Cairo University
Faculty of Engineering

Department of Computer
Engineering

Big Data – Lab 5 Requirement
Logistic Regression

Team Info

Name	Code	Sec	BN
Abdelrahman Hamdy Ahmed	9202833	1	38
Abdelrahman Noaman Loqman	9202851	2	4

Date: 4/08/2024

Presented to: Eng. Omar Samir

Answers to Theoretical Questions
With Plots

Q1) **Write the variable pairs that are not correlated at all to each other.**

Correlation Matrix

	Price	Income	Age
Price	1	0.00000000	0.00000000
Income	0	1.00000000	0.09612083
Age	0	0.09612083	1.00000000

Variable pairs that have no correlation are the ones yielding 0 in the matrix which are:

- Price & Income
- Price & Age

Q2) **Are there any highly correlated variables in this dataset?**

The only correlated variables in the dataset are Age and Income.
Their correlation is 0.09612083 which is really low, meaning that there are no highly correlated variables in this dataset.

Q3) **How many categories are there for the Price variable?**

	MYDEPV	
Price	0	1
10	115	135
20	137	113
30	174	76

Three categories → Represented by 2 dummy/indicator variables

Q4) **Why is it divided into two entries only in the model?**

Okay so how we would normally do it is to use 3 entries like the following:

10	20	30	
1	0	0	→ 10
0	1	0	→ 20
0	0	1	→ 30

We can simply cross out the third column since we can represent the third category by setting the first and second to zero, therefore we know that this will output 30 as the other categories already have 1s in their cells. It'll become like this

x1	x2	Category
0	1	10
1	0	20
0	0	30

Therefore, 2 entries are enough for 3 categories.

Q5.1) **The AUC score: tells you how well the model predicts. Write the value of this expression (just the number)**

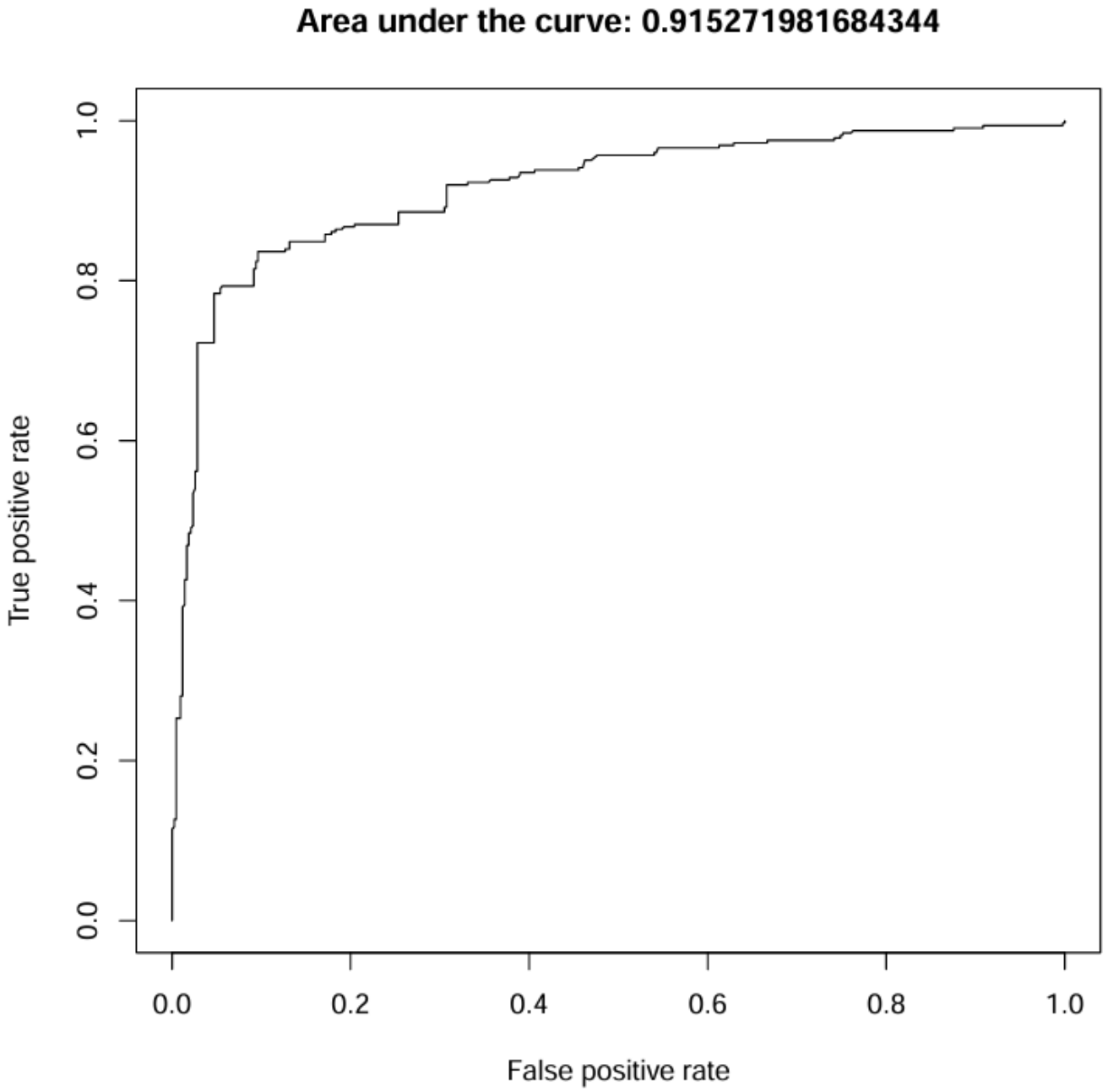
0.915272

Q5.2) **What is the maximum value of AUC (ideal case)?**

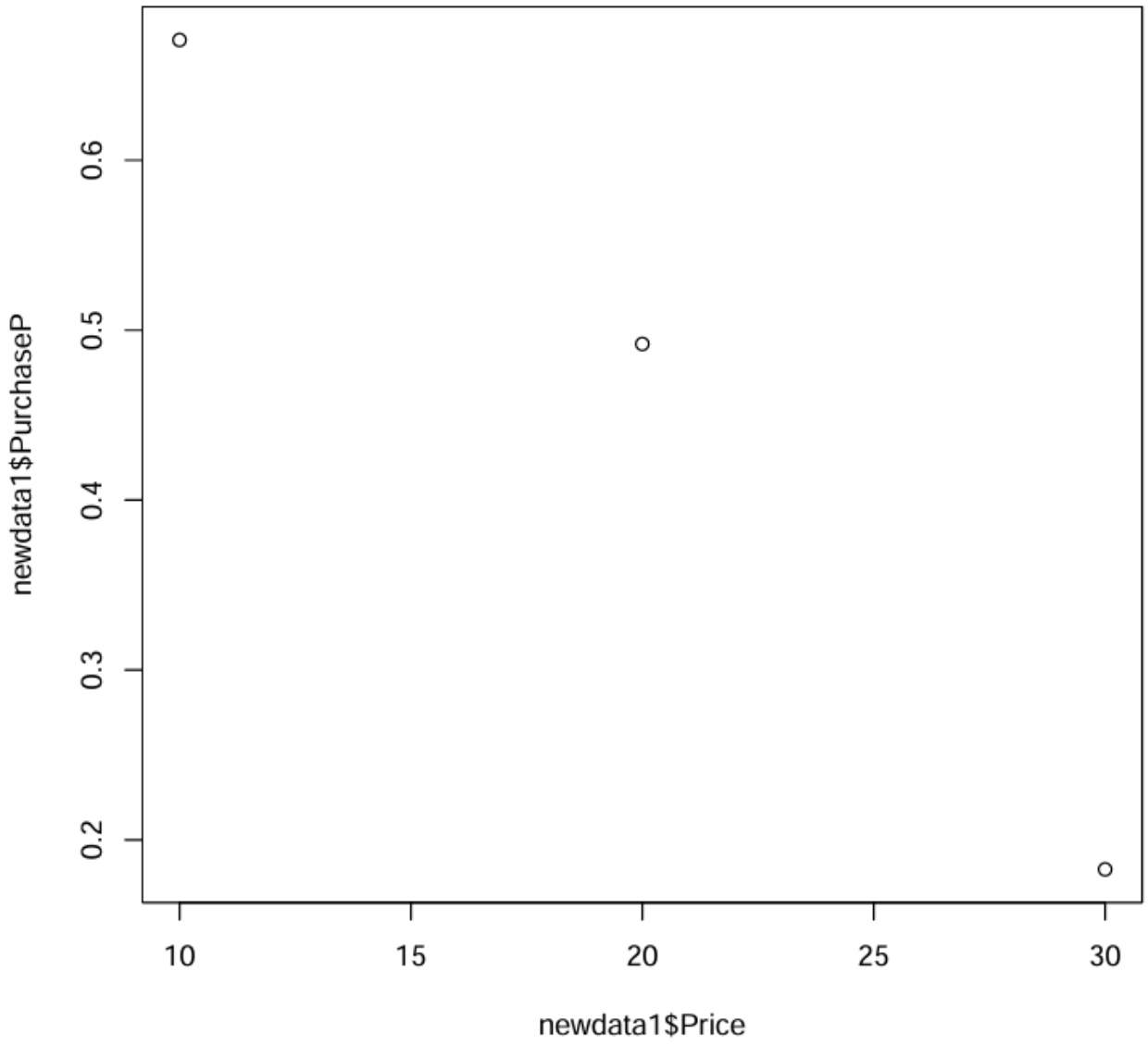
AUC: Area Under the Receiver Operating Characteristic Curve
The maximum value of the AUC is 1.0.
In an ideal case, where the model perfectly predicts the outcome, the AUC would be 1.0, indicating perfect discrimination between the positive and negative classes. This means the model has a perfect ability to distinguish between the two classes, with no false positives or false negatives.

Q6) **What does each point in the ROC graph represent? In other words, what is the value that changes and drives TPR and FPR to change too from one point to another in the graph?**

The threshold value defines the point at which a data point is classified as positive if its predicted probability exceeds the threshold and negative otherwise.

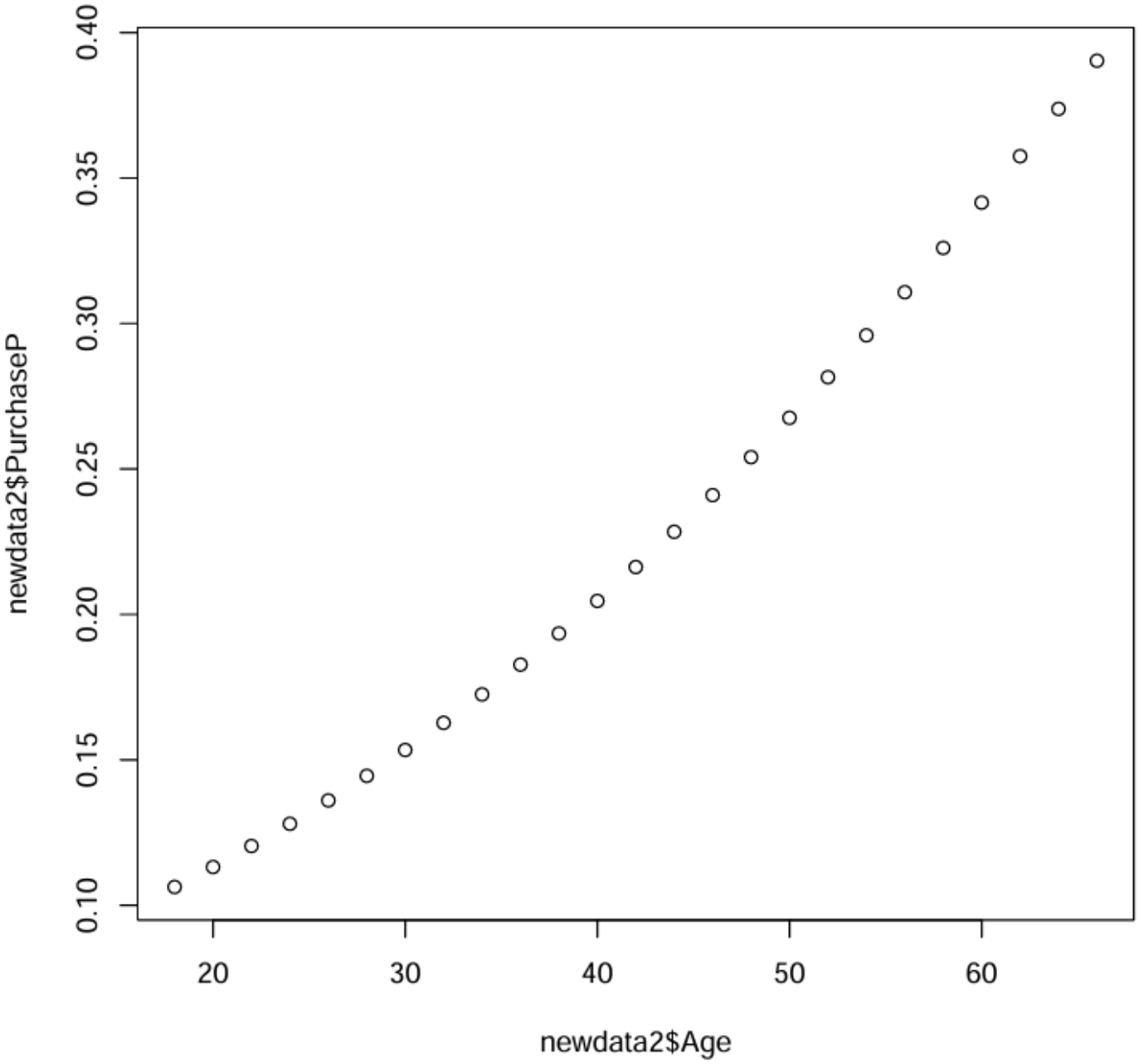


Q7) **How is the predicted probability affected by changing only price holding all other variables constant?**



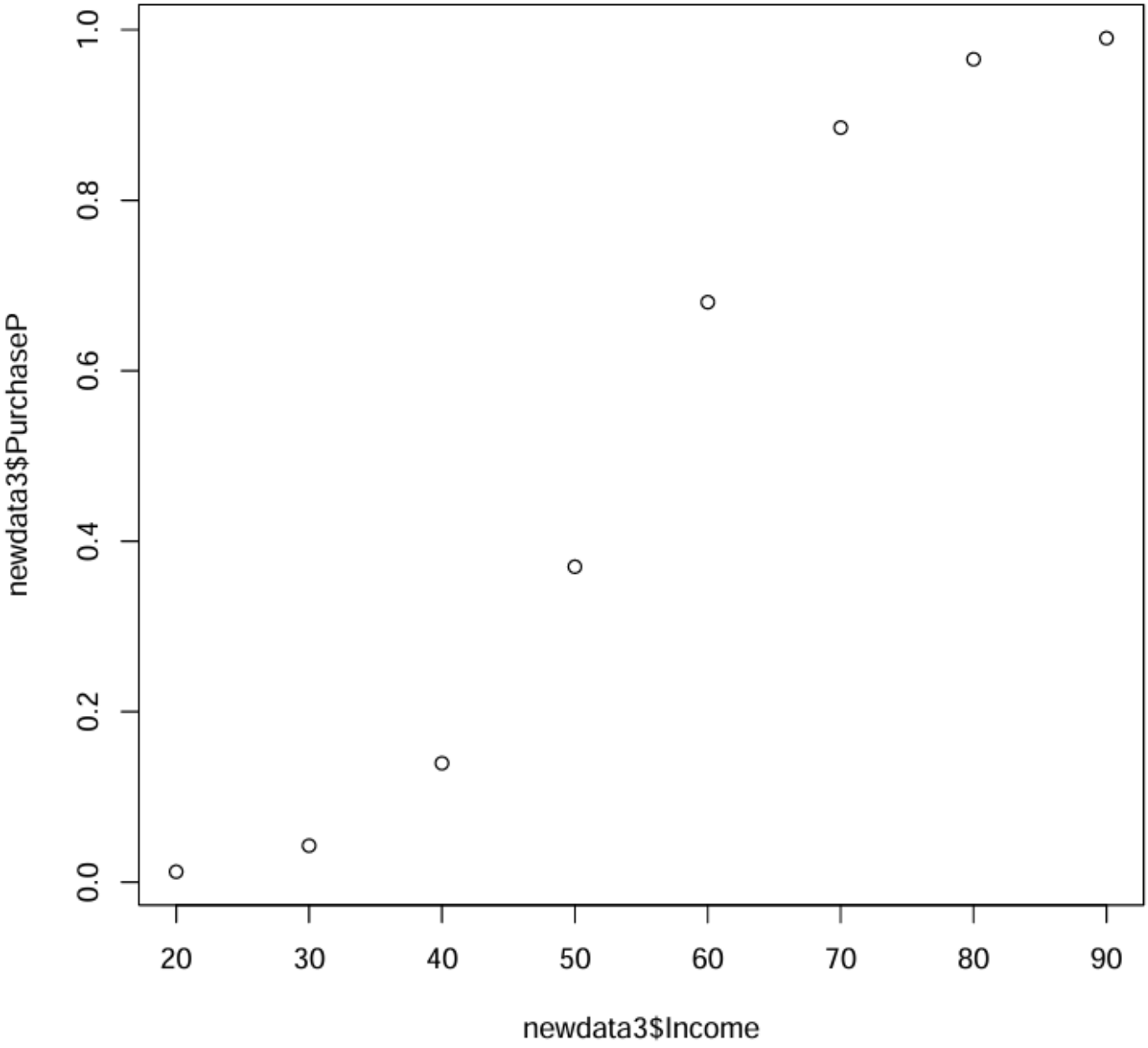
As we can see from the slope of the points, Price and PurchaseP show a negative correlation relationship: The higher the price, the lower the predicted probability of purchase. This is also logically makes sense since if the price of an item is too high, we don't usually expect many people to buy it, hence the decrease in probability.

Q8) **How is the predicted probability affected by changing only age holding all other variables constant?**



Positive Correlation: The graph shows an exponential increase in the predicted probability of purchase as the age increases.

Q9) **How is the predicted probability affected by changing only income holding all other variables constant?**



Positive Correlation: We already know what this shape looks like… Yes it's a sigmoid function.
Which means that as income increases, the probability increases but with a varying rate, depending on how close we are to 1.