Business Analytics: Homework 2

Qingyuan Dong (UNI: qd2145)

12/01/2018

1 Question 1 (Classification)

(a) Should DogBark send pamphlets to all the potential customers included in the dataset `dog.csv`?

DogBark should NOT send pamphlets to all the potential customers in `dog.csv`, because the expected profit is negative, as showed below:

# dog owner	1148	
# not dog owner	2145	
	dog owner	not dog owner
send	0.5	-1
not send	0	0
expected profit	1148*0.5+2145*-1=	-1571

Figure 1: Expected profit of sending to all potential customers

(b) Load the data from `dog.csv` and split your sample into training (75%) and validation (25%). We will not use a test dataset for this exercise. Use the command `set.seed(4650)` to set the randomizer's seed. Print the summary of the training data.

```
> dog = read.csv("dog.csv")
> attach(dog)
> set.seed(4650)
> train = sample(1:nrow(dog), 0.75*nrow(dog))
> test = (1:nrow(dog)) [ ( (1:nrow(dog)) %in% train) == FALSE]
>
> dog.train = dog[train, ]
> dog.test = dog[test, ]
> summary(dog.train)

  dog         pub_dist        supermaket_dist     laundry_dist
 No :1611   Min.   :   7.782   Min.   :   1.502   Min.   :   1.33
 Yes: 858   1st Qu.: 176.556   1st Qu.: 323.726   1st Qu.: 203.25
            Median : 407.435   Median : 546.770   Median : 553.16
```

```
           Mean   : 671.102   Mean   : 572.407   Mean   : 601.03
           3rd Qu.: 980.241   3rd Qu.: 794.781   3rd Qu.: 905.40
           Max.   :2000.000   Max.   :1746.658   Max.   :2000.00
   park_dist        neigh_density_score   tree_score
 Min.   :   1.018   Min.   :3.001       Min.   :  1.329
 1st Qu.: 630.520   1st Qu.:4.723       1st Qu.: 31.119
 Median : 929.864   Median :6.545       Median : 51.371
 Mean   : 964.001   Mean   :6.516       Mean   : 63.796
 3rd Qu.:1256.990   3rd Qu.:8.333       3rd Qu.:102.668
 Max.   :2000.000   Max.   :9.995       Max.   :149.949
```

(c) A first idea is to use `tree_score` to classify potential customers (the hypothesis being be that dog owners might seek residences located close to a locale with many trees). To use this idea, we would set a threshold for this variable such that we send pamphlets only to customers whose `tree_score` is above (or below) that threshold.
Find the optimal threshold for the training data, and construct the confusion matrix for the training data. What would the resulting profit have been if DogBark inc. had used this method to target potential customers in the training data with this method.

I used the following codes to find the optimal threshold:

```
> range(tree_score)
[1]   1.329249 149.949300
>
> profits =  c(c(0,-1),c(0,0.5))
> profitPerThreshold = vector("numeric",149)
> for (s in 2:149)
+ {
+   dogpred = tree_score > s
+   # table to see error rates
+   classficationTable = table(predict = dogpred,
+                              truth = dog.train$dog )
+   profitPerThreshold[s] = sum(classficationTable * profits)
+ }
> plot(2:149,profitPerThreshold[2:149],pch = 15, xlab = "Threshold")
> best_s = which.max(profitPerThreshold[2:149]) +1
> best_s
[1] 149
```

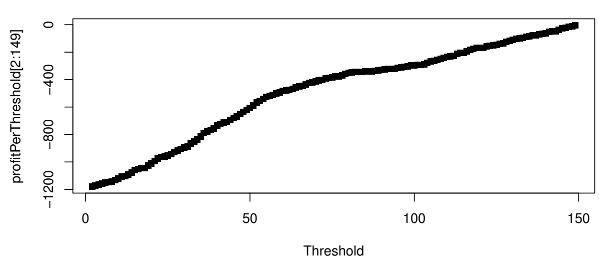

Figure 2: Net profit for different thresholds

So the optimal threshold is 149, and the confusion matrix and the resulting profit is:

```
> dogpred = tree_score > best_s
> classficationTable = table(predict = dogpred,
+                            truth = dog.train$dog)
> classficationTable
       truth
predict   No  Yes
  FALSE 1602  848
  TRUE     9   10

> sum(classficationTable * profits)
[1] -4
```

(d) Construct a confusion matrix and calculate expected profit for the classifier from the previous question on the validation data. Is the performance better or worse? Explain why.

Confusion matrix:

```
> classficationTable
       truth
predict   No  Yes
  FALSE 2131 1138
  TRUE    14   10

> sum(classficationTable * profits)
[1] -9
```

The expected profit is worse than that on the training data. That's because our model was trained by the training data by choosing the optimal threshold based on the training data. So there's some variance when testing on different data.

(e) If DogBark inc. had a perfect classifier, what would its profit be on the validation data? Is this higher or lower than the performance on the previous part? Explain why.

# dog owner	1148						
# not dog owner	2145						
Confusion matrix				Cost matrix			
		Actual dog owner				Actual dog owner	
		1	0			1	0
Predict dog owner	1	1148	0	Predict dog owner	1	0.5	-1
	0	0	2145		0	0	0
Expected profit	574						

Figure 3: Expected profit of a perfect classifier

If Dogbark had a perfect classifier, the net expect would be 574, which is definitely higher than the performance of our prediction on the last part, because our prediction has errors which will results in sending pamphlets to customers who don not own dogs (increase no revenue but cost) while a perfect classifier can make the company only sending pamphlets to dog owners which ends up with a higher net profit.

(f) Fit a logistic regression to the training data using all the covariates to predict dog ownership. Print the estimated coefficients and interpret them.

```
> lgfit = glm(dog ~., data = dog.train, family = binomial)
> summary(lgfit)

Call:
glm(formula = dog ~ ., family = binomial, data = dog.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3880  -0.9511  -0.7306   1.2266   2.2229

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.715e-01  2.144e-01   2.199   0.0279 *
pub_dist            -5.525e-04  7.422e-05  -7.444 9.76e-14 ***
supermaket_dist      1.801e-04  1.349e-04   1.335   0.1820
laundry_dist        -1.170e-04  9.811e-05  -1.192   0.2331
park_dist           -8.172e-04  9.179e-05  -8.903  < 2e-16 ***
neigh_density_score -9.564e-03  2.125e-02  -0.450   0.6527
tree_score           6.214e-04  1.046e-03   0.594   0.5526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3189.4  on 2468  degrees of freedom
Residual deviance: 3037.4  on 2462  degrees of freedom
AIC: 3051.4

Number of Fisher Scoring iterations: 4
```

Interpretation: Among all the features, `pub_dist` and `park_dist` have very significantly negative impact on probability of whether a customer has a dog, which indicates that if one lives very far to pubs and parks, then he/she is much less likely to have a dog. And distance to supermarkets, laundries and neighborhood density and number of trees nearby are not statistically important predictors on identifying dog owners

(g) Use the output of the logistic regression to create a classifier. What is the threshold that maximizes DogBark inc.'s profits, based on the training set? What is the confusion matrix on the validation set? What is the profit on the validation set?

```
> lgPrediction = predict(lgfit, newdata = dog.train, type = "response")
> range(lgPrediction)
[1] 0.08453272 0.62652852

> threshold = seq(0.09,0.62,0.01)
> lgprofit = numeric(54)
> for(i in 1:54){
+   lgDecision = ifelse(lgPrediction > threshold[i],1,0)
+   classficationTable = table(predict = lgDecision,
+                              truth = dog.train$dog)
+   lgprofit[i] = sum(classficationTable * profits)
+ }
```

```
> best_s = which.max(lgprofit)
> threshold[best_s]
[1] 0.56
```

The optimal threshold on the training data is 0.56, meaning when predicted probability of a customer owning a dog is larger than 0.56, we conclude him/her as a predicted dog owner. And here's the confusion matrix and profit on validation data:

```
> lgPrediction = predict(lgfit, newdata = dogdata, type = "response")
> lgDecision = ifelse(lgPrediction > threshold[best_s],1,0)
> classficationTable = table(predict = lgDecision,
+                            truth = dogdata$dog)
> classficationTable
       truth
predict   No  Yes
      0 2122 1037
      1   23  111

> sum(classficationTable * profits)
[1] 32.5
```

(h) Fit a decision tree to the training data to predict who is a dog owner. Use cross-validation to find the best tree and plot it. How would you use your tree to decide whom to send pamphlets to (based on the training data). What is the confusion matrix on the validation data? What is the profit on the validation set?

```
> tree.dog = tree(dog ~ ., data = dog.train)
> set.seed(123)
> cv.dog = cv.tree(tree.dog)
> best_size = cv.dog$size[which(cv.dog$dev == min(cv.dog$dev))][1]
> prune.dog = prune.misclass(tree.dog, best=best_size)
>
> par(mfrow = c(1, 1))
> plot(prune.dog)
> text(prune.dog,pretty=0)
```

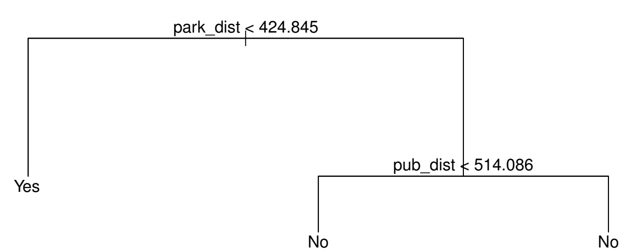

Figure 4: Best tree on training data

In this case, I will pick `park_dist` $= 424.845$ as a threshold. If it's larger than this threshold, I will consider the customer as a dog owner and send pamphlet to him/her. Here's the confusion matrix and profit on the validation data:

```
> attach(dogdata)
> lgDecision = ifelse(park_dist < 424.845,1,0)
> classficationTable = table(predict = lgDecision,
+                            truth = dogdata$dog)
> classficationTable
       truth
predict   No  Yes
      0 1983  869
      1  162  279

> sum(classficationTable * profits)
[1] -22.5
```

(i) Which method would you recommend using?

Based on comparing the profit on validation data from the three classifier in previous problems, I will recommend using logistic regression due to highest profit.

(j) Suppose that DogBark inc. got stuck with a large inventory of dog toys, and wishes to change the goal of the market campaign. Instead of maximizing profits, DogBark inc. wishes to use the marketing campaign to get 1,000 purchases by sending the minimal number of pamphlets. DogBark inc. has mailing information and the indexes included in the dataset about 1,000,000 potential customers. The data in dog.csv is a representative sample of the larger dataset. Using the classifier you selected in the previous question, how many pamphlets (on average) would DogBark inc. have to send in order to get 1,000 purchases?

```
> # 1j-number of pamphlets need to send in order to get 1000 purchases
> 1000/(0.05 * mean(lgPrediction))
[1] 57436.91
```

2 Question 2 (Quality of Classification)

(a) Given these ROC test results, what is your estimate of the total readmissions and CareTracker costs for AMI patients for the past three years if Tahoe had used the Xalta system? Explain your estimate.

```
> xaltra = read.csv("xaltra.csv")
> attach(xaltra)

> x = round(998 * True.Positive.Rate..Xaltra.)
> y = round(3384 * False.Positive.Rate)
> pred.readmission = x+y
> pred.readmission
  [1] 4382 4382 4372 4297 4157 3970 3776 3590 3392 3210 3040 2898 2739 2612
 [15] 2490 2356 2269 2172 2064 1984 1905 1822 1742 1670 1600 1545 1492 1433
 [29] 1388 1336 1300 1266 1233 1195 1159 1124 1092 1056 1024  991  944  902
 [43]  860  836  810  770  730  706  672  638  606  581  561  543  520  492
```

```
 [57]  468  443  429  400  387  367  331  311  300  276  258  236  226  206
 [71]  188  178  163  153  141  134  118  106  101   97   84   74   64   53
 [85]   42   36   34   24   20   16   10    7    6    6    3    0    0    0
 [99]    0    0

> xaltra.cost = numeric(100)
> cost.matrix = c(c(6000, 8000), c(1200, 0))
> for (i in 1:100){
+   confusion.table = c(c(x[i], 998-x[i]), c(y[i], 3384-y[i]))
+   xaltra.cost[i] = sum(confusion.table * cost.matrix)
+ }
> best= which(xaltra.cost == min(xaltra.cost))
> c(x[best],y[best])
[1] 639 352
> best.pred.adm = x[best]+y[best]
> best.pred.adm
[1] 991

> min(xaltra.cost)
[1] 7128400
```

Therefore, using Xaltra, the estimated total readmissions is 991, and Ct total cost is $7,128,400.

(b) What is the reduction in cost relative to Tahoe's current system? Do the savings justify the fees Xaltra is charging? Why or why not.

```
> reduct.cost = 7984000 - min(xaltra.cost)
> reduct.cost
[1] 855600
>
> xaltra.fee = 250000 +45000*3
> xaltra.fee
[1] 385000
>
> Xaltra.net.benefit = reduct.cost - xaltra.fee
> Xaltra.net.benefit
[1] 470600
```

As shown in this result, the reductionin cost relevant to Tahoe's current system is $855,600, and the Xaltra fee is $385,000. So the savings does justify the fees, resulting in a net benefit of involving Xaltra as $470,600.

3 Question 3 (Skill vs. Luck and DiD)

(a) Using 2011 as the "before" period and 2012 as the "after" period, perform a difference-in-difference analysis on the change in the average test scores of the SIS students. Based on your DiD estimate, what is the increase in test scores from SIS?

Students Group	Average 2011 ST	Average 2012 ST	Difference
SIS	8.88	12.88	4.00
no SIS	16.69	16.13	-0.56
		DiD	4.56

Figure 5: DiD estimated increase from SIS

(b) You suspect the results in part (a) may be overly optimistic because of the effects of regression to the mean. That is, because only the students who performed poorly on the 2011 exam were enrolled in SIS, some increase in their 2012 scores would be expected due simply to regression to the mean. To test this idea, consider the performance of the students between 2010 and 2011. Use the data from 2010 and the data from 2011 to determine whether there was regression to the mean. If so, what is the shrinkage coefficient?

```
> hs = read.csv("hillside_data.csv")
> attach(hs)

> reg = lm(X2011.ST ~ X2010.ST, data = hs)
> summary(reg)

Call:
lm(formula = X2011.ST ~ X2010.ST, data = hs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0721 -2.2934 -0.1828  2.1512  8.4852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.13356    1.17241   5.232 9.55e-07 ***
X2010.ST     0.61066    0.07374   8.281 6.34e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 3.342 on 98 degrees of freedom
Multiple R-squared:  0.4117,    Adjusted R-squared:  0.4057
F-statistic: 68.58 on 1 and 98 DF,  p-value: 6.34e-13

> shrink.coef = reg$coefficients[2]
> shrink.coef
 X2010.ST
0.6106589
```

As seen from the result above, there did exists regression to the mean due to the high significance of coefficient of 2010 score, and the shrinkage coefficient is 0.61.

(c) Using the shrinkage coefficient you obtained in part (b), construct a shrinkage estimate of 2012 scores based on the 2011 test results. What is the RMSE of your predictions?

```
> new.x <- data.frame(
+   x = hs$X2011.ST
```

```
+ )
> pred = predict(reg, newdata = new.x)
> rmse = sqrt(mean((pred - hs$X2012.ST)^2))
> rmse
[1] 4.218803
```

(d) Now, use the results from (b),(c) to correct the DiD analysis so it accounts for the shrinkage effect. To do that, compute the average of the estimated and actual 2012 scores for both the SIS students and non-SIS students. Considering the estimated 2012 scores as the "before" scores and the actual 2012 scores as the "after" scores, perform another DiD analysis of the SIS program. With this correction for shrinkage, what is your new estimate of the increase in test scores from SIS? Make sure you completely understand this technique before you apply it.

Students Group	Average pred.2012	Average 2011 ST	Difference
SIS	11.44	12.88	1.44
no SIS	16.20	16.13	-0.07
		New DiD	1.51

Figure 6: New DiD with correction for shrinkage

(e) Briefly comment on what was 'wrong' with the first method, and how the second method 'fixed' this problem.

The first method did not consider the shrinkage factor of the exam score, which can be interpreted as luck, those who got a score under 11 might only be unlucky and not able to show their true skill. This is a confounding variable that both cause them to be in the treatment group (SIS) as well as get a much higher score in the next exam. So what the second method did is to control this confounding variable and mute its effect on the DiD estimate, which did fix the problem.

(f) How might you use a regression discontinuity framework to estimate the effect of SIS on grades? Obtain such an estimate using the data provided.

Here's the DiD result using only students data with scores 10, 11, 12, 13 in year 2011, which are close to the threshold 11. Those part of treatment group and control group can be treated as satisfying the parallel trends assumption.

Row Labels	Average of Y2011	Average of Y2012	Difference	
0	12.375	13.0625	0.6875	
1	10.6	13.3	2.7	
Grand Total	11.69230769	13.15384615	2.0125	DiD

Figure 7: New DiD using regression discontinuity

4 Question 4 (Clustering and PCA)

(a) First, consider each country only in terms of red meat and white meat protein (ignore all other data). Cluster the countries into three clusters based on these data, and plot the resulting cluster memberships.

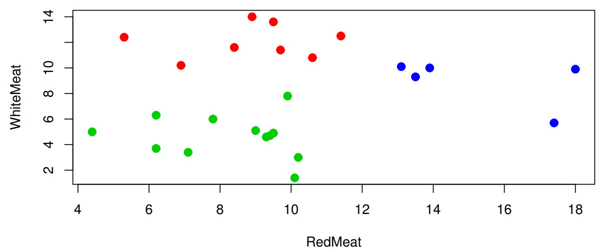

Figure 8: Clustering into 3 groups using kmeans

(b) Now, consider all the data. Pick an appropriate number of clusters to group these countries. Print out cluster memberships.

```
> wss = c()
> for (i in 1:20)
+ {
+   wss[i] = kmeans(pt, centers = i, nstart = 10)$tot.withinss
+ }
> plot(1:20, wss)
>
> # Five clusters seems fair
> km.fit = kmeans(pt, centers = 5, nstart = 20)
>
> # PRint the states in each cluster
> for (i in 1:nrow(km.fit$centers))
+ {
+   print(paste("Cluster", i))
+   print(names(km.fit$cluster)[km.fit$cluster == i] )
+ }
[1] "Cluster␣1"
[1] "Austria"     "Belgium"     "France"      "Ireland"     "Netherlands"
[6] "Switzerland" "UK"          "W␣Germany"
[1] "Cluster␣2"
[1] "Albania"        "Czechoslovakia" "Greece"         "Hungary"
[5] "Italy"          "Poland"         "USSR"
[1] "Cluster␣3"
[1] "Bulgaria"    "Romania"     "Yugoslavia"
[1] "Cluster␣4"
[1] "Denmark" "Finland" "Norway"  "Sweden"
[1] "Cluster␣5"
[1] "E␣Germany" "Portugal"  "Spain"
```

(c) Perform hierarchical clustering on these data, and plot the results. Are the results as you would expect?

```
pt = scale(pt)
hc.complete = hclust( dist(pt), method = "complete" )
plot(hc.complete)
```

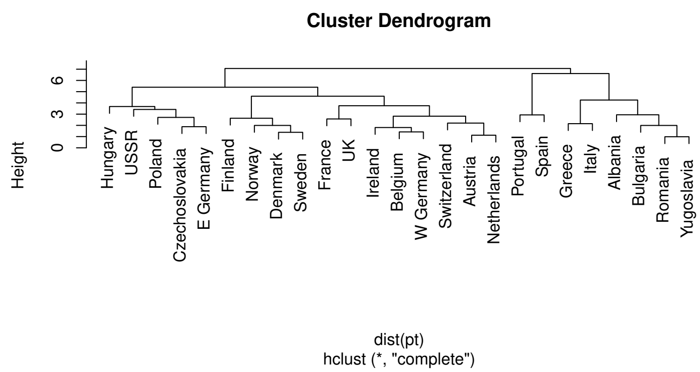

Figure 9: hierarchical clustering

(d) Perform a principal component analysis on these data, and plot the cities in terms of their first two principal components. How much of the variance is explained by these two principal components?

```
> biplot(pr.out)
> pr.out = prcomp(pt, center = TRUE, scale=TRUE)
> pr.var = pr.out$sdev^2
> pr.var = pr.var / sum(pr.var)
>
> biplot(pr.out)
> sum(pr.var[1:2])
[1] 0.6268263
```

Based on the result as above, the variance explained by the first two principal components are 62.7%, and the plot is shown by Figure 10 on the next page.

5 Question 5 (Simulation)

(a) Simulate 10,000 days of total demand and create a histogram of daily demand. What is the 10th and 90th percentile for the demand?

```
> N = 10000
> x = rnorm(N, mean = 50, sd = 10)
> y = numeric(N)
> u = runif(N)
> for (i in 1:N){
+   if (u[i] <= 0.4){
+     y[i] = runif(1) * 30 + 20
```

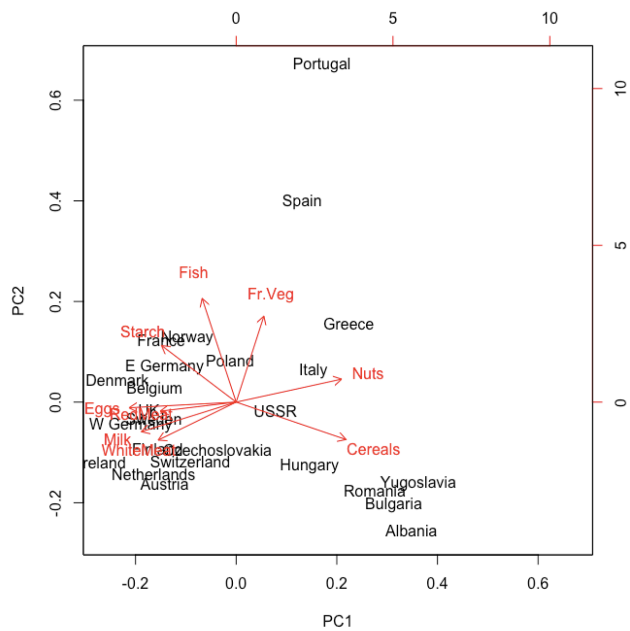

Figure 10: First 2 Principal Components Analysis

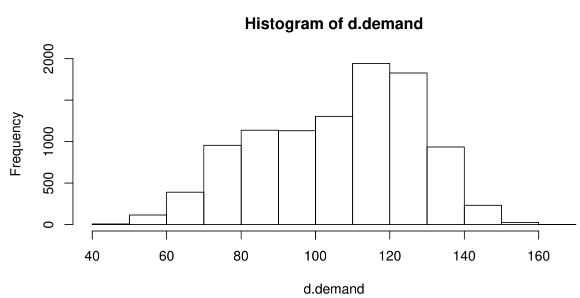

Figure 11: Histogram of daily demand

```
+   }
+   else{
+     y[i] = runif(1) * 20 + 60
+   }
+ }
> d.demand = x+y
> hist(d.demand)
> quantile(d.demand, probs = c(0.1, 0.9))
      10%       90%
 75.77404 131.39820
```

(b) Assume that each croissant costs \$1 to make, and sells for \$4. All croissants are made daily before the store opens. What is the expected profit if 120 croissants are made every day?

Denote morning demand as X, and afternoon demand as Y, and assume X and Y are independent random variables. Given the information we have, the expected daily demand is:

$$\begin{aligned} E(daily\ demand) &= E(X+Y) = E(X) + E(Y) \\ &= 50 + (0.4 \times 35 + 0.6 \times 70) \\ &= 106 \end{aligned}$$

So the expected daily profit is:

$$E(daily\ profit) = 4E(daily\ demand) - 120 = 304$$

(c) What is the optimal number of croissants to make every day? What is the corresponding optimal profit? Use simulation to find your answer.

```
# find optimal average daily profit
> for (k in 1:max(d.demand)){
+   for (i in 1:N){
+     if (k > d.demand[i]){
+       d.profit[i] = 4 * d.demand[i] - k
+     }else{
+       d.profit[i] = 3 * k
+     }
+   }
+   ave.profit[k] = mean(d.profit)
+ }
> best.k = which(ave.profit == max(ave.profit))
> best.k
[1] 123
> ave.profit[best.k]
[1] 293.0668
```

So based on our simulated data, the optimal number of croissants to make every day is 123, and the corresponding optimal daily profit is approximately \$293.

6 Question 6 (Optimization)

(a) Write down a mathematical formulation to optimize the total net profit. Is it linear, nonlinear, or discrete?

Denote L_s, L_c as the amount of production of standard and customized Laptops, and D_s, D_c as the amount of production of standard and customized desktops. Here's the formulation of the optimization problem:

$$\begin{aligned} Max\ \ & 100min\{L_s, 1200\} + 200min\{L_c, 1000\} + 150min\{D_s, 700\} + 400min\{D_c, 400\} \\ s.t.\ \ & L_s + L_c = 1500 \\ & D_s + D_c = 1000 \\ & L_c + D_c \leq 500 \end{aligned}$$

As we can see, it s a nonlinear program.

(b) Solve this problem and describe the optimal strategy and the optimal net profit.

Primal Problem:							
	Ls	Lc	Ds	Dc	Optimal Objective		
Production	1400	100	600	400	390000		
Demand	1200	1000	700	400			
Min{demand, production}	1200	100	600	400			
Maximize	100	200	150	400			
Contraints	1	1	0	0	1500	=	1500
	0	0	1	1	1000	=	1000
	0	1	0	1	500	<=	500

Figure 12: Primal optimization

(c) What is the benefit of being able to customize 200 more machines?

Customize 200 more machines:							
	Ls	Lc	Ds	Dc	Optimal Objective	Changes	
Production	1199	301	601	399	429850	39850	
Demand	1200	1000	700	400			
Min{demand, production}	1199	301	601	399			
Maximize	100	200	150	400			
Contraints	1	1	0	0	1500	=	1500
	0	0	1	1	1000	=	1000
	0	1	0	1	700	<=	700

Figure 13: Optimization with 200 more customization capacity

(d) What is the benefit from being able to sell 300 more desktops?

For selling 300 more desktops, I tends to increase the demand for desktop, but I don't know it's for standard or customized desktops. So to simplify the problem, I just assume we are able to sell 300 more customized desktops.

Sell 300 more desktops (assume for customized desktops):							
	Ls	Lc	Ds	Dc	Optimal Objective	Changes	
Production	1500	0	500	500	395000	5000	
Demand	1200	1000	700	700			
Min{demand, production}	1200	0	500	500			
Maximize	100	200	150	400			
Contraints	1	1	0	0	1500	=	1500
	0	0	1	1	1000	=	1000
	0	1	0	1	500	<=	500

Figure 14: Optimization with 300 more sold customized desktops

(e) What happens if we manufacture 100 fewer laptops?

Manufacture 100 fewer laptop:							
	Ls	Lc	Ds	Dc	**Optimal Objective**	Changes	
Production	1300	100	601	399	389750	-250	
Demand	1200	1000	700	400			
Min{demand, production}	1200	100	601	399			
Maximize	100	200	150	400			
Contraints	1	1	0	0	1400	=	1400
	0	0	1	1	1000	=	1000
	0	1	0	1	499	<=	500

Figure 15: Optimization with 100 fewer laptop manufacturing