MATH 4432 Mini-Project 1 Report

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

This report studies the binary classification of switching unsafe wells based on various given features (independent variables) about the situations. We attempt to select the best model with the assistance of various techniques such as fitting with z-values; estimating Type I & II Errors; computing ROC and AUC and choosing by validation.

1 Introduction

1.1 Background

Our project focuses on the decision making process of switching off unsafe wells that may have been polluted by arsenic in Bangladesh. We are trying to model the switching-off decisions with several potential influential factors provided in the data set. Our ultimate goal is to undetstand the underlying decision making process by treating it as a binary classifcation problem and to evaluate whether this decision making process in Bangladesh is practical.

1.2 Data Description

In the data set given in the .csv file given on the website:

https://github.com/yuany-pku/data/blob/master/wells.csv

we have eight variables *switch, arsenic, unsafe, distance, x, y, community, education*. We can find that *switch* and *unsafe* are boolean (binary) variables taking the value of either TRUE or FALSE. *education* serves as a catagorical variable and the others are real-valued features.

According to the description of this data set given in the project description, our response variable would be *switch*, which is the variable that we try to model and predict. And other 7 variables together form the candidate pool of features for the modeling of *switch*.

1.3 Methodologies

Since our response variable is binary, a usual and intuitive model on which the response is regressed would be logistic regression. Other models such as linear discriminant analysis (LDA) and K nearest neighbor (KNN) can also be potential candidates. As specified in the project description, the variable *education* is set to be a catagorical variable.

We first use a modified validation set approach to perform a preliminary exploration about which model (logsitic regression, LDA, QDA, KNN) gives the highest performance.To evaluate the goodness of model fitting, we also evaluate the models by their z-values. Then we apply validation set approach to estimate the misclassification error and confusion matrix, which are also known as type I and type II errors, respectively. Finally we compute the ROC curve to illustrate the diagnostic ability of a

Submitted to 30th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

binary classifier system as its discrimination threshold is varied, along with Area-Under-Curve (AUC) to evaluate how well a parameter can distinguish between two diagnostic groups.

2 Model Fitting and Evaluation

2.1 Preliminary explorations

The meanings of variables *x, y* in reality are not clear in the data description and thus, we decide to ignore these two variables and focus on other variables that may have more implications of properties on the pollution conditions of the wells.

2.2 Individual Correlation Analysis Against Response Variable

The threshold method we apply is to plot each independent variables against *switch* using boxplot. The results are shown below:

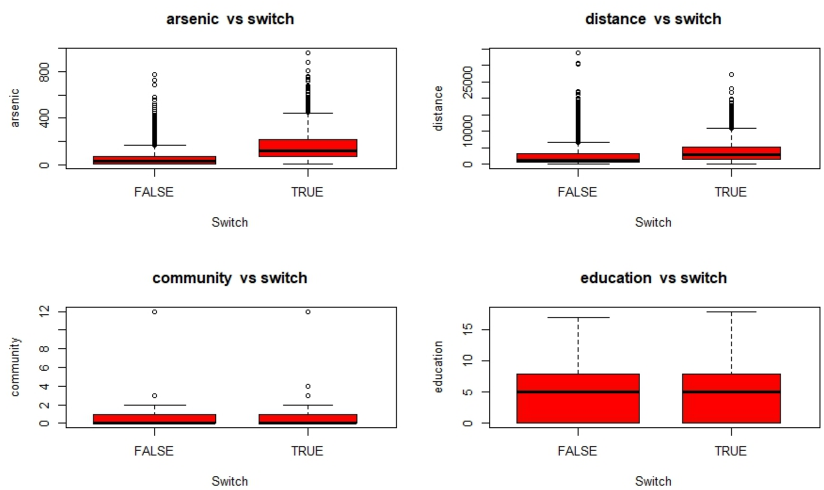

Figure 1: Relationships between *switch* and other variables

Note that we do not plot *unsafe* against *switch* because plotting one dummy variable against another may not visually reflect the significance of the dummy independent variable. So we use *table()* function to output the frequency of two variables. The results are shown below:

	FALSE	TRUE
FALSE	2638	1284
TRUE	432	2094

Figure 2: Relation between *unsafe* and *switch*

We can notice from Figure 1 that *arsenic* is the most significant variable, while *community* and *education* barely shows any difference between two response conditions. Combining resutls in Figure 2, we continue to consider *arsenic, distance* and *unsafe* as candidate independent variables (predictors). For simplicity, we will start from the most significant variable *arsenic* to look for appropriate method based on our model selection criteria, which is elaborated in the next subsection.

2.3 Threshold Exploration Using LDA, Logistic Regresion, QDA and KNN

Before the results of exploration using such methods are introduced, we would like to state how we apply validation set approach as resampling method, which are inherent in the computataion of error rates of each model selection method. As written in the code we attach, we randomly divide the data set into 100 pairs of training and validation data sets. We then iteratively fit LDA, logistic regression, QDA and KNN model to each training data set and obtain error rates of each model by using the validtion set. By randomly spliting the data set into multiple pairs of training and validation data sets, the uncertainty caused by each single choice of validation set can be accounted for. We do not choose K-fold cross validation because 1) as a method for preliminary exploration, K-fold cross validation, especially leaving one out cross validation, can be too computationally expensive 2) our data set is reaosnably large with 6448 observations in total and thus there is no need to use K-fold cross validation to ensure every observation has been used for training.

In this subsection, we will use the results from this modified validation set approach to demonstrate our ideas and results.

First of all, we plot the error rates of the model with *arsenic* as the single predictor of *switch* using LDA, Logistic Regression, QDA and KNN, repsectively. As stated above, since 100 pairs of training and validation data sets have been used, 100 error rates are calculated for each model. The results are shown below:

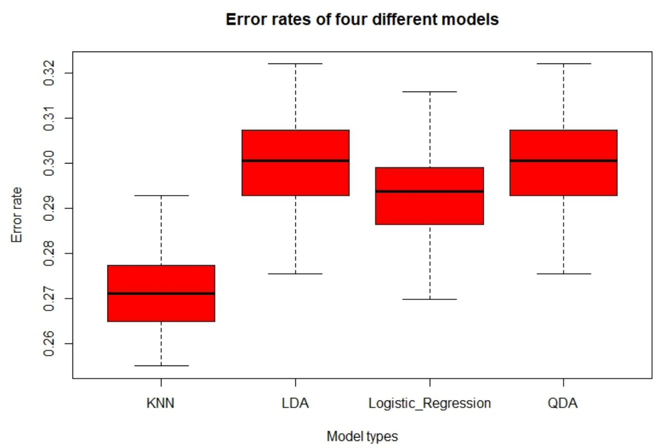

Figure 3: Error rates of four methods

As seen in Figure 3, KNN performs the best among all methods, which indicates that there may be a significant non-linear relationship between *arsenic* and *switch*.

Also, LDA and QDA performs significantly worse than other methods. The reason is that both methods assume the arsenic level in both classes (switch and not swicth) to follow normal distribution, which is not the case as shown in the normality check in Figure 4. This could explain the fact that logistic regression of binary response performs better than LDA and QDA because it does not require the Guassian assumption.

Not requiring on Guassian assumption also enables KNN method to outperform LDA and QDA methods. Moreover, the implication of KNN has inspired us that we may improve the performance of logistic regression by including non-linearity into logistic regression, which also explains why the error rate of logistic regression lies between KNN and LDA (QDA). Next up would be the inclusion of non-linearity of the model.

2.4 Including Non-Linearity Using Multiple Transformations

We try to find the type of non-linearity to appropriately and efficiently reflect the relationship between *arsenic* and *switch* using various types of transformation.

The potential choices of transformation we use are polynomial transformation of order 1 to 5, log transformation and square-root transformation. By applying these transformation, we plot the updated error rates against each type of transformation. The results is shown in Figure 5:

Based on the plot, we choose log transformation to be our transformation method. The resons are two-fold: first, its error rate is very close to the minimum error rate among these seven methods and due to the inherent randomness of our validation set approach, the subtle difference in error

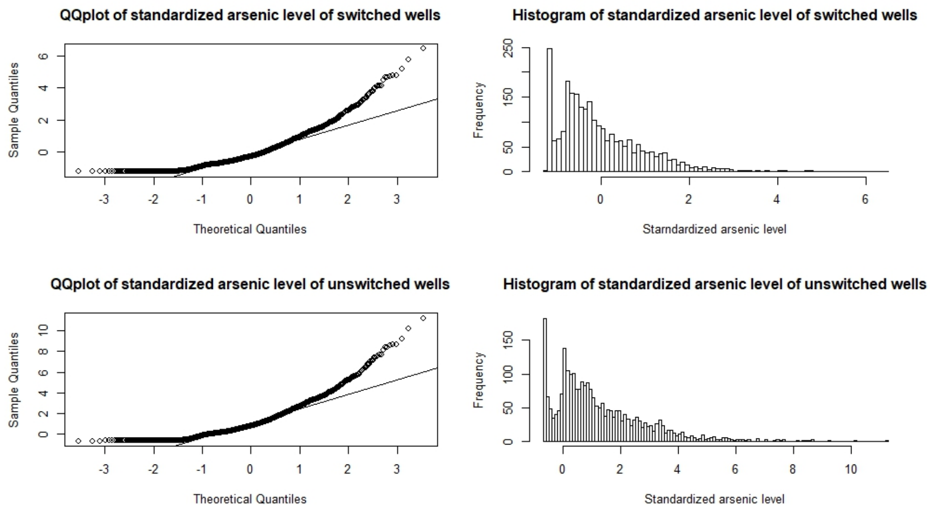

Figure 4: Normality assumptions of LDA and QDA has failed

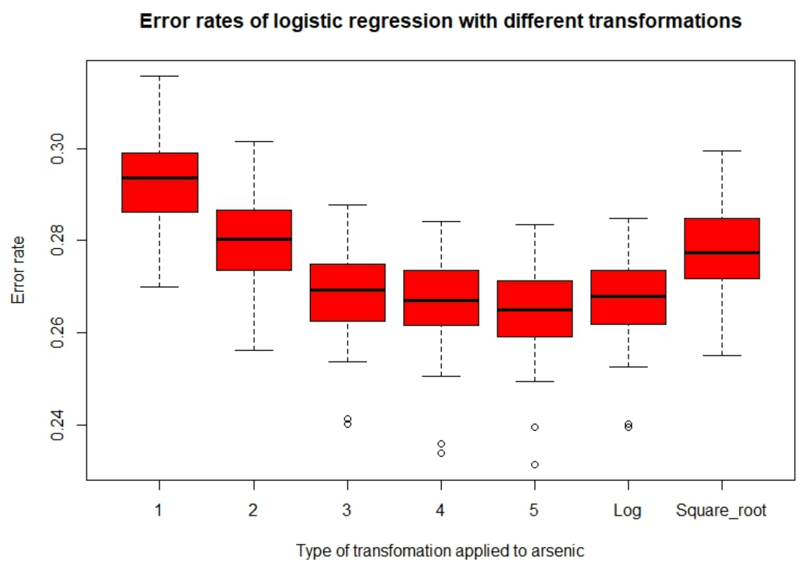

Figure 5: Performance of each transformation

rates cannot imply difference in performance. Secondly, logistic regression with log transformed arsenic predictor only requires two variables to be inferred while the seemingly best-performing transformation (5th order polynomial transformation requires six). As said earlier, our selection values both accuracy and efficiency.

After the transformation, we fit the model and the fitting results are shown below in Figure 6:

The fitted model is:

$$\log \frac{P(switch)}{1 - P(switch)} = -3.64163 + 0.81627 * \log{(arsenic)}$$

After this, we compute an estimate of the confusion matrix of the regression model above, which is shown in Figure 7:

```
Call:
glm(formula = switch ~ log(arsenic), family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9733  -0.9002  -0.4314   0.9636   2.7083

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.64163    0.11899  -30.60   <2e-16 ***
log(arsenic)  0.81627    0.02763   29.55   <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: 6465.4  on 4835  degrees of freedom
Residual deviance: 5246.1  on 4834  degrees of freedom
AIC: 5250.1

Number of Fisher Scoring iterations: 4
```

Figure 6: Model fitting with log transformation on *arsenic*

```
glm.log.results FALSE TRUE
          FALSE   752  203
          TRUE    217  440
```

Figure 7: Confusion matrix of the model fitting with log transformation on *arsenic*

3 Model Selection

Even though this section also includes some model fitting, we generally name it as Model Selection due to its emphasis on the choices between models.

3.1 Adding Other Predictors

Continued from Section 2.2, we are exploring if adding any combination of *unsafe* and *distance* would improve our regression. By fitting the models that add *unsafe*, *distance* and both, we are able to see that the estimates of regression coefficients are statistically significant. Due to the limitation of report length, we are not including the glm results in the report. The relevant code are in the Part(F) of the attached code.

Though all regression coefficients are significant, a immediate question that we need to answer is that should we include both *unsafe* and *distance* then. To answer this, we need to plot the error rates of each model. The results are shown in Figure 8.

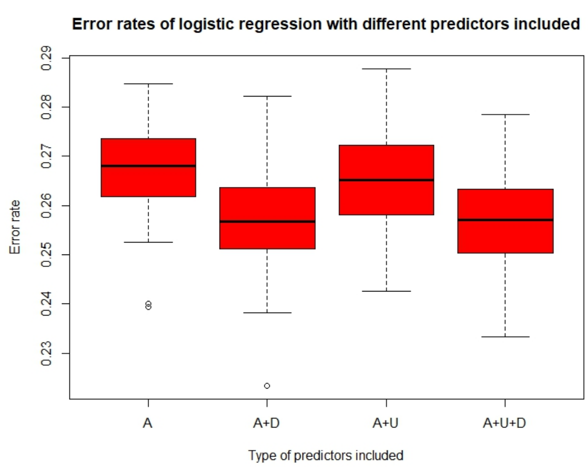

Figure 8: Error rates of models with different predictors

We find that the model performance improves significantly when *distance* is added, and that not much is changed when *unsafe* is added and thus we arrive to the conclusion that we should add *distance* but not *unsafe*.

The reason why *unsafe* is excluded is made clear if we examine its correlation with *arsenic*. As shown in Figure 9, the correlation between them are high so adding *unsafe* to a model with *arsenic* already included will contribute little. In this case, *unsafe* is simply redundant:

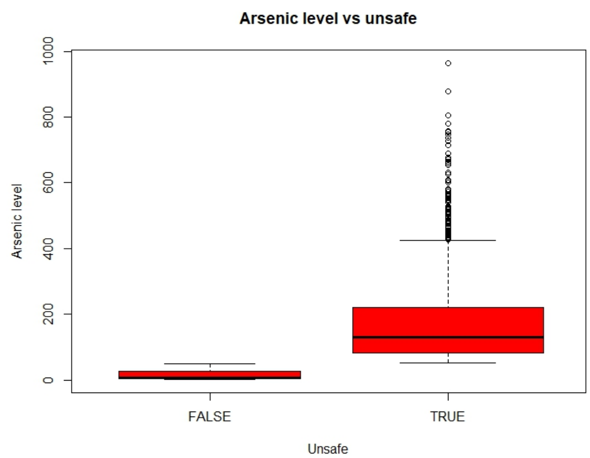

Figure 9: *arsenic* Vs. *unsafe*

3.2 One Last Evaluation

We compute the ROC curve and calculate the AUC of our finalized model (logistic regression with distance and log transformed arsenic as predictors).

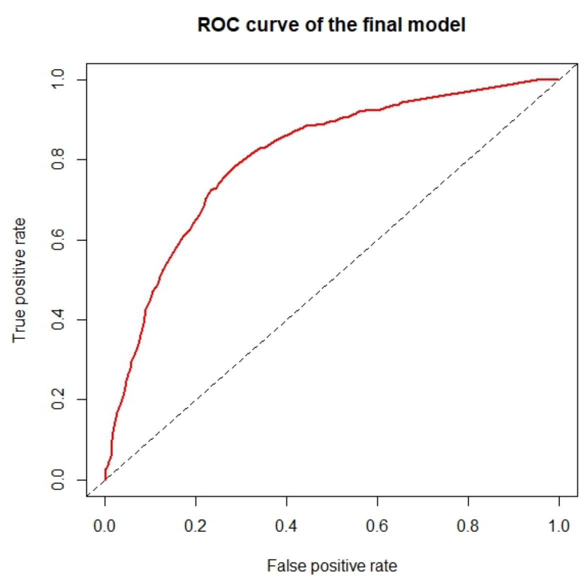

Figure 10: ROC

The AUC is 0.743, according to R output.

4 Model interpretation

The final fitting results are in Figure 11 below:

```
Call:
glm(formula = switch ~ log(arsenic) + distance, family = "binomial",
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0273  -0.9263  -0.4186   0.9739   2.7770

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.781e+00  1.245e-01 -30.364  < 2e-16 ***
log(arsenic)  9.052e-01  3.202e-02  28.270  < 2e-16 ***
distance     -6.618e-05  1.045e-05  -6.333  2.4e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6467.2  on 4835  degrees of freedom
Residual deviance: 5228.3  on 4833  degrees of freedom
AIC: 5234.3

Number of Fisher Scoring iterations: 4
```

Figure 11: Model with *arsenic* and *distance*

The final fitted model is:

$$\log \frac{P(switch)}{1 - P(switch)} = -3.781 + 0.9052 * \log{(arsenic)} - 6.618^{-5} * distance$$

As can be seen from the modle, higher arsenic level renders the corresponding well more likely to be switched off while longer distance from community can render the corresponding well less likely to be switched off. The negative correlation between distance and probability of switching off the well can be explained in the way that wells far away from community are less accessible to the public and thus may not be handled with top priority by local government. The log transformation of arsenic level means that at low arsenic level, the probability of switching off the well is very sensitive to increase in arsenic level, but as the arsenic level goes higher, the probability will become less sentitive to increase in arsenic level. This also has practical meaning because, for example, an increase in arsenic level by 90 from 10 to 100 can mean the difference between safe and unsafe while an increase by 90 from 310 to 400 does not make a difference because wells with arsenic pollution at this order of magnitude are all extremely unsafe and should not be drunk by the public. Therefore, it is reasonable to conclude that local government in Bangladesh is using a relatively systematic and practical strategy to control the hazard caused by arsenic pollution in drinking water. However, any well, no matter how far away from community it is, can potentially be the water source of people, especially when nearby wells have been switched off, and thus whenever goverment has enough administrative power and resources, it should consider excluding distance from their decision making process.

5 Conclusion

We first obtain a pool of potential predictors including *unsafe*, *arsenic* and *distance*. By analyzing the most significant one *arsenic*, we decide to transform it using log transformation for a relatively better fit. We then add *distance* to the model and exclude *unsafe* because of its redundancy with *arsenic*.

The final model we choose is a logistic regression model with distance and log transformed arsenic level as predictors. The performance of the model, as demonstrated by its error rate, ROC curve and AUC, is reasonably high. Also, as a parametric approach, the logistic regression can be interpreted to reflect the underlying decision making process of switching unsafe wells in Bangladesh. As revealed by the model, arsenic level and distance are both taken into consideration by the administrator in Bangladesh. By quantifying a decision making process into a parametric model, this model can potentially assist Bangladesh in future decision making or strategy evaluation.

6 Contributions

Coding in R: Xingbo SHANG;

Code Commenting: Xingbo SHANG, Kao ZHANG;

Project Report: Kao ZHANG.