Logistic Regression Assignment

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

Problem Statement:

Census-income data plays the most important role in the democratic system of government, highly affecting the economic sectors. Census-related figures are used to allocate federal funding by the government to different states and localities. Census data is also used for post census residents estimates and predictions, economic and social science research, and many other such applications. Therefore, the importance of this data and its accurate predictions is very clear to us. The main aim is to increase awareness about how the income factor actually has an impact not only on the individual lives of citizens but also an effect on the nation and its betterment. You will have a look at the data pulled out from the 1994 Census bureau database, and try to find insights into how various features have an effect on the income of an individual.
The data contains approximately 32,000 observations with over 15 variables.
The strategy is to analyze the data and perform a predictive task of classification to predict whether an individual makes over 50K a year or less by using a logistic regression algorithm.

Column Names	Description
Age	Age of the individual
Workclass	department of the working individual
fnlwgt	Final weight of the individual
education	The education degree of the individual
education-num	Number of years of education
marital-status	Marital status of the individual
occupation	Occupation of the individual
relationship	Relation value
race	Ethnicity of the individual
sex	Female, Male
capital-gain	capital gain of the individual
capital-loss	capital loss of the individual

hours-per-week	number of working hours
native-country	The native country of the individual
Annual-Income	Annual income either >50K or <=50K

1. How many types of occupations do we have?
    a. 13
    b. 14
    c. 15
    d. 11
2. How many people are working as tech support and have an annual income greater than 50k?
    a. 278
    b. 389
    c. 289
    d. 934
3. How many total missing values are present in the dataset?
    a. 4262
    b. 5000
    c. 5349
    d. 4302
4. If there are missing values in the Marital Status column, which option among the following should be used for replacing the missing values:
    a. Mean
    b. Median
    c. Mode
    d. All of the above
5. How many people are having private work classes and are not from the United States of America?
    a. 2151
    b. 2300
    c. 2000
    d. 2190
6. How many people are either having Annual Income(last column) less than or equal to 50k or their working hours is greater than or equal to 40 hrs:
    a. 23008

b. 23448
c. 29505
d. 25903

7. Which of the following methods can you use for handling outliers?
   a. Interquartile Range(IQR) Method
   b. Z Score method
   c. Both of the above methods
   d. None of the above
8. Chi-square is used to analyze:
   a. Determine the relationship b/w the variables
   b. Compare observed results with expected results
   c. both a and b
   d. None of the above

9. What is VIF?

   a. It can detect multicollinearity
   b. If the VIF value is greater than 10, then there is no correlation between the independent variables
   c. It stands for Variance Impact Factor
   d. VIF is when there is no correlation between one predictor and the other predictors in a model.

10. What predict_proba will tell you?
   a. It will predict the class probabilities
   b. It will tell you the target value
   c. Both are correct
   d. None of the above
11. Logistic regression is useful for regression problems:
   a. True
   b. False
12. In logistic regression, if the predicted logit is 0, what's the transformed probability?
   a. 0.5
   b. 0.05
   c. Both of the above
   d. None of the above
13. Which variant of logistic regression is recommended when you have

a categorical dependent variable with more than two values?

a. Multiple Logistic regression
b. Multinomial logistic regression
c. Ordered logit regression
d. Poisson regression

Perform the following tasks for answering the remaining questions

- Rename the last column as Annual Income
- Remove the missing values from the dataset
- Change the labels of categorical data into numerical data using Label Encoder.
- Split the dataset into a train and test of proportions 70:30 and set the random state to 0.
- Build a Logistic Regression Model on the data.

Answer the following questions with the help of the above-created model.

14. What is the accuracy score of the above model?
    a. 0.60 to 0.70
    b. 0.40 to 0.60
    c. 0.70 to 0.85
    d. None of the above
15. What is the specificity of the above model?
    a. 0.20 to 0.30
    b. 0.30 to 0.40
    c. 0.50 to 0.60
    d. None of the above
16. What is the model's precision when the target is False?
    a. 0.60 to 0.70
    b. 0.40 to 0.60
    c. 0.70 to 0.80
    d. None of the above
17. What is the total support value from the above model?
    a. 9049
    b. 9032
    c. 10000
    d. 9847

18. What is the f1 score of the above model when the target is True?

a. 0.30 to 0.40
b. 0.40 to 0.50
c. 0.60 to 0.70
d. 0.90 to 0.99

19. How many records are correctly classified by the model?

a. 7173
b. 7043
c. 7000
d. None of the above