ML_output_fig

Steph Blain

2025-07-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(forcats)
library(scico)
```

```
## Warning: package 'scico' was built under R version 4.4.3
```

```
theme_set(theme_classic())

shap1<-read_csv("C:/Users/Steph/GitHub/thrush_hybrids/migratory_traits/rf_shap_dependence_data_202507.cs
```

```
## Rows: 359 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (28): fall_detectDay1, fall_bearing1, distal, p10, bodyCondition, tarsus...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
features1<-read_csv("C:/Users/Steph/GitHub/thrush_hybrids/migratory_traits/rf_feature_importances_20250
```

```
## Rows: 14 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Feature, Category
## dbl (1): Importance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
knn1<-read_csv("C:/Users/Steph/GitHub/thrush_hybrids/migratory_traits/knn_imputed_full_dataset.csv")
```

```
## Rows: 479 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): set
## dbl (15): fall_detectDay1, fall_bearing1, distal, p10, bodyCondition, tarsus...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
renameFeatures1<-
  data.frame(Feature=c("release_year","fall_detectDay1","aims_heterozygosity",
                       "bodyCondition","sex_binary","fall_bearing1","kipps",
                       "tail.length","distal","releaseDay","wing.cord","p10",
                       "aims_ancestry","tarsus.length"),
             FeatureRenamed=c("Release year","Fall timing","Heterozygosity",
                       "Body condition","Sex","Fall orientation","Kipps (wing)",
                       "Tail length","Distal (wing)","Release day","Wing cord","p10 (feather)",
                       "Ancestry","Tarsus length"),
             FeatureCategory=c("Other","Behaviour","Genetics",
                       "Morphology","Other","Behaviour","Morphology",
                       "Morphology","Morphology","Behaviour","Morphology","Morphology",
                       "Genetics","Morphology"))

features1<-features1%>%select(-Category)%>%left_join(renameFeatures1)
```

```
## Joining with `by = join_by(Feature)`
```

```
#get expected row count for remodeled df
ncol(shap1)/2*nrow(shap1)
```

```
## [1] 5026
```

```
shap1<-shap1%>%
  pivot_longer(!starts_with("SHAP"),names_to="Feature",values_to="FeatureValue")%>%
  pivot_longer(starts_with("SHAP"),names_to="Feature2",values_to="SHAPValue")%>%
  mutate(Feature2=gsub("SHAP_","",Feature2))%>%
  filter(Feature==Feature2)%>%
  left_join(renameFeatures1)
```

```
## Joining with `by = join_by(Feature)`
```

```
nrow(shap1)
```

```
## [1] 5026
```

- verify that features importances sum to 1 and all features are unique
- extract traits with a feature importance greater than expected

```
round(sum(features1$Importance),6)==1
```

```
## [1] TRUE
```

```
unique(features1$Feature)==features1$Feature
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

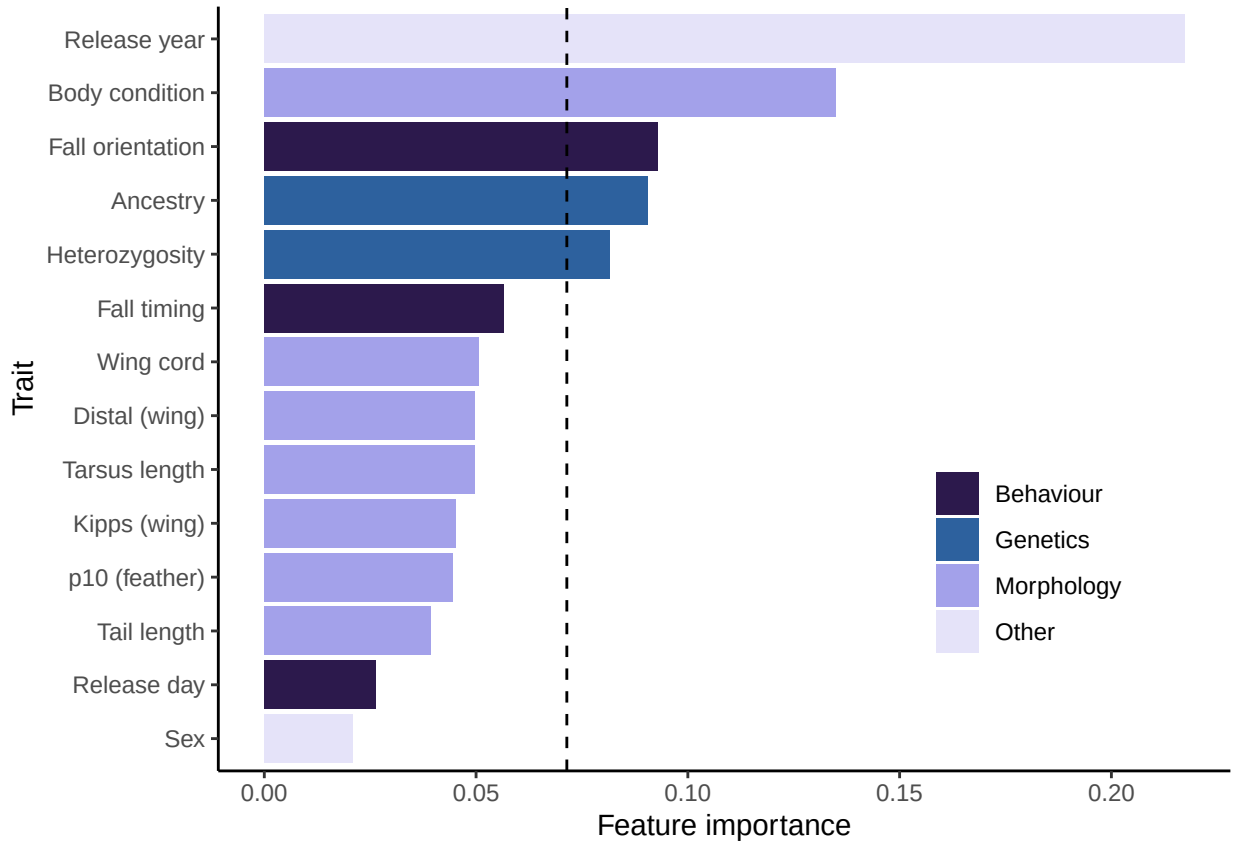
```
nullImportance=1/nrow(features1)

topFeatures<-features1%>%
  filter(Importance>nullImportance&
           Feature!="release_year")%>%
  pull(FeatureRenamed)
```

```
gg1<-ggplot(features1,
       aes(x=Importance,
           y=fct_reorder(FeatureRenamed,Importance),
           fill=FeatureCategory))+
  geom_bar(stat='identity')+
  geom_vline(xintercept=nullImportance,linetype=2)+
  scale_fill_manual(values=scico(4,palette='devon',categorical=F,end=0.88),
                    name='')+
  ylab('Trait')+xlab('Feature importance')+
  theme(legend.position = c(.8, .3))
```

```
## Warning: A numeric ‘legend.position‘ argument in ‘theme()‘ was deprecated in ggplot2
## 3.5.0.
## i Please use the ‘legend.position.inside‘ argument of ‘theme()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

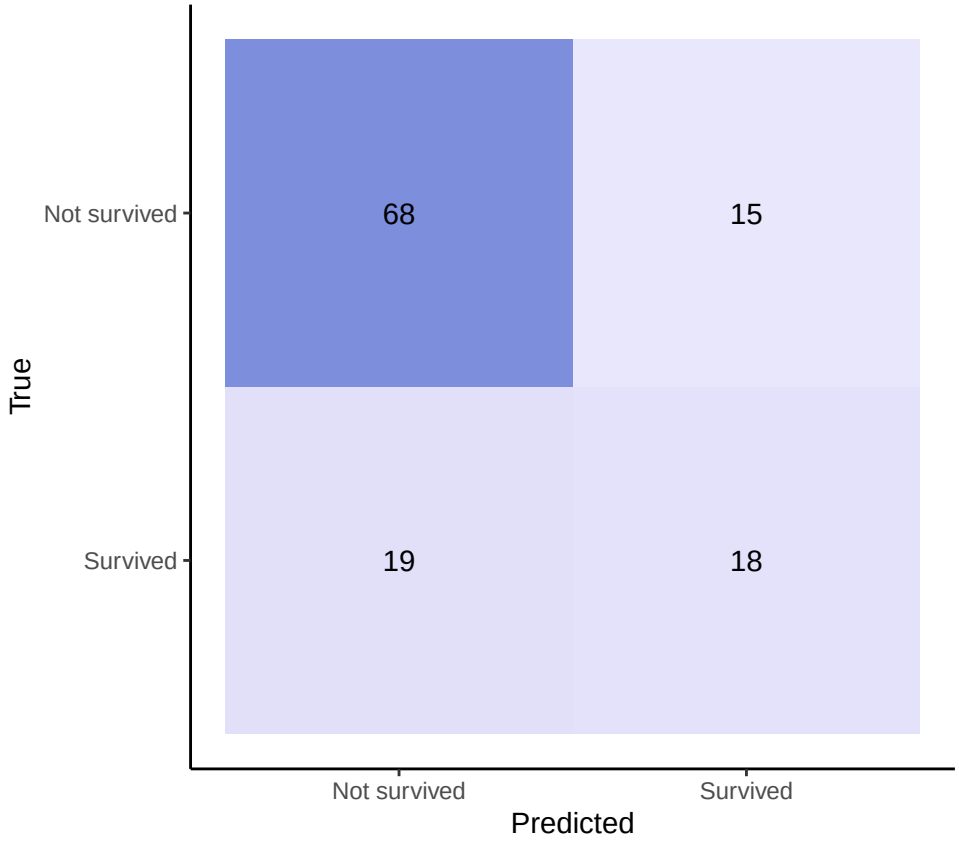
```
gg1
```




```
 #68 true negatives, 15 false positives, 19 false negatives, and 18 true positives

confusion1<-data.frame(TrueLabel=factor(c("Survived","Survived","Not survived","Not survived"),
                                        levels=c("Survived","Not survived")),
           PredictedLabel=factor(c("Survived","Not survived","Survived","Not survived"),
                                        levels=c("Not survived","Survived")),
           Count=c(18,19,15,68))

gg3<-ggplot(confusion1,aes(y=TrueLabel,x=PredictedLabel,fill=Count))+
  geom_tile()+
  geom_text(aes(label=Count))+
  scale_fill_scico(palette='devon',end=0.9,begin=0.5,guide='none',direction=-1)+
  coord_equal()+
  xlab("Predicted")+ylab("True")
gg3
```

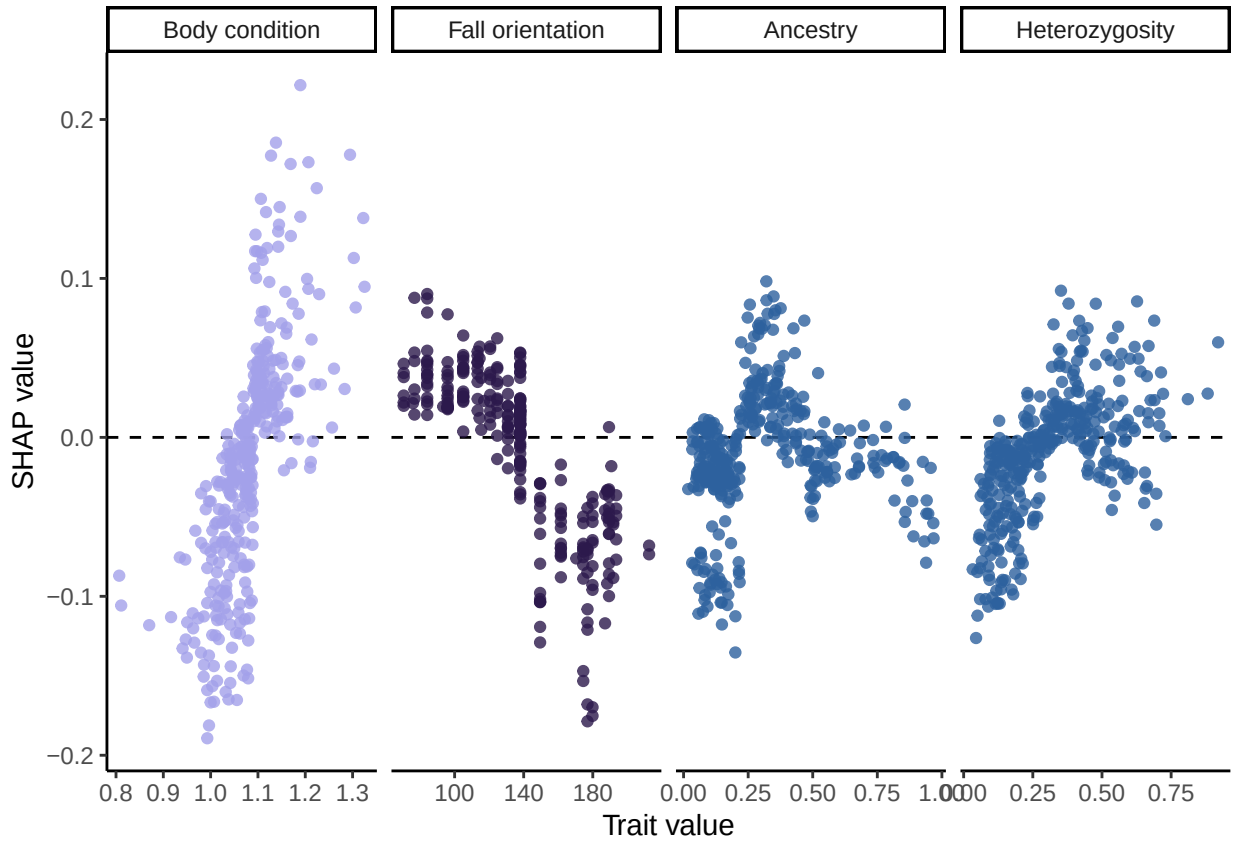



```
shap1topFeatures<-shap1%>%
  filter(FeatureRenamed%in%topFeatures)%>%
  mutate(FeatureRenamed=factor(FeatureRenamed,levels=topFeatures))

gg2<-ggplot(shap1topFeatures,
       aes(x=FeatureValue,y=SHAPValue,colour=FeatureCategory))+
  geom_hline(yintercept=0,linetype=2)+
  geom_point(alpha=0.8)+
  facet_grid(cols=vars(FeatureRenamed),scales='free')+
  scale_colour_manual(values=scico(4,palette='devon',categorical=F,end=0.88),
                      guide='none')+
  ylab('SHAP value')+xlab('Trait value')

gg2
```

```
## Warning: Removed 105 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- SHAP sample size

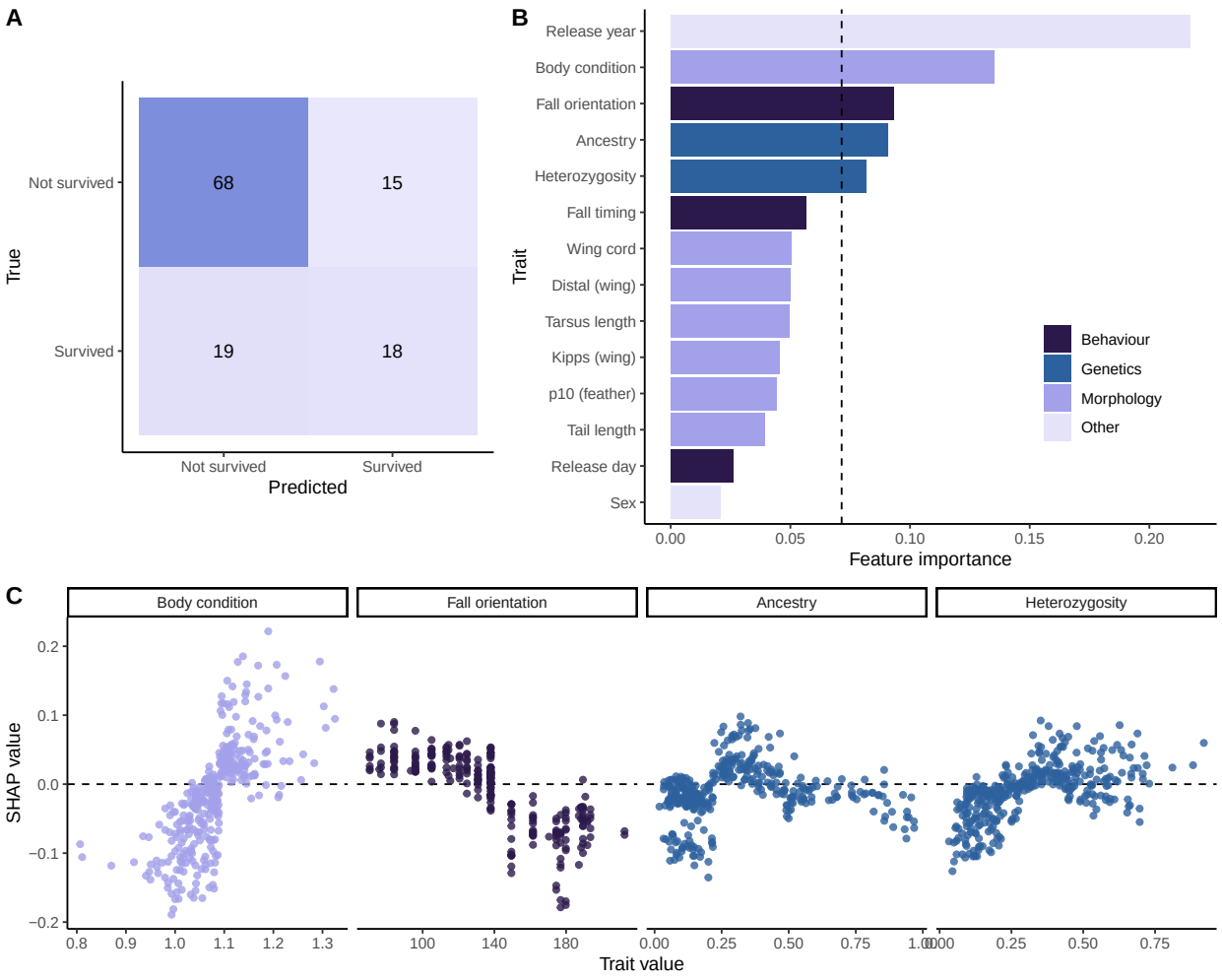
```
nrow(shap1)/length(unique(shap1$Feature))
```

```
## [1] 359
```

```
gg4=ggpubr::ggarrange(ggpubr::ggarrange(gg3,gg1,nrow=1,ncol=2,widths=c(0.7,1),labels=c("A","B")),
                      gg2,nrow=2,labels=c("","C"),heights=c(1,0.7))
```

```
## Warning: Removed 105 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

```
gg4
```




```
#ggsave("C:/Users/Steph/GitHub/thrush_hybrids/migratory_traits/Fig2.pdf",
#       plot=gg4,
#       height = 8,width=10,bg='white')
```