Toronto Metropolitan University

R course

Final Project

Authors
Laya Ahmadi

Lecturer
Dr. Cody Ross

July 21, 2022

Contents

1. Introduction

Different fertilizers are needed to support plant growth. Nitrogen, phosphorus, and potassium are the main macronutrients that all plants require to survive. Specialized fertilizers with different nutrient combinations are also used to satisfy the needs of many different plants. Although fertilizer use is absolutely necessary for most crops' well-being and increasing the efficiency of agricultural activities, the application of fertilizer can lead to adverse outcomes. Much of the fertilizer applied to fields is not absorbed into the soil but is instead lost to runoff or leached into the soil profile, which eventually contaminates water resources.

For instance, Lake Erie is subjected to a pervasive and persistent nutrient-driven algal bloom problem and its effects. Micronutrient pollution leads to issues that compromise the Lake's ecosystem's ability to continue providing potential opportunities and benefits. Although agricultural activities are not the only source of nutrients entering Erie Lake, it is the dominant one.

To combat the growing threat of toxic algal development and the expansion of hypoxic zones in Lake Erie, the United States and Canada agreed to work together toward a 40 percent reduction from the 2008 baseline in the amount of phosphorus entering Lake Erie's Western Basin. The agreement is based on various mandatory and voluntary action plans and strategies to reduce nutrient loss. As mentioned earlier, agricultural activities, as the main component of non-point sources, are the most challenging part. Best management practices(BMPs) should be implemented to control the movement of potential contaminants from the source. However, the current state of agricultural practices should be assessed and understood well.

1.1. Objectives

Based on the literature review, the methods farmers use to apply fertilizer and the timing of the application matter in how much micronutrients will spread in surface water, and the objective is to know these factors better and quantify their effects.

For example, if the fertilizer is incorporated into the soil, it is less likely to be washed by runoff 1.1.

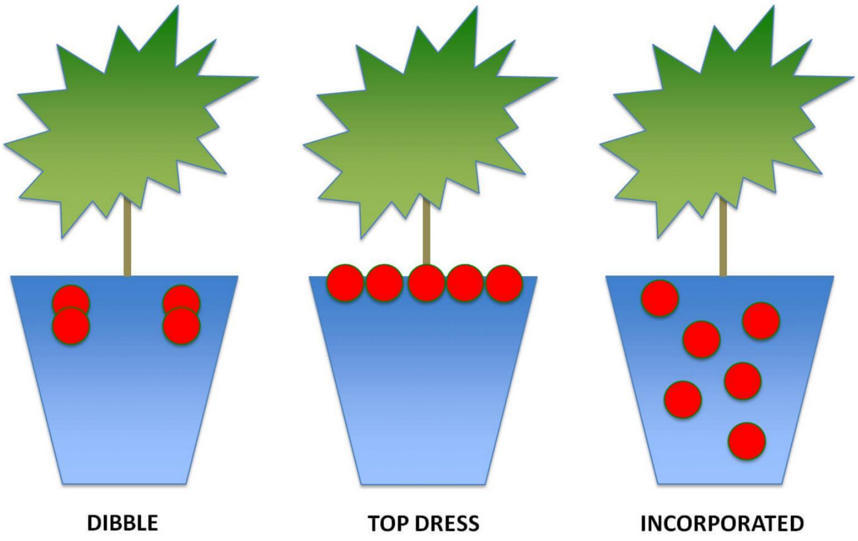


Figure 1.1.: Various fertilizer application

Farmers use different methods to apply fertilizer and manure to the agricultural fields. Different factors influence the chosen method for fertilization, such as crop type, field area, machinery availability, farmer attributes, season, plant's growth stage, type of fertilizer, and farmers' knowledge of conservation.

A survey dataset is available that collected data from 281 farmers that own 1261 farm fields in 11 watersheds on the western side of Lake Erie. The dataset entails data on fertilization events for five consecutive years, from 2015 to 2019, for each field. It also includes data about manure application, farmer attributes, field tillage practices, and cover crops.

The main idea is to determine the predictors that influence the outcome. In the first part, where I used logistic regression, the primary outcome is whether a field has been fertilized or not, and the aim is to find the potential predictors of whether a field has been fertilized or not.

In the second part, I used the multinomial logistic regression to determine the application that is most probable to be used to fertilize a field. Based on the dataset, we have four main methods that are used for applying fertilizer:

- Applied on living crop
- broadcast with no incorporation
- broadcast with incorporation
- injected

1.2. Literature Review

Multiple research studies study the effect of different factors such as farms' and farmers' attributes on farmers' behavior related to fertilization.

Filson et al. (2009) used a logistic regression model to predict the best management practices adoption rate index in a paper titled "beneficial management practice adoption in five southern Ontario watersheds." The survey data of 481 landowners from five Ontario watersheds used for this study and found that size of the land affected the adoption rate, but other variables such as farmer's education and age, watershed, farmer's gender, and off-farm income did not have a significant effect [1].

Liu et al. (2020) published a paper titled "best management practices and nutrient reduction: An integrated economic-hydrologic model of the Western Lake Erie Basin" and used an integrated economic model of farmers' field-level best management practices and SWAT model to assess the cost-effectiveness of different management scenarios in the Western Lake Erie Basin. They wanted to determine the probability of subsurface placement adoption using an ordered logistic regression model. The outcome categories are "unlikely to adopt," "likely to adopt," and "already adopted." They found that significant predictors are perceived efficacy of BMP, farmer age, farm income, cost of implementing BMP, and field size [2].

2. Methods

2.1. Logistic Regression

Logistic regression is a classification algorithm used to find the probability of a binary outcome. If the outcome is binary, meaning has two possible discrete, not continuous values; then the basic logistic regression analysis can be used. The logistic regression model is simply a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped function similar to the standard normal distribution but easier to work with in most applications. The logit distribution constrains the estimated probabilities to lie between 0 and 1. Here are the formulations used to calculate odds, log odds, and odds ratio [3].

Odds: odds are ratios in logistic regression.

$$odds = \frac{p}{1-p}$$

Log Odds: The natural log of the odds is also known as a logit.

$$logodds = logit = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

Odds Ratio: Odds ratios are actually ratios of ratios.

$$oddsratio = \frac{odds1}{odds2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Computing Odds Ratio from Logistic Regression Coefficient, odd ratios are obtained by exponentiating the coefficients.

$$oddsratio = \exp(b)$$

Computing Probability from Logistic Regression Coefficients, where Xb is the linear predictor

$$probability = \exp(Xb)/(1+\exp(Xb))$$

It means that if the probability of outcome-one of some event is p. Then the probability of outcome-two is 1 – p. The odds of outcome-one are defined as the ratio of the probability of happening outcome-one over the probability of happening outcome-two. That is to say that the odds of outcome-one are

$$odds = \frac{p}{1-p}$$

to 1.

The transformation from probability to odds is an exponential transformation, meaning the odds increase faster as the probability increases. Probability ranges from 0 and 1, and consequently, odds range from 0 and positive infinity.

There are different packages in R that contains a function for logistic regression. However, I used the function from LessR package [4] since it also creates a confusion matrix that shows the precision of the model for a binary outcome.

The following code shows how to fit a logistic regression model in R using the Logit function that comes with the LessR package:

$$Model\texttt{<-}Logit(outcomeVariable \sim Predictor1{+}Predictor2, data = dataset, prob\_cut = [0,1])$$

Here is how to interpret the values in the Pr(>|z|) column: a single asterisk (*) next to the p-value of a specific predictor means the p-value is statistically significant at $\alpha = 0.05$. The p-values in the output also give us an idea of how effective each predictor variable is at predicting the probability of the outcome [5].

For example, we might say that observations with a probability greater than or equal to 0.5 will be classified as ”1” and all other observations will be classified as ”0.” The Logit function’s cut-off probability is ”0.5” by default, but the user can change it. Using this threshold(the cut-off probability), we can create a confusion matrix that shows our predictions compared to the actual values. Using the confusion matrix, we can see the number of false positives and false negatives that are predicted using the Logistic model; in other words, the precision of the model is also calculated.

The coefficients in the output indicate the average change in log odds of defaulting. For example, a one unit increase in Predictor1 is associated with an average increase/decrease of ”reported coefficient” in the log odds of defaulting [6].

2.2. Multinomial Logistic Regression

Multinomial logistic regression, or simply multinomial regression, is sometimes considered an extension of binomial logistic regression to allow for a dependent variable with more than two categories, assume K categories, where the log odds of the outcomes are modeled as a linear combination of the predictor variables. In this regression model, we need to specify the reference category of our dependent variable, which is called pivot in this matter. To arrive at the multinomial logit model, K-1 independent binary logistic regression models would be running; the K-1 outcomes are separately regressed against the pivot outcome or the reference outcome. This would proceed as follows if outcome 1 (the first outcome) is chosen as the pivot [7]:

$$\ln \frac{Pr(Y_i = 2)}{P(Pr(Y_i = 1)} = \beta_2.X_i$$

$$\ln \frac{Pr(Y_i = 3)}{P(Pr(Y_i = 1)} = \beta_3.X_i$$

$$...$$

$$\ln \frac{Pr(Y_i = K)}{P(Pr(Y_i = 1)} = \beta_k.X_i$$

For each dependent variable category, except for the reference category, odds ratios are determined for all independent variables [8].

3. Results

For this analysis, the following packages are used.

- readxl [9]: to read the excel file(the survey data we have) and return a dataset in R environment.
- LessR [4]: for the Logit function, analyzing a model with a binary response variable. The output includes the confusion matrix and various classification fit indices.
- nnet [10]: for the multinom function
- forcats [11]: manage the categorical data fields such as crop type
- ggplot2 [12]: To plot the figures

The results for binary and multinomial logistic regression are presented in the following sections.

3.1. Logistic Regression

For this analysis, the data from the chemical fertilization database is used to predict whether a field has been fertilized or not based on the area of the field, "HA," and the crop type, "Crop." The S-shaped graph for

$$Model\texttt{<-}Logit(HasFertilizer \sim HA, dataset = fertiliser, prob\_cut = 0.8)$$

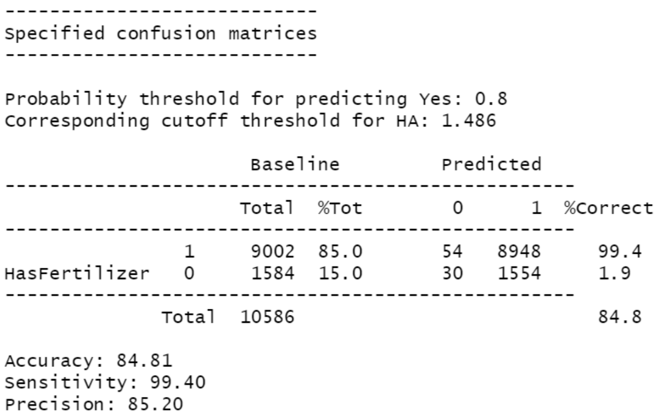
is below but there is no similar graph for

$$Model\texttt{<-}Logit(HasFertilizer \sim HA, dataset = fertiliser)$$

since Crop is a discrete data type and the regression coefficients calculated for each crop type. The confusion matrices for these two models are shown in figure 3.1 and 3.3 respectively.

The results for barley, corn, edible beans, hay_establish, rye, and wheat are significant and show when the crop type is

- barley, the odds of being fertilized (versus not being fertilized) increase by 8.43.



```
----------------------------
Specified confusion matrices
----------------------------

Probability threshold for predicting Yes: 0.8
Corresponding cutoff threshold for HA: 1.486

                      Baseline        Predicted
---------------------------------------------------
                     Total  %Tot       0      1  %Correct
---------------------------------------------------
                1     9002  85.0      54   8948     99.4
HasFertilizer   0     1584  15.0      30   1554     1.9
---------------------------------------------------
              Total  10586                          84.8

Accuracy: 84.81
Sensitivity: 99.40
Precision: 85.20
```

Figure 3.1.: confusion matrices, area of the field as a predictor

- corn, the odds of being fertilized (versus not being fertilized) increase by a factor of 87.53.
- edible beans, the odds of being fertilized (versus not being fertilized) increase by 14.64.
- hay_establish, the odds of being fertilized (versus not being fertilized) increase by a factor of 4.13.
- rye, the odds of being fertilized (versus not being fertilized) increase by 48.75.
- wheat, the odds of being fertilized (versus not being fertilized) increase by a factor of 55.40.

The results for the area of the field show that when the area increases by one unit, the odds of being fertilized (versus not being fertilized) increase by a factor of 1.026, so it is not significant.

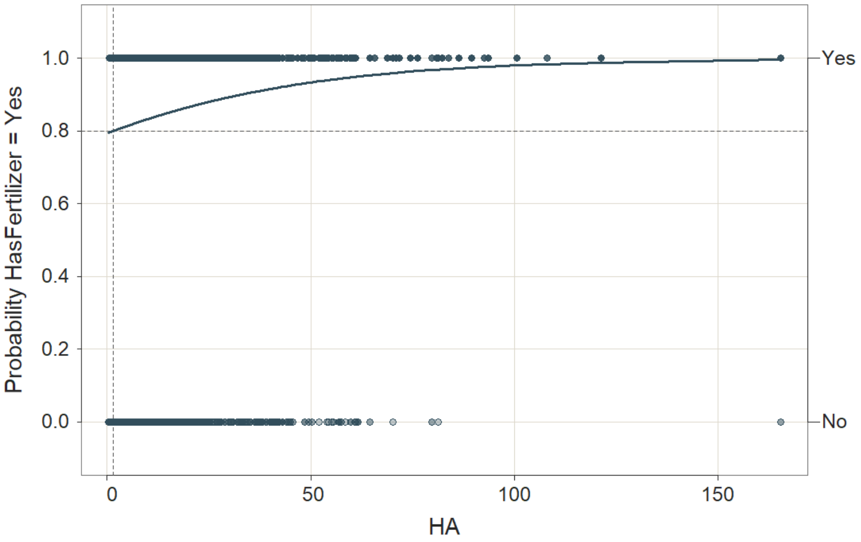
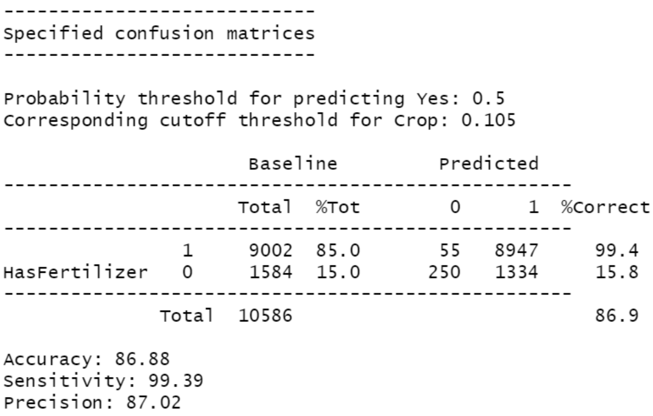


Figure 3.2.: incomplete S-shaped curve, area of the field as a predictor



```
----------------------------
Specified confusion matrices
----------------------------

Probability threshold for predicting Yes: 0.5
Corresponding cutoff threshold for Crop: 0.105

                      Baseline          Predicted
---------------------------------------------------
                     Total  %Tot         0       1  %Correct
---------------------------------------------------
                1     9002  85.0        55    8947     99.4
HasFertilizer   0     1584  15.0       250    1334     15.8
---------------------------------------------------
               Total  10586                             86.9

Accuracy: 86.88
Sensitivity: 99.39
Precision: 87.02
```

Figure 3.3.: confusion matrices, crop type as a predictor

3.2. Multinomial Logistic Regression

The data set contains variables on 7864 chemical fertilizer application incidents and 4510 manure application incidents. The outcome variable is ”Application,” the application that is used for fertilization for that specific incident. In section one of this part, the outcome is the manure application method, and in the second part, the outcome variable is the chemical fertilizer application method. The predictor variables are crop type, a three-level categorical variable, and field area, a continuous variable for both analyses.

I estimate a multinomial logistic regression model by using the ”multinom” function from the nnet [10] package. This analysis's pivot or reference value is ”M1: applied on living crop.” Here is the various binary logistic regression formula that would be considered for this analysis.

- M1: applied on living crop
- M2: broadcast with no incorporation
- M3: broadcast with incorporation
- M4: injected

The interpretation of β values in these formulation are described in the sections 3.2.1 and 3.2.2.

$$\ln \frac{P(Application = M2)}{P(Application = M1)} = \beta_{20} + \beta_{21}(Crop = Soybeans) + \beta_{22}(Crop = Wheat) + \beta_{23} * area$$

$$\ln \frac{P(Application = M3)}{P(Application = M1)} = \beta_{30} + \beta_{31}(Crop = Soybeans) + \beta_{32}(Crop = Wheat) + \beta_{33} * area$$

$$\ln \frac{P(Application = M4)}{P(Application = M1)} = \beta_{40} + \beta_{41}(Crop = Soybeans) + \beta_{42}(Crop = Wheat) + \beta_{43} * area$$

3.2.1. Manure Application method

Here are the results for multinomial logistic regression model on the manure dataset that shows the probability of using different application methods over a range of field sizes and three dominant crop types 3.4.

$$Model{<}{-}multinom(Application2 \sim Crop + HA, data = manure)$$

- β_{23} A one-unit increase in the variable ”area” is associated with the increase in the log odds of using the ”broadcast no incorporation” method vs. ”applying on living crops” in the amount of 0.0027, although this coefficient is not significant.

- β_{33} A one-unit increase in the variable ”area” is associated with the decrease in the log odds of using the ”broadcast with incorporation” method vs. ”applying on living crops” in the amount of 0.0211, although this coefficient is not significant.

- β_{43} A one-unit increase in the variable ”area” is associated with the increase in the log odds of using the ”injected” method vs. ”applying on living crops” in the amount of 0.0228, although this coefficient is not significant.

- β_{22} The log odds of using the ”broadcast no incorporation” method vs. ”applying on living crops” method will increase by 0.1862 if changing from crop = ”corn” to crop = ”wheat.”

- β_{21} The log odds of using the ”broadcast no incorporation” method vs. ”applying on living crops” method will increase by 1.0075 if changing from crop = ”corn” to crop = ”soybeans.”

- β_{32} The log odds of using the ”broadcast with incorporation” method vs. ”applying on living crops” method will decrease by 2.6611 if changing from crop = ”corn” to crop = ”wheat.”

- β_{31} The log odds of using the ”broadcast with incorporation” method vs. ”applying on living crops” method will decrease by 0.7908 if changing from crop = ”corn” to crop = ”soybeans.”

- β_{42} The log odds of using the ”injected” method vs. ”applying on living crops” method will decrease by 20.9497 if changing from crop = ”corn” to crop = ”wheat.”

- β_{41} The log odds of using the ”injected” method vs. ”applying on living crops” method will decrease by 12.1739 if changing from crop = ”corn” to crop = ”soybeans.”

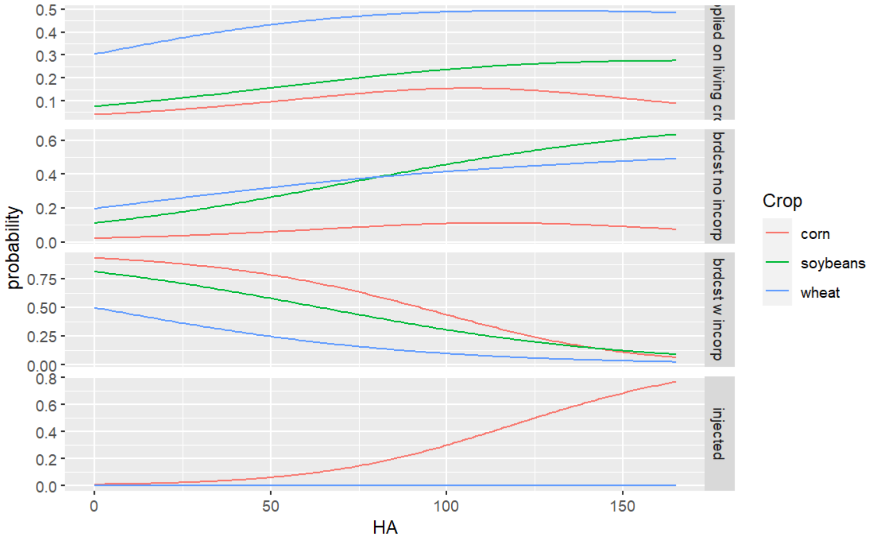


Figure 3.4.: Probability of using different fertilization method for manure for different crop types

3.2.2. Chemical fertiliser Application method

Here are the results for having multinomial logistic regression analysis on the chemical fertilizer dataset that shows the probability of using different application methods over a range of field sizes and three dominant crop types 3.5.

$$Model{<}{-}multinom(Application2 \sim Crop + HA, data = fertiliser)$$

- β_{23} A one-unit increase in the variable "area" is associated with the decrease in the log odds of using the "broadcast no incorporation" method vs. "applying on living crops" in the amount of 0.0032, although this coefficient is not significant.
- β_{33} A one-unit increase in the variable "area" is associated with the decrease in the log odds of using the "broadcast with incorporation" method vs. "applying on living crops" in the amount of .0042, although this coefficient is not significant.
- β_{43} A one-unit increase in the variable "area" is associated with the decrease in the log odds of using the "injected" method vs. "applying on living crops" in the amount of .0088, although this coefficient is not significant.
- β_{22} The log odds of using the "broadcast no incorporation" method vs. "applying on living crops" method will decrease by 0.2000 if changing from crop = "corn" to crop = "wheat."
- β_{21} The log odds of using the "broadcast no incorporation" method vs. "applying on living crops" method will increase by 3.9691 if changing from crop = "corn" to crop = "soybeans."

- β_{32} The log odds of using the ”broadcast with incorporation” method vs. ”applying on living crops” method will decrease by 2.7220 if changing from crop = ”corn” to crop = ”wheat.”
- β_{31} The log odds of using the ”broadcast with incorporation” method vs. ”applying on living crops” method will increase by 2.7393 if changing from crop = ”corn” to crop = ”soybeans.”
- β_{42} The log odds of using the ”injected” method vs. ”applying on living crops” method will decrease by 1.8200 if changing from crop = ”corn” to crop = ”wheat.”
- β_{41} The log odds of using the ”injected” method vs. ”applying on living crops” method will increase by 1.6421 if changing from crop = ”corn” to crop = ”soybeans.”

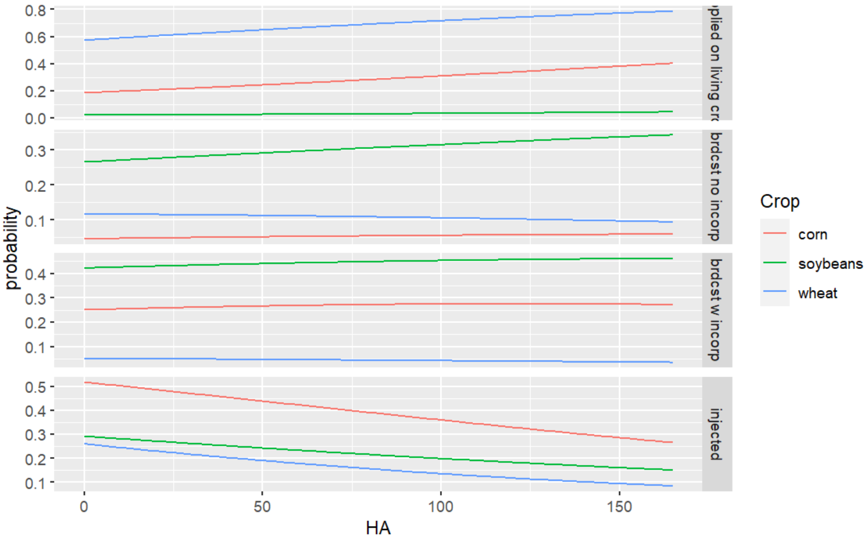


Figure 3.5.: Probability of using different fertilization method for chemical fertilizer for different crop types

4. Discussion

The main idea at first was to aggregate categorize the application methods to a binary outcome based on the easiness of application. I did aggregate "applied on living crop" and "broadcast with no incorporation" in one group and "broadcast no incorporation" and "injected" in another group. The results were not promising.

For the analysis of whether a field has been fertilized or not, the following analysis was carried out as well; since the results did not improve compared to single predictor ones, I did not include them in the results section.

$$Model\texttt{<-}Logit(HasFertilizer \sim HA + Crop, data = fertilizer)$$

$$Model\texttt{<-}Logit(HasFertilizer \sim HA + HA : log(HA), data = fertilizer)$$

The two figures produced for the probability of different methods used for manure or chemical fertilizer application show how different these two fertilizers are applied. Apparently, other factors are involved in decision-making rather than the field size or crop type.

In the logistic regression section, 3.1 there is an incomplete s-shaped curve. I tried to understand what is the reason for this, I have some assumptions, and I listed them based on their importance:

- The predictor value, area of the field, is not a significant predictor. There is a different range of field sizes for both outcomes.
- The ratio of the data for two outcomes is not appropriately distributed. The ratio of the fertilized to the not fertilized incidents is 9002:1584.

It seems that having a high percentage of precision in the confusion matrices shown in figure 3.1 is because most of the data available in the "fertilizer" dataset has been fertilized regardless of field size(9002:1584 ratio.)

5. Conclusion

The main hope for working on both manure and fertilizer dataset was to see if there is a correlation between the methods used for applying manure application versus chemical fertilizer, since there are other survey data available for other watersheds around Lake Erie that entail manure application methods but not include chemical fertilizer application. If there was a correlation between the chosen methods for manure and chemical fertilizer, the model could be used for predicting the methods.

Most of the research on this matter considers the farmers' attributes as important predictors in this matter. I will explore hierarchical logistic regression models or multilevel regression models to consider farmers' and fields' attributes at different levels to improve the results.

Bibliography

[1] G. C. Filson, S. Sethuratnam, B. Adekunle, and P. Lamba, "Beneficial management practice adoption in five southern ontario watersheds," *Journal of Sustainable Agriculture*, vol. 33, no. 2, pp. 229–252, 2009. DOI: 10.1080/10440040802587421. eprint: https://doi.org/10.1080/10440040802587421. [Online]. Available: https://doi.org/10.1080/10440040802587421.

[2] H. Liu, W. Zhang, E. Irwin, J. Kast, N. Aloysius, J. Martin, and M. Kalcic, "Best management practices and nutrient reduction: An integrated economic-hydrologic model of the western lake erie basin," *Land Economics*, vol. 96, no. 4, pp. 510–530, Nov. 2020. DOI: 10.3368/wple.96.4.510. [Online]. Available: https://doi.org/10.3368/wple.96.4.510.

[3] *Faq: How do i interpret odds ratios in logistic regression?* https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/, Accessed: 2022-07-01.

[4] P. S. U. David Gerbing The School of Business. (2022). Lessr: Less code, more results, [Online]. Available: https://cran.r-project.org/web/packages/lessR/index.html.

[5] *Interpreting the output of a logistic regression model*, https://rpubs.com/raoulbia/interpreting_glm_logistic_regression_output, Accessed: 2022-07-01.

[6] *Logit regression, r data analysis examples*, https://stats.oarc.ucla.edu/r/dae/logit-regression/, Accessed: 2022-07-01.

[7] Wikipedia contributors, *Plagiarism — Wikipedia, the free encyclopedia*, [Online; accessed 18-July-2022], 2004. [Online]. Available: https://en.wikipedia.org/wiki/Multinomial_logistic_regression.

[8] *Multinomial logistic regression, r data analysis examples*, https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/, Accessed: 2022-07-01.

[9] J. B. Hadley Wickham. (2022). Readxl: Read excel files, [Online]. Available: https://cloud.r-project.org/web/packages/readxl/index.html.

[10] W. V. Brian Ripley. (2022). Nnet: Feed-forward neural networks and multinomial log-linear models, [Online]. Available: https://cran.r-project.org/web/packages/nnet/index.html.

[11] R. Hadley Wickham. (2021). Forcats: Tools for working with categorical variables (factors), [Online]. Available: `https://cran.r-project.org/web/packages/forcats/index.html`.

[12] W. C. Hadley Wickham. (2022). Ggplot2: Create elegant data visualisations using the grammar of graphics, [Online]. Available: `https://cloud.r-project.org/web/packages/ggplot2/index.html`.

A. Appendix

The Github link to the project and the source code.

Github - R Project

Link to the presentation file: Presentation - R Project