DEEP LOGIC

Advanced Python Programing Class

Logistic Regression Assignment

Instructions: *Try and answer all the following questions to the best of your ability, all by yourself. If you need any assistance, please reach out to me so I can help you decipher the fair details or interact with fellow classmates for group discussions. Your development environment of choice should be any IDE you would prefer, but the best IDE for this work would be Jupyter Notebook which will allow you to download and save the completed work as an HTML file which you will eventually share with me via your GitHub account. You also can use Google Colab and share your work and this would be ONLY good during the development stages.*

Background: In this assignment you will conduct binary logistic regression on a hypothetical Titanic data set, with survival as the target, utilizing machine learning methods. The variables for analysis are:

- Pclass: passenger class (1 = 1st, 2 = 2nd, and 3 = 3rd)
- SexNum: gender of the passengers
- Age: age of the passengers
- SibSp: number of siblings/spouses aboard
- Parch: number of parents/children aboard
- Fare: passenger fares
- Embarked: port of embarkation (C = Cherbourg, Q = Queenstown, and S = Southampton)
- Survived: 0 = No and 1 = Yes

Python: Using Python, load the data set and complete the work below. The variable Survived is the binary outcome variable.

1. Describe the dataset
2. Display the counts for the number nulls in the data & drop any missing data.
3. Create a sns counplot of the Survived column
4. Describe the age column
5. Create a scatterplot for age and survived
6. Create a FacetGridwith plot having col='Survived', row='Pclass'. Plot Age on the same grid & add a legend.
7. Select the features as df[['Pclass','Age','SibSp','Parch','Fare']] and the target is Survived
8. Use train_test_split to create the x & y training and testing sets
9. Using statsmodel to generate & fit a logistic regression model with the formula:
   formula = 'Survived ~ C(Pclass) + SexNum + Age + SipSp'
10. Display a summary of the model
11. Calculate the y predicted values with your logistic model
12. Round y_pred and save the values in y_pred.
13. Calculate the model residuals.
14. Create and display a confusion matrix using Pandas crosstab.
15. Plot the confusion matrix with a heatmap
16. Print the value for TN, TP, FN and FP
17. Calculate and print the accuracy score
18. Plot the ROC curve
19. Using your model, plot the distribution of the probability of survival with respect to the age of the passengers