Applied probability and statistics
Assignment -2

Name: Nikhil S61
Reg: 2410580024
Big Data Analytics

1) a) How often does she forecast rain?

forecast rain = P(forecast rain & Actual rain) + P(forecast rain & no rain)

= 0.4 + 0.2 ⇒ 0.6

no of times she predicted that forecast rain.

b) How often does she make a mistake?

no of times she predicted forecast and it is not a match with the actual truth.

= P(forecast rain & no rain) + P(forecast no rain & Actual rain)

= 0.2 + 0.15 ⇒ 0.35

c) Given that she just forecast rain, what is the chance that it will actually rain? (conditional probability):

$$P(\text{Actually rain}/\text{forecast rain}) = \frac{P(\text{Actually rain} \cap \text{forecast rain})}{P(\text{forecast rain})}$$

$$= \frac{0.4}{0.6} = 0.67$$

d) Given that it rains today, what is the probability that she had forecast rain in last night's broadcast?

$$P(\text{forecast rain}/\text{actual rain}) = \frac{P(\text{forecast rain} \cap \text{actually rain})}{P(\text{actually rain})}$$

P(actually rain) = P(forecast rain & Actual rain) + P(forecast no rain & Actual rain)

P(actually rain) = 0.4 + 0.15 = 0.55

$$\Rightarrow \frac{0.4}{0.55} \Rightarrow 0.72$$

2) F ⇒ 52% M ⇒ 48% CS ⇒ 5% finc ⇒ 2%

a) The student is female given that the student is majoring in computer science

P(F) = 52% ⇒ 0.52 P(CS) = 5% ⇒ 0.05 P(F ∩ CS) = 2% = 0.02

$$P(F/CS) = \frac{P(CS/F) * P(F)}{P(CS)} \Rightarrow P(CS/F) = \frac{P(CS \cap F)}{P(F)} = \frac{0.02}{0.52}$$

P(CS/F) = 0.038

$P(F/CS) = \frac{P(CS/F) \times P(F)}{P(CS)} \Rightarrow \frac{0.038 \times 0.52}{0.05} = 0.395 \simeq 0.40$

b) This student is majoring in computer science given that the student is female.

$P(CS/W) = 0.038\%$ calculated in the above problem.

3) $P(BR) = 20\% \Rightarrow 0.20$ $P(AR) = 40\% \Rightarrow 0.40$ $P(GR) = 40\% \Rightarrow 0.40$

$P(mh/BR) = 25\% \Rightarrow 0.25$ $P(mh/AR) = 20 \Rightarrow 0.20$ $P(mh/GR) = 10\% \Rightarrow 0.1\%$

$P(mh^c/Risk) = 1 - P(Mh/Risk) \Rightarrow P(mh/BR) + P(mh/AR) + P(mh/GR)$

$\Rightarrow \frac{(mh \cap Risks)}{P(Risks)} \Rightarrow \frac{(0.25 \times 0.2) + (0.4 \times 0.2) + (0.4 \times 0.1)}{0.05}$ ($P(BR)$, $P(AR)$, $P(GR)$)

$0.05 + 0.08 + 0.04 \Rightarrow 0.17$

$\Rightarrow 1 - 0.17 = \boxed{0.83}$

4) $P(S) = 0.002$ $P(e) = 0.002$ $P(CS) = 0.01$ $P(he) = 0.001$

$P(Cr/S) = 0.25$ $P(Cr/e) = 0.30$ $P(Cr/CS) = 0.90$ $P(Cr/he) = 0.10$

~~$P(CS/Cr)$~~ $P(CS/Cr) = \frac{P(Cr/CS) \times P(CS)}{P(Cr)} \Rightarrow \frac{0.90 \times 0.01}{P(Cr)} =$

Law of total probability

$P(Cr) = (P(S) \times P(Cr/S)) + (P(e) \times P(Cr/e)) + (P(CS) \times P(Cr/CS)) + (P(he) \times P(Cr/he))$

$(0.002 \times 0.25) + (0.002 \times 0.30) + (0.01 \times 0.90) + (0.001 \times 0.10)$

$0.0005 + 0.0006 + 0.009 + 0.0001$

$P(Cr) = 0.0102$

$\Rightarrow \frac{0.90 \times 0.01}{0.0102} = \boxed{0.882}$

5) $P(L_1) = 0.80$ $P(L_2) = 0.20$

$P(W/L_1) = 0.2$ $P(W/L_2) = 0.9$

i) $P(L_1/W) = \frac{P(W/L_1) \times P(L_1)}{P(W)}$

$P(W) = P(W/L_1) \times P(L_1) + P(W/L_2) \times P(L_2)$

$= 0.8 \times 0.2 + 0.2 \times 0.9$

$P(W) = 0.16 + 0.18 = 0.34$

$$P(L_1/W) = \frac{P(W/L_1) \times P(L_1)}{P(W)} = \frac{0.2 \times 0.8}{0.34} = 0.470.$$

$$P(L_2/W) = \frac{P(W/L_2) \times P(L_2)}{P(W)} = \frac{0.9 \times 0.2}{0.34} = 0.529$$

there is a 47% that the robot is in location L_1 and there is a 53% that the robot is in location L_2.

6) a) Random donor and Random receiver.
Select the donor and check for all recipient that can take it. for all the donors.

P (Donor (All) and recipient (who all can take))

mutual exclusive events and Donor and recipient are Independent. so that add all the Donors and recipients.

$P(O^- \cap (\text{who all can take})) + P(O^+) + P(A^- \cap \_) + P(A^+ \cap \_)$
$P(B^- \cap \_) + P(B^+ \cap \_) + P(AB^- \cap \_) + P(AB^+ \cap \_)$

* O^- ∩ who all can take
$0.066\,[0.066 + 0.374 + 0.063 + 0.357 + 0.015 + 0.085 + 0.006 + 0.034] = 0.066$

* O^+ ∩ who all can take.
$0.374\,[0.374 + 0.357 + 0.085 + 0.034] = 0.3179$

* A^- ∩ who all can take.
$0.063\,[0.063 + 0.357 + 0.006 + 0.034] = 0.0289$

* A^+ ∩ who all can take
$0.357\,[0.357 + 0.034] = 0.1395.$

* B^- ∩ who all can take.
$0.015\,[0.015 + 0.085 + 0.006 + 0.034] = 0.0021$

* B^+ ∩ who all can take
$0.085\,[0.085 + 0.034] = 0.0101$

* AB^- ∩ who all can take
$0.006\,[0.006 + 0.034] = 0.0002$

* AB^+ ∩ who all can take
$0.034\,[0.034] = 0.0011$

P(donor(All) and (who all) can take) = 0.0667 + 0.3179 + 0.0289 + 0.1395 + 0.0021 + 0.0101 + 0.0002 + 0.0011 ≈ 0.56

≈ 56%

b) O^- is the universal donor the blood drives should focus more on O^- blood group. the transfusion policies is prioritize in keeping stocks of O^- blood for emergencies. blood transfusion priority should be according to the. Rear blood types. likes $AB-$ and look for other matching or receiving groups here if the blood is in stock of O^- if the recipient need O^- it should be prioritized first.

c) The given case where we have limited time to transfusion of blood. because of the limited time and only A^+ blood type is in availability. it is best to do the blood sampling for the wounded solider. Here we can get out any blood group. but if the blood type is A^+ or AB^+ of the wounded. we can immediate start transfusion if the blood group comes out to be of different blood type. we are helpless in this situations. and wait for any miracle to happen.

7) In a future society we assume that the law enforcement is strong and even if a crime is committed it can be easily or can be detected early after crime commitement. now in this case to ensure no innocent people are imprisoned. we should make the False positive near to zero so that any Innocent people should be not test / classified as positive (a crime maker) by the machine, so the false positive is low. so does the ~~recall, TPR, sensitivity~~ precision. is high for the machine.

8) a) Recall is more relevant performance metric. here because the false negative should be less so that the Recall is high [If false negative is high we see that model is predicting a spam mail as not spam and our data may be lost) so FN ↓ Recall ↑ (TPR, sensitivity) ↑.

Increasing the classification threshold generally decreases FP

c) when the classification threshold increases, precision definitly increases.

d) when the classification threshold increases, quantity of TP (True positive) decreases.

e) when the classification threshold is decreased, the quantities TN and FN both uniformly, decrease.

f) decreasing the classification threshold generally decreases FN

g) when the classification threshold is decreased, recall definitly Increases

h) when the classification threshold is decreased the quantities #P and FP both uniformly, Increases

9) a) An expensive and critical hydro metric turbine operates 23 hours a day. An ML model evalutes vibration patterns and predicts when the turbine is operating without anomaly with an accuracy 99.99%. This the ML model that its accuracy value high suggest that the model is doing good job, because assuming out of 1000 turbine the model is accurately predicting that 99.9 times that the turbine are working correctly it can not be less than 99.99% because if it less then this then their might be a damage in the system and may led to the crash and a huge property loss & life loss may happen so this should be high and here the model predicates it is a good model.

10) a) If a model A has better precision and better recall than model B, then model A is probably better. as the F1 score is high for model A so it is doing better.

11) An ROC curve is a plot of True positive rate vs False positive rate for different thresholds.

12) lowering the threshold classifies more items as positive, thus increase in both True positive & False positive values

13) c) No change. AUC only cares about relative prediction probabilities.

14) It is a good model, because every time it predicts something the opposite happens. So we can think of the opposite decisions on prediction, we can help the model to go from zero to hero by telling the opposite or negation of the model as the model will mostly predicate the exact wrong prediction.

15) a) Dashed black line represents random classification (True)
b) ROC curve for any model can't fall below the dotted black line (False)
c) The model represented by solid blue line is better than that represent by solid line (False)

16) TN → True negative metric does not play a role in forming the precision recall curve, because recall = $\frac{TP}{TP+FN}$, precision = $\frac{TP}{TP+FP}$
So here we clearly see that the value of TN is not need to create this recall and precision metrics.

17) a) shows the overfits model.
b) shows the good fit model.
c) shows the underfit model.

18) a) metric = area under ROC curve because classify. the class as positive can buy, negative will not buy.
b) metric: precision - recall curves are more good with imbalanced datasets and focus mostly on positive classes.