CIS 8695: Homework 2

Predicting System Administrator Performance

Professor Name: **Dr. Ling Xue**
Student Name: **Polli Fakhretdinova**
Assignment Due Date: **02/02/2018**

INTRODUCTION

Logistic regression - regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Binary logistic regression major assumptions:

1. The dependent variable should be binary.
2. There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores and removing values below -3.29 or greater than 3.29.
3. There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors.

At the center of the logistic regression analysis is the task estimating the log odds of an event. Mathematically, logistic regression estimates a multiple linear regression function defined as:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

BACKGROUND

A management consultant is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. In particular, she is interested in discriminating between administrators who are able to complete given tasks and those who are not. Data are collected on the performance of 75 randomly selected administrators. They are stored in the file **SystemAdminstrations.xlsx**.

The variable Experience measures the months of full-time system administrator experience, while Training measures the number of relevant training credits. The dependent variable **Completed** is **either Yes or No**, according to whether or not the administrator completed the tasks.

VARIABLES & DATA

The data consists of categorical dependent variable of **Completed** which is either **Yes** or **No**, and two continuous independent variables **Experience** and **Training**.

Dependent Variable

Completed_Yes
Completed_No

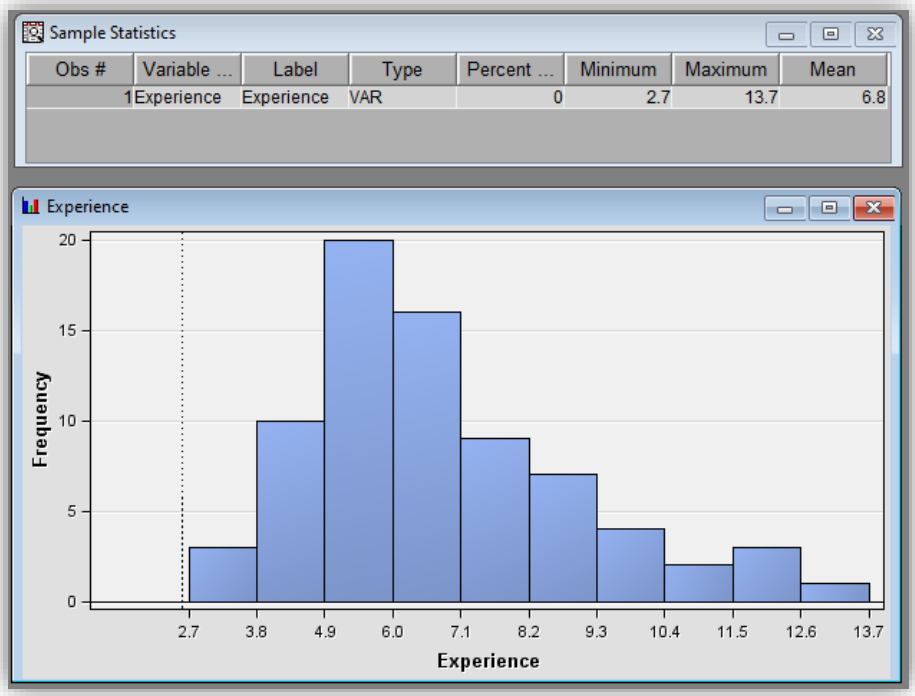
Independent Variables

Experience
Training

Completed	Experience	Training
Yes	10.9	4
Yes	9.9	4
Yes	10.4	6
Yes	13.7	6
Yes	9.4	8
Yes	12.4	4
Yes	7.9	6
Yes	8.9	4
Yes	10.2	6
Yes	11.4	4

EXPLORING DATA WITH SAS & R

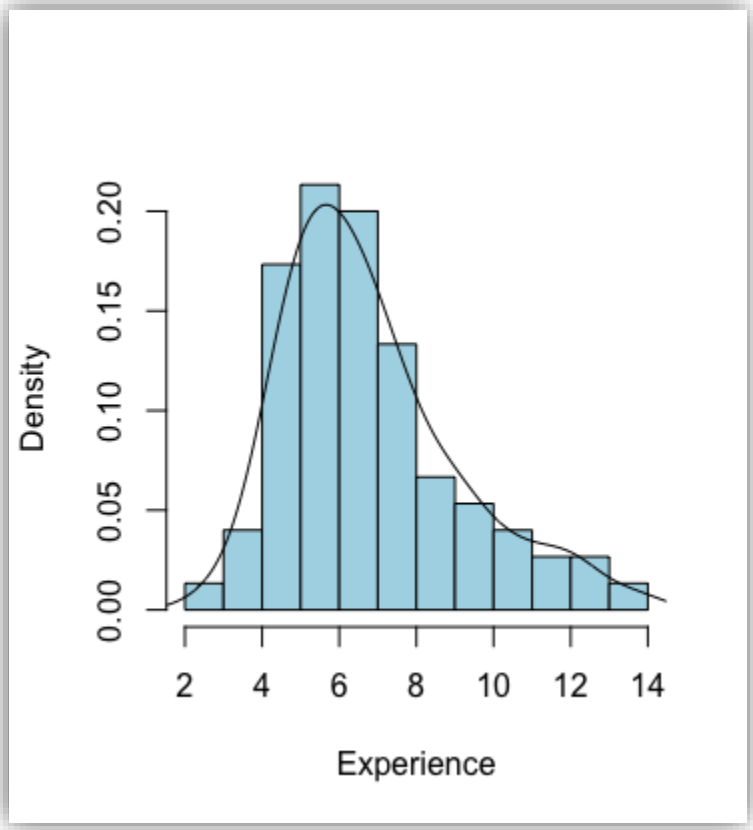
To understand the patterns in the data and the behavior of the data, it is important to explore the data. The histogram for the independent variable **'Experience'** along with descriptive statistics information is demonstrated below.



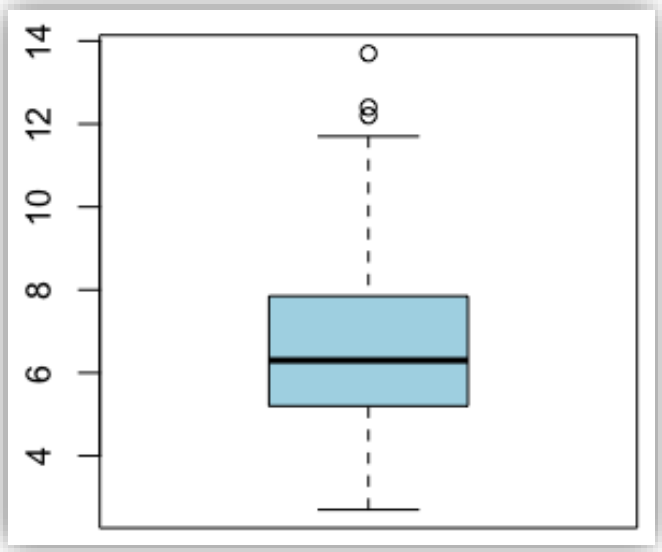
System administrator(s) with maximum months of full-time system administrator experience has 13.7 months.

Whereas, system administrator(s) with minimum months of full-time system administrator experience has 2.7 months of experience.

The average experience measured in months of full-time system administrator experience is 6.8 months.

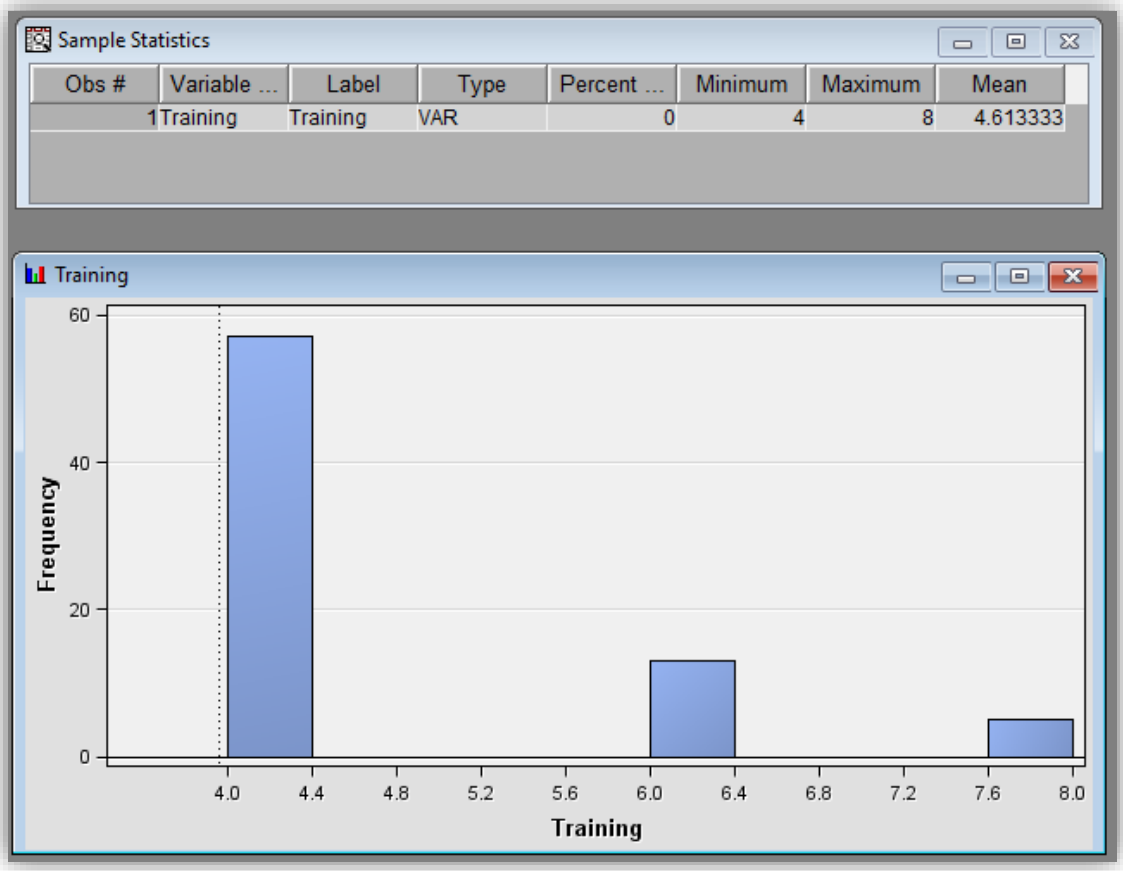


Based on the histogram distribution, there are 3 system administrators with experience between 2.7 – 3.8 months, 12 system administrators with experience between 3.8 – 4.9 months, 18 system administrators with experience between 4.9 – 6.0 months, 15 system administrators with experience between 6.0 – 7.1 months, 10 system administrators with experience between 7.1 – 8.2 months, 7 system administrators with experience between 8.2 – 9.3 months, 4 system administrators with experience between 9.3 – 10.4 months, 2 system administrators with experience between 10.4 – 11.5 months, 3 system administrators with experience between 11.5 – 12.6 months, and 1 system administrator with experience of 13.7 months. **We can conclude that most of the system administrators have experience between 5 – 6 months.**

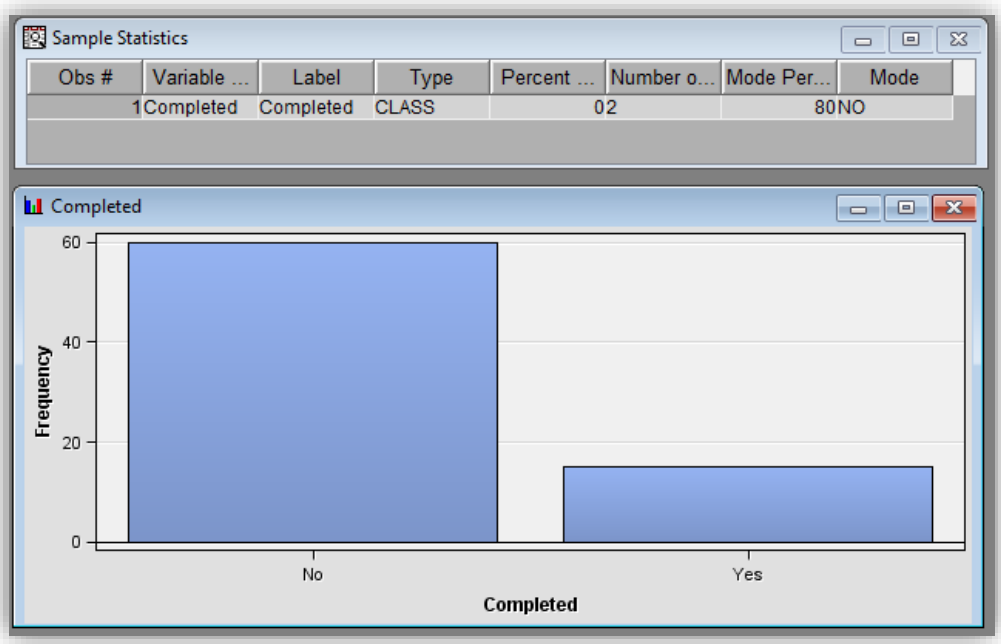


The boxplot of the independent variable 'Experience' tells that the average experience system administrators have is between 6 – 7 months. The boxplot also shows that there are three outliers – two system administrators with experience of over 12 months and one system administrator with experience of over 13 months.

The histogram for the independent variable **'Training'** along with descriptive statistics information is demonstrated below. System administrator(s) with maximum credits for training has 8 credits. System administrator(s) with minimum credits for training has 4 credits. The average training credits system administrators have is 4.6 credits. Additionally, we can conclude that most of the system administrators (57 of the 75 system administrators) had 4 credits for Training. Whereas, 13 of the system administrators had 6 credits for Training, and 5 system administrators had 8 credits for Training.

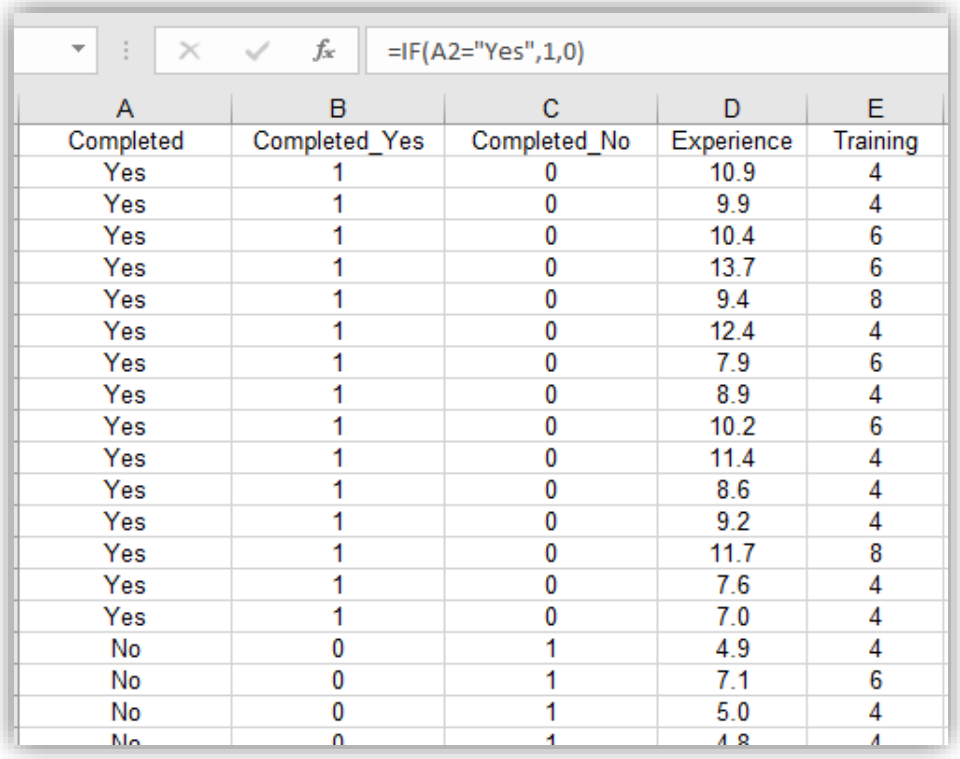


The histogram for the dependent variable **'Completed_Yes'** is demonstrated below. Based on the histogram, we can conclude that 15 of the system administrators could complete the task and 60 of the system administrators could not complete the task.



ANALYSIS OF VARIABLES

Transform the Completed variable into dummy variables: Completed_Yes and Completed_No.



A	B	C	D	E
Completed	Completed_Yes	Completed_No	Experience	Training
Yes	1	0	10.9	4
Yes	1	0	9.9	4
Yes	1	0	10.4	6
Yes	1	0	13.7	6
Yes	1	0	9.4	8
Yes	1	0	12.4	4
Yes	1	0	7.9	6
Yes	1	0	8.9	4
Yes	1	0	10.2	6
Yes	1	0	11.4	4
Yes	1	0	8.6	4
Yes	1	0	9.2	4
Yes	1	0	11.7	8
Yes	1	0	7.6	4
Yes	1	0	7.0	4
No	0	1	4.9	4
No	0	1	7.1	6
No	0	1	5.0	4
No	0	1	4.8	4

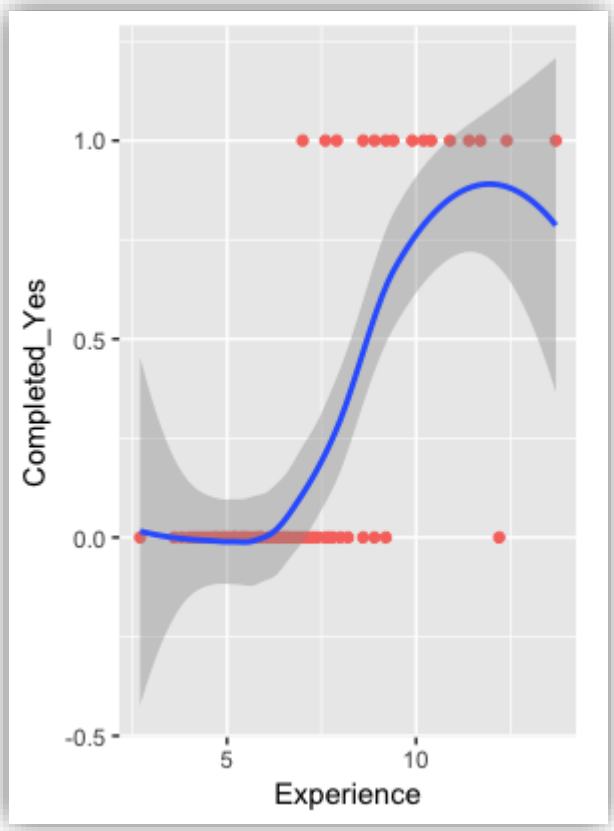
In order to create dummy variables for the categorical dependent variable, I used a simple formula in Excel in order assign Yes as 1 and No as 0. The same process can be done in SAS as well.

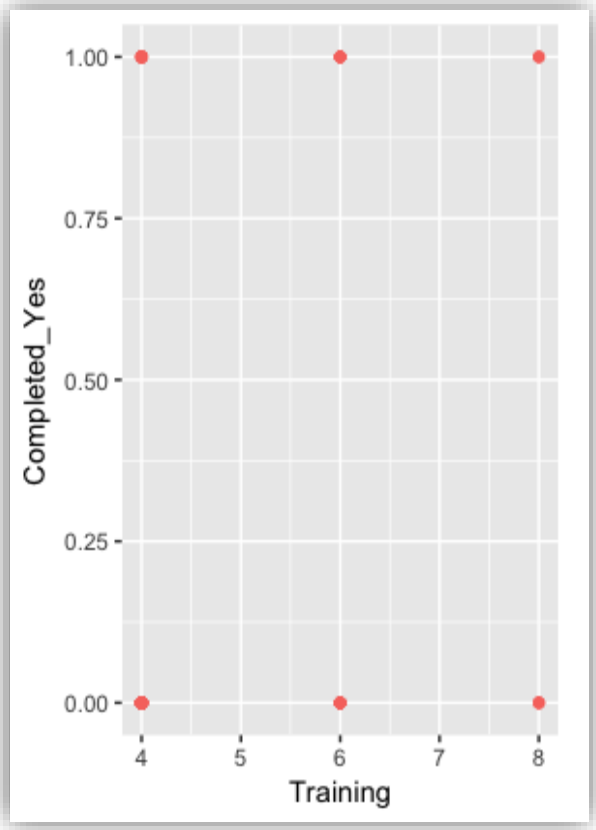
After transforming the dependent variable into dummies, imported the data into SAS and R in order to perform descriptive and applied statistics.

The charts and analysis results are demonstrated and listed below.

Develop a scatter plot of Completed_Yes versus Experience and a scatter plot of Completed_Yes versus Training. Based on your observation, which predictors appear potentially useful for classifying task completion?

Based on my observation by looking at the scatterplots of independent variables **'Experience'** and **'Training'**, I can conclude that the predictor **'Experience'** appear potentially useful for classifying task completion. The S-shaped curve indicates that there is positive association between the dependent and independent variable. In other words, as **Experience** increases the probability that a system developer completed the task increases. Scatterplot for **Experience** is more useful because a perfect relationship represents a perfectly curved S.





CORRELATION MATRIX

Based on the correlation matrix below, we can state that the independent variable **'Experience'** and the dependent variable **'Completed_Yes'** are highly correlated, meaning that it is very likely that probability of a system administrator completing the task would increase if he or she has longer period of experience. On the other hand, the independent variable **'Training'** has very little correlation with the dependent variable, meaning training credit amount may not be a good predictor of the dependent variable.

	Completed_Yes	*Experience*	*Training*
Completed_Yes	1		
Experience	0.696647285	1	
Training	0.192684823	0.227848747	1

LOGISTIC REGRESSION

Use the entire dataset as training data. Fit a logistic regression using Completed_Yes as the outcome variable, and Experience and Training as predictors. Report the logistic regression model result. Explain what the prediction model looks like (using the coefficient estimates)

Based on the **Analysis of Maximum Likelihood Estimates** demonstrated below, we can state that the independent variable or predictor **Experience** has a **P-value** of **0.0001** indicating that it has an impact on the dependent variable. Whereas, the independent variable **Training** has **P-value** of **0.5940** indicating that it has NO impact on the dependent variable or it is not a good predictor of system administrator performance.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-10.9813	2.8920	14.42	0.0001		0.000
Experience	1	1.1269	0.2909	15.01	0.0001	1.4126	3.086
Training	1	0.1805	0.3386	0.28	0.5940	0.1179	1.198

The prediction model for predicting the system administrators' performance on completing a task is demonstrated below. The model tells us that the probability of completing a task for a system administrator would increase if he or she has more experience. The model also tells that the probability of completing a task for a system administrator would increase if he or she has more training credits.

$$\text{logit}(p) = -10.9813 + 1.1269 \times \text{Experience} + 0.1805 \times \text{Training}$$

Report the classification tables that show the prediction performance.

According to the **Event Classification Table** demonstrated below, 5 system administrators were classified as people who FAILED complete the task falsely and 2 system administrators were classified as people who COULD complete the task falsely. On the other hand, 58 system administrators were correctly classified as True Negatives and 10 system administrators were correctly classified as True Positives.

Event Classification Table

Data Role=TRAIN Target=Completed_Yes Target Label=Completed_Yes

False Negative	True Negative	False Positive	True Positive
5	58	2	10

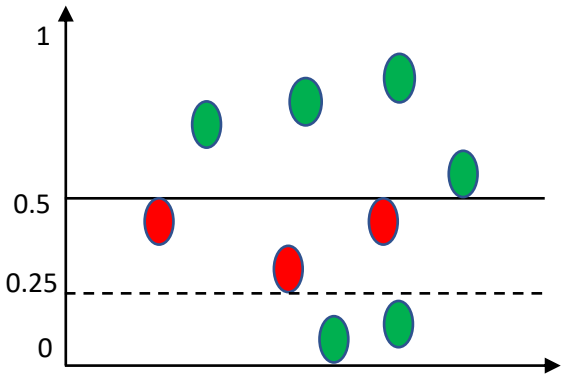
QUESTIONS & EXPLANATIONS

Based on the regression results, among those who complete the task, what is the percentage of administrators who were successfully completed the task but are incorrectly classified as failing to complete the task? Show the calculation steps that you use to answer this question.

True Positive → 10 system administrators who could complete the task and correctly were classified
False Negative → 5 system administrators who could complete the task and incorrectly were classified
Total → 15 system administrators who could compete the task

Calculation → **5 / (10 + 5) = 33.34 %** The **33.34 %** of system administrators who successfully completed the task but incorrectly were classified as failing to complete the task.

In question (3), the administrators are predicted to fail if the predicted probability is less than a cutoff level of 0.5. If we want to decrease the percentage in question (3), should the cutoff level be increased or decreased? Why?



As it is demonstrated in the figure above, in order to decrease the percentage of system administrators incorrectly classified as people who could not complete the task we need to decrease the cutoff level. For example, if we decrease the cutoff level to 25 % the number of False Negatives decreases.

Based on the prediction model you obtain in question (2), how much experience credit must be earned by a programmer who has earned 4 training credits before s/he is classified as being able to complete the task? Show the calculation steps that you use to answer this question.

METHOD 1 & OUTPUT: With a cutoff level of 50%, a system administrator who earned 4 training credits needs to have at least 9.1040 months of experience in order to be considered as a person who completed the task.

p = 1 / 1 + e^-(−10.9813 + 1.1269 x Experience + 0.1805 x Training)
0.5= 1 / 1 + e^-(−10.9813 + 1.1269 x Experience + 0.1805 x 4)
0.5= 1 / 1 + e^-(−10.9813 + 1.1269 x Experience + 0.722)
log(0.5 / 1 – 0.5) = −10.9813 + 1.1269 x Experience + 0.722
0 = −10.9813 + 1.1269 x Experience + 0.722
-1.1269 * Experience = −10.9813 + 0.722
-1.1269 * Experience = −10.2593
Experience = 9.1040

METHOD 2 → Look up the p value in the table below in order to find the corresponding value for Experience
OUTPUT: With a cutoff level of 50%, a system administrator who earned 4 training credits needs to have at least 9.1040 months of experience in order to be considered as a person who completed the task.

logit(p) = −10.9813 + 1.1269 x Experience + 0.1805 x Training
logit(p) = −10.9813 + 1.1269 x **9.1040** + 0.1805 x 4
logit(p) = −10.9813 + 10.2593 + 0.722
logit(p) = −10.9813 + 10.9813
logit(p) = 0.00
p = 0.00 → 50%

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$$p = \frac{1}{1 + e^{-logit(p)}}$$

p	logit(p)	p	logit(p)	p	logit(p)	p	logit(p)
0.01	-4.5951	**0.26**	-1.0460	**0.51**	0.0400	**0.76**	1.1527
0.02	-3.8918	**0.27**	-0.9946	**0.52**	0.0800	**0.77**	1.2083
0.03	-3.4761	**0.28**	-0.9445	**0.53**	0.1201	**0.78**	1.2657
0.04	-3.1781	**0.29**	-0.8954	**0.54**	0.1603	**0.79**	1.3249
0.05	-2.9444	**0.30**	-0.8473	**0.55**	0.2007	**0.80**	1.3863
0.06	-2.7515	**0.31**	-0.8001	**0.56**	0.2412	**0.81**	1.4500
0.07	-2.5867	**0.32**	-0.7538	**0.57**	0.2819	**0.82**	1.5163
0.08	-2.4423	**0.33**	-0.7082	**0.58**	0.3228	**0.83**	1.5856
0.09	-2.3136	**0.34**	-0.6633	**0.59**	0.3640	**0.84**	1.6582
0.10	-2.1972	**0.35**	-0.6190	**0.60**	0.4055	**0.85**	1.7346
0.11	-2.0907	**0.36**	-0.5754	**0.61**	0.4473	**0.86**	1.8153
0.12	-1.9924	**0.37**	-0.5322	**0.62**	0.4895	**0.87**	1.9010
0.13	-1.9010	**0.38**	-0.4895	**0.63**	0.5322	**0.88**	1.9924
0.14	-1.8153	**0.39**	-0.4473	**0.64**	0.5754	**0.89**	2.0907
0.15	-1.7346	**0.40**	-0.4055	**0.65**	0.6190	**0.90**	2.1972
0.16	-1.6582	**0.41**	-0.3640	**0.66**	0.6633	**0.91**	2.3136
0.17	-1.5856	**0.42**	-0.3228	**0.67**	0.7082	**0.92**	2.4423
0.18	-1.5163	**0.43**	-0.2819	**0.68**	0.7538	**0.93**	2.5867
0.19	-1.4500	**0.44**	-0.2412	**0.69**	0.8001	**0.94**	2.7515
0.20	-1.3863	**0.45**	-0.2007	**0.70**	0.8473	**0.95**	2.9444
0.21	-1.3249	**0.46**	-0.1603	**0.71**	0.8954	**0.96**	3.1781
0.22	-1.2657	**0.47**	-0.1201	**0.72**	0.9445	**0.97**	3.4761
0.23	-1.2083	**0.48**	-0.0800	**0.73**	0.9946	**0.98**	3.8918
0.24	-1.1527	**0.49**	-0.0400	**0.74**	1.0460	**0.99**	4.5951
0.25	-1.0986	**0.50**	0.0000	**0.75**	1.0986