STA 663: Final Project

Hugh Ford

2025-05-07

This project aims to predict seasonal flu vaccine uptake using data from the National 2009 H1N1 Flu Survey (NHFS). Although seasonal flu vaccines are widely available in the U.S., not everyone chooses to get vaccinated. We first apply several imputation methods to handle missing data, ultimately opting for a complete case analysis. Next, we use logistic regression, classification trees, and random forest models to predict vaccination status based on 35 individual characteristics. Our comparison of these techniques indicates that the random forest consistently outperforms the others across key metrics. Feature importance analysis highlights perceived risk and belief in the vaccine's effectiveness as significant predictors of vaccination status. Based on these findings, we recommend that public health campaigns focus on these factors to improve seasonal flu vaccination uptake.

Table of contents

List of Figures

List of Tables

Section I: Introduction

This project aims to predict whether individuals received the seasonal flu vaccine using data from the National 2009 H1N1 Flu Survey (NHFS). The NHFS was conducted by the National Center for Immunization and Respiratory Diseases (NCIRD), the National Center for Health Statistics (NCHS), and the Centers for Disease Control and Prevention (CDC). It was a one-time survey "designed specifically to monitor vaccination during the 2009-2010 flu season in response to the 2009 H1N1 pandemic." (DrivenData 2020) The dataset was made available through DrivenData's "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines" competition. (DrivenData 2020)

We carry out a binary classification task to predict whether an individual received the seasonal flu vaccine or not. The NFHS dataset used in our analysis contains 26,707 observations across 37 variables, including participant ID, a binary indicator for seasonal flu vaccination, and 35 predictive features. These features describe the sampled individuals' health, behaviors, and opinions–for example, whether an individual has a chronic health condition, whether they frequently wash their hands, and their level of trust in vaccine safety. The dataset also includes demographic variables such as gender, race, age group, and employment status. Although the dataset originally included an outcome variable for H1N1 vaccination, we decided to exclude it from our analysis to focus exclusively on seasonal flu prediction.

We begin by cleaning the data and conducting an exploratory data analysis (EDA). We then implement and evaluate two prediction approaches: logistic regression and tree-based methods. Finally, we compare the performance of the techniques, discuss limitations of our analysis, and provide recommendations for future inquiry.

Section II: Data Cleaning and EDA

II.A: Data Cleaning

We begin by preparing the dataset for the prediction task. The data are provided in two separate CSV files, which we merge by matching participant IDs. As noted in the introduction, the dataset includes information on whether a participant received the H1N1 vaccine. We remove this variable, as our analysis focuses exclusively on predicting seasonal flu vaccination.

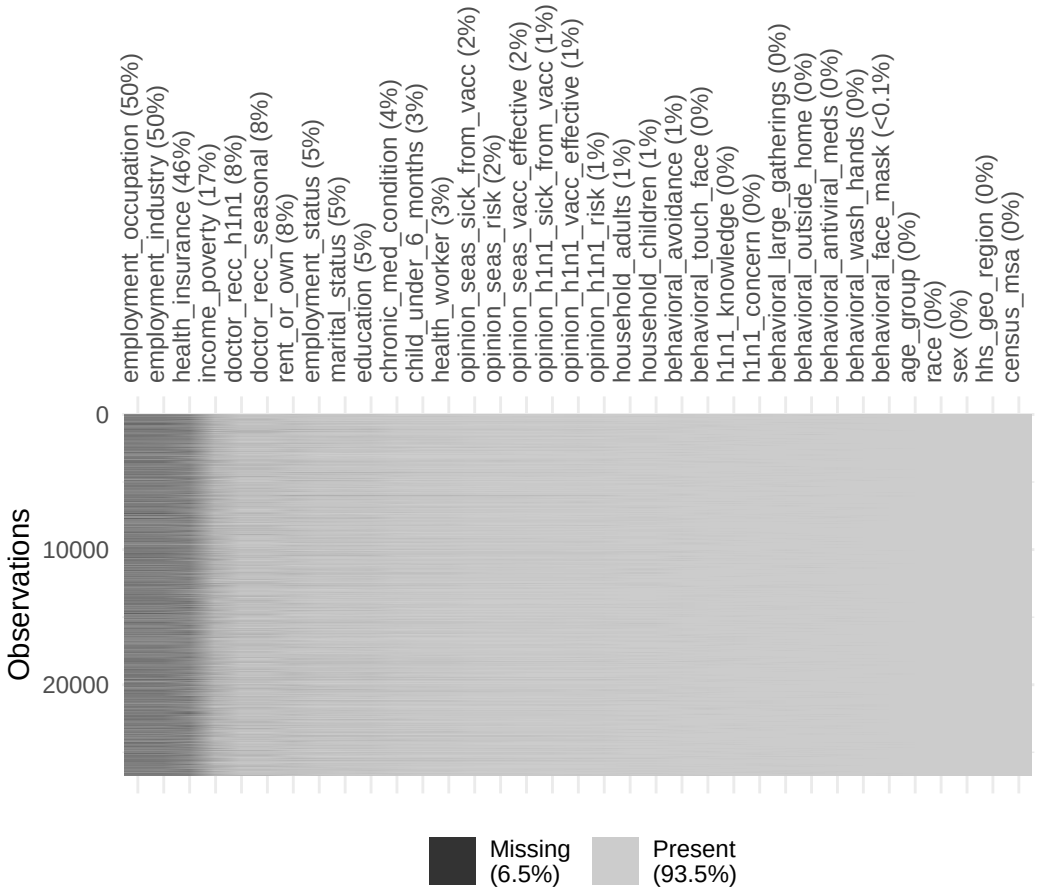

Figure 1: Missing Data by Feature (Pre-processing)

We then inspect the dataset for missingness. Overall, approximately 6.5% of the data are missing. Of the 35 features, 22 have missingness greater than or equal to 1%, as shown in **Figure 1**.

A key complication regarding missingness arises in the variables related to employment– specifically employment industry and occupation. As some participants are unemployed or not in the workforce, their industry and occupation data are missing due to their employment status, while for others, the data are missing for unknown reasons. We address this issue by distinguishing between these two types of missingness. We also hypothesize that the latter type may not be missing completely at random, and thus could be informative. Consequently, we retain these values in the dataset and designate them as "Unknown."

We suspect that missingness may also be informative for several variables that relate to more sensitive topics. The most concerning is health insurance status, which has the highest proportion of missing data (46%). We hypothesize that participants without insurance may have been reluctant to respond to this question on the survey. Similarly, participants may have

hesitated to answer questions about whether their doctor recommended the seasonal flu shot or H1N1 vaccine, especially if they do not regularly see a doctor or did not follow their medical advice. These two variables each have about 8% missing data. Lastly, some participants may have been unwilling to disclose whether they have a chronic health condition, although this variable had a lower rate of missingness (4%). Given the potential informativeness of this missing data, we again chose to label these values as "Unknown" rather than exclude the variables from the dataset or attempt to impute values.

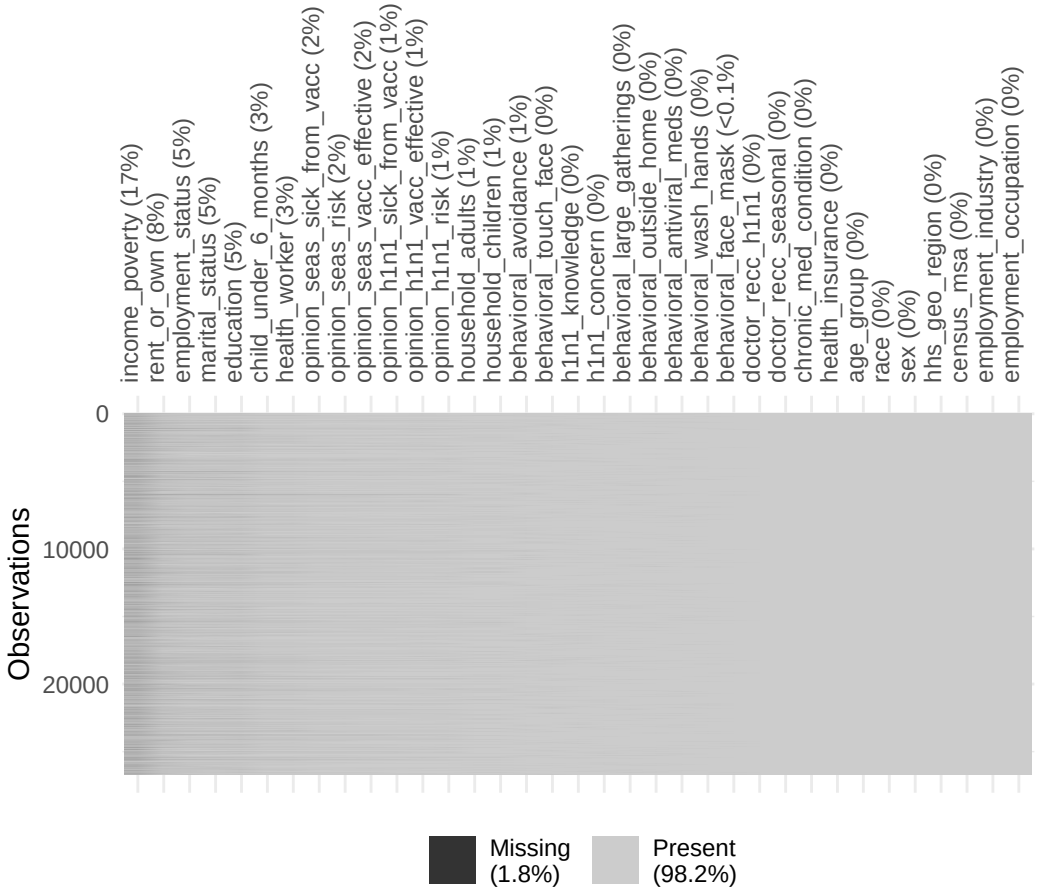

Figure 2: Missing Data by Feature - Iteration 2

After addressing this informative missing data, approximately 1.8% of the data remains missing and unaccounted for. As shown in **Figure 2**, this missingness is primarily concentrated in economic features such as income level (17% missing), home ownership status (8%), employment status (5%), and education level (5%).

We adopt two approaches to handling the remaining missingness. First, we repeat the process used earlier by assuming that the missing data in these four features is informative and labeling their missing values as "Unknown." Then, given the relatively low proportion of missingness

in the other features, we assume the remaining data are missing completely at random and perform a complete case analysis–removing any observation with missing values. This dataset serves as our baseline for evaluating the effectiveness of data imputation techniques.

For the second approach, we conduct Chi-squared tests of independence and find that the features with missingness are not independent of the other variables in the dataset at a 95% confidence level. As a result, we can leverage data from the other features to predict the missing values. We consider four imputation techniques: classification trees, random forest, chained equations, and k-Nearest Neighbors (k-NN). We detail and evaluate each approach in the following sections.

II.B.1: Classification Trees for Imputation

Introduction

The first imputation method we explore is classification trees. All of the features in the dataset are either categorical or discrete numeric with five or fewer unique values. We convert the discrete numeric features to ordered categorical features, allowing us to use classification-based techniques to impute missing values across all features.

Method

Classification trees are a type of decision tree used to predict categorical outcomes. They work by partitioning the data into clusters based on the values of certain features and assigning the most common outcome within each cluster as the predicted value.

A tree begins at a root node, which simply assigns the most frequent value of the target variable in the entire dataset. This prediction is refined by splitting the data into two subgroups using the feature that best predicts the outcomes. For our purposes, in order to determine the best splits, we use the Gini Index, which is a measure of the stability of the predictions in the tree. The Gini Index is a weighted average of the Gini Impurity score at each end node–or leaf. A low Gini Score indicates the node has a high concentration of one predicted outcome, and thus a results in a more stable prediction.

For example, suppose we are predicting the home ownership status of a participant based on the splitting rule of there being one or fewer children in the household. If 99% of participants in such households rented their home, while only 1% owned their home, this split would have a very low Gini Score, indicating a stable prediction. Conversely, suppose we are predicting home ownership status based on whether a participant has taken antiviral medication. If 51% of those who took the medication own their home and 49% do not, this split would yield a high Gini Score, indicating a poor, unstable prediction.

At each split, we select the feature and threshold that minimize the overall Gini Index of the resulting tree. The tree continues growing until a stopping condition is met–in our case, a minimum reduction in the Gini Index.

In using classification trees for imputation, we grow classification trees for each feature with missing values, using remaining features as predictors (except for the overall outcome–seasonal flu vaccination–and participant ID). We begin with the feature with the highest proportion of missing data: income level (17% missing). Using the remaining features, we build a classification tree to predict income levels for the observations with missing values for income. We then proceed to the feature with the next most missing data–home ownership status (8%)–this time excluding income level to avoid imputing based on imputed values. We repeat this process until all missing values are imputed.

Results

To evaluate the performance of the classification tree imputation method, we fit a logistic regression model to predict the outcome of interest: whether each participant received the seasonal flu vaccine. For simplicity, we assess the performance of the model using in-sample predictions–that is, predictions made on the same data used for imputation. Although logistic regression is involved in the chained equations method, it is not employed as a standalone imputation technique in our analysis. As a result, we consider it a relatively neutral predictive model for comparing our imputation methods. For a detailed discussion of logistic regression for classification, see **Section III**.

We assess the model's performance on the imputed data using several metrics: accuracy, classification error rate (CER), sensitivity, specificity, the geometric mean of sensitivity and specificity, and the F1 score. *Accuracy* is the proportion of correct predictions out of the total number of predictions, while *CER* is the proportion of incorrect ones. *Sensitivity* measures the proportion of correctly predicted vaccine recipients out of all participants who received the vaccine (i.e., true positive rate). Conversely, *specificity* is the proportion of correctly predicted unvaccinated participants out of all unvaccinated participants (i.e., true negative rate). The *geometric mean* and the *F1 score* measure the balance between sensitivity and specificity.

The performance metrics for the classification tree imputation are presented in **Table 1**. When trained on the dataset completed by classification tree imputation, the logistic regression model is 78.6% accurate in correctly predicting whether a participant received the seasonal flu vaccine. Among participants who actually received the vaccine, the model predicts correctly 75.0% of the time, while among participants who did not receive the vaccine, the model predicts correctly 81.7% of the time.

Table 1: Classification Tree Imputation

Measure	Result
Accuracy	0.7855

CER	0.2145
Sensitivity	0.7498
Specificity	0.8166
Geometric Mean of Sens. and Spec.	0.7825
F1	0.7757

Note:
Measures based on in-sample prediction using logistic regression

II.B.2: Random Forest Imputation

Introduction

The second method we investigated was random forest imputation. As an extension of the decision tree methodology, random forests can similarly be used to predict missing values for categorical variables and are thus suited to imputing the missing data in our dataset.

Method

A random forest consists of a set of decision trees, with the final prediction determined by the majority "vote" across all individual trees. Since all variables in our dataset are categorical, the decision trees in the forest are classification trees, which make predictions for missing values as described in **Section II.B.1**. However, unlike a single classification tree, each tree in a random forest is grown from a bootstrap sample. A bootstrap sample is a sample that is the same size as the original dataset, drawn with replacement from the original sample. As a result, some observations may appear more than once or not at all.

Additionally, at each split in a tree, rather than evaluating all features to determine a splitting rule, the random forest only considers a random subset of features. A common rule of thumb is to use the square root of the total number of features, rounded down to the nearest integer. In our case, out of thirty-five original features, we choose from five features at each split. This random selection introduces diversity among the trees and often improves the predictive performance of the random forest. As with single trees, the Gini Index is used to evaluate the best splits and determine the stopping rule.

For computational efficiency, we limit our forest to 100 trees and cap each tree at a maximum of five leaf nodes.

Results

We evaluate the performance of the random forest imputation using the same approach we used for the classification tree: fitting a logistic regression model to predict whether each participant received the seasonal flu vaccine. As before, we use in-sample predictions.

The results from the random forest imputation are displayed in **Table 2**. When trained on the dataset completed by random forest imputation, the logistic regression model is 78.5% accurate in correctly predicting whether a participant received the seasonal flu vaccine. Among participants who actually received the vaccine, the model predicts correctly 74.9% of the time, while among participants who did not receive the vaccine, the model predicts correctly 81.7% of the time.

Table 2: Random Forest Imputation

Measure	Result
Accuracy	0.7860
CER	0.2140
Sensitivity	0.7501
Specificity	0.8173
Geometric Mean of Sens. and Spec.	0.7830
F1	0.7762

Note:
Measures based on in-sample prediction using logistic regression

II.B.3: Chained Equations Imputation

Introduction

Our third imputation method uses chained equations. Chained equation imputation is a flexible technique that can handle categorical and numeric variables. As a result, it is suited to impute the missing data in our dataset.

Method

Chained equations imputation is an iterative process that preserves conditional relationships between features in the dataset. The procedure begins with the feature with the least missingness. In our case, this feature is the indicator of whether or not the participant has purchased a face mask. We impute values for the missing data in this feature using the observed values of all other features. The process then continues to the variable with the next least missingness. In this way, each subsequent imputation is conditioned on the previous imputations.

Once all features with missing data have been imputed, we return to the first variable and repeat the imputation procedure, updating the values for the data that was originally missing conditioned on previous imputations. This cycle continues until the imputations converge–that is, when changes between successive iterations are no longer significant. Due to computational limitations, we cap the maximum number of iterations at 50.

The technique is flexible because the imputation method used for each feature can be specified according to its variable type. In our implementation, we use logistic regression to impute missing values for binary categorical features, such as whether a participant has a chronic health condition. We use polytomous logistic regression for unordered categorical features with more than two levels, such as employment status (employed, unemployed, or not in the workforce). Lastly, for ordered categorical variables, such as income levels, we impute using a proportional odds model. Detailed discussions of polytomous logistic regression and proportional odds modeling are beyond the scope of this project (see Appendix for linked resources), while further information on logistic regression can be found in **Section III**.

Results

We evaluate the performance of the chained equations imputation method using the same technique and metrics as in the previous approaches. The results are summarized in **Table 3**. When trained on the dataset completed by chained equations imputation, the logistic regression model is 78.8% accurate in correctly predicting whether a participant received the seasonal flu vaccine. Among participants that actually received the vaccine, the model correctly predicts 75.3% of the time, while among participants that did not receive the vaccine, the model correctly predicts 81.8% of the time.

Table 3: Chained Equations Imputation (Monotone with 50 Iterations)

Measure	Result
Accuracy	0.7878
CER	0.2122
Sensitivity	0.7530
Specificity	0.8181
Geometric Mean of Sens. and Spec.	0.7849
F1	0.7784

Note:
Measures based on in-sample prediction using logistic regression.

II.B.4: k-Nearest Neighbor Imputation

Introduction

Lastly, we evaluated k-Nearest Neighbor (k-NN) imputation. k-NN is a non-parametric algorithm, meaning it makes no assumptions regarding the distribution of the data. It is flexible and can be adapted to both numeric and categorical data, making it suitable for imputing the missing values in our dataset.

Method

The idea behind k-NN imputation is to estimate a missing value by identifying the k most similar observations and using their values to make a prediction. For computational efficiency, we set $k = 5$. That is, for each missing value, we identify the five most similar observations and impute using the most frequent value among them.

To determine similarity between observations, we use Gower's distance, which is a weighted average of distances across variables. For binary variables–such as whether or not a participant purchased a face mask–the distance is 0 if the values match and 1 otherwise. For example, two individuals who both purchased a face mask would have a distance of 0 on this variable, while an individual who did not purchase a mask would have a distance of 1 from the other two. For unordered categorical features–such as occupation–we use the Dice distance. We link to a discussion of Dice distance in the Appendix. For ordered categorical variables–such as income level–the distance is calculated as the positive number of steps between the two levels, divided by the total possible steps. For example, if one participant earns below the poverty line and another earns under $75,000 but above the poverty line, they are one step apart. Since there are three income levels in the dataset, the total number of possible steps is two, so the distance between the observations in terms of income is $\frac{1}{2}$.

Once all distances are computed, the five observations with the smallest Gower's distance to the observation in question are selected, and their mode (i.e. most common value) is used to impute the missing value.

Results

We evaluate the performance of k-NN imputation on our dataset using the same technique and metrics as in previous methods. The results are displayed in **Table 4**. When trained on the dataset completed by 5-NN imputation, the logistic regression model is 78.7% accurate in correctly predicting whether a participant received the seasonal flu vaccine. Among participants that actually received the vaccine, the model correctly predicts 75.2% of the time, while among participants that did not receive the vaccine, the model correctly predicts 81.7% of the time.

Table 4: 5-Nearest Neighbor Imputation

Measure	Result
Accuracy	0.7868
CER	0.2132
Sensitivity	0.7518
Specificity	0.8173
Geometric Mean of Sens. and Spec.	0.7839
F1	0.7773

Note:
Measures based on in-sample prediction using logistic regression.

II.B.5: Comparison of Imputation Techniques

After testing the four imputation techniques, we compare their performances using the metrics discussed in **Section II.B.1**. We also evaluate these metrics on the complete case analysis (CCA) dataset and report the results in **Table 5**. A visual comparison across all methods, including CCA, is shown in **Figure 3**. We exclude the CER from the plot, as it distorts the scale and can be inferred from *Accuracy*.

From the figure, we observe that all five datasets perform similarly across most metrics. However, the logistic regression model achieves the best performance on the CCA dataset for every metric. Among the imputation methods, Chained Equations Imputation performs the best, followed closely by k-NN. Random Forest Imputation has a slight edge on the Classification Tree method, with the latter performing the worst across all metrics. It is possible that a larger forest could yield better results and be more comparable to the other imputation methods.

Given that CCA yields the best performance across the five metrics, we proceed with the CCA dataset for the prediction portion of the project. Due to the large overall sample size and prior treatment of informative missingness as discussed in **Section II.A**, we believe there is minimal risk of introducing bias in our results by excluding incomplete observations. In total, CCA removes 5,175 observations with missing data from the original 26,707, resulting in a final dataset of 21,532 observations across 37 variables.

Table 5: Complete Case Analysis

Measure	Result
Accuracy	0.7899
CER	0.2101
Sensitivity	0.7577
Specificity	0.8183
Geometric Mean of Sens. and Spec.	0.7874
F1	0.7814

Note:
Measures based on in-sample prediction using logistic regression.

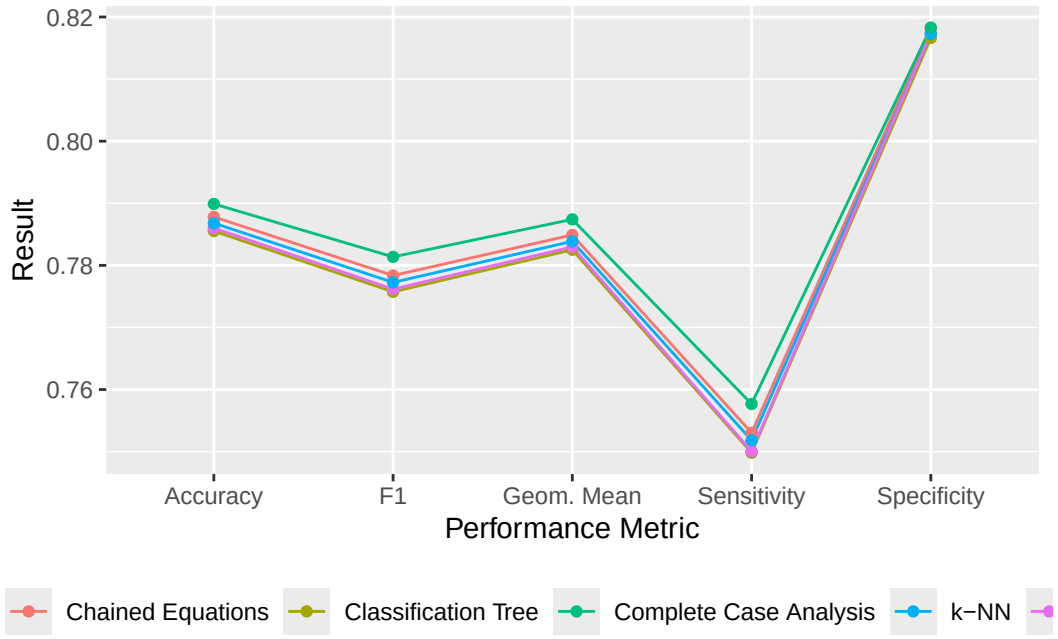

Figure 3: Comparison of Imputation Techniques across Performance Metrics

II.C: Exploratory Data Analysis

Using the CCA dataset, we begin with a brief exploratory data analysis (EDA) to better understand the distributions and relationships between key variables. We focus on six variables of particular interest: age group, race, education level, income level, health insurance status, and residential area (major city, minor city, or rural area, as defined by the Census Metropolitan Statistical Area guidelines). Within each of these variables, we look at the split between male and female participants.

We display the distributions using bar plots. In **Figure 4**, we observe that the most common age group in the sample is 65 and older, followed by 18-34, then 55-64, 45-54, and 35-44. We also display the proportion of male and female participants within each age group. Across all age groups, female participants outnumber male participants, which aligns with the overall distribution of sex in the sample.

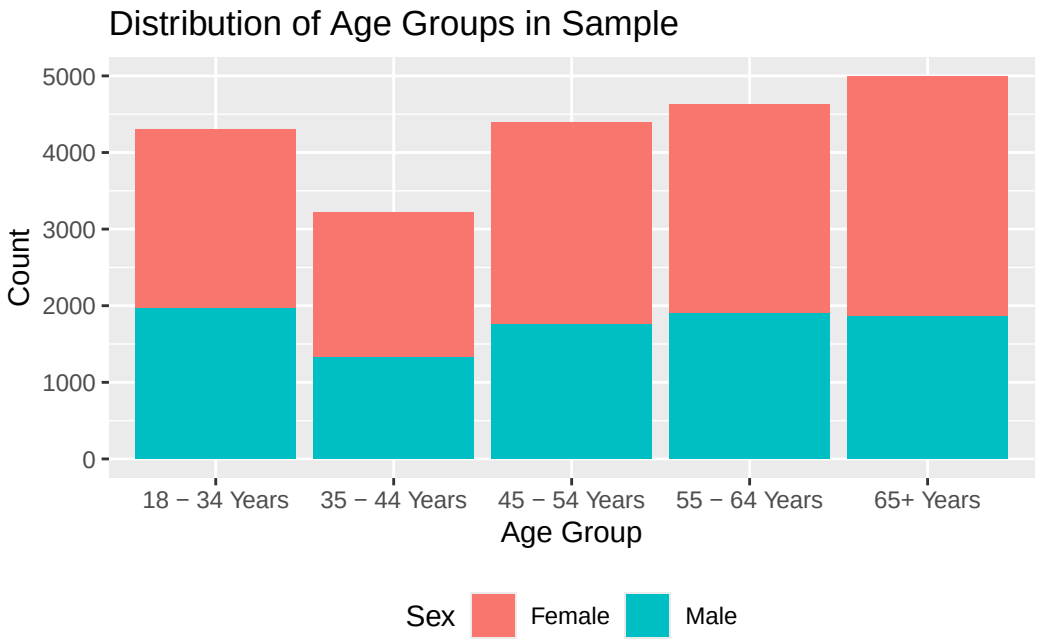

Figure 4: Distribution of Age Groups in Sample

In **Figure 5**, we see that the vast majority of participants identify as white, with nearly 12,000 individuals in this group. The next most common racial groups are Black, Hispanic, and Other or Multiracial, each with fewer than 1,500 participants. Given historical disparities in healthcare access and trust in the public health system across racial groups in the U.S., we anticipate that race may be a meaningful predictor of seasonal vaccine uptake. However, the substantial racial imbalance in the daataset may limit the model's ability to make accurate predictions for underrepresented racial groups.

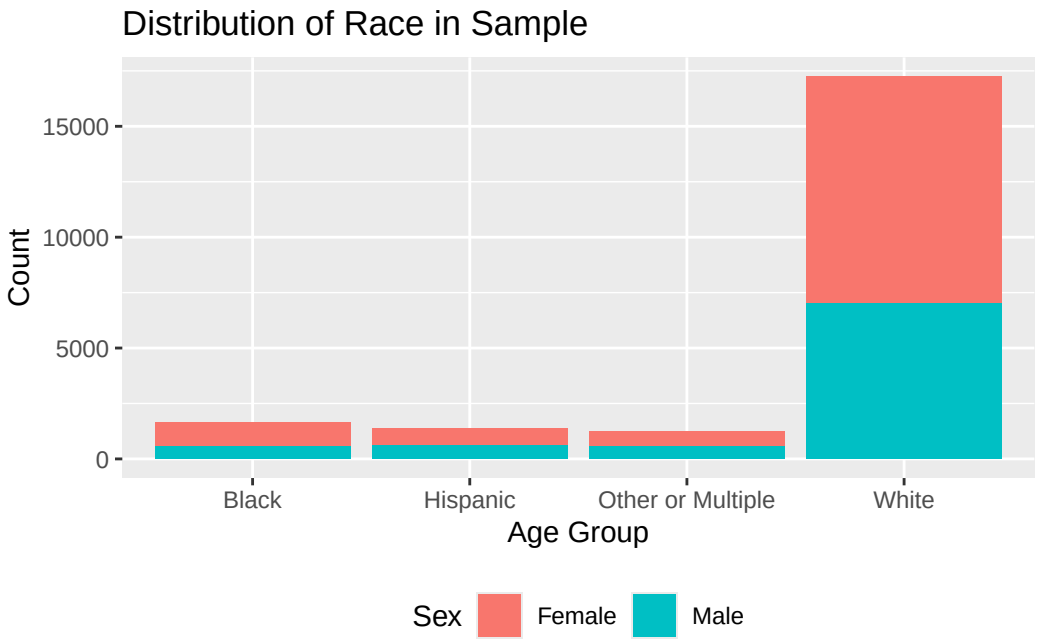

Figure 5: Distribution of Races in Sample

In **Figure 6**, we observe that most participants fall into three education categories–college graduates, those with some college education, and individuals with a high school diploma or equivalent–each with over 4,000 participants. In contrast, the group with fewer than 12 years of education is much smaller, with approximately 1,600 individuals. We hypothesize that education levels may be associated with seasonal vaccine uptake, as individuals with more education may be more aware of health risks.

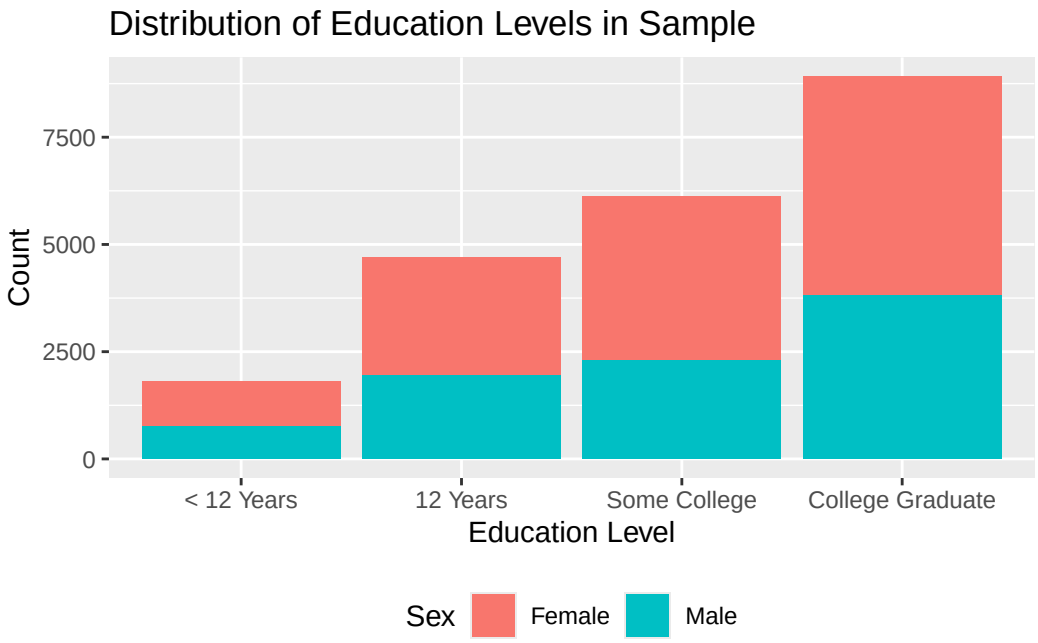

Figure 6: Distribution of Education Levels in Sample

Figure 7 shows the distribution of annual income and poverty status among participants in the sample. Income is categorized into three groups: those earning above $75,000, those earning at or below $75,000 but above the poverty line, and those below the poverty line. The middle-income group is the largest, comprising nearly 12,500 individuals. The high-income group follows with about 6,500 participants, while approximately 2,500 participants fall below the poverty line. We expect income to be a strong predictor of seasonal vaccine uptake, as higher income is often associated with better access to healthcare.

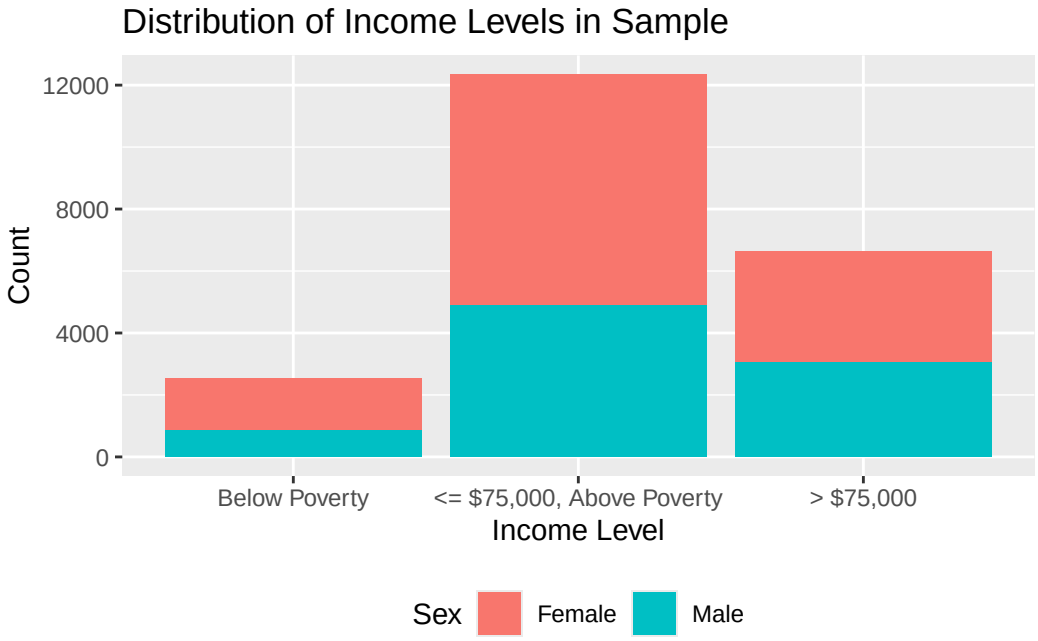

Figure 7: Distribution of Income Levels in Sample

Turning to health insurance status, **Figure 8** shows that the majority of participants–over 11,500–report having health insurance, while 1,500 do not have coverage. Because insurance often covers seasonal immunizations, we expect this feature to be a strong predictor of seasonal flu vaccine uptake. Nevertheless, insurance status is missing for more than 9,000. To retain these observations, we previously categorized their insurance status as "unknown," which may capture useful information. However, this high level of missingness may weaken the predictive power from this feature.

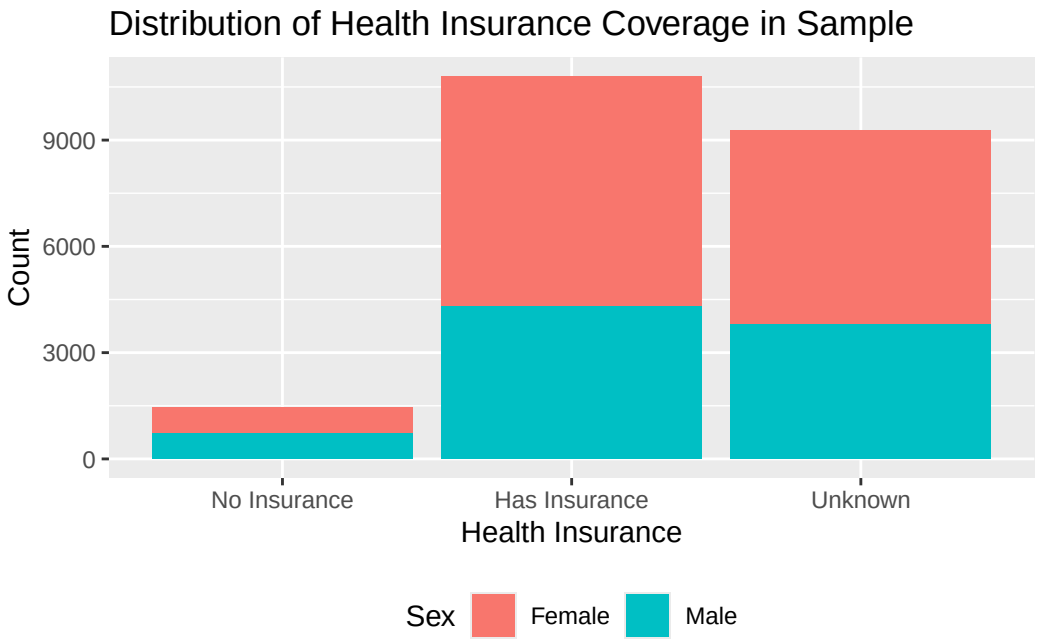

Figure 8: Distribution of Health Insurance Coverage in Sample

Lastly, **Figure 9** shows the distribution of participants across census-defined metropolitan areas. In our sample, over 9,000 participants live in non-principal metropolitan statistical areas (MSA). About 6,250 participants live in a principal city, while around 6,000 live in non-MSAs, i.e., less densely populated regions. Since access to healthcare is typically better in urban areas, we expect geographic area to also be an important predictor of seasonal vaccine uptake.

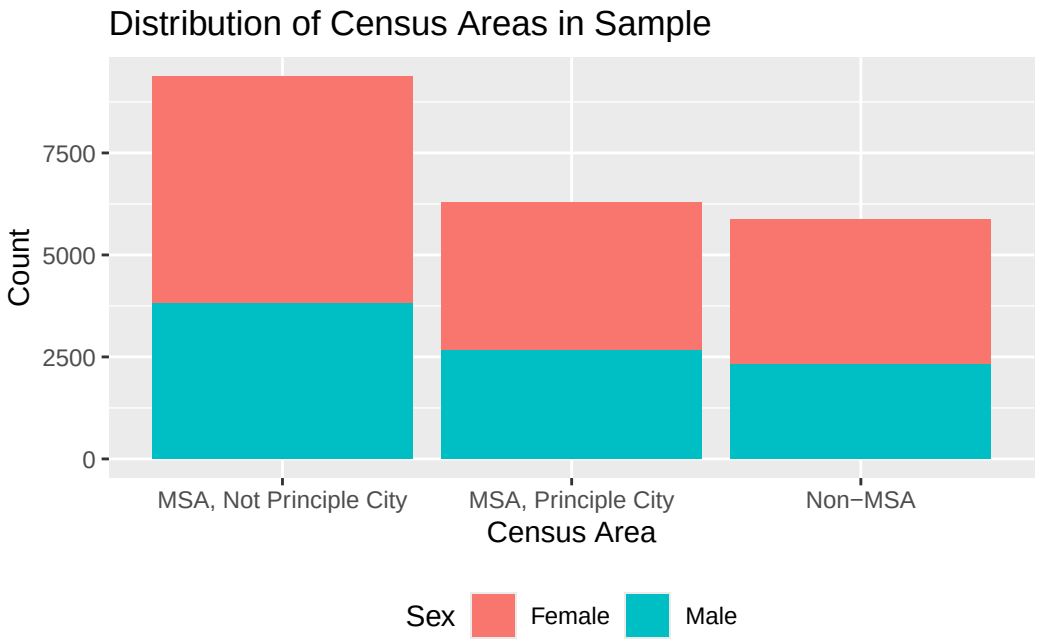

Figure 9: Distribution of Census Areas in Sample

To summarize the six features, we display the sample characteristics for these variables–along with sex–partitioned by their seasonal vaccination status in **Table 6**. We conduct Chi-squared tests of independence to assess the association between each feature and vaccine status. As shown in the table, all tests yield p-values below the significance level $\alpha = 0.05$, supporting our belief that these variables may be useful predictors of vaccination status. With that in mind, we turn to the prediction task, beginning with logistic regression-based methodology.

III. Method 1: Logistic Regression + Penalized Regression

III.A. Introduction

We begin the prediction portion of the project by investigating logistic regression and penalized logistic regression as methods for predicting seasonal flu vaccination status. Logistic regression is a generalized linear model, which predicts a binary response based on a set of features. As a result, it is well-suited to predict our outcome of interest: whether or not a participant received the seasonal flu vaccine.

Logistic regression is a parametric model and, therefore, makes certain assumptions about the relationship between the predictors and the outcome. In particular, it assumes a linear relationship between the features and the log-odds of the response. To evaluate whether this assumption holds for our data, we generate empirical log-odds plots, which plot the numeric

Table 6: Sample Characteristics by Seasonal Vaccine Status

Characteristic	**Not vaccinated** N = 11,442	**Vaccinated** N = 10,090	**Overall** N = 21,532	**p-value**
Age Group				<0.001
18 - 34 Years	3,052 (27%)	1,245 (12%)	4,297 (20%)	
35 - 44 Years	2,013 (18%)	1,202 (12%)	3,215 (15%)	
45 - 54 Years	2,576 (23%)	1,818 (18%)	4,394 (20%)	
55 - 64 Years	2,213 (19%)	2,421 (24%)	4,634 (22%)	
65+ Years	1,588 (14%)	3,404 (34%)	4,992 (23%)	
Education Level				<0.001
< 12 Years	1,082 (9.5%)	723 (7.2%)	1,805 (8.4%)	
12 Years	2,616 (23%)	2,077 (21%)	4,693 (22%)	
Some College	3,386 (30%)	2,733 (27%)	6,119 (28%)	
College Graduate	4,358 (38%)	4,557 (45%)	8,915 (41%)	
Income Level				<0.001
Below Poverty	1,638 (14%)	914 (9.1%)	2,552 (12%)	
<= $75,000, Above Poverty	6,466 (57%)	5,876 (58%)	12,342 (57%)	
> $75,000	3,338 (29%)	3,300 (33%)	6,638 (31%)	
Race				<0.001
Black	1,056 (9.2%)	577 (5.7%)	1,633 (7.6%)	
Hispanic	930 (8.1%)	477 (4.7%)	1,407 (6.5%)	
Other or Multiple	729 (6.4%)	518 (5.1%)	1,247 (5.8%)	
White	8,727 (76%)	8,518 (84%)	17,245 (80%)	
Insured Status				<0.001
No Insurance	1,112 (9.7%)	342 (3.4%)	1,454 (6.8%)	
Has Insurance	5,008 (44%)	5,786 (57%)	10,794 (50%)	
Unknown	5,322 (47%)	3,962 (39%)	9,284 (43%)	
Census MSA				0.012
MSA, Not Principle City	4,873 (43%)	4,500 (45%)	9,373 (44%)	
MSA, Principle City	3,409 (30%)	2,891 (29%)	6,300 (29%)	
Non-MSA	3,160 (28%)	2,699 (27%)	5,859 (27%)	
Sex				<0.001
Female	6,368 (56%)	6,368 (63%)	12,736 (59%)	
Male	5,074 (44%)	3,722 (37%)	8,796 (41%)	

[1] n (%)

[2] Pearson's Chi-squared test

features against the log-odds of vaccination. **Figure 10** shows an example of an empirical log-odds plot for participants' perceived risk of seasonal flu without a vaccine. The linear trend of the plot suggests that the linearity assumption is reasonable for this feature.

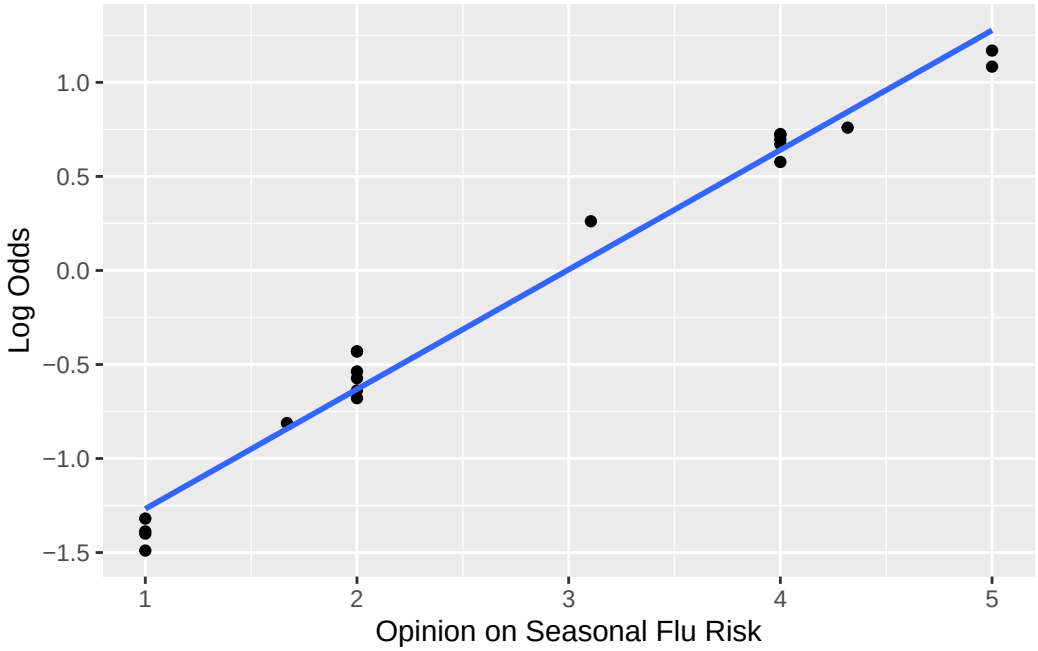

Figure 10: Empirical Log-odds plot of Perceived Seasonal Flu Risk

To improve upon the standard logistic regression model, we also investigate incorporating elastic net into the model. Elastic net combines two forms of penalized regression–ridge and lasso–which can help improve estimates in the presence of multicollinearity (i.e. when features are highly correlated with one another). Given the large number of features in our dataset and the potential for correlation between them, we believe it reasonable to apply elastic net regression.

III.B. Method

In logistic regression, the model predicts the logit (or log-odds) of a binary outcome by selecting coefficients that minimize deviance. The probability of the outcome is a function of the logit, so based on a probabilistic threshold (0.5 in our case), the model predicts the most likely outcome for an observation from the values of its features.

The elastic net model functions in much the same way. However, instead of minimizing only the deviance to select coefficients, it also minimizes a penalty term. This penalty term combines ridge and lasso penalties, which helps reduce the impact of multicollinearity among features.

III.C. Results

In fitting our elastic net model, we observe that the value of lambda that minimizes the error in our regression is zero. Since lambda is the constant coefficient of the penalization term, in this case, there is no penalty applied to the model. Without a penalty term, elastic net is equivalent to logistic regression, so we only evaluate the performance of the fitted logistic regression model.

To assess the predictive power of the model, we use 10-fold cross-validation. This method trains the model on $\frac{9}{10}$ of the dataset, evaluates it on the remaining $\frac{1}{10}$, and repeats this process for each 9-1 split. The most common predicted values across each fold are used as the final prediction.

We use the same performance metrics as we did with our imputation methods–accuracy, CER, sensitivity, specificity, their geometric mean, and the F1 score. Results for the logistic regression are displayed in **Table 7**. We observe that logistic regression has an accuracy of approximately 78.6%. The model performs slightly worse when predicting participants who actually received the seasonal flu shot, with a sensitivity of 75.4%. However, it performs slightly better when predicting participants who did not receive the flu shot, with a specificity of 81.4%.

Table 7: Logistic Regression Prediction Performance

Measure	Result
Accuracy	0.7861
CER	0.2139
Sensitivity	0.7543
Specificity	0.8142
Geometric Mean of Sens. and Spec.	0.7837
F1	0.7776

IV. Method 2: Classification Trees and Random Forest

IV.A. Introduction

We next examine two tree-based methods for prediction–classification trees and random forests. As discussed in **Sections II.B.1** and **II.B.2**, these approaches are well-suited for predicting categorical outcomes. Since seasonal flu vaccination status is a binary categorical variable, both methods are appropriate for our task.

IV.B. Method

The methodologies for classification trees and random forests are outlined in detail in **Sections II.B.1** and **II.B.2**, respectively. However, it is important to note that, unlike logistic regression, these methodologies are non-parametric, meaning they do not assume an underlying shape to the relationship between the predictors and the response.

In **Figure 11**, we display the fitted classification tree. The root node indicates that approximately 53 percent of the sample did not get the vaccine, while about 47 percent did. The first split is based on the participant's perceived risk of contracting the seasonal flu without a vaccine–rated on a scale from 1 (least risk) to 5 (most risk). The tree splits at a threshold of less than 2.5, meaning that participants who rated their risk as 1 or 2 were more likely not to get vaccinated, while those who rated their risk as 3 or higher were more likely to get the vaccine. For the low-risk participants, the next split occurs based on whether their doctor recommended the seasonal flu vaccine. If no recommendation was given, participants were less likely to have been vaccinated. Conversely, those who did receive a recommendation from their doctor were more likely to get the vaccine.

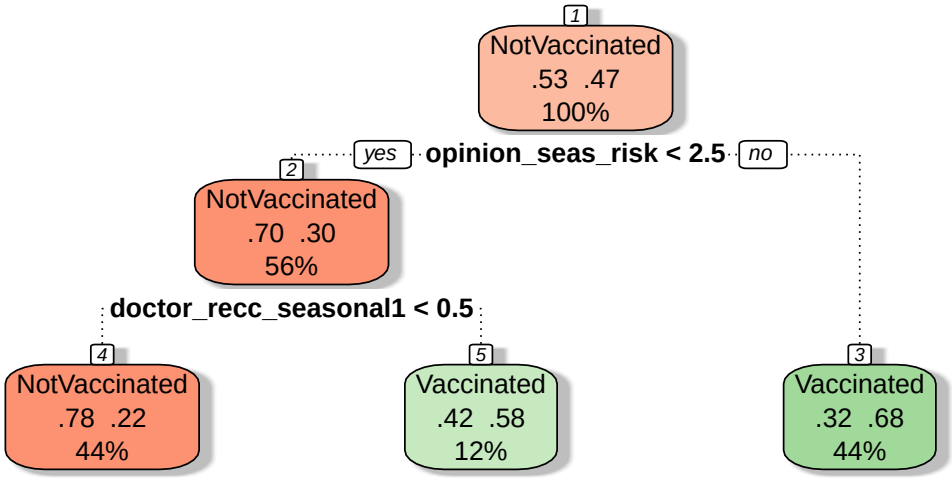

Figure 11: Classification Tree for Predicting Seasonal Flu Vaccine

Turning to the random forest model, we present the ten most important and ten least important features ranked by the mean decrease in Gini Index in **Figure 12** and **Figure 13**, respectively. Similar to the classification tree, the most important feature is the participant's perceived risk of illness without the seasonal flu vaccine. This feature is followed closely by the participant's belief in the effectiveness of the vaccine. These results are unsurprising, as we would expect both perceptions to figure strongly into a person's decision to get vaccinated. In contrast, the least important features are whether a participant has taken anti-viral medications and whether they have purchased a face mask–behaviors that, at least in the context of 2009, are less directly related to vaccination decisions.

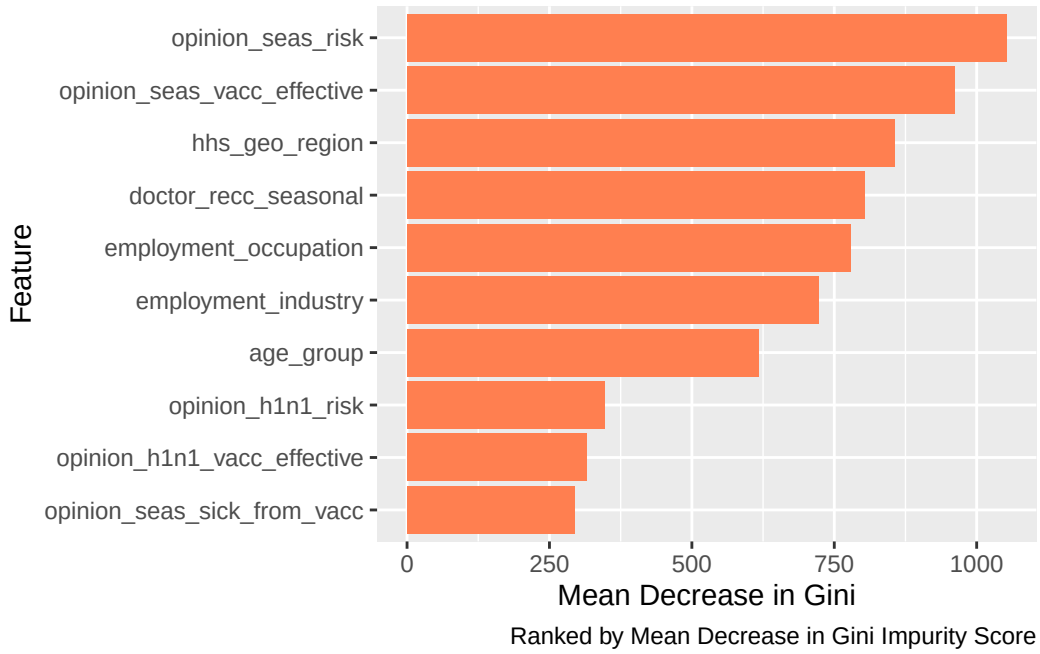

Figure 12: 10 Most Important Features in Random Forest

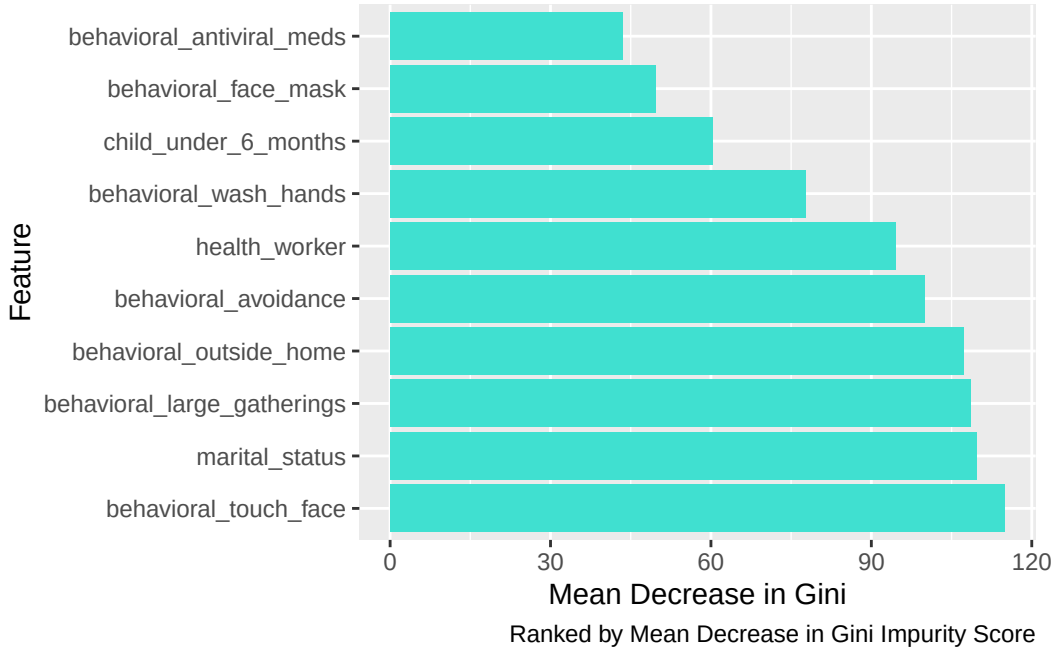

Figure 13: 10 Least Important Features in Random Forest

IV.C Results

As with the logistic regression model, we use 10-fold cross-validation to assess the performance of the classification tree, applying the same performance metrics as before. These results are summarized in **Table 8**.

For the random forest, however, we take advantage of out-of-bag (OOB) data to evaluate the technique without needing 10-fold cross validation. OOB data refers to the subset of observations left out of each tree during the bootstrapping process. By aggregating the most common predictions made on OOB data across trees, we generate final predictions and compute performance metrics, which are displayed in **Table 9**.

The classification tree achieves an overall accuracy of approximately 68.3%. In comparison, the random forest has a significantly higher predictive accuracy of 78.6%, which is expected given the ensembling involved in the random forest, discussed in **Section II.B.2**. The classification tree has a sensitivity of about 60.0%, whereas the random forest correctly predicts vaccination for 76.1% of vaccinated participants. While the classification tree performs reasonably well in terms of specificity (75.6%), the random forest outperforms it again, correctly identifying unvaccinated participants 81.0% of the time.

Table 8: Classification Tree Prediction Performance

Measure	Result
Accuracy	0.6829
CER	0.3171
Sensitivity	0.6002
Specificity	0.7559
Geometric Mean of Sens. and Spec.	0.6736
F1	0.6509

Table 9: Random Forest Prediction Performance

Measure	Result
Accuracy	0.7854
CER	0.2146
Sensitivity	0.7605
Specificity	0.8091
Geometric Mean of Sens. and Spec.	0.7845
F1	0.7795

V. Concluisions

The primary goal of this project was to predict an individual's seasonal flu vaccination status based on demographic information and health-related characteristics. We used the National 2009 H1N1 Flu Survey data, which contains 26,707 observations across 37 variables. Our analysis began by addressing missing data in the dataset. Notably, over half of all observations (14,913) contained missing values.

We evaluated four imputation techniques to handle this missingness: classification trees, random forests, chained equations, and k-nearest neighbors (k-NN). Each imputed dataset was assessed using a neutral logistic regression model and in-sample prediction. We compared these results to a complete case analysis (CCA) and found that the CCA version performed the best across all five performance metrics.

Using the CCA dataset, we then applied three statistical learning methods–logistic regression, classification trees, and random forests–to predict vaccination status. Of these techniques, the random forest performed the best, achieving the highest accuracy (78.6%), sensitivity (76.1%), and specificity (81.0%). In practice, we would recommend using the random forest due to its strong predictive performance, flexibility with a variety of data types, and lack of reliance on parametric assumptions. The main drawback of the technique is that, unlike logistic regression or a classification tree, it cannot be easily interpreted in terms of associations between the outcome and the features. However, by using feature importance analysis, we can attempt to identify key predictive variables.

Our feature importance analysis of the random forest indicated that the most important predictors of vaccination were a participant's perceived risk of illness without the seasonal flu vaccine and their belief in the vaccine's effectiveness. In contrast, behavioral features like antiviral medication usage and mask purchasing were far less predictive.

Several limitations may affect the validity of our findings. Most notably, due to computational constraints, we restricted the complexity of our imputation techniques. For instance, in the random forest–which only performed slightly better than a single classification tree–we capped the number of trees at 100 and limited the number of terminal nodes per tree to 5. Similarly, for the chained equations imputation, we only ran 50 iterations and performed a single imputation. Increased processing power would have allowed us to run more iterations and aggregate multiple imputations. Lastly, we limited k-NN imputation to $k = 5$. With more computing power, we could test higher values of k to achieve the best performance from the imputation.

We also originally intended to use a k-NN algorithm for prediction but were again limited by computing resources. Future work could incorporate k-NN as well as other computationally expensive prediction techniques.

Nevertheless, while none of our prediction techniques achieved perfect classification, the results suggest that personal beliefs about flu risk and vaccine efficacy are strong predictors

of vaccination behavior. These insights could be used to direct public health messaging and encourage seasonal vaccine uptake. Future analyses could apply these prediction techniques to other vaccines–such as the H1N1 vaccine included in the dataset–to assess their generalizability. Additionally, the models could be applied to more recent datasets–especially in light of the rise in vaccine skepticism following the COVID-19 pandemic–to determine whether they remain effective in the current context.

References

DrivenData. 2020. "Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines."

Appendix

Polytomous logistic regression

https://online.stat.psu.edu/stat504/lesson/8/8.1

Proportional odds modeling

https://peopleanalytics-regression-book.org/ord-reg.html

Dice distance:

https://distancia.readthedocs.io/en/latest/Dice.html