1. Logistic Regression, Linear Regression or Linear Discriminant Analysis? Write your thoughts in one
page summary.

Logistic regression and LDA are both classifiers, but they use very different techniques. Logistic regression and linear regression actually use pretty similar techniques, but they are used for different purposes. Linear regression is used to predict a metric dependent variable from several independent variables, whereas logistic regression is used as a binary classifier. However, both use a similar method of fitting the data.

Linear regression fits a linear model to the data to predict the outcome, logistic regression predicts the probability of a sample belonging to one class or another. Since probabilities are never actually 0 or 1, a linear model is inappropriate–we need something which will asymptotically approach 0 and 1, which is why we use a logistic curve. We then assign the sample to the class for which the model predicts a higher probability of membership.

Logistic regression and LDA are both classifiers, so their use cases overlap–one would frequently need to decide whether to use logistic regression or LDA for a problem. Logistic regression in its basic form is a binary classifier, but it can be extended (e.g. with 1 vs all classification) to multiclass scenarios, very similar to LDA. LDA, however, does not use a regression style formula to make predictions, rather, it simply creates a decision boundary which minimizes the within class scatter and maximizes the between class scatter. This gives each method properties which make them better suited to some cases than others.

Logistic regression is much more robust to violations of the assumptions of normality and heteroskedasticity, so in cases where the data is not normally distributed, or verifying that it is normally distributed is not feasible, logistic regression may yield better results. Logistic regression can also easily include nonmetric dependent variables through one-hot encoding.

LDA may seem more limited, but when all independent variables are metric and we can reasonably assume that they are normally distributed,

LDA can yield superior results. LDA can also be used for dimensionality reduction by projecting the data into a lower dimensional space after it has fit the data.

2. For the data set associated with this homework (HBAT) Using X4 as the non-metric response variable
and (X6 up to X15) as the metric variables:

a. Apply forward selection binary logistic regression (1 is the level of interest with single non-cross effects) and report what variable is entered into the model after each step. (Use 0.05 significance level). Report the final summary of the regression model and the ROC curve and the area under the ROC curve after each step.

Only one variable was entered, x11. SAS detected quasi-complete separation, so the fit is questionable. Area under the ROC curve = 0.9584,

b. Apply backward selection binary logistic regression (1 is the level of interest with single non-cross effects) and report what variable is eliminated from the model after each step. (Use 0.05 significance level). Report the final summary of the regression model and the ROC curve and the area under the ROC curve after each step.

SAS removed x9-x15 as redundant. SAS then removed x5 and detected complete separation in the model. SAS then removed x6, again detecting complete separation. After this, SAS removed x7, and only detected quasi-complete separation.While this throws a wrench in our analysis, it could mean that we are able to perfectly predict x4 using x7-x8. This complete separation leads to an area under the ROC curve of 1.

c. Which selection method from (a) or (b) provides a better model? Explain.