An Excellent Classifier

Madeline Gillman

Read in the data

```
# Get sheet names corresponding to study
sheets <- readxl::excel_sheets("data/nature24473_MOESM5_survival.xlsx")

# Read in and merge data, add column for study
outcomes <- bind_rows(lapply(
  1:length(sheets),
  function(i) {
    readxl::read_excel(
      "data/nature24473_MOESM5_survival.xlsx",
      sheet = sheets[i],
      col_types = c("guess", "guess", "numeric")
    ) %>%
      mutate(study = str_replace(sheets[i], "Survival ", ""))
  }
))

# Now do the same for neoantigens data
sheets <- readxl::excel_sheets("data/nature24473_MOESM4_neoantigens.xlsx")

neoantigens <- bind_rows(lapply(
  1:length(sheets),
  function(i) {
    readxl::read_excel(
      "data/nature24473_MOESM4_neoantigens.xlsx",
      sheet = sheets[i]
    ) %>%
      mutate(study = sheets[i])
  }
))
```

```r
head(outcomes)
```

```
# A tibble: 6 x 4
  Sample Months Status study
  <chr>   <dbl>  <dbl> <chr>
1 Pat02   53.7        0 VanAllen et al.
2 Pat03    3.29       1 VanAllen et al.
3 Pat04   32.4        0 VanAllen et al.
4 Pat06    5.29       1 VanAllen et al.
5 Pat08    4.60       1 VanAllen et al.
6 Pat100  11.8        1 VanAllen et al.
```

```r
head(neoantigens)
```

```
# A tibble: 6 x 10
     ID MUTATION_ID     Sample WT.Peptide MT.Peptide MT.Allele WT.Score MT.Score
  <dbl> <chr>           <chr>  <chr>      <chr>      <chr>        <dbl>    <dbl>
1     1 1_1563747_C_T_~ Pat02  NHREVAQIL  NYREVAQIL  C0702          701       70
2     2 1_17087582_G_A~ Pat02  SPSNDFQVL  SPLNDFQVL  B0801          739      202
3     3 1_17087582_G_A~ Pat02  SPSNDFQVL  SPLNDFQVL  B0702           32       37
4     4 1_21806573_A_G~ Pat02  LLDEKEPEV  LLDEKGPEV  A0201            8        9
5     5 1_46073697_C_T~ Pat02  KPGQEAPVL  KPGQEASVL  B0702           90       31
6     6 1_75693438_C_T~ Pat02  YFELQQTWF  YFKLQQTWF  C0702          311      213
# i 2 more variables: HLA <chr>, study <chr>
```

Engineer features

First, a bit more data tidying. Let's split up some columns and make indicator columns to make generating features easier.

```r
dat <- neoantigens %>%
  separate_wider_delim(HLA, delim = ",", names = c(paste0("HLA_", 1:6)),
                       too_few = "align_end", too_many = "drop") %>%
  mutate(tmp = strsplit(as.character(WT.Peptide), "")) %>%
  unnest() %>%
  group_by(WT.Peptide, ID) %>%
  mutate(n = 1:n()) %>%
  spread(n, tmp) %>%
  ungroup() %>%
  rename_with(~ paste0("WT_peptide_pos_", .x, recycle0 = TRUE),
              .cols = c(16:24)) %>%
  mutate(tmp = strsplit(as.character(MT.Peptide), "")) %>%
  unnest() %>%
  group_by(MT.Peptide, ID) %>%
  mutate(n = 1:n()) %>%
  spread(n, tmp) %>%
  ungroup() %>%
  rename_with(~ paste0("MT_peptide_pos_", .x, recycle0 = TRUE),
              .cols = c(25:33)) %>%
  mutate(
    mutation_chr = str_split_i(MUTATION_ID, "_", 1),
    mutation_pos = str_split_i(MUTATION_ID, "_", 2),
    mutation_a1 = str_split_i(MUTATION_ID, "_", 3),
    mutation_a2 = str_split_i(MUTATION_ID, "_", 4),
    a1_C = case_when(
      mutation_a1 == "C" ~ 1,
      TRUE ~ 0
    ),
    a1_G = case_when(
      mutation_a1 == "G" ~ 1,
      TRUE ~ 0
    ),
    a1_A = case_when(
      mutation_a1 == "A" ~ 1,
      TRUE ~ 0
    ),
    a1_T = case_when(
      mutation_a1 == "T" ~ 1,
      TRUE ~ 0
    ),
    a2_C = case_when(
      mutation_a2 == "C" ~ 1,
      TRUE ~ 0
    ),
    a2_G = case_when(
      mutation_a2 == "G" ~ 1,
      TRUE ~ 0
    ),
    a2_A = case_when(
      mutation_a2 == "A" ~ 1,
      TRUE ~ 0
    ),
    a2_T = case_when(
      mutation_a2 == "T" ~ 1,
      TRUE ~ 0
    )
  )
```

Making our first set of features: summing up the number of amino acids overall and in each peptide position for each patient.

```
aa_list <- c("L", "V", "F", "T", "A", "I", "M", "C", "G", "Y", "H", "K",
             "S", "E", "R", "W", "P", "Q", "D", "N")

for (aa in aa_list) {
  dat[paste0("count_WT_", aa)] <- sapply(strsplit(dat$WT.Peptide, ""),
                                         function(x) sum(x == aa))
  dat[paste0("count_MT_", aa)] <- sapply(strsplit(dat$MT.Peptide, ""),
                                         function(x) sum(x == aa))

  dat[paste0("count_MT_pos1_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_1, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos2_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_2, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos3_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_3, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos4_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_4, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos5_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_5, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos6_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_6, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos7_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_7, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos8_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_8, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_MT_pos9_", aa)] <- sapply(strsplit(dat$MT_peptide_pos_9, ""),
                                              function(x) sum(x == aa))

  dat[paste0("count_WT_pos1_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_1, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos2_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_2, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos3_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_3, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos4_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_4, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos5_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_5, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos6_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_6, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos7_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_7, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos8_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_8, ""),
                                              function(x) sum(x == aa))
  dat[paste0("count_WT_pos9_", aa)] <- sapply(strsplit(dat$WT_peptide_pos_9, ""),
                                              function(x) sum(x == aa))
}
```

Our second set of features: summing up how many mutations on each chromosome for each patient.

```
chr_list <- c(1:22, "X", "Y")
for (chr in chr_list) {
  dat[paste0("count_chr", chr)] <- sapply(strsplit(dat$mutation_chr, ""),
                                          function(x) sum(x == chr))
}
```

Now let's collapse by subject and join with the outcomes data for analysis:

```
dat2 <- dat %>%
  mutate(across(c(mutation_pos), as.numeric)) %>%
  group_by(Sample) %>%
  summarise(across(c(6, 7, 34, 37:468), sum, .names = "sum_{.col}")) %>%
  left_join(outcomes)
```

```
Joining with `by = join_by(Sample)`
```

Finding "useful" features

And by useful, I mean they generate a significant split in the data. I'm going to be looking for these using the entire dataset.

```r
# create a dataframe to retain features
df_with_good_features <- data.frame(Sample = dat2$Sample)
# Loop through each column and create a new binary feature
# based on mean or median
for (i in c(2:436)) {
  # print(i)
  feature <- colnames(dat2)[i]
  mean_i <- mean(dat2[[i]])
  if (mean_i == 0) { # If there's no variance in the feature, move on
    next
  }
  median_i <- median(dat2[[i]])
  temp <- dat2 %>%
    select(Sample, all_of(i), Months, Status) %>%
    mutate(
      above_mean = case_when(.data[[colnames(dat2)[i]]] > mean_i ~ 1, TRUE ~ 0),
      above_median = case_when(.data[[colnames(dat2)[i]]] > median_i ~ 1, TRUE ~ 0)
    )
  # Calculate log rank p value
  km_diff <- survdiff(Surv(Months, Status) ~ above_mean, data = temp)
  if (km_diff$pvalue < 0.05) {
    print(paste0("column number ", i, " has a p-value of ",
                 km_diff$pvalue, " (mean)"))

    # Add feature to dataframe to use for model
    df_with_good_features <- temp %>%
      select(1, 5) %>%
      rename_with(~ paste0(feature, "_above_mean"),
                  matches("above_mean")) %>%
      left_join(df_with_good_features)
  }

  km_diff <- survdiff(Surv(Months, Status) ~ above_median,
                      data = temp)

  if (km_diff$pvalue < 0.05) {
    print(paste0("column number ", i, " has a p-value of ",
                 km_diff$pvalue, " (median)"))

    # Add feature to dataframe to use for model
    df_with_good_features <- temp %>%
      select(1, 6) %>%
      rename_with(~ paste0(feature, "_above_median"),
                  matches("above_median")) %>%
      left_join(df_with_good_features)
  }
}
```

```
[1] "column number 2 has a p-value of 0.0447453383518615 (mean)"
[1] "column number 4 has a p-value of 0.0230009910888821 (mean)"
[1] "column number 5 has a p-value of 0.0140795919784464 (median)"
[1] "column number 6 has a p-value of 0.0378925812415947 (median)"
[1] "column number 8 has a p-value of 0.0169561822903697 (mean)"
[1] "column number 10 has a p-value of 0.00603375042635501 (mean)"
[1] "column number 10 has a p-value of 0.0455276963520781 (median)"
[1] "column number 12 has a p-value of 0.0366373424892781 (mean)"
[1] "column number 12 has a p-value of 0.027189989336928 (median)"
[1] "column number 13 has a p-value of 0.0263147277311545 (median)"
[1] "column number 15 has a p-value of 0.0486894319758719 (median)"
[1] "column number 17 has a p-value of 0.0403729129119503 (median)"
[1] "column number 19 has a p-value of 0.0480989651626312 (median)"
[1] "column number 20 has a p-value of 0.0259842727873005 (median)"
[1] "column number 22 has a p-value of 0.0322336083467359 (median)"
[1] "column number 27 has a p-value of 0.0415357098983233 (median)"
[1] "column number 29 has a p-value of 0.0224751276259443 (median)"
[1] "column number 31 has a p-value of 0.0265862034849499 (median)"
[1] "column number 33 has a p-value of 0.00642078797761971 (median)"
[1] "column number 34 has a p-value of 0.00488742772665154 (median)"
[1] "column number 35 has a p-value of 0.014188990624516 (median)"
[1] "column number 37 has a p-value of 0.0143536444687929 (median)"
[1] "column number 40 has a p-value of 0.0261754018779639 (median)"
[1] "column number 41 has a p-value of 0.0288970518466814 (median)"
[1] "column number 43 has a p-value of 0.0210473880997189 (median)"
[1] "column number 45 has a p-value of 0.0193909743572345 (median)"
[1] "column number 46 has a p-value of 0.0156935269402163 (median)"
[1] "column number 48 has a p-value of 0.0466905669788288 (mean)"
[1] "column number 48 has a p-value of 0.00373259520326279 (median)"
[1] "column number 49 has a p-value of 0.0285561412732493 (median)"
```

```
[1] "column number 50 has a p-value of 0.0498894872498977 (median)"
[1] "column number 52 has a p-value of 0.0188397917249204 (median)"
[1] "column number 55 has a p-value of 0.0276758602893867 (median)"
[1] "column number 56 has a p-value of 0.0333765781274191 (median)"
[1] "column number 58 has a p-value of 0.0434213032624256 (median)"
[1] "column number 59 has a p-value of 0.0217648502470677 (median)"
[1] "column number 60 has a p-value of 0.0404863796868527 (mean)"
[1] "column number 64 has a p-value of 0.0438155079361048 (median)"
[1] "column number 65 has a p-value of 0.0151077376085789 (median)"
[1] "column number 68 has a p-value of 0.0150413856280221 (median)"
[1] "column number 71 has a p-value of 0.0499153747552886 (mean)"
[1] "column number 71 has a p-value of 0.0356207782879299 (median)"
[1] "column number 72 has a p-value of 0.0314561932942931 (mean)"
[1] "column number 74 has a p-value of 0.0188602230864867 (median)"
[1] "column number 75 has a p-value of 0.0352708564291483 (mean)"
[1] "column number 76 has a p-value of 0.00246767350611709 (median)"
[1] "column number 77 has a p-value of 0.00908365532251165 (median)"
[1] "column number 78 has a p-value of 0.0327750675891761 (median)"
[1] "column number 79 has a p-value of 0.0373389374819942 (median)"
[1] "column number 80 has a p-value of 0.036121843737979 (median)"
[1] "column number 81 has a p-value of 0.0310585100680771 (median)"
[1] "column number 82 has a p-value of 0.0128286910124298 (median)"
[1] "column number 85 has a p-value of 0.00108275716724668 (median)"
[1] "column number 86 has a p-value of 0.0324965167217493 (median)"
[1] "column number 87 has a p-value of 0.0253629076757833 (median)"
[1] "column number 88 has a p-value of 0.0246701853229943 (median)"
[1] "column number 91 has a p-value of 0.00318144424402079 (median)"
[1] "column number 96 has a p-value of 0.0230598388328513 (median)"
[1] "column number 98 has a p-value of 0.0432183401474321 (median)"
[1] "column number 100 has a p-value of 0.0432157221671721 (mean)"
[1] "column number 100 has a p-value of 0.0360240159713785 (median)"
[1] "column number 101 has a p-value of 0.0366728498780792 (mean)"
[1] "column number 103 has a p-value of 0.0256659786761097 (mean)"
[1] "column number 104 has a p-value of 0.0168155275146806 (median)"
[1] "column number 105 has a p-value of 0.0275416581569317 (median)"
[1] "column number 107 has a p-value of 0.0370176512224733 (median)"
[1] "column number 109 has a p-value of 0.0404501105406317 (mean)"
[1] "column number 114 has a p-value of 0.0377123134823165 (mean)"
[1] "column number 117 has a p-value of 0.0246516267557402 (median)"
[1] "column number 119 has a p-value of 0.0164818869797081 (median)"
[1] "column number 121 has a p-value of 0.047570064580219 (median)"
[1] "column number 122 has a p-value of 0.0162260481811393 (median)"
[1] "column number 126 has a p-value of 0.0325568582512993 (median)"
```

```
[1] "column number 127 has a p-value of 0.00638881001468882 (median)"
[1] "column number 128 has a p-value of 0.0153911452685828 (median)"
[1] "column number 130 has a p-value of 0.0171450779152572 (median)"
[1] "column number 131 has a p-value of 0.028293060448571 (median)"
[1] "column number 132 has a p-value of 0.0371312083144478 (median)"
[1] "column number 133 has a p-value of 0.032582888034464 (median)"
[1] "column number 134 has a p-value of 0.0334227223607583 (median)"
[1] "column number 135 has a p-value of 0.0435705687867458 (mean)"
[1] "column number 135 has a p-value of 0.0158273284983683 (median)"
[1] "column number 136 has a p-value of 0.0209826853488029 (median)"
[1] "column number 137 has a p-value of 0.0478757488827675 (mean)"
[1] "column number 137 has a p-value of 0.0343904967586552 (median)"
[1] "column number 139 has a p-value of 0.0492675531065005 (mean)"
[1] "column number 142 has a p-value of 0.0039789619070661 (median)"
[1] "column number 144 has a p-value of 0.0412729680802678 (mean)"
[1] "column number 144 has a p-value of 0.0077378944867684 (median)"
[1] "column number 145 has a p-value of 0.0104969767482946 (mean)"
[1] "column number 145 has a p-value of 0.018901948188801 (median)"
[1] "column number 146 has a p-value of 0.0310121044499547 (median)"
[1] "column number 150 has a p-value of 0.0361676027135812 (median)"
[1] "column number 151 has a p-value of 0.00160002255211911 (median)"
[1] "column number 152 has a p-value of 0.0456455900741041 (median)"
[1] "column number 154 has a p-value of 0.0491833908712004 (median)"
[1] "column number 155 has a p-value of 0.0314356408148394 (mean)"
[1] "column number 155 has a p-value of 0.0235670208309632 (median)"
[1] "column number 158 has a p-value of 0.014078839779343 (mean)"
[1] "column number 158 has a p-value of 0.00960833276949841 (median)"
[1] "column number 162 has a p-value of 0.0291385435336191 (median)"
[1] "column number 163 has a p-value of 0.020416651301711 (median)"
[1] "column number 164 has a p-value of 0.0176337680209662 (median)"
[1] "column number 166 has a p-value of 0.010749789531042 (median)"
[1] "column number 167 has a p-value of 0.0207307516205597 (mean)"
[1] "column number 167 has a p-value of 0.000751973828563669 (median)"
[1] "column number 170 has a p-value of 0.0374602867194796 (median)"
[1] "column number 171 has a p-value of 0.021160339810806 (median)"
[1] "column number 177 has a p-value of 0.00795012191742886 (median)"
[1] "column number 178 has a p-value of 0.0208604780269284 (median)"
[1] "column number 182 has a p-value of 0.00555291728826353 (median)"
[1] "column number 185 has a p-value of 0.00301521776373538 (median)"
[1] "column number 188 has a p-value of 0.0165431879045406 (median)"
[1] "column number 191 has a p-value of 0.0103820906137256 (median)"
[1] "column number 193 has a p-value of 0.0306145352508821 (median)"
[1] "column number 194 has a p-value of 0.0273661536161626 (mean)"
```

```
[1] "column number 194 has a p-value of 0.0193528417397252 (median)"
[1] "column number 195 has a p-value of 0.0363470850039664 (median)"
[1] "column number 196 has a p-value of 0.00925217711083624 (median)"
[1] "column number 198 has a p-value of 0.0365404779977031 (median)"
[1] "column number 199 has a p-value of 0.0174869289811364 (median)"
[1] "column number 200 has a p-value of 0.0120140254491872 (median)"
[1] "column number 201 has a p-value of 0.027361407624224 (median)"
[1] "column number 202 has a p-value of 0.0230186967209104 (median)"
[1] "column number 203 has a p-value of 0.023130415064836 (median)"
[1] "column number 204 has a p-value of 0.0136788217071076 (median)"
[1] "column number 207 has a p-value of 0.0377310706775507 (mean)"
[1] "column number 207 has a p-value of 0.0132927274931815 (median)"
[1] "column number 208 has a p-value of 0.00719634436577168 (median)"
[1] "column number 209 has a p-value of 0.047506559465955 (median)"
[1] "column number 210 has a p-value of 0.0489432206083825 (median)"
[1] "column number 212 has a p-value of 0.0377043653359417 (median)"
[1] "column number 213 has a p-value of 0.0264715511585915 (median)"
[1] "column number 215 has a p-value of 0.00435361093197296 (median)"
[1] "column number 216 has a p-value of 0.0491221932230368 (median)"
[1] "column number 221 has a p-value of 0.00943289548356734 (median)"
[1] "column number 224 has a p-value of 0.0486299602978995 (median)"
[1] "column number 225 has a p-value of 0.0308311817217452 (median)"
[1] "column number 226 has a p-value of 0.0193337965263306 (median)"
[1] "column number 228 has a p-value of 0.0470900221006419 (median)"
[1] "column number 229 has a p-value of 0.0348064570118452 (median)"
[1] "column number 230 has a p-value of 0.0152104393081591 (median)"
[1] "column number 232 has a p-value of 0.00922826537977141 (median)"
[1] "column number 236 has a p-value of 0.0313490953658195 (median)"
[1] "column number 237 has a p-value of 0.0250873830242445 (median)"
[1] "column number 239 has a p-value of 0.0420444807757053 (median)"
[1] "column number 240 has a p-value of 0.00340611148040854 (median)"
[1] "column number 245 has a p-value of 0.00597863174069088 (median)"
[1] "column number 249 has a p-value of 0.0359265842357015 (mean)"
[1] "column number 249 has a p-value of 0.00557311769378889 (median)"
[1] "column number 251 has a p-value of 0.0138470257006989 (median)"
[1] "column number 255 has a p-value of 0.02276034947249 (median)"
[1] "column number 256 has a p-value of 0.00823506427908482 (median)"
[1] "column number 262 has a p-value of 0.0453388083313867 (median)"
[1] "column number 263 has a p-value of 0.0467731463582647 (mean)"
[1] "column number 263 has a p-value of 0.00211519987404041 (median)"
[1] "column number 264 has a p-value of 0.0133385925690425 (median)"
[1] "column number 265 has a p-value of 0.0259580390416383 (median)"
[1] "column number 271 has a p-value of 0.0347132571024541 (mean)"
```

```
[1] "column number 271 has a p-value of 0.0379984273776235 (median)"
[1] "column number 273 has a p-value of 0.016783876274487 (median)"
[1] "column number 274 has a p-value of 0.00491029158074473 (median)"
[1] "column number 275 has a p-value of 0.0309977660173118 (mean)"
[1] "column number 275 has a p-value of 0.0151423961980858 (median)"
[1] "column number 276 has a p-value of 0.0084850094624533 (mean)"
[1] "column number 276 has a p-value of 0.0368100510671225 (median)"
[1] "column number 278 has a p-value of 0.00581598177324026 (median)"
[1] "column number 279 has a p-value of 0.00504919806591094 (median)"
[1] "column number 280 has a p-value of 0.0192942949919402 (median)"
[1] "column number 281 has a p-value of 0.00651651007103843 (median)"
[1] "column number 282 has a p-value of 0.00640169824054608 (median)"
[1] "column number 283 has a p-value of 0.00813178392700535 (mean)"
[1] "column number 283 has a p-value of 0.0193512417872866 (median)"
[1] "column number 284 has a p-value of 0.0123166393014003 (mean)"
[1] "column number 284 has a p-value of 0.0365551393042241 (median)"
[1] "column number 285 has a p-value of 0.00216836147389301 (mean)"
[1] "column number 285 has a p-value of 0.0167833386191503 (median)"
[1] "column number 287 has a p-value of 0.0120285562164111 (median)"
[1] "column number 288 has a p-value of 0.00509004187761208 (median)"
[1] "column number 289 has a p-value of 0.0288584724060528 (median)"
[1] "column number 290 has a p-value of 0.0455431653206433 (median)"
[1] "column number 291 has a p-value of 0.00375178102570676 (median)"
[1] "column number 293 has a p-value of 0.0347400987266384 (median)"
[1] "column number 296 has a p-value of 0.025675565691281 (median)"
[1] "column number 298 has a p-value of 0.0165255396122532 (mean)"
[1] "column number 298 has a p-value of 0.00210246293550769 (median)"
[1] "column number 300 has a p-value of 0.0186761356036222 (median)"
[1] "column number 304 has a p-value of 0.0189037370990262 (median)"
[1] "column number 305 has a p-value of 0.0340350483160456 (median)"
[1] "column number 306 has a p-value of 0.0442207814869785 (median)"
[1] "column number 307 has a p-value of 0.0312046277246967 (mean)"
[1] "column number 307 has a p-value of 0.00165846875622055 (median)"
[1] "column number 313 has a p-value of 0.0244706871063301 (mean)"
[1] "column number 313 has a p-value of 0.034793036611283 (median)"
[1] "column number 314 has a p-value of 0.0277570714623071 (mean)"
[1] "column number 314 has a p-value of 0.0332005147083119 (median)"
[1] "column number 318 has a p-value of 0.030199180948194 (median)"
[1] "column number 320 has a p-value of 0.00711162922001093 (mean)"
[1] "column number 320 has a p-value of 0.00270522102489838 (median)"
[1] "column number 323 has a p-value of 0.00383989838207014 (mean)"
[1] "column number 323 has a p-value of 0.017522050172816 (median)"
[1] "column number 329 has a p-value of 0.00351490570687437 (mean)"
```

```
[1] "column number 329 has a p-value of 0.00663668838476085 (median)"
[1] "column number 330 has a p-value of 0.0392012125235197 (median)"
[1] "column number 332 has a p-value of 0.000932873511024279 (mean)"
[1] "column number 332 has a p-value of 0.0117725509621966 (median)"
[1] "column number 334 has a p-value of 0.0460241308180831 (median)"
[1] "column number 335 has a p-value of 0.0480930467444569 (median)"
[1] "column number 337 has a p-value of 0.0174677983338767 (median)"
[1] "column number 339 has a p-value of 0.0271101078403613 (median)"
[1] "column number 340 has a p-value of 0.0466017354220169 (mean)"
[1] "column number 341 has a p-value of 0.00676373949985338 (median)"
[1] "column number 342 has a p-value of 0.0273953206087948 (median)"
[1] "column number 346 has a p-value of 0.0362232945848884 (median)"
[1] "column number 350 has a p-value of 0.00846765339596218 (median)"
[1] "column number 352 has a p-value of 0.018965540354664 (median)"
[1] "column number 355 has a p-value of 0.00943126939205987 (median)"
[1] "column number 356 has a p-value of 0.0422995178719737 (median)"
[1] "column number 357 has a p-value of 0.0126168102849029 (median)"
[1] "column number 358 has a p-value of 0.0204050563332224 (median)"
[1] "column number 359 has a p-value of 0.0163062103716259 (median)"
[1] "column number 362 has a p-value of 0.0044516892436591 (median)"
[1] "column number 363 has a p-value of 0.00978932339242685 (mean)"
[1] "column number 363 has a p-value of 0.00144380064816933 (median)"
[1] "column number 364 has a p-value of 0.0309148121511298 (median)"
[1] "column number 365 has a p-value of 0.0374422019752844 (median)"
[1] "column number 366 has a p-value of 0.037206762211237 (median)"
[1] "column number 367 has a p-value of 0.025249159999801 (median)"
[1] "column number 368 has a p-value of 0.00209913492096078 (median)"
[1] "column number 369 has a p-value of 0.0229355596948034 (median)"
[1] "column number 370 has a p-value of 0.0402789842962606 (median)"
[1] "column number 371 has a p-value of 0.0300562087708294 (median)"
[1] "column number 372 has a p-value of 0.00453275875957716 (mean)"
[1] "column number 372 has a p-value of 0.00453275875957716 (median)"
[1] "column number 373 has a p-value of 0.00409473993581522 (median)"
[1] "column number 374 has a p-value of 0.00656391251592189 (median)"
[1] "column number 375 has a p-value of 0.00607150952221326 (median)"
[1] "column number 377 has a p-value of 0.00345283158480548 (mean)"
[1] "column number 377 has a p-value of 0.00917196668171005 (median)"
[1] "column number 379 has a p-value of 0.0119334628315369 (median)"
[1] "column number 380 has a p-value of 0.00308829951699931 (median)"
[1] "column number 383 has a p-value of 0.0460071827678907 (mean)"
[1] "column number 383 has a p-value of 0.0460071827678907 (median)"
[1] "column number 384 has a p-value of 0.00295374326090106 (median)"
[1] "column number 385 has a p-value of 0.0312728944220104 (mean)"
```

```
[1] "column number 385 has a p-value of 0.00929239091739398 (median)"
[1] "column number 386 has a p-value of 0.0116786319944399 (mean)"
[1] "column number 386 has a p-value of 0.00729974674645371 (median)"
[1] "column number 387 has a p-value of 0.0422168260164299 (mean)"
[1] "column number 387 has a p-value of 0.0122178126943353 (median)"
[1] "column number 388 has a p-value of 0.00950277882463817 (median)"
[1] "column number 389 has a p-value of 0.0259996276133903 (median)"
[1] "column number 390 has a p-value of 0.0324139494577777 (median)"
[1] "column number 391 has a p-value of 0.0142760577821536 (median)"
[1] "column number 392 has a p-value of 0.0063750879035417 (mean)"
[1] "column number 392 has a p-value of 0.0063750879035417 (median)"
[1] "column number 393 has a p-value of 0.00795345656696159 (mean)"
[1] "column number 393 has a p-value of 0.00612616295172423 (median)"
[1] "column number 394 has a p-value of 0.0278078045135304 (mean)"
[1] "column number 394 has a p-value of 0.014292575661505 (median)"
[1] "column number 396 has a p-value of 0.00454645788240679 (median)"
[1] "column number 397 has a p-value of 0.0202187034166393 (median)"
[1] "column number 398 has a p-value of 0.0213364996652339 (mean)"
[1] "column number 398 has a p-value of 0.0309700990359792 (median)"
[1] "column number 399 has a p-value of 0.00301644238089071 (mean)"
[1] "column number 399 has a p-value of 0.00908444554763544 (median)"
[1] "column number 400 has a p-value of 0.0107851822155993 (median)"
[1] "column number 401 has a p-value of 0.0374798938289689 (mean)"
[1] "column number 401 has a p-value of 0.0167228045006762 (median)"
[1] "column number 402 has a p-value of 0.00781335097802567 (median)"
[1] "column number 405 has a p-value of 0.00122597143280465 (median)"
[1] "column number 406 has a p-value of 0.0141573863800106 (median)"
[1] "column number 407 has a p-value of 0.0232086048965759 (mean)"
[1] "column number 408 has a p-value of 0.0111305874434686 (mean)"
[1] "column number 408 has a p-value of 0.00819345065144906 (median)"
[1] "column number 409 has a p-value of 0.04346623032276 (mean)"
[1] "column number 409 has a p-value of 0.00694898649143727 (median)"
[1] "column number 410 has a p-value of 0.00201981745847408 (median)"
[1] "column number 411 has a p-value of 0.00864107195981286 (median)"
[1] "column number 414 has a p-value of 0.0207198395793902 (median)"
[1] "column number 415 has a p-value of 0.0476998219532624 (median)"
[1] "column number 417 has a p-value of 0.0345065644526089 (median)"
[1] "column number 420 has a p-value of 0.00025338596131657 (median)"
[1] "column number 421 has a p-value of 0.0029532495378425 (median)"
[1] "column number 435 has a p-value of 0.0265899310775519 (mean)"
```

Select features

Now that we have found > 200 potentially significant features, let’s use feature selection to keep the meaningful ones.

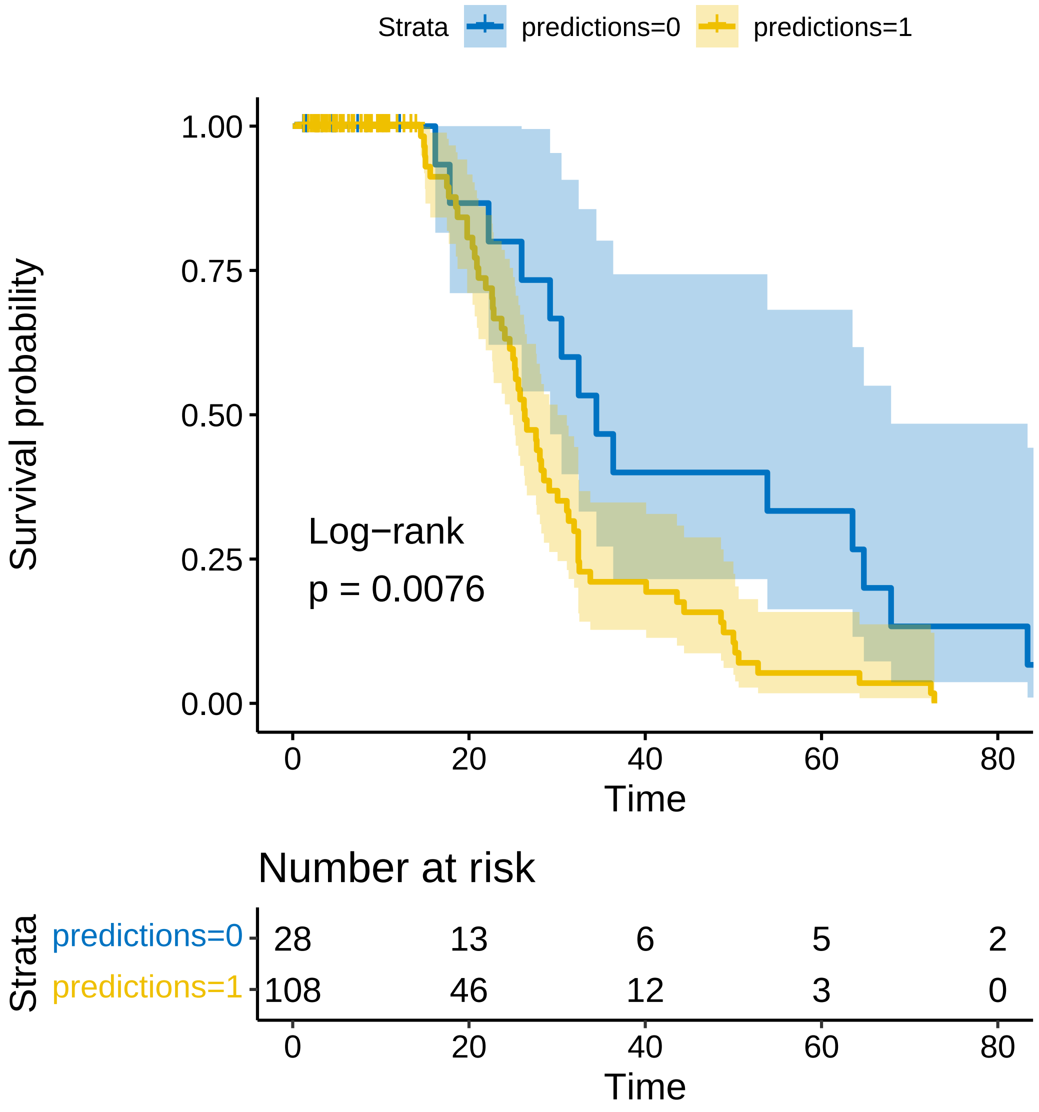
```
df_with_good_features <- df_with_good_features %>%
  left_join(outcomes) %>%
  mutate(Status2 = factor(case_when(
    Months >= 14 ~ 1,
    TRUE ~ 0
  ))) %>%
  na.omit()

x <- as.matrix(df_with_good_features[2:286])
colnames(x) <- colnames(df_with_good_features)[2:286]
y <- as.numeric(as.character(df_with_good_features$Status))
lasso_model <- cv.glmnet(x = x, y = y)

coef_matrix <- coef(lasso_model, s = lasso_model$lambda.min)
selected_features <- colnames(x)[which(coef_matrix != 0)]
print(selected_features)
```

```
[1] "sum_count_chrX_above_mean"       "sum_count_MT_pos8_N_above_median"
[3] "sum_count_MT_pos9_Q_above_mean"  "sum_count_WT_pos7_W_above_median"
[5] "sum_count_WT_pos1_E_above_median"
```

Add the selected features to our model and test it with a genuine train/test split. Note we are using a different outcome for this because why not □

```
set.seed(35)
# Define training/test datasets
sample <- sample(c(TRUE, FALSE),
                 nrow(df_with_good_features),
                 replace = TRUE,
                 prob = c(0.3, 0.7))
dat_train <- df_with_good_features[sample, ]
dat_test <- df_with_good_features[!sample, ]

# Generate classification model
# With 14 month outcome
default_glm_mod <- caret::train(
```

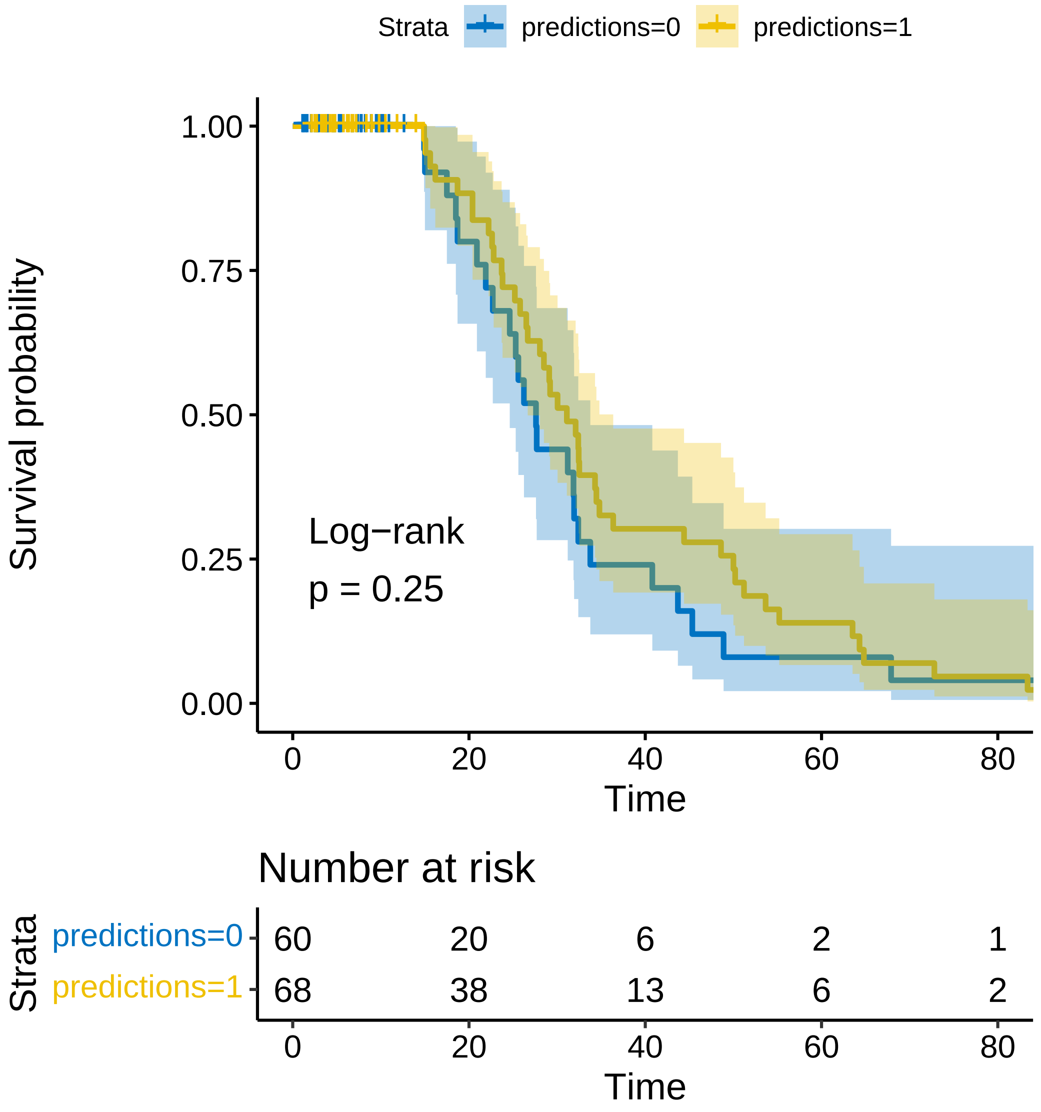
```r
  form = Status2 ~ sum_count_chrX_above_mean +
    sum_count_chr5_above_median +
    sum_count_MT_pos8_N_above_median +
    sum_count_MT_pos9_Q_above_mean +
    sum_count_WT_pos7_W_above_median +
    sum_count_WT_pos1_E_above_median,
  data = dat_train %>% select(all_of(c(2:286, 290))),
  method = "glm",
  family = "binomial"
)

predictions <- predict(default_glm_mod, newdata = dat_test)

dat2 <- cbind(dat_test, predictions)

km_fit <- survfit(Surv(Months, as.numeric(as.character(Status2))) ~ predictions,
                  data = dat2)

ggsurvplot(km_fit,
  data = dat2,
  risk.table = TRUE,
  pval = TRUE, conf.int = TRUE, palette = "jco", pval.method = TRUE
)
```



This is also quite sensitive to which individuals end up in the test/train groups. If we update `set.seed()`, the resulting p-value is not longer significant:

```
set.seed(34)

# Define training/test datasets
sample <- sample(c(TRUE, FALSE),
                 nrow(df_with_good_features),
                 replace = TRUE,
                 prob = c(0.3, 0.7))
```

```
dat_train <- df_with_good_features[sample, ]
dat_test <- df_with_good_features[!sample, ]

# Generate classification model
# With 14 month outcome
default_glm_mod <- caret::train(
  form = Status2 ~ sum_count_chrX_above_mean +
    sum_count_chr5_above_median +
    sum_count_MT_pos8_N_above_median +
    sum_count_MT_pos9_Q_above_mean +
    sum_count_WT_pos7_W_above_median +
    sum_count_WT_pos1_E_above_median,
  data = dat_train %>% select(all_of(c(2:286, 290))),
  method = "glm",
  family = "binomial"
)

predictions <- predict(default_glm_mod, newdata = dat_test)

dat2 <- cbind(dat_test, predictions)

km_fit <- survfit(Surv(Months, as.numeric(as.character(Status2))) ~ predictions,
                  data = dat2)

ggsurvplot(km_fit,
           data = dat2,
           risk.table = TRUE,
           pval = TRUE, conf.int = TRUE, palette = "jco", pval.method = TRUE)
```


Strata
predictions=0
predictions=1
Survival probability
1.00
0.75
0.50
0.25
0.00
Log−rank
p = 0.25
0
20
40
60
80
Time
Number at risk
Strata
predictions=0
predictions=1
60
20
6
2
1
68
38
13
6
2
0
20
40
60
80
Time