MSiA 421, Data Mining
Assignment 3: CLV and Churn Modeling
Due: Thursday, February 6 by 1pm

1. (Based on Gupta and Lehmann (2003)) In Aug 1994, Jason and Matthew Olim launched CDNow in the basement of their parents' house in Ambler, Pennsylvania. Within a year, revenues reached $2 million. Like most Web-based startup companies, CDNow focused heavily on acquiring new customers. By 1998, after acquiring a rival company, CDNow had a customer base of more than 3 million customers. The company highlighted the number of new customers in its reports to financial analysts. Was their emphasis on acquisition misplaced? The answer to this question depends on whether CLV exceeds acquisition costs.

   From 1998–2000 the average customer acquisition cost ranged from $30–$55, according to company reports. During the same time, the annual gross margin per customer ranged from $10–$20 and averaged about $15. The retention rate ranged from 51–68%. Assume an annual discount rate of 12%.

   (a) Estimate CLV (not counting acquisition costs) for the following scenarios assuming payments at the end of the year (ordinary):
      i. Pessimistic: $m = \$10$, $r = .51$ *Answer:* $10 \times 0.51/(1.12 - 0.51) = 8.360656$
      ii. Average: $m = \$15$, $r = .595$ *Answer:* $15 \times 0.595/(1.12 - 0.595) = 17$
      iii. Optimistic: $m = \$20$, $r = .68$ *Answer:* $20 \times 0.68/(1.12 - 0.68) = 30.90909$

   (b) Compare your answers with the reported acquisition costs ($30–$50). Was the acquisition strategy profitable? Continue to assume that payments come at the end of the period. *Answer: Only at the lowest cost ($30) and most optimistic assumptions ($m = 20, r = 0.68$) would it be profitable.*

2. At the end of 1999 CDNow reported a loss of over $100 million and in Mar 2000 it publicly announced that it had only enough cash to sustain operations for 6 months. Soon after, the media giant Bertelsmann entered negotiations to acquire CDNow. Bertelsmann bought CDNow for $117 million in Jul 2000. Did Bertelsmann overpay? This is usually a complex question depending on many different factors. CDNow, however, had few physical assets and its major asset was its customers. Assuming the total number of customers (in Jun 2000) was 3.29 million customers, an average annual margin of $15 and an annual discount rate of 12%, find customer equity (assuming payments at the end of the year) under the following scenarios:

   (a) Average: $r = .595$ *Answer:* $3.29 \times 15 \times 0.595/(1.12 - 0.595) = 55.93$

   (b) Optimistic: $r = .68$ *Answer:* $3.29 \times 15 \times 0.68/(1.12 - 0.68) = 76.26818$

   (c) Really optimistic: $r = .8$ (due to Bertelsmann having better management, market position and more cash) *Answer:* $3.29 \times 15 \times 0.8/(1.12 - 0.8) = 123.375$. *This is the only case where it would be profitable.*

3. A company that offers cable television service acquires customers who pay a monthly fee of \$60 until they cancel the service. Customers may cancel at any time. You may assume the event that a customer cancels during any specific time period is statistically independent of the event that the customer cancels during any other time period. Assume an immediate annuity, where the first payment occurs at time 0, i.e., the customer pays for the service at the beginning of the month and if the customer cancels before the end of the month the customer does not receive any refund. Suppose that the company retains 82% of its customers each month and that this retention rate is constant across all months and customers.

   (a) What is the expected and median time of attrition in months? That is, if T is a discrete random variable that represents the time in years that a customer cancels, what is the mean and median of T? *Answer:* $E(X) = 1/(1 - .82) = 5.56$.

   (b) What is the expected life-time revenue of a customer assuming a monthy discount rate of 1%? *Answer:* $CLV = 60 \times 1.01/(1.01 - .82) = 318.9474$.

   (c) *Customer equity* is the sum of CLV across customers. Suppose that the company recently acquired 1000 customers as described above. What is the customer equity of these 1000 customers? (total CLV across 1000 customers) *Answer:* $1000 \times 318.9474 = \$318,947.4$

   (d) The remaining parts discuss the effects of *unobserved heterogeneity* — .

   Suppose that the retention rate is not constant across customers, i.e., the retention rate is heterogeneous. Assume that customers are from two segments: half the customers acquired during a particular month have a monthly retention rate of 92% (the "loyalists") while the other half have a monthly retention rate of 72% (non-loyalists). Note that the aggregate retention rate during the first month is 82% (averaged across all customers), as in the original problem. Compute CLV for a single loyalist and separately for a single non-loyalist. *Answer: For loyalists,* $CLV = 60 \times 1.01/(1.01 - .92) = 673.33$. *For non-loyalists,* $CLV = 60 \times 1.01/(1.01 - .72) = 208.9655$.

   (e) Suppose the company acquires 500 loyalists and 500 non-loyalists during a given month. Find the customer equity of the 1000 customers. Compare this with the customer equity obtained in part (c). *Answer:* $500 \times 673.33 + 500 \times 208.966 = 441,149.4$. *By not accounting for unobserved heterogeneity you underestimate CLV and CE.*

4. The Comcast "Digital Starter" plan costs \$29.99 for the first six months and \$60.48 thereafter. Payments come at the beginning of the month, i.e., the first payment is made when the technician connects the household. There are no cancelation fees. Suppose that the retention rate is 95% during the four months ($r_1 = \cdots = r_4 = .95$), $r_5 = 70\%$ during the month prior to the first higher payment, $r_6 = 60\%$ during the next month, and $r_t = 90\%$ thereafter ($t \geq 7$). Let random variable T be the month of

cancelation, i.e., customers are retained for $T-1$ months and cancel during month T. For example, if $T=1$ then the customer makes one payment at the time of installation and cancels before the second payment, receiving no refund for the unused portion of the first month. Assume Comcast does not spend any money on marketing to retain customers.

(a) Find the probability that a customer defaults during month 8. *Answer:* $P(T=8)=.0308$;

(b) Find the probability that a customer will cancel after the first year (i.e., has made 12 or more payments)? *Answer:* $S(12)=.2020$;

(c) Find the probability that a customer is retained for the first two years. (i.e., makes at least 24 payments) *Answer:* $S(24)=.0571$;

(d) Find the probability that a customer cancels before the end of the first year (i.e., $T<12$). *Answer:* $1-S(12)=1-.2020$;

(e) Find the expected month of attrition (i.e., the mean of T). *Answer:* $\sum_t S(t)=8.52$;

(f) Graph the hazard, PDF and Survival functions. *Answer: See Excel;*

(g) Find expected CLV, assuming an annual discount rate of 12%. *Answer: $339.39.*

5. (39 points) The `cell.csv` data set has a sample of customers from a leading US cell phone provider.

(a) (2 points) The variable `billmonth` is a string variable ($6.) that tells the "current" month. It ranges from Jun 2007 through May 2008. The `churn` variable indicates whether a customer churned during the "current" month (1 = churn, 0 = not churn). Run a crosstab of these two variables. Describe the apparent sampling plan. *Answer: From this you should deduce that the company drew an approximate balanced sample from each month (e.g., ≈ 50% churners and 50% non-churners). You should also conclude that there are about 12 "current months" with approximately equal sample sizes.*

```
> table(cell$churn, cell$billmonth)
   Apr   Aug   Dec   FEB   JAN   Jul   Jun   Mar   May   Nov   Oct   Sep
    08    07    07    08    08    07    07    08    08    07    07    07
0  209   207   246   244   191   205   207   203   207   206   190   209
1  205   213   224   191   229   194   217   205   183   203   189   223
```

(b) (4 points) How many people canceled? How many opportunities were there to cancel? What is the retention rate? Use the `t2` variable for this part, which gives the difference in months between `billmonth` and `contractdt`. The `contractdt`

variable gives the start date of the current contract, so `t2` will tell us the time until churn or censoring.

Answer: 2414 churns, 95624 opportunities to churn, $\hat{r} = .975$ and $E(T) = 1/(1-.975) = 39.61$ months. There are missing values and some duplicated cucustomers in the file. Give full credit if they are following the right steps and their answer is close.

```
> ok = !is.na(cell$t2) & !is.na(cell$churn)
> sum(cell$t2[ok])
[1] 95624
> sum(cell$churn[ok])
[1] 2414
> rhat = 1-sum(cell$churn[ok])/sum(cell$t2[ok])
> rhat  # estiamted retention rate
[1] 0.9747553
> 1/(1-rhat) # E(T)
[1] 39.61226
```

(c) (2 points) Briefly discuss the effect of the sampling plan on your estimates of the retention rate? *Answer: Since the sample was pulled to have half churners and we do not have sampling weights, estimates of retention rates will be biased in an unknown direction. This was done intentionally by the company to disguise the data.*

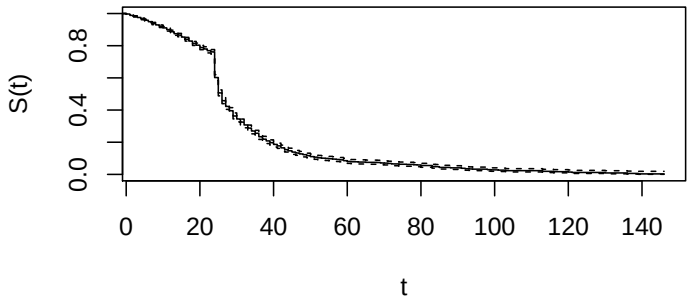
(d) (4 points) Compute average monthly revenue using the `TOTAL_REV_AMT` variables. Assume payments come at the beginning of the month. Compute expected lifetime revenue using your retention rate estimate, assuming a monthly discount rate of 1%. *Answer: $m = \$55.27$ and $E(CLV) = \$1591.12$.*

```
> m = mean(as.matrix((cell[,7:21]), ncol=1), na.rm=T)
> m    # average payment/month
[1] 55.52319
> m*1.01/(1.01-rhat)
[1] 1591.116
```

(e) (3 points) Run a life table analysis and generate hazard and survival function plots for the `T2` variable. Submit only the hazard or survival plot and briefly describe what it tells you about churning.

Answer: The survivor curve decreases slowly (indicating that they are slowly losing customers) until month 24, when it decreases sharply (indicating that many customers start to cancel).

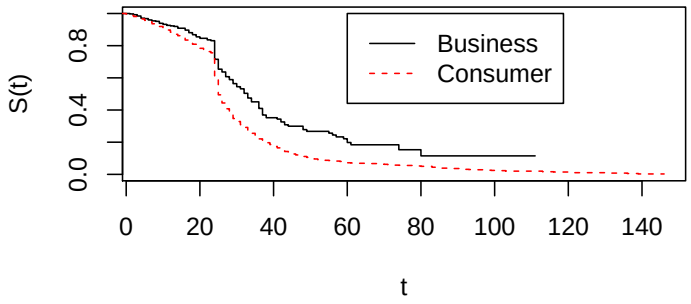
```
library(survival)
fit = survfit(Surv(t2, churn) ~ 1, cell)
plot(fit)
```



(f) (3 points) Generate life tables stratifying on accounts of type "B" (business) and "I" (individual). Use `account_type` variable. Submit only a survival or hazard plot and indicate which group is more likely to churn?

Answer: I find it easier to interpret the survival plot. The pattern is similar for the two, but business customers survive longer. I'm not sure how to interpret line=0, but the effect size goes in the right direction.

```
fit = survfit(Surv(t2, churn) ~ ACCOUNT_TYPE, cell)
plot(fit, col=1:2, lty=1:2)
legend(60,1,c("Business", "Consumer"), col=1:2, lty=1:2)
```

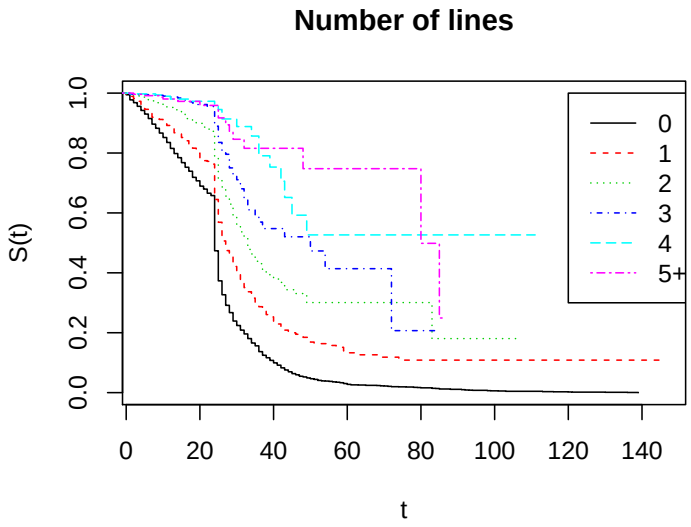


(g) (3 points) Generate separate life tables (or KM) and survival curves for different numbers of lines (line_count), after capping line_count at 5 (set all values greater than 5 equal to 5. What is the story? How does the number of lines affect retention rates? Submit only the survival or hazard plot.

Answer: The patterns are similar, but the more lines, the lower the hazard of churning.

```
fit = survfit(Surv(t2, churn) ~ ACCOUNT_TYPE, cell)
```

```
plot(fit, col=1:2, lty=1:2)
legend(60,1,c("Business", "Consumer"), col=1:2, lty=1:2)
```



(h) Which of the following factors affect the likelihood of churn: account type, line count, time out of contract? You will build a discrete-time survival model using logistic regression to answer this question. First, we will study the effect of time out of contract after controlling for some other variables.

i. (5 points) Create a variable `timeout` giving the number of months that someone is out of his/her contract, e.g., -1= 1 month until the end, 0 = contract just ended, $+1$ = one month out of contract, etc. Make a new variable that equals timeout, and then cap it at 7 and -7 (any values greater than 7 should be set equal to 7 and values less than -7 should be set to -7). Using logistic regression, predict the variable "event" (=1 if customer cancels during the current month, 0 otherwise) on your capped timeout variable, treating it as categorical, `account_type` and `line_count` (capped at 5).
Answer: Full credit if they decide to drop line=0.

ii. (2 points) Interpret the effects for `account_type` and `line_count`. Do they tell you the same story as the life tables?
Answer: After controlling for line count and contract status, there is no significant difference ($P = .3940$) between businesses and individuals. Line count is still a strong predictor, with more lines implying lower risk of churn.

iii. (4 points) Describe how the probability of canceling depends on `timeout`. Create a graph showing how the log-odds of canceling depends on `timeout`.

Answer: One way to do this is to copy the values into R and use plot. The effect is small for negative values of timeout (in contract). There is a large spike at 0, which decreases thereafter.

iv. (2 points) Which variables are most important in explaining cancelations?

Answer: Contract status and line count. One could quantify importance by looking at LRT test statistics.

(i) (5 points) Make a new copy of `timeout` and cap it at 4 and -1. Create a variable that indicates whether or not there was an overage charge during the previous month (`lagover`). Use logistic regression to predict `event` from the new capped version of `timeout`, `account_type`, `line_count`, and lagged overage. Submit the estimated coefficients. Which variables seem to be good predictors of churning?
Answer: Same as above. Overage is not significant.

```
/* part j */
proc logistic data=cell descending;
  class account_type line_count timeout / param=ref;
  model event = account_type line_count timeout lagover;
  format line_count linecnt. timeout timeoutj.;
run;
```

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.5056	0.2695	417.3613	<.0001
ACCOUNT_TYPE	B	1	0.1956	0.0980	3.9856	0.0459
LINE_COUNT	0	1	2.7021	0.2684	101.3593	<.0001
LINE_COUNT	1	1	1.9965	0.2728	53.5459	<.0001
LINE_COUNT	2	1	1.2894	0.2788	21.3906	<.0001
LINE_COUNT	3	1	0.5512	0.3008	3.3584	0.0669
LINE_COUNT	4	1	0.00165	0.3511	0.0000	0.9962
timeout	0	1	2.3809	0.0717	1101.2159	<.0001
timeout	1	1	2.4002	0.0846	804.8199	<.0001
timeout	2	1	1.8055	0.1114	262.5699	<.0001
timeout	3	1	1.5527	0.1317	138.8942	<.0001
timeout	4+	1	1.1944	0.0549	473.8506	<.0001
lagover		1	0.1372	0.0482	8.1112	0.0044