Homework 3

Andey Nunes, Mengyu Li, Jordan Hilton, Peter Boss

2/2/2019

Document and Exercise Set Up

```
Claims_and_Days <- read_csv("Claims_and_Days.csv",
col_types = cols(CharlsonIndex = col_character()))

names(Claims_and_Days)
```

```
##  [1] "ID"                    "MemberID"
##  [3] "ProviderID"            "vendor"
##  [5] "pcp"                   "Year"
##  [7] "specialty"             "placesvc"
##  [9] "paydelay"              "LengthOfStay"
## [11] "dsfs"                  "PrimaryConditionGroup"
## [13] "CharlsonIndex"         "sex"
## [15] "AgeAtFirstClaim"       "DaysInHospital_Y2"
```

```
Claims_and_Days <- Claims_and_Days %>%
  mutate(CharlsonIndex = case_when(
    CharlsonIndex == "0" ~ "0" ,
    CharlsonIndex == "2-Jan" ~ "1-2",
    CharlsonIndex == "4-Mar" ~ "3-4",
    CharlsonIndex == "5+" ~ "5+"),
    AgeAtFirstClaim = case_when(
      AgeAtFirstClaim == "19-Oct" ~ "10-19",
      AgeAtFirstClaim != "19-Oct" ~ AgeAtFirstClaim),
    LengthOfStay = if_else(is.na(LengthOfStay), "0 or unknown", LengthOfStay)) %>%
  mutate(CharlsonIndex = ordered(CharlsonIndex, levels =
                                   c("0","1-2", "3-4", "5+")),
         AgeAtFirstClaim = ordered(AgeAtFirstClaim, levels = c(
           "0-9", "10-19", "20-29", "30-39", "40-49", "50-59", "60-69", "70-79",  "80+")),
         LengthOfStay = ordered(LengthOfStay, levels = c(
           "0 or unknown", "1 day", "2 days", "3 days","4 days", "5 days", "6 days", "1- 2 weeks", "2- 4
)
```

Tidy 1 R Algorithm: Homework 3

Quantize the answer field: `Risk_Level`

```
VlookupSim <- Claims_and_Days %>%
  mutate(Risk_Level = case_when(
    DaysInHospital_Y2 == 0 ~ "A",
    DaysInHospital_Y2 == 1 ~ "B",
    DaysInHospital_Y2 == 2 | DaysInHospital_Y2 == 3 ~ "C",
    DaysInHospital_Y2 == 4 | DaysInHospital_Y2 == 5 ~ "D",
    DaysInHospital_Y2 >= 6 ~ "E"),
      Risk_Level_label = case_when(
```

```
        Risk_Level == "A" ~ "no risk",
        Risk_Level == "B" ~ "very low risk",
        Risk_Level == "C" ~ "low risk",
        Risk_Level == "D" ~ "medium risk",
        Risk_Level == "E" ~ "high risk")) %>%
  mutate(Risk_Level_label = ordered(Risk_Level_label, levels = c(
    "no risk", "very low risk", "low risk", "medium risk", "high risk")))

str(VlookupSim)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    644706 obs. of  18 variables:
##  $ ID                    : num  2 3 4 5 6 7 1 8 9 10 ...
##  $ MemberID              : num  25872 25872 25872 25872 25872 ...
##  $ ProviderID            : num  376108719 171278567 171278567 171278567 171278567 ...
##  $ vendor                : num  5024957 7891165 7891165 7891165 7891165 ...
##  $ pcp                   : num  294037 294037 294037 294037 294037 ...
##  $ Year                  : chr  "Y1" "Y1" "Y1" "Y1" ...
##  $ specialty             : chr  "Laboratory" "Internal" "Internal" "Internal" ...
##  $ placesvc              : chr  "Independent Lab" "Office" "Office" "Office" ...
##  $ paydelay              : num  23 16 19 21 21 11 22 37 23 30 ...
##  $ LengthOfStay          : Ord.factor w/ 13 levels "0 or unknown"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ dsfs                  : chr  "0- 1 month" "1- 2 months" "2- 3 months" "3- 4 months" ...
##  $ PrimaryConditionGroup: chr  "MSC2a3" "RESPR4" "RESPR4" "RESPR4" ...
##  $ CharlsonIndex         : Ord.factor w/ 4 levels "0"<"1-2"<"3-4"<..: 1 2 2 2 2 2 2 1 1 1 ...
##  $ sex                   : chr  "F" "F" "F" "F" ...
##  $ AgeAtFirstClaim       : Ord.factor w/ 9 levels "0-9"<"10-19"<..: 2 2 2 2 2 2 2 9 9 9 ...
##  $ DaysInHospital_Y2     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Risk_Level            : chr  "A" "A" "A" "A" ...
##  $ Risk_Level_label      : Ord.factor w/ 5 levels "no risk"<"very low risk"<..: 1 1 1 1 1 1 1 1 1 1 .
```

Create pivot tables

proportion of observations for risk groups

```
prop_obs <- function(df = VlookupSim, x,
                     y = VlookupSim$Risk_Level_label) {
   prop <- with(df, table(x, y))
   prop <- cbind(prop, total = rowSums(prop))
   high_prop <- prop[,5]/prop[,6]*100
   prop <- cbind(prop, high_risk_proportion = high_prop)
   #proptibble <- tibble(prop) %>%
    #  arrange(desc(high_risk_proportion))

   return(prop)
}
```

Primary Condition Group Counts and Proportion of High Risk

```
pcg <- prop_obs(x = VlookupSim$PrimaryConditionGroup)
kable(pcg)
```

	no risk	very low risk	low risk	medium risk	high risk	total	high_risk_proportion
AMI	5919	713	483	588	1188	8891	13.4

	no risk	very low risk	low risk	medium risk	high risk	total	high_risk_proportion
APPCHOL	3805	461	207	213	355	5041	7.0
ARTHSPIN	54575	6096	3087	3080	4873	71711	6.8
CANCRA	866	80	19	35	101	1101	9.2
CANCRB	7343	685	346	421	944	9739	9.7
CANCRM	170	21	11	7	22	231	9.5
CATAST	309	33	16	19	68	445	15.3
CHF	2004	322	130	153	563	3172	17.8
COPD	8516	1007	409	575	1199	11706	10.2
FLaELEC	899	104	45	55	144	1247	11.6
FXDISLC	6951	710	289	262	470	8682	5.4
GIBLEED	21002	2755	1343	1222	2524	28846	8.8
GIOBSENT	2186	230	132	94	189	2831	6.7
GYNEC1	8793	775	734	553	379	11234	3.4
GYNECA	1847	162	163	91	106	2369	4.5
HEART2	8446	1124	542	716	1591	12419	12.8
HEART4	5142	612	327	270	726	7077	10.3
HEMTOL	4513	461	177	286	757	6194	12.2
HIPFX	655	106	82	50	130	1023	12.7
INFEC4	17064	1864	783	691	1139	21541	5.3
LIVERDZ	595	57	20	30	48	750	6.4
METAB1	752	95	35	56	86	1024	8.4
METAB3	56338	5377	2383	2662	5040	71800	7.0
MISCHRT	24141	2585	994	1264	2437	31421	7.8
MISCL1	968	125	64	46	98	1301	7.5
MISCL5	9330	1098	496	524	867	12315	7.0
MSC2a3	89147	8331	3899	3769	6353	111499	5.7
NEUMENT	35789	3936	1524	1713	3343	46305	7.2
ODaBNCA	10050	844	368	397	657	12316	5.3
PERINTL	128	9	8	6	10	161	6.2
PERVALV	618	61	32	54	85	850	10.0
PNCRDZ	179	40	0	20	20	259	7.7
PNEUM	1905	238	81	131	226	2581	8.8
PRGNCY	5336	411	1048	609	429	7833	5.5
RENAL1	58	35	9	4	18	124	14.5
RENAL2	1323	178	76	97	429	2103	20.4
RENAL3	9657	1070	496	554	972	12749	7.6
RESPR4	29576	3274	1355	1266	2062	37533	5.5
ROAMI	8884	1200	622	676	1453	12835	11.3
SEIZURE	3530	625	271	278	458	5162	8.9
SEPSIS	76	12	5	10	17	120	14.2
SKNAUT	21217	2150	847	903	1784	26901	6.6
STROKE	1474	207	122	110	236	2149	11.0
TRAUMA	14868	1752	710	607	1023	18960	5.4
UTI	7817	822	403	433	680	10155	6.7

Charlson Index Group Counts and Proportion of High Risk

```
ci <- prop_obs(x = VlookupSim$CharlsonIndex)
kable(ci)
```

	no risk	very low risk	low risk	medium risk	high risk	total	high_risk_proportion

	no risk	very low risk	low risk	medium risk	high risk	total	high_risk_proportion
0	299915	27731	13408	12086	16051	369191	4.3
1-2	186875	24094	11320	12828	28407	263524	10.8
3-4	7136	902	414	627	1701	10780	15.8
5+	835	126	51	59	140	1211	11.6

Length of Stay Counts and Proportion of High Risk

```
los <- prop_obs(x = VlookupSim$LengthOfStay)
kable(los)
```

Simulate the 1R Algorithm

```
data1R <- filter(VlookupSim, Risk_Level != "A")
```

Ignoring Risk Level A build the 1R Rule for each of

- "PrimaryConditionGroup"
- "CharlsonIndex"
- "LengthOfStay"

  Select the Best Rule and Paste in here

Questions:

Q4.1

Why does our selected rule work better (reference error rate)

Q4.2

For the 1R Rule is it better to have a lower or higher cardinality?

Q4.3

We are ignoring risk level A because it is a subset of the data that is independent of the group we are trying to predict and it consists of nearly 77% of the original observations. By working with a smaller subset that contains only the observations with any risk level other than "no risk", we have a smaller search space for our algorithm.

Extra Credit

Ignoring Risk Level A rebin the risk level buckets and rebuild the 1R Rule and see if there is a difference.

Appendix

Risk Group Counts

graphical inspection function

```
risk_group_counts <- ggplot(VlookupSim, aes(x = Risk_Level_label)) +
   geom_text(aes(label = ..count..), stat = "count", vjust = -0.25) +
   geom_bar()
```

```
risk_group_counts +
  facet_wrap(~PrimaryConditionGroup)
```

```
risk_group_counts +
  facet_wrap(~CharlsonIndex)
```

```
risk_group_counts +
  facet_wrap(~LengthOfStay)
```