

Math 4211 Final Report

Lavinia Xu

May 8, 2024

1 Confidentiality Agreement

I understood and agreed to that the data used in this report must be kept confidential and must not be shared or distributed outside this class. The statistical analysis, the result and the report were agreed to be only used in this class and will not be posted or published without the authorization of Prof. Jimin Ding.

2 Introduction

In September 2022, the Pew Research Center conducted a comprehensive survey exploring U.S. public opinions on a range of topics including scientists, religion, and COVID policies. This survey was part of the American Trends Panel (ATP), a nationally representative online platform that draws from a pool of over 10,000 adults randomly selected from across the United States. More information about the ATP and its methodologies can be found at [The American Trends Panel](#).

One main focus of the survey was the uptake of COVID vaccinations, captured in the variable `COVID_vaccinated`. The survey is composed of nearly 100 questions, and plenty of factors that can potentially be related to vaccination uptake are included, such as age, race, education, income, etc. In this report, we want to delve into the survey data to further investigate and address potential barriers to COVID vaccination uptake.

The variable names of the survey data is designed in a way that computers can read them, so they might be less intuitive and informative to human. Thus, I will briefly describe some variables that will be frequently mentioned in the rest of the report.

- `F_METRO`: Metropolitan area indicator
- `F_REGION`: Census region
- `F_AGECAT`: Age category
- `F_PARTYSUMIDEO_FINAL`: Combining ideology and party identification
- `DEVICE_TYPE_W114`: The device used in the latest access of the survey link
- `CONF_a_W114`: Confidence in elected officials
- `CONF_f_W114`: Confidence in medical scientists
- `CONF_g_W114`: Confidence in scientists
- `CVD_PROBS_NOCOMPLY_W114`: Opinions on the relationship between ordinary Americans failing to follow public health and the problems the country has faced dealing with the coronavirus outbreak recommendations
- `MPX_VAX_W114`: Opinions on vaccines to prevent monkeypox
- `VAXBOOST3_W114`: Opinions on a new booster shot designed for recent variants of COVID-19

Other variables not listed above will be briefly explained when they first appear in the report.

3 Data Analysis

Before we delve into all features and build a model with all features related to COVID-19 vaccination uptake, it is crucial to first establish a foundational understanding of the relationships between individual variables and vaccination status. To facilitate this initial exploration, we have employed a contingency table to succinctly summarize the interplay between trust in elected officials and the uptake of COVID-19 vaccinations.

	High Trust		Low Trust			High Trust		Low Trust	
	Vaccinated	Unvaccinated	Vaccinated	Unvaccinated		Vaccinated	Unvaccinated	Vaccinated	Unvaccinated
##	2558	6079	##	89.06685	81.01013	##	10.93315	18.98987	
##	314	1425	##			##			

Figure 1: Contingency Table of Trust in Elected Officials and Vaccination Uptake

From the table, we can see that people with high trust in elected officials are more likely to take COVID vaccinations: around 89% of them took vaccinations. On the contrary, only 81% people with low trust were vaccinated. We further conducted a test on the association between vaccination and trust, and our hypotheses are:

- H_0 : There is no association between COVID vaccination status and trust in elected officials.
- H_a : There is an association between COVID vaccination status and trust in elected officials.

In this case we chose to do a Chi-square test of independence. This tests if $X = E(X)$. The expected value of each portion can be represented by the following equation:

$$\text{Expected frequency} = \frac{\text{Grand total}}{\text{Row total} \times \text{Column total}}$$

From the test, we have a p-value of 1.112×10^{-22} , which is far below 0.05, meaning we have sufficient evidence to reject the null hypothesis that there is no association between COVID vaccination status and trust in elected officials. In practicality, this means that there is likely an association between COVID vaccination status and trust in elected officials. Keeping the association in mind, we want to redo the test by stratifying by age group since it is natural to assume that age groups might play a key role in both vaccination status and trust in the legal authorities. We again built contingency tables categorized by age (on top of Page 3). From the table, we might surprisingly find out that as high as 17% of people under 49 did not get vaccinations even with their high trust in elected officials, while as people getting older, they are more likely to be vaccinated even if they have low trust in the officials. However, we can still expect an association between trust and vaccination uptake for all age groups based on the percentage. We re-ran the Chi-square test for each age group, and below is the table that summarizes p-values for all the age groups.

	Age_Group	P_Value
##	18-29	18-29 5.827860e-01
##	30-49	30-49 2.435949e-05
##	50-64	50-64 3.565592e-17
##	65+	65+ 1.284771e-06
##	Refused	Refused 1.000000e+00

Figure 2: Summary of p-values for Age Groups

As we can see from the results, we have sufficient evidence to reject the null hypothesis for age groups 30-49, 50-64, and 65+. However, we do not have sufficient evidence to reject the null hypothesis for the age group of 18-29 and those who refused to give age information. This means that for individuals 30+, there is likely an association between trust and vaccination. However, for those in the range of 18-29 there is not sufficient evidence to make the same conclusion. One thing to note is that insignificant p-values for young adults might be resulted from small sample size. Only 857 young adults participated in the survey, while all other age groups have at least 3000 people involved.

Age Group: 18-29	High Trust Low Trust	Vaccinated 174 501	Age Group: 18-29	High Trust Low Trust	Vaccinated 82.07547 80.03195
		Unvaccinated 38 125			Unvaccinated 17.92453 19.96805
Age Group: 30-49	High Trust Low Trust	Vaccinated 779 1855	Age Group: 30-49	High Trust Low Trust	Vaccinated 83.22650 76.46331
		Unvaccinated 157 571			Unvaccinated 16.77350 23.53669
Age Group: 50-64	High Trust Low Trust	Vaccinated 740 1696	Age Group: 50-64	High Trust Low Trust	Vaccinated 91.811414 78.373383
		Unvaccinated 66 468			Unvaccinated 8.188586 21.626617
Age Group: 65+	High Trust Low Trust	Vaccinated 859 2007	Age Group: 65+	High Trust Low Trust	Vaccinated 94.291987 88.570168
		Unvaccinated 52 259			Unvaccinated 5.708013 11.429832
Age Group: Refused	High Trust Low Trust	Vaccinated 6 20	Age Group: Refused	High Trust Low Trust	Vaccinated 85.714286 90.909091
		Unvaccinated 1 2			Unvaccinated 14.285714 9.090909

Figure 3: Contingency Table of Trust and Vaccination Uptake Stratified by Age Groups

4 Data Visualization

Completing our initial data analysis, we now want to investigate more on intuitively important features in the dataset in order for future works. We can visualize some features (and their distributions) of the data, as well as their relationship with vaccination uptake.

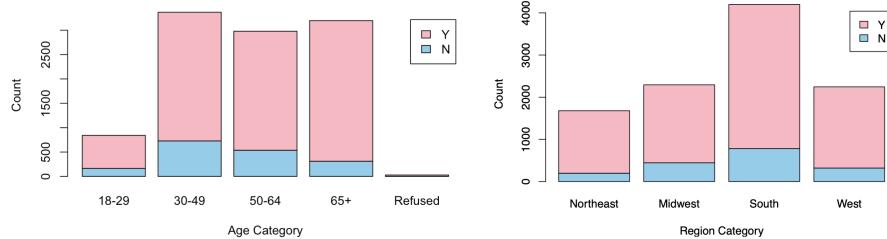


Figure 4: Relationships between Age / Census Region and Vaccinations

For the stacked bar plot on the left, this illustration reflects the contingency table above in a more straightforward way. I chose to illustrate the age variable since the findings above are a bit surprising to me and would like to further investigate this variable. On the other hand, the graph representing the relationship between census region and vaccination shows that people from South answered the survey the most. Comparing to the other regions, people from Midwest have higher percentage of people not taken vaccinations.

5 Data Preprocessing

From the age visualization, we notice a very small section of people answered "Refused", and compared to the sample size of other age groups, the size of people who refused to answer is too small to be considered as an independent category. From the summary of the survey data, we can see for each variable, there are values of 98 and 99, both representing missing values. Thus, before we head to the model setup, we want to do some data preprocessing to deal with missing values in the dataset.

Before the imputation, we first removed QKEY (representing unique ID) and time related features that indicate how long the interview took. Another important thing worth mentioning is that we combined the variable about confidence in medical scientists with the variable about confidence in scientists. There were two different sets of survey, one asking people's opinion on medical scientists while the other one asking scientists. Since the two questions are essentially designed for the same purpose, we combined them for better regression.

In the preprocessing phase of our analysis, we chose to impute missing data using the k-Nearest Neighbors (kNN) technique. We selected this imputer over others because it can effectively handle categorical data by finding the nearest neighbors based on similar data points. By scaling the dataframe columns to a [0, 1] range, we accommodate the distance-based mechanics of kNN. Given that the VIM library does not inherently support parallelization, we used a kNN imputer where we manually parallelized over the CPU cores. This setup allowed us to efficiently apply the kNN algorithm with 3 neighbors to the scaled data and subsequently restore the original scale using the maximum values from the unaltered dataset.

However, it is crucial to acknowledge a limitation in our approach: the use of Euclidean distance in the kNN algorithm. While Euclidean distance is widely used for its simplicity and effectiveness in many scenarios, it can be less ideal for categorical data because it treats categories as equidistant, potentially misrepresenting the true relationships between different categories. This aspect of the methodology could potentially affect the accuracy of our imputation. Future improvements might include exploring alternative distance metrics that better capture the nuances of categorical data.

6 Model Choice

Now, after we have imputed all missing values, we can finally start to build our model. We chose logistic regression model, which is specifically designed for binary classification tasks. It models the probability that a given input belongs to a particular category – in our case, vaccinated or not vaccinated. Its high interpretability is also a reason why we chose logistic regression model. The coefficients of the model can be directly understood in terms of odds ratio, giving clear insights into how each predictor affects the likelihood of a particular outcome.

As we mentioned above, we have nearly 100 features in our original dataset, and it is both computationally costly and less informative if we use all the features to predict non-vaccination. To address this problem, we performed a Chi-square test for significance on every feature in the dataset and selected only 20 features with highest Chi-square statistic:

```
## [1] "COVID_VAXDMOD_W114"      "WEIGHT_W114"
## [3] "WEIGHT_W84_W114"          "WEIGHT_W64_W66_W83_W114"
## [5] "VAXBOOST3_W114_imp"       "COVID_BOOST_W114"
## [7] "DEVICE_TYPE_W114_imp"     "LANG_W114_imp"
## [9] "FORM_W114_imp"            "F_METRO_imp"
## [11] "F_CREGION_imp"           "F_CDIVISION_imp"
## [13] "F_PARTYSUM_FINAL_imp"    "F_PARTYSUMVIDEO_FINAL_imp"
## [15] "WEIGHT_W114_imp"          "trust_imp"
## [17] "MPX_VAX_W114"             "CVD_PROBS_NOCOMPLY_W114"
## [19] "FAUCI_W114"               "CONF_f_W114"
```

Figure 5: Features with Top 20 Chi-square Statistic

Except for booster and weight related variables, which are intuitively inappropriate in predicting non-vaccination, I will briefly discuss why each feature is not selected into our model:

- **LANG_W114:** It has very imbalanced classes with 10195 people answering in English and 393 people answering in Spanish.
- **FORM_W114:** It only indicates which form people answered.
- **F_CDIVISION:** This feature representing census division is just a more detailed version of census region variable.
- **F_PARTYSUM_FINAL:** This feature representing party summary is a more general version of ideology and party identification variable. We chose the other one since it is more comprehensive.
- **FAUCI_W114:** This variable indicates people's opinion on a medical advisor, so it can potentially have very high collinearity with the variable of confidence in medical scientists.

Based on both Chi-square statistic and hand selection, we narrowed down to eight features, but we want to manually add the age variable even if it might not be statistically significant. Thus, our baseline logistic regression model has nine independent variables, and we set the response variable to non-vaccination. Below is a summary of our baseline model:

```

## 
## Call:
## glm(formula = I(COVID_vaccinated == "N") ~ factor(trust) + factor(DEVICE_TYPE_W114) +
##       factor(F_METRO) + factor(F_CREGION) + factor(F_PARTYSUMIDEO_FINAL) +
##       factor(MPX_VAX_W114) + factor(CVD_PROBS_NOCOMPLY_W114) +
##       factor(CONF_f_W114) + factor(F_AGECAT), family = binomial,
##       data = X_train)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -4.30923   0.31303 -13.766 < 2e-16 ***
## factor(trust)Y                0.30344   0.09173   3.308 0.000940 ***
## factor(DEVICE_TYPE_W114)2        0.40204   0.08294   4.847 1.25e-06 ***
## factor(DEVICE_TYPE_W114)3        0.69100   0.20845   3.315 0.000917 ***
## factor(F_METRO)2                  0.51679   0.09787   5.280 1.29e-07 ***
## factor(F_CREGION)2                  0.37869   0.12421   3.049 0.002299 **
## factor(F_CREGION)3                  0.49071   0.11310   4.339 1.43e-05 ***
## factor(F_CREGION)4                  0.28392   0.12772   2.223 0.026220 *
## factor(F_PARTYSUMIDEO_FINAL)2      -0.09281   0.10057  -0.923 0.356116
## factor(F_PARTYSUMIDEO_FINAL)3      -0.16288   0.11262  -1.446 0.148091
## factor(F_PARTYSUMIDEO_FINAL)4      -0.75011   0.15070  -4.977 6.44e-07 ***
## factor(F_PARTYSUMIDEO_FINAL)9      0.57700   0.15722   3.670 0.000243 ***
## factor(MPX_VAX_W114)2                  0.32512   0.26062   1.247 0.212217
## factor(MPX_VAX_W114)3                  0.85996   0.24074   3.572 0.000354 ***
## factor(MPX_VAX_W114)4                  2.42956   0.23697  10.252 < 2e-16 ***
## factor(MPX_VAX_W114)5                  1.17521   0.34075   3.449 0.000563 ***
## factor(CVD_PROBS_NOCOMPLY_W114)2      0.35752   0.10376   3.446 0.000569 ***
## factor(CVD_PROBS_NOCOMPLY_W114)3      0.78230   0.11055   7.077 1.48e-12 ***
## factor(CVD_PROBS_NOCOMPLY_W114)4      1.37808   0.12470  11.051 < 2e-16 ***
## factor(CONF_f_W114)2                  0.65487   0.11591   5.650 1.61e-08 ***
## factor(CONF_f_W114)3                  1.30385   0.13021  10.013 < 2e-16 ***
## factor(CONF_f_W114)4                  1.96360   0.17177  11.431 < 2e-16 ***
## factor(F_AGECAT)2                  -0.32592   0.13067  -2.494 0.012625 *
## factor(F_AGECAT)3                  -0.80454   0.13668  -5.886 3.95e-09 ***
## factor(F_AGECAT)4                  -1.50168   0.14709 -10.209 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7786.5  on 8470  degrees of freedom
## Residual deviance: 5233.2  on 8446  degrees of freedom
## AIC: 5283.2
##
## Number of Fisher Scoring iterations: 6

```

Figure 6: Baseline Logistic Regression Model Summary

We attempted to enhance our baseline model with `stepAIC` function, but it did not reduce any features, meaning that our model has reached its minimum AIC value with this set of independent variables. Thus, we will stick with our baseline model and evaluate its performance in the next section.

During this process, we noticed that the variable representing device type is unexpectedly a very promising predictor for whether the individual will more (less) likely take the vaccination. However, this variable seems to be very random and not at all appear to have correlation intuitively with vaccination status. Thus, we decided to investigate how a seemingly random variable would contribute significantly predicting the response variable. We theorize that it is correlated with trust in elected officials, confidence in scientists, journalists, and business leaders, and family income. We are only interested in the different between people who used a phone or a laptop because all other device type is too low in number.

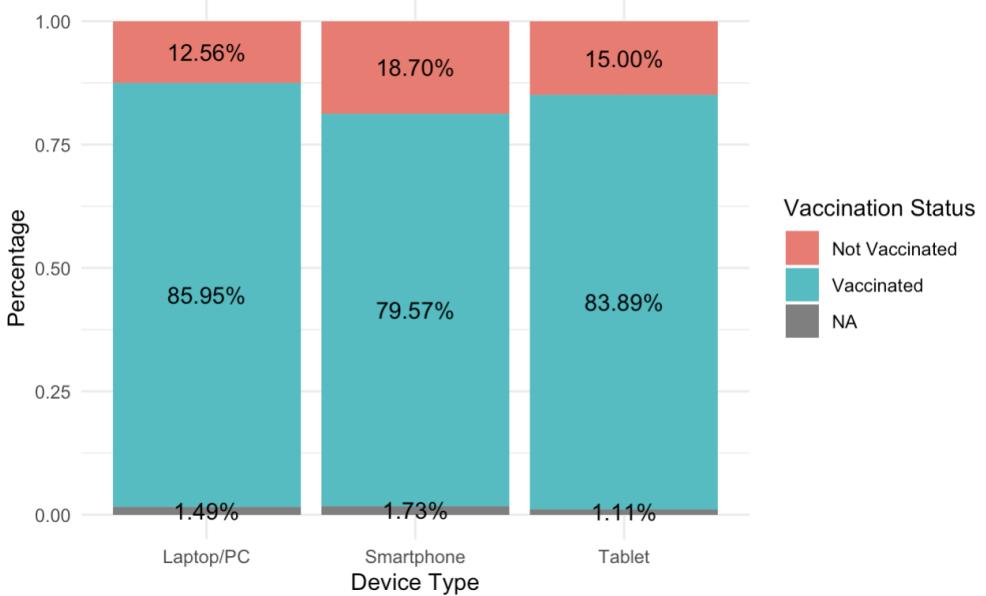


Figure 7: Mosaic Plot of Device Type vs. Vaccination Status

The summary of the model is as below:

```

##           Variable      P_Value
## trust          trust 9.193714e-01
## F_INC_SDT1    F_INC_SDT1 9.697487e-25
## CONF_b_W114   CONF_b_W114 4.282490e-06
## CONF_e_W114   CONF_e_W114 5.832190e-02
## CONF_f_W114   CONF_f_W114 6.283165e-04

```

Figure 8: Summary of Device Type Model

This table shows that selection between different device types is significantly correlated with family income and people's trust in scientists and journalists, all with a very small p-value indicating statistical significance. These three features are seemingly more related to vaccination uptakes, and their correlations with device type might explain why it contributes significantly in our baseline model.

7 Model Evaluation

As we return to our model analysis, having established and understood the relationships among our variables, we now want to evaluate our model. This stage is essential for determining the effectiveness and reliability of the model in making accurate predictions. Since we have a binary classification task, ROC curve is a good evaluation metric as it plots False Positive Rate against True Positive Rate.

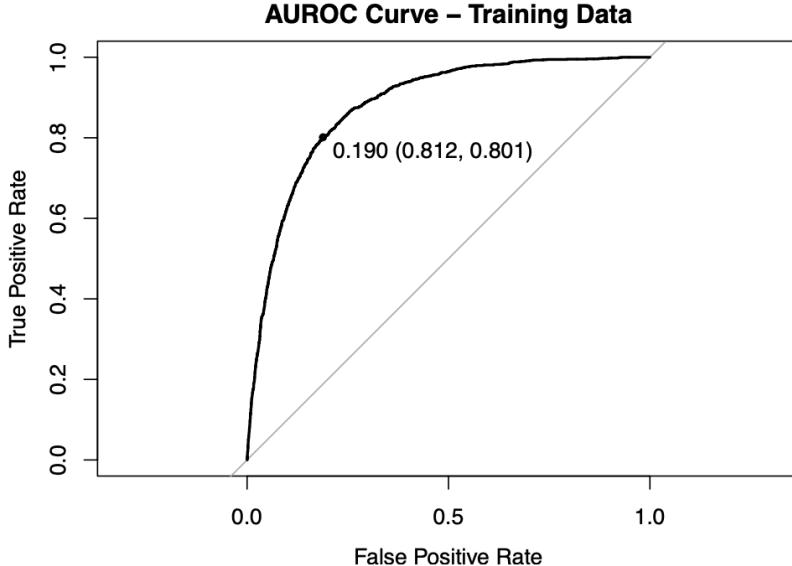


Figure 9: ROC curve

Based on the curve, we can tell that our curve is very close to a perfect ROC curve which reaches the point (0.0, 1.0), so it indicates that our model is very promising.

To further evaluate the goodness of fit of the model on the train data. We use a Hosmer and Lemeshow GOF test, where the null hypothesis is that the model fits the data reasonably well, with the alternative hypothesis that the model does not fit the data. HL test is a generalized version of Pearson's Chi-square test for the goodness of fit of a logistic regression. The key idea under this test is to group observations and predictions into larger groups.

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: as.numeric(X_train$COVID_vaccinated == "N"), train_probs  
## X-squared = 23.28, df = 8, p-value = 0.003023
```

Figure 10: HL Test Results on Training Data

From the test results, at $\alpha = 0.05$, we fail to reject the null hypothesis that our model fits the data reasonably well. Therefore, our goodness of fit is relatively acceptable.

Now, we can evaluate our model on the test data, which is 20% of the data randomly chosen from our original (imputed) dataset. We have an accuracy of **81.77%**, a precision of **96.27%**, and a recall of randomly chosen **81.10%**. The three values are all very promising, indicating that we have a fairly decent model.

8 Conclusion

Having developed a model with good performance, it is important to interpret our findings in a broader context. Our objective was to identify potential barriers to COVID vaccination uptake. Given that our model consistently forecasts an individual's non-vaccination status with some reliability, we can take a closer look at the predictors and their associated coefficients to infer possible risk factors preventing people from receiving vaccination.

From the summary of our model along with the coefficients for each feature category, we have pinpointed several factors that may contribute to vaccine hesitancy, including:

- Distrust in medical scientists
- Frustration with the non-compliance of others regarding COVID safety measures
- Negative opinions on receiving the Monkeypox vaccine
- Preference for using smartphones for survey participation
- Hesitation to disclose political party or ideological affiliation

Although our model demonstrates commendable accuracy, it is impossible to be flawless. Therefore, it is essential to recognize that this model does not universally apply to every scenario and should primarily serve as an informative tool.