Imbalance Class

GS

4/8/2021

Load datasets

```
train_raw <- read_csv("train.csv", guess_max = 1e5) %>%
  mutate(damaged = if_else(damaged > 0, "damaged", "no damaged"))

test_raw <- read_csv("test.csv", guess_max = 1e5)
```

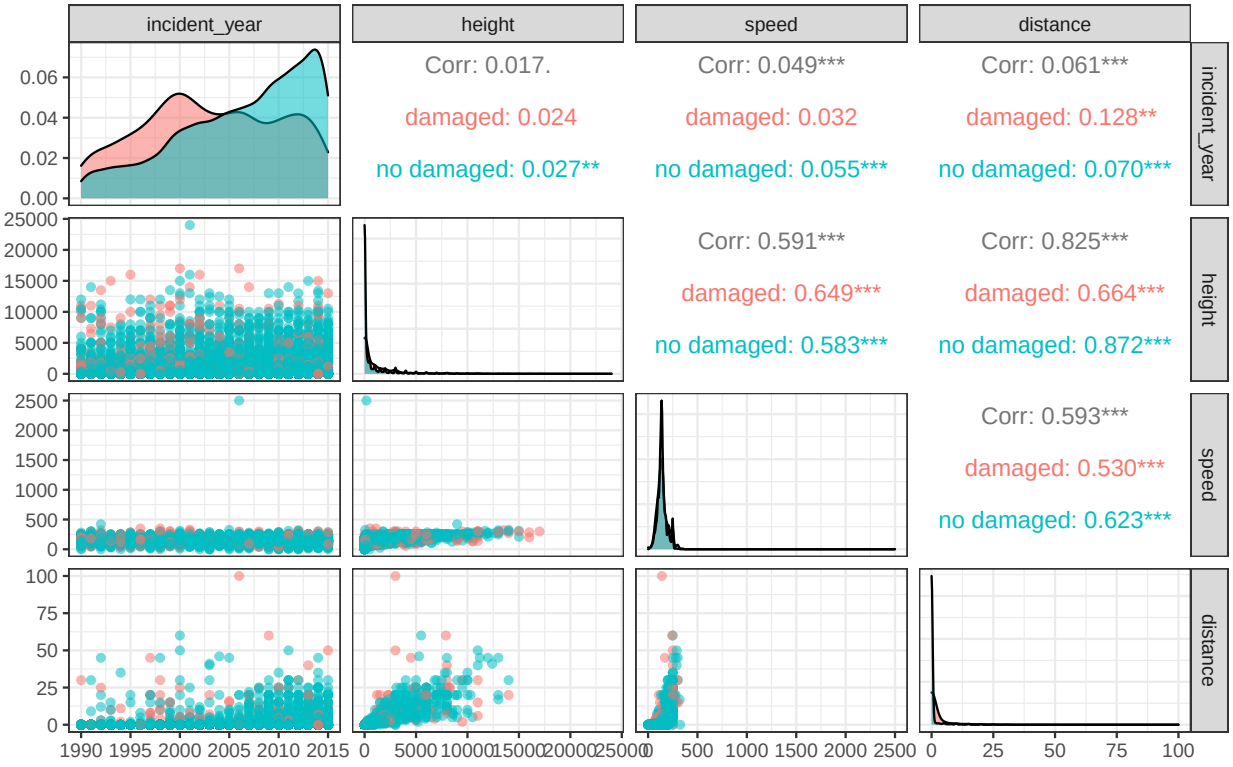
EDA

```
glimpse(train_raw)
```

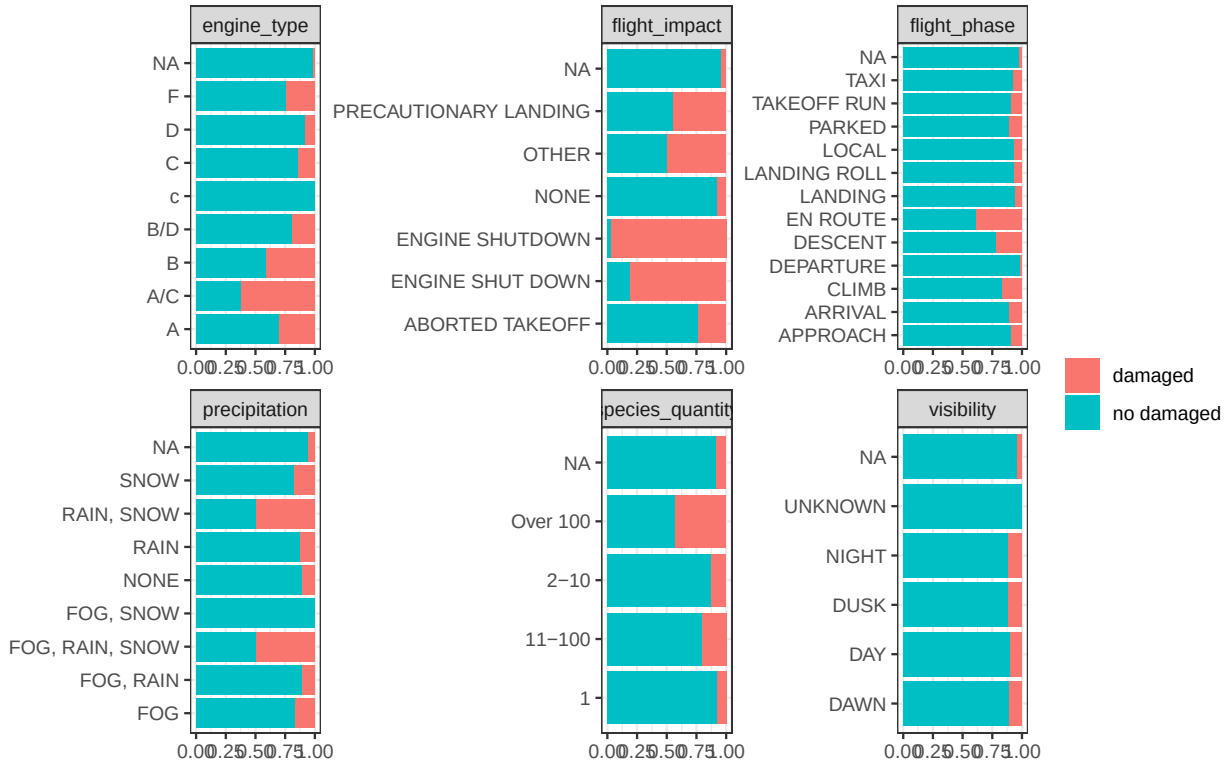
```
Rows: 21,000
Columns: 34
$ id               <dbl> 23637, 8075, 5623, 19605, 15142, 27235, 12726, 20781,~
$ incident_year    <dbl> 1996, 1999, 2011, 2007, 2007, 2013, 2002, 2013, 2015,~
$ incident_month   <dbl> 11, 6, 12, 9, 9, 5, 5, 5, 7, 8, 10, 9, 11, 7, 5, 3, 3~
$ incident_day     <dbl> 7, 26, 1, 13, 13, 28, 4, 19, 22, 22, 21, 7, 2, 7, 20,~
$ operator_id      <chr> "MIL", "UAL", "SWA", "SWA", "MIL", "UNK", "UAL", "BUS~
$ operator         <chr> "MILITARY", "UNITED AIRLINES", "SOUTHWEST AIRLINES", ~
$ aircraft         <chr> "T-1A", "B-757-200", "B-737-300", "B-737-700", "KC-13~
$ aircraft_type    <chr> "A", "A", "A", "A", "A", NA, "A", "A", NA, "A", "A", ~
$ aircraft_make    <chr> "748", "148", "148", "148", NA, NA, "148", "226", NA,~
$ aircraft_model   <dbl> NA, 26, 24, 42, NA, NA, 97, 7, NA, NA, 14, 22, 37, NA~
$ aircraft_mass    <dbl> 3, 4, 4, 4, NA, NA, 4, 1, NA, 1, 3, 4, 4, NA, 4, 4, 4~
$ engine_make      <dbl> 31, 34, 10, 10, NA, NA, 34, 7, NA, NA, 1, 34, 34, NA,~
$ engine_model     <chr> "1", "40", "1", "1", NA, NA, "46", "10", NA, NA, "10"~
$ engines          <dbl> 2, 2, 2, 2, NA, NA, 2, 1, NA, 2, 2, 2, 2, NA, 2, 2, 2~
$ engine_type      <chr> "D", "D", "D", "D", NA, NA, "D", "A", NA, "C", "D", "~
$ engine1_position <dbl> 5, 1, 1, 1, NA, NA, 1, 7, NA, 3, 5, 5, 5, NA, 1, 1, 5~
$ engine2_position <dbl> 5, 1, 1, 1, NA, NA, 1, NA, NA, 3, 5, 5, 5, NA, 1, 1, ~
$ engine3_position <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ engine4_position <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ airport_id       <chr> "KLBB", "ZZZZ", "KOAK", "KSAT", "KGFK", "KMDT", "KJFK~
$ airport          <chr> "LUBBOCK PRESTON SMITH INTL ARPT", "UNKNOWN", "METRO ~
$ state            <chr> "TX", NA, "CA", "TX", "ND", "PA", "NY", "IL", "NJ", "~
$ faa_region       <chr> "ASW", NA, "AWP", "ASW", "AGL", "AEA", "AEA", "AGL", ~
$ flight_phase     <chr> "LANDING ROLL", NA, "LANDING ROLL", "APPROACH", "APPR~
$ visibility       <chr> "DAY", NA, "DAY", "NIGHT", "NIGHT", NA, NA, "NIGHT", ~
```

```
$ precipitation    <chr> NA, NA, "NONE", "NONE", NA, NA, NA, "FOG", NA, NA, "N~
$ height           <dbl> 0, NA, 0, 300, NA, NA, NA, 2700, NA, 0, 3500, 1400, 0~
$ speed            <dbl> 80, NA, NA, 130, 140, NA, NA, 110, NA, NA, 180, 170, ~
$ distance         <dbl> 0, NA, 0, NA, NA, 0, NA, NA, 0, 0, NA, NA, 0, 0, 0, 0~
$ species_id       <chr> "UNKBM", "UNKBM", "ZT002", "UNKBS", "ZT105", "YI005",~
$ species_name     <chr> "UNKNOWN MEDIUM BIRD", "UNKNOWN MEDIUM BIRD", "WESTER~
$ species_quantity <chr> "1", "1", "1", "1", NA, "1", "1", "1", "1", "1", "1",~
$ flight_impact    <chr> NA, NA, "NONE", "NONE", NA, NA, NA, "NONE", NA, NA, "~
$ damaged          <chr> "no damaged", "damaged", "no damaged", "no damaged", ~
```

```
train_raw %>%
  select(damaged, incident_year, height, speed, distance) %>%
  ggpairs(columns = 2:5, aes(color = damaged, alpha = .5))
```




```
  train_raw %>%
  select(damaged, precipitation, visibility, engine_type,
         flight_impact, flight_phase, species_quantity) %>%
  pivot_longer(precipitation : species_quantity) %>%
  ggplot(aes(y = value, fill = damaged)) +
  geom_bar(position = "fill") +
  facet_wrap(vars(name), scales = "free") +
  labs(x = NULL, y = NULL, fill = NULL)
```




```
bird_df <- train_raw %>%
  select(damaged, flight_impact, precipitation, visibility, flight_phase,
         engines, incident_year, incident_month, species_id, engine_type,
         aircraft_model, species_quantity, height, speed)
```

Build model

```
set.seed(123)
bird_folds <- vfold_cv(train_raw,
                       v = 5,
                       strata = damaged)

bird_metrics <- metric_set(mn_log_loss, accuracy,
                           sensitivity, specificity)
```

```
bird_rec <- recipe(damaged ~ ., data = bird_df) %>%
  step_novel(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.01) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_impute_median(all_numeric_predictors()) %>%
  step_zv(all_predictors())

bird_rec
```

```
Data Recipe
```

```
Inputs:

      role #variables
   outcome          1
 predictor         13

Operations:

Novel factor level assignment for all_nominal_predictors()
Collapsing factor levels for all_nominal_predictors()
Unknown factor level assignment for all_nominal_predictors()
Median Imputation for all_numeric_predictors()
Zero variance filter on all_predictors()
```

```
bird_rec %>% prep() %>% juice()
```

```
# A tibble: 21,000 x 14
   flight_impact precipitation visibility flight_phase engines incident_year
   <fct>         <fct>         <fct>      <fct>          <dbl>         <dbl>
 1 unknown       unknown       DAY        LANDING ROLL       2          1996
 2 unknown       unknown       unknown    unknown            2          1999
 3 NONE          NONE          DAY        LANDING ROLL       2          2011
 4 NONE          NONE          NIGHT      APPROACH           2          2007
 5 unknown       unknown       NIGHT      APPROACH           2          2007
 6 unknown       unknown       unknown    unknown            2          2013
 7 unknown       unknown       unknown    APPROACH           2          2002
 8 NONE          FOG           NIGHT      DESCENT            1          2013
 9 unknown       unknown       unknown    unknown            2          2015
10 unknown       unknown       NIGHT      LANDING ROLL       2          2007
# ... with 20,990 more rows, and 8 more variables: incident_month <dbl>,
#   species_id <fct>, engine_type <fct>, aircraft_model <dbl>,
#   species_quantity <fct>, height <dbl>, speed <dbl>, damaged <fct>
```

```
bird_df %>% count(damaged) # imbalance
```

```
# A tibble: 2 x 2
  damaged        n
  <chr>      <int>
1 damaged     1799
2 no damaged 19201
```

```
library(baguette)

bag_spec <-
  bag_tree(min_n = 10) %>%
  set_engine("rpart", times = 25) %>%
  set_mode("classification")

bag_spec
```

```
Bagged Decision Tree Model Specification (classification)
```

```
Main Arguments:
  cost_complexity = 0
  min_n = 10

Engine-Specific Arguments:
  times = 25

Computational engine: rpart
```

```
imb_wf <-
  workflow() %>%
  add_recipe(bird_rec) %>%
  add_model(bag_spec)

imb_wf
```

```
== Workflow ====================================================================
Preprocessor: Recipe
Model: bag_tree()

-- Preprocessor ----------------------------------------------------------------
5 Recipe Steps

* step_novel()
* step_other()
* step_unknown()
* step_impute_median()
* step_zv()

-- Model -----------------------------------------------------------------------
Bagged Decision Tree Model Specification (classification)

Main Arguments:
  cost_complexity = 0
  min_n = 10

Engine-Specific Arguments:
  times = 25

Computational engine: rpart
```

```
fit(imb_wf, data = bird_df)
```

```
== Workflow [trained] ==========================================================
Preprocessor: Recipe
Model: bag_tree()

-- Preprocessor ----------------------------------------------------------------
5 Recipe Steps

* step_novel()
* step_other()
```

```
* step_unknown()
* step_impute_median()
* step_zv()

-- Model ---------------------------------------------------------------------
Bagged CART (classification with 25 members)

Variable importance scores include:

# A tibble: 13 x 4
   term             value std.error  used
   <chr>            <dbl>     <dbl> <int>
 1 flight_impact    480.       6.81    25
 2 aircraft_model   363.       4.97    25
 3 incident_year    354.       5.51    25
 4 species_id       337.       4.62    25
 5 height           332.       5.45    25
 6 speed            297.       4.82    25
 7 incident_month   285.       6.18    25
 8 flight_phase     246.       4.41    25
 9 engine_type      213.       3.31    25
10 visibility       196.       3.82    25
11 precipitation    136.       3.23    25
12 engines          117.       2.67    25
13 species_quantity  83.7      3.12    25
```

Resample & compare models

```
doParallel::registerDoParallel()
set.seed(123)

imb_res <- fit_resamples(
  imb_wf,
  resamples = bird_folds,
  metrics = bird_metrics
)

imb_res
```

```
# Resampling results
# 5-fold cross-validation using stratification
# A tibble: 5 x 4
  splits               id    .metrics          .notes
  <list>               <chr> <list>            <list>
1 <split [16800/4200]> Fold1 <tibble [4 x 4]> <tibble [0 x 1]>
2 <split [16800/4200]> Fold2 <tibble [4 x 4]> <tibble [0 x 1]>
3 <split [16800/4200]> Fold3 <tibble [4 x 4]> <tibble [0 x 1]>
4 <split [16800/4200]> Fold4 <tibble [4 x 4]> <tibble [0 x 1]>
5 <split [16800/4200]> Fold5 <tibble [4 x 4]> <tibble [0 x 1]>
```

```
collect_metrics(imb_res)
```

```
# A tibble: 4 x 6
  .metric     .estimator  mean     n  std_err .config
  <chr>       <chr>      <dbl> <int>    <dbl> <chr>
1 accuracy    binary     0.925     5 0.00221  Preprocessor1_Model1
2 mn_log_loss binary     0.212     5 0.00511  Preprocessor1_Model1
3 sens        binary     0.278     5 0.00941  Preprocessor1_Model1
4 spec        binary     0.986     5 0.000843 Preprocessor1_Model1
```

```
library(themis)

bal_rec <-
  bird_rec %>%
  step_dummy(all_nominal_predictors()) %>%
  step_smote(damaged)

bal_rec
```

```
Data Recipe

Inputs:

      role #variables
   outcome          1
 predictor         13

Operations:

Novel factor level assignment for all_nominal_predictors()
Collapsing factor levels for all_nominal_predictors()
Unknown factor level assignment for all_nominal_predictors()
Median Imputation for all_numeric_predictors()
Zero variance filter on all_predictors()
Dummy variables from all_nominal_predictors()
SMOTE based on damaged
```

```
bal_rec %>% prep() %>%  juice()
```

```
# A tibble: 38,402 x 52
   engines incident_year incident_month aircraft_model height speed
     <dbl>         <dbl>          <dbl>          <dbl>  <dbl> <dbl>
 1       2          1996             11             22      0    80
 2       2          1999              6             26     50   137
 3       2          2011             12             24      0   137
 4       2          2007              9             42    300   130
 5       2          2007              9             22     50   140
 6       2          2013              5             22     50   137
 7       2          2002              5             97     50   137
 8       1          2013              5              7   2700   110
 9       2          2015              7             22     50   137
10       2          2007              8             22      0   137
```

```
# ... with 38,392 more rows, and 46 more variables: flight_impact_NONE <dbl>,
#   flight_impact_OTHER <dbl>, flight_impact_PRECAUTIONARY.LANDING <dbl>,
#   flight_impact_other <dbl>, flight_impact_unknown <dbl>,
#   precipitation_NONE <dbl>, precipitation_RAIN <dbl>,
#   precipitation_other <dbl>, precipitation_unknown <dbl>,
#   visibility_DAY <dbl>, visibility_DUSK <dbl>, visibility_NIGHT <dbl>,
#   visibility_other <dbl>, visibility_unknown <dbl>, flight_phase_CLIMB <dbl>,
#   flight_phase_DESCENT <dbl>, flight_phase_EN.ROUTE <dbl>,
#   flight_phase_LANDING.ROLL <dbl>, flight_phase_TAKEOFF.RUN <dbl>,
#   flight_phase_other <dbl>, flight_phase_unknown <dbl>,
#   species_id_K5114 <dbl>, species_id_N5111 <dbl>, species_id_NE1 <dbl>,
#   species_id_O2111 <dbl>, species_id_O2205 <dbl>, species_id_UNKB <dbl>,
#   species_id_UNKBL <dbl>, species_id_UNKBM <dbl>, species_id_UNKBS <dbl>,
#   species_id_YH004 <dbl>, species_id_YI005 <dbl>, species_id_YL001 <dbl>,
#   species_id_ZT001 <dbl>, species_id_ZX3 <dbl>, species_id_other <dbl>,
#   species_id_unknown <dbl>, engine_type_C <dbl>, engine_type_D <dbl>,
#   engine_type_F <dbl>, engine_type_other <dbl>, engine_type_unknown <dbl>,
#   species_quantity_X2.10 <dbl>, species_quantity_other <dbl>,
#   species_quantity_unknown <dbl>, damaged <fct>
```

```
bal_rec %>% prep() %>%  juice() %>% count(damaged) # balanced
```

```
# A tibble: 2 x 2
  damaged        n
  <fct>      <int>
1 damaged    19201
2 no damaged 19201
```

```
bal_wf <-
  workflow() %>%
  add_recipe(bal_rec) %>%
  add_model(bag_spec)

set.seed(123)

bal_res <- fit_resamples(
  bal_wf,
  resamples = bird_folds,
  metrics = bird_metrics
)

bal_res
```

```
# Resampling results
# 5-fold cross-validation using stratification
# A tibble: 5 x 4
  splits               id    .metrics          .notes
  <list>               <chr> <list>            <list>
1 <split [16800/4200]> Fold1 <tibble [4 x 4]> <tibble [0 x 1]>
2 <split [16800/4200]> Fold2 <tibble [4 x 4]> <tibble [0 x 1]>
3 <split [16800/4200]> Fold3 <tibble [4 x 4]> <tibble [0 x 1]>
4 <split [16800/4200]> Fold4 <tibble [4 x 4]> <tibble [0 x 1]>
5 <split [16800/4200]> Fold5 <tibble [4 x 4]> <tibble [0 x 1]>
```

```
collect_metrics(bal_res)
```

```
# A tibble: 4 x 6
  .metric     .estimator  mean     n  std_err .config
  <chr>       <chr>      <dbl> <int>    <dbl> <chr>
1 accuracy    binary     0.919     5 0.00252  Preprocessor1_Model1
2 mn_log_loss binary     0.225     5 0.00604  Preprocessor1_Model1
3 sens        binary     0.317     5 0.0129   Preprocessor1_Model1
4 spec        binary     0.976     5 0.000847 Preprocessor1_Model1
```