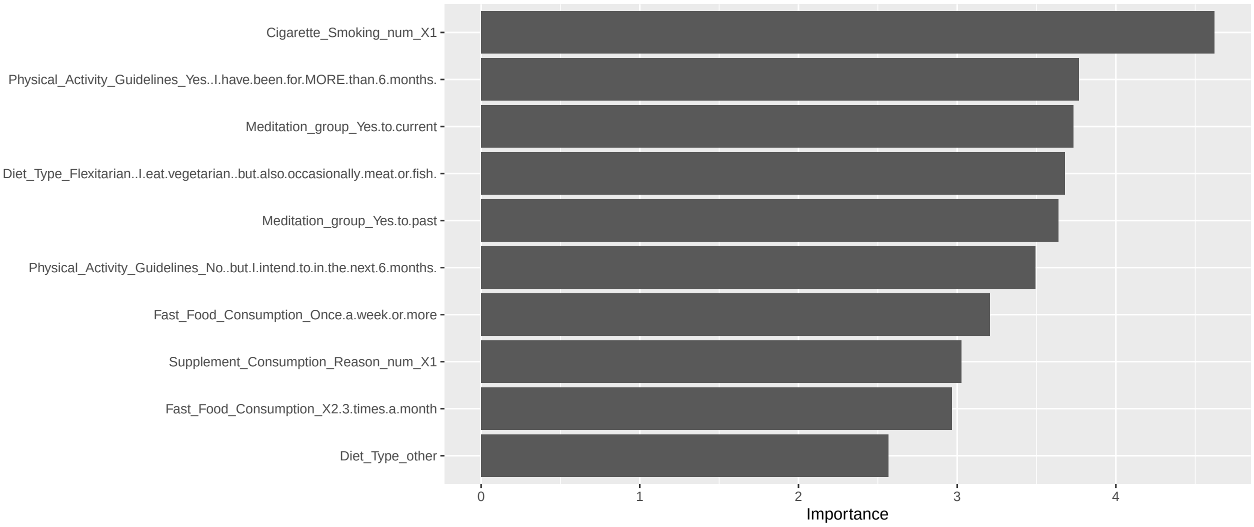


Table 39: Random Forest Classification Model Performance (Health Behavior Predictors)

Metric	Value
accuracy	0.946
kap	0.000

Table 40: Confusion Matrix

Truth	0	1
0	1975	113
1	0	0

Table 41: ROC AUC Score

Metric	Value
ROC AUC	0.472

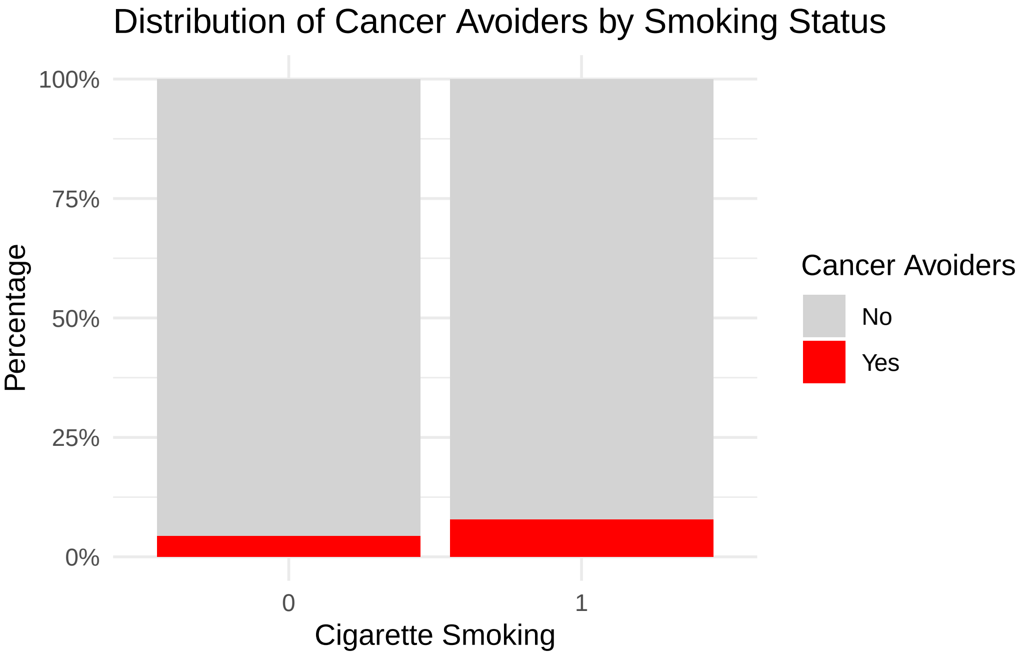
- High accuracy (0.946) simply reflects predicting majority class.
- ROC AUC = 0.472, which means no better than random
- The random forest model shows that `Cigarette_Smoking` is the most predict variable.
- The logistic regression shows a statistically significant positive relationship between smokers and the probability of being a cancer avoider (β = 0.62694, p = 1.05e-06). Smokers have odds of being a cancer avoider that are 1.87 times higher (e^0.62694 = 1.872) compared to non-smokers, representing an 87.2% increase in odds
- However, the model shows modest improvement over the null model, with a 21.7 point reduction in deviance (null deviance = 2709.8, residual deviance = 2688.1), indicating that Facebook usage explains modest of the variation in cancer avoider.

Table 42: Logistic Regression: Cigarette Smoking Predicting Cancer Avoiders

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1004153	0.06452494	-48.049874	0.00e+00
Cigarette_Smoking_num1	0.6269373	0.12839376	4.882927	1.05e-06

```
Area under the curve: 0.5509
```

- ROC AUC = 0.55, which means slightly better than random



- The stacked bar chart displays the percentage distribution of cancer avoiders and non-avoiders across smoking status (0 = Non-smoker, 1 = Smoker). Smokers display a larger red segment at approximately 8-9% cancer avoiders. This visual pattern confirms the logistic regression finding, smokers are roughly twice as likely to be cancer avoiders compared to non-smokers.
- However, it's important to note that despite this significant association, cancer avoiders remain a minority in both groups, with over 90% of both smokers and non-smokers classified as non-avoiders.

3.8.1.5 Other Model

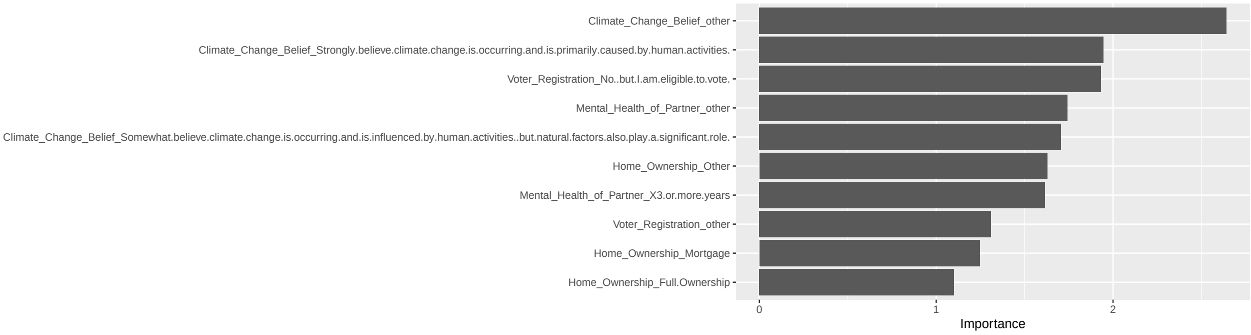


Table 43: Random Forest Classification Model Performance (Other Predictors)

Metric	Value
accuracy	0.959
kap	0.000

Table 44: Confusion Matrix

Truth	0	1
0	2106	91
1	0	0

Table 45: ROC AUC Score

Metric	Value
ROC AUC	0.42

- High accuracy (0.96) simply reflects predicting majority class.
- ROC AUC = 0.42, which means no better than random
- The random forest model shows that `Climate_Change_Belief`(other) is the most predict variable.

- The logistic regression shows significant negative associations between accepting human caused climate change and the probability of being a cancer avoider. Individuals who strongly believe climate change is primarily caused by human activities have 71% lower odds (1 - $e^{-}1.236$ = 0.71) of being cancer avoiders (β = -1.236, p = 1.87e-10) compared to climate dennier.
- The model shows modest improvement over the null model with a 50.1 point deviance reduction (null deviance = 2781.4, residual deviance = 2731.3)

```
Area under the curve: 0.5879
```

- ROC AUC = 0.58, which means slightly better than random

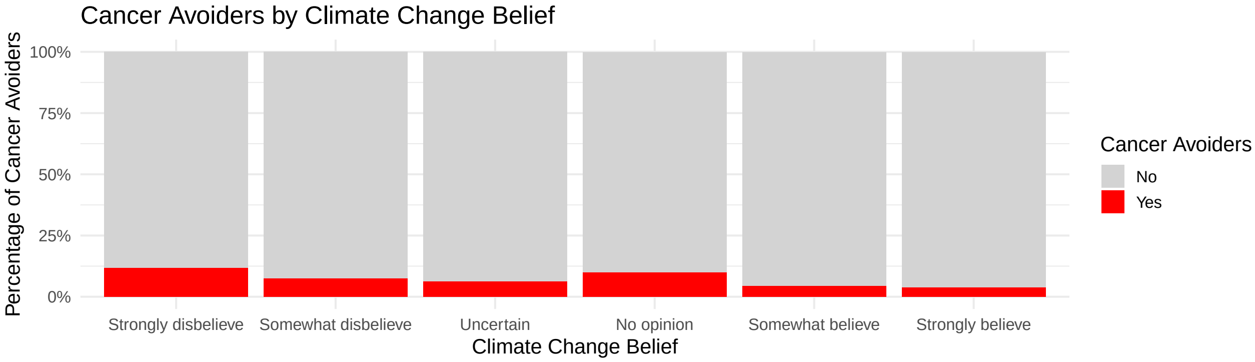


Table 46: Logistic Regression: Climate Change Belief Predicting Cancer Avoiders

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change, believing
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human activities, bu
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by human activitie

- The stacked bar chart displays the percentage distribution of cancer avoiders across six climate change belief categories. This visual pattern demonstrates that individuals who deny about human-caused climate change are 3-4 times more likely to be cancer avoiders compared to those who strongly accept about human-caused climate change.

3.9 Comprehensive Model with All Predictors

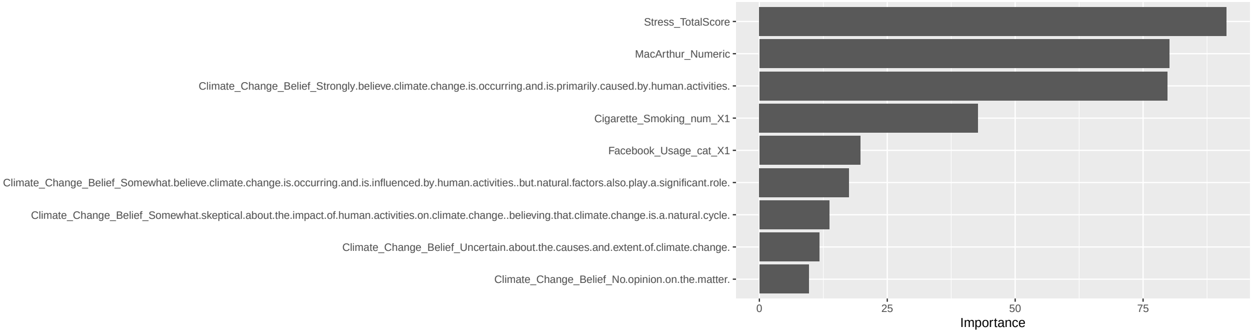


Table 47: Comprehensive Random Forest Model Performance (Regression)

Metric	Value
RMSE	0.620
MAE	0.516
R-squared	0.088
Correlation	0.296

- RMSE (~0.62) and MAE (~0.52) are pretty close, meaning errors aren't heavily dominated by outliers.
- Correlation = 0.3 is low. But it imporve a little bit compare to all the regression models. It basically means that predicted values do not track the actual values at all. So the model is not capturing the pattern in the test data.
- If predictors have very weak relationships with the outcome, the model will predicts values near the mean of the training set.
- The random forest model shows that `Stress_TotalScore` is the most predict variable.

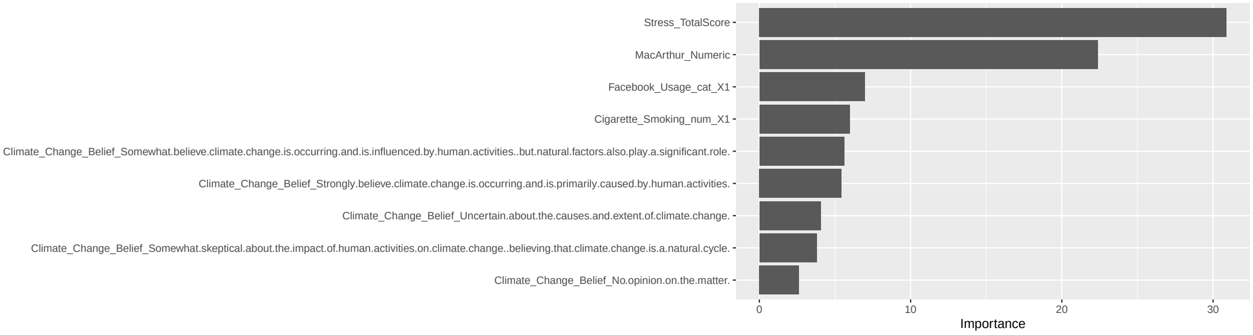


Table 48: Performance Metrics for Binary Outcome Model

Metric	Estimate
accuracy	0.952
kap	0.000

- High accuracy (0.95) simply reflects predicting majority class.
- ROC AUC = 0.23, which means no better than random, and is worse than all the classification models.
- The random forest model shows that `Stress_TotalScore` is the most predict variable.

4 Discussion

This analysis examined predictors of cancer avoidance scores across demographic, media-related, health, behavioral, and other domains using regression, classification, and MARS models. Across all approaches, the findings point to the same conclusion is that many predictors reached significance, but none meaningfully predicted cancer avoidance.

Regression models explained less than 3% of variance ($R^2 < 0.03$), and the full model reached only a modest correlation of 0.30. Classification models showed high accuracy due to class imbalance, but poor distinction, with ROC AUC values between 0.23 and 0.49, at below chance. The full classification model performed worst (AUC = 0.23), reinforcing that adding strong predictors does not improve performance.

MARS models found some nonlinear patterns, but the negative cross-validated R^2 shows they likely don't hold up and reflect overfitting. Overall, the small predictive value suggests important factors are missing. The low number of cancer avoiders (4–5%) also makes prediction harder. The unexpected links with smoking and climate beliefs may come from unmeasured factors or differences in how people understand cancer avoidance.

4.1 Limitations

- Cross-sectional design: Because the data were collected at one point in time, we cannot tell what causes what. Other unmeasured factors may also affect the results.

- Self-reported outcome: People may interpret “cancer avoidance” differently, which means the score may not fully reflect their real behaviors.
- Small effect sizes: Many predictors were statistically significant but had tiny effects, likely because the sample size was large.
- Inconsistent samples across models: Missing data led to different subsets being used, limiting direct comparisons.
- Random forest issues: The variable importance results were not consistent with simpler models and may exaggerate weak patterns.
- Limited generalizability: Results may not generalize to other groups or contexts.

4.2 Future Directions

Future studies should: (1) use longitudinal data to better understand cause-and-effect; (2) explore factors that might explain or change the relationships found; (3) oversample cancer avoiders to address class imbalance; (4) include qualitative work to understand unexpected patterns; and (5) validate these findings in independent samples.

4.3 Acknowledgments

We thank the following people and organizations for their guidance, support, and resources in this project:

- **Dr. Shane McCarty** (Binghamton University) – Principal Investigator and mentor
- **Dr. Heather Orom** (University at Buffalo) – Principal Investigator and mentor
- **Dr. Kargin Vladislav** (Binghamton University) – Principal Investigator and mentor
- **Cloud Research** – Owner of the Health Avoiders dataset and provider of access to de-identified survey data

4.4 Appendix A: Detailed Statistical Output

4.4.1 Linear Regression

4.4.1.1 MacArthur Scale vs cancer-avoidance score

```r
MacArthur_cancer_linear <- lm(Cancer_Avoidance_Mean ~ MacArthur_Numeric, data = demo_data)
summary(MacArthur_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ MacArthur_Numeric, data = demo_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83605 -0.58208 -0.08208  0.44371  2.34450 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.861837   0.021558  86.363  < 2e-16 ***
MacArthur_Numeric -0.025792   0.003998  -6.451 1.18e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6312 on 8043 degrees of freedom
Multiple R-squared:  0.005147,	Adjusted R-squared:  0.005023 
F-statistic: 41.61 on 1 and 8043 DF,  p-value: 1.178e-10
```

4.4.1.2 Age vs cancer-avoidance score

```r
agegroup_cancer_linear <- lm(Cancer_Avoidance_Mean ~ AgeGroup, data = demo_data)
summary(age_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ AgeBand, data = demo_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76075 -0.57167 -0.07167  0.42833  2.30333 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.69667    0.01024  165.72  < 2e-16 ***
AgeBand35+   0.06408    0.01411    4.54  5.7e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.632 on 8043 degrees of freedom
Multiple R-squared:  0.002557,	Adjusted R-squared:  0.002433 
F-statistic: 20.62 on 1 and 8043 DF,  p-value: 5.695e-06
```

4.4.1.3 Age vs cancer-avoidance score

```r
ageband_cancer_linear <- lm(Cancer_Avoidance_Mean ~ AgeBand, data = demo_data)
summary(age_cancer_linear)
```

```

Call:
lm(formula = Cancer_Avoidance_Mean ~ AgeBand, data = demo_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.76075 -0.57167 -0.07167  0.42833  2.30333 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.69667    0.01024  165.72  < 2e-16 ***
AgeBand35+   0.06408    0.01411    4.54  5.7e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.632 on 8043 degrees of freedom
Multiple R-squared:  0.002557,	Adjusted R-squared:  0.002433 
F-statistic: 20.62 on 1 and 8043 DF,  p-value: 5.695e-06
```

4.4.1.4 Age over 35 vs cancer-avoidance score

```r
model_over35 <- lm(Cancer_Avoidance_Mean ~ Education_Level + Income +
                   MacArthur_Numeric + Political_Party,
                   data = dplyr::filter(demo_data, AgeBand == "35+"))

summary(model_over35)
```

```

Call:
lm(formula = Cancer_Avoidance_Mean ~ Education_Level + Income + 
    MacArthur_Numeric + Political_Party, data = dplyr::filter(demo_data, 
    AgeBand == "35+"))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.09262 -0.59425 -0.05984  0.45032  2.45083 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       2.160298   0.044233  48.839  < 2e-16 ***
Education_Level                  -0.011298   0.008378  -1.349 0.177554    
Income                           -0.015754   0.006892  -2.286 0.022306 *  
```

```
MacArthur_Numeric                -0.029331   0.006519  -4.499 7.00e-06 ***
Political_PartyDemocrat          -0.225468   0.028414  -7.935 2.67e-15 ***
Political_PartyIndependent       -0.131147   0.030024  -4.368 1.28e-05 ***
Political_PartySomething else    -0.138479   0.040765  -3.397 0.000688 ***
Political_PartyPrefer not to say -0.093357   0.057189  -1.632 0.102659
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6599 on 4226 degrees of freedom
Multiple R-squared:  0.02841,   Adjusted R-squared:  0.0268
F-statistic: 17.65 on 7 and 4226 DF,  p-value: < 2.2e-16
```

4.4.1.5 Influencer Following vs cancer-avoidance score

```
influencer_cancer_linear <- lm(Cancer_Avoidance_Mean ~ Influencer_Following, data = media_data)
summary(influencer_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ Influencer_Following, data = media_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7519 -0.6157 -0.1157  0.3843  2.2593

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.75190    0.02045  85.664   <2e-16 ***
Influencer_FollowingUnsure -0.03774    0.05751  -0.656    0.512
Influencer_FollowingYes    -0.01121    0.02215  -0.506    0.613
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6428 on 6821 degrees of freedom
Multiple R-squared:  7.698e-05, Adjusted R-squared:  -0.0002162
F-statistic: 0.2626 on 2 and 6821 DF,  p-value: 0.7691
```

4.4.1.6 Stress Score vs cancer-avoidance score

```
stress_cancer_linear <- lm(Cancer_Avoidance_Mean ~ Stress_TotalScore, data = health_condition_
summary(stress_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ Stress_TotalScore, data = health_condition_data)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.8190 -0.6173 -0.1122  0.4629  2.2679 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.726935   0.014247 121.214   <2e-16 ***
Stress_TotalScore 0.005116   0.002852   1.794   0.0729 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6563 on 6658 degrees of freedom
Multiple R-squared:  0.000483,  Adjusted R-squared:  0.0003329 
F-statistic: 3.217 on 1 and 6658 DF,  p-value: 0.0729
```

4.4.1.7 Smoking vs cancer-avoidance score

```
smoking_cancer_linear <- lm(Cancer_Avoidance_Mean ~ Cigarette_Smoking_num, data = health_behavi
summary(smoking_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ Cigarette_Smoking_num, data = health_behavior_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.89432 -0.58335 -0.08335  0.41665  2.29165 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.708348   0.008468  201.75   <2e-16 ***
Cigarette_Smoking_num1 0.185977   0.020992    8.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6463 on 6955 degrees of freedom
Multiple R-squared:  0.01116,   Adjusted R-squared:  0.01102 
F-statistic: 78.49 on 1 and 6955 DF,  p-value: < 2.2e-16
```

4.4.1.8 Climate Change Belief vs cancer-avoidance score

```
voter_cancer_linear <- lm(Cancer_Avoidance_Mean ~ Climate_Change_Belief, data = other_data)
summary(voter_cancer_linear)
```

```
Call:
lm(formula = Cancer_Avoidance_Mean ~ Climate_Change_Belief, data = other_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9943 -0.5506 -0.0506  0.4494  2.3484 

Coefficients:

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
0.02146
Climate_Change_BeliefUncertain about the causes and extent of climate change.
0.06294
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
0.12172
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
0.27071

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
0.463
Climate_Change_BeliefUncertain about the causes and extent of climate change.
1.315
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
3.165
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
7.291

(Intercept)
16
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
13
```

```
(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6363 on 7315 degrees of freedom
Multiple R-squared:  0.02481,   Adjusted R-squared:  0.02415
F-statistic: 37.23 on 5 and 7315 DF,  p-value: < 2.2e-16
```

4.4.2 Mars

4.4.2.1 Demographic model

```
Call: earth(formula=Cancer_Avoidance_Mean~Ethnicity+Political_Party+Ge...),
            data=train)

                                              coefficients
(Intercept)                                     1.74763211
Political_PartyDemocrat                        -0.11313097
Job_ClassificationBlue Collar                   0.12744504
Job_ClassificationProfessional                 -0.09341490
Job_ClassificationUnemployed/Student/Parent    -0.07383926
RaceWhite                                       0.08027776
h(56-Age)                                      -0.00278826
h(Age-56)                                      -0.01447206
h(Income-2)                                    -0.01704592
h(9-MacArthur_Numeric)                          0.01881144
h(MacArthur_Numeric-9)                          0.27468651

Selected 11 of 14 terms, and 8 of 37 predictors
Termination condition: RSq changed by less than 0.001 at 14 terms
Importance: Political_PartyDemocrat, MacArthur_Numeric, RaceWhite, ...
Number of terms at each degree of interaction: 1 10 (additive model)
GCV 0.3877556    RSS 2167.197    GRSq 0.02086902    RSq 0.02781319


                                              nsubsets   gcv    rss
Political_PartyDemocrat                             10 100.0  100.0
MacArthur_Numeric                                    9  75.8   81.0
RaceWhite                                            8  63.7   71.0
Job_ClassificationProfessional                       7  52.7   61.9
Job_ClassificationUnemployed/Student/Parent          6  46.6   55.8
Job_ClassificationBlue Collar                        5  37.7   48.1
```

```
Age                                                    4  32.5   42.3
Income                                                 2  19.7   28.1
```

4.4.2.2 Cross validation demographic

```r
library(earth)

mars_model <- earth(
  Cancer_Avoidance_Mean ~ Ethnicity + Political_Party + Gender4 + Job_Classification +
    Education_Level + Age + Income + Race + MacArthur_Numeric,
  data = train,
  degree = 2,       # allow up to 2-way interactions
  nfold = 10,       # 10-fold CV
  keepxy = TRUE
)
summary(mars_model)
```

```
Call: earth(formula=Cancer_Avoidance_Mean~Ethnicity+Political_Party+Ge...),
            data=train, keepxy=TRUE, degree=2, nfold=10)

                                                                                      coefficients
(Intercept)                                                                             1.51853878
Political_PartyDemocrat                                                                -
0.12354920
RaceWhite                                                                               0.18407526
h(56-Age)                                                                               0.00899561
h(MacArthur_Numeric-9)                                                                  0.91738017
Political_PartyDemocrat * Job_ClassificationIT                                          0.12512063
Job_ClassificationFreelance/Gig * RaceWhite                                             0.12053036
EthnicityNo, not of Hispanic, Latino, or Spanish origin * h(MacArthur_Numeric-
9)  -0.95687380
Political_PartyDemocrat * h(2-Education_Level)                                          0.37967090
h(3-Gender4) * RaceWhite                                                                0.03914996
Job_ClassificationBlue Collar * h(9-MacArthur_Numeric)                                  0.03785557
Job_ClassificationProfessional * h(9-MacArthur_Numeric)                                -
0.02389413
h(56-Age) * RaceWhite                                                                  -
0.00475175
h(Age-56) * RaceWhite                                                                  -
0.02318944
h(Income-2) * RaceWhite                                                                -
0.03875043
h(56-Age) * h(5-Income)                                                                -
0.00128766
h(Age-26) * h(9-MacArthur_Numeric)                                                      0.00144843
```

```
Selected 17 of 30 terms, and 12 of 37 predictors
Termination condition: RSq changed by less than 0.001 at 30 terms
Importance: Political_PartyDemocrat, Age, MacArthur_Numeric, RaceWhite, ...
Number of terms at each degree of interaction: 1 4 12
GCV 0.3849115  RSS 2135.989  GRSq 0.02805093  RSq 0.04181287  CVRSq 0.007724538

Note: the cross-validation sd's below are standard deviations across folds

Cross validation:   nterms 20.20 sd 3.19    nvars 12.70 sd 1.77

     CVRSq    sd     MaxErr    sd
     0.008 0.019       2.47 0.106
```

```r
summary(mars_model) %>% .$coefficients %>% head(10)
```

```
                                                                               Cancer_Avoidance
(Intercept)                                                                              1.5185
Political_PartyDemocrat                                                                 -
0.123549198
h(MacArthur_Numeric-9)                                                                   0.9173
h(56-Age)                                                                                0.0089
h(Age-26)*h(9-MacArthur_Numeric)                                                         0.0014
RaceWhite                                                                                0.1840
Job_ClassificationProfessional*h(9-MacArthur_Numeric)                                   -
0.023894133
h(Income-2)*RaceWhite                                                                   -
0.038750430
EthnicityNo, not of Hispanic, Latino, or Spanish origin*h(MacArthur_Numeric-
9)          -0.956873797
Job_ClassificationBlue Collar*h(9-MacArthur_Numeric)                                     0.0378
```

4.4.2.3 Health condition Model

```r
# need to drop NA to get accuracy
health_condition_data <- selectdata %>%
  drop_na(
    Cancer_Avoidance_Mean, Stressful_Events_Recent, Current_Depression, Anxiety_Severity_num, I
    Health_Depression_Severity_num, Stress_TotalScore
  )

# split into training and testing sets
set.seed(123)
train_idx <- sample(seq_len(nrow(health_condition_data)), size = 0.7 * nrow(health_condition_da
train <- health_condition_data[train_idx, ]
test  <- health_condition_data[-train_idx, ]
```

```r
# Fit MARS model
mars_model_health_condition <- earth(
    Cancer_Avoidance_Mean ~ Stressful_Events_Recent + Current_Depression + Anxiety_Severity_nu
    PTSD5_Score +  Health_Depression_Severity_num + Stress_TotalScore,
  data = train
)

# Predict on test set
pred <- predict(mars_model_health_condition, newdata = test)

summary(mars_model_health_condition)
```

```
Call: earth(formula=Cancer_Avoidance_Mean~Stressful_Events_Recent+Curr...),
            data=train)

                          coefficients
(Intercept)                 1.76031430
h(Anxiety_Severity_num-3)   0.14374629
h(1-PTSD5_Score)           -0.08783213

Selected 3 of 5 terms, and 2 of 7 predictors
Termination condition: RSq changed by less than 0.001 at 5 terms
Importance: Anxiety_Severity_num, PTSD5_Score, ...
Number of terms at each degree of interaction: 1 2 (additive model)
GCV 0.4286923    RSS 1994.279    GRSq 0.007299509    RSq 0.00900262
```

```r
# Variable importance
evimp(mars_model_health_condition)
```

```
                     nsubsets   gcv    rss
Anxiety_Severity_num        2 100.0  100.0
PTSD5_Score                 1  61.6   63.4
```

4.4.2.4 Cross validation on Health Condition

```r
library(earth)

mars_model_health_condition <- earth(
    Cancer_Avoidance_Mean ~ Stressful_Events_Recent + Current_Depression + Anxiety_Severity_nu
    PTSD5_Score +  Health_Depression_Severity_num + Stress_TotalScore,
  data = train,
  degree = 2,       # allow up to 2-way interactions
  nk = 100,
  nfold = 10,       # 10-fold CV
  keepxy = TRUE
)
summary(mars_model_health_condition)
```

```
Call: earth(formula=Cancer_Avoidance_Mean~Stressful_Events_Recent+Curr...),
            data=train, keepxy=TRUE, degree=2, nfold=10, nk=100)

                                                    coefficients
(Intercept)                                           1.76029852
h(Anxiety_Severity_num-3)                             0.19063475
h(1-PTSD5_Score)                                     -0.08778390
h(Anxiety_Severity_num-3) * h(Stress_TotalScore-10)   0.15364202
h(Anxiety_Severity_num-3) * h(Stress_TotalScore-5)   -0.06007696

Selected 5 of 10 terms, and 3 of 7 predictors
Termination condition: RSq changed by less than 0.001 at 10 terms
Importance: Anxiety_Severity_num, PTSD5_Score, Stress_TotalScore, ...
Number of terms at each degree of interaction: 1 2 2
GCV 0.4286946  RSS 1989.154  GRSq 0.007294166  RSq 0.01154922  CVRSq -0.003463359

Note: the cross-validation sd's below are standard deviations across folds

Cross validation:   nterms 6.00 sd 1.63    nvars 3.10 sd 0.57

     CVRSq    sd     MaxErr     sd
    -0.003 0.012       2.35 0.0772
```

```r
pred <- predict(mars_model_health_condition, newdata = test)
rmse <- sqrt(mean((pred - test$Cancer_Avoidance_Mean)^2))
cor(pred, test$Cancer_Avoidance_Mean)
```

```
                           [,1]
Cancer_Avoidance_Mean 0.1452411
```

4.4.3 Logistic Regression

4.4.3.1 MacArthur Scale vs cancer-avoidance score

```r
MacArthur_cancer_logistic <- glm(Cancer_Avoiders01 ~ MacArthur_Numeric, data = demo_data, fami
summary(MacArthur_cancer_logistic)
```

```
Call:
glm(formula = Cancer_Avoiders01 ~ MacArthur_Numeric, family = binomial,
    data = demo_data)
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -2.60016    0.16150 -16.100  < 2e-16 ***
MacArthur_Numeric -0.10846    0.03165  -3.427 0.000611 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2785.6  on 8044  degrees of freedom
Residual deviance: 2773.8  on 8043  degrees of freedom
AIC: 2777.8

Number of Fisher Scoring iterations: 6
```

4.4.3.2 Facebook usage vs cancer-avoidance score

```
facebook_cancer_logistic <- glm(Cancer_Avoiders01 ~ Facebook_Usage_cat, data = media_data, fami
summary(facebook_cancer_logistic)
```

```
Call:
glm(formula = Cancer_Avoiders01 ~ Facebook_Usage_cat, family = binomial, 
    data = media_data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -3.1041     0.0870 -35.680   <2e-16 ***
Facebook_Usage_cat1   0.2358     0.1141   2.067   0.0388 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2654.9  on 6823  degrees of freedom
Residual deviance: 2650.6  on 6822  degrees of freedom
AIC: 2654.6

Number of Fisher Scoring iterations: 5
```

4.4.3.3 Stress score vs cancer-avoidance score

```
stress_cancer_logistic <- glm(Cancer_Avoiders01 ~ Stress_TotalScore, data = health_condition_da
summary(stress_cancer_logistic)
```

```
Call:
glm(formula = Cancer_Avoiders01 ~ Stress_TotalScore, family = binomial,
    data = health_condition_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.00919    0.09944 -30.260   <2e-16 ***
Stress_TotalScore  0.01475    0.01937   0.761    0.447
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2638.3  on 6659  degrees of freedom
Residual deviance: 2637.8  on 6658  degrees of freedom
AIC: 2641.8

Number of Fisher Scoring iterations: 5
```

4.4.3.4 Smoking vs cancer-avoidance score

```r
smoking_cancer_logistic <- glm(Cancer_Avoiders01 ~ Cigarette_Smoking_num, data = health_behavi
summary(smoking_cancer_logistic)
```

```
Call:
glm(formula = Cancer_Avoiders01 ~ Cigarette_Smoking_num, family = binomial,
    data = health_behavior_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.10042    0.06452 -48.050  < 2e-16 ***
Cigarette_Smoking_num1  0.62694    0.12839   4.883 1.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2709.8  on 6956  degrees of freedom
Residual deviance: 2688.1  on 6955  degrees of freedom
AIC: 2692.1

Number of Fisher Scoring iterations: 6
```

4.4.3.5 Climate Change vs cancer-avoidance score

```r
climate_cancer_logistic <- glm(Cancer_Avoiders01 ~ Climate_Change_Belief, data = other_data, fa
summary(climate_cancer_logistic)
```

```
Call:
glm(formula = Cancer_Avoiders01 ~ Climate_Change_Belief, family = binomial,
    data = other_data)

Coefficients:

(Intercept)
2.0239
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
0.4822
Climate_Change_BeliefUncertain about the causes and extent of climate change.
0.6842
Climate_Change_BeliefNo opinion on the matter.
0.1903
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
1.0698
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
1.2358

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu

(Intercept)
11.570
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
1.945
Climate_Change_BeliefUncertain about the causes and extent of climate change.
2.528
Climate_Change_BeliefNo opinion on the matter.
0.559
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
5.182
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
6.372

(Intercept)
16
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
```

```
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
07
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
10

(Intercept)
Climate_Change_BeliefSomewhat skeptical about the impact of human activities on climate change
Climate_Change_BeliefUncertain about the causes and extent of climate change.
Climate_Change_BeliefNo opinion on the matter.
Climate_Change_BeliefSomewhat believe climate change is occurring and is influenced by human ac
Climate_Change_BeliefStrongly believe climate change is occurring and is primarily caused by hu
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2781.4  on 7320  degrees of freedom
Residual deviance: 2731.3  on 7315  degrees of freedom
AIC: 2743.3

Number of Fisher Scoring iterations: 6
```

Chae, J., Lee, C.-J., & Kim, K. (2019). Prevalence, Predictors, and Psychosocial Mechanism of Cancer Information Avoidance: Findings from a National Survey of U.S. Adults. *Health Communication*, *35*(3), 322–330. https://doi.org/10.1080/10410236.2018.1563028

Dattilo, T. M., Roberts, C. M., Traino, K. A., Bakula, D. M., Fisher, R., Basile, N. L., Chaney, J. M., & Mullins, L. L. (2022). Illness stigma, health anxiety, illness intrusiveness, and depressive symptoms in adolescents and young adults: A path model. *Stigma and Health*, *7*(3), 311–317. https://doi.org/10.1037/sah0000390

Emanuel, A. S., Kiviniemi, M. T., Howell, J. L., Hay, J. L., Waters, E. A., Orom, H., & Shepperd, J. A. (2015). Avoiding cancer risk information. *Social Science & Medicine*, *147*, 113–120. https://doi.org/10.1016/j.socscimed.2015.10.058

Gigerenzer, G., & Garcia-Retamero, R. (2017). Cassandra's regret: The psychology of not wanting to know. *Psychological Review*, *124*(2), 179–196. https://doi.org/10.1037/rev0000055

Ho, E. H., Hagmann, D., & Loewenstein, G. (2021). Measuring Information Preferences. *Management Science*, *67*(1), 126–145. https://doi.org/10.1287/mnsc.2019.3543

Howell, J. L., Lipsey, N. P., & Shepperd, J. A. (2020). Health Information Avoidance. *The Wiley Encyclopedia of Health Psychology*, 279–286. https://doi.org/10.1002/9781119057840.ch77

Kelly, Christopher. A., & Sharot, T. (2021). Individual differences in information-seeking. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-27046-5

O'Brien, A. G., Meese, W. B., Taber, J. M., Johnson, A. E., Hinojosa, B. M., Burton, R., Ranjan, S., Rodarte, E. D., Coward, C., & Howell, J. L. (2024). Why do people avoid health risk information? A qualitative analysis. *SSM - Qualitative Research in Health*, *6*, 100461. https://doi.org/10.1016/j.ssmqr.2024.100461

Orom, H., Schofield, E., Kiviniemi, M. T., Waters, E. A., & Hay, J. L. (2020). Agency beliefs are associated with lower health information avoidance. *Health Education Journal*, *80*(3), 272–286.

https://doi.org/10.1177/0017896920967046

Song, S., Yao, X., & Wen, N. (2021). What motivates Chinese consumers to avoid information about the COVID-19 pandemic?: The perspective of the stimulus-organism-response model. *Information Processing & Management*, *58*(1), 102407. https://doi.org/10.1016/j.ipm.2020.102407

Soroya, S. H., & Faiola, A. (2023). Why did people avoid information during the COVID-19 pandemic? Understanding information sources' dynamics among Pakistani Z generation. *Library Hi Tech*, *41*(1), 229–247. https://doi.org/10.1108/lht-02-2022-0113

Sultana, T., Dhillon, G., & Oliveira, T. (2023). The effect of fear and situational motivation on online information avoidance: The case of COVID-19. *International Journal of Information Management*, *69*, 102596. https://doi.org/10.1016/j.ijinfomgt.2022.102596

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information Avoidance: Who, What, When, and Why. *Review of General Psychology*, *14*(4), 340–353. https://doi.org/10.1037/a0021288

Zhao, X., & Cai, X. (2009). The Role of Risk, Efficacy, and Anxiety in Smokers' Cancer Information Seeking. *Health Communication*, *24*(3), 259–269. https://doi.org/10.1080/10410230902805932