Solutions Online Test 2

Instructions: The test is worth up to 5 points. Each correct response is worth 1 point. There is a bonus question. The questions are based on the churn reading/data discussed in class. Download the files "customerchurn.csv", "DataAnalyticsFunctions.R" and "OnlineTest2.R" to your working directory for this assignment. Open and execute the code "OnlineTest2.R" to answer the following questions.

1. There is substantial value in understanding what influences churn rates. In particular, customer stickiness (the nature of your customers to continue to use your products or services, to "stick" with you) is a relevant aspect to consider. Which of the following visualizations provides evidence of "customer stickiness"?

a) plot(factor(Churn) ~ factor(gender), data=churndata, col=c(8,2), ylab="Churn Rate", xlab="Gender")

b) plot(factor(Churn) ~ factor(SeniorCitizen), data=churndata, col=c(8,2), ylab="Churn Rate", xlab="Senior Citizen")

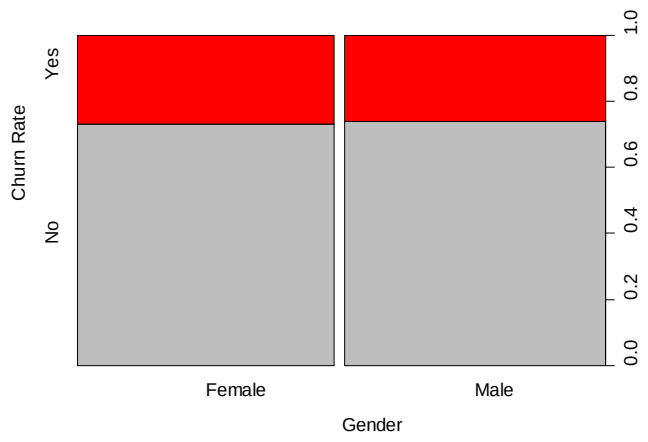
c) plot(factor(Churn) ~ MonthlyCharges, data=churndata, col=c(8,2), ylab="Churn Rate", xlab="Monthly charges")

d) ✓plot(factor(Churn) ~ tenure, data=churndata, col=c(8,2), ylab="Churn Rate", xlab="Tenure (months)")

In the plots below we are looking at how churn rates changes across different demographics. In this question we need to provide a visualization that provides evidence of customer stickiness. In this case tenure seems to provide a way to verify that. In contrast, gender, age, and monthly charges are not capturing "stickiness" as they do not have a time dimension. Next we discuss each visualization.

The command in (a) plots Churn (as a factor) by gender (as a factor)

```
> plot(factor(Churn) ~ factor(gender), data=churndata, col=c(8,2), ylab
="Churn Rate", xlab="Gender")
```



This visualization suggests that there is no substantial difference on churn rates between female or male customers overall. It is not telling anything about how these demographics are continuing (or not) to use the services.

R Note: It is always convenient to inform R when some the variable is a factor instead of numbers. In the case above, Churn can be "Yes" or "No" while gender can be "Female" or "Male" so R automatically treats them as factors. Indeed you would get the same output from the command

```
> plot(factor(Churn) ~ factor(gender), data=churndata, col=c(8,2), ylab
="Churn Rate", xlab="Gender")
```

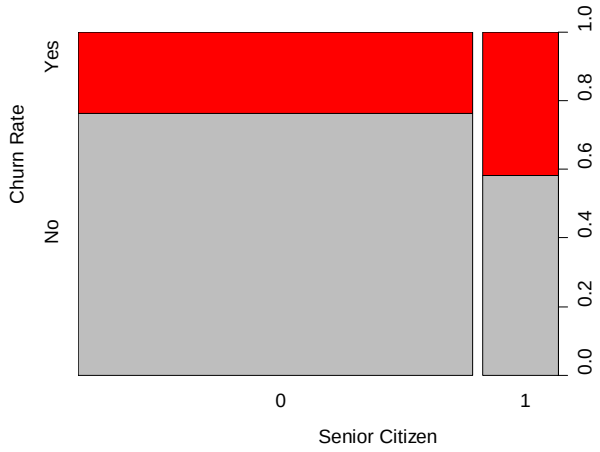
if we did not apply "factor" to Churn and gender as follows

```
> plot( Churn ~ gender, data=churndata, col=c(8,2), ylab="Churn Rate",
xlab="Gender")
```

However, it is good practice to use it. (In the next option things are different.)

The line command (b) plots churn as it changed by age group (senior citizen or not senior citizen).

```
> plot(factor(Churn)~factor(SeniorCitizen), data=churndata, col=c(8,2),
ylab="Churn Rate", xlab="Senior Citizen")
```

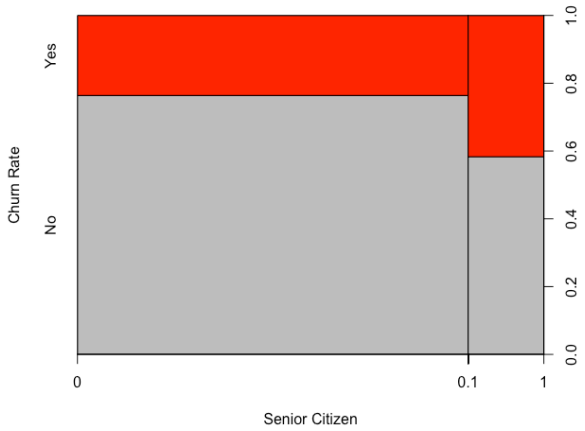


This visualization suggests that senior citizens are more likely to churn than non-senior citizens. (By nearly 20%.) Again this does not indicate how these customers are continuing to use the service. Thus it does not portrait stickiness properly.

R Note: It is always convenient to inform R when some variables are factor instead of numbers. This is the case for the dummy variable SeniorCitizen which takes values 0 or 1. The figure above was generated with `factor(SeniorCitizen)` *in the call which informs R to treat the variable* `SeniorCitizen` *as a factor. Otherwise the command*

```
> plot(factor(Churn)~SeniorCitizen, data=churndata, col=c(8,2), ylab="C
hurn Rate", xlab="Senior Citizen")
```

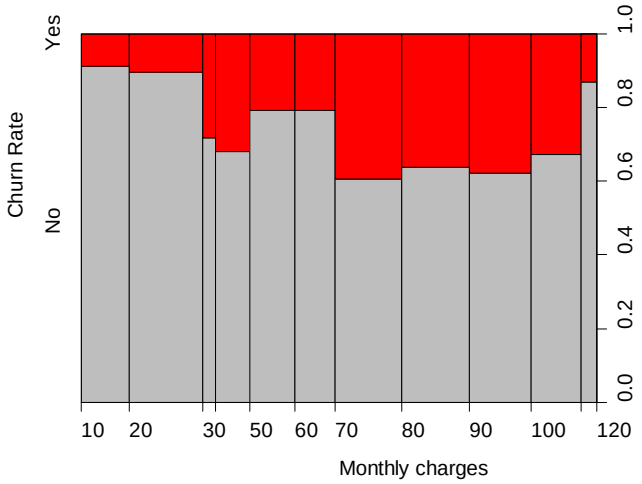
yields the following output



which has some strange features. In particular it uses an artificial threshold .1 to split the values of Senior Citizen. This should definitely be avoided (by simply using the factor command.)

Command (c) plots how churn rates varies across different pricing points

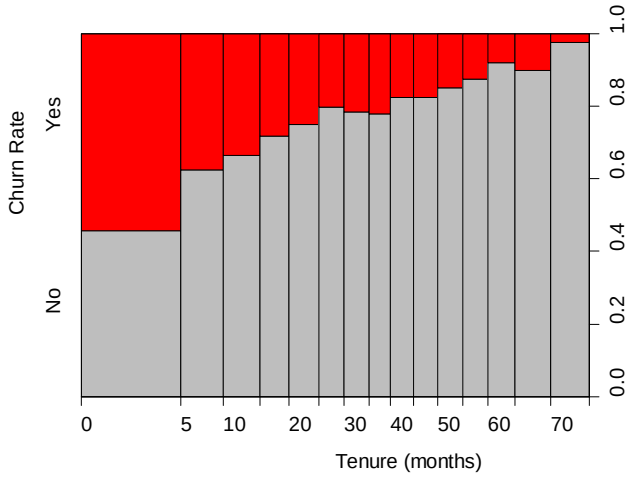
```
> plot(factor(Churn) ~ MonthlyCharges, data=churndata, col=c(8,2), ylab
="Churn Rate", xlab="Monthly charges")
```



Not a clear relationship. This suggests the potentially existence of different customers segments. Thus, this visualization is more suitable to display customers' price sensitivity (it can be very useful for pricing strategies, determining optimal pricing, etc) but one needs to be concerned with the fact that prices were by the offering of different services, and were not set at random which creates endogeneity problems (relates to "Causal Analysis"). For instance, the most expensive bucket has a small churn rate. This might be because they are paying for many services and that price is fair or that they were given such high monthly charges because they were identified as customers that are not likely to churn. All interesting points but does not help with the customer stickiness.

Finally, command (d) plots how churn rates change with customer tenure.

```
> plot(factor(Churn) ~ tenure, data=churndata, col=c(8,2), ylab="Churn Rate", xlab="Tenure (months)")
```



Clear an inverse relation between tenure and churn rate. Indeed this shows that churn rate is decreasing as customers stay longer. The longer a customer stays the lower is the change of him to churn. This is providing evidence that the company does enjoy some sort of stickiness as customers become more and more likely to continue to use their services.

2. We would like to know if long term customers (with higher tenure) are receiving discounts in their monthly fees or not. Which of the following provides the best argument?

a) Long term customers are more loyal to the company and therefore they are willing to pay more to stay with the company. This is exemplified by the positive correlation between the variables
cor(churndata$tenure,churndata$MonthlyCharges)
plot(MonthlyCharges~tenure, data=churndata, xlab='Monthly charges (dollars)', ylab='Tenure (months)', main='Churn')

b) Long term customers have contracts which are older and hence more expensive as the technology industry reduces costs over time.
res_tenure.simple <- glm(MonthlyCharges~tenure, data=churndata)
summary(res_tenure)
This is confirmed by the positive coefficient which is statistically significant.

c) ✓The company seems to price only based on service features and does not seem to price discriminate based on customer's characteristics (as those coefficients are not statistically significant).

res_tenure <- glm(MonthlyCharges~.-customerID-Churn-TotalCharges, data=churndata)

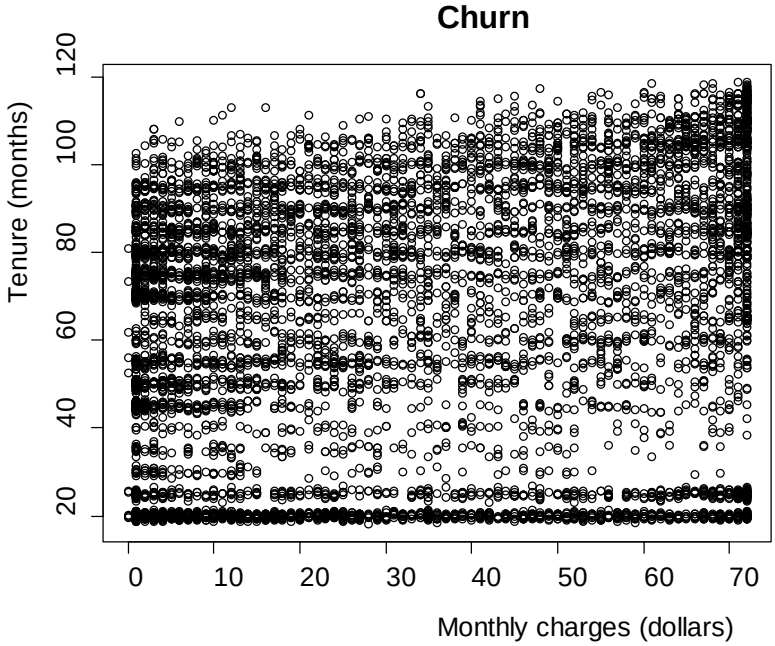
d) After accounting for a quadratic trend, and adding a new variable to the model, we see that the company also seems to price discriminate based on tenure (although with decreasing impact).

```
churndata$tenure.sq <- churndata$tenure^2
res_tenure.sq <-glm(MonthlyCharges~.-customerID-Churn-TotalCharges, data=churndata)
summary(res_tenure.sq)$coef[,c(1,4)]
churndata<-churndata[,(names(churndata)!="tenure.sq")]
```

Option a) argues that long term customers are more loyal and are willing to pay more to stay. Correlation information suggests a positive relation

```
> cor(churndata$tenure,churndata$MonthlyCharges)
[1] 0.2478999
```

And a plot of tenure by monthly charges does not reveal much insight on the claim. In particular, the picture is quite unappealing. (Also keep in mind that correlation is easy to understand but does not quantify impact of one variable on the other.) Overall the argument that long term customers are more loyal and are willing to pay more to stay is inconclusive.



Option b) uses a simple regression with MonthlyCharges being explained by tenure

```
> res_tenure.simple <- glm(MonthlyCharges~tenure, data=churndata)
> summary(res_tenure.simple)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.92978    0.57476   95.57   <2e-16 ***
tenure       0.30372    0.01415   21.47   <2e-16 ***
```

R Note: Recall that glm will run a linear regression if no parameter "family" is specified. `MonthlyCharges~tenure` *tells R that* `MonthlyCharges` *is used as the Y variable and tenure is used as the X variable. The command* `summary(res_tenure.simple)` *simply displays the output of the regression (coef, SE, p-value, etc.)*

The issue with this approach is potential confounding factors. In particular, this does not account for the different services being provided. For example, customers with high tenure might be asking for more services (on average) and therefore are paying more (not because of tenure but because of additional service). Keep in mind that one is limited by the data it uses (so spend time thinking about the good and bad aspects of the available data).

Option c) decides to run a multiple linear regression to explain monthly charges using many variables and displays the coefficient of tenure:

```
> res_tenure <- glm(MonthlyCharges~.-customerID-Churn-TotalCharges, dat
a=churndata)
> summary(res_tenure)$coef[,c(1,4)]
                                      Estimate  Pr(>|t|)
tenure                          1.454668e-04 0.8621363
```

R Note: Recall that glm will run a linear regression if no parameter "family" is specified in the call. Using `MonthlyCharges~.-customerID-Churn-TotalCharges` *we are telling R to use* `MonthlyCharges` *as the response Y variable. Thepiece "*`~.`*" tells R to use all the available columns of the data (*`data=churndata`*) but we chose to remove 3 of them. The removal of the column from the regression is simply "-" followed by the name of he column. Here* `~.-customerID-Churn-TotalCharges` *excludes customerID (which is an unique identifier for each observation), Churn, and Total charges.*

On average (holding all else constant), after controlling for type of service and other information, tenure is not statistically significant. (Indeed, the corresponding p-value of 0.862 is very large.) The company does not seem to be exploiting tenure in its pricing strategy.

Finally option d) aims to add the flexibility of a non-linear trend with respect to tenure by adding "tenure.sq" (= tenure squared) in the model. It is always possible that the linear specification can be improved. However, when the new model is fitted, we obtain

```
                                       Estimate  Pr(>|t|)
tenure                           -9.006395e-04 0.6726508
tenure.sq                         1.526173e-05 0.5935532
```

where the p-values of both coefficients are large. This means that they are not statistically significant which seems to support the same conclusion as in option c). Not the new conclusion stated in option d).

R Note: The code for this option

```
> churndata$tenure.sq <- churndata$tenure^2
>res_tenure.sq<-glm(MonthlyCharges~.-customerID-Churn-TotalCharges,data=churndata)
> summary(res_tenure.sq)$coef[,c(1,4)]
> churndata<-churndata[,(names(churndata)!="tenure.sq")]
```

works as follows. The first line created a new column in the dataframe churndata (which is our data). The column is called tenure.sq. The second line runs a linear regression. The third line displays the coefficients of the regression. Finally, the last line removes the created column from the data.

3. Based on the logistic regression model with all variables (except customerID) discussed in class,

```
result.logistic<-glm(Churn~.-customerID, data=churndata,family="binomial")
```

among the customers 101 to 110, what is the highest probability of churn and how much is it?

```
predict(result.logistic,newdata=churndata[101:110,])
which.max(predict(result.logistic,newdata=churndata[101:110,],type="response"))
max(predict(result.logistic,newdata=churndata[101:110,],type="response"))
```

a) It is customer number 110 and it is below 1%

b) ✓**It is customer number 106 and it is near 45%**

c) It is customer number 110 and it is above 20%

d) It is customer number 106 and it is below 45%

We fixed a model, logistic regression to explain churn based on all the variables (we need to remove customerID since it is a factor and we have a different value for each customer).

```
result.logistic<-glm(Churn~.-customerID, data=churndata,family="binomial")
```

We can use this model to predict the churn probabilities of customers 101 to 110

```
> predict(result.logistic,newdata=churndata[101:110,])
  101   102   103   104   105   106   107   108   109   110
-1.39 -1.13 -1.54 -3.35 -2.37 -0.16 -1.78 -2.60 -1.13 -6.76
```

Recall that if you just ask to predict the software return the log odds of churn. To obtain the actual probabilities you need to use type="response" in the command

```
> predict(result.logistic,newdata=churndata[101:110,],type="response")
   101    102    103    104    105    106    107    108    109    110
0.198 0.243 0.176 0.033 0.085 0.458 0.144 0.068 0.242 0.001
```

The highest probability is then (by inspection) customer 106 with 0.458 (highlighted). We can ask R to tell us which is the customer (which is quite important when you have lists with 1000s of customers to choose from).

This is what the command below does:

```
> which.max(predict(result.logistic,newdata=churndata[101:110,],type="r
esponse"))
106
  6
```

The command `which.max` *returns the index of the vector that has the maximum value. The actual probability value can be obtained by calling* `max` *instead of* `which.max` *as shown below*

```
> max(predict(result.logistic,newdata=churndata[101:110,],type="respons
e"))
[1] 0.4589779
```

4. Run the logistic regression model with all variables (except customerID) discussed in class

```
result.logistic<-glm(Churn~.-customerID, data=churndata,family="binomia
l")
```

Next, we will use the model to classify using different thresholds on the predicted probability (not necessarily .5). We will use a function in FPR_TPR.R that computes the true positive rate and false positive rate. The code below plots several choices.

```
plot( c( 0, 1 ), c(0, 1), type="n", xlim=c(0,1), ylim=c(0,1), bty="n",
xlab = "False positive rate", ylab="True positive rate")
lines(c(0,1),c(0,1), lty=2)
for ( val in seq(from = 0, to = 1, by = 0.05)  ){
values <- FPR_TPR( (result.logistic$fitted >= val),result.logistic$y)
points( values$FPR , values$TPR )
}
```

Which of the following is not true?

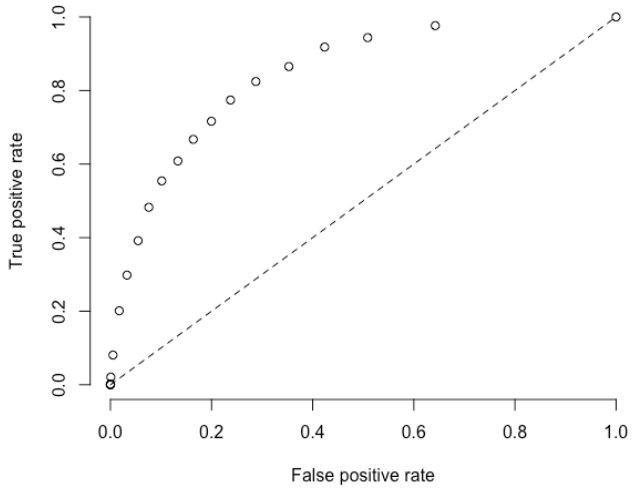
a) Good performance should be above the diagonal.

b) Although 100% accuracy seems impossible, using this predictive model we can achieve a false positive rate that is three times smaller than the true positive rate.

c) The points (0,0) and (1,1) are not interesting as they correspond to "always predict negative (no churn)" and "always predict positive (churn)" independently of the customer.

d) ✓If we choose the predicted probability threshold properly we can achieve FPR=.1 and TPR=.9 or better.

By running the commands above, we compute the ROC curve:

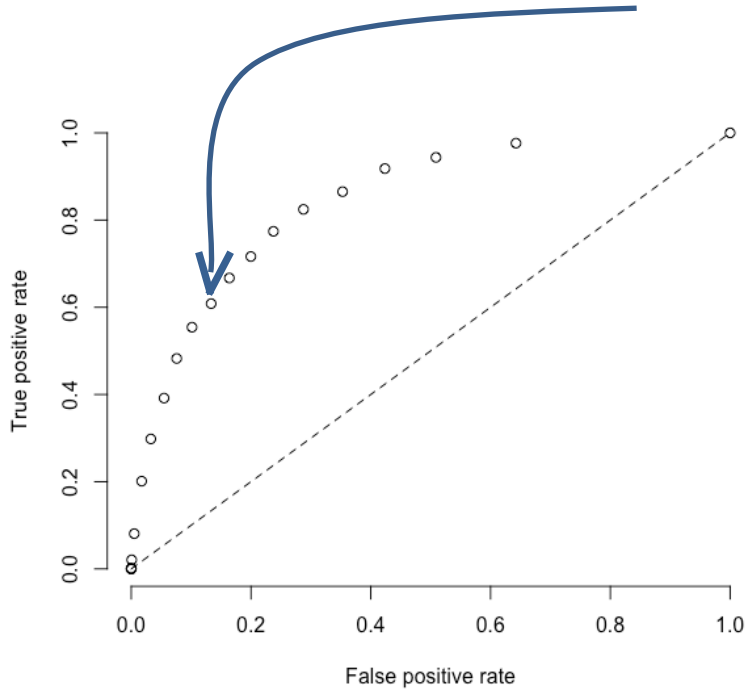


The question requests the "false" statement among the options.

Option a) is true. Indeed, in the picture moving vertically up is good (i.e. given a fixed false positive rate we prefer to have a higher false positive rate). Similarly, moving to the left is good (i.e. for a fixed true positive rate, reduce the false positive rate). Thus to be above the diagonal line is good. Recall that diagonal line corresponds to random guessing so we would like to be able to perform better than that.

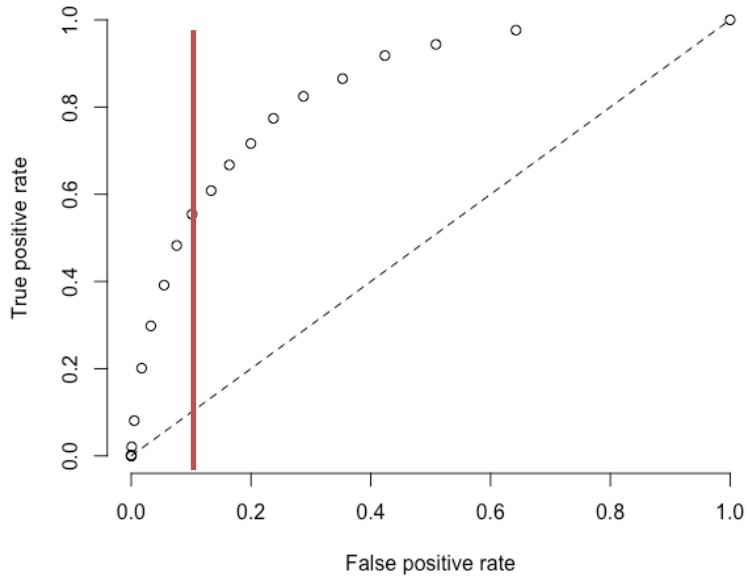
Option b) is true. 100% accuracy would correspond to 100% TPR (true positive rate) and 0% FTP (false positive rate). This would be the upper left corner of the picture (for which we do not have a circle). This 100% accuracy seems impossible.

Regarding the other statement we can pick a circle for which FPR < 0.2 and TPR > 0.6.



Option c) is true. Indeed, the point (1,1) says that we always get it right when somebody is going to churn, and we always get it wrong when somebody does not churn. Therefore, this corresponds to saying "everybody will churn" which is not very informative. Similarly for the point (0,0) which corresponds to "nobody will churn."

Option d) is false because when we restrict to FPR = 0.1 (see the red vertical line in the plot below), the maximum TPR we can achieve is below 0.6. Thus the claim of FPR=0.1 and TPR >.9 is false.



5. You are modeling the churn problem in your company. Currently, to propose an offer to avoid churn costs $5. If the offer is accepted the customer stays but you incur an additional $45 in costs. A customer has an expected value of $1000 if he/she stays with the company. Which of the following cost-benefit matrix models the setting you are concerned with?

a)

	Churn	No Churn
Offer	45	1000
No Offer	0	0

b)

	Churn	No Churn
Offer	50	1000
No Offer	0	1000

c) ✓

	Churn	No Churn
Offer	-5	950
No Offer	0	1000

d)

	Churn	No Churn
Offer	50	950
No Offer	0	1000

The relevant actions we can take are: make an offer to a customer ("Offer") or do not make an offer ("No Offer"). The potential random outcomes are: the customer churns ("Churn") or the customer stays with the company ("No Churn"). Thus we have a 2x2 cost-benefit matrix.

The proposal of the offer costs $5. If a customer stays the expected value of our gains is $1000. Thus if we do not make an offer, we get $1000 if he stays (this corresponds to the cell "No Offer, No Churn"). Moreover, if we do not make an offer and the customer leaves we make $0 (this corresponds to the cell "No Offer, Churn").

Moreover, if the offer is accepted an additional cost of $45 is incurred and the customer stays. Therefore if we make an offer and the customer leaves, we just lose $5 (that is, we have -5 in the cell "Offer, Churn"). Finally, if we make an offer and it is accepted we obtain $1000 - $5 -$45 = $950 (which corresponds to the cell "Offer, No Churn").

6. In the Churn problem discussed and modeled in class, our decision is whether or not to make an offer to a customer while our prediction is whether a customer will churn or not. There are offers which are made available to all customers (e.g. via TV advertisement) and other offers which are exclusive (e.g. via phone calls). There are fundamental differences between these strategies. Which of the following is not true?

a) The "exclusive offers" allow the company to target most profitable customers while the "all-in offer" allows customers to self-select. The latter can attract a large number of "lemons" (i.e. bad clients) if not properly designed.

b) The "exclusive offer" requires personal information and its deployment has a more limited reach compared to the "all-in offer." Thus the latter can be interesting to broad the customer base.

c) ✓Both strategies cannot be offered simultaneously as they cannibalize each other.

d) Even though offers could in principle cannibalize each other data analytics tools can be used to set prices and discounts to help reduce the impact of cannibalization.

We are searching for the false statement.

Option (a) discusses that "all-in offer" on TV need care as people can select into them. In contrast, the "exclusive offers" are target to customers that the company can choose. (Customers do not select to get the call.)

Option (b) states that the "exclusive offer" requires some personalized information which can potentially limit the "reach" of the campaign in contrast to the "all-in offer" which can reach customers outside your base.

Option (c) is false. Although the offers can cannibalize each other (as people who received the call can choose which one to pick), this is not pose a barrier that prevent them to be applied simultaneously. Indeed, one can design offers to distinguish different types of customers. Moreover, a slightly better "exclusive offer" might make the customer feel special once he has the knowledge about the "all-in offer" that provides an anchor/baseline for the customer's decision.

Option (d) explains a possible way to circumvent conflict (cannibalization) between offers.