Feature engineering

Data Science in a Box

datasciencebox.org

Feature engineering

Feature engineering

- We prefer simple models when possible, but **parsimony** does not mean sacrificing accuracy (or predictive performance) in the interest of simplicity
- Variables that go into the model and how they are represented are just as critical to success of the model
- **Feature engineering** allows us to get creative with our predictors in an effort to make them more useful for our model (to increase its predictive performance)

Same training and testing sets as before

```r
# Fix random numbers by setting the seed
# Enables analysis to be reproducible when random numbers are used
set.seed(1116)

# Put 80% of the data into the training set
email_split <- initial_split(email, prop = 0.80)

# Create data frames for the two sets:
train_data <- training(email_split)
test_data  <- testing(email_split)
```

A simple approach: `mutate()`

```
train_data %>%
  mutate(
    date = lubridate::date(time),
    dow  = wday(time),
    month = month(time)
    ) %>%
  select(time, date, dow, month) %>%
  sample_n(size = 5) # shuffle to show a variety
```

```
## # A tibble: 5 × 4
##   time                date         dow month
##   <dttm>              <date>     <dbl> <dbl>
## 1 2012-03-15 20:51:35 2012-03-15     5     3
## 2 2012-03-03 16:24:02 2012-03-03     7     3
## 3 2012-01-18 18:55:23 2012-01-18     4     1
## 4 2012-02-25 06:08:59 2012-02-25     7     2
## 5 2012-01-11 15:18:51 2012-01-11     4     1
```

Modeling workflow, revisited

- Create a **recipe** for feature engineering steps to be applied to the training data
- Fit the model to the training data after these steps have been applied
- Using the model estimates from the training data, predict outcomes for the test data
- Evaluate the performance of the model on the test data

Building recipes

Initiate a recipe

```r
email_rec <- recipe(
  spam ~ .,          # formula
  data = train_data  # data to use for cataloguing names and types of variables
  )

summary(email_rec)
```

```
## # A tibble: 21 × 4
##    variable     type    role      source  
##    <chr>        <chr>   <chr>     <chr>   
##  1 to_multiple  nominal predictor original
##  2 from         nominal predictor original
##  3 cc           numeric predictor original
##  4 sent_email   nominal predictor original
##  5 time         date    predictor original
##  6 image        numeric predictor original
##  7 attach       numeric predictor original
##  8 dollar       numeric predictor original
##  9 winner       nominal predictor original
## 10 inherit      numeric predictor original
## 11 viagra       numeric predictor original
## 12 password     numeric predictor original
## 13 num_char     numeric predictor original
## 14 line_breaks  numeric predictor original
## 15 format       nominal predictor original
## 16 re_subj      nominal predictor original
## 17 exclaim_subj numeric predictor original
## 18 urgent_subj  nominal predictor original
## 19 exclaim_mess numeric predictor original
## 20 number       nominal predictor original
## 21 spam         nominal outcome   original
```

Remove certain variables

```
email_rec <- email_rec %>%
  step_rm(from, sent_email)
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variables removed from, sent_email
```

Feature engineer date

```r
email_rec <- email_rec %>%
  step_date(time, features = c("dow", "month")) %>%
  step_rm(time)
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variables removed from, sent_email
## Date features from time
## Variables removed time
```

Discretize numeric variables

```r
email_rec <- email_rec %>%
  step_cut(cc, attach, dollar, breaks = c(0, 1)) %>%
  step_cut(inherit, password, breaks = c(0, 1, 5, 10, 20))
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variables removed from, sent_email
## Date features from time
## Variables removed time
## Cut numeric for cc, attach, dollar
## Cut numeric for inherit, password
```

Create dummy variables

```r
email_rec <- email_rec %>%
  step_dummy(all_nominal(), -all_outcomes())
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variables removed from, sent_email
## Date features from time
## Variables removed time
## Cut numeric for cc, attach, dollar
## Cut numeric for inherit, password
## Dummy variables from all_nominal(), -all_outcomes()
```

Remove zero variance variables

Variables that contain only a single value

```
email_rec <- email_rec %>%
  step_zv(all_predictors())
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         20
##
## Operations:
##
## Variables removed from, sent_email
## Date features from time
## Variables removed time
## Cut numeric for cc, attach, dollar
## Cut numeric for inherit, password
## Dummy variables from all_nominal(), -all_outcomes()
## Zero variance filter on all_predictors()
```

All in one place

```
email_rec <- recipe(spam ~ ., data = email) %>%
  step_rm(from, sent_email) %>%
  step_date(time, features = c("dow", "month")) %>%
  step_rm(time) %>%
  step_cut(cc, attach, dollar, breaks = c(0, 1)) %>%
  step_cut(inherit, password, breaks = c(0, 1, 5, 10, 20)) %>%
  step_dummy(all_nominal(), -all_outcomes()) %>%
  step_zv(all_predictors())
```

Building workflows

Define model

```r
email_mod <- logistic_reg() %>%
  set_engine("glm")

email_mod
```

```
## Logistic Regression Model Specification (classification)
##
## Computational engine: glm
```

Define workflow

Workflows bring together models and recipes so that they can be easily applied to both the training and test data.

```
email_wflow <- workflow() %>%
  add_model(email_mod) %>%
  add_recipe(email_rec)
```

```
## == Workflow ==========================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------------
## 7 Recipe Steps
##
## • step_rm()
## • step_date()
## • step_rm()
## • step_cut()
## • step_cut()
## • step_dummy()
## • step_zv()
##
## -- Model -----------------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Computational engine: glm
```

Fit model to training data

```
email_fit <- email_wflow %>%
  fit(data = train_data)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
tidy(email_fit) %>% print(n = 31)
```

```
## # A tibble: 32 × 5
##    term                 estimate  std.error statistic  p.value
##    <chr>                   <dbl>      <dbl>     <dbl>    <dbl>
##  1 (Intercept)          -0.707      0.252    -2.80    5.04e- 3
##  2 image                -1.65       0.949    -1.74    8.27e- 2
##  3 viagra                2.42     300.        0.00806 9.94e- 1
##  4 num_char              0.0475     0.0243    1.95    5.11e- 2
##  5 line_breaks          -0.00514    0.00138  -3.72    2.03e- 4
##  6 exclaim_subj         -0.205      0.277    -0.740   4.59e- 1
##  7 exclaim_mess          0.00879    0.00186   4.72    2.31e- 6
##  8 to_multiple_X1       -2.56       0.354    -7.24    4.61e-13
##  9 cc_X.1.68.           -0.289      0.490    -0.590   5.55e- 1
## 10 attach_X.1.21.        2.03       0.369     5.51    3.67e- 8
## 11 dollar_X.1.64.        0.246      0.216     1.14    2.56e- 1
## 12 winner_yes            2.15       0.430     5.00    5.64e- 7
## 13 inherit_X.1.5.      -10.5     1241.       -0.00843 9.93e- 1
## 14 inherit_X.5.10.       2.48       1.47      1.69    9.16e- 2
## 15 password_X.1.5.      -1.73       0.747    -2.31    2.08e- 2
## 16 password_X.5.10.   -13.5       776.       -0.0174  9.86e- 1
## 17 password_X.10.20.  -14.9      1322.       -0.0112  9.91e- 1
## 18 password_X.20.22.  -15.0      1697.       -0.00886 9.93e- 1
## 19 format_X1            -0.904      0.159    -5.69    1.29e- 8
## 20 re_subj_X1           -2.89       0.437    -6.63    3.37e-11
## 21 urgent_subj_X1        3.50       1.07      3.28    1.05e- 3
## 22 number_small         -0.892      0.167    -5.34    9.41e- 8
## 23 number_big           -0.183      0.250    -0.731   4.65e- 1
## 24 time_dow_Mon         -0.340      0.295    -1.15    2.49e- 1
## 25 time_dow_Tue         -0.00277    0.275    -0.0101  9.92e- 1
## 26 time_dow_Wed         -0.223      0.269    -0.830   4.06e- 1
## 27 time_dow_Thu         -0.328      0.277    -1.18    2.36e- 1
## 28 time_dow_Fri         -0.0534     0.270    -0.198   8.43e- 1
## 29 time_dow_Sat          0.0536     0.290     0.185   8.53e- 1
## 30 time_month_Feb        0.800      0.181     4.42    9.85e- 6
## 31 time_month_Mar        0.587      0.181     3.24    1.18e- 3
## # … with 1 more row
```

Make predictions for test data

```r
email_pred <- predict(email_fit, test_data, type = "prob") %>%
  bind_cols(test_data)
```

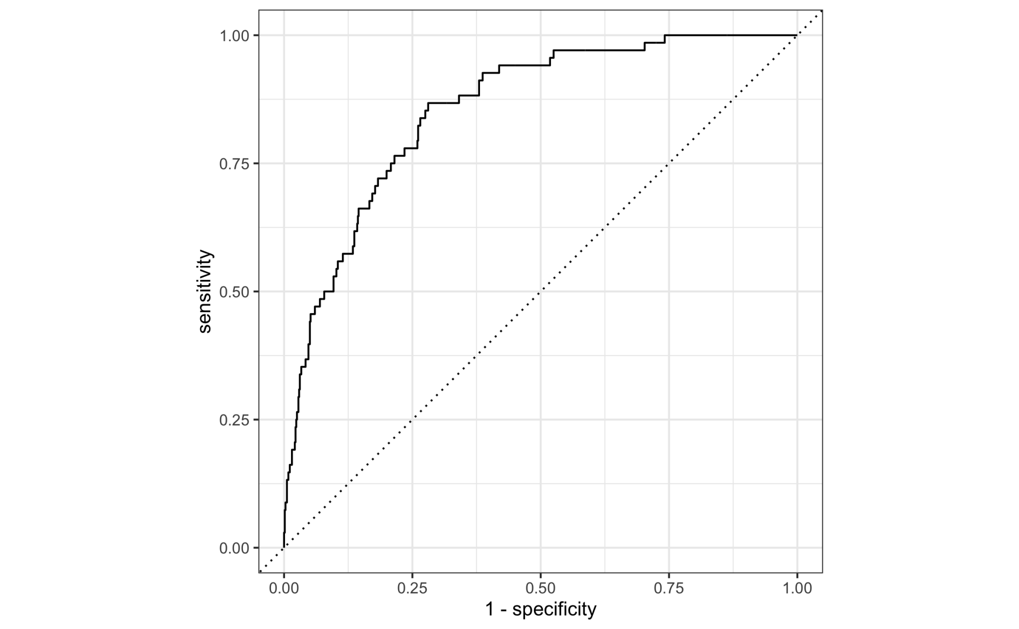
```
## Warning: There are new levels in a factor: NA
```

```r
email_pred
```

```
## # A tibble: 785 × 23
##   .pred_0  .pred_1 spam  to_multiple from     cc sent_email
##     <dbl>    <dbl> <fct> <fct>       <fct> <int> <fct>
## 1   0.994 0.00602  0     1           1         0 1
## 2   0.998 0.00164  0     0           1         1 1
## 3   0.972 0.0281   0     0           1         0 0
## 4   0.999 0.000652 0     0           1         1 0
## 5   0.995 0.00546  0     0           1         4 0
## 6   0.881 0.119    0     0           1         0 0
## # … with 779 more rows, and 16 more variables: time <dttm>,
## #   image <dbl>, attach <dbl>, dollar <dbl>, winner <fct>,
## #   inherit <dbl>, viagra <dbl>, password <dbl>, num_char <dbl>,
## #   line_breaks <int>, format <fct>, re_subj <fct>,
## #   exclaim_subj <dbl>, urgent_subj <fct>, exclaim_mess <dbl>,
```

Evaluate the performance

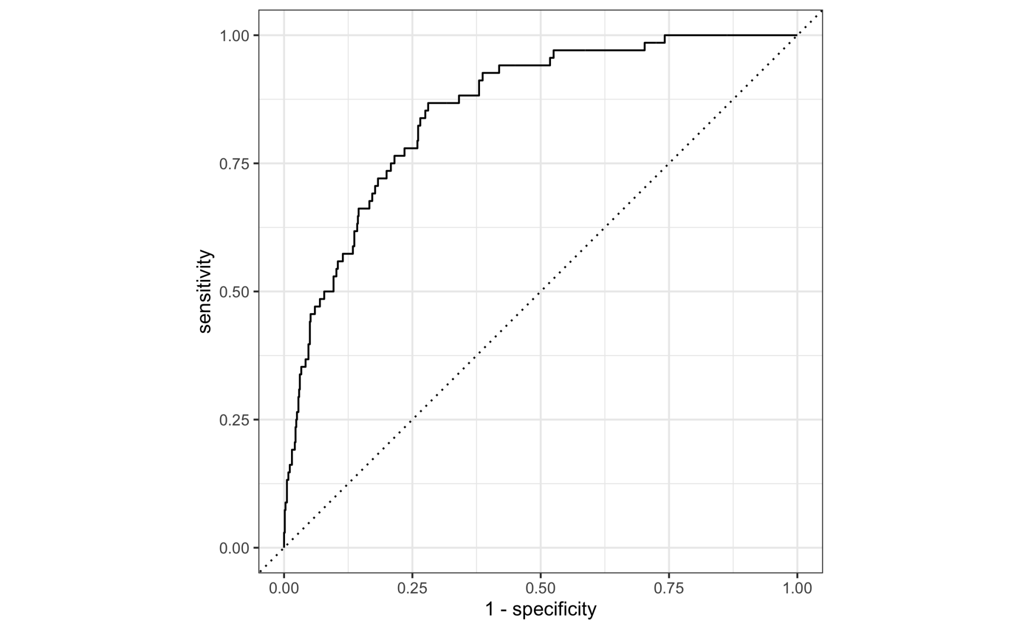
```r
email_pred %>%
  roc_curve(
    truth = spam,
    .pred_1,
    event_level = "second"
  ) %>%
  autoplot()
```



Evaluate the performance

```r
email_pred %>%
  roc_auc(
    truth = spam,
    .pred_1,
    event_level = "second"
  )
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.856
```



Making decisions

Cutoff probability: 0.5

Output Code

Suppose we decide to label an email as spam if the model predicts the probability of spam to be **more than 0.5**.

	Email is not spam	Email is spam
Email labelled not spam	705	57
Email labelled spam	11	11
NA	1	NA

Cutoff probability: 0.5

Output | Code

```
cutoff_prob <- 0.5
email_pred %>%
  mutate(
    spam      = if_else(spam == 1, "Email is spam", "Email is not spam"),
    spam_pred = if_else(.pred_1 > cutoff_prob, "Email labelled spam", "Email labelled not
    ) %>%
  count(spam_pred, spam) %>%
  pivot_wider(names_from = spam, values_from = n) %>%
  kable(col.names = c("", "Email is not spam", "Email is spam"))
```

Cutoff probability: 0.25

Output Code

Suppose we decide to label an email as spam if the model predicts the probability of spam to be **more than 0.25**.

	Email is not spam	Email is spam
Email labelled not spam	660	34
Email labelled spam	56	34
NA	1	NA

Cutoff probability: 0.25

Output | Code

```
cutoff_prob <- 0.25
email_pred %>%
  mutate(
    spam      = if_else(spam == 1, "Email is spam", "Email is not spam"),
    spam_pred = if_else(.pred_1 > cutoff_prob, "Email labelled spam", "Email labelled not
    ) %>%
  count(spam_pred, spam) %>%
  pivot_wider(names_from = spam, values_from = n) %>%
  kable(col.names = c("", "Email is not spam", "Email is spam"))
```

Cutoff probability: 0.75

Output Code

Suppose we decide to label an email as spam if the model predicts the probability of spam to be **more than 0.75**.

	Email is not spam	Email is spam
Email labelled not spam	715	65
Email labelled spam	1	3
NA	1	NA

Cutoff probability: 0.75

Output Code

```
cutoff_prob <- 0.75
email_pred %>%
  mutate(
    spam      = if_else(spam == 1, "Email is spam", "Email is not spam"),
    spam_pred = if_else(.pred_1 > cutoff_prob, "Email labelled spam", "Email labelled not
    ) %>%
  count(spam_pred, spam) %>%
  pivot_wider(names_from = spam, values_from = n) %>%
  kable(col.names = c("", "Email is not spam", "Email is spam"))
```