Balaji Vijayaraj

V22balvi@du.se

**Home Exercise 1**

Statistical Learning (AMI22T)

**Question 1**

**Introduction:**

The question requires us to identify the type of people who did not trust CDC during the COVID-19 pandemic in the US. We will use the conspiracy data taken from a survey conducted in the US in 2020. The analysis will be conducted using a logistic regression model to predict trust in the CDC using several demographic and media-related variables.

**Method:**

1. I loaded the dataset into R and removed the columns cons\_biowpn, cons\_covax, cons\_biowpn\_dummy, cons\_covax\_dummy, and weight as per the instruction.
2. I checked the summary and structure of the dataset to get the number of missing data and to know the data types.
3. I coded the values 1 and 2 in the trust\_1 column as 1, indicating distrust, and 3 and 4 as 0, indicating trust. I now have a new column called Distrust. Then I removed the trust\_1 column.
4. I fitted the logistic model with trust as the response variable and the remaining variables as predictors(I fitted the predictors individually). I chose the logistic model because the response variable was binary.

**Results:**

Based on the coefficients and p-values, we can see that there are several variables that are statistically significant in predicting distrust in the Centers for Disease Control and Prevention (CDC)

Graphical user interface

Description automatically generated

**Discussion:**

The variables "populism\_2", "cov\_beh\_sum", "white" (whether the respondent identifies as White), "md\_radio" ,"md\_national" , "md\_broadcast" , "md\_localpap", "local\_tv", "ms\_news" and "md\_fox" all have statistically significant negative coefficients, indicating that higher levels of these variables are associated with **lower levels of distrust in the CDC.**

The variable "**populism\_5**" has a statistically significant positive coefficient, indicating that higher levels of this variable are associated with higher levels of distrust in the CDC.

The variables "populism\_1", "populism\_3", "populism\_4", "age", "gender", "hhi", "hispanic", "highered", "idlg", "pid2", "rw\_news" and "pid3" do not have statistically significant coefficients, indicating that they are not strongly associated with distrust in the CDC.

**Conclusion:**

Therefore, it seems that people who marked high in populism\_2, cov\_beh\_sum , and consume new from certain sources like radio, national, broadcast, local\_tv, ms\_news and local newspapers are less likely to have distrust in the CDC, while **those who marked high in populism\_5 are more likely to have distrust in the CDC**.

**Limitations:**

Limitations of the methods used include the assumption of linearity in the logistic regression model, which may not hold in reality, and the assumption of normality in the correlation analysis.

Finally, the results of this analysis are specific to the sample used in the survey and may not generalize to other populations.

**Question 2**

**Method:**

1. For this task, I followed the same procedure as the previous one, but with a slight modification. I recoded the responses to the CDC trust question, such that 1, 2, and 3 were considered as 0, and 4 was considered as 1, representing full trust in the CDC.
2. I plotted the correlation matrix to find out the predictors that are highly correlated.
3. From the plot, I found that rw\_news with md\_con and md\_fox were highly correlated (more than 0.8), so I removed rw\_news as it was highly correlated with the other two, and md\_con and md\_fox were not highly correlated compared to rw\_news.
4. I found that ms\_news with md\_radio, md\_national, md\_localpap, md\_agg, and md\_broadcast were highly correlated (more than 6.8). So, I removed ms\_news as it was highly correlated with the other predictors and the other predictors were less correlated with themselves.
5. I found that pid2 was highly correlated with pid3 (0.88) and removed pid2.
6. I removed idlg as it was correlated with pid3 (0.54) and had a large number of missing values (150).
7. I removed md\_localtv as it was correlated with md\_broadcast.
8. After removing these columns, I removed missing values. I removed missing values after finding correlated variables to prevent unwanted removal of observations.

**Feature Selection:**

**Backward stepwise selection** was used for subset selection in this task because it is a computationally efficient method that can handle a large number of predictors. With 26 predictors in this analysis, it would be time-consuming to consider all possible combinations of predictors. Additionally, backward stepwise selection is less prone to overfitting than other methods, and it can handle missing data and correlated features by removing one feature at a time. It starts with a model that includes all predictors and removes them iteratively until the most important predictors are identified, which is suitable for this analysis where I want to identify the most important predictors for fully trusting the CDC. Then I used the resulting predictors in my model.

Text

Description automatically generated

I chose to use a simple prediction models to predict if someone fully trusts CDC based on the variables selected from a backward selection process. The dataset I used had a balanced number of observations for both classes(497 for class 0 and 424 for class 1). I split the data into 70% for training and 30% for testing. This is because a simple model can overfit with too much training data. If the model were more complex, I could have used more training data.

I decided to use logistic regression, LDA, QDA, KNN, and Naive Bayes models because my response variable is binary, and my predictors are a mix of categorical and numeric variables. Finally, I will pick the best model among these.

**Assumptions:**

**Logistic Regression:**

1. Linearity: The relationship between predictors and the logit of the outcome should be linear.
2. Independence: The observations should be independent of each other.
3. No multicollinearity: There should be no perfect multicollinearity between the predictors.
4. Large sample size: Logistic regression requires a large sample size for accurate estimates of coefficients.

**Linear Discriminant Analysis (LDA):**

1. Normality: The predictors are normally distributed in each class.
2. Homogeneity of variance: The variance of each predictor is the same across all levels of the outcome variable.
3. Independence: The predictors should be independent of each other.

**Quadratic Discriminant Analysis (QDA):**

1. Normality: The predictors are normally distributed in each class.
2. Homogeneity of variance: The variance of each predictor is the same across all levels of the outcome variable.
3. Independence: The predictors should be independent of each other.

**K-Nearest Neighbor (KNN):**

1. No assumptions of distribution: KNN does not assume a specific distribution of the predictors.
2. Independence: The predictors should be independent of each other.
3. Similarity: The predictor values should be similar for similar outcomes.

**Naive Bayes:**

1. Conditional independence: The predictors are conditionally independent given the class.
2. No multicollinearity: There should be no perfect multicollinearity between the predictors.
3. Large sample size: Naive Bayes requires a large sample size for accurate estimates of probabilities.

**Result:**

|  |  |
| --- | --- |
| Text  Description automatically generated with medium confidence | Text  Description automatically generated |
|  |  |
| Text  Description automatically generated with low confidence | Text  Description automatically generated |
|  |  |
| A picture containing text  Description automatically generated |  |

**Discussion:**

To determine whether someone fully trusts the CDC, the most important metric would be **precision**. This metric measures how accurate our model is at identifying people who truly have full trust in the CDC. While **accuracy and F1 score** would provide an overall assessment of how well the model is performing. Difference between training accuracy and test accuracy should also be lowest as possible.

**Conclusion:**

Among the five models I tried, **LDA and Logistic** models performed the best in terms of accuracy and precision. However, **I would choose LDA as it is simple and has a slightly higher F1 score than the Logistic model.**

**Limitations:**

One limitation of the backward selection method is that it assumes that the predictors that are left in the model are not related to each other, but this may not always be true.

The LDA model may not work well if the relationship between predictors and the response variable is not linear. LDA is also a model that assumes the data is distributed in a particular way, so if the data doesn't follow that pattern, LDA may not be a good option. Lastly, if there are many predictors in the dataset, LDA may be slow and not practical.

**Question 3:**

**Critical Review of Prediction, Estimation, and Attribution by Bradley Efron:**

Bradley Efron's "Prediction, Estimation, and Attribution" highlights key differences between prediction algorithms and traditional statistical methods. While prediction algorithms offer impressive accuracy and scalability, Efron contends they lack interpretability and theoretical grounding in comparison to traditional methods.

One issue that Efron raises is the "black box" nature of prediction algorithms. While these algorithms can produce highly accurate predictions, it is often difficult to understand how they arrived at those predictions. Efron argues that this lack of transparency makes it difficult to critique or improve these algorithms. To address this issue, he suggests that researchers develop methods for interpreting the output of prediction algorithms and measuring their statistical sufficiency.

Another issue that Efron raises is the lack of a theoretical framework for prediction algorithms. Traditional statistical methods are grounded in maximum likelihood theory, which provides a lower bound on the accuracy of an estimate and a practical way of nearly achieving it. In contrast, Efron argues that we do not have an optimality theory for prediction algorithms. This means that we cannot say with confidence whether a given prediction algorithm is performing optimally, or if there is another algorithm that could do much better.  
  
Overall, Efron's paper highlights important limitations and challenges of prediction algorithms, while also acknowledging their potential benefits. One counter-example to Efron's arguments, however, might be the field of deep learning. Deep learning algorithms, which are used in tasks such as image and speech recognition, have been shown to produce highly accurate predictions while also offering some interpretability through techniques such as feature visualization. Additionally, recent work has shown that deep learning models can be theoretically grounded in concepts such as information theory and Bayesian inference. These examples suggest that while there are certainly challenges to developing accurate and interpretable prediction algorithms, it is possible to address some of these challenges through continued research and innovation.  
  
While Efron acknowledges that prediction algorithms offer certain advantages, he argues that traditional methods are still important for providing interpretability and theoretical grounding. However, it is possible that as prediction algorithms continue to improve, traditional methods may become less relevant for certain applications. Additionally, it may be possible to combine the strengths of both approaches to develop new methods that offer both accuracy and interpretability.

**Questions for the author :**

1. How might the lack of a theoretical optimality framework for prediction algorithms impact the development and assessment of these algorithms?
2. What are some of the most promising research directions for improving the interpretability and theoretical grounding of prediction algorithms? Are there any specific techniques or approaches that are particularly promising?
3. Do you think that traditional statistical methods will continue to be relevant in the face of advances in prediction algorithms, or do you see the two approaches as complementary? If yes, then how?