Exercise 1

Exercise 1:

1) $IG(\text{Attend Class}) = 1 - \left(\frac{2}{6} H\left(\frac{2}{2}, \frac{0}{2}\right) + \frac{4}{6} H\left(\frac{1}{4}, \frac{3}{4}\right)\right)$

$$= 1 - \left(\frac{4}{6}\left(-\frac{1}{4}\log_2 \frac{1}{4} - \frac{3}{4}\log_2 \frac{3}{4}\right)\right)$$

$$= 0.459$$

$$IG(\text{Ends 3-year}) = 1 - \left(\frac{5}{6} H\left(\frac{2}{5}, \frac{3}{5}\right) + \frac{1}{6} H\left(\frac{1}{1}, \frac{0}{1}\right)\right)$$

$$= 1 - \left(\frac{5}{6}\left(-\frac{2}{5}\log_2 \frac{2}{5} - \frac{3}{5}\log_2 \frac{3}{5}\right)\right)$$

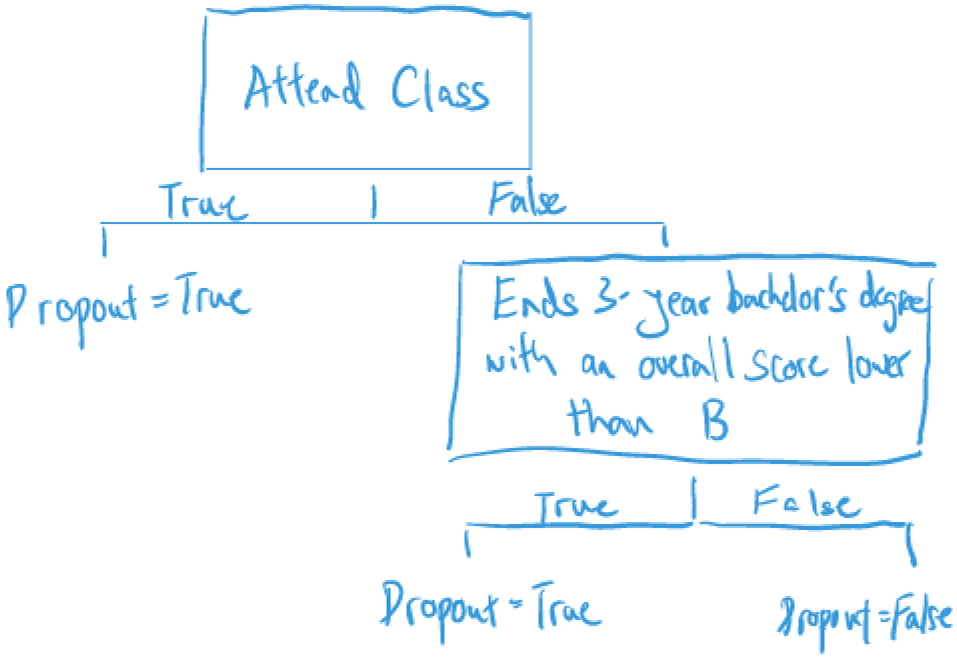
$$= 0.191$$

$$IG(\text{Less than 2 jobs}) = 1 - \left(\frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right) + \frac{1}{2} H\left(\frac{1}{3}, \frac{2}{3}\right)\right)$$

$$= 1 - \left(\frac{1}{2}\left(-\frac{2}{3}\log_2 \frac{2}{3} - \frac{1}{3}\log_2 \frac{1}{3}\right) + \frac{1}{2}\left(-\frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3}\right)\right)$$

$$= 0.0817$$

$\therefore$, since IG of Less than 2 part-time jobs is close to zero, we disregard it when creating our decision tree thus, we have the following.

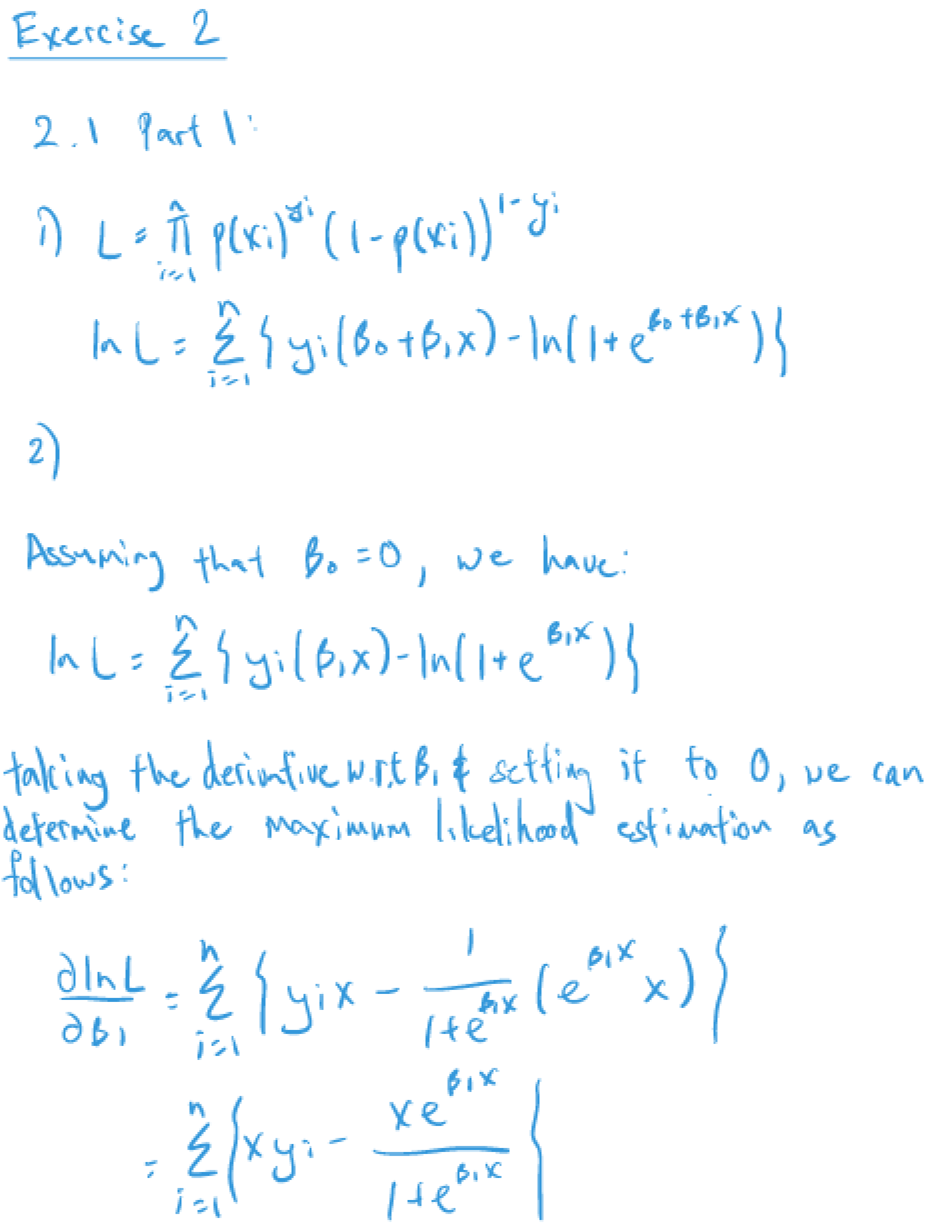


2) Dropout = True

3) Dropout = True

Exercise 2

2.1 Part 1



Exercise 2

2.1 Part 1:

1) $L = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$

$$\ln L = \sum_{i=1}^{n} \{ y_i(\beta_0+\beta_1 x) - \ln(1+e^{\beta_0+\beta_1 x}) \}$$

2)

Assuming that $\beta_0 = 0$, we have:

$$\ln L = \sum_{i=1}^{n} \{ y_i(\beta_1 x) - \ln(1+e^{\beta_1 x}) \}$$

taking the derivative w.r.t β_1 & setting it to 0, we can determine the maximum likelihood estimation as follows:

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{i=1}^{n} \left\{ y_i x - \frac{1}{1+e^{\beta_1 x}}(e^{\beta_1 x} x) \right\}$$

$$= \sum_{i=1}^{n} \left\{ x y_i - \frac{x e^{\beta_1 x}}{1+e^{\beta_1 x}} \right\}$$

$$\therefore \sum_{i=1}^{n} \left\{ x y_i - \frac{x e^{\beta_1 x}}{1+e^{\beta_1 x}} \right\} = 0$$

$$x \sum_{i=1}^{n} \left\{ y_i - \frac{e^{\beta_1 x}}{1+e^{\beta_1 x}} \right\} = 0$$

$$\sum_{i=1}^{n} \left\{ y_i - \frac{e^{\beta_1 x}}{1+e^{\beta_1 x}} \right\} = 0$$

$\therefore$ using the hint & taking the limit of $g(x)$ as $\beta \to \infty$ gives:

$$\lim_{\beta_1 \to \infty} \sum_{i=1}^{n} \left\{ y_i - \frac{e^{\beta_1 x}}{1+e^{\beta_1 x}} \right\} = 0$$

Since $y_i = 1$ if $x > 0$ & $y_i = 0$ if $x < 0$, we have:

For $x > 0$: $\lim_{\beta_1 \to \infty} \sum_{i=1}^{n} \left\{ 1 - 1 \right\} = 0$

For $x < 0$: $\lim_{\beta_1 \to \infty} \sum_{i=1}^{n} \left\{ 0 - \frac{0}{1} \right\} = 0$

Therefore, the maximum likelihood estimation does not converge in this case

2.2 Part 2

2.2 Part 2:

1) $$P(Y=k \mid X=x) = \frac{P(X=x \mid Y=k)\,P(Y=k)}{P(X=x \mid Y=k)\,P(Y=k) + P(X=x \mid Y=k')\,P(Y=k')}$$

allowing $f_k(x) = P(X=x \mid Y=k)$ & $\pi_k = P(Y=k)$, we have

$$P(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{a=1}^{K} \pi_a f_a(x)}$$

$$P(Y=k \mid X=x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}}{\sum_{a=1}^{K} \pi_a \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_a)\right)^2}}$$

assuming $f_k(x)$ is a normal distb.

$$P(Y=k \mid X=x) \propto \pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}$$

expressing as a proportionality constant

$$\log(P(Y=k \mid X=x)) \propto \log \pi_k - \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(x-\mu_k)^2$$

$$\log(P(Y=k \mid X=x)) \propto \log \pi_k - \frac{1}{2\sigma^2}(x-\mu_k)^2$$

$-\log(\sqrt{2\pi}\sigma)$ goes to proportionality constant as it does not depend on k

$$\log(P(Y=k|X=x)) \propto \log \pi_k - \frac{1}{2\sigma^2}\left(x^2 + \mu_k^2 - 2\mu_k x\right)$$

$$\log(P(Y=k|X=x)) \propto \log \pi_k + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$ # $\frac{x^2}{2\sigma^2}$ goes to the proportionality constant as it does not depend on k

$$\therefore \delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

□

2)

$$P(Y=1 \mid X=x) = \frac{\pi_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)}}{\sum_{a=1}^{K} \pi_a \frac{1}{\sqrt{2\pi}}\sigma_1 e^{\left(-\frac{1}{2\sigma_1^2}(x-\mu_a)\right)^2}} \quad — ①$$

$$P(Y=0 \mid X=x) = \frac{\pi_0 \frac{1}{\sqrt{2\pi}\sigma_0} e^{\left(-\frac{1}{2\sigma_0^2}(x-\mu_0)^2\right)}}{\sum_{a=1}^{K} \pi_a \frac{1}{\sqrt{2\pi}}\sigma_0 e^{\left(-\frac{1}{2\sigma_0^2}(x-\mu_a)\right)^2}} \quad — ②$$

$$P(Y=1 \mid X=x) \propto \pi_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)} \quad — ①$$

$$P(Y=0 \mid X=x) \propto \pi_0 \frac{1}{\sqrt{2\pi}\sigma_0} e^{\left(-\frac{1}{2\sigma_0^2}(x-\mu_0)^2\right)} \quad — ②$$

expressing ① & ② as pr ortionality c stants

$$\therefore\ P(Y=1 \mid X=x) = P(Y=0 \mid X=x)$$

$$\rightarrow\ \pi_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right)} = \pi_0 \frac{1}{\sqrt{2\pi}\sigma_0} e^{\left(-\frac{1}{2\sigma_0^2}(x-\mu \quad ^2\right)}$$

Taking the log of both sides we have.

$$\log(\pi_1) - \log\sqrt{2\pi}\,\sigma_1 - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 = \log(\pi_0) - \log\sqrt{2\pi}\,\sigma_0 - \frac{1}{2\sigma_0^2}(x-\mu_0)^2$$

$$\rightarrow \log(\pi_1) - \log(\pi_0) - \log\sqrt{2\pi}\,\sigma_1 + \log\sqrt{2\pi}\,\sigma_0 - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 + \frac{1}{2\sigma_0^2}(x-\mu_0)^2 = 0$$

$$\rightarrow \log(\pi_1) - \log(\pi_0) - \log\sqrt{2\pi}\,\sigma_1 + \log\sqrt{2\pi}\,\sigma_0 - \frac{1}{2\sigma_1^2}(x^2+\mu_1^2-2x\mu_1) + \frac{1}{2\sigma_0^2}(x^2+\mu_0^2-2x\mu_0)$$

$$\rightarrow \log(\pi_1) - \log(\pi_0) - \log\sqrt{2\pi}\,\sigma_1 + \log\sqrt{2\pi}\,\sigma_0 + x^2\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) + x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right) + \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2} = 0$$

$\therefore$, we have determined Bayes discriminant points as shown above

3)

$$P(X=x|Y=1) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right)}$$

$$P(X=x|Y=0) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_0)^2\right)}$$

$$\therefore \frac{P(X=x|Y=1)}{P(X=x|Y=0)} = \frac{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right)}}{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} e^{\left(-\frac{1}{2\sigma^2}(x-\mu_0)^2\right)}}$$

$$= \frac{e^{\left(-\frac{1}{2\sigma^2}(x-\mu_1)^2\right)}}{e^{\left(-\frac{1}{2\sigma^2}(x-\mu_0)^2\right)}}$$

$$= \frac{e^{\left(-\frac{1}{2\sigma^2}(x^2+\mu_1^2-2x\mu_1)\right)}}{e^{\left(-\frac{1}{2\sigma^2}(x^2+\mu_0^2-2x\mu_0)\right)}}$$

$$= \frac{e^{\left(-\frac{1}{2\sigma^2}\left(x^2+\mu_1^2-2x\mu_1\right)\right)}}{e^{\left(-\frac{1}{2\sigma^2}\left(x^2+\mu_0^2-2x\mu_0\right)\right)}}$$

$$= \frac{\cancel{e^{-\frac{1}{2\sigma^2}x^2}}\; e^{-\frac{1}{2\sigma^2}\mu_1^2}\; e^{\frac{1}{2\sigma^2}2x\mu_1}}{\cancel{e^{-\frac{1}{2\sigma^2}x^2}}\; e^{-\frac{1}{2\sigma^2}\mu_0^2}\; e^{\frac{1}{2\sigma^2}2x\mu_0}}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}\mu_1^2}\; e^{\frac{1}{2\sigma^2}2x\mu_1}}{e^{-\frac{1}{2\sigma^2}\mu_0^2}\; e^{\frac{1}{2\sigma^2}2x\mu_0}}$$

$$= e^{-\frac{1}{2\sigma^2}\mu_1^2+\frac{1}{2\sigma^2}2x\mu_1+\frac{1}{2\sigma^2}\mu_0^2-\frac{1}{2\sigma^2}2x\mu_0}$$

$$= e^{\frac{1}{2\sigma^2}\left(2x\mu_1-2x\mu_0+\mu_0^2-\mu_1^2\right)}$$

$$= e^{\frac{1}{2\sigma^2}\left(2x\left(\mu_1-\mu_0\right)+\mu_0^2-\mu_1^2\right)}$$

4)

Since we have equal proir, the odds ratio for the LDA model is:

$$e^{\frac{1}{2\sigma^2}(2x(\mu_1-\mu_0)+\mu_0^2-\mu_1^2)}$$

The odds ratio of the logistic model is:

$$\frac{P(Y=1|X)}{P(Y=0|X)} = e^{\beta_0+\beta_1 x}$$

Hence, the log odds are:

$$\frac{1}{2\sigma^2}(2x(\mu_1-\mu_0)+\mu_0^2-\mu_1^2) \quad \text{—— LDA}$$

$$\beta_0+\beta_1 x \quad \text{—— Logistic}$$

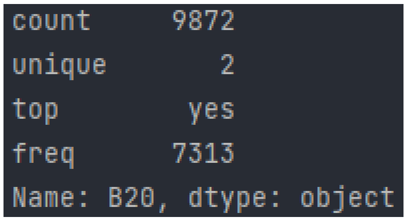
Therefore, both the LDA & logistic log odds are linear functons of x which allows them to produce linear decision boundaries.

5)

The difference between the two models is where the coefficients & intercepts (β_0 & β_1) of the logistic model are estimated using maximum likelihood whereas the coefficients & intercepts of the LDA model are computed using the estimated mean & variance from a normal distributon. Furthermore, the LDA also assumes that observations are drawn from a normal distribution. Thus, when this assumption holds, it provides a better model fit than the logistic regression otherwise, the logistic regression performs better than the LDA if the normality assumption of the observations is violated

Exercise 3

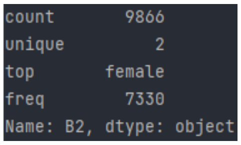
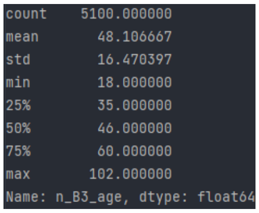
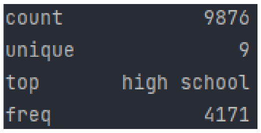
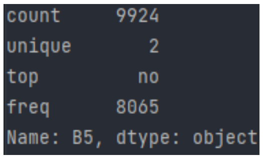
2)

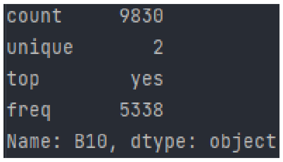
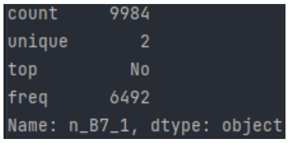
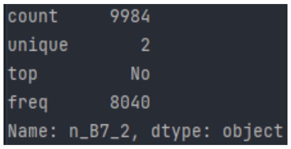
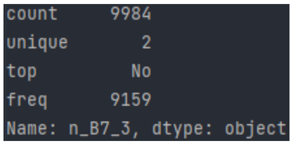
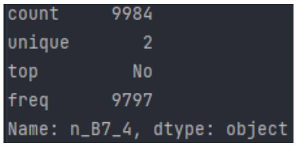
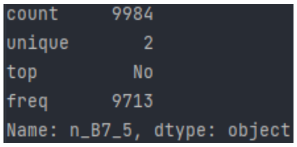
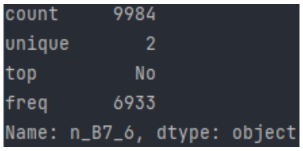
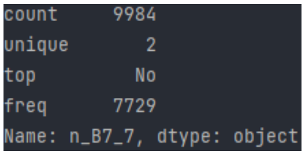


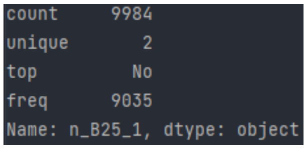
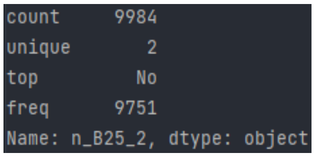
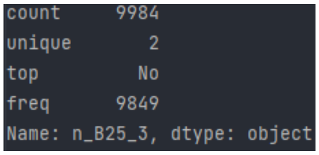
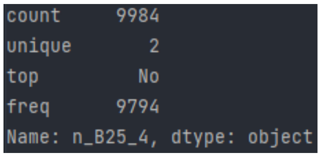
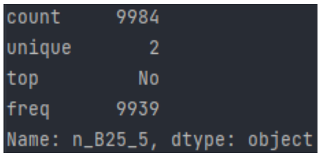
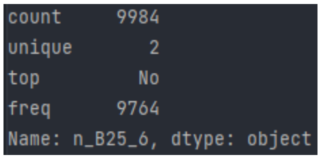
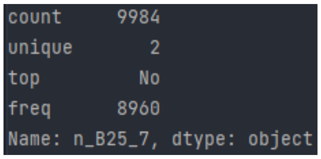
```
count     9872
unique       2
top        yes
freq      7313
Name: B20, dtype: object
```

Using the command .describe() in Python, I am able to determine that there are a total of 9872 observations of the outcome variable B20, with the mode being “yes” that the individuals have a bank account. Out of the 9872 observations, 7313 of the individuals answered yes.

3)
Variables chosen to predict the likelihood of having a bank account:
'B2', 'n_B3_age', 'B4', 'B5', 'B10', 'n_B7_1', 'n_B7_2', 'n_B7_3', 'n_B7_4', 'n_B7_5', 'n_B7_6', 'n_B7_7', 'n_B25_1', 'n_B25_2', 'n_B25_3', 'n_B25_4', 'n_B25_5', 'n_B25_6', 'n_B25_7'.

Variables	**Why this variable matters**	**Descriptive Statistics**
'B2'	Men could more likely be breadwinners and thus, having a bank account when compared to women.	 count 9866 unique 2 top female freq 7330 Name: B2, dtype: object
'n_B3_age'	People that are of young or very old age i.e. below 18 or above 85 years old are less likely to have a bank account.	 count 5100.000000 mean 48.106667 std 16.470397 min 18.000000 25% 35.000000 50% 46.000000 75% 60.000000 max 102.000000 Name: n_B3_age, dtype: float64
'B4'	People who are better educated are more likely to have a job and get paid thus, having a bank account for their salary.	 count 9876 unique 9 top high school freq 4171
'B5'	People who have received a loan from a bank or financial institution would most likely have a bank account in order to receive the loan.	 count 9924 unique 2 top no freq 8065 Name: B5, dtype: object

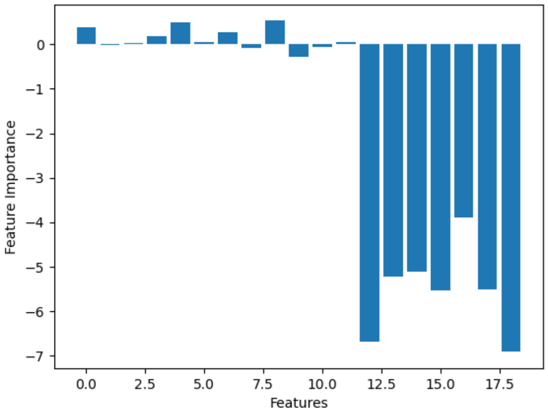
‘B10’	BancoEstado is a public bank in Chile thus if the individual is informed about the bank products, they are more likely to have a bank account created with them.	 count 9830 unique 2 top yes freq 5338 Name: B10, dtype: object
'n_B7_1'	The following variables indicate that if an individual agrees with most of these variables being advantages of having a bank account, it implies that they are more likely to have a bank account since they recognise the benefits of having a bank account.	 count 9984 unique 2 top No freq 6492 Name: n_B7_1, dtype: object
'n_B7_2'		 count 9984 unique 2 top No freq 8040 Name: n_B7_2, dtype: object
'n_B7_3'		 count 9984 unique 2 top No freq 9159 Name: n_B7_3, dtype: object
'n_B7_4'		 count 9984 unique 2 top No freq 9797 Name: n_B7_4, dtype: object
'n_B7_5'		 count 9984 unique 2 top No freq 9713 Name: n_B7_5, dtype: object
'n_B7_6'	If an individual agrees with this variable that having a bank account provides no advantages, they are less likely to have a bank account.	 count 9984 unique 2 top No freq 6933 Name: n_B7_6, dtype: object
'n_B7_7'	If an individual agrees with other advantages of having a bank account, it implies that they are more likely to have a bank account since they recognise the other benefits of having a bank account.	 count 9984 unique 2 top No freq 7729 Name: n_B7_7, dtype: object

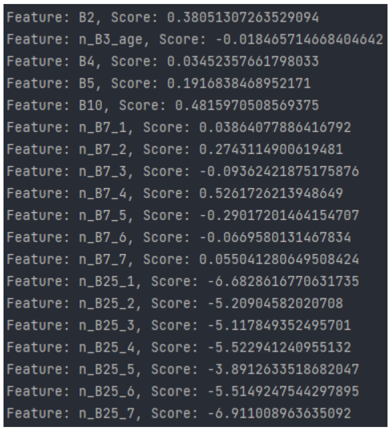
'n_B25_1'	If individuals agree that they would use a bank account due to most of these situations, it would indicate that they are more likely to have a bank account as they recognise the purpose of having a bank account	 count 9984 unique 2 top No freq 9035 Name: n_B25_1, dtype: object
'n_B25_2'		 count 9984 unique 2 top No freq 9751 Name: n_B25_2, dtype: object
'n_B25_3'		 count 9984 unique 2 top No freq 9849 Name: n_B25_3, dtype: object
'n_B25_4'		 count 9984 unique 2 top No freq 9794 Name: n_B25_4, dtype: object
'n_B25_5'		 count 9984 unique 2 top No freq 9939 Name: n_B25_5, dtype: object
'n_B25_6'	If individuals agree with this variable, it indicates that they do not know what to use a bank account for this, indicating that they are less likely to have a bank account.	 count 9984 unique 2 top No freq 9764 Name: n_B25_6, dtype: object
'n_B25_7'	If there are other reasons that an individual would use a bank account, it would indicate that there are particular features of having a bank account that entices them thus, indicating that they are more likely to have a bank account.	 count 9984 unique 2 top No freq 8960 Name: n_B25_7, dtype: object

5)
AUC scores for knn, logistic, decision tree and random forest respectively are:
[0.853, 0.990, 0.965, 0.985]

Therefore, the best classifier amongst these four classifiers is the Logistic Classifier.

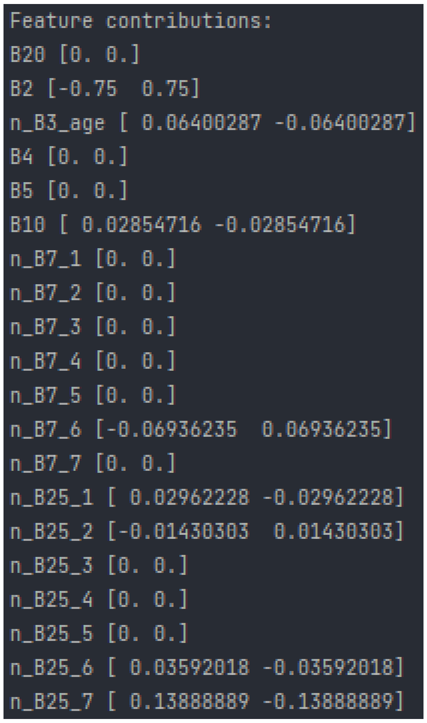
6)
The feature importance includes both positive and negative values since the positive scores indicate a feature that predicts class 1 (i.e. yes), whereas the negative scores indicate a feature that predicts class 0 (i.e. no).




```
Feature: B2, Score: 0.38051307263529094
Feature: n_B3_age, Score: -0.018465714668404642
Feature: B4, Score: 0.03452357661798033
Feature: B5, Score: 0.1916838468952171
Feature: B10, Score: 0.4815970508569375
Feature: n_B7_1, Score: 0.03864077886416792
Feature: n_B7_2, Score: 0.2743114900619481
Feature: n_B7_3, Score: -0.09362421875175876
Feature: n_B7_4, Score: 0.5261726213948649
Feature: n_B7_5, Score: -0.29017201464154707
Feature: n_B7_6, Score: -0.0669580131467834
Feature: n_B7_7, Score: 0.055041280649508424
Feature: n_B25_1, Score: -6.6828616770631735
Feature: n_B25_2, Score: -5.20904582020708
Feature: n_B25_3, Score: -5.117849352495701
Feature: n_B25_4, Score: -5.522941240955132
Feature: n_B25_5, Score: -3.8912633518682047
Feature: n_B25_6, Score: -5.5149247544297895
Feature: n_B25_7, Score: -6.911008963635092
```

From the output in the figure on the right, we determine that the top 3 most important features are n_B25_7, n_B25_1 and n_B25_6 which are “If you had a bank account, what would you use it for?: Other”, “If you had a bank account, what would you use it for?: To save up for emergency” and “If you had a bank account, what would you use it for?: Don’t know”.

7)
The feature contributions of married individuals without a bank account are as follows:



```
Feature contributions:
B20 [0. 0.]
B2 [-0.75  0.75]
n_B3_age [ 0.06400287 -0.06400287]
B4 [0. 0.]
B5 [0. 0.]
B10 [ 0.02854716 -0.02854716]
n_B7_1 [0. 0.]
n_B7_2 [0. 0.]
n_B7_3 [0. 0.]
n_B7_4 [0. 0.]
n_B7_5 [0. 0.]
n_B7_6 [-0.06936235  0.06936235]
n_B7_7 [0. 0.]
n_B25_1 [ 0.02962228 -0.02962228]
n_B25_2 [-0.01430303  0.01430303]
n_B25_3 [0. 0.]
n_B25_4 [0. 0.]
n_B25_5 [0. 0.]
n_B25_6 [ 0.03592018 -0.03592018]
n_B25_7 [ 0.13888889 -0.13888889]
```

Therefore, the top 3 characteristics of married individuals without a bank account are B2, n_B25_7 and n_B7_6 which are the “Respondent’s gender”, “If you had a bank account, what would you use it for?: Other” and “Main advantages of having a bank account: None/no advantages” respectively, from the most important to the least important characteristic

8)
Based on the data provided and variables used, I determined that the features mentioned in 6) are of the highest importance when trying to promote the use of bank account in the population. With this information, we can conduct further surveys on these variables to determine the individuals without a bank account and encourage these individuals to open a bank account through methods such as targeted advertising. Furthermore, the top 3 characteristics of married individuals without a bank account would indicate the general consensus to the survey questions of married individuals without a bank account. Thus, we can target these married individuals with the questions they answered and encourage these individuals to sign up for a bank account.