STATISTICAL LEARNING (AMI22T)

ASSIGNMENT 1

DALARNA
UNIVERSITY

Author:

Maha Vajeeshwaran

INTRODUCTION:

This assignment consists of two tasks and firstly I performed the task with the help of R studio. I was provided survey data from the 2016 Prudential Election campaign in the United States.

1.a. Recode the variable Trump as follows. Denote Slightly liberal to Extremely liberal (levels 1-3) as “Liberal”, and Moderate to Extremely conservative (levels 4-7) as “Conservative”. Is there any personal characteristics of the individuals that determines whether someone would consider Donald Trump as Liberal (or conservative)? Motivate your methods and interpret your results.

1.b. Build a suitable prediction model to predict an individual’s party identification using the respective individual’s personal, and family characteristics. Experiment with different methods, and model specifications, and motivate your choice.

The second task is to write a critical summary of the following article: Efron, B. (2020), Prediction, Estimation, and Attribution.

METHODS:

Task:1

1.a In this task datasets provided with the categorical values and separate .txt document is provided to show the clear understanding of data.

- In this task data is imported in the R studio firstly checked the summary of the given data. It is found that presence of missing values in the data. By analysing, it is found that missing values are very less so I removed all the missing values.
- Now data cleaning process is performed. I removed all the data like -9 and -8 which is “Refused” and “Don’t Know”. As this doesn’t provide any valuable information to perform prediction.
- Now recoded the Trump variable data as Liberal and Conservative as per the given task

Firstly I check the correlation of all the feature by using cor() function and heatmap. It is found that some variables are more correlated like Education and Education2, Partner & Martial, Age & Marital. So, I felt like it is wrong to come with a solution by feeding all variables to model.

I factored the Trump variable to make it as categorical and I performed with the logistic regression model for all predictor with the Trump as response. But this is not the correct method because of multi collinearity the result will be not reliable for accessing p values so I performed another method.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.723015   0.353895  -4.869 1.12e-06 ***
Media       -0.039571   0.024070  -1.644 0.100173
FamSize     -0.029292   0.046006  -0.637 0.524325
Hillary      0.567200   0.029144  19.462  < 2e-16 ***
Age          0.003096   0.003175   0.975 0.329518
Education   -0.076412   0.022211  -3.440 0.000581 ***
Employment   0.004065   0.021516   0.189 0.850146
Birthplace   0.048852   0.044090   1.108 0.267859
GBirth      -0.081630   0.042040  -1.942 0.052168 .
Dependent    0.135381   0.061470   2.202 0.027638 *
Housing     -0.017403   0.052640  -0.331 0.740936
Income      -0.020198   0.006920  -2.919 0.003517 **
Partner     -0.111461   0.065737  -1.696 0.089971 .
SpouseEdu   -0.033184   0.014281  -2.324 0.020147 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3722.1  on 3755  degrees of freedom
Residual deviance: 3097.9  on 3742  degrees of freedom
AIC: 3125.9
```

By the presence of multi collinearity we cannot get the correct relationship of individuals personal characteristics and Trump. I also cannot remove the variable which multi collinear.

Secondly, I factored the Trump variable to make it as categorical and I performed with the logistic regression model for each predictor with the Trump as response. By using the P value its relationship with the Trump is accessed. Below Table shows the how well each variable related with the Trump variable. * Symbol represents the significance codes. How strongly the Trump variables has the relationship with the predictors. It is categorized as *** Very strong, ** Low and * Very low.

***(p<0.001	**(p<0.01)	*(p<0.05)
Media	Famsize	Age
Hillary	Employment	Gberth
Education		
Birthplace		
Dependent		
Housing		
Income		
Partner		
SpuseeEdu		

It is found that Media, Hillary, Education, Birthplace, Dependent, Housing, Income, Partner, SpuseeEdu has very strong relationship with trump. Famsize and Employment has low relationship with trump. Age and Gberth has very low relationship with Trump.

Task 1.b:

As like the first question 1.a data is imported and data cleaning process is performed by removing missing values and unwanted values like -9 and -8. In this task response variable is PartyID and predictors are remaining fields except ID.

Method 1:

In this multinom() method is performed as it is basic model because my response variable contains more than two categories so I used this method.

- I performed cross validation by using validation set approach. Divided the 80% of data as training data and remaining 20% data as test data.
- I factored the PartyID and performed prediction using multinom() and checked the accuracy by using mean() function, predicted values with the test data and I got the result of 0.492 or 49.2%.

```
> pred.fit<-predict(multinom.fit, newdata=US_data_test)
> accuracy <- mean(pred.fit == US_data_test$PartyID)
> accuracy
[1] 0.4920213
```

Method 2:

In this method to go with the subset selection and checking the model I used best subset selection method. I found that following features are necessary for the prediction with the response PartyID variable. Media, Hillary, Trump, Age, Partner, SpouseEdu, Birthplace, GBirth now I performed prediction with the multinom method. It gives the accuracy of 0.49468. it is found that after feature selection accuracy is slightly improved. It shows this model predicted correctly at 49.46%

```
> accuracy <- mean(pred.fit == US_data_test$PartyID)
> accuracy
[1] 0.4946809
```

Method3:

By using the best subset selection method features I used LDA model for the prediction. LDA is a multi-class classification approach that may be used to do dimensionality reduction automatically. It is found that the accuracy by using mean() function, predicted values with the test data and I got the result of 0.4853723. It shows this model predicted correctly at 48.53%.

```
> table(lda.class, US_data_test$PartyID)

lda.class   1   2   3   4
        1 155  41  86  11
        2  52 153  85  18
        3  58  29  57   7
        4   0   0   0   0
> mean(lda.class == US_data_test$PartyID)
[1] 0.4853723
```

Method4:

By using the best subset selection method features I used QDA model for the prediction. The LDA variant quadratic discriminant analysis (QDA) allows for non-linear data separation. It is found that the accuracy by using mean() function, predicted values with the test data and I got the result of 0.4281915. It shows this model predicted correctly at 42.8%

```
qda.class   1   2   3   4
        1  76  34  41   7
        2  73 152  92  17
        3 115  37  94  12
        4   1   0   1   0
> mean(qda.class == US_data_test$PartyID)
[1] 0.4281915
```

Method 5:

By using the best subset selection method features I used KNN model for the prediction and it is found that the accuracy by using mean() function, predicted values with the test. It shows this model predicted correctly at 44.4% when K=10 and 41.7% when K=5 and 39.7% when K=3.

Conclusion:

Here as per the above model analysis it is found that the Multinomial model (multinom()) perform better for this data with the best subset selection variables which shows the accuracy rate of 49.46% when compared to the LDA, QDA, KNN. Through this analysis it is found that the LDA and logistic regression techniques tend to perform well when the true decision boundaries are linear. QDA may provide better outcomes when the boundaries are significantly non-linear. A non-parametric technique like KNN can be preferable for much more complicated decision boundaries.

Code for the above test is attached in the Zip folder while submitting.

Task 2 (Question2)

Summary:

The research paper published on 2020 examines what are the differences between modern machine learning algorithms and traditional regression methods such as ordinary least squares or logistic regression? Several significant inconsistencies will be investigated in this paper, with a focus on the distinctions between prediction and estimation, as well as prediction and attribution (significance testing). The analysis is concentrated on tiny data sets. In his paper author highlighted

The research paper published on 2020 examines what are the differences between modern machine learning algorithms and traditional regression methods such as ordinary least squares or logistic regression? Several significant inconsistencies will be investigated in this paper, with a focus on the distinctions between prediction and estimation, as well as prediction and attribution (significance testing). The analysis is concentrated on tiny data sets. In his paper author highlighted six differences in the last part of his paper. The Author emphasis that mathematical equations in surface plus noise models, can create continuous and smooth functions which used in traditional regression models. Whereas Pure prediction models, on the other hand, are really not built on any scientifically verified methodology.

Pure prediction models are subject to changing environments since they do not have a scientific foundation. Regression models, according to the author, do not have this issue and are designed to last a long time. Every variable has a coefficient in a classical model like linear regression, which describes how much each variable impacts the result. whereas a pure predictive algorithm is often a black box, without any understanding of how the prediction was made. Traditional models have a significant advantage in terms of parsimonious modeling. The researchers can determine which factors are most essential based on the results. This is in sharp contrast to pure prediction models, which usually favor more variables.

The author performed an experiment in which he removed half of the factors that pure prediction model believed to be the most essential. The model's accuracy did not vary much, implying that no meaningful inferences regarding the value of any variable can be drawn. Traditional models chose homogeneous data, which contains few predictors but numerous samples, whereas prediction algorithms prefer the opposite. Because correlation with the predictors can lead to overfitting. To determine how good a model is, pure prediction models rely on training and test paradigm (common task framework) whereas classical model uses theory of optimal inference.

Discussion:

The Professor Efron said that twenty-first century has witnessed the birth of a new generation of what may be termed pure prediction algorithms. In his research paper, the author has presented us with thought-provoking research on the link between prediction, estimate, and attribution. I agree with most of Professor Efron's ideas in his research paper, for example, addresses random splits when using the train/test approach, which is common in the pure prediction algorithms. In particular, he points out that in some cases, such as with time-dependent data, a random split will be significantly too optimistic rather than offering an accurate estimate of the error. He claims that in such cases, a more realistic split is still not random, but instead an early or late split, where the early data is used to train and latter data is employed for testing. In many circumstances, a random

split guarantees that the training and testing dataset are distributed equally, However, it falls short of ensuring that they are independent.

Thus, a random split bears little resemblance to the link between the data for training an algorithm and the data it will be used in the future. While the author uses this example to emphasize how prediction is simpler than extrapolation or interpolation based on the observation that in time-dependent data, a random split achieves better test-set results than early or late split. The train or test model, in my opinion, should not be focused on randomization, but rather on reflecting how the prediction algorithm can be utilized. The testing data should be indicative of the new dataset where the algorithm will be applied, while the training dataset should be relevant upon the current data to train the algorithm. While I agree with the author that a random split is not always suitable, I believe that the representation of the split is critical for the pure prediction model, not its randomness of split [Bin Yu et al.,].

I agree with the author that interpretable prediction models can occasionally be useful in getting closer to the truth. Simple interpretable models are sometimes more reliable and resilient. It's also easy to identify and maybe fix an issue when a basic model stops operating. Predictive model interpretation is a popular subject in research right now. It's important to understand why a prediction model looks to be working. Sometimes deep learning can fool us for example like classification of image between dog and tiger as shown below Fig:1.

Fig:1 False prediction of image between dog and tiger

The Author uses different criteria in Table 5 to show his views on the difference between pure prediction and classic regression approaches. For example, he sees traditional regression models as focusing on scientific truth, whereas pure prediction methods focus on prediction accuracy; he sees traditional regression models as focusing on low-dimensional problems, whereas pure prediction methods focus on high-dimensional problems; and he sees traditional regression problems using the optimal inference theory like the maximum likelihood and Neyman–Pearson, whereas pure prediction methods focus on training or test paradigms.

Many modern machine learning books incorporate traditional statistical concepts like linear regression, maximum likelihood and logistic regression. The argument that conventional regression models emphasis upon scientific truth whereas pure prediction techniques focused on prediction accuracy, it seems to have some value, however the fact that classic regression models supposedly convey is rarely justified or confirmed. According to the author pure prediction algorithms, "concentrate on prediction, to the exclusion of estimation and attribution." While p-values related with model coefficients are used in classic regression methods, several pure

prediction approaches, like random forest, have their own attribution mechanisms as in form of variable importance (gini scores). The author addresses this theory but criticizes current attribution methods as inadequate, stating the pure prediction method's emphasis on prediction employing numerous weak learners who do not favor specific strong predictors as a reason. Unlike the author, I do not believe that prediction, estimate, and attribution are wholly incompatible. They have different objectives, but both are necessary.

Modern prediction algorithms, in my opinion, go hand in hand with more classic estimating techniques. There are Machine Learning examples all over the place. It's how Facebook identifies a friend's face in a digital snapshot or how google translate used to translate the language, or how a customer care professional can tell if you'll be satisfied with the services before you really fill out a customer satisfaction survey.

Traditional regression machine learning algorithms have the advantage of being machine-like, despite their complexity. They require a great deal of subject expertise, as well as human assistance that is effective of doing what it was built to accomplish. Whereas Deep learning holds more potential for AI creators and the developing world in this aspect. Most of the applicable features in conventional statistical approaches must be identified by a domain expert in order to minimize data complexity and make patterns more obvious to learning techniques to work. As previously said, the most significant advantage of Deep Learning algorithms is that they attempt to understand high-level characteristics from data in an incremental manner. hard-core extraction of features and domain expertise are no longer required.

Another significant distinction between Deep Learning and traditional Machine Learning is the problem-solving strategy. Deep Learning strategies address problems from beginning to finish, but traditional regression machine learning techniques require the research problems to be broken down into separate sections, which must be solved separately before the findings can be combined at the end. When the data amount is high, Deep Learning outperforms conventional approaches. When it comes to complicated issues like natural language processing, picture classification and speech recognition, it truly shines. And also, when there is a lack of domain awareness for feature analysis, Deep Learning approaches shine since feature engineering is less of a concern.

Professor Efron concludes that modern machine learning algorithms require considerable further improvement before they are suitable for regular scientific use. Such methodologies, are already being successfully applied in ordinary scientific applications. I agree with many of the author points and of course, I believe that further improvement is always important.

Contributions:

This paper examines the key differences in current machine learning methods with the traditional regression methods by small data sets. Furthermore, this research paper contributes by providing meaningful insights of the use of classical statistical methods. Much of this article has been concerned with what the prediction algorithms cannot do, at least not in their present formulations. Author also advises the further research and development in modern machine learning methods.

Conclusion:

Professor Efron said in his research paper that the twenty-first century has witnessed the birth of a new generation of pure prediction algorithms. He states that prediction algorithms can be stunningly successful and that the emperor has wonderful clothes, but they are not fit for every occasion is something I completely agree with. I illustrated and discussed how current prediction methods are successful along the same lines and I firmly believe the most fundamental principles of twentieth-century statistics modeling, estimation, and inference will play a pivotal role in laying the mathematical basis for current data science and reaching its full potential for various applications.

Question to Author:

1. What will be your approach if the data for the image classification comes like Fig:1 during production stage by using traditional statistics?

2. What you will prefer between pure prediction model and traditional in the future AI research for eg, space research, satellite image analysis, etc and why?

3. As you said traditional model is long term and reliable why many sectors are not using it like Google, Amazon, etc?

3. What is your advice to the people to get rid of cancer from your medical research?

4. Do you think pure prediction model is good or bad?

Reference:

1. Min-ge Xie and Zheshi Zheng: Discussion of Professor Bradley Efron's Article on "Prediction, Estimation, and Attribution": JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION: VOL. 115, NO. 530, 667–671, 2020.

2. Jerome Friedmana , Trevor Hastiea,b, and Robert Tibshirania": JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION: VOL. 115, NO. 530, 665–666: Comment 2020.

3. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019), "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn," Nature Communications, 10, 1096.