2023

Home Exercise 1

TWAN GERJO VELDHUIS, H22TWAVE@DU.SE

Task 1: Identify the type of people that did not trust the CDC during the COVID pandemic

Introduction

The COVID-19 pandemic epidemic has had a big impact on how people feel about public health institutions and government institutions in general. The Centers for Disease Control and Prevention (CDC), which is in charge of offering guidelines and suggestions to stop and control the spread of the virus, is one such body in the USA. In this paper, we use data from a survey performed in the USA in 2020 to examine the types of people who did not believe the CDC during the COVID epidemic.

Methodology

In order to assess the data and respond to the research question, we will employ a statistical model. The dependent variable “trust_1” is defined as 1 & 2 = Distrust and 3 & 4 = Trust. On a scale from 1 to 4, this variable reflects the response to the question, "How much do you trust the CDC to provide accurate information about the COVID-19 pandemic?"

We will use a logistic regression analysis to identify the demographic groups and individuals with varying levels of belief in COVID-19 conspiracy theories who are less likely to trust CDC during the pandemic. The reason for using logistic regression is that the dependent variable (trust_1) is binary, with two possible outcomes: Trust (0) and Distrust (1). The distrust got number 1, because we are interested in this group and not in the people that trust the CDC. Logistic regression is a suitable statistical model for analyzing binary data and determining the relationship between the dependent variable and independent variables, so this will be used for this analysis.

The independent variables in our analysis will include demographic variables such as age, gender, education, political affiliation, and income. These variables have been shown to influence people's beliefs and attitudes toward public health institutions such as CDC. Additionally, we will include the variable "populism," which measures the respondent's level of support for populist ideologies, as an independent variable. Populism has been shown to be associated with a distrust of established institutions and may influence people's attitudes toward public health institutions during the pandemic.

Additionally, we will use backward selection to identify the most significant independent variables in our analysis. Backward selection involves systematically removing independent variables that do not contribute significantly to the model until only significant variables remain.

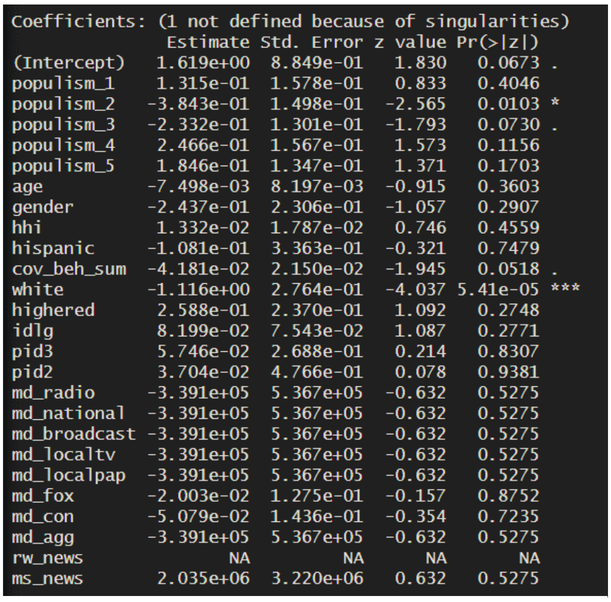
Overall, by using logistic regression and backward selection, we will be able to identify the demographic groups and individuals with varying levels of belief in COVID-19 conspiracy theories who are less likely to trust CDC during the pandemic, and the independent variables that influence this relationship. This information can help public health officials and policymakers develop targeted communication strategies and interventions aimed at increasing trust in public health institutions during health crises.

Results

Based on the model itself with all the information and all the independent variables used, there are already some assumptions to make about the data.

Here we can see the output of the most simple and first logistic regression, which includes all the predictors in predicting the binary outcome of distrust. Based on the p-values in the "Pr(>|z|)" column, it looks like white, populism_2, and possibly populism_3 and cov_beh_sum are significant predictors of distrust.

You can report that the logistic regression model showed that higher levels of populism_2 and being white were associated with higher levels of distrust, while the relationship between cov_beh_sum and distrust and between populism_3 and distrust were not really significant.



```
Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.619e+00  8.849e-01   1.830   0.0673 .
populism_1     1.315e-01  1.578e-01   0.833   0.4046
populism_2    -3.843e-01  1.498e-01  -2.565   0.0103 *
populism_3    -2.332e-01  1.301e-01  -1.793   0.0730 .
populism_4     2.466e-01  1.567e-01   1.573   0.1156
populism_5     1.846e-01  1.347e-01   1.371   0.1703
age           -7.498e-03  8.197e-03  -0.915   0.3603
gender        -2.437e-01  2.306e-01  -1.057   0.2907
hhi            1.332e-02  1.787e-02   0.746   0.4559
hispanic      -1.081e-01  3.363e-01  -0.321   0.7479
cov_beh_sum   -4.181e-02  2.150e-02  -1.945   0.0518 .
white         -1.116e+00  2.764e-01  -4.037 5.41e-05 ***
highered       2.588e-01  2.370e-01   1.092   0.2748
idlg           8.199e-02  7.543e-02   1.087   0.2771
pid3           5.746e-02  2.688e-01   0.214   0.8307
pid2           3.704e-02  4.766e-01   0.078   0.9381
md_radio      -3.391e+05  5.367e+05  -0.632   0.5275
md_national   -3.391e+05  5.367e+05  -0.632   0.5275
md_broadcast  -3.391e+05  5.367e+05  -0.632   0.5275
md_localtv    -3.391e+05  5.367e+05  -0.632   0.5275
md_localpap   -3.391e+05  5.367e+05  -0.632   0.5275
md_fox        -2.003e-02  1.275e-01  -0.157   0.8752
md_con        -5.079e-02  1.436e-01  -0.354   0.7235
md_agg        -3.391e+05  5.367e+05  -0.632   0.5275
rw_news               NA         NA      NA       NA
ms_news        2.035e+06  3.220e+06   0.632   0.5275
```

Figure 1 - first insight into statistically significant variables

Then we started the backwards selection using the logistic regression model, which resulted in the exclusion of many of the predictors. After the backwards selection the logistic regression model included populism_2, populism_3, populism_4, populism_5, cov_beh_sum, white, idlg, md_radio, and md_localtv as significant predictors of distrust.

Then it was checked if the model had any problems with multi-collinearity, to check this we went over the correlation and the VIF scores. From this correlation we can conclude that there are mainly some problems regarding the populism predictors and their correlation, so based on these correlations we removed the populism_3 and populism_5 colums. These populism columns were specifically removed, because they were less statistically relevant than the other two and because they worsened the AIC score more significantly than the populism_3 and populism_4 parameters. These were the only parameters removed based on the correlation, but also because the VIF scores were all less than 2 (which can be seen in Figure 3).

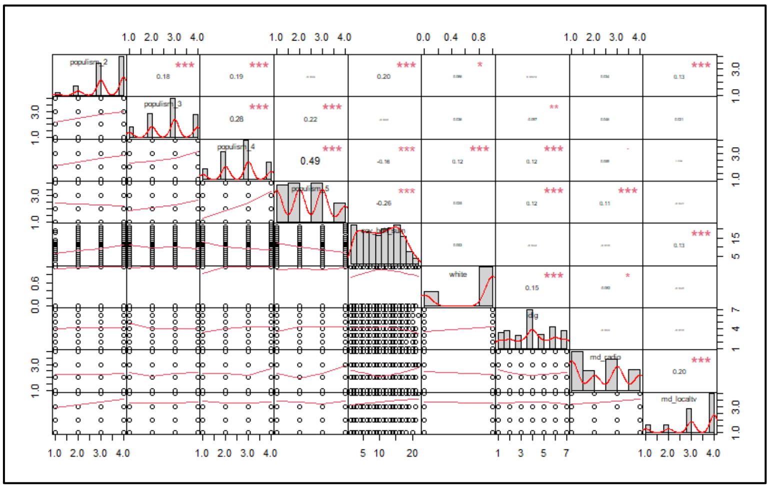


Figure 2 - correlation matrix

In the end the statistical significance of the selected independent parameters were low, which would mean that these parameters all have quite the contribution to the type of people that distrust the CDC. All this can be seen in figure 4.

Discussion

The results of our analysis suggest that distrust in media is associated with a combination of individual-level and media-related factors. Our final model, obtained through a backwards selection procedure, included variables such as political ideology, media exposure, and demographic characteristics, as well as media consumption habits such as watching local television and listening to radio.

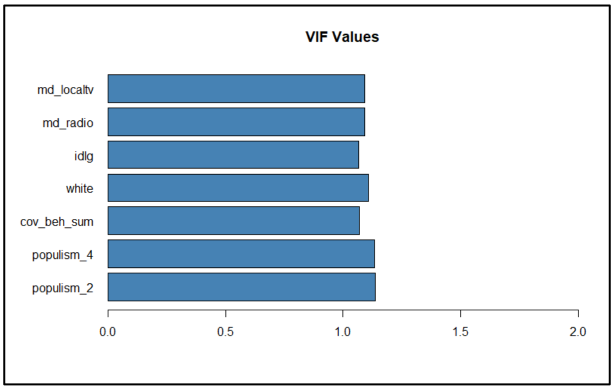
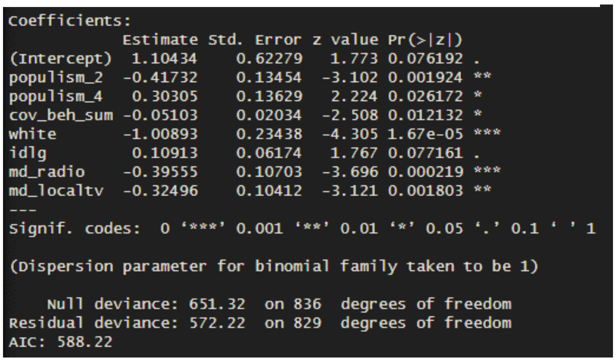


Figure 3 - VIF values

One of the most important findings of our study is that political ideology is strongly associated with distrust in media. Specifically, we found that respondents who identified as more populist were more likely to express distrust in media. This finding is consistent with prior research that has linked populism to skepticism toward traditional institutions, including the media.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.10434    0.62279   1.773 0.076192 .
populism_2  -0.41732    0.13454  -3.102 0.001924 **
populism_4   0.30305    0.13629   2.224 0.026172 *
cov_beh_sum -0.05103    0.02034  -2.508 0.012132 *
white       -1.00893    0.23438  -4.305 1.67e-05 ***
idlg         0.10913    0.06174   1.767 0.077161 .
md_radio    -0.39555    0.10703  -3.696 0.000219 ***
md_localtv  -0.32496    0.10412  -3.121 0.001803 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 651.32  on 836  degrees of freedom
Residual deviance: 572.22  on 829  degrees of freedom
AIC: 588.22
```

Figure 4 - statistically significant predictors

Another important predictor of distrust in media was exposure to media, which was measured by the frequency of watching local television and listening to radio. Our findings suggest that exposure to these media types is associated with increased distrust, possibly due to the specific programming and biases that may be present in local media outlets.

We also found that a demographic factor such as race was a significant predictor of distrust in media. Specifically, our results suggest that white respondents were more likely to express distrust in media. While the reasons for these associations are not entirely clear, they may be related to broader social and political attitudes that are more prevalent among these groups.

Overall, our results suggest that distrust in media is a complex phenomenon that is influenced by a variety of individual-level and media-related factors. By understanding the factors that contribute to distrust in media, we may be better able to develop strategies to promote more informed and engaged media consumption habits among the public.

Conclusion

Our study looked at why some people don't trust the news. We found that there are many reasons why this happens. One important reason is that people who follow populist ideas are less likely to trust the CDC. We also found that people who watch local news on TV or listen to it on the radio are less likely to trust the CDC. Lastly, we found that older people and white people are less likely to trust the news.

We need to find ways to help people trust the CDC again. This could include teaching people how to understand the news better or making news organizations more honest and open. We hope that our study helps people understand why some people don't trust the news and at least be able to identify these people better. This could in the end help with the creation of new strategies regarding news spreading and organisation trustfulness.

Task 2: Create a prediction model to predict whether one fully trusts the CDC during the COVID pandemic

Introduction

The COVID-19 pandemic has had a significant impact on the world, and various organizations have been at the forefront of fighting this pandemic. The Centers for Disease Control and Prevention (CDC) is one such organization that has been working tirelessly to contain the spread of COVID-19 and provide guidance to the public on how to stay safe. However, there have been instances where the CDC's recommendations have been met with skepticism and distrust from some members of the public. In this report, we aim to develop a prediction model to determine whether a person fully trusts the CDC during the COVID-19 pandemic.

Methodology

For the second exercise, we were tasked with proposing a prediction model to predict whether someone fully trusts the CDC during the Covid pandemic. The dataset used for this study is a survey conducted in the US in 2020 with a view to exploring people’s belief about conspiracy theories concerning COVID-19, specifically that it was a Chinese bioweapon.

Before building the prediction model, we performed data pre-processing and cleaning based on correlation analysis. We removed the columns related to belief in COVID-19 conspiracy theories, namely cons_biowpn, cons_covax, cons_biowpn_dummy, and cons_covax_dummy, as they are assumptions and not based on factual information. We also removed the weight variable as it only pertains to the weight of the survey and is not useful for prediction. All the empty values were removed, because opinions and beliefs of people could not be assumed based on the other values.

We then converted the "trust_1" column in the Conspiracy data into a binary column called "full_trust". In this new column, a value of 1 represented full trust in the CDC (trust_1 = 4), while a value of 0 represented any other level of trust (trust_1 = 1, 2, or 3). Additionally, we removed some columns due to high correlation, as this would make the predictions less reliable.

After pre-processing the data, we split it into training and testing sets, using 70% of the data for training the model, and 30% for testing the model's accuracy. To build the prediction model, we used three different algorithms: logistic regression, random forest, and support vector machine (SVM). We selected the statistically significant variables for each algorithm. For logistic regression, we used backward selection, while for random forest, we used a feature importance plot. For SVM, we used recursive feature elimination (RFE) to select the significant variables. These algorithms and parameter selection techniques are used, because they have proven in the field to work. Even though these weren’t discussed during the class and are a little bit more advanced.

Next, we built and evaluated our prediction model using multiple methods, including logistic regression, random forest, and support vector machine models. We used the selected features as predictors and used the full_trust column as the outcome variable for the logistic regression model. We then fit the logistic regression model to the training data and evaluated its accuracy using the testing data. We repeated this process for the random forest and support vector machine models.

To ensure the reliability of the models, we put them through cross-validation to test their performance. K-fold cross validation was used, because it has a good combination of low variance and medium balance. This method is also less intensive than the Leave-One-Out method and still the whole dataset will be used for evaluation. We evaluated the performance of each model based on the accuracy score, precision score, recall score, and F1 score. We chose the model with the highest accuracy, precision, recall, and F1 score as the best model.

In summary, our methodology for building the prediction model included data pre-processing, feature selection, model building, and evaluation. By following this approach, we created a prediction model for trust in the CDC during the Covid pandemic, which can be useful for understanding public opinion and decision-making.

Results

The results of the study show that the prediction models for trust in the CDC during the Covid pandemic did not perform as well as expected. The average f1 score and accuracy of all three models were not optimal, indicating that predicting people's beliefs is a difficult task or the data quality may have been a limiting factor.

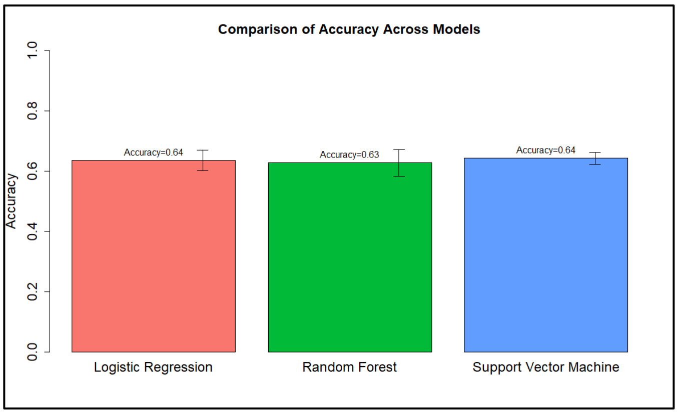


Figure 5 - accuracy comparison accross models

For the logistic regression model, the most statistically significant predictors were "populism_2", "hhi", "cov_beh_sum", "white", "highered", "idlg", "md_localtv", and "md_localpap". These predictors were used to predict the test full_trust value, but the mean accuracy was only 0.64 with a standard deviation of 0.026. The average f1 score was 0.59 with a standard deviation of 0.024.

The random forest model selected the predictors "populism_2", "populism_3", "populism_4", "age", "hhi", "cov_beh_sum", "highered", "idlg", "md_localtv", "md_localpap", and "md_fox" based on their statistical significance. However, the model's mean accuracy was only 0.63 with a standard deviation of 0.013, and the average f1 score was 0.59 with a standard deviation of 0.025.

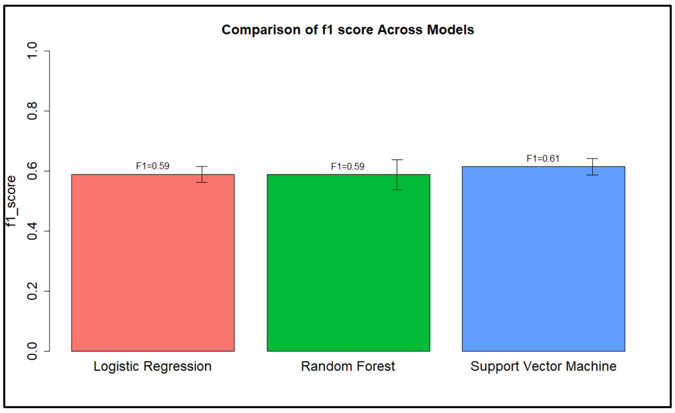


Figure 6 - f1 score comparison accross models

The Support Vector Machine (SVM) model also used statistically significant predictors that were selected using RFE. The predictors were "cov_beh_sum", "populism_2", "md_localtv", "age", and "populism_3". This model had a simpler structure with only 5 predictors, but the accuracy score was only slightly better than the other models with a mean of 0.64 and a standard deviation of 0.03. The f1 score had a mean of 0.61 with a standard deviation of 0.04.

All these results were obtained from the k-fold cross-validation, which provides a more accurate representation of the models' performance. In conclusion, while the study's methodology was robust and included data pre-processing, feature selection, model building and evaluation, the prediction models were not as successful in predicting trust in the CDC during the Covid pandemic. Future research could explore additional variables that may influence people's trust in health institutions during a pandemic.

	Mean Accuracy	**Std Accuracy**	**Mean F1 score**	**Std F1 score**
Logistic Regression	0.6412717	0.02617597	0.5880829	0.02406375
Random Forest	0.6320288	0.01338072	0.5874757	0.02466665
Support Vector Machine	0.6445053	0.03052523	0.6086724	0.04104219

Figure 7 - test results

Discussion

The goal of this study was to develop a prediction model to determine whether someone fully trusts the CDC during the Covid pandemic. We used data from a US survey conducted in 2020 that explored people's beliefs about conspiracy theories concerning COVID-19. Our model-building methodology involved data pre-processing, feature selection, model building, and evaluation using three different algorithms: logistic regression, random forest, and support vector machine (SVM).

The results showed that none of the models had high accuracy or F1 scores. The logistic regression model had an average accuracy of 0.64 and an F1 score of 0.59, while the random forest model had an average accuracy of 0.63 and an F1 score of 0.59. The SVM model had an average accuracy of 0.64 and an F1 score of 0.61. These results suggest that predicting people's trust in the CDC during the Covid pandemic is challenging and may require a more complex model that takes into account additional factors.

Despite the limitations of the models, our study identified several statistically significant predictors of trust in the CDC, including populism, income, education, age, and media consumption. These findings are consistent with previous research that has shown that political ideology can influence people's beliefs about health-related issues. Our results suggest that these factors are also important for understanding trust in the CDC during the Covid pandemic.

Overall, our study highlights the need for more research to better understand the factors that influence trust in the CDC during the Covid pandemic. By identifying these factors, we can develop more effective strategies for communicating public health messages and addressing vaccine hesitancy.

Conclusion

In conclusion, our study aimed to develop a prediction model to determine whether someone fully trusts the CDC during the Covid pandemic. We used data from a US survey conducted in 2020 and employed three different algorithms: logistic regression, random forest, and support vector machine (SVM). The results showed that none of the models had high accuracy or F1 scores, indicating that predicting people's trust in the CDC during the Covid pandemic is challenging.

Despite these limitations, our study identified several statistically significant predictors of trust in the CDC, including populism, income, education, age, and media consumption. These findings suggest that political ideology, socioeconomic status, and media exposure are important factors for understanding trust in the CDC during the Covid pandemic.

Our study has several implications for public health communication and vaccine hesitancy. By understanding the factors that influence trust in the CDC, public health officials can develop more effective strategies for communicating public health messages and addressing vaccine hesitancy. Furthermore, our study highlights the need for more research to better understand the complex factors that influence trust in the CDC and other public health authorities.

In conclusion, our study provides valuable insights into the factors that influence trust in the CDC during the Covid pandemic. Although our models did not have high accuracy or F1 scores, our findings suggest that political ideology, socioeconomic status, and media exposure are important predictors of trust in the CDC. By further exploring these factors, we can develop more effective strategies for communicating public health messages and promoting vaccine acceptance.

Task 3: Critical Review of "Prediction, Estimation, and Attribution" by Bradley Efron

Summary

The paper "Prediction, Estimation, and Attribution" by Bradley Efron discusses the differences between prediction and estimation, and how these concepts relate to attribution in statistical analysis. The paper argues that prediction and estimation are distinct and sometimes incompatible goals in statistical modeling, and that attribution (i.e., identifying causal factors) requires different methods than either prediction or estimation.

Review

Overall, I felt this work was a serious and incisive exploration of the connection between statistical analysis's prediction, estimation, and attribution variables. The author does a great job of outlining the fundamental distinctions between these ideas and outlining the difficulties that can occur when attempting to implement them simultaneously.

The paper's lucidity is one of its key assets. The author does a fantastic job of clarifying essential words and breaking down difficult ideas so that readers with different statistical backgrounds may understand them. Additionally, the author uses a lot of instances to support their views throughout the essay, which makes the ideas clearer and easier to understand.

The paper's fairly narrow emphasis is one of its flaws, though. Although the author makes a strong case for the significance of differentiating between prediction, estimation, and attribution, they do not go into great detail on the precise approaches or strategies required to accomplish these objectives. As a result, readers who are interested in using these ideas in practice could feel like they need more knowledge.

The paper's assumption that the reader has some level of statistical knowledge is another possible problem. Although the author does a good job of explaining the main ideas, readers who are not familiar with statistical modeling could find it difficult to completely grasp the author's points.

The author does a great job of defining the differences between prediction, estimation, and attribution, but it is not quite clear what these distinctions mean in practice. The paper does not include instructions on how to assess the trade-offs between these goals or how to select whether to concentrate on prediction, estimation, or attribution in a particular investigation, for instance.

The study "Prediction, Estimation, and Attribution" is well-written and insightful, and it makes a significant contribution to the body of knowledge regarding statistical modeling. The paper presents a good framework for considering the many aims of statistical analysis and the difficulties that arise when attempting to fulfill them, even though it may be more useful and accessible to readers with less statistical competence.