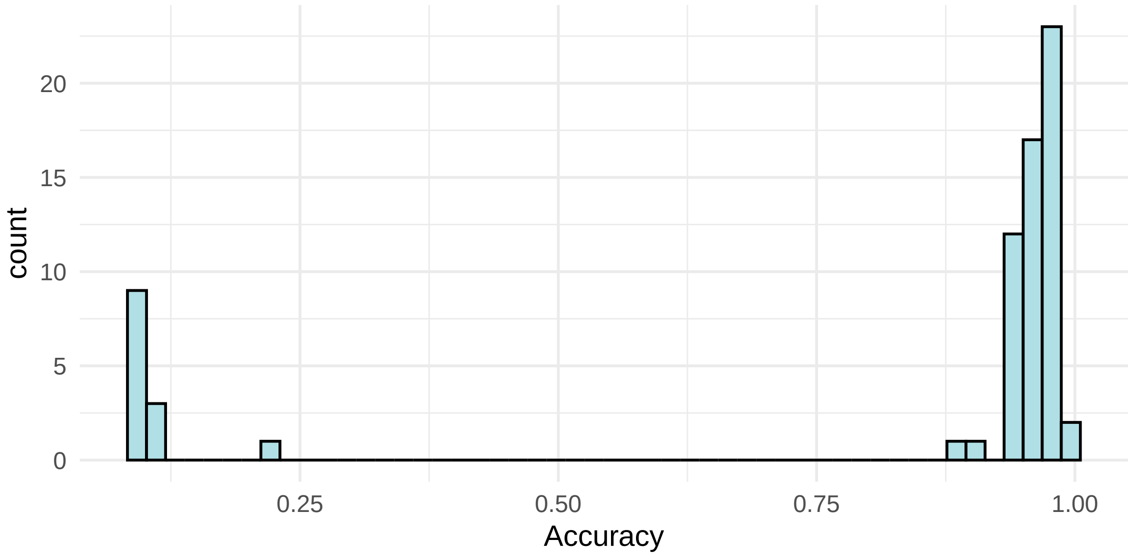
HW3 Results

Table of contents

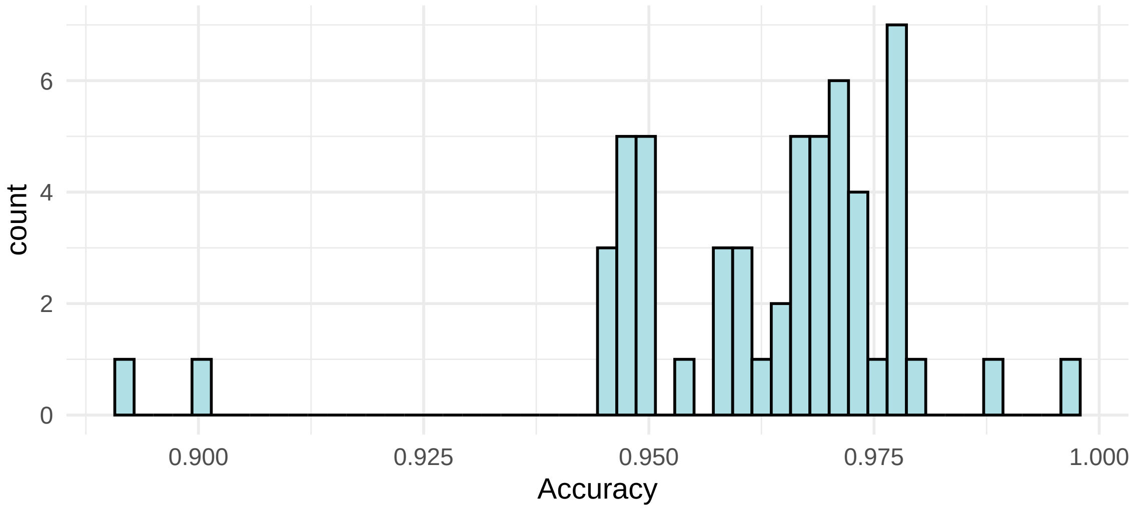
1 Digits

```
Error in readRDS(fnames[i]) : unknown input format
Error in readRDS(fnames[i]) : unknown input format
Error in readRDS(fnames[i]) : unknown input format
```

1.1 Distribution of Accuracies



Zoom in on upper

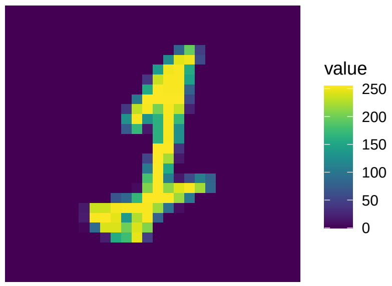


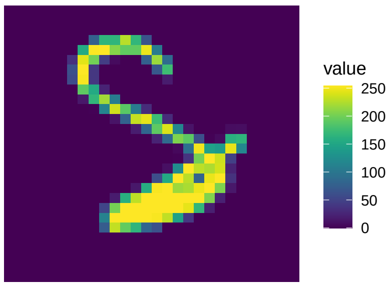
1.2 Top 20 Leaderboard (following all directions and no missing predictions)

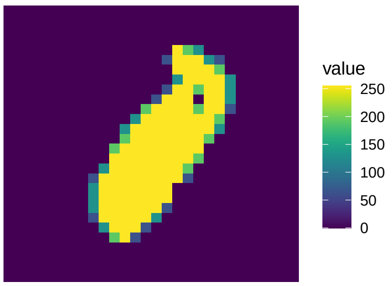
username	acc	dp	digits_notes
gubowen2	0.99736	0	
luoguangze	0.97843	0	
cheny0501	0.97829	0	
Tony-Xiayi-Ding	0.97793	0	
mkline1	0.97779	0	
kieranptodd	0.97744	0	
rirusso	0.97744	0	
Yiyannnnn	0.97451	0	
yyy1229	0.97379	0	
Jonajarro	0.97358	0	
vshao2000	0.97351	0	
xgulib	0.97351	0	
Leacavalli	0.97129	0	
maihantrinh	0.97129	0	
valeriad1610	0.97129	0	
bcardona0	0.97122	0	
Eva-Rumpler	0.97108	0	
joannakennedyharvard	0.97044	0	
newche	0.96929	0	
tzhang1hsph	0.96887	0	

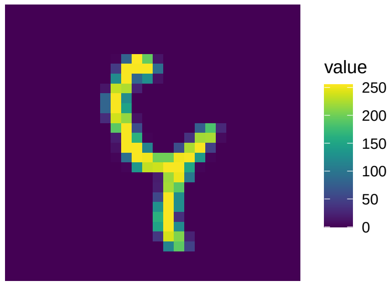
1.3 For fun

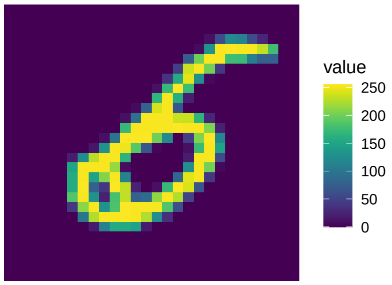
1.3.1 Digits that nobody got correct (only 5!!!)


Label: 2
Most common prediction: 1


Label: 8
Most common prediction: 5


Label: 0
Most common prediction: 1

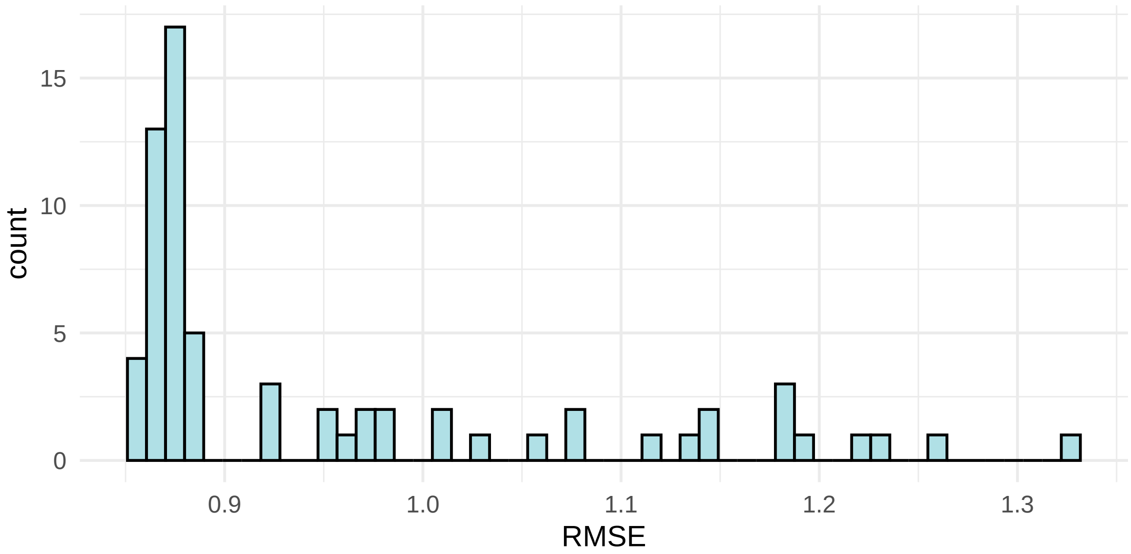

Label: 9
Most common prediction: 4


Label: 5
Most common prediction: 6

2 Movies

```
Error in readRDS(fnames[i]) : unknown input format
Error in readRDS(fnames[i]) : unknown input format
Error in readRDS(fnames[i]) : unknown input format
```

2.1 Distribution of RMSEs



2.2 Top 20 Leaderboard (following all directions and no missing predictions)

username	rmse	mp	movies_notes
Pratibha533	0.85781	0	
vsrip1	0.86780	0	
mkline1	0.86786	0	
davidolander	0.86789	0	
Yiyannnnn	0.86917	0	
carriecheng0924	0.87142	0	
bcardona0	0.87157	0	
Tony-Xiayi-Ding	0.87157	0	
xgulib	0.87157	0	
yyy1229	0.87567	0	
isabellayuxinliu	0.87827	0	
Jonajarro	0.87869	0	
MajedaAlzaydan	0.87869	0	
valeriad1610	0.87869	0	
ryan-hdez	0.87873	0	
MinyeZhou429	0.87881	0	
tzhang1hsph	0.87881	0	
zoe-love	0.87881	0	
Tianxiuli	0.88181	0	
mian3322	0.88360	0	

3 Code

```r
library(dplyr)
library(ggplot2)
library(stringr)
library(viridis)
library(data.table)
library(kableExtra)
library(grid)
library(gridExtra)
studentnames <- fread('~/Documents/BST260/StudentNames.csv')
# real values
digits <- readRDS('digits_ta.RDS')
y <- digits$test$labels
fnames <- list.files(path = '~/Documents/BST260/bst260_hw',
                     pattern = 'digit_predictions',
                     full.names = T, recursive = T)
usernames <- gsub('-2022HW3', '', str_extract(fnames, '([a-z]|[A-Z]|[0-9]|-)+-2022HW3'))
predictions <- matrix(data = NA, nrow = length(fnames), ncol = length(y))
predictions_binary <- matrix(data = NA, nrow = length(fnames), ncol = length(y))
results_digits <- data.frame(username = usernames, acc = NA, dp = 0, digits_notes = '')
for (i in seq_along(fnames)) {
  y_hat <- try(readRDS(fnames[i]))
  if (is(y_hat, 'try-error')) {
    results_digits$digits_notes[i] <- 'your file was not properly saved as an .rds'
    results_digits$dp[i] <- results_digits$dp[i]-5
    test <- try(load(fnames[i]))
    if (is(test, 'try-error')) {
      next
    }
    y_hat <- digit_predictions
  }
  if (length(class(y_hat))==2) {
    results_digits$digits_notes[i] <- paste(
      results_digits$digits_notes[i],
      '; you stored your predictions incorrectly as a matrix -
      we wanted the actual predicted value.
      assuming your matrix is a matrix of probabilities'
    )
    results_digits$dp[i] <- results_digits$dp[i]-5
    y_hat <- as.factor(c(0:9)[apply(y_hat,1,which.max)])
  } else if (length(class(y_hat))==1) {
    if (class(y_hat)!='factor') {
      results_digits$digits_notes[i] <- paste(
        results_digits$digits_notes[i],
        '; your predictions were not saved as a vector of factors'
      )
      results_digits$dp[i] <- results_digits$dp[i]-5
      if (class(y_hat)=='data.frame') {
        y_hat <- y_hat[,1]
      }
      if (class(y_hat)=='array' & is.numeric(y_hat)) {
        y_hat <- as.factor(as.character(y_hat))
      }
      if (class(y_hat)=='character') {
        y_hat <- as.factor(y_hat)
      }
    }
  } else if (length(class(y_hat))==3) {
    y_hat <- y_hat %>% pull() %>% as.character() %>% as.factor()
  }
  if (any(is.na(y_hat))) {
    results_digits$digits_notes[i] <- paste(
      results_digits$digits_notes[i],
      '; some values of your prediction are NA'
    )
    results_digits$dp[i] <- results_digits$dp[i]-3
  }
  acc <- mean(y==y_hat)
  results_digits$acc[i] <- acc
  if (length(y_hat)!=length(y)) {
    results_digits$digits_notes[i] <- paste(
      results_digits$digits_notes[i],
      '; your predictions are length', length(y_hat),
      ', which is not the length ', length(y), 'of the test set'
    )
    results_digits$dp[i] <- results_digits$dp[i]-10
    next
  }
  predictions[i,] <- as.numeric(as.character(y_hat))
  predictions_binary[i,] <- y_hat==y
}
```

```
results_digits |>
  ggplot(aes(x = acc)) +
  geom_histogram(bins=50, color = 'black', fill = 'powderblue') +
  theme_minimal() +
  xlab('Accuracy')
results_digits |>
  filter(acc > 0.75) |>
  ggplot(aes(x = acc)) +
  geom_histogram(bins=50, color = 'black', fill = 'powderblue') +
  theme_minimal() +
  xlab('Accuracy') +
  ggtitle('Zoom in on upper')
results_digits |>
  filter(digits_notes == '') |>
  arrange(desc(acc)) |>
  dplyr::slice(1:20) |>
  knitr::kable(format = 'latex', digits = 5) |>
  column_spec(4, width = "7cm")
no_correct <- which(colSums(predictions_binary, na.rm=T)==0)
plotlist <- list()
for (i in no_correct) {
  real_label <- digits$test$labels[i]
  most_common_prediction <- summary(as.factor(as.character(predictions[,i])))
  most_common_prediction <- names(most_common_prediction)[which(most_common_prediction==ma
  p <- grob(digits$test$images[i,] |>
    matrix(nrow=28, ncol=28) |>
    reshape2::melt() |>
    ggplot(aes(x = Var1, y = Var2)) +
    geom_raster(aes(fill = value)) +
    scale_fill_viridis() +
    scale_y_reverse() +
    theme_void() +
    xlab('') +
    ylab('') +
    ggtitle(paste('Label:', real_label, '\nMost common prediction:', most_common_predictio
  plotlist[as.character(i)] <- p
}
grid.arrange(grobs = plotlist, nrow = 2, ncol = 3)

# real values
y <- readRDS('mv_ta.RDS')$test$rating
```

```
fnames <- list.files(path = '~/Documents/BST260/bst260_hw', pattern = '(rating|mv|movie|mo
                     full.names = T, recursive = T)
usernames <- gsub('-2022HW3', '', str_extract(fnames, '([a-z]|[A-Z]|[0-9]|-)+-2022HW3'))
predictions <- matrix(data = NA, nrow = length(fnames), ncol = length(y))
results_movies <- data.frame(username = usernames, rmse = NA, mp = 0, movies_notes = '')
for (i in seq_along(fnames)) {
  y_hat <- try(readRDS(fnames[i]))
  if (is(y_hat, 'try-error')) {
    results_movies$movies_notes[i] <- 'your file was not properly saved as an .rds'
    results_movies$mp[i] <- results_movies$mp[i]-5
    test <- try(load(fnames[i]))
    if (is(test, 'try-error')) {
      next
    }
    y_hat <- rating_predictions
  }
  if (is.factor(y_hat)) {
    results_movies$movies_notes[i] <- paste(
      results_movies$movies_notes[i],
      '; you saved your predictions as a factor when they should have been numeric'
    )
    results_movies$mp[i] <- results_movies$mp[i]-5
    y_hat <- as.numeric(as.character(y_hat))
  }
  if (length(class(y_hat))==3 | length(class(y_hat))==4) {
    results_movies$movies_notes[i] <- paste(
      results_movies$movies_notes[i],
      '; you saved your whole data frame instead of a vector of numeric values'
    )
    results_movies$mp[i] <- results_movies$mp[i]-5
    # for people who put it in another column.... >:(........
    if (!any(!is.na(y_hat$rating))) {
      results_movies$movies_notes[i] <- paste(
        results_movies$movies_notes[i],
        '; your ratings column is all NA, which means it was difficult to find the correct
      )
      results_movies$mp[i] <- results_movies$mp[i]-5
      if ('rating_predictions'%in%colnames(y_hat)) {
        y_hat <- y_hat$rating_predictions
      } else if ('pred' %in% colnames(y_hat)) {
        y_hat <- y_hat$pred
      }
    } else {
      y_hat <- y_hat$rating
    }
  } else if (length(class(y_hat))==1 & class(y_hat)=='data.frame') {
    y_hat <- y_hat[,1]
  }
  if (any(is.na(y_hat))) {
    results_movies$movies_notes[i] <- paste(
      results_movies$movies_notes[i],
      '; some values of your prediction are NA'
    )
    results_movies$mp[i] <- results_movies$mp[i]-3
  }
  rmse <- sqrt(mean((y-y_hat)**2, na.rm=T))
  results_movies$rmse[i] <- rmse
  if (length(y_hat)!=length(y)) {
    results_movies$movies_notes[i] <- paste(
      results_movies$movies_notes[i],
      '; your predictions did not include all values of the test set'
    )
    results_movies$mp[i] <- results_movies$mp[i]-10
    next
  }
  predictions[i,] <- y_hat
}
results_movies |>
  ggplot(aes(x = rmse)) +
  geom_histogram(bins=50, color = 'black', fill = 'powderblue') +
  theme_minimal() +
  xlab('RMSE')
results_movies |>
  arrange((rmse)) |>
  filter(movies_notes == '') |>
  dplyr::slice(1:20) |>
  knitr::kable(format = 'latex', digits = 5) |>
  column_spec(4, width = "8cm")
```