International Journal of Innovation in Science and Mathematics
Volume 3, Issue 6, ISSN (Online): 2347–9051

Comparison of Logistic regression model, Neural network model and Decision tree model on an Epidemiological Study

***Renhao Jin, Fang Yan, and Tao Liu**
School of Information, Beijing Wuzi University, Beijing, China
***Correspondent Author's Email: Renhao.Jin@outlook.com**

***Abstract* – This paper applies three predictive models (logistic regression model, neural network model, and decision tree model) to an epidemiological study. The binary target variable is whether the herd is in the restricted status, which is defined by whether any bovine tuberculosis (bTB) reactor is detected in the herd. To compare the fitting performance of the three models, the observations are divided into three parts of Training data set (50%), Validation data set (30%), and Test data set (20%). The model performances are mainly based on the results of test dataset, and the decision tree model is the best model on this study. By analysis on the lift charts on test data set, the built decision tree model can be used to enhance practice efficiency.**

***Keywords* – Logistic regression model, Neural network model, Decision Tree model, Bovine tuberculosis, Spearman's rank correlation, Lift chart.**

I. INTRODUCTION

Numerous tools are available for developing predictive models. Some use statistical methods such as linear regression and logistic regression model, and others use non-statistical methods, such as neural networks, decision tree model, and even support vector machine. Much debate range about which are the best methods [1]. In this paper, the performance of three wildly used models: logistic regression model, neural network model and decision tree model are compared based on an epidemiological study of bovine tuberculosis (bTB) in cattle herds, which has a binary target variable indicating whether a herd had a bTB occurrence.

Logistic regression is firstly developed by statistician D.R. Cox in 1958 as a statistical method, and after that it is used widely in many fields, including the medical and social sciences. Unlike regression model, which is used to model or predict a continuous target variable, the logistic model is applied for binary target variable [2]. Recent years as the big data society comes, Logistic regression model is also extensively used in many data mining applications, such as, credit risk models in banking industry, customer preference models in retails, and even segment of customers in all areas of business. For example, it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Logistic regression is a direct probability model, which is also called as logit regression or logit model.

Logistic regression measures the relationship between the binary dependent variable and one or more independent variables (explanation variables), which are usually (but not necessarily) continuous, by estimating probabilities. More detailed, for a binary dependent variable y with value 0 or 1, corresponding to the status of an observation, the logit model is written as

$$\text{logit}(\text{Ey}) = \beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 \qquad (1)$$

where Ey is the expect value of y, i.e., the probability of y=1, logit(Ey) = log(Ey/((1-Ey)), and the part of $\beta_0+\beta_1 \text{X}_1+\beta_2 \text{X}_2$ is the linear part of independent variables. The equation (1) is also can be transformed to be

$$\text{Ey} = \frac{\exp(\beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2)}{1 + \exp(\beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2)} \qquad (2)$$

The important feature of the logit function is that it produces values that lie only between 0 and 1, just as the probability of response and non-response dependent variables, as shown in equation (2).

The approach to fit logit model is often based on maximum likelihood method, as logit model predicts probabilities rather than just classes. For each observation with explanation variables marked as X_i and let $\text{Ey}_i = p(Xi)$. The likelihood function for n observations can be written as

$$L(\beta) = \prod_{i=1}^{n} p(x_i)^{yi} (1 - p(x_i))^{1-yi} .$$

With the procedure of maximum likelihood estimation, the parameters' estimates and their variances all can be obtained from likelihood method [1,2].

Neural network model is also regarded as artificial neural networks, and it is widely used in a lot of fields as a statistical learning model. The neural model is inspired by the biological neural net, and it is a predictive method with higher precision and larger computation costs comparing with logistic regression model and decision tree model. Neural network models estimate weights and functions in the network depending on a large number of inputs and outputs. The flow chart of neural model are shown in Figure 1 and Figure 2. In the figure 1, the input and output are similar to the independent and dependent variable in regression models respectively. The hidden layer is the unique part of neural model, and it may have several hidden layers in a neural model. Each circle in the hidden layer is called a hidden layer node. In general, one hidden layer is adequate for the estimation precision. More hidden layers may increase the prediction precision but the computation cost increases exponentially [1].

The tree model is a simple but powerful form of multiple variable analysis. It provides unique capabilities to supplement, complement and substitute for a variety of data mining tools and techniques, such as linear regression model, logistic regression model, and neural network model [1,3,4]. A tree model represents a hierarchical segmentation of data. The original segment is the entire data set, and it is called the root node of the tree. The original segment is first portioned into two or more segments by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of an input for that observation. In a similar fashion, each resulting segment is further portioned into sub-segments; each sub-segment is further

portioned into more sub-segments, and so on. This process continues until no more portioning is possible. This process of segmenting is called recursive portioning, and it results in a hierarchy of segments within segments. This hierarchy is called a tree, and each segment or sub-segment is called a node [5]. For continuous target, the tree model is called regression tree, and for categorical target it is called decision tree model. Once the relationship is extracted, then one or more decision rules that describe the relationship between inputs and targets can be derived. Rules can be selected and used to display the tree, which provides a means to visually examine and describe the tree-like network of relationships that characterize the input and target values [6].

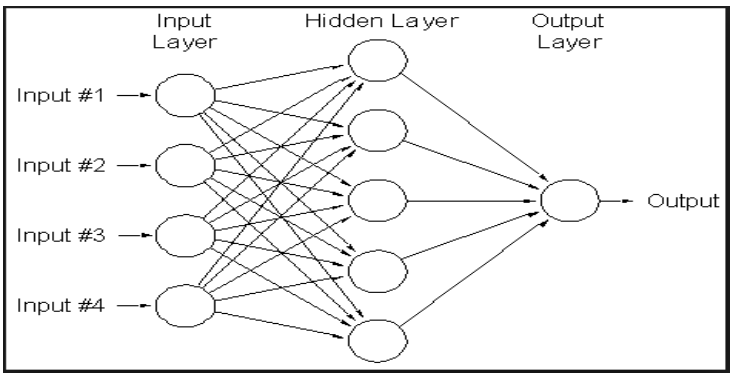

Figure 1. The general flow chart of neural network model with input layer, hidden layer and output layer.

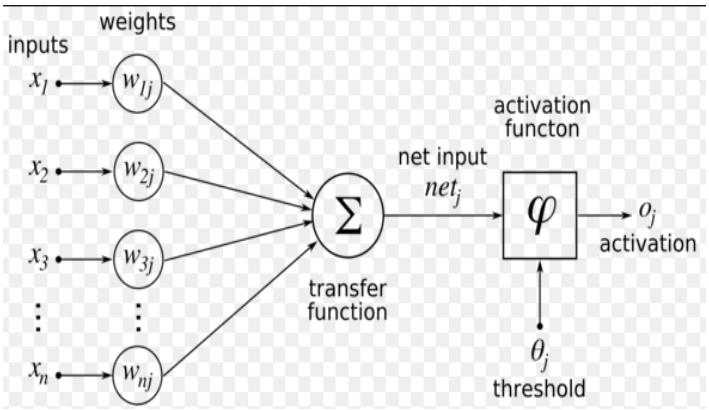

Figure 2. The mathematical functions in the link of all inputs, a hidden layer node and output in a neural network model.

The epidemiological study in this paper is based on aggregated bovine tuberculosis (bTB) data in cattle herds from 2010 to 2014, together with well-established risk factors in the area known as West Wicklow, in the east of Ireland. The target variable in this study is binary, and thus the logit model, neural network model and decision model could be used for the modelling.

Bovine tuberculosis (bTB), caused by infection with Mycobacterium bovis [7], affects approximately 0.3% of cattle annually in Ireland [8, 9]. This has major financial implications both for the farmer whose herd is restricted from trading and cattle slaughtered, and for the exchequer that compensates the farmer and implements measures to control the disease. Data for the bTB study were obtained from three sources: herd data from the Irish national databases of bTB testing herd and animal history (Animal Health Computer System, AHCS); land usage from Herdfinder, a unique multi-layered purpose built spatial mapping system whereby farms shapes submitted by farmers to DAFM and weather data from Met Éireann, all for West Wicklow. Both AHCS and Herdfinder databases use the same herd ID number so that farm, geographic location and testing data may be linked [9,10]. The spatial distribution of herds and rainfall stations, and the study areas is shown in Figure 3.

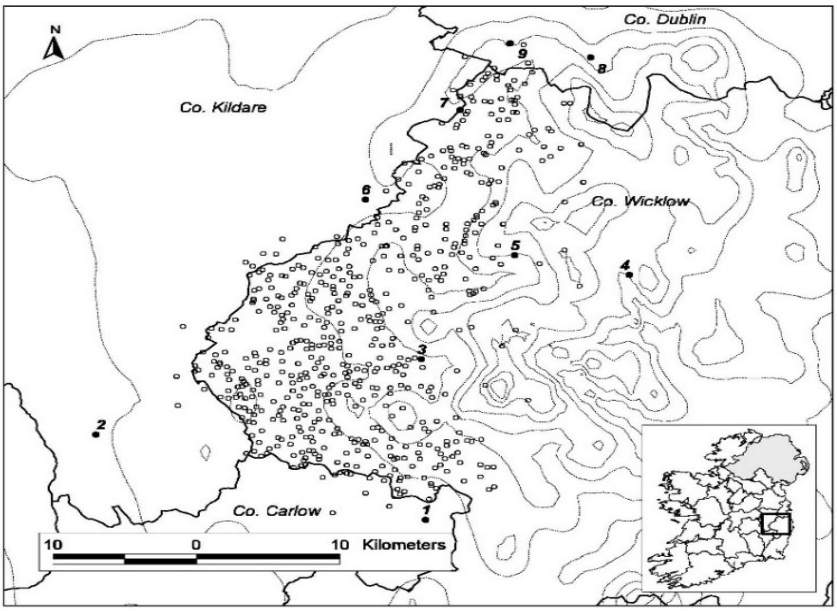

Figure 3. The spatial distribution of herd observations and rainfall stations. The herd observations and rainfall stations are marked as grey points and black points respectively. The dashed lines in the map are contour lines with 100 meters intervals, i.e. joined points are equal to 100 meters height above the sea level.

II. MODELING

Logistic regression model, neural network model and decision tree model are built to predict bTB incidence in cattle herds based on the potential risk factors (explanatory variables) from 2010 to 2014. The herd target variable is binary, indicating whether any bTB reactor is detected in the herd, and the herd with target value 1 is with restriction status. So, the response variable Y_{ij} is binary: restriction status of the ith herd in year j (1=restricted, 0=not restricted), where $j = 1, \cdots, 5$ denoting the years 2010 to 2014. There are more than 30 explanatory variables and it is unreasonable to put all them in model building. The associations between the response variable Y and each explanatory variable in a univariate analysis are firstly examined using Spearman's rank correlation coefficient. Many explanatory variables were skewed and outliers were present and Spearman's rank correlation was chosen as it is not sensitive to outliers, also Spearman's correlation is more suitable for categorical variables. Explanatory variables are considered for inclusion in the modelling if an association significant at the 0.1 level was found from the univariate analysis [11].

To develop a tree model and a neural network model, two data sets are needed. The first is training data for training the model, and the second is validation data for fine-tuning the model. In a statistical way, validation data is no need to fit a logistic model, but here the validate date are also used to choose parameters in logistic model to match the other two models. A third data set (Test data) is taken for an independent assessment of the three models. There are no theoretical rules for allocation the percentage of the model data set [1]. In this study, the partition of Training data set (50%), Validation data set (30%), and Test data set (20%) are used. Reserving more data for the training data generally results in more stable parameter estimate. The training data set is used for initial model fitting, and the validation data set is used to evaluate the models and choose the best parameters in the iterations on training data but assessed on validate data. The process on the validate data is to prevent the models from over-fitting problems. Since both the Training and Validation data sets are used for parameter estimation and parameter selection, respectively, an additional holdout data set is required for an independent assessment of the models. The Test data set is set aside for this purpose and it is used for an independent assessment of the three final models [1].

In the model fitting of logistic model and neural network model, the following Bernoulli error function is used:

$$E = -2\sum_{i,j}\{y_{ij}\ln\frac{\pi_{ij}}{y_{ij}} + (1-y_{ij})\ln\frac{1-\pi_{ij}}{1-y_{ij}}\} \quad (3)$$

Each iteration yields a set of weights, and each set of weight defines a model. Validation data set are used to choose the models defined by training data. The average squared Bernoulli error are set to be model selection criterion, and the algorithm selects the set of weights that results in the smallest average squared Bernoulli error where the Bernoulli error is calculated from the Validation data set.

For decision tree model, in the computation on training data set, in order to split any segment or sub-segment of the data set at a node, the worth of all candidate splitting values for each input are calculated, and the split with the highest

worth are chosen. As the target variable in this study is binary, the p-value of the Pearson Chi-Square test are used to determine the best split. The threshold significant level is set to be 0.2 to control the tree growth, which means a node stops splitting to sub-node if p-values of Chi-Square of all inputs are all larger than this threshold value. Having grown the largest possible tree under the rules described in training process, the validate data are used to prune the maximal tree to the right size. The assessment measure is set to be misclassification, and the misclassification rate is defined as 1 minus validation accuracy. The sub-tree with the minimum misclassification rate on the validation data is chosen to be final tree.

The model fitting results of these three models are compared on test data, and the cumulative lift is taken as the criterion for model comparisons. All the computations are done by SAS® 9.4

III. Results

From 2010 to 2014, there were 609 distinct herds in the study, giving 2666 observations. Table 1 presents the number of herds and the percentage restricted on an annual basis, and the total herds and percentage restricted for each year keep stable and are around 540 and 4% respectively. In the univariate analysis, herd bTB restriction status was significantly associated with 16 explanatory variables (Table 2). The remaining variables which were not significantly associated herd bTB restriction status are deleted from next model fitting.

Table 1. Number of herds and percentage of these herds with confirmed restrictions for tuberculosis in West Wicklow, Ireland from 2010 to 2014. A herd was considered restricted if any bTB reactor was found on any bTB test in the year.

`	Total herds	Number of restricted herds	Percentage restricted*
2010	555	25	4.5%
2011	550	22	4%
2012	530	17	3.2%
2013	517	29	5.6%
2014	514	29	5.6%

*Percentage restricted= Number of herds restricted/ Total number of herds.

Table 2. Spearman's rank correlation between explanatory variables and herd bTB restriction status (1=restricted, 0=not restricted). The variables with p value<0.1 are listed in the table. The four herd bTB history variables were all binary indicating whether bTB reactors were found in the herd in the past. Herd bTB history 1, 2, 3 denoted the herd bTB status for the previous 1, 2, 3 years, respectively while herd bTB history of past 3 years denotes whether bTB reactors were found in any of the past 3 years. A1, A2, A3 and P1, P2, P3 are the amplitude and the phase of monthly variables respectively.

Explanatory variables	Spearman's correlation coefficient	P value
Herd size	0.14	<.0001
Presence /absence of commonage	0.03	0.08
Total farm area	0.11	<.0001
Total farm perimeter	0.12	<.0001
Herd bTB history 1	0.09	<.0001
Herd bTB history 2	0.09	<.0001
Herd bTB history 3	0.08	<.0001
Herd bTB history of past 3 years	0.12	<.0001
Annual total rainfall	0.05	0.01
Annual max monthly rainfall	0.04	0.03
Annual mean monthly temperature	-0.04	0.04
Temperature.A3	0.04	0.04
Annual mean monthly VPD	-0.04	0.05
No. of bought	0.13	<.0001
VPD.A2	-0.04	0.03
VPD.P3	0.04	0.05

Table 3. Estimates from the best fitting logit model, explaining herd bTB occurrence in West Wicklow, Ireland from 2010 to 2014.

Fixed effect	Estimate	95% C.I.		p Value
		Lower	Upper	
Intercept	-5.606	-6.863	-4.349	<0.0001*
Log (herd size)	0.631	0.432	0.83	<0.0001*
Annual total rainfall	0.077	0.0185	0.135	0.02
Presence versus absence of herd bTB in past 3 years	0.841	0.42	1.261	<0.0001*
presence versus absence of commonage	0.397	-0.056	0.85	0.07

C.I., confidence interval; *p value<0.05.

Explanatory variables are considered for inclusion in the three models if an association significant at the 0.1 level was found from the univariate analysis. Tables 3 shows the results of the best fitting logit model (stepwise procedure are used to further variable selection). In the final model, herd bTB occurrence was positively associated with log (herd size), annual total rainfall, herd bTB history of past three years, and presence /absence of commonage. The final fitted decision tree model is shown in Figure 4, the 5 variables (Herd size, total farm area, Annual max monthly rainfall, Herd bTB history of past 3 years and No. of bought) are determined to be splitting nodes in the decision tree model. It concludes that these variables are more relative to herd restricted status. As the final tree are pruned by validate data, the statistics calculated by validate data are shown in the right column on every node text panel. In node ID 15 and 16, the herd bTB rates on validate data are 0, while in node ID 10, the herd bTB rate is relative high to be 42.86%. In other leaf nodes, the herd bTB rates on validate data all turn to larger or smaller than overall herd bTB rate. Following the results of the decision tree model, the accurate inference of herd bTB incidence can be obtained. However, the neural network model fitting parameters are not shown in this paper, as the complex of is model results.

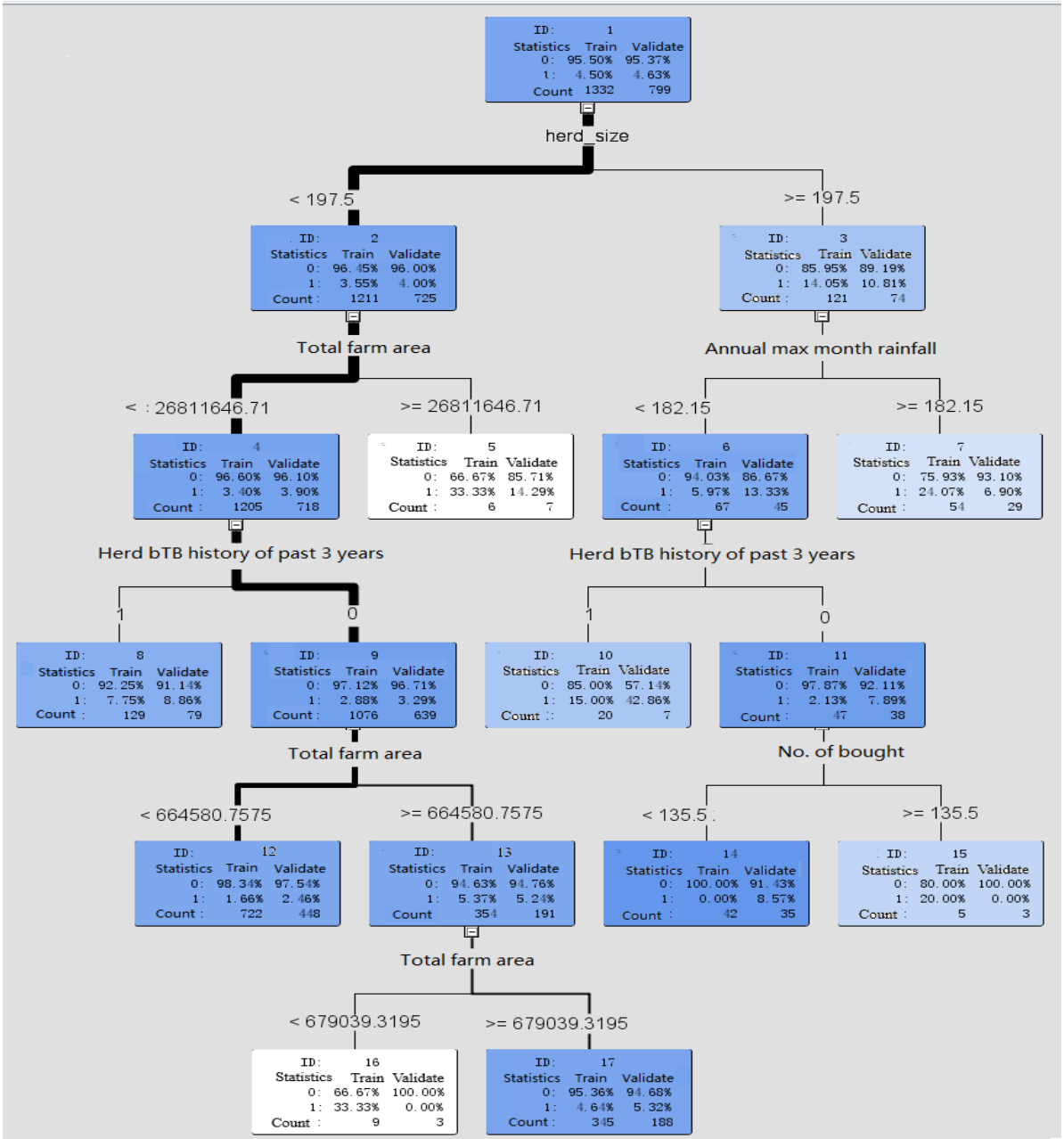

Figure 4. The final Decision Tree model built on training and validate data set.

In order to assess the predictive performances of logistic regression model, decision tree model and neural network model, the lift charts for the Training, Validation, and Test data sets are shown in Figure 5. The lift capture rates calculated from the Test data set are used for evaluating the models or comparing the models because the Test data set is not used in training or fine-tuning the model [1]. To create lift chart, the model is used to calculate the probability of

getting herd restricted status for each observation, then the observations are sorted descending by their probability. Then it divides the data set into 20 equal segments called Percentiles. Since the percentiles are created from the sorted data set based on the computed probabilities, the first percentile (called the top percentile) has the herds with the highest mean probability of bTB reactors. The lift in a given percentile is the actual observed herd bTB rate in that percentile divided by the overall actual herd restricted rate.

It can be seen from the Figure 5 that on the train data the tree model and neural network model generally have the highest lift values; on the validate data the logistic regression model generally has the highest lift value; but on the test dataset the decision tree model and neural network model generally have the highest lift values. As noted above, the model performances are mainly based on the results of test dataset, thus the decision tree model and neural network model are the best models on this study. However, decision tree model is more readable comparing with the other two models, so decision tree model is preferred in the case of this paper. More specifically on the result of decision model on the test data, in the first 5% of observations, the herd restricted rate is 18.8% comparing with 4.7% of overall restricted rate. The fitted decision tree model has a high lift value on test data set, and it can be used to enhance work efficiency. For example, in a prevention project of herd bTB, based on time and economic consideration, the Irish government may not examine all the herds in the country. Instead, if they only select 5% of herds and detect the bTB incidence, with the decision tree model, they could select 4 times herds with bTB reactors more than by random select, which is very useful for the prevention project

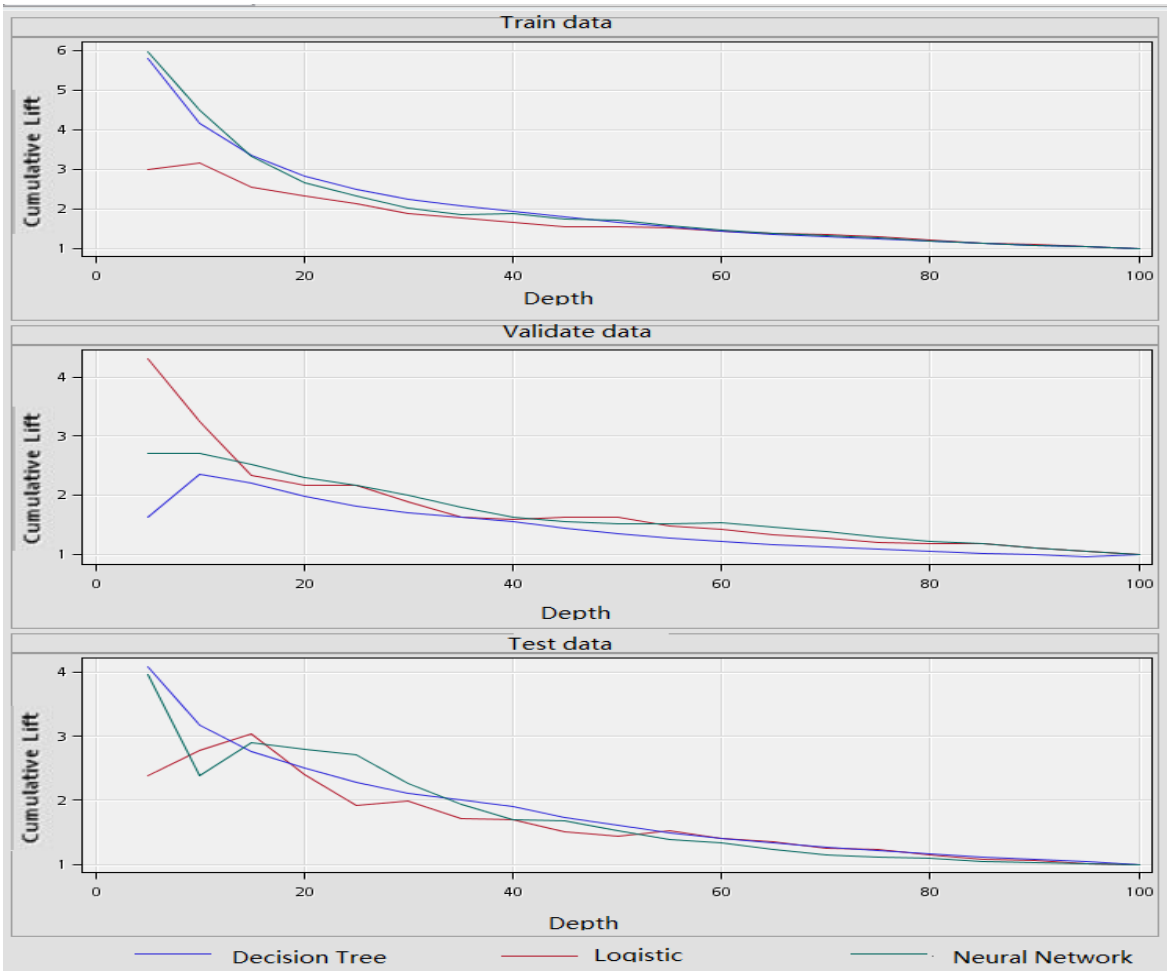

Figure 5. The cumulative lift charts of fitting results of Decision Tree model, logistic model and neural network model on Training, Validate and Test data set.

IV. Conclusion and Discussion

This paper applies three predictive models (logistic regression model, neural network model, and decision tree model) on an epidemiological study. The binary target variable is whether the herd is in the restricted status. To compare the performance of the three model, the observations are divided into three part of Training data set (50%), Validation data set (30%), and Test data set (20%). The training data set is used for initial model fitting, and the validation data set is used to evaluate the models and choose the best parameters in the iterations. The test data set is set aside for purpose of an independent assessment of the three final models. The model performances are mainly based on the results of test data; as shown in Figure 6, decision tree model and neural network model perform well on test data,

and tree model is easy to understand and apply, so it is the preferred model in this paper. By the decision tree model results, government could select more bTB reactors than by random select in a sampling survey, which is very useful for the bTB prevention project.

In the general practices of data mining, neural network model often has the best model fits but with bad interpretation, while tree model is often weak in fitting but strong in interpretation. The logit model takes the benefits of the above two models, with reasonable fitting and explanatory. However, in this paper, logit model works worst comparing with the other models, and it might be caused by the data uncertainty. If more fixed and random effects added, and considering more correlation structures among the target variables, logit model might show higher fitting performance. However, it is out of the topic of this paper. Also, these settings of logit model are more complicated, and difficult for non-statistical researchers. While, neural network model and tree model are easier to handle, and that is why these two models are wildly used in data practices. In all, neural network model, tree model and logit model all have their own advantages, the choice of which to modeling is highly dependent on the data and the knowledge structure of researchers.

ACKNOWLEDGMENT

The first author wishes to thank to DAFM for providing all the herd data, and Met Éireann for providing weather data. This paper is funded by the project of National Natural Science Fund, Logistics distribution of artificial order picking random process model analysis and research (Project number: 71371033); and funded by intelligent logistics system Beijing Key Laboratory (No.BZ0211); and funded by scientific-research bases---Science & Technology Innovation Platform---Modern logistics information and control technology research (Project number: PXM2015_014214_000001); and funded by University Cultivation Fund Project of 2014-Research on Congestion Model and algorithm of picking system in distribution center (0541502703).

REFERENCES

[1] Sarma, K.S. (2013). *Predictive Modeling with SAS Enterprise Miner Practical Solutions for Business Applications Second Edition*. NC: SAS Institute Inc, Cary.

[2] Collett, D., (2002). *Modelling binary data.* Chapman & Hall/CRC, London.

[3] Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* New York: Oxford University Press.

[4] Afifi, A.A and Clark (2004). *Computer Aided Multivariate Analysis,* Virginia: CRC Press.

[5] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

[6] Freedman, D.A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

[7] Gordejo, R.F.J., Vermeersch, J.P. (2006). *Towards eradication of bovine tuberculosis in the European Union.* European Union Veterinary Microbiology 112, 101-109.

[8] Jin, R., Good, M., More, S. J., Sweeney, C., Mcgrath, G., & Kelly, G. E. (2013). *An association between rainfall and bovine tb in wicklow, ireland.* Veterinary Record, 173(18), 452.

[9] Griffin, J.M., Williams, D.H., Kelly, G.E., Clegg, T.A., O'Boyle, I., Collins, J.D., More, S.J. (2005). *The impact of badger removal on the control of tuberculosis in cattle herds in Ireland.* Preventive Veterinary Medicine 67, 237–266.

[10] Kelly, G. E., & More, S. J. (2011). *Spatial clustering of tb-infected cattle herds prior to and following proactive badger removal.* Epidemiology & Infection, 139(8), 1220-1229.

[11] Ma, E., Lam, T., Wong, C., Chuang, S.K. (2010). *Is hand, foot and mouth disease associated with meteorological parameters.* Epidemiology and Infection 138, 1779-1788.

Renhao Jin
Lecturer in Statistics, School of Information, Beijing Wuzi University, Beijing, China.
Current research areas:
Spatial and Spatio-temporal statistics, Data Mining and Statistical Modelling.
Email: Renhao.Jin@outlook.com