Introduction to feature engineering in R

FEATURE ENGINEERING IN R

Jose Hernandez

Data Scientist, University of Washington

Feature engineering in this course

Representing raw predictors by:

- Adjusting raw features
- Combining raw features
- Decomposing raw features into meaningful subsets

Outcome or target variable:

```
adult_incomes %>%
   select(income) %>%
   table()
```

```
 <=50K   >50K
24720   7841
```

Existing features:

```
adult_incomes %>%
   select(-income) %>%
   glimpse()
```

```
Observations: 32,561
Variables: 14
$ age             <int> 39, 50,...
$ workclass       <chr> "State-gov",...
$ fnlwgt          <int> 77516, 83311,...
$ education       <fct>  Bachelors,  Bachelors,...
$ educational_num <int> 13, 13,...
$ marital_status  <fct>  Never-married,...
$ occupation      <fct>  Adm-clerical,...
$ relationship    <fct>  Not-in-family,  Husband,...
$ race            <fct>  White,  White,...
$ gender          <fct>  Male,  Male,...
$ capital_gain    <int> 2174, 0, 0, 0,...
$ capital_loss    <int> 0, 0, 0,...
$ hours_per_week  <int> 40, 13, 40,...
$ native_country  <fct>  United-States,  United-State...
```

Finding meaning from raw data

```
adult_incomes %>%
  select(income, workclass) %>%
  head()
```

```
  income workclass
1  <=50K State-gov
2  <=50K     other
3  <=50K   Private
4  <=50K   Private
5  <=50K   Private
6  <=50K   Private
```

Finding meaning from raw data

```
adult_incomes %>%
    group_by(workclass) %>%
    summarise(totals = n())
```

```
# A tibble: 8 x 2
  workclass     totals
  <chr>          <int>
1 Federal-gov      960
2 Local-gov       2093
3 Never-worked       7
4 Private        22696
5 Self-emp-inc    1116
6 Self-emp-not-inc   2541
7 State-gov       1298
8 Without-pay       14
```

Useful functions

```
adult_incomes %>%
 mutate(new_workclass = ifelse(workclass == "Federal-gov", 1, 0))
```

```
library(caret)

new_data <- dummyVars("~ gender", data = adult_incomes)
```

```r
# One-hot encoding
 adult_incomes %>%
    mutate(federal_gov = ifelse(workclass == "Federal-gov", 1, 0),
       local_gov = ifelse(workclass == "Local-gov", 1, 0),
       state_gov = ifelse(workclass == "State-gov", 1, 0),
       private = ifelse(workclass == "Private", 1, 0),
       self_employed_inc = ifelse(workclass == "Self-emp-inc", 1, 0),
       self_employed_not_inc = ifelse(workclass == "Self-emp-not-inc", 1, 0),
       without_pay = ifelse(workclass == "Without-pay", 1, 0),
       never_worked = ifelse(workclass == "Never-worked", 1, 0)) %>%
    select(federal_gov:never_worked) %>%
    glimpse()
```

```
Observations: 32,561
Variables: 8
$ federal_gov           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ local_gov             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ state_gov             <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0...
$ private               <dbl> 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1...
$ self_employed_inc     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ self_employed_not_inc <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ without_pay           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ never_worked          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

Let's practice!

FEATURE ENGINEERING IN R

Binning encoding: content driven

FEATURE ENGINEERING IN R

Jose Hernandez

Data Scientist, University of Washington

A closer look at the categories

```
Observations: 32,561
Variables: 10
$ income                <fct>  <=50K,  <=50K,  <=50K,  <=50K,  <=50K,  <=50K,  <=50K,  >50K,  >50K,  >50K,  >50K,  >50K,  ...
$ workclass             <chr> "State-gov", "Self-emp-not-inc", "Private", "Private", "Private", "Private", "Private", "Sel...
$ federal_gov           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, NA, 0, 0, 0...
$ local_gov             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, NA, 0, 0, 1...
$ state_gov             <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0...
$ private               <dbl> 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, NA, 1, 1, 0...
$ self_employed_inc     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0...
$ self_employed_not_inc <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0...
$ without_pay           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0...
$ never_worked          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0...
```

Looking for similar categories

```
adult_incomes %>%
    select(workclass) %>%
    table()
```

```
     Federal-gov        Local-gov     Never-worked          Private
             960             2093                7            22696
    Self-emp-inc Self-emp-not-inc        State-gov      Without-pay
            1116             2541             1298               14
```

Public	**Private**	**Self-employed**	**Unemployed**

```
adult_incomes %>%
  mutate(new_workclass =
           case_when(workclass == "Federal-gov" ~ "public",
                     workclass == "Local-gov" ~ "public",
                     workclass == "State-gov" ~ "public",
                     workclass == "Self-emp-inc" ~ "self_empl",
                     workclass == "Self-emp-not-inc" ~ "self_empl",
                     workclass == "Without-pay" ~ "unemployed",
                     workclass == "Never-worked" ~ "unemployed",
                     TRUE ~ as.character(workclass)))
```

```
adult_income %>%
    select(new_workclass) %>%
    table()
```

```
   Private     public  self_empl unemployed
     22696       4351       3657         21
```

```
adult_incomes %>%
   mutate(public = ifelse(new_workclass == "public", 1, 0),
          private = ifelse(new_workclass == "Private", 1, 0),
          self_empl = ifelse(new_workclass == "self_empl", 1, 0),
          unemployed = ifelse(new_workclass == "unemployed", 1, 0))
```

```
# A tibble: 32,561 x 7
   income   workclass     new_workclass public private self_empl unemployed
   <fct>    <chr>         <chr>          <dbl>   <dbl>     <dbl>      <dbl>
 1 " <=50… State-gov     public             1       0         0          0
 2 " <=50… Self-emp-no… self_empl          0       0         1          0
 3 " <=50… Private       Private            0       1         0          0
 4 " <=50… Private       Private            0       1         0          0
 5 " <=50… Private       Private            0       1         0          0
 6 " <=50… Private       Private            0       1         0          0
 7 " <=50… Private       Private            0       1         0          0
 8 " >50K" Self-emp-no… self_empl          0       0         1          0
 9 " >50K" Private       Private            0       1         0          0
10 " >50K" Private       Private            0       1         0          0
# ... with 32,551 more rows
```

Let's practice!

FEATURE ENGINEERING IN R

Binning encoding: data driven

FEATURE ENGINEERING IN R

Jose Hernandez

Data Scientist, University of Washington

Education levels

```
adult_incomes %>%
    select(education) %>%
    table()
```

```
        10th         11th         12th      1st-4th      5th-6th
         933         1175          433          168          333
     7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
         646          514         1067         1382         5355
   Doctorate      HS-grad      Masters    Preschool  Prof-school
         413        10501         1723           51          576
Some-college
        7291
```

```
ed_table <- adult_incomes %>%
    select(education, income) %>%
    table()
```

```
prop_results <- as_tibble(prop.table(ed_table, 1))
```

```
                   <=50K       >50K
   10th         0.93354770 0.06645230
   11th         0.94893617 0.05106383
   12th         0.92378753 0.07621247
   1st-4th      0.96428571 0.03571429
   5th-6th      0.95195195 0.04804805
   7th-8th      0.93808050 0.06191950
   9th          0.94747082 0.05252918
   Assoc-acdm   0.75164011 0.24835989
   Assoc-voc    0.73878437 0.26121563
   Bachelors    0.58524743 0.41475257
   Doctorate    0.25907990 0.74092010
   HS-grad      0.84049138 0.15950862
   Masters      0.44341265 0.55658735
   Preschool    1.00000000 0.00000000
   Prof-school  0.26562500 0.73437500
   Some-college 0.80976546 0.19023454
```

Leveraging data to inform groupings

```
prop_results %>%
  filter(income == " >50K") %>%
  arrange(n)
```

```
# A tibble: 16 x 3
   education        income        n
   <chr>            <chr>     <dbl>
 1 " Preschool"     " >50K" 0
 2 " 1st-4th"       " >50K" 0.0357
 3 " 5th-6th"       " >50K" 0.0480
 4 " 11th"          " >50K" 0.0511
 5 " 9th"           " >50K" 0.0525
 6 " 7th-8th"       " >50K" 0.0619
 7 " 10th"          " >50K" 0.0665
 8 " 12th"          " >50K" 0.0762
 9 " HS-grad"       " >50K" 0.160
10 " Some-college" " >50K" 0.190
```

low-education	**0% - 10%**
medium-education	**10% - 30%**
high-education	**30% - 100%**

Encoding meaning using ranges

```
inner_join(adult_incomes, prop_results,
        by = "education" = "ed_span") %>%
    select(education, income, n) %>%
    head()
```

```
   education income            n
1  Bachelors    <=50K 0.41475257
2  Bachelors    <=50K 0.41475257
3    HS-grad    <=50K 0.15950862
4       11th    <=50K 0.05106383
5  Bachelors    <=50K 0.41475257
6    Masters    <=50K 0.55658735
```

```r
adult_incomes %>%
   mutate(education_levels =
            case_when(prop >= 0 & prop < .10 ~ "low_education",
                      prop >= .10 & prop < .30 ~ "medium_education",
                      prop >= .30 & prop < 1 ~ "high_education")) %>%
   head()
```

```
# A tibble: 32,561 x 4
   income   education      prop education_levels
   <fct>    <chr>         <dbl> <chr>
 1 " <=50K" " Bachelors" 0.415  high_education
 2 " <=50K" " Bachelors" 0.415  high_education
 3 " <=50K" " HS-grad"   0.160  medium_education
 4 " <=50K" " 11th"      0.0511 low_education
 5 " <=50K" " Bachelors" 0.415  high_education
 6 " <=50K" " Masters"   0.557  high_education
 7 " <=50K" " 9th"       0.0525 low_education
 8 " >50K"  " HS-grad"   0.160  medium_education
 9 " >50K"  " Masters"   0.557  high_education
10 " >50K"  " Bachelors" 0.415  high_education
# ... with 32,551 more rows
```

Let's practice!

FEATURE ENGINEERING IN R