Digital Innovations & Contemporary Research in SCIENCE, ENGINEERING & TECHNOLOGY
Vol. 10. No. 1, 2022

Article Citation Format
Onoghojobi, B., Fadugba, S.E., Kekana, M., Ozioko, A. L & Audu, S. O. (2022): Decision Tree Algorithm in Fuzzy Logistic Regression Analysis.. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology.
Vol. 9, No. 1. Pp 110-1118
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V10N1P10

Article Progress Time Stamps
Article Type: Research Article
Manuscript Received:19th Dec, 2021
Review Type: Blind
Final Acceptance: 16th February, 2022

Decision Tree Algorithm in Fuzzy Logistic Regression Analysis

Onoghojobi, B., Fadugba, S.E., Kekana, M., Ozioko, A. L & Audu, S. O.
[1]Department of Statistics, Federal University Lokoja, Nigeria
[2]Dept of Physical Sciences, Mathematics Prog, Landmark University, Omu-Aran, Nigeria
[3]Dep of Mathematics and Statistics, Tshwane University of Tech, Pretoria, South Africa
[4]Department of Mathematics, Federal University Lokoja, Lokoja, Kogi State Nigeria
E-maills: benson.onoghojobi@fulokoja.edu.ng fadugba.sunday@lmu.edu.ng kekanamc@tut.ac.za arinze.luke@fulokoja.edu.ng audusamuel01@yahoo.com

ABSTRACT

A novel approach to fitting Fuzzy Logistic Regression model which incorpo- rates the machine learning algorithms is derived. The properties of both the Decision Tree Classifier which is a type of the supervised machine learning algorithm and the method of estimation of its parameters are discussed. It identifies the most informative variables for the Fuzzy Logistic Regression model, reducing the risk of over fitting and enhancing the interpretabil- ity of the results. It is applied to a fuzzy benchmark datasets of student with bivariate dependent pass/fail variable. The results demonstrate that our method outperforms the traditional Fuzzy Logistic Regression models in terms of predictive accuracy, with improved classification performance and reduced prediction errors. It is proposed for situations where there is need for reduced dependence on expert knowledge or automating the parameter selection process in various fields, such as medical diagnosis, credit scoring, and risk analysis.

Keywords: Fuzzy Logistic, Tree Classifier, Supervised Machine Learning, Bivariate Dependent, Reduced Prediction

1. INTRODUCTION

The regression model is a widely used tool in fields like economics, medicine, and engineering, with applications in fields like finance, healthcare, and decision support systems [1][2]. It is a variation of classical regression analysis, with two types: Tanaka linear programming approach and fuzzy least square approach [3][4]. The fuzzy set theory, developed by Iranian Azerbaijani mathematician Lofti Zadeh [5], is used to represent data with vagueness, allowing for mathematical expression of non-precise data. Fuzzy logic is a form of many-valued logic, with the truth value of a variable being any real number between 0 and 1.

Tanaka et al introduced fuzzy linear regression (FLR) in 1982 to model causal relationships in systems with ambiguity or human judgment [3]. Fuzzy Logistic Regression is a modeling approach that represents ambiguity or fuzziness in data, particularly suitable for situations with inherent fuzziness or uncertainty [7][8]. It has been used in finance, healthcare, and decision support systems for modeling stock market volatility, disease diagnosis and prognosis, and predicting patient outcomes [9][10].

Fuzzy logistic regression is an extension of traditional logistic regression that incorporates fuzzy set theory to handle uncertain or imprecise data [11][12]. It is particularly useful for binary dependent variables with vagueness and ambiguity. Fuzzy membership functions represent the imprecise nature of data, assigning membership degrees to data points. This allows for a more nuanced and flexible modeling approach compared to standard logistic regression. Fuzzy Logistic Regression involves data analysis, classification, and churn prediction [13].

However, if the fuzzy data are binary, we need to consider a fuzzy logic based approach. The main contribution of this paper is to derive and implement a machine learning algorithm to fit the Fuzzy Logistic Regression model. This paper contains the following sections; section 2 is the derivation of the linearized fuzzy logistic regression model. In section 3, we have machine learning algorithm for logistic regression model. Finally, the outcome and findings were addressed in section 4.

Derivation Of The Linearized Fuzzy Logistic Regression Model

Human situations often involve categorizing data in ordinal textual form, but statistical methods struggle with vagueness and uncertainty. To address this, fuzzy logic and classical logistic regression are combined, creating the "Fuzzy Logistic Regression Model." [14]. Fuzzy Logistic Regression Analysis integrates fuzzy regression and logistic regression, modeling dependent vari- ables in imprecise data without logistic regression assumptions [12].

In vague binary dependent observations, the probability of an observation be- longing to category 1 (p = P(y=1)) and the odd ratio (p/(1-p)) and odd ratio is not directly calculated. Instead, observations are compared to confirmed criteria for a specific class, and Logistic Regression is used to model independent variables against calculated probabilistic odds [15]. This gives rise to what the researcher calls the "Linearized Fuzzy Logistic Regression Model". Whenever a categorical variable is the dependent variable, the Logistic Regression becomes the most appropriate model. While the linear regression estimates the parameters of the model based on the numerical contribution of each of the explanatory variables, the Logistic Regression considers the possibility of occurrence in the estimation of its parameters.

The Logistic Regression Model is of the form

$$y_i = \frac{1}{(a\beta^{x_i} + g)}$$

Where y_i is the dependent categorical variable, x_i represents the independent or explanatory variables that are crisp values or fuzzy numbers and β_i are the constants parameters and g is the error term.

$$y_i = (a\beta^{x_i} + g)^{-1}$$

We apply the triangular fuzzy number (TFN) given by the central tendency and its left and right spread (l, c, u)

$$(y_l, y_c, y_u) = (a\beta^{((X_c - S_x),(X_c + S_x))} + g)^{-1}$$

Where y_l is the lower bound of y, y_u is the upper bound of y and y_c is the central tendency of y. x_l is lower bound of x, x_u is the upper bound of x, x_c is the central tendency of x We are now going to apply the assumption of the symmetric Triangular Fuzzy Number (sTFN) where the left spread and right spread are equal. We therefore denote the spread of y as Sy and the spread of x as Sx. Hence the sTFN is represented in interval form as

$$((Y_c - S_y), (Y_c + S_y)) = (a\beta^{((x_c - S_x),(X_c + S_x))} + g)^{-1}$$

Let $g^* = g^{-1}$ since the inverse of a constant is a constant

$$((Y_c - S_y), (Y_c + S_y)) = (a\beta^{((X_c - S_x),(X_c + S_x))})^{-1} + g^*$$

the logarithm of both sides of the above expression, we have;

$$\log((Y_c - S_y), (Y_c + S_y)) = \log(a\beta^{((x_c - S_x),(X_c + S_x))})^{-1} + g^*$$

Applying the rules of logarithm,

$$\log((Y_c - S_y), (Y_c + S_y)) = -(\log(a\beta^{((x_c - S_x),(X_c + S_x))})) + g^*$$

$$\log((Y_c - S_y), (Y_c + S_y)) = -(\log(a) + \log\beta^{((x_c - S_x),(X_c + S_x))}) + g^*$$

$$\log((Y_c - S_y), (Y_c + S_y)) = -(\log(a) + ((X_c - S_x), (X_c + S_x))\log(\beta)) + g^*$$

$$\log((Y_c - S_y), (Y_c + S_y)) = -\log(a) - ((X_c - S_x), (X_c + S_x))\log(\beta) + g^*$$

Let $\log((Y_c - S_y), (Y_c + S_y)) = ((Y_c - S_y), (Y_c + S_y))^*$, $\log(a) = a^*$, $\log(\beta) = \beta^*$, and g^* as the error term, the Linearized Fuzzy Logistic Regression Model can then be written as;

$$((Y_c - S_y), (Y_c + S_y))^* = -a^* - ((X_c - S_x), (X_c + S_x))\beta^* + g^*$$

The fitting of any logistic model is for the purpose of control, prediction of the category based on the estimated probability value from the model when given a set of input variables. Hence, the model has to be at its optimum best to make veritable estimates. This research is driven at using a machine learning approach to fit the best fuzzy logistic regression model involves us- ing Decision Tree algorithms and techniques to automatically select the most suitable model.

Fuzzy logic: In fuzzy logic, variables are represented using fuzzy sets. A fuzzy set is a set of elements that have a degree of membership to the set. The degree of membership is a value between 0 and 1, where 0 represents no membership and 1 represents full membership.

Where

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Membership functions: Define the membership functions for each input variable. A membership function maps the input data to a degree of member- ship in the corresponding fuzzy set. Common membership functions include triangular (applied earlier), trapezoidal, and Gaussian.

2. MACHINE LEARNING ALGORITHM FOR LOGISTIC REGRESSION MODEL

The decision tree classifier is a widely used machine learning algorithm for classification and regression tasks, handling both numerical and categorical data [16]. It constructs a tree-like model of decisions and their consequences by recursively partitioning data based on input features. Splits are deter- mined using criteria like Gini impurity and information gain. The decision tree classifier can be combined with fuzzy logistic regression to improve performance and interpretability, introducing fuzzy concept for handling uncertainty and imprecision in noisy or ambiguous data [17]. Combining a decision tree classifier with fuzzy logistic regression allows for capturing complex patterns and interactions among input features, while fuzzy logistic regression provides probabilistic interpretations and handles uncertain data. The decision tree extracts derived features, which are then input to the fuzzy logistic regression model while learning fuzzy relationships be- tween derived features and the target variable. Combining decision tree and fuzzy logistic regression improves model performance and interpretability. The decision tree provides insights into the important features and their interactions, while the fuzzy logistic regression enhances the model's ability to handle uncertainty and provides a probabilistic view of the results [18].

In using machine learning algorithms, the researcher applies the maximum likelihood estimation process using the step wise estimation approach involv- ing the following steps

1. Data preprocessing involving the numerical coding of the categorical independent input variables
2. The dataset is divided into training set and test set
3. The logarithmic transformation of the binary dependent variable
4. The logistic regression model is fitted using all available variables
5. Evaluation of the significance of each input variable to the fitted logistic regression model
6. Repetition of steps (4) and (5) until an optimum model is fitted

The objective is to minimize the difference between the predicted probabilities (classification) and the actual values. The research utilized a virtual student pass/fail fuzzy dataset for model train- ing and evaluation. The dataset was divided into training and test groups, with 80% for training and 20% for testing. Decision Tree machine learning was used to optimize model performance on the test set. The fuzzy logistic regression model's performance is evaluated using metrics like accuracy and precision, with cross-validation techniques for optimal results. Excel software and Python programming along with its fuzzy packages and libraries, were the essential tools for achieving the goal.

3. NUMERICAL ILLUSTRATION

Eighty percent of the dataset was allocated into the training set and twenty percent into the test set, creating two separate datasets. Using both the conventional fuzzy logistic regression model and the decision tree optimized fuzzy logistic regression model, the fuzzy logistic regression models were fitted (i.e. trained) on the training dataset and in both cases, tested on the test dataset. The following results were obtained:

Central tendency of the Traditional Fuzzy Logistic Regression Model: Outcome = 8.78865858 - 1.19571818***HOS*** + 0.81013402***COS*** - 0.31596209***TM***

Central tendency of the Decision Tree Fuzzy Logistic Regression Model:
Outcome = -0.22744687 - 2.11124719***HOS*** + 2.11124731***COS***

Where **HOS** is the hours of study
COS is Course of Study **TM**
is Teaching Method

The first model above is the result of the fitted traditional Fuzzy Logistic Regression model while the second is that of the Decision Tree Optimized Fuzzy Logistic Regression model. The models were used in the estimation of the probability value and classification into pass or fail using a bench mark probability of (p = 0.5), the predicted probability value of the models were close estimates of the observed probability values.

The observed values

1011100001

These are the observed student success (1) and failure (0) outcomes in the examination. We then went ahead to classify the student outcome using both models. The classification of both models are as found below

Table 1: Confusion Matrix for Predicted Student Pass and Fail with Traditional Fuzzy Logistic Regression Model

	Predicted Outcome		
Observed Outcome	Fail	Pass	Total
Fail	3	2	5
Pass	1	4	5
Total	4	6	10

The predicted values by the Traditional Fuzzy Logistic Regression Model

1111110000

The predicted values by the Machine Learning Decision Tree Op- timized Model
1011100001

The contingency table 1 shows the observed classification and predicted clas- sification of the 10 observations in the test dataset shows the efficiency of the model traditional Fuzzy Regression Model with an accuracy of 70%. The contingency table 2 shows the observed classification and predicted clas- sification of the 10 observations in the test dataset shows the efficiency of the model using the Decision Tree Machine Learning Classifier Model with an accuracy of 100%.

Table 2: Confusion Matrix for Predicted Student Pass and Fail with Machine Learning Optimized Fuzzy Logistic Regression Model

	Predicted Outcome		
Observed Outcome	Fail	Pass	Total
Fail	5	0	5
Pass	0	5	5
Total	5	5	10

4. CONCLUSION

Logistic regression analysis relies on assumptions, but can be challenging due to data uncertainty. Fuzzy statistic theory, a combination of fuzzy set and classical methods, was used to address this issue. The Decision Tree algo- rithm was integrated into Fuzzy logistic regression analysis. This paper uses Fuzzy Logistic Regression based on an iterative machine learning algorithm. It uses Python's step-wise model fitting technique and a student pass/fail data set. The model's success is demonstrated using predictions on a test dataset. The Decision Tree Algorithm, an iterative machine learning decision tree, improves predictions and classification accuracy.

More so, the new model can identify and use only the most informative vari- ables in the dataset for the Fuzzy Logistic Regression model. This helps to reduce the risk of over fitting where insignificant variables are factored into the model, hence enhancing the interpretability and accuracy of the results. More so, the new model can identify and use only the most informative vari- ables in the dataset for the Fuzzy Logistic Regression model. This helps to reduce the risk of over fitting where insignificant variables are factored into the model, hence enhancing the interpretability and accuracy of the results.

REFERENCES

1. Harrell Jr., F. E. (2015). Regression Modeling Strategies: With Appli- cations to Linear Models, Logistic Regression, and Survival Analysis. Springer.
2. Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. United Kingdom: Wiley. 293
3. Cengiz Kahraman, Ahmet Beskese, Tunc Bozbura F. (2007) Fuzzy Ap- plications in Industrial Engineering. Germany: Springer Berlin Heidel- berg. 589
4. Tanaka, H. (1982). Linear regression analysis with uncertain data. IEEE Transactions on Systems, Man, and Cybernetics, 12(6), 903-907.
5. Tanaka, H. (1984). Fuzzy regression analysis based on linear program- ming. Fuzzy Sets and Systems, 13(1), 29-46.
6. Lee, H. and Tanaka, H. (1999) Fuzzy approximations with non-symmetric fuzzy parameters in fuzzy regression analysis. Journal of the Opera- tions Research Society Japan 42: 98-112.

7. Chang, Yun-Hsi & Ayyub, Bilal. (2001). Fuzzy regression methods – A comparative assessment. Fuzzy Sets and Systems. 119. 187-203. 10.1016/S0165-0114(99)00091-3.
8. Tseng, F. M., Lee, C. W., & Tzeng, G. H. (2008). Fuzzy regression model for forecasting stock market volatility. Expert Systems with Applications, 35(1-2), 93-98.
9. Tsipouras, M. G., Fotiadis, D. I., Konitsiotis, S., Tsalikakis, D. G., Vlachos, M., & Naka, K. K. (2016). Fuzzy modeling of disease diagnosis based on an extended set of fuzzy rules. Knowledge-Based Systems, 109, 247-259.
10. Pedrycz, W., & Gomide, F. (2007). Fuzzy systems engineering: Theory and practice. John Wiley & Sons.
11. Pourahmad, Saeedeh & Ayatollahi, Seyyed Mohammad Taghi & Taheri, S.Mahmoud. (2011). Fuzzy logistic regression: A new possibilistic model and its application in clinical vague status. Iranian Journal of Fuzzy Systems. 8.
12. ATALIK, G., & Senturk, S. (2018). A new approach for parameter es- timation in fuzzy logistic regression. Iranian Journal of Fuzzy Systems, 15(1), 91-102.
13. Gavrilova, M. L. (2009). Transactions on Computational Science VI. Germany: Springer Berlin Heidelberg.
14. Ahmadini, A. A. H. (2022). A novel technique for parameter estimation in intuitionistic fuzzy logistic regression model. Ain Shams Engineering Journal, 13(1), 101518.
15. Pourahmad, S., Ayatollahi, S. M. T., Taheri, S. M., & Agahi, Z. H. (2011). Fuzzy logistic regression based on the least squares approach with application in clinical studies. Computers & Mathematics with Applications, 62(9), 3353-3365.
16. Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.
17. Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non- linear decision-tree effects. European Journal of Operational Research, 297(3), 1178-1192.
18. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer
19. Nagar, P. and Srivastava, S. (2008) Adaptive fuzzy regression model for the prediction of dichotomous response variables using cancer data: a case study, Journal of Applied Mathematics, Statistics and Informat- ics, 4(2), 183-191.

APPENDIX

Python Code for Traditional Fuzzy Logistic Regression

```
# -*- coding: utf-8 -*- """
Spyder Editor
This is a temporary script file. """
#installing necessary packages #!pip
install scikit-fuzzy #!pip install numpy
#!pip install pandas #!pip
install sklearn
#importing packages import
numpy as np import pandas as pd
import skfuzzy as fuzz
from sklearn.model selection import train test split from
sklearn.metrics import accuracy score
from sklearn.linear model import LogisticRegression from
sklearn.metrics import confusion matrix
#importing dataset
data = pd.read csv('studat.csv') df =
pd.DataFrame(data)
# Convert categorical variables to numerical values df['COS'] =
pd.factorize(df['COS'])[0]
df['TM'] = pd.factorize(df['TM'])[0] df['Outcome'] =
pd.factorize(df['Outcome'])[0]
# Split the dataset into training and testing sets X =
df[['HOS', 'COS', 'TM']]
y = df['Outcome']
X train, X test, y train, y test = train test split(X, y, test size=0.2, ran-
dom state=42)
# Fit the fuzzy logistic regression model model =
LogisticRegression() model.fit(X train, y train)
# Get the coefficients and intercept coefficients =
model.coef
intercept = model.intercept
# Print the coefficients and intercept print("Coefficients:",
coefficients) print("Intercept:", intercept)
# Make predictions
y pred = model.predict(X test) print(y
pred)
# Calculate accuracy
accuracy = accuracy score(y test, y pred) print("Accuracy:", accuracy)
# Assuming you have the following variables:
# y test: the actual values of the target variable from the test set
# y pred: the predicted values of the target variable from the model
# Create the confusion matrix
cm = confusion matrix(y test, y pred)
# Print the confusion matrix
print("Confusion Matrix:") print(cm)
print(y test, y pred)
```

Python Code for Decision Tree Optimized Fuzzy Logistic Regres- sion Model

```
# -*- coding: utf-8 -*-
""" Created on Thu Jun 15 03:09:27 2023
@author: user """
import numpy as np import
pandas as pd import skfuzzy as
fuzz
from sklearn.model selection import train test split from
sklearn.tree import DecisionTreeClassifier
from sklearn.linear model import LogisticRegression from
sklearn.metrics import accuracy score
from sklearn.metrics import confusion matrix
#importing dataset
data = pd.read csv('studat.csv') df =
pd.DataFrame(data)
# Convert categorical variables to numerical values df['COS'] =
pd.factorize(df['COS'])[0]
df['TM'] = pd.factorize(df['TM'])[0] df['Outcome'] =
pd.factorize(df['Outcome'])[0]
# Split the dataset into training and testing sets X =
df[['HOS', 'COS', 'TM']]
y = df['Outcome']
X train, X test, y train, y test = train test split(X, y, test size=0.2, ran- dom
state=42)
# Fit the decision tree classifier decision tree =
DecisionTreeClassifier() decision tree.fit(X train,
y train)
# Extract features from the decision tree train tree features = decision tree.predict proba(X
train) test tree features = decision tree.predict proba(X test)
# Fit the fuzzy logistic regression model using decision tree features fuzzy logreg =
LogisticRegression()
fuzzy logreg.fit(train tree features, y train)
# Get the coefficients and intercept coefficients =
fuzzy logreg.coef intercept = fuzzy
logreg.intercept
# Print the coefficients and intercept
print("Coefficients:", coefficients) print("Intercept:",
intercept)
# Make predictions using the combined model
train predictions = fuzzy logreg.predict(train tree features) test
predictions = fuzzy logreg.predict(test tree features)
# Create the confusion matrix
cm = confusion matrix(y test, test predictions)
# Print the confusion matrix
print("Confusion Matrix:") print(cm)
# Calculate accuracy train accuracy = accuracy score(y train, trai predictions) test accuracy
= accuracy score(y test, test predictions)
print("Training Accuracy:", train accuracy) print("Testing
Accuracy:", test accuracy)
```