Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Mar 4th 2019

Please note that assignments in this class may be submitted to www.turnitin.com, a web- based anti-plagiarism system, for an evaluation of their originality.

1. **Define binary and dummy variables for the categorical variables in this dataset and report it in an excel file (20%)**

Answer:

Non-numeric variables: x1, x4, x9, x10, x12, x13 and Y

Binary variables are x1, x9, x10, x10 and Y

And there are 3 possible outcomes for each x4 and x13

Set dummy values for Non-numeric variables, the result dataset is in Excel file

x1	x2	x3	x4	x8	x9	x10	x11	x12	x13	x14	x15	Y
0	30.83	0	0	1.25	0	0	1	1	0	202	0	1
1	58.67	4.46	0	3.04	0	0	6	1	0	43	560	1
1	24.5	0.5	0	1.5	0	1	0	1	0	280	824	1
0	27.83	1.54	0	3.75	0	0	5	0	0	100	3	1
0	20.17	5.625	0	1.71	0	1	0	1	1	120	0	1
0	32.08	4	0	2.5	0	1	0	0	0	360	0	1
0	33.17	1.04	0	6.5	0	1	0	0	0	164	31285	1
1	22.92	11.585	0	0.04	0	1	0	1	0	80	1349	1
0	54.42	0.5	1	3.96	0	1	0	1	0	180	314	1
0	42.5	4.915	1	3.165	0	1	0	0	0	52	1442	1
0	22.08	0.83	0	2.165	1	1	0	0	0	128	0	1
0	29.92	1.835	0	4.335	0	1	0	1	0	260	200	1
1	38.25	6	0	1	0	1	0	0	0	0	0	1
0	48.08	6.04	0	0.04	1	1	0	1	0	0	2690	1
1	45.83	10.5	0	5	0	0	7	0	0	0	0	1
0	36.67	4.415	1	0.25	0	0	10	0	0	320	0	1
0	28.25	0.875	0	0.96	0	0	3	0	0	396	0	1

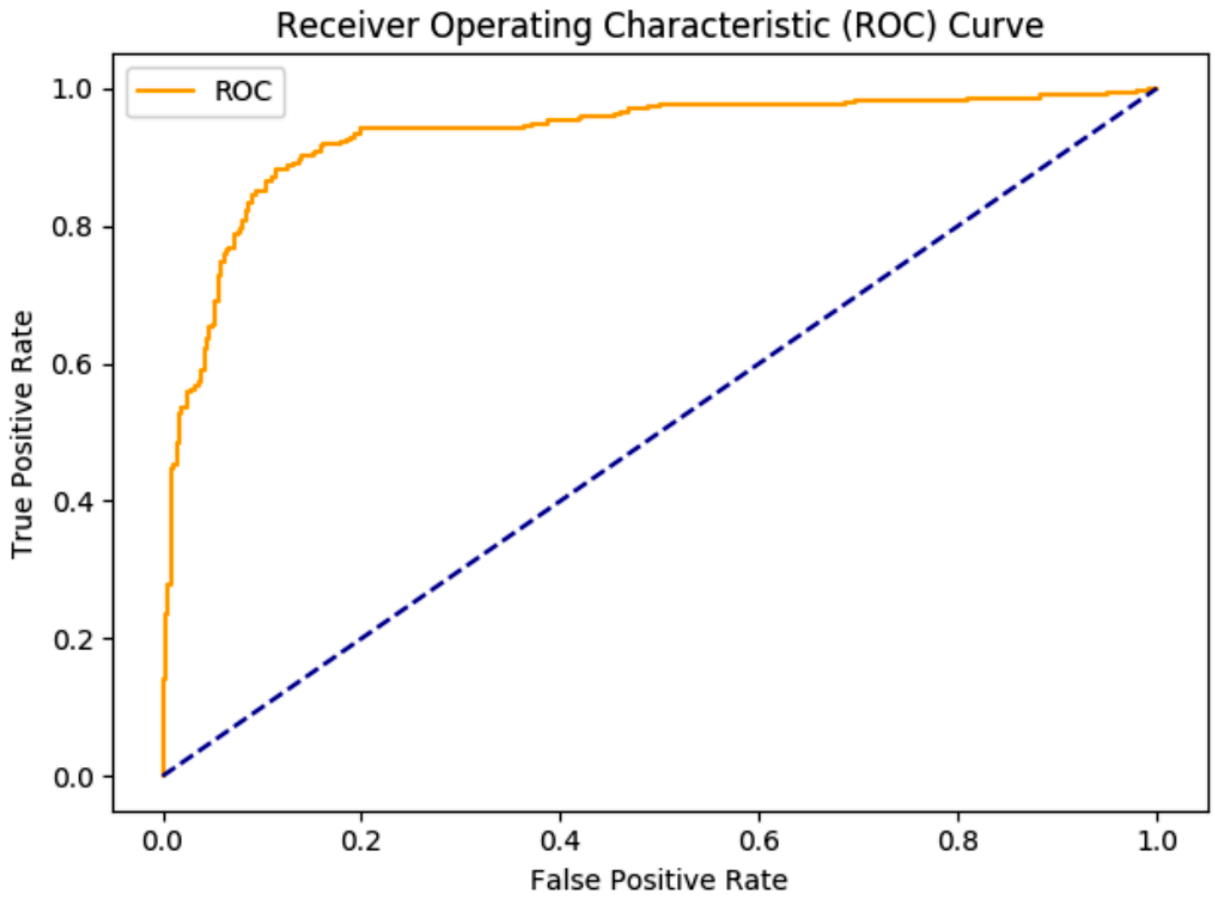
	More than 2 outcomes' variables
	Binary variables
x1	change a to 1, b to 0
x9	change t to 0, f to 1
x10	change t to 0, f to 1
x12	change t to 0, f to 1
Y	change + to 1, - to 0
x4	change u to 0, y to 1, l to 2
x13	change g to 0, s to 1, p to 2

2. **Develop a linear discriminant analysis model to predict Y (+ or -) and report the ROC curve (40%)**

Answer:

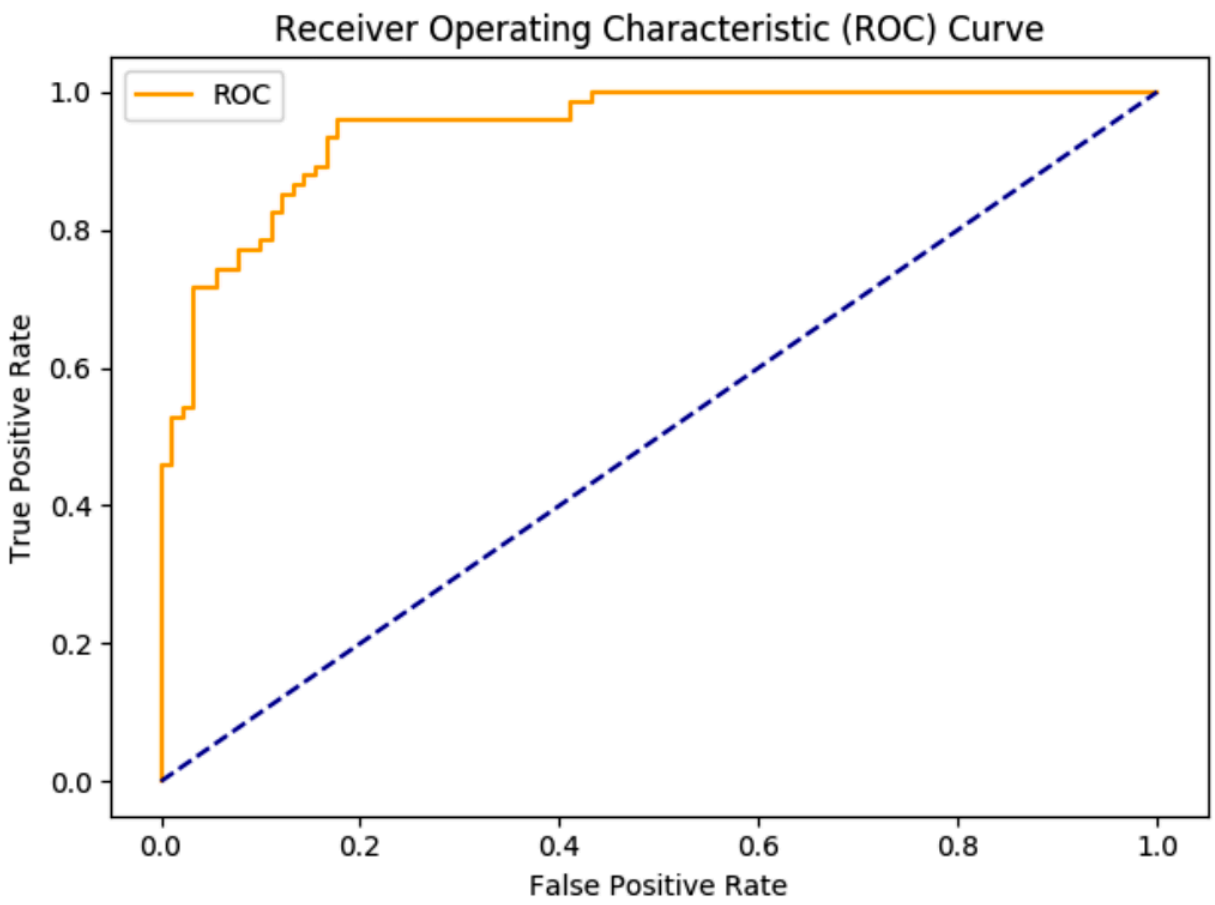
Following is the LDA model predict score and ROC curve for whole dataset (without split training and validation)

```
The score for the Linear discriminant analysis model is  0.863914373088685
The AUC value for the ROC curve is  0.9307243696210177
```



Following is the LDA model predict score and ROC curve by split the dataset to 75% training data and 25% validation data.

```
The score for the Linear discriminant analysis model is  0.8841463414634146
The AUC value for the ROC curve is  0.9459459459459459
```


Receiver Operating Characteristic (ROC) Curve
ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

3. **Find the k in the range of 3 to 10 that provides the most accurate KNN model to classify Y and report its confusion matrix (40%)**

Answer:

When we split the data to 75% training and 25% validation, the KNN model result and confusion matrix is following.

KNN:

```
When k = 3 the score is 0.6951219512195121
When k = 4 the score is 0.7012195121951219
When k = 5 the score is 0.7317073170731707
When k = 6 the score is 0.725609756097561
When k = 7 the score is 0.725609756097561
When k = 8 the score is 0.6951219512195121
When k = 9 the score is 0.6829268292682927
When k = 10 the score is 0.7012195121951219
```

We can find when we have 75% training data and 25% validation data, k = 5 provides the most accurate KNN model.
The confusion matrix of this is:

```
[[70 24]
 [20 50]]
```

If we change our data to 85% training data and 15% validation data, we can find k = 8 provides the most accurate KNN model, and the confusion matrix is:

```
[[52 17]
 [ 6 24]]
```